High-Throughput Atomistic Modeling of Biomolecular Structure and Association

Zur Erlangung des akademischen Grades eines

DOKTORS DER NATURWISSENSCHAFTEN

von der Fakultät für Physik des
Karlsruher Institut für Technologie (KIT)

genehmigte

DISSERTATION

von

Dipl.-Phys. Timo Strunk
aus Herne

Tag der mündlichen Prüfung: 02.11.2012
Referent: Prof. Dr. Wolfgang Wenzel
Korreferent: Prof. Dr. G. Ulrich Nienhaus

Deutsche Zusammenfassung

Proteine regeln viele biologische Prozesse des menschlichen Körpers. Neben unzähligen weiteren Prozessen nehmen sie an der Katalyse von biochemischen Reaktionen teil, regulieren Ionenkonzentrationen oder vermitteln die Immunantwort. Mutationen und Fehlfaltungen von Proteinen werden häufig in Zusammenhang mit Krankheiten, wie zum Beispiel Alzheimer oder Krebs gebracht. Die meisten Proteine falten eigenständig in ihre eindeutige, funktionale, dreidimensionale Struktur, welche durch die Abfolge der Aminosäuren, der Bausteine der Proteine, bestimmt ist. Obwohl die Proteinsequenz häufig durch Analyse der DNA zugänglich ist, existieren bei weitem nicht für alle medizinisch und biologisch relevanten Proteine 3D Strukturen ihrer gefalteten dreidimensionalen Struktur. Die Diskrepanz zwischen der Anzahl von Millionen von bekannten Proteinsequenzen und den etwa 80.000 bekannten, experimentell aufgelösten Proteinstrukturen, zeigt die Schwierigkeiten, welche mit der experimentellen Bestimmung einer Proteinstruktur verbunden sind.

Methoden zur theoretischen Vorhersage von Proteinstrukturen generieren häufig durchwachsene Resultate. Proteine mit mindestens 40% Sequenzübereinstimmung zu einer bekannten Proteinstruktur können zwar häufig mit angemessener Qualität vorhergesagt werden, für Proteine ohne oder mit nur sehr geringer Sequenzähnlichkeit gibt es jedoch noch keine zuverlässige Methode zur Proteinstrukturvorhersage. Die durchwachsene Qualität vieler Proteinmodelle spiegelt sich in der geringen Akzeptanz vieler vorhergesagter Strukturen in den biologischen und medizinischen Wissenschaften wider und führte zur kompletten Entfernung aller theoretischer Vorhersagen aus der Datenbank der bekannten Proteinstrukturen. Gerade in den Lebenswissenschaften würde ein fundiertes Proteinmodell einen enormen Erkenntnisgewinn bei der Analyse von Krankheitsbildern oder allgemeinen biologischen Prozessen darstellen.

Die Forschung in den Lebenswissenschaften kommt immer mehr zu der Erkenntnis, dass Proteine nicht als statische Objekte sondern als dynamische Maschinen verstanden werden müssen. Da die Aufklärung von Proteindynamik experimentell noch wesentlich aufwändiger ist, als die Bestimmung der Struktur, ergibt sich hier eine Chance für Simulationsverfahren die experimentellen Untersuchungen zu ergänzen. Hierzu werden überwiegend Verfahren der Molekulardynamik eingesetzt, die jedoch aus fundamentalen Gründen in der Länge der behandelbaren Zeitskala und damit der Komplexität der zu Grunde liegenden Prozesse beschränkt sind.

Vor kurzem konnte mit der Methode der Molekulardynamik der Faltungsprozeß eines aus 56 Aminosäuren bestehenden Proteins über $1\,ms$ lang simuliert werden. Diese Simulationen wurden auf einer kostspieligen, spezialisierten Computerarchitektur durchgeführt, welche bislang nur einer einzigen Forschergruppe zur Verfügung steht. Aufgrund von atomaren Schwingungen im Femtosekundenbereich benötigt eine Molekulardynamiksimulationen einen Integrationsschritt in der gleichen Größenordnung. Der hohe Aufwand einer Molekulardynamiksimulation

begründet sich daher in der großen Diskrepanz der Zeitskalen: Für die Simulation eines Proteins über eine Millisekunde werden 10^{12} Simulationsschritte benötigt.

Monte-Carlo Algorithmen charakterisieren das thermodynamische Ensemble einer Proteinstruktur, ohne eine direkte Simulation der Kinetik zu benötigen. Da sehr schnelle Bewegungen des Systems dadurch innerhalb der Simulation eliminiert werden können, erlauben Monte-Carlo Simulationen die Charakterisierung von großen Konformationsübergängen oder Selbstassemblierung; Prozessen, die typischerweise auf langen Zeitskalen ablaufen.

Bisherige Monte-Carlo Strategien, welche in unserer Gruppe implementiert wurden, erlaubten die Simulation von Proteinfaltungsprozessen, die Vorhersage von Protein-Liganden Komplexen und die Simulation der Morphologie vieler, für die Materialwissenschaften relevanter, Materialien. Bis heute gibt es jedoch kein allgemein anwendbares, effizientes Monte-Carlo Simulationspaket.

Mein Beitrag zur Entwicklung von Monte-Carlo basierten Systemen umfasste:

- Die Implementierung Monte-Carlo basierter Algorithmen zur Simulation von Proteinen, dem Protein-Ligandendocking und allgemeiner Nanosysteme in einer Vielzahl von Kraftfeldern.
- Die Implementierung und Verifikation einer Methode zur absoluten Qualitätskontrolle von Proteinmodellen.
- Die Vorhersage von Proteinstrukturen und Komplexen in Übereinstimmung mit experimentellen Ergebnissen.
- Die Simulation nanoskaliger Morphologien neuartiger Materialien.

Ein Großteil dieser Dissertation umfasste die Entwicklung des Monte-Carlo basierten Simulationspaketes SIMONA: **SI**mulation of **MO**lecular and **NA**noscale systems. Das Programmpaket ist Teil dieser Dissertation und erhältlich auf *http://www.int.kit.edu/nanosim/*.

Absolute Qualitätskontrolle von Proteinmodellen

Biologische Prozesse auf der Nanoskala können häufig durch die Kenntnis von den am Prozess beteiligten Proteinstrukturen erklärt werden. Obwohl nur eine begrenzte Menge an experimentell aufgelösten Proteinstrukturen von biologischer Relevanz bekannt ist, finden theoretisch vorhergesagte Proteinstrukturen nur selten Akzeptanz in den biologischen und medizinischen Wissenschaften. In vielen Fällen kann für Proteinsequenzen jedoch eine korrekte Proteinstruktur auf Basis der Homologiemodellierung vorhergesagt werden. Selbst in Abwesenheit von Homologie zu einer bekannten, experimentell aufgelösten, Proteinstruktur ist es in Einzelfällen möglich eine korrekte Proteinstruktur nahe der nativen Proteinstruktur zu generieren. Jedoch ist es gerade für Proteine, bei denen keine, oder nur wenig, Homologie zu einer experimentell aufgelösten Proteinstruktur besteht, nicht möglich Aussagen über die Qualität eines Proteinmodells zu treffen. Dies gilt in besonderem Maße für Proteinstrukturen, welche von vollautomatischen Strukturvorhersageservern generiert wurden. Die Entwicklung einer Methode für die Qualitätskontrolle

von solchen Proteinmodellen könnte die Akzeptanz theoretisch vorhergesagter Proteinmodelle in den Lebenswissenschaften drastisch verbessern.
In dieser Arbeit entwickelten wir daher einen statistischen Test zur a-priori Bestimmung der Qualität von Proteinmodellen. Da Proteine in ihrer funktionalen Form nur marginal stabil sind, vermuteten wir, dass auch einzelne Aminosäuren einen optimalen Beitrag zur Energie der globalen Proteinstruktur in ihrer aktiven, biologischen Konformation leisten. Daher erfassten wir Statistiken für diese Energiebeiträge für einen Satz hochaufgelöster experimenteller Proteinstrukturen und entwickelten einen N-dimensionalen statistischen Test, welcher die Qualität eines Modells durch Vergleich mit den erfassten Statistiken überprüft. Die Energiestatistiken von Aminosäuren im gefalteten Zustand unterschieden sich, wie vermutet, von den Energien der Aminosäuren von Proteinmodellen in vorhergesagten Strukturbibliotheken. Durch die Anwendung des statistischen Tests auf Aminosäuretriplets, war es uns möglich die Spezifizität des Tests zu erhöhen, so dass die Proteinstrukturen schlechter Qualität für insgesamt 93% der getesteten 160 Proteinmodelle korrekt identifiziert werden konnten. Die verbleibenden 7%, welche nicht korrekt identifiziert wurden, waren entweder Oligomere, nicht globulare Proteine, oder an Kofaktoren gebunden, welche in unserem Kraftfeld nicht berücksichtigt wurden. Mit Hilfe von Methoden der Bioinformatik lässt sich allerdings vor Durchführung unseres Tests und nur mit Kenntnis der Proteinsequenz bestimmen, ob ein Protein diesen Proteinklassen angehört und zuverlässig bewertet werden kann. Durch Kombination beider Techniken hoffen wir, dass unser Test als ein Prototyp für die Entwicklung weiterer statistischer Tests zur Qualitätskontrolle von Proteinmodellen dient und die Akzeptanz von Proteinstrukturvorhersagen für positiv evaluierte Proteinmodelle steigert.

Genetische Modifikation eines Hydrophobins zur Veredelung von Implantaten

Durch die allgemein steigenden Lebenserwartungen von Implantatträgern, welche ihr erstes Implantat häufig bereits im Alter von unter 50 Jahren bekommen, ist es immer häufiger notwendig, Implantate in hohem Alter auszutauschen. Für viele ältere Patienten bedeutet dies ein hohes Risiko. Es ist daher wichtig, Implantatmaterialien zu entwickeln, welche länger als zwei bis drei Jahrzehnte im Körper verweilen können. Eine Möglichkeit zur Entwicklung haltbarer Implantate ist die Entwicklung von biokompatiblen Oberflächenbeschichtungen, welche eine Anhaftung von neuem Zellmaterial, idealerweise Knochenstammzellen, ermöglicht, ohne eine Ausbildung von Biofilmen zu fördern.
In einer Kollaboration zwischen zwei Arbeitsgruppen am KIT und dem Universitätsklinikum Heidelberg entwickelten wir ein genetisch modifiziertes Protein zur biokompatiblen Implantatbeschichtung. Hydrophobine sind pilzstämmige Proteine mit interessanten physikalisch-chemischen Eigenschaften, da sie eine hydrophile Oberfläche durch Hydrophobinbeschichtung in eine hydrophobe Oberfläche umwandeln können, welche somit weniger anfällig für die Ausbildung von Biofilmen ist. Pilotstudien zeigten allerdings, dass Hydrophobine keine ausgeprägte Zellbindung ermöglichen. Wir entwickelten daher eine Strategie zur genetischen Modifikation von Hydrophobinen, welche eine Zellbindung ermöglicht, ohne die Hydrophobizität der Hydrophobine zu beeinträchtigen. Vor der Durchführung unseres Projekts existierte keine experimentelle Proteinstruktur des untersuchten Hydrophobinkomplexes. Mittels Proteinstrukturvorhersagemethoden, entwickelten wir daher ein atomistisches Modell des Hydrophobins

und identifizierten eine Wasser-zugewandte Stelle, welche sich für eine genetische Modifikation eignete. Durch Aufnahme der zellbindenden Sequenzmotive RGD und LG3 an besagter Stelle, konnten wir die Bindung von Zellen auf Hydrophobinschichten immens verbessern, ohne dabei eine Biofilmbildung zu begünstigen.

Strukturelles Modell zur Erklärung der Ausbildung von Gasvesikeln

Seit der Entdeckung von Krankheiten, welche in Zusammenhang mit Proteinaggregation stehen, wie zum Beispiel die Formierung von Amyloid Strukturen bei Alzheimer Patienten, ist die Erforschung von aggregierten Proteinstrukturen ein zentrales wissenschaftliches Thema. Strukturelle Informationen über aggregierte Proteinzustände sind nur schwer experimentell zu charakterisieren, da unstrukturierte Proteinkomplexe für Kristallographietechniken ungeeignet, und viele Proteinaggregate zu groß für die Studie mit NMR Methoden sind. Ein interessantes Beispiel sind Gasvesikel, welche den Auftrieb vieler wasserstämmiger Bakterien regulieren. Mit zwei Kollaborationspartnern der Universität Darmstadt entwickelten wir Modelle für Proteinaggregate, die in der Ausbildung von Gasvesikeln in Bakterien wichtig sind. Für das innerhalb der Gasvesikelwand befindliche Protein GvpA gibt es bislang keine experimentelle Proteinstruktur. Durch de-novo Modellierung erstellten wir ein Proteinmodell, welches in den experimentell arbeitenden Gruppen bestätigt werden konnte. Mittels dieses Modells konnten wir den strukturellen Grund zur Ausbildung von Proteinaggregaten durch das Protein GvpA erklären und zeigen, wie Gas durch eine hydrophobe Oberfläche innerhalb des Vesikels eingeschlossen wird.
Die durch in-silico Protein-Protein Docking erstellte Struktur ist sowohl mit früheren solid-state NMR Ergebnissen, als auch mit Mutageneseexperimenten kompatibel, welche spezielle Kontaktstellen innerhalb der Aggregatstruktur untersuchten.

Hochdurchsatzvorhersage unbekannter Peptidstrukturen

Seit 1970 wurden nur drei neue Klassen Antibiotika für die klinische Nutzung zugelassen. Antimikrobielle, antifungale und antibiotisch wirkende Peptide werden daher als Hoffnungsträger der Medizin gesehen, um der immer weiter fortschreitenden Immunität einiger Bakterienstämme gegen bekannte Antibiotika entgegenzuwirken. Experimentelle Hochdurchsatzscreenings zur Überprüfung der Wirksamkeit dieser Peptide sind sehr aufwändig und ressourcenintensiv, da sie die Synthese vieler verschiedener Peptidsequenzen erfordern. Die 3D-Struktur einer Peptidsequenz könnte zur Entwicklung von Struktur-Funktions-Modellen genutzt werden und so eine verbesserte Möglichkeit zur Identifikation funktionaler Peptide darstellen.
Homologie- und andere wissensbasierte Methoden schlagen bei der Peptidstrukturvorhersage häufig fehl, da Peptidsequenzen zu kurz sind, um eine aussagekräftige Homologie dafür abzuleiten. Im Vergleich zu größeren Proteinen führen Punktmutationen in Peptiden häufiger zu großen Änderungen in der Tertiärstruktur. In diesem Projekt implementierten wir daher ein Protokoll zur de-novo Peptidstrukturvorhersage auf POEM@HOME, dem von mir entwickelten, weltweit verteilten, Rechnernetz für Proteinsimulation, das eine Hochdurchsatzvorhersage von Peptidstrukturen ermöglicht. Wir verifizierten das Protokoll durch die Vorhersage der Strukturen von vier experimentell aufgelösten Peptiden. Weiterhin untersuchten wir die niederenergetischen Strukturcluster der Energielandschaft der Peptide und erklärten damit eine Scherungsbewegung,

mit der ein β-Peptid zwischen seinen niederenergetischen Zuständen wechseln kann.

In-silico Untersuchung von Interaktionshotspots in Protein-Protein Interfaces

Eine Vielzahl biologischer Signalprozesse wird durch die Assoziation von Proteinen in Protein-Protein Komplexen vermittelt. Das Verständnis dieser Interaktionen kann daher Einblick in diese Prozesse geben und möglicherweise ihre Manipulation ermöglichen. Protein-Protein Interfaces sind daher vielversprechende Ziele für die Entwicklung neuer Medikamente. Im Vergleich zu Bindungstaschen, welche zur Medikamentenentwicklung mit niedermolekularen Liganden inhibiert werden, sind die Bindungsflächen von Protein-Protein Komplexen weit ausgedehnt. Eine Inhibierung der Protein-Protein Bindung konnte mit mehreren experimentellen Methoden, wie zum Beispiel der Bindung mit Antikörpern, erreicht werden. Eine Inhibierung mit kleinen Molekülen ist aufgrund von besserer Bioverfügbarkeit und geringerem Preis jedoch wünschenswerter. Hierfür muss jedoch bekannt sein, welche Aminosäuren auf der Bindungsfläche den größten Einfluss auf die Bindung haben, damit ein niedermolekularer Ligand für genau diese Bindungsstelle generiert werden kann. Zur Bestimmung dieser sogenannten Hotspots kann experimentell ein Alanin Mutagenese Screening durchgeführt werden, in welchem jede Aminosäure in der Nähe der Bindungstasche durch Alanin ersetzt wird. Wir haben den experimentellen Aufbau in-silico nachempfunden und verifizierten ihn anhand von zwei experimentell charakterisierten Systemen. Dieses Projekt wurde in Zusammenhang mit der Young Investigator Group von Dr. Katja Schmitz (ehemals KIT) durchgeführt, welche Bindungshotspots am System der Chemokine analysierte.

De-novo Protein-Protein und Protein-Liganden Docking

In-silico drug-design untersucht Liganden aus umfassenden Liganden-Datenbanken und dockt sie in bekannte Bindungstaschen von pharmazeutisch relevanten Proteinstrukturen. In SIMONA wurden ebenfalls Parametrisierungen für Liganden implementiert. Diese testeten wir anhand von Benchmarksystemen und erhielten in allen Fällen die korrekte Dockingpose. Ein identisches Dockingprotokoll führten wir auch für die in der vorherigen Sektion untersuchten Protein-Protein Komplexe durch. Wiederum reproduzierten wir die Benchmarkstrukturen innerhalb der experimentellen Genauigkeit.
Da sich die Benchmarksimulationen nicht für Hochdurchsatzscreenings von Protein-Liganden Interaktionen eignen, entwickelten wir ein hierarchisches Dockingprotokoll, welches mit einem Bruchteil des Aufwandes auskommt und verifizierten es an weiteren sechs experimentell bekannten Protein-Liganden Systemen. Diese Methode kann nun zum effizienten in-silico Protein-Ligandenscreening verwendet werden.

Morphologiesimulationen von Pentacenclustern

Die in der vorherigen Studie implementierten Kraftfelder für allgemeine organische Moleküle wurden in diesem Projekt zur Simulation der Selbstorganisation von Pentacenclustern verwendet. Pentacen ist ein vielversprechendes Material zur Verwendung in der emittierenden Schicht von OLED Displays. Bevor jedoch elektronische Betrachtungen dieser Schicht durchgeführt

werden können, muss eine entsprechende amorphe Morphologie der Schicht erstellt werden. Die simulierten Pentacencluster nukleiren zu Strukturen, wie sie auch der experimentell bestimmte native Pentacenkristall aufweist. Die teilweise auch amorphen Strukturen sind jedoch in der Größe limitiert. Wir stellen daher abschließend eine Methode vor, welche das Verschmelzen zweier Pentacencluster stark beschleunigt und so Simulationen von nahezu unbegrenzten Clustergrößen ermöglicht.

Sortierung von Kohlenstoffnanoröhren durch spezifische Polymerwicklung

Aufgrund ihrer vielfältigen Eigenschaften sind Kohlenstoffnanoröhren ein äußerst vielversprechendes Material für zahlreiche Anwendungen der Elektronik und Mechanik. Weil die intrinsischen Eigenschaften der Nanoröhre von ihren Chiralitätsindizes (n,m) abhängen, müssen diese für den gezielten industriellen Einsatz nach ihrer Produktion erst getrennt und dann sortiert werden. Bestimmte Polymere mit Fluorengruppen binden präferentiell an Nanoröhren mit definierter Chiralität. Wir konnten diese Präferenz in all-atom Monte-Carlo Simulationen sowohl für Polymere mit Fluorengruppen als auch für eine weitere Polymerart mit Carbazolgruppen bei zwei Nanoröhren unterschiedlicher Chiralität nachweisen. Die so erlangten Erkenntnisse decken sich mit den experimentellen Dispersionsspektren und klären die dominanten Bindungsmodii zwischen Nanoröhre und Polymer auf. Sie stellen den ersten Schritt zum zielgerichteten Design neuer spezifischer Polymere zur Dispersion anderer Chiralitätskonfigurationen dar.

Ich bedanke mich herzlich bei der Carl-Zeiss Stiftung, die dieses Dissertationsprojekt über den Zeitraum von 3 Jahren ermöglichte.

Contents

1. Introduction

1.1. Overview

Proteins constitute the nano-scale machinery carrying out most of the biological processes in the human body. Among many other functions, they catalyze biochemical reactions, exert mechanical force or possess a structural function by providing the building blocks of cells. Sequencing techniques were developed to obtain the primary amino acid sequences of many proteins: the human genome project, which mapped most of the human DNA, in particular the protein-coding sequences, has been completed in 2003[1]. However, knowledge of the genetic code yields all but a glimpse at biological function. Most proteins spontaneously assume a unique structure, determined by their amino acid sequence, in which they function. Where available, the medical and biological sciences benefit greatly from such structural information, as it often offers key insights into biological processes in general and diseases in particular. Unfortunately, to date, three-dimensional structures are available only for a fraction of the interesting proteins. This is particularly true for some pharmaceutically relevant protein classes, such as transmembrane proteins. Often the experimental determination of the 3D-structure of these proteins is either a very difficult and resource-intensive task or even not amenable to presently available techniques.

Modeling methods increasingly attempt to predict the structure of proteins, for which no experimental structures exist, but the results of these methods have been mixed. For proteins with high sequence similarity to a structurally resolved protein, adequate models can be generated. Occasionally modeling even succeeds in the absence of a good template[2], but none of the presently available methods can decide with certainty, whether a proposed model or an experimentally determined structure is correct. This problem hinders the acceptance of theoretical protein models in the life-sciences. In 2006 the database of all known protein models (RCSB)[3] was pruned of all previously deposited theoretical predictions[4].

Prior to the deposition of an experimentally resolved protein structure in the RCSB database, quality control software is applied to the protein structure, as even experimental structures can feature local unphysical conformations, such as steric clashes, missing atoms or unphysical sidechain conformers[5]. A similar local quality control method for theoretical protein models does not help to identify low quality models, which have an incorrect global fold, but seem to be correct locally. The development of a method for absolute quality control of protein models, one of the tasks presented in this thesis, would significantly contribute to increase the acceptance of theoretical protein models in the life science research.

Presently Molecular Dynamics simulations dominate modeling attempts in the life and materials science. The high computational cost incurred in these simulations stems from the small timescales, which need to be simulated due to atomic vibrations; a fundamental problem which severely limits the timescales that can be realistically modeled by these techniques. Recently Molecular Dynamics algorithms were optimized to work on a highly parallel architecture to fold multiple small proteins repeatedly[6]. These simulations, which spanned more than $1\,ms$ folding time, took multiple months on a very expensive high performance computer available to only a single research group.

The timescale problem is not specific to the simulation of proteins, but to many systems, which form their nano-scale structure in complex processes, such as polymer glass transitions[7]. Considering this situation it is surprising, that Monte-Carlo techniques, which historically preceded Molecular Dynamics simulations in the material sciences, have received considerably less attention in the last decades. Monte-Carlo algorithms permit computation of thermodynamic expectation values by generating the thermodynamically relevant ensemble of conformations and do not exhibit the timescale problem. Previous Monte-Carlo investigations in our group demonstrated their applicability for several different problems, including the protein folding process[8, 9], protein-ligand docking[10] and morphology simulations of nanostructures in the material sciences[11]. Unfortunately, in contrast to Molecular Dynamics, there is no code-basis for adaptable and efficient Monte-Carlo simulations readily available.

The overwhelming fraction of the method development, required for the investigations reported in this thesis, went into the development of a novel Monte-Carlo based simulation package, which we named SIMONA: SImulation of MOlecular and NAnoscale systems, downloadable from *http://www.int.kit.edu/nanosim/*.

In the work presented here, I contributed to the development of Monte-Carlo based simulation for nano-scale systems by:

- Implementing Monte-Carlo algorithms using a variety of forcefields for proteins, protein-ligand docking and general nanosystems.
- Implementing and testing a method for absolute quality control of protein models.
- Predicting protein structures and complexes and verifying the results experimentally.
- Simulating the nano-scale morphologies of novel materials.

I would like to express my gratitude to the Carl-Zeiss Stiftung, which funded my Ph.D project over the course of three years.

1.2. Outline

I structured the thesis into two method and four application chapters:

In chapter 2, I introduce the structural basis for protein simulations, which included applications to protein folding (section 4 and section 5), protein-protein interaction and protein-ligand docking (section 6). In addition to biomolecular systems I have implemented forcefields and simulation protocols for nanomaterials, including carbon nanotubes and pentacene (section 7).

In chapter 3, I discuss Monte-Carlo methods, required to simulate the systems studied in this thesis, and the forcefields, required for the simulation. Section 3.2.1 presents a general N-Body (Lennard-Jones, Electrostatics) algorithm for emerging high-performance computer graphics cards, which accelerate some of the simulations, in particular for the materials sciences over 100-fold compared to standard CPUs. I continue the chapter with the discussion of implicit solvent models in section 3.2.2 used to simulate most of the biomolecular systems and conclude with implementation specific details of the development of SIMONA in section 3.3.

Method for absolute Quality Control of Protein Structures

In chapter 4, I present a method for absolute quality control of protein models. Knowledge of a protein's 3D structure allows insight into its function and possible avenues to modulate its function. However the reliability of many protein models arising from protein-structure-prediction methods, in particular those generated by fully automated servers, is unclear, which has led to a low acceptance of theoretical protein models in the life-science community. In this chapter, I report on the development of a novel approach for absolute quality assessment of protein models for the important subclass of globular proteins.
As proteins are only marginally stable I hypothesized that not only the complete global protein structure is in the state of lowest free-energy, but that also the distribution of energies of single amino acids is characteristic for the folded state. I therefore collected statistics of energy contributions of individual amino acids, using a biomolecular forcefield developed previously in the group from a set of experimental high-resolution protein structures and derived a N-dimensional statistical test to assess the quality of a single protein structure by comparing against the experimental data. Our method does not rely on extensive sampling of the protein's conformational space, as it is able to judge the quality a-priori. Applied to decoy sets of globular proteins, we could identify the low quality structures in the sets in 93% of the cases.

Chapter 5 presents three applications of several projects for protein-structure prediction, in part in collaboration with experimental groups.

Design of genetically engineered proteins for biocompatible surfaces

Development of novel materials for implants has become an urgent issue in medicine. The longer life-expectancy of patients, who often get their first implant before the age of 50, increasingly necessitates implantreplacement at very high age, when operations carry a highly increased risk. Development of implant materials that last longer than two or three decades would obviate the need for many of these operations. One possibility to develop such implants is research into surfaces with engineered biocompatibility with respect to cell adhesion. Ideally the surfaces should attract bone stem-cells, but at the same time repel bacteria and other cells that do not participate in strong adhesion between the newly grown bone and the implant.

In section 5.2, I report a joint investigation of groups at the KIT and the orthopedic clinic at the University of Heidelberg to develop genetically-modified proteins for biocompatible implant coating. Hydrophobins, which are fungal proteins with interesting physico-chemical properties, are one possible candidate for the coating of implants as they are able to turn hydrophilic surfaces hydrophobic and therefore less susceptible to the adhesion of bacteria. However, initial studies of hydrophobin coated surfaces did not show improved cell adhesion. We have therefore developed a strategy to design genetically modified hydrophobins, which inherit the hydrophobic and immunologically inert characteristics of the parent protein, but improve the differential adhesion of cells. After designing the protein with computational tools its characteristics are verified experimentally.

To date there is no experimental structure of the hydrophobin-fusion-protein complex which we investigated. We therefore developed an atomistic model of the protein using structure prediction. With this model we identified a solvent-exposed surface suitable for genetic modification. By fusing RGD and LG3 binding motifs into the exposed site, we improved the cell adhesion properties of the hydrophobin: Stem-cells now adhere better to the surface coated with the genetically modified hydrophobin, while adhesion of bacteria and fibroblasts is suppressed. As one possible avenue to improve these coatings even further we investigated assembly mechanisms of hydrophobins into very stable rodlets at air-water interfaces to determine, which parts of the protein are essential for the rodlet formation.

Structural model of the development of gas-vesicles in aqueous bacteria

The study of protein aggregation has been an active research topic for many years, which increased further with the discovery of aggregation-related diseases, such as the formation of amyloid fibrils in patients with Alzheimer's disease. Presently structural information about many aggregates is difficult to obtain, because aggregations of unstructured assemblies are difficult to investigate with x-ray crystallography, while the complexes are too large for investigation with NMR. An example for such protein aggregates are gas-vesicles, which are not soluble and tend to form amorphous aggregates. Gas vesicles are present in many bacteria living in water and used by these organisms to regulate their buoyancy. Although the genetic information of the proteins participating in gas-vesicle formation have been identified, it has not been possible to resolve the three-dimensional structure of some of the most important proteins in gas-vesicle formation. Together with an experimental group at the University of Darmstadt, I have developed a first model for gas-vesicle formation in aqueous bacteria, which I report in

section 5.4.
Using protein-structure prediction methods we developed a nano-scale structure of the gas-vesicle protein GvpA, a major component of the gas-vesicle wall and its assembly into extended structures. Our results provide insight into the mechanism by which gas-vesicles are formed and how gas is trapped inside the vesicle. This investigation was complicated by the fact, that no structural information of similar (homologous) proteins was available. We therefore resorted to model the protein using de-novo structure prediction methods based on secondary structure analysis. The predicted structure of the protein monomer could be used to model the aggregation of the β-sheets by docking monomers to form an extended superstructure, which fits well with the observed rib-like shape of the gas-vesicles. We were able to provide experimental validation for the predicted structure by mutagenesis and protease cleavage experiments validating the most important contact sites predicted in the model and ATR-FTIR to validate the secondary structure content.

High-throughput prediction of peptide structures

In section 5.5, we extended our protein structure prediction efforts to design a method for the high-throughput prediction of peptides. Antimicrobial, antibiotical and antifungal peptides are emerging as novel drugs, to combat the growing immunity of some bacterial strains against current antibiotics. The experimental high-throughput screening of the activity of these peptides is a very costly and demanding task and requires the synthesis of thousands of different peptide sequences. This work could be significantly reduced by the knowledge of the peptide's 3D structure which could be used as the basis for structure-function relationships.
Homology or knowledge based methods for peptide structure prediction frequently fail to predict the correct peptide structure as peptide sequences are too short to find a viable homolog. In contrast to large proteins point mutations lead more often to large structural change. In this study, we implemented a de-novo prediction protocol on the POEM@HOME server to perform high-throughput prediction of peptide structures. We verified this protocol by predicting the structure of four experimentally known peptides of different topology. We further introduce a clustering algorithm to elucidate the low-energy landscape of the peptides and identify a shearing mechanism, by which a β-peptide transitions between its low energy states.

Chapter 6 presents two applications investigating protein-protein and protein-ligand interfaces.

Computational Alanine Screening

Many biological signaling processes are mediated by protein-protein association; many proteins function only as part of a complex assembly of subunits. For this reason protein-protein interfaces are emerging as novel drug targets. In comparison to complexes of proteins and small-molecule ligands, protein-protein interfaces are more extended. Several experimental techniques to target these extended interfaces, e.g. with antibodies have been demonstrated, but these remain costly and limited in the range of applicability. In particular small-molecule ligands are

highly desirable, because of their lower cost, ease of handling and better bioavailability. Due to the usually large area of protein-protein interactions, it is even more difficult to develop viable ligands that modify protein-protein binding. One of the established experimental techniques to identify the binding hotspots within protein-protein interfaces is alanine screening, which requires genetic modification of the protein sequence for every amino acid. Because this is costly, development of computational alanine screening protocols, which we report in section 6.1, can aid understanding the experiments and reduce the associated effort. We have developed an in-silico alanine-screening protocol to pre-screen specific interaction sites and verified its precision by investigating the binding hotspots of two experimentally known protein-protein complexes. This work was carried out in cooperation with an experimental group, the young investigator group of Dr. Katja Schmitz, to identify hotspots in the binding of chemokines, which are important target molecules in inflammation, to their receptors. The validation of the predictions by experimental methods is still ongoing.

De-novo protein-protein and protein-ligand interactions

In section 6.2, we present a strategy for protein-protein and protein-ligand docking. Methods for in-silico alanine screening (presented in the previous project) rely on the existence of a model of the protein complex, which is sometimes not available, even if structures for the binding partners exist. For this reason we implemented an algorithm to predict the binding pose of protein-protein complexes and verified it by docking three different protein-protein complexes. We were able to transfer the protein-protein docking protocol to dock also small-molecule ligands into protein receptors for drug design. We further introduce a method for high-throughput screening of protein-ligand complexes for drug development, which was verified by docking of six pharmaceutically relevant small molecule ligands to their protein receptors.

In chapter (Ch. 7), I report the first morphology simulations using the new SIMONA code for two nanoscale systems of interest to the materials sciences:

Morphology Simulations of Pentacene Clusters

In section 7.1, we introduce a method to model the morphology of amorphous pentacene clusters. Pentacene is a widely used material for the electroluminescent layer of OLED, as mobilities of pentacene composites rival those of amorphous silicone. Simulating the morphology of amorphous pentacene is the first step towards electronic structure calculations to modeling electronic transport in these systems, which is required for the development of computer-aided strategies to optimize the properties of OLED/OPV materials. We identify nucleation centers of pentacene stacking in a herringbone conformation, which have the same local order as pentacene crystals.

Dispersion of Single-Walled Carbon Nanotubes by Chiral Index

In section 7.2, we implement a strategy to investigate the wrapping of polymers around

nanotubes. Mechanical and electronic properties of nanotubes vary depend strongly on their chiral index. Efficient exploitation of these properties in industrial applications requires an efficient and inexpensive sorting methods to obtain a homogeneous nanotube material. A promising approach to nanotube sorting was the discovery of polymers, which bind selectively to nanotubes with specific chiral indices. Since many polymers can be envisioned for this purpose understanding of the mechanism of selective binding is essential. In agreement with the experimental results we modeled selective binding of polymers to nanotubes the different diameter and chiral index.

Chapter 8 concludes this thesis and summarizes the main insights won during the various projects.

2. Biomolecular Systems

Proteins incorporate many different functions vital to the human body. They catalyze chemical reactions, transcribe DNA, exert physical force by muscle contraction and mediate the immune response among many other functions[12]. Many proteins assume a unique, functional 3D structure, determined only by the sequence of amino acids they consist of. The misfolding and mutation of proteins is the cause of many serious and often fatal diseases. To understand biomolecular function and malfunction, the first step is to understand the structure of biomolecules.
In this chapter, we characterize the structure of biomolecules. In section 2.1, we introduce the components of proteins from the protein's sequence of amino acids, to local and global structural motifs. In section 2.1.3, we report folding mechanisms, by which some proteins fold into their tertiary structure. In section 2.2, we briefly discuss hypotheses explaining the marginal stability of proteins.

2.1. Biomolecular Structure

Proteins are polymers of amino acids, most of which fold spontaneously into a unique 3D structure. The amino acid chain of a protein is comprised of an alphabet of twenty different amino acids[12]. The DNA carries the information for all primary amino acid sequences in the human body. It is transcribed in multiple intermediate steps from an alphabet of four nucleotides of the DNA into an alphabet of twenty amino acids.

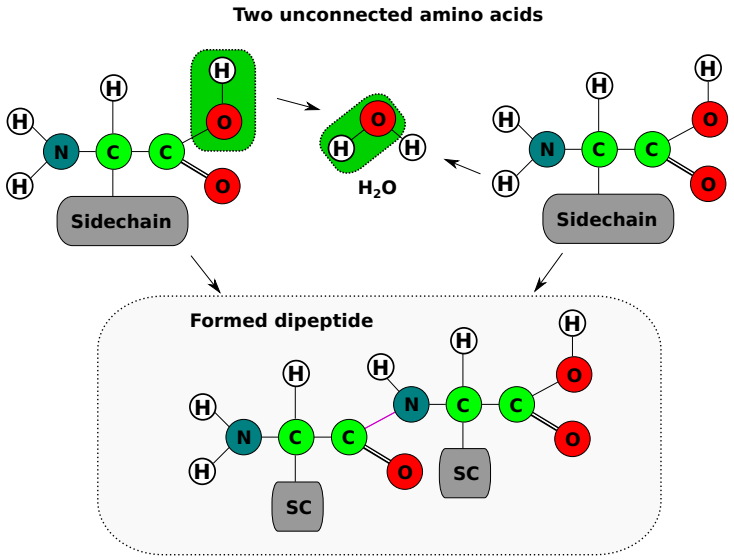


Fig. 2.1.: *Formation of a peptide bond between two amino acids. The terminal OH group from the N-terminal amino acid and a hydrogen atom from the C-terminal amino acid form a water molecule. The now open bond positions are bonded together by a new peptide bond between carbon and nitrogen.*

2.1.1. Primary Structure – Amino acid sequence

Amino acids are molecules containing both a functional amide- and a carboxyl-group[12]. Amino acids can be linked by covalent peptide bonds illustrated in Fig. 2.1. The terminal carboxyl- and amino- groups of the respective amino acids are linked together and form a water molecule. Twenty amino acids, which feature identical backbone and distinct sidechain atoms, exist in the human body. The carbon atom bound to the sidechain is usually denoted as C_α, the other carbon in the mainchain is denoted as C'.
The polypeptide of linked amino acids of a protein is called the primary structure. The informational content stored in this sequence is equivalent to the corresponding strand of DNA. Because of the planarity of the peptide bond shown in Fig. 2.1, two consecutive C_α atoms lie almost in the same plane. Two backbone dihedral angles define the relative orientation of the peptide planes. The dihedral angle ϕ involves the backbone atom C' of the preceding amino acid in the chain and N,C_α, and C' of the following amino acid, while the dihedral angle ψ involves the backbone atoms N,C_α,C' and N of the next amino acid in the chain. Shear within the peptide plane is denoted with the angle ω, which involves the atoms C_α,C' from one and N and C_α of the successive amino acid. Two different amino-acid isomers exist: Most amino acids occur in the trans-isomer ($\omega \approx 180°$) in the folded state, with an exception being Proline, which also frequently occurs in the cis-isomer ($\omega \approx 0°$). The definitions of the three dihedral angles, ϕ,ψ,ω are shown in Fig. 2.2. According to their sidechains, amino acids can be grouped by their various chemical properties into charged, polar and hydrophobic amino acids. Twelve of the twenty amino acids can be synthesized by the human body, the eight remaining amino acids are called essential amino acids and need to be supplied externally. The properties of the twenty amino acids can be found in Fig. 2.3[13].

2.1.2. Secondary Structure – Formation of locally stable segments

Pauling and Corey proposed two locally stable segments in proteins: β-sheets[14] and α-helices[15]. Both motifs are stabilized by the development of multiple mainchain hydrogen

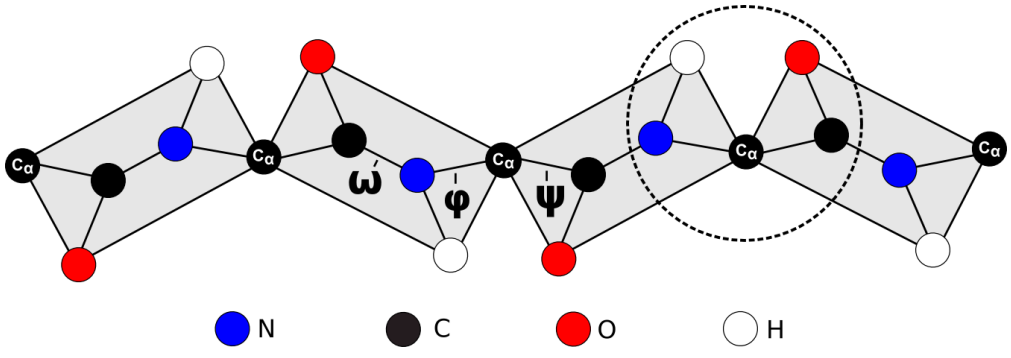


Fig. 2.2.: *Illustration of three mainchain dihedral angles* (ϕ, ψ, ω). *The dihedral angle* ϕ *involves* C', N,C_α *of one and* C' *of the following amino acid.* ψ *involves the backbone atoms* N,C_α,C' *of one and* N *of the next amino acid. The dihedral angle* ω *involves* C_α,C' *from one and* N *and* C_α *from the following amino acid. As the dihedral angle* ω *is close to* $180°$ *in the trans isomer,* C_α *atoms lie in plane (gray areas). The circle denotes the atoms of a single amino acid from N- to C-terminal. Sidechains are not shown.*

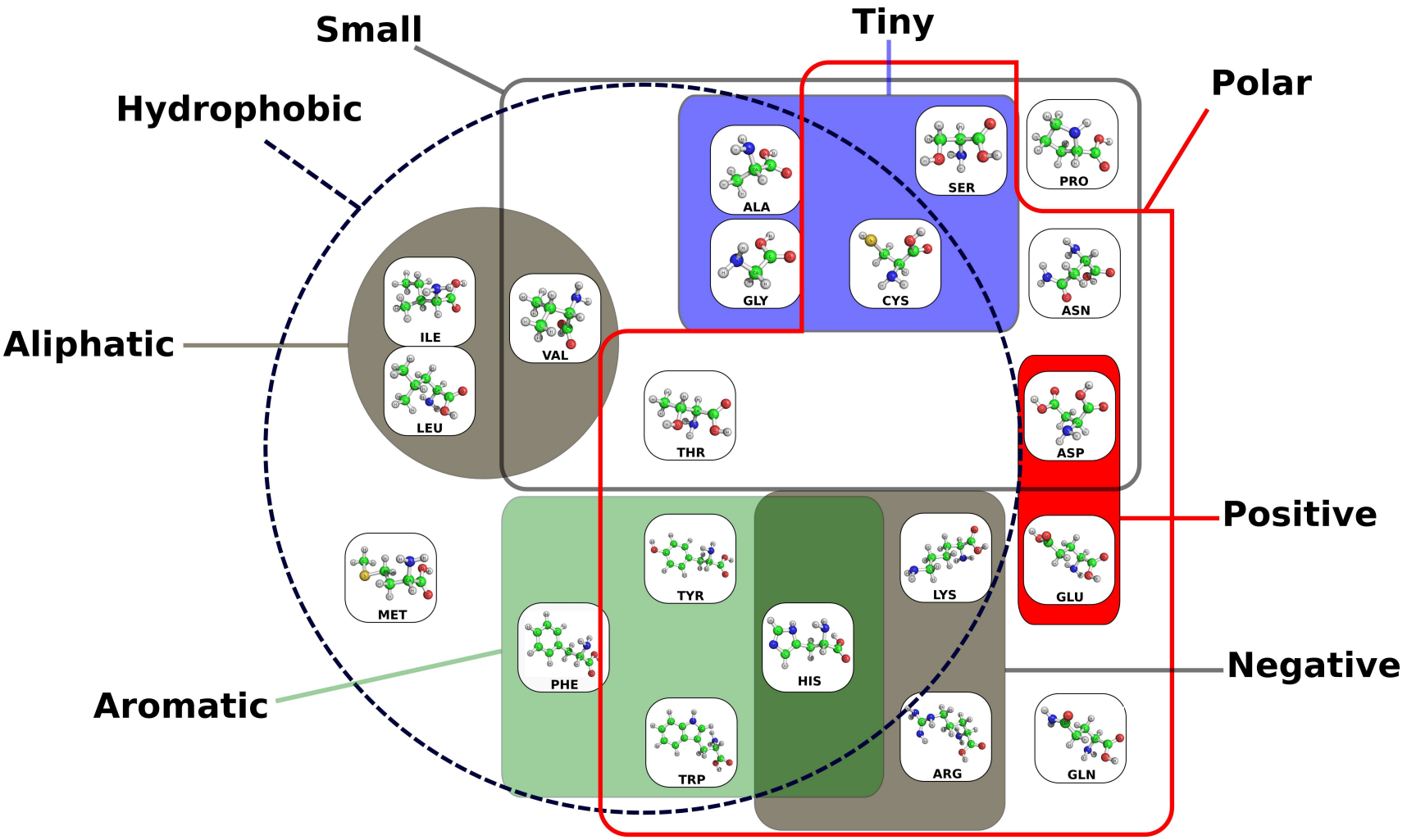
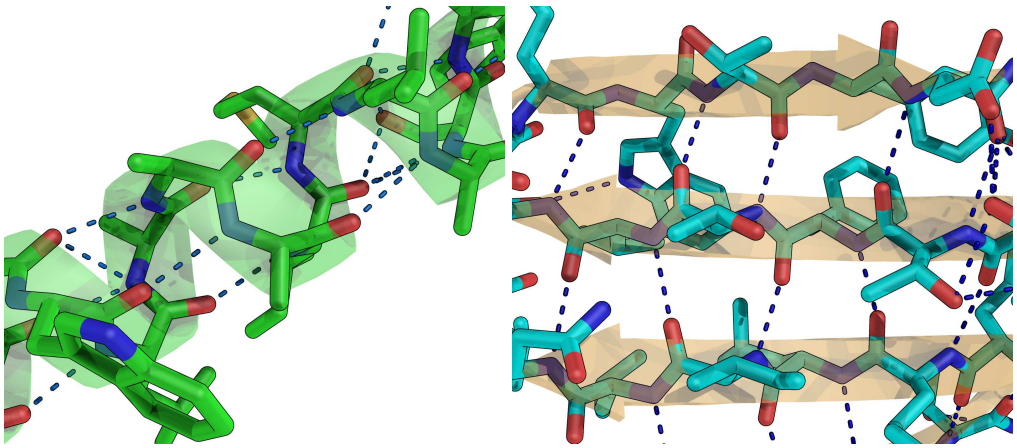


Fig. 2.3.: *Classification of amino acids properties. Image adapted from Livingstone and Barton[13].*



(a) α-Helix (b) β-Sheet

Fig. 2.4.: *Two major secondary structure motifs: α-helix and β-sheet. Both motifs are stabilized by development of multiple hydrogen bonds. While the α-helix is stabilized by hydrogen bonds (blue dashed lines) between every turn, the β-sheet is stabilized by hydrogen bonds between two bridges.*

bonds. The α-helix winds the mainchain helically, while pointing the sidechain outwards as seen in Fig. 2.4a. The mainchain oxygen group of each amino acid a_i develops a hydrogen bond to the nitrogen group of the amino acid a_{i+4}. β-sheets are pleated conformations of the mainchain exposing the sidechains of following amino acids on opposite sides (Fig. 2.4b). The N and O groups are exposed perpendicular to the sheets direction and form hydrogen bonds with adjacent β-bridges. Depending on the bond order, sheets are classified as antiparallel (two amino acids on opposing side share two hydrogen bonds) or parallel (N and O groups of one amino acid bond to two amino acids on the other sheet, which are separated in sequence by another amino acid).

2.1.3. Tertiary Structure – Development of long range native contacts

The assembly of the local secondary structure into a global fold is called tertiary structure. Anfinsen's thermodynamic hypothesis states that a protein finds its native conformation as the global minimum of its free-energy $G = H - T\,S$[16]. The protein's unique 3D structure is therefore defined only by its primary amino acid sequence. Several folding theories exist, which explain how a protein reaches this unique conformation. Levinthal ruled out a simple random search through conformational space by estimating the time to find the unique folded conformation using just random sampling to exceed the age of the universe[17]. One of the most prevalent folding hypotheses is the theory of a folding funnel[18], illustrated in Fig. 2.5. In this theory, a trade-off between internal energy U and entropy S occurs in the folding process. While during the initial stages the protein has a large conformational freedom (high entropy) the chain collapses into a molten globule state afterwards reducing in energy and entropy. During this stage the first native contacts form, leading to the transition state, after which the protein may or may not pass discrete folding intermediates towards the final native conformation (see Fig. 2.5). Literature reports three different generalizations of this funnel scenario:

- *Hydrophobic collapse model*: Proteins in the *hydrophobic-collapse* model first develop

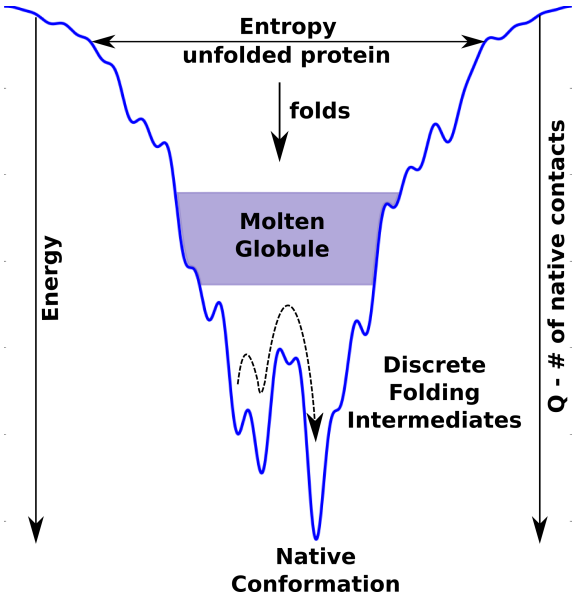


Fig. 2.5.: *Protein folding in the funnel view. The entropy is shown as the width; the energy as the depth of the funnel. The protein has large conformational freedom in its extended conformation when starting to fold (top). It drops down in energy and entropy into a molten globule collapsed state in which the first native contacts start forming. At the end of the folding process the protein undergoes several discrete intermediate conformations until it rests at the global free energy minimum. Figure based on work by Onuchic et al[18].*

a hydrophobic protein core by enclosing hydrophobic amino acids within a hydrophilic shell. Secondary structure development is proposed to occur afterwards[19].

- *Framework model*: The framework model predicts secondary structure development as the initial stage of folding. The stable secondary structure elements then diffuse through conformational space until native contacts stabilize the complete tertiary structure[20].

- *Nucleation condensation model*: In the *nucleation* model locally-stable segments form first. The rest of the structure then forms incrementally around this nucleation point. The three sheet WW domain protein FIP35, which was recently studied in extensive Molecular Dynamics simulations[6], showed this behavior: Before the second sheet was able to fold, the first β-sheet had to be created from the nucleated β-turn. A further specialization exists in the *nucleation-condensation* model, where nucleation points fold and unfold during the folding process.

Very often one cannot attribute the folding mechanism of a specific protein to a single of the three theories, but rather to a mix of the above[21]. Fig. 2.6 shows the folding of the fructose repressor DNA-binding domain 1UXD, which featured properties of both the hydrophobic collapse model and the framework model.

2.2. Protein stability

It is widely accepted that most globular proteins are marginally stable with ΔG ranging from 5 to 15 $kcal/mol$[23]. Experimental and theoretical studies successfully increased the thermo-

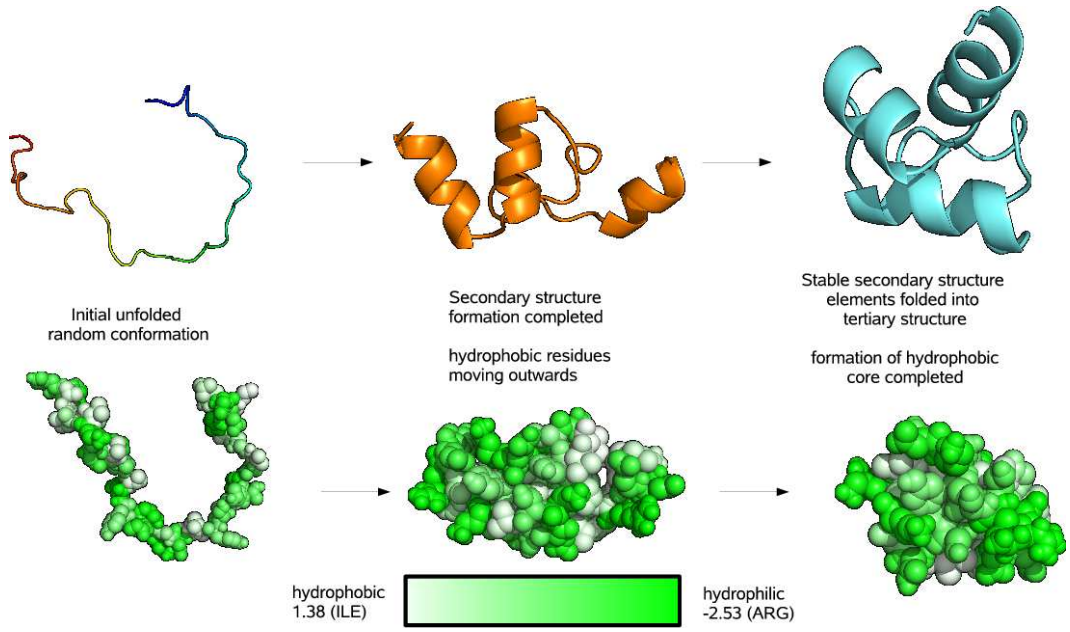


Fig. 2.6.: *Three snapshots of a folding simulation of the fructose repressor DNA-binding domain 1UXD. During the initial stage, the protein undergoes both development of secondary structure (upper row) and hydrophobic collapse (lower row). After initial helix development, the third helix diffuses into the gaps left by the other two helices (upper row, second and third image). During this process most of the hydrophobic surface (white, lower row) remains covered after the initial collapse. The hydrophobicity (Eisenberg-Scale[22]) is encoded in shades of green. Fig. published in Strunk[21].*

dynamic stability of proteins. The most popular method for increasing thermostability was the optimization of hydrogen bond networks and internal hydrophobic packing by mutation[24]. While it seems obvious that optimizing hydrogen bonding will lead to increased thermostability, it was expected that this optimization should have already occurred in nature due to evolutionary pressure, i.e. the protein structures should favor conformations of maximum thermostability. Multiple hypotheses exist to explain, why proteins are only marginally stable[25, 26]. They either propose an active reason implying that low thermostability is required for the correct protein function or state a passive reason attributing the low stability to a side effect of evolution.

Active Hypotheses

- The protein's low stability results in high flexibility required for proper biophysical function[27].
- Protein degradation is rendered impossible if proteins are too thermostable[27, 28].
- High protein stability could extend folding time or trap the protein in intermediate trapped states[25].
- Low thermostability might be needed for structural uniqueness[25].

Passive Hypotheses

- An (energetical) stability threshold exists above which random mutations dominate.

- The mutational random walk below the stability threshold is biased towards destabilizing mutations, as stabilizing mutations occupy only a minimal fraction of the whole configurational space[25].

The consensus of these theories is that proteins are marginally stable over the vast configurational space sampled by random mutations, which permit the correct function of the protein.

3. SIMONA: Simulation of Molecular and Nanoscale Systems

The exponential growth of the available computational resources increasingly permits molecular simulation methods to complement experiment and theory in understanding and predicting the properties of molecular and nanoscale systems in the life- and materials sciences [6, 29–31]. While Monte-Carlo algorithms are widely used to estimate thermodynamic properties in condensed matter physics[32], the majority of simulations in the life-sciences use Molecular Dynamics methods as the main workhorse[33–35].

For many processes in biological systems and materials development, it is sufficient to consider the equations of motion that describe the time evolution of a system of particles in a classical forcefield. The simulation of the trajectory of these particles can be carried out using Molecular Dynamics methods. Molecular Dynamics methods solve the Newton equations of motion (Eq. 3.1) using various numerical integration techniques.

$$m_i \frac{\partial^2 \vec{r}_i(t)}{\partial t^2} = \vec{F}_i(t), \qquad i = 1, ..., N \tag{3.1}$$

$$\vec{F}_i = -\frac{\partial V}{\partial \vec{r}_i} \qquad . \tag{3.2}$$

Most of these techniques discretize the equations of motion at timesteps Δt. One of the most popular integration techniques is the Verlet-Störmer method[36]:

$$\vec{r}(t + \Delta t) = 2\,\vec{r}(t) - \vec{r}(t - \Delta t) + \frac{\vec{F}(t)}{m}\,(\Delta\, t)^2 + O(\Delta t^4) \tag{3.3}$$

$$\vec{v}(t) = \frac{1}{2\,\Delta t}\,(\vec{r}(t + \Delta t) - \vec{r}(t - \Delta t)) + O(\Delta t^2) \tag{3.4}$$

The local error of $\vec{r}(t)$ in a Verlet-integration (Eq. 3.3) is therefore of fourth order in Δt. As the local errors accumulate over time, the global error of a Verlet integration is $O(\Delta t^2)$ for both $\vec{r}(t)$ and $\vec{v}(t)$. To guarantee stability of the simulations, the timestep Δt has to be chosen commensurate with the shortest timescale occurring in the system: atomic oscillations happen on femtosecond timescales, therefore timesteps of $1\,fs$ are common in most Molecular Dynamics simulations. This discretized timestep is many orders of magnitude smaller than the timescales of most relevant processes, which occur on ms to s timescales. A millisecond simulation at a femtosecond timestep requires about 10^{12} evaluations of the force F for all particles in the system. Investing in this effort, the reproducible folding of small proteins was reported with simulations spanning trajectories of more than a millisecond[6]. These simulations were only possible using a specialized computing architecture currently available only to a single research group worldwide. The de-novo folding of large proteins or simulations of large conformational change are still out of reach, even with these specialized architectures.

In comparison most Metropolis Monte Carlo techniques, and variants thereof, do not suffer

from the timescale problem and provide equivalent thermodynamic information of the system. A Monte-Carlo algorithm generates a chain of conformations, by perturbing the old conformation and accepts or rejects the proposed conformation based on its physical distribution. In this approach, motions on short timescales (oscillations) can be averaged out and do not need to be modeled explicitly. Unlike Molecular Dynamics algorithms, most Monte-Carlo algorithms do not offer kinetic information about the system, although Kinetic Monte-Carlo generalizations exist for special systems[37]. Presently few efficient code packages for Monte-Carlo simulations are available, such that the full potential of this method family is not realized.
This chapter comprises the development of SIMONA, a general-purpose tool for Monte-Carlo simulations. Section 3.1 reports the Metropolis Monte-Carlo method. In section 3.2, we will discuss the forcefields needed to parametrize the systems under study and explain their implementation into the SIMONA simulation package developed in section 3.3. Parts of this chapter were previously published in Strunk et al.[38]. I thank the publisher John Wiley and Sons and all co-authors for the possibility to republish these materials as part of my thesis.

3.1. Monte-Carlo simulation techniques

Monte-Carlo methods generate a chain of conformations, such that the time-average of the sampled conformations equals the ensemble average in the limit of long simulation time. The most widely used algorithm employs Metropolis Monte-Carlo sampling to evaluate thermodynamic expectation values according to an equilibrium distribution (Eq. 3.5)[39].

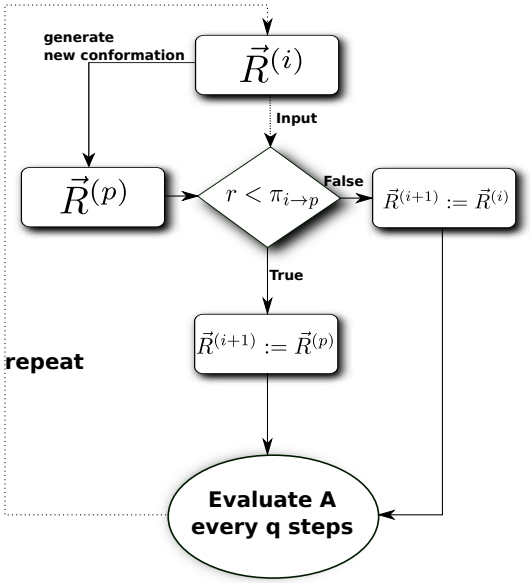
$$\langle A \rangle_\rho = \frac{\mathbf{Tr}(A\, e^{-\beta\, H})}{\mathbf{Tr}(e^{-\beta\, H})} \qquad . \tag{3.5}$$

Explicit evaluation of Eq. 3.5 is not numerically feasible for most non-trivial systems, because of the size of the configuration space. However, for many systems, the partition function Z is dominated by a much smaller subspace, which is sampled by means of importance sampling in Metropolis Monte-Carlo simulations.

Metropolis Monte-Carlo

Metropolis Monte-Carlo methods implement a Markov-Chain process, in which a proposed conformation $\vec{R}^{(i)}$ is generated from the previous conformation $\vec{R}^{(i-1)}$ without memory of prior conformations[39]. The general Markov-chain algorithm follows the scheme shown in Fig. 3.1. The transition probability $\pi_{i\to j}$ between conformations i and j is chosen such that the distribution of states $\rho^{(i)}$ converges towards the target equilibrium distribution $\rho = \rho^{(\infty)}$. If we denote the distribution of states at step i with $\rho^{(i)}$, the distribution in step $i+1$ can be obtained by multiplying with the transition probability matrix $\Pi(i,j) = \pi_{i\to j}$ (Eq. 3.6):

$$\rho^{(i+1)} = \rho^{(i)}\Pi \qquad . \tag{3.6}$$



The Metropolis Monte-Carlo algorithm:

1. We start with an initial conformation $\vec{R}^{(i)}$.
2. We generate another conformation by a random perturbation $\vec{R}^{(p)}$.
3. We draw a new random number $r \in [0, 1]$.
4. If r is smaller than the transition probability $\pi_{i\to p}$, we accept $\vec{R}^{(p)}$ as the new $\vec{R}^{(i+1)}$; otherwise we continue with $\vec{R}^{i}$ as the new $\vec{R}^{(i+1)}$.
5. Every q steps, we evaluate our observable A and update the mean value $\langle A \rangle_\rho$.

Fig. 3.1.: *Generalized Markov-Chain Monte-Carlo simulation flow: New conformations are generated and accepted according to the transition probability $\pi_{i\to p}$.*

In equilibrium, the distribution is invariant under the same transformation (Eq. 3.7):

$$\rho^{(\infty)} = \Pi\rho^{(\infty)} \quad . \tag{3.7}$$

Hence $\rho^{(\infty)}$ has to be eigenvector to Π with eigenvalue 1. The probability to observe the system in any of the states i is $\sum \rho_i = 1$ (the system is definitely in one state). As the probability content $\sum \rho_i$ must not increase or decrease in any step $i \to i + 1$, the sum of all transition probabilities in a row of the matrix has to also be 1. A matrix with these features is called a stochastic matrix. One example to fulfill Eq. 3.7 is:

$$\frac{\pi_{i\to j}}{\pi_{j\to i}} = \frac{\rho_i}{\rho_j} \quad . \tag{3.8}$$

Eq. 3.8 is called the detailed balance condition, which needs to be fulfilled to converge towards the equilibrium distribution. The most popular choice for $\pi_{i\to j}$ is the Metropolis criterion[39] given in Eq. 3.9.

$$\pi_{i\to j,M} = \begin{cases} M\ \exp(-\beta\,(E_j - E_i)) & E_j > E_i \\ M & \text{otherwise} \end{cases} \quad . \tag{3.9}$$

Evolutionary algorithm

Often Metropolis Monte-Carlo techniques are not efficient enough to reach the thermodynamic equilibrium for large systems. Sometimes, for example for the problem of protein structure prediction and refinement, the detailed balance criterion does not need to be fulfilled, as one is only interested in the conformations of lowest energy, but not the complete equilibrium ensemble. Extensions of Metropolis Monte-Carlo algorithms exist to reach the equilibrium faster and sample a larger area of the conformational space in parallel. In this section we present an

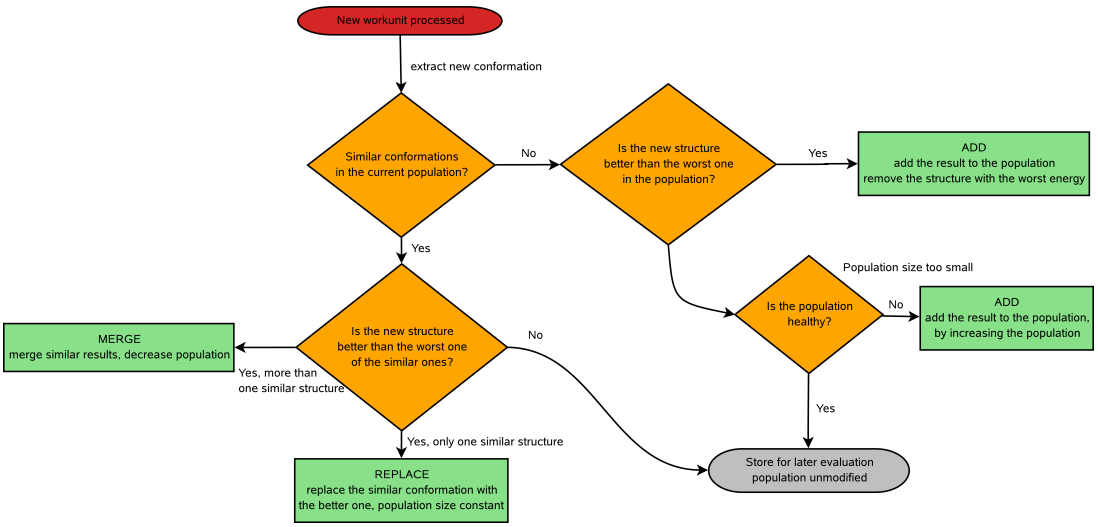


Fig. 3.2.: *Flowchart of the Evolutionary Algorithm. The Evolutionary Algorithm focuses on evolving structures in a population by balancing structural diversity and energy optimization. After randomly annealing a single structure from the population, the conformation is extracted and compared to the population. In case of similar conformations worse of energy, these are discarded from the population. If the new conformation is energetically viable and structurally dissimilar to the existing ones it is included in the population.*

evolutionary algorithm, which evolves many protein conformations inside a population towards the global energy minimum, while balancing population diversity (see Fig. 3.2). Balancing the structural diversity of the population of conformations allows the parallel sampling of different parts of the free-energy landscape, also those inaccessible by barriers. In comparison, multiple Metropolis Monte-Carlo simulations started from N similar replica, would sample mostly the area around the thermodynamic equilibrium, where they were started.
One iteration of this algorithm consists of the following steps:

1. Metropolis Monte-Carlo simulations are started from the members of the population (usually as many as the number of available CPUs).
2. A fixed amount of Metropolis Monte-Carlo steps is simulated.
3. The structures are assembled at the end of the simulations and sorted into the population:
   a) If no similar conformations are found within the population:
      i. The structure is accepted, if the population consists of not enough members.
      ii. The structure is accepted, if it has a better energy than the worst energy structure currently in the population.
      iii. The structure is rejected otherwise.
   b) If similar conformations are found inside the population:
      i. If the conformation has the best energy among all similar ones, all similar structures are deleted, the new conformation is retained.
      ii. The structure is rejected otherwise.

As a similarity criterion one can use the RMSD between two structures. If the RMSD is below a threshold, conformations are regarded similar and removed from the population. This algorithm ameliorates the problem of Monte-Carlo to freeze the conformation on a rugged energy landscape, where a Monte Carlo algorithm can get stuck in local minima. The evolutionary algorithm solves this problem in a similar way to Tabu search algorithms[40, 41]. In Tabu search algorithms, a single simulation is carried out, which keeps a memory of visited conformations. These visited conformations must not be revisited by the algorithm in subsequent simulation steps, they are "taboo". If the simulation has once visited a local minimum, subsequent simulation steps will sample the vicinity of the minimum and ultimately be able to leave region of the conformational space near the minimum, as all conformations in the vicinity will be forbidden.
In comparison, the evolutionary algorithm allows a single member of the population to be "stuck" in a local minimum. Due to the similarity criterion, other members of the population are not allowed to reside near the same conformation and are therefore forced to sample other regions of the energy landscape.
If the population size is too small, the evolutionary algorithm can still freeze and get stuck in many local minima. We have implemented a multi-temperature generalization, which evolves several populations at different temperatures. A temperature is assigned to each replica, when a new simulation is started. This temperature can either be the temperature of the population it resides in, or the temperature of the population one step higher or one step lower in temperature. Using this algorithm, which is akin to parallel tempering[42–45], barriers in the energy landscape can be overcome, by switching into another population.

3.2. Forcefields

Many processes in nature, which do not break or form covalent bonds, can be described by classical mechanics. The potential functions that approximate the energy of the systems under study are called forcefields. Many forcefield terms are now completely standard and many specific parametrizations of forcefields exist for the systems of interest[46–50]. This section introduces the types of forcefields implemented in SIMONA. In section 3.2.1, we first introduce generalized distance dependent forcefields and discuss the parallel evaluation of these energy terms for N-body systems. In section 3.2.2, we present methods for implicit treatment of solvation effects.

3.2.1. Pairwise interactions

Pairwise interactions are modeled using forcefields of the form:

$$E_P = \sum_{i=0}^{N} \sum_{j=0, j\neq i}^{N} f(\vec{r}_i, \vec{r}_j) \qquad . \tag{3.10}$$

Most systems simulated in this thesis require a potential that models Pauli-exclusion and dipole-induced London attraction. These are modeled using a standard 6-12 Lennard-Jones poten-

tial[51]. The most common form of the Lennard-Jones potential is:

$$U = 4\,\epsilon \left[\left(\frac{\sigma}{R} \right)^{12} - \left(\frac{\sigma}{R} \right)^{6} \right] \quad . \tag{3.11}$$

Electrostatic interactions are most often described using partial-charge Coulomb electrostatics[52]:

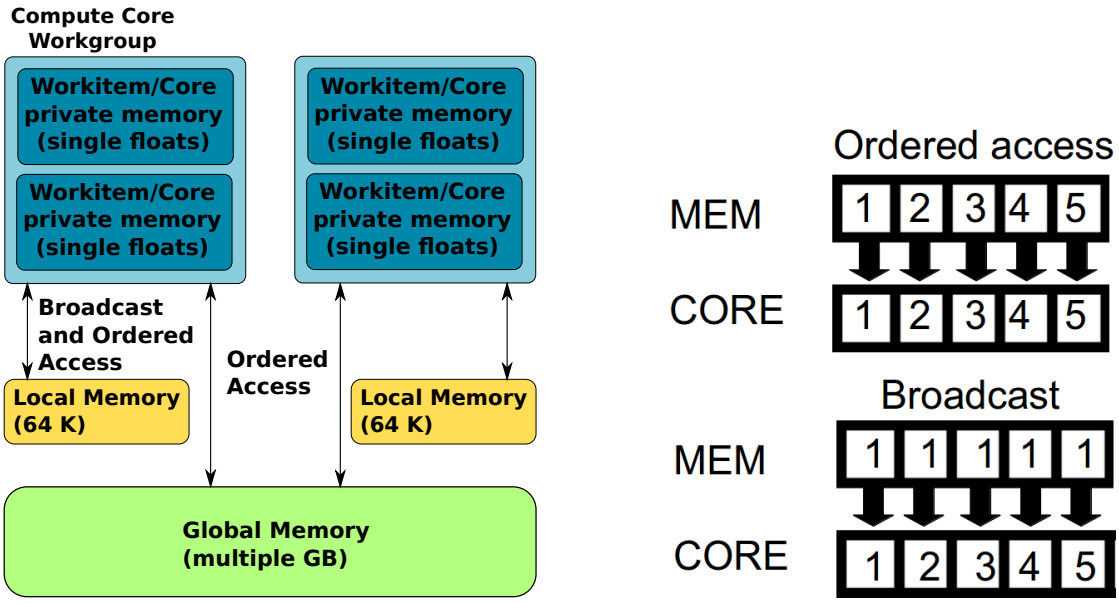
$$E = \frac{q_1\, q_2}{4\,\pi\,\epsilon_0} \frac{1}{|\vec{r_i} - \vec{r_j}|} \quad . \tag{3.12}$$

With increasing system size, the *N-Body* problem, i.e. the evaluation of pairwise potentials for a system of N atoms, constitutes the most computing intensive component. For systems with $50,000$ atoms or more, 95% of the time are spent in the $O(N^2)$ evaluation loop of the forcefield evaluation in Eq. 3.10, e.g. for the electrostatic energy. We therefore developed a parallel implementation to evaluate general N-Body terms on GPU architectures using the OpenCL framework[53]. Previous implementations of efficient pairwise potentials often evaluated only a fraction (usually $O(N \log(N))$ of the interactions by a spatial (tree-) decomposition of the atom coordinates to optimize the simulation of large systems[54]. Short-range interactions, such as Lennard-Jones use cutoffs, while long-range interactions, most notably electrostatic interactions, use multipole-expansions[55]. The work presented here is based on work by Elsen et al.[56] and was previously published in Strunk et al.[57]. It uses an algorithm of higher order $O(N^2)$ than treecode algorithms, but is easier to parallelize and therefore faster for the systems of intermediate size ($10^4 - 10^5$ atoms)[58].

GPU optimization of N-Body forcefield evaluations

To port the evaluation of Eq. 3.10 efficiently, we have to keep the limitations of modern SIMT (Single Instruction, Multiple Threads) architectures, like GPUs, in mind[59]. An off-the-shelf 2012 single-chip graphics card can have up to 2048 cores grouped into 32 compute units (Data from AMD Radeon 7970). Each compute unit is only able to execute the same codepath at the same time using different data as input. Every branch (if, else statement) inside an OpenCL kernel doubles the execution time of the branches, as one branch is idle, waiting for the other branch to finish.

As different compute units can execute different code branches at the same time, workitems (also called tasks or threads) are grouped into workgroups, which are executed on different compute cores: The size of one workgroup has to be divisible by the size of a compute unit[59]. Most GPUs nowadays adopt a three-stage memory hierarchy (see Fig. 3.3). Global off-chip memory is the largest memory area with sizes of $2\,GB$ for consumer products to $24\,GB$ for compute cluster solutions. Access to this memory is generally slow compared to the on-chip cache and should only occur in an ordered fashion, i.e. 64 cores should access 64 consecutive memory positions, of which the first is aligned to the on-chip memory (see Fig. 3.3 b). Whenever an unordered memory access occurs, an extra read instruction is generated: 16 ordered reads result in one read instruction, 16 unordered reads result in 16 read instructions. All cores have access to the whole segment of global memory.



(a) OpenCL memory models

(b) Memory access patterns

Fig. 3.3.: *Memory hierarchy and access patterns of SIMT architectures[53]. (a) Current GPUs adopt a three-staged memory model. A large segment (multiple GB) of global memory is placed off chip, i.e. not within the GPU. This memory allows efficient ordered reads from all cores. Workgroups can access a faster shared local memory segment. Local memory sizes differ between architectures but are usually less than $100\,K$. This memory segment allows ordered and broadcast reads. Every core has a very small segment of private memory, which usually holds only a few floating point values. These segments allow for random access. (b) Ordered (up) and Broadcast (down) memory access patterns. In a broadcast read, every workitem of a single workgroup reads the same memory position. In an ordered read, consecutive cores read consecutive memory positions.*

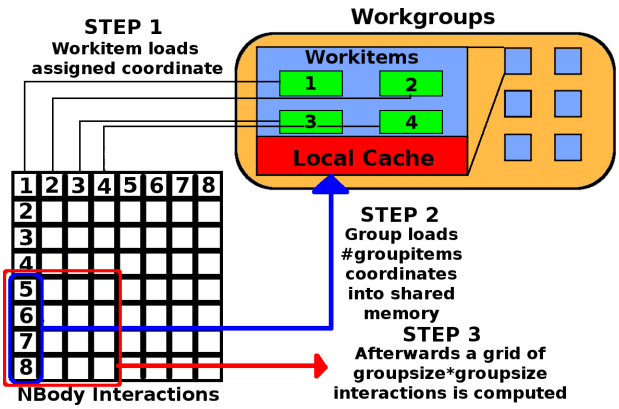


Fig. 3.4.: *N-Body parallelization strategy. Step 1: Every core loads its assigned coordinate ordered into the private memory of a core. Step 2: A coordinate segment is loaded into the local memory (ordered load, ordered store). Step 3: All cores iterate over the loaded coordinates by broadcast loads and calculate the interaction energy between their assigned coordinates and the loaded one. This is repeated for all coordinate segments.*

Access to local memory is faster as it is usually situated on-chip and shared between workgroups. A common local memory size is $64\,KB$. Local memory only persists for a single execution of a kernel in a workgroup; afterwards it is cleared. Similar to the global memory, only ordered reads can be carried out in parallel. Additionally local memory supports a broadcast in one operation. During a broadcast the workgroup can read one memory position at the same time. Random access is only possible from register memory local to the single core. Only a few floating point numbers fit into this memory.
Keeping the GPU and memory limitations in mind we designed the N-Body kernel in the following way; N denotes the number of atoms and also the number of started kernels, W_S denotes the workgroup size:

1. For each atom one workitem (kernel) is started grouped into N/W_S workgroups.
2. Each workitem loads its coordinate in an *ordered* manner.
3. The sequence of all N coordinates is split into segments of length W_S. The last segment might have an overhang containing empty elements.
4. The following instructions are carried out for each integer segment number between 1 and N/W_S.
   a) A segment i of length W_S is loaded by each workgroup ordered and stored in local memory.
   b) The workitems iterate over all $(i \cdot W_S, i+1 \cdot W_S]$ coordinate members in this segment and load the matching coordinate via a broadcast read from local memory.
   c) Each core calculates the interaction energy with its own assigned coordinate and accumulates it locally.
5. Once all segments are computed, the sum over all N energies is returned.

This evaluation strategy is illustrated in Fig. 3.4. The reader should note that no memory access

in this algorithm is more expensive than a single instruction, as only ordered reads are used from global memory and ordered and broadcast reads and writes are used from local memory. The drawback of this strategy is however that segments have to be loaded multiple times for multiple coordinates. This strategy scales perfectly for architectures with exactly N cores. Benchmarks could show a speedup of 150 compared to a single CPU core for a 240 core graphics adapter running at $1.5\,Ghz$ compared to a single $2\,Ghz$ CPU for systems of $30,000$ atoms (data in Strunk et al.[57]).

3.2.2. Implicit treatment of solvent molecules

Simulations of biomolecules in aqueous solution have to include up to ten times more water atoms, than are present in the main constituent of the simulation. To alleviate the complexity of the simulation very often implicit solvent models are employed. To this end, the Hamiltonian is split into three separate parts, describing the energy of the main constituent (the protein) $H_S(\vec{R})$, the solvent $H_W(\vec{W})$ and interaction $H_{SW}(\vec{R},\vec{W})$. Here, $\vec{R}$ denotes the configuration of the main system, $\vec{W}$ the configuration of the water molecules[60, 61].

$$H_C(\vec{R},\vec{W}) = H_S(\vec{R}) + H_W(\vec{W}) + H_{SW}(\vec{R},\vec{W}) \qquad . \tag{3.13}$$

Using Eq. 3.13 the partition function can be split into two terms as shown in Eq. 3.16.

$$\begin{aligned} Z &= \int\limits_{\mathcal{R},\mathcal{W}} \exp(-\beta\, H_C(\vec{R},\vec{W}))\, d\vec{R}\, d\vec{W} \qquad , & (3.14)\\ &= \int\limits_{\mathcal{R},\mathcal{W}} \exp(-\beta\, H_S(\vec{R}))\, \exp(-\beta\, H_W(\vec{W}) - \beta\, H_{SW}(\vec{R},\vec{W}))\, d\vec{R}\, d\vec{W} \qquad , & (3.15)\\ &= \int\limits_{\mathcal{R}} \exp(-\beta\, H_S(\vec{R}))\, \left[\int\limits_{\mathcal{W}} \exp(-\beta\, H_W(\vec{W}) - \beta\, H_{SW}(\vec{R},\vec{W}))\, d\vec{W}\right] d\vec{R} & (3.16) \end{aligned}$$

If we define an effective Hamiltonian $H_{W,eff}$, we can simplify Eq. 3.16.

$$H_{W,eff}(\vec{R}) = -\frac{1}{\beta} \ln \left[\int\limits_{\mathcal{W}} \exp(-\beta\, H_W(\vec{W}) - \beta\, H_{SW}(\vec{R},\vec{W}))\, d\vec{W}\right] \qquad . \tag{3.17}$$

Equations 3.16 and 3.17 then simplify to Eq. 3.18:

$$Z = \int\limits_{\mathcal{R}} \exp\left(-\beta[H_S(\vec{R}) + H_{W,eff}(\vec{R})]\right) \qquad . \tag{3.18}$$

$H_{W,eff}$ is called an implicit solvent model. The effective Hamiltonian $H_{W,eff}$ can be interpreted as the free energy $\Delta\, G_W$ of the solvent[62]. An analytical solution for Eq. 3.17 is too complicated for simulations of the size of biomolecules, but many models exist in literature. In the following section, we present two implicit solvent models: a solvent accessible surface area model and a generalized Born model.

Solvent Accessible Surface Area Model

In a Solvent Accessible Surface Area (SASA) model, one popular implicit solvent model, the solvent contribution to the free-energy is modeled linear in the solvent accessible surface area A_i of the atoms i (Eq. 3.19), as proposed by Eisenberg and McLachlan[60].

$$\Delta G_W = \sum_{i=0}^{N} \sigma_i A_i \qquad . \tag{3.19}$$

As the evaluation of the solvent exposed surfaces A_i take a significant amount of computational time, SIMONA employs an efficient evaluation protocol for the solvent accessible surface, as described in Klenin et al.[63]. The σ_i can be determined empirically by fitting to the known transfer energies of single amino acids or GLY-X-GLY tripeptides from octanol to water[61].

Generalized-Born electrostatics

The linear solvation model does not completely account for polarizability of the solvent. Solvent free-energy contributions ΔG_W can be split into polar $\Delta G_{W,\text{polar}}$ and non-polar $\Delta G_{W,\text{non-polar}}$ contributions. The polar contribution to the effective water Hamiltonian could, in theory, be evaluated by solving the Poisson-Boltzmann equation[64]. As this would be too expensive to compute in each simulation step, approximations have been available in the form of generalized Born models[65]. Generalized Born models introduce effective Born radii R_i describing the burial of charge i inside the low-dielectric medium. The expression for the generalized Born energy reads:

$$G_{\text{GB}} = \frac{1}{8\,\pi\,\epsilon_0}\left(\frac{1}{\epsilon_W} - \frac{1}{\epsilon_P}\right)\sum_{i,j}^{N} \frac{q_i\,q_j}{\sqrt{r_{ij}^2 + R_i\,R_j\,\exp(-r_{ij}^2/4\,R_i\,R_j)}} \qquad . \tag{3.20}$$

The variables ϵ_P and ϵ_W represent the dielectric constant of the protein and the solvent, R_i is the Born radii of atom i. Various definitions of the Born radii can be found in Still et al.[65] and elsewhere. After evaluating the Born radii, the evaluation of the energy constitutes a pairwise interaction term and can be parallelized with the strategy described in section 3.2.1.
The combined GB/SA implicit solvent model combines a linear non-polar solvent accessible surface area contribution to model non-polar free energy $\Delta G_{W,\text{non-polar}}$ with the generalized Born term to model polar contributions $\Delta G_{W,\text{polar}}$. It is a popular choice to describe solvent effects and used in the implicit solvent version of the Amber99SB forcefield[46, 47, 66].

Protein forcefield PFF02

The forcefield PFF02[9] (Eq. 3.21) models the internal free energy of protein conformations. It is comprised of five terms modelling electrostatics, angle-dependent hydrogen bonding,

Lennard-Jones, solvent interactions and mainchain torsions.

$$
\begin{aligned}
E_{\text{PFF02}} &= V_{\text{lj}} + V_{\text{ele}} + V_{\text{hb}} + V_{\text{pse}} + V_{\text{tor}} \qquad (3.21)\\
V_{\text{lj}} &: \text{Lennard-Jones potential}\\
V_{\text{ele}} &: \text{Electrostatics}\\
V_{\text{hb}} &: \text{Hydrogen bonding}\\
V_{\text{pse}} &: \text{Linear solvation contribution}\\
V_{\text{tor}} &: \text{Torsional potential}
\end{aligned}
$$

PFF02 was shown to select the near-native conformations for all 32 monomeric proteins from the ROSETTA decoy set[67]. It was also used to fold a set of 27 proteins with up to 72 amino acids with helical, sheet and mixed secondary-structure from extended conformations[9, 68]. A full forcefield specification can be found in Verma et al.[9]. The electrostatics model V_{ele} includes polarization effects by scaling the electrostatic interaction with the exposed surface area of the group. Solvent is treated by a linear implicit solvent accessible surface area term as presented in section 3.2.2.

3.3. Implementation of the general purpose Monte-Carlo simulation package SIMONA

Proteins are nanoscale structures consisting of an amino acid chain of an alphabet of twenty amino acids, i.e. well known components. One of the primary aims of developing SIMONA was to ease the parametrization of these standard systems for the user as much as possible. At the same time, we aimed to simplify simulation of other components of biological system, like small-molecule ligands, and material simulations on the nanoscale. The simulation of these systems requires either different parametrizations in one of the existing forcefields, or the development of completely novel forcefields.
SIMONA provides a two-stage concept to implement the simulation: A Python-preprocessor splits the input information into abstract coordinates (the Configuration object), assigns forcefield information (radii, forcefield parameters) and stores them in a Forcefield object, detects degrees of freedom (Moves object) and implements a temporal simulation hierarchy for the simulations (Algorithm). These four distinct sections (Configuration, Moves, Forcefield and Algorithm) are stored in a user editable XML file and read by the SIMONA kernel written in C++ to run the simulation (Fig. 3.5). The Python preprocessor is implemented as a GUI application; the C++ SIMONA kernel is a command-line application compatible with x86 and PowerPC CPU architectures on Linux and x86 architectures on Mac and Windows.
The generation of the input XML can be controlled by the user at various levels, in the following listed in the order of the complexity of the required changes of the protocol or program.

- For standardized applications (proteins, amino acids) the user can set up the simulation with the graphical interface by employing several tutorials for selected applications.
- Non-standard interactions or parameter values can be defined by changing assignments in

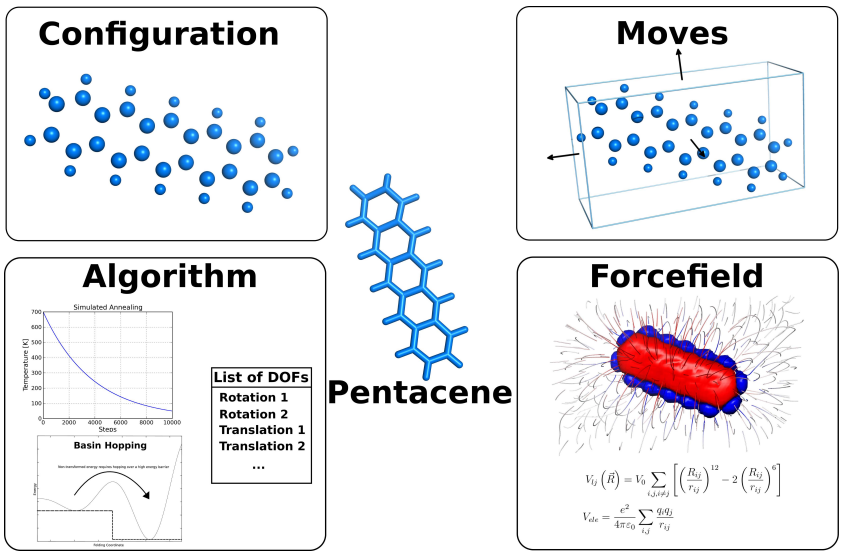


Fig. 3.5.: *Definition of a system for use with SIMONA. The SIMONA preprocessor splits the information required to simulate a system into four categories: 1. The* Configuration *section contains only the coordinates of the atoms. 2. The* Moves *section contains specifications of degrees of freedom, which can be perturbed during the simulation (dihedral angles, rigid body movements) 3. The* Algorithm *section contains the protocol, which is used to perturb the system using the specifications in the* Moves *section (Metropolis Monte-Carlo, Parallel Tempering). 4. The* Forcefield *section implements all interactions between the atoms in an energy model.*

the preprocessors.

- Complex algorithms, which are key to many specialized MC methods, can be encoded using a XML-based programming language.
- Class derivation on the Python and C++ side permit the expert user to implement novel methods, while inheriting all existing features.

Input definitions of the four XML sections (Configuration, Forcefield, Moves and Algorithm) can be found in the code documentation[38].

Algorithm section of a SIMONA XML file

An exemplary input file for the *Algorithm* section is shown in Fig. 3.6. Due to the modularity of the XML file, transformations and acceptance criterions can be replaced with other classes to customize the simulation. Simulation strategies currently implemented in this way include Parallel Tempering[42–45] and a multiple-try Monte-Carlo scheme[69]. As a more complex example, we implemented the evolutionary algorithm previously introduced in section 3.1. The XML code for this algorithm is shown in Fig. 3.7. Comparison with Fig. 3.6 for the standard Metropolis Monte-Carlo algorithm illustrates that the much more complex algorithm can be implemented very easily.

Implementation of novel potentials

Similar to the modularity of transformations carried out during the simulations, forcefields can also be easily adapted inside the XML. All the forcefields presented in section 3.2 are available

```
<Algorithm>
  <RepeatedMove>                                  for i in 0..12000 do
    <repeats>12000</repeats>
    <tstart>450.0</tstart> <tend>150.0</tend>       #geometrical temperature scaling
    <tscaling>geometric</tscaling>                  temperature = temperature_before*temp_factor
    <TransformationSequence>                        Do all transformations in a row
      <ConditionalTransformation>                     Create a copy of the current configuration
                                                      Apply following transformations to copy
        <TransformationChoice>                          Choose one transformation of
          <SetTranslationRandom wt="1.0">                 A random translation (with a weight of 1)
          <SetRotationRandom wt="1.0">                    A random rotation (with a weight of 1)
          <TransformationChoice wt="20.0">                Choose one (dihedral perturbation) (weight 20)
            <SetDihedralRelativeRandom wt="1.0">              A dihedral perturbation (weight 1)
            <SetDihedralRelativeRandom wt="1.0">              A dihedral perturbation (weight 1)
            <SetDihedralRelativeRandom wt="1.0">              A dihedral perturbation (weight 1)
              ....                                            ... (complete List of dihedral angles)
          </TransformationChoice>
        </TransformationChoice>
       <MetropolisAcceptanceCriterion>              Evaluate energy difference between copy and original
          <energymodel_nr>0</energymodel_nr>        Accept or reject based on Metropolis criterion
          <kB>0.0019858775</kB>
       </MetropolisAcceptanceCriterion>
      </ConditionalTransformation>
      <TransformationSequence>                      Do all transformations in a row
        <EnergyOutput>                                Output the energy
        <ConfigurationOutput>                         Output the snapshot for a trajectory
      </TransformationSequence>
    </TransformationSequence>
  </RepeatedMove>
</Algorithm>
```

Fig. 3.6.: *Algorithm section of a Metropolis Monte Carlo implementation translated into pseudocode. In every repeat of the RepeatedMove a Transformation is carried out (in this case a displacement, rotation or dihedral rotation). The resulting state is accepted or rejected by the Metropolis acceptance criterion. Right: Pseudocode, Left: Actual XML Code*

```
<Algorithm>
  <MultiEA>                                     Perform a MultiEA simulation with settings:
    <mea_cycles>10</mea_cycles>                 do 10 MultiEA steps
    <rmsd_threshold>3.0</rmsd_threshold>        RMSD threshold of 3.0 Å
    <jobs_per_node>10</jobs_per_node>           generate 10 new configurations per step on each MPI-node
    <maxpopsize>1000</maxpopsize>               keep population size below 1000
    <filename>mea.log</filename>                write verbose output into file 'mea.log'
    <temperature>450.0</temperature>            only take one population with temperature 450K
    <RepeatedMove>                              rule, how to generate new configurations
     <tscaling>none</tscaling>                  here: normal Metropolis MC run with 1000 steps
     <repeats>1000</repeats>                    for i in 0..1000 do
    <TransformationSequence>                      Do all transformations in a row
      <ConditionalTransformation>                   Create a copy of the current configuration
                                                    Apply following transformations to copy
        <TransformationChoice>                        Choose one transformation of
          <SetTranslationRandom wt="1.0">               A random translation (with a weight of 1)
          <SetRotationRandom wt=”1.0">                  A random rotation (with a weight of 1)
          <TransformationChoice wt="20.0">              Choose one (dihedral perturbation) (weight 20)
            <SetDihedralRelativeRandom wt="1.0">            A dihedral perturbation (weight 1)
            <SetDihedralRelativeRandom wt="1.0">            A dihedral perturbation (weight 1)
            <SetDihedralRelativeRandom wt="1.0">            A dihedral perturbation (weight 1)
              ....                                          ... (complete List of dihedral angles)
          </TransformationChoice>
        </TransformationChoice>
       <MetropolisAcceptanceCriterion>              Evaluate energy difference between copy and original
          <energymodel_nr>0</energymodel_nr>        Accept or reject based on Metropolis criterion
          <kB>0.0019858775</kB>
       </MetropolisAcceptanceCriterion>
      </ConditionalTransformation>
    </TransformationSequence>
    </RepeatedMove>
  </MultiEA>
</Algorithm>
```

Fig. 3.7.: *XML specification of the SIMONA Algorithm section for a simulation using the multiple population evolutionary algorithm. In comparison to the specification of a Metropolis Monte-Carlo simulated annealing simulation in Fig. 3.6, only a single XML class (MultiEA) had to be added with various configuration parameters. Right: Pseudocode, Left: Actual XML Code*

```
<DistanceConstraint>
     <first_id>30</first_id>
     <second_id>55</second_id>
     <distribution type="UserDefinedDist">
          <function>5*(r-7)^2</function>
     </distribution>
</DistanceConstraint>
```

Fig. 3.8.: *XML code required to implement a custom distance constraint potential. In this case the potential* $f(x) = 5\,(r-7)^2$ *was implemented, where* $r = |\vec{r}_{30} - \vec{r}_{55}|$*. The* <function> *object is an arbitrary string, which is compiled, when SIMONA reads the XML file.*

including parametrizations for protein systems. The implementation of novel forcefields is possible using the functionparser interface[70]. To allow for rapid prototyping, these forcefields are integrated into the XML code and work without recompilation of the SIMONA binary. Multi body and constraint potentials of the form $f(\vec{r}_1, \vec{r}_2, \cdots, \vec{r}_N)$ can be implemented in cleartext in the XML. The XML code for a single parabolic potential modeling a distance constraint of the coordinates of two particles is shown in Fig. 3.8. This scheme is extensible and was implemented for sets of atoms using the XML class MultiBodyPotential. As the implementation of the MultiBodyPotential derives the number of variables, i.e. the number of atom indices used in one evaluation of the forcefield, at runtime, the potential can be used to test very complex potentials to model for example reactive forcefields or coordination chemistry for transition metals.

3.4. Conclusions

In this chapter we presented the Monte-Carlo method and various forcefields that are widely used for the simulations of biomolecules. Monte-Carlo techniques avoid the timescale problem of Molecular Dynamics and provide thermodynamic information for the systems studied, but few implementations are presently available. We have also briefly described the implementation of a novel Monte-Carlo based simulation package, called SIMONA, to permit efficient simulation of molecular and nanoscale systems. The main paradigm in the design of SIMONA was to allow the easy extension to conduct simulations not considered during the development. The modular structure of the XML input file and the system agnostic implementation of the C++ code, allow for the parametrization of novel systems without changing the C++ code. The performance of many of the underlying forcefield components was proven to be very efficient[63, 71].
Given the wide range of possible Monte-Carlo methods and applications, we hope that other groups build upon our framework and extend the versatility of SIMONA. SIMONA is free for academic use and available at http://www.int.kit.edu/nanosim/simona.php .

4. Absolute Quality Assessment of Protein Structures

Modeling methods increasingly attempt to close the gap between the number of known protein-coding sequences[72] to that of structurally resolved proteins[3], but the results of these methods have been mixed. Adequate models can be built for proteins with high sequence similarity to a structurally resolved protein[73]. Occasionally modeling even succeeds in the absence of a good template[74], but none of the presently available methods can decide with certainty, whether a proposed model is of correct topology. For this reason acceptance of protein structure prediction has been low. While the development of accurate prediction methods for proteins with low homology is a long term goal[75], better acceptance of models from structure prediction would be achieved with methods that rank the likelihood that a particular structure represents an accurate model.
Current quality assessment methods can be grouped into two different categories: statistical (knowledge-based) and physics-based energy methods and machine-learning methods.

- *DOPE – Discrete Optimized Protein Energy*
  The *DOPE* (Discrete Optimized Protein Energy) scoring function is a statistical atomic distance dependent scoring function used in the *Modeller* program package[76]. As many similar approaches it is derived from a joint probability density function of the Cartesian coordinates of the protein atoms in a set of 1472 high-resolution crystallographic sample structures.
  By relating the statistical score to a single amino acid, DOPE obtains a per-residue statistical assessment indicating the “nativeness“ of the state of a single amino-acid. Although the *DOPE* potential was successfully verified by decoy assessment, only few applications of the score are reported in literature[73, 77].

- *SVM – Support Vector Machines*
  Meta-servers draw the predicted native conformation from a population of structures obtained from various different sources. One approach to construct a relative measure of protein quality to assess structures from different sources employs a support vector machine (SVM), a concept originating from machine learning.
  In one implementation by Eramian et al.[78] 24 different individual scoring functions were included in a SVM. By optimizing the weights in a linear combination of all single functions in respect to the RMSD error of the model, a unified model score could be constructed, which outperformed the 24 individual scores.

However, such quality assessment scores deliver only a relative score value for a set of structures. The statistical ”average“ value of the scoring function obtained for a single specific structure is compared to energies of other models to identify the native structure. Therefore these methods are unable to judge the ”nativeness“ from a single model alone.

We investigate a novel approach for absolute quality assessment and elucidate, whether such a measure can be devised. We concentrate on monomeric, globular proteins, as an important subclass of proteins that avoids additional difficulties encountered in oligomers or membrane-bound proteins. In section 4.1, we introduce the main idea behind our approach for quality assessment. In section 4.2, we derive a N-dimensional statistical test, present our choice of the energy model and introduce the training- and decoy-sets. We conclude by analyzing the success of our method in section 4.3 and provide an outlook and other application areas of our data in the discussion section (section 4.4).[1]

4.1. Introduction

Many protein structure prediction methods rate protein structures using an established scoring function[73, 77, 80, 81] by comparing the energies of an ensemble of structures and choosing the lowest energy members of said ensemble as the prediction. Here we investigate an approach to provide an a-priori estimator of the quality of a protein model without comparing it to a competing ensemble using a free-energy scoring function.
The free-energy G of a protein's macrostate can be formally divided into energy contributions g_i by amino acid i as shown in Eq. 4.1.

$$G = \sum_{i=0}^{N} g_i \qquad . \tag{4.1}$$

Pair- or higher-order contributions are evenly split among the participating atoms. For a particular protein conformation $\vec{R}$, the single contributions g_i for amino acid AA can now be understood as random values drawn from the distribution $\rho_{AA}(g_i)$. The probability of the total free-energy $G = g_1 + g_2$ of a two amino acid protein would then be the fold of the two probability distributions ρ_{AA1} and ρ_{AA2} as in Eq. 4.2.

$$\rho_{AA1,AA2} = \rho_{AA1} * \rho_{AA2} \qquad . \tag{4.2}$$

This can be generalized to sequences of N amino acids by folding the native distributions of the N amino acids. The resulting Gaussian distribution is a distribution for the total free-energy G of the folded protein, which can be used as a quality measure in a standard statistical test. We recall that according to C.B. Anfinsen's thermodynamic hypothesis the distribution of the competing unfolded macrostates states must differ on a protein-by-protein basis from the free energies of the folded macro-states in order to stabilize the latter thermodynamically[16]. Although we will also investigate this testvalue, the main statistical test introduced in this chapter uses the values of all single amino acid energies g_i. It is explained in the next section.

[1] Parts of this section will be published as part of Strunk et al.[79]. I thank all co-authors for the opportunity to publish it as part of my thesis.

4.2. Methods

N-Dimensional hypothesis tests

The basis of our method consists of a N-dimensional hypothesis test. In the following section, we derive an expression for the smallest N-dimensional confidence interval for a specific confidence level α. The hypothesis of this test is:

- Null-Hypothesis: The per amino acid contributions $\vec{G} = g_1, ..., g_N$ of the protein are drawn from the energy distribution of native structures.
- Alternative Hypothesis: The per amino acid energy contributions are incompatible with the native distributions: The protein model cannot be a native protein structure.

Let us now assume to know all single distributions of our test parameter $\vec{G} = (g_1, \ldots, g_N)$ and call them $\rho_i(g_i)$. For a confidence value α (usually 0.95 to 0.99), we will derive an equation for defining an optimal confidence interval Ξ. We define the optimal interval to be the smallest confidence interval treating all amino acid contributions equal. If the estimated free energy $\vec{G}_M$ of a model is outside of the optimal interval Ξ, we discard the protein structure. The probability to incorrectly classify a protein structure from the native conformation, i.e. a Type-I error, is $1 - \alpha$. The definition of the confidence level is:

$$\alpha = \int \cdots \int_{\Xi} \rho_1(g_1) \ldots \rho_N(g_N)\, dg_1 \ldots dg_N \qquad . \tag{4.3}$$

We assume the single distributions ρ_i to be of Gaussian type, resulting in:

$$\alpha = \frac{1}{(2\pi)^{\frac{N}{2}} \prod_i \sigma_i} \int \cdots \int \prod_{i=0}^{N} e^{-\frac{1}{2}\left(\frac{g_i - \mu_i}{\sigma_i}\right)^2} dx_i \qquad . \tag{4.4}$$

Introducing new coordinates $\xi_i = \frac{g_i - \mu_i}{\sigma_i}$ will simplify the integral borders. The integration is carried out over N-dimensional spherical coordinates, $d\,\Omega$ denotes the angular component:

$$\alpha = \frac{1}{(2\pi)^{\frac{N}{2}}} \int \cdots \int \int_0^{R_\kappa} d\Omega_{N-1}\, dR\, R^{N-1}\, e^{-\frac{R^2}{2}} \qquad , \tag{4.5}$$

$$= \frac{\Omega_{N-1}}{(2\pi)^{\frac{N}{2}}} \int_0^{R_\kappa} R^{N-1}\, e^{-\frac{R^2}{2}}\, dR \qquad . \tag{4.6}$$

The integral in Eq. 4.6 can be simplified, as shown in Eq. 4.8

$$I_N := \int_0^{R_\kappa} R^N\, e^{-\frac{R^2}{2}}\, dR \tag{4.7}$$

$$= \begin{cases} (N-1)!!\, I_0 - \sum\limits_{i=1}^{\frac{N}{2}} \frac{(N-1)!!}{(2i-1)!!} f_{2i-1}\Big|_0^{R_\kappa} & N \quad \text{even} \\ (N-1)!!\, I_1 - \sum\limits_{i=1}^{\frac{N-1}{2}} \frac{(N-1)!!}{(2i)!!} f_{2i}\Big|_0^{R_\kappa} & N \quad \text{odd} \end{cases} \tag{4.8}$$

$$\text{with:} \quad f_N = R^N\, e^{-\frac{R^2}{2}} \qquad . \tag{4.9}$$

Here $A!!$ denotes the double factorial with $A!! = A \cdot (A-2)!!$ with $0!! = 1!! = 1$. The proof for Eq. 4.8 can be found in appendix A.1. I_1 can be evaluated analytically:

$$I_1 = \int_0^{R_\kappa} R\, e^{-\frac{R^2}{2}}\, dR = -e^{-\frac{R^2}{2}}\Big|_0^{R_\kappa} \qquad . \tag{4.10}$$

I_0 is usually tabulated as the error function erf:

$$\mathrm{erf}(x) := \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\, dt \qquad , \tag{4.11}$$

$$I_0(R_\kappa) = \int_0^{R_\kappa} e^{-\frac{t^2}{2}}\, dt = \sqrt{2} \int_0^{\frac{R_\kappa}{\sqrt{2}}} e^{-x^2}\, dx \qquad , \tag{4.12}$$

$$= \sqrt{\frac{\pi}{2}}\, \mathrm{erf}\left(\frac{R_\kappa}{\sqrt{2}}\right) \qquad . \tag{4.13}$$

The integral over the angular part Ω_N evaluates to:

$$\Omega_{N-1} = \frac{N\, \pi^{\frac{N}{2}}}{\Gamma\left(1+\frac{N}{2}\right)} \qquad . \tag{4.14}$$

Using these identities an equation for α can be given, which is only dependent on the R_κ:

$$\alpha = \frac{\Omega_{N-1}}{(2\pi)^{\frac{N}{2}}} I_{N-1} \tag{4.15}$$

$$= \begin{cases} \frac{N}{\Gamma\left(1+\frac{N}{2}\right) 2^{\frac{N}{2}}} \left((N-2)!! \sqrt{\frac{\pi}{2}}\, \mathrm{erf}\left(\frac{R_\kappa}{\sqrt{2}}\right) - \sum_{i=1}^{\frac{N-1}{2}} \frac{(N-2)!!}{(2i-1)!!} f_{2i-1} \Big|_0^{R_\kappa} \right) & N \quad \text{odd} \\ \frac{N}{\Gamma\left(1+\frac{N}{2}\right) 2^{\frac{N}{2}}} \left((N-2)!! \left(1 - e^{-\frac{R_\kappa^2}{2}}\right) - \sum_{i=1}^{\frac{N-2}{2}} \frac{(N-2)!!}{(2i)!!} f_{2i} \Big|_0^{R_\kappa} \right) & N \quad \text{even} \end{cases} . \tag{4.16}$$

This equation can be solved numerically, resulting in the $R_K(N)$ shown in Fig. 4.1 for a confidence level of 95%.

Residue-specific energy model

As we have no access to the ideal free energy of a protein model, the test must be developed with respect to an estimate, for which we selected the protein free-energy force-field PFF02 (section 3.2.2). The energy model was broken down into per-amino-acid contributions as shown in Fig. 4.2. As the dihedral potential was already amino-acid specific only the SASA and pairwise terms needed to be split into amino acid contributions as shown in Eq. 4.17 (SASA)

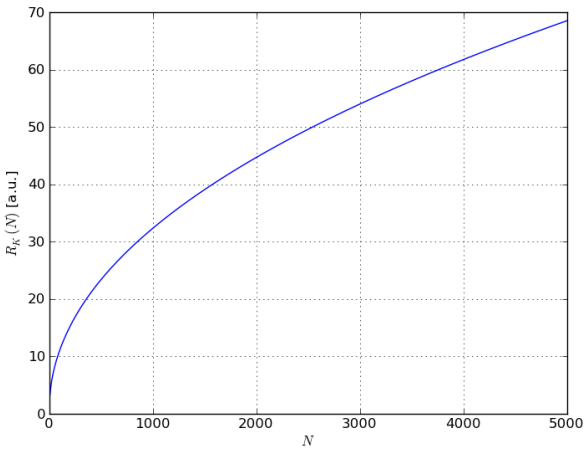
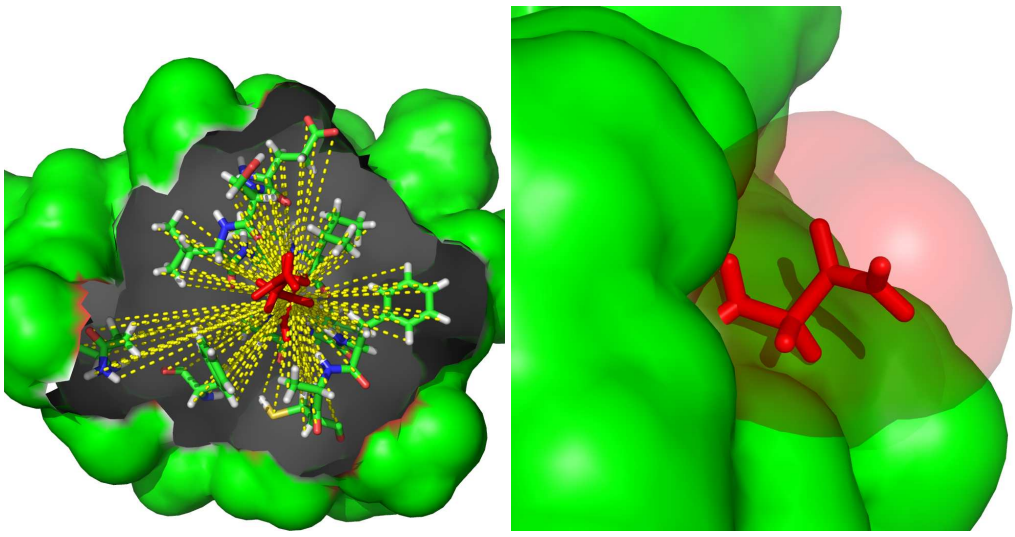


Fig. 4.1.: *Critical values of $R_K(N)$ for a confidence level of $\alpha = 0.95$. The critical values were estimated using Eq. 4.16. Due to the close proximity of individual data points, a line was plot through the discrete values of $R_K(N)$ to guide the eye.*



(a) Pairwise terms (b) SASA

Fig. 4.2.: *Illustration of the derivation of per-amino acid energies. a) For generalized pairwise terms, interactions are broken down pair-wise and accumulated for every amino acid. b) The solvent accessible surface area was evaluated and stored for every amino acid separately.*

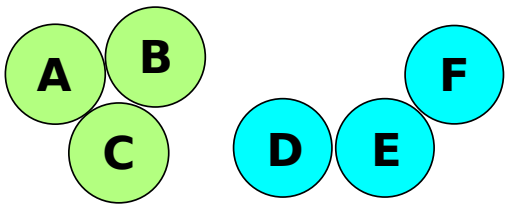


Fig. 4.3.: *Criterion used to decompose amino acid triplet energies for adjacent amino acids. The circles A,B,C are considered adjacent, as all three Van der Waals sphere touch. D,E,F are considered non-adjacent as D and F are not in contact. Note that amino acids need not follow each other in the sequence.*

and Eq. 4.20 (Pairwise).

$$E_{\text{SASA}} = \sum_{i=0}^{N_{\text{at}}} \sigma_i \, A_i = \sum_{i=0}^{N_{\text{amino}}} \sum_{j=0}^{N_{\text{at}} \in AA_i} \sigma_j \, A_j = \sum_{i=0}^{N_{\text{amino}}} e_{i,\text{SASA}} \quad . \tag{4.17}$$

$$E_{\text{Pairwise}} = \sum_{i=0}^{N_{\text{at}}} \sum_{j=0, i \neq j}^{N_{\text{at}}} f(r_i, r_j) \quad , \tag{4.18}$$

$$= \sum_{i=0}^{N_{\text{amino}}} \sum_{j=0, i \neq j}^{N_{\text{amino}}} \sum_{k=0}^{N_{\text{at}} \in AA_i} \sum_{l=0}^{N_{\text{at}} \in AA_j} \delta_{ij,kl} \, f(r_{i,k}, r_{j,l}) \quad , \tag{4.19}$$

$$= \sum_{i=0}^{N_{\text{amino}}} e_{i,\text{Pairwise}} \quad . \tag{4.20}$$

The per-amino acid energy functions were evaluated for a set of 256 high-resolution protein structures compiled by Dunbrack et al. into the CulledPDB dataset[82]. The PDB IDs used can be found in appendix A.2. We excluded membrane or multimer proteins, because their free energy contributions per amino acid cannot be reliably estimated on the basis of the monomer structure. Before the energies were calculated, missing atoms were added and bond lengths were normalized using the Rosetta idealization protocol[80]. Afterwards a relaxation simulation was started for 100 copies of each structure on POEM@HOME using fixed temperature simulations at $250\,K$. The low temperature was selected to keep structures near the native minimum.
We then extracted the contributions for the single amino acid energies of all individual PFF02 terms. Furthermore energies for neighboring amino acid pairs and triplets were extracted and accumulated. As the extracted pairs and triplets were not sequence dependent, i.e. $\rho_{\text{ALA,CYS}} = \rho_{\text{CYS,ALA}}$, only 1540 triplet and 210 pair distributions needed to be recorded. Information for triplets was restricted to amino acids where all three amino acids were in close geometric proximity (atom centers of the neighboring amino acids were closer than $5\,$Å), as illustrated in Fig. 4.3. The TM-Score, which ranges from 0 to 1, was used to quantify the quality of a protein structure in comparison with the native structure[83]. A score below 0.4 can be viewed as a random prediction and denotes decoys of a bad quality. TM-Scores higher than 0.9 are close to the experimental structure. In principle the RMSD-value could also be used to quantify the quality of a protein model, but it is difficult to identify a uniform RMSD threshold for proteins of different size.

Decoy Sets

We selected 160 random decoy sets of monomeric proteins from the database assembled by

Rajgaria et al.[84] as a test set. Membrane proteins were not considered. Structures binding metal ions or containing heme groups were removed from the population, as these interactions are not modeled in PFF02. All models in the sets were relaxed three times each using the same protocol as in the training set. Additionally the sets were enriched with three copies of the native structure. Per-amino acid energies were extracted and the testvector T^2_{abs} (Eq. 4.21) was calculated individually for all PFF02 energy contributions and the unweighted solvent accessible surface area.

$$T^2_{\text{abs}} = \sum_{i=0}^{N_{\text{DOF}}} \left(\frac{e_i - \mu_i}{\sigma_i} \right)^2 \tag{4.21}$$

The T^2_{abs} value was used to test for the model quality using the threshold $R_K(N)$ derived in Eq. 4.16. In the following investigation, we always show the score T^2 normalized to the critical threshold (Eq. 4.22):

$$T^2 = \frac{T^2_{\text{abs}}}{R_K(N)^2} \qquad . \tag{4.22}$$

Using the normalized score, structures passing the test have a T^2 value below 1.0; rejected structures have a T^2 value above 1.0. For large N, most of the area of the integral in the derivation of the threshold equation (Eq. 4.16) lies directly beneath the surface of the N-dimensional shell. Most of the structures drawn from native distributions are presumed to lie close to the shell. We therefore relaxed the critical T^2 value to 1.05 to allow for small errors in the numerical derivation of the R_K.

4.3. Results

The energy statistics for all 256 high resolution structures of the training set and the 160 decoy-sets are shown in Fig. 4.4. We always observe an overlap between decoy and native structures. A slight shift towards higher energies is observed for the distributions of the decoy structures (green histograms in Fig. 4.4). Especially for the hydrophobic amino acids ILE, PHE, TYR, VAL and LEU, we observe a large difference in the mean energy of the histogram between native and decoy structures. One possible explanation is that decoy structures are very often not perfectly packed; hydrophobic amino acids could still remain exposed, which leads to an unfavorable energy. The distribution of exposed surface areas for these amino acids shows that roughly one to two times more hydrophobic amino acids are buried in the native structures than in the decoy set. Most of the distributions of hydrophobic amino acids also feature a skew towards higher energies. The same observation holds for the distribution of exposed surface areas; hydrophobic amino acids are sometimes exposed even in the native state, leading to large contributions in energy.

Charged amino acids (ARG, ASP, GLU, HIS, LYS) cover a large energy range due to their long-range electrostatic contribution. Especially the mainchain hydrogen bonding energies showed a large deviation between native and non-native energies: All histograms present three major peaks for one, two or no main-chain hydrogen bonds. Native conformations incorporated two main-chain hydrogen bonds far more often than decoy structures. It is likely that the protocols

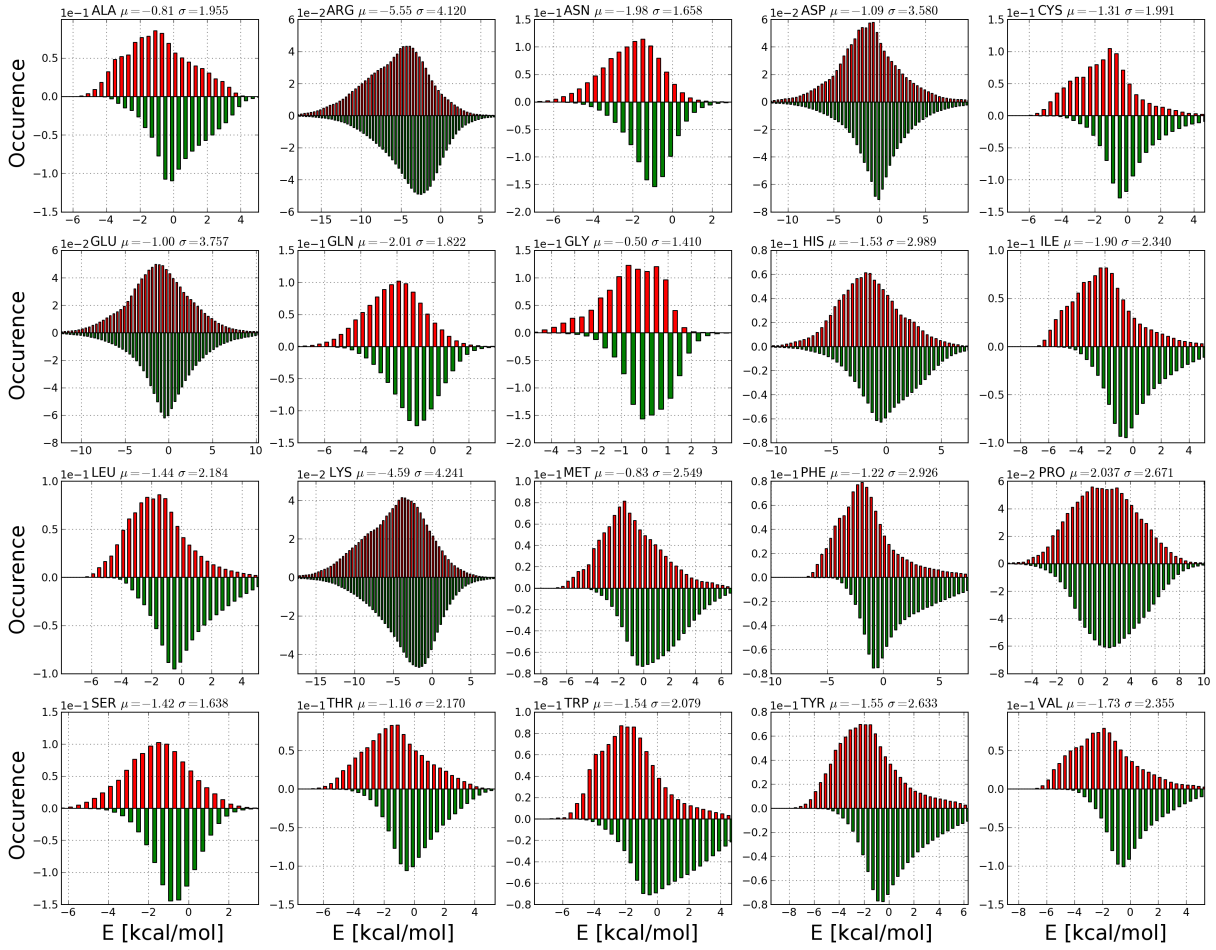


Fig. 4.4.: *Statistics for the amino-acid specific PFF02 contributions of all 20 amino acids. Positive occurrences relate to energy distributions of the set of native protein structures (red); negative occurrences relate to the decoy energy distributions. Only a slight deviation in the statistics between native and decoy statistics can be observed. This deviation is especially pronounced in the case of hydrophobic amino acids ILE, PHE, TYR, VAL and LEU and can be accounted to higher solvent exposure of hydrophobic amino acids in decoy structures. Mean value μ and standard deviation σ relate to the distribution of native structures.*

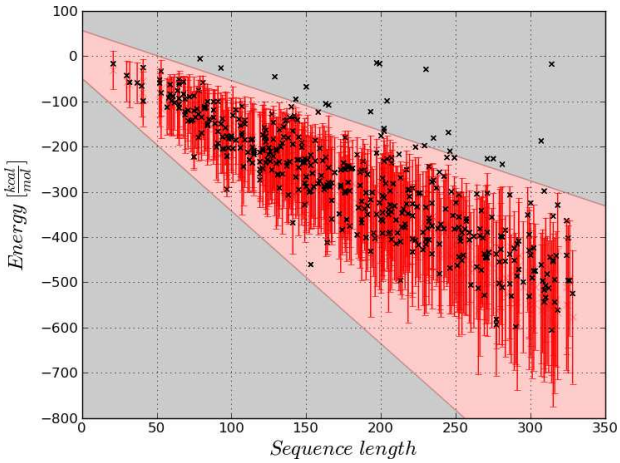


Fig. 4.5.: *Comparison of the sampled PFF02 energies with the energy ranges calculated for the specific convolutions. Confidence intervals ($3\,\sigma$) of the convolution histograms are shown as red error bars; the actual energies are shown as black crosses.* 418*, of the* 450 *structures, lie in the confidence intervals. Only one outlier lies below the error bars: The large positive energy contribution of the* 31 *remaining outliers can be accounted to steric overlap in the relaxation simulations.*

used to construct the decoys did not optimize the hydrogen bonding network fully or simply did not discover a structure with an optimal hydrogen bond network. Figures for the SASA, electrostatics and hydrogen bonding contributions can be found in appendix A.3. As previously mentioned, a per-amino acid energy e in a globular protein can be interpreted as a random value drawn from the sampled distributions $\rho_{AA1}(e_1)$. The energy of protein structure with a specific sequence SEQ is then drawn from the convolution of all single amino acid distributions (Eq. 4.23):

$$\rho_{SEQ}\left(\sum_i e_i\right) = \rho_{AA1}(e_1) * \rho_{AA2}(e_2) * \cdots * \rho_{AAN}(e_N) \quad . \tag{4.23}$$

Testing the total energy as a quality control test, we applied this convolution to the sequences of an extended set of 450 high-resolution structures of the Dunbrack PDB database and evaluated the complete PFF02 energy for them and extracted the mean μ and the three σ radius of the convolutions. Results are shown in Fig. 4.5.

Of the 450 structures the energies of 418 structures lie within the confidence range. All 32 outliers, except for a single one, lie above the confidence intervals. Sometimes a high energy can be attributed to a simulation artefact: A single steric collision within the structure during the relaxation simulation could have caused a large energy contribution for these structures. These results show that the sampled histograms are consistent with the energies of proteins of various topologies.

Statistical test for single amino acid statistics

Due to the large overlap in the energy distributions it was unlikely that a quality assessment method using only single amino acid energy distributions can succeed. Fig. 4.6 shows four characteristic distributions for the single amino acid energy criterion. For most of the analyzed structures, the bulk of the decoys cannot be separated. Although the initial hypothesis that

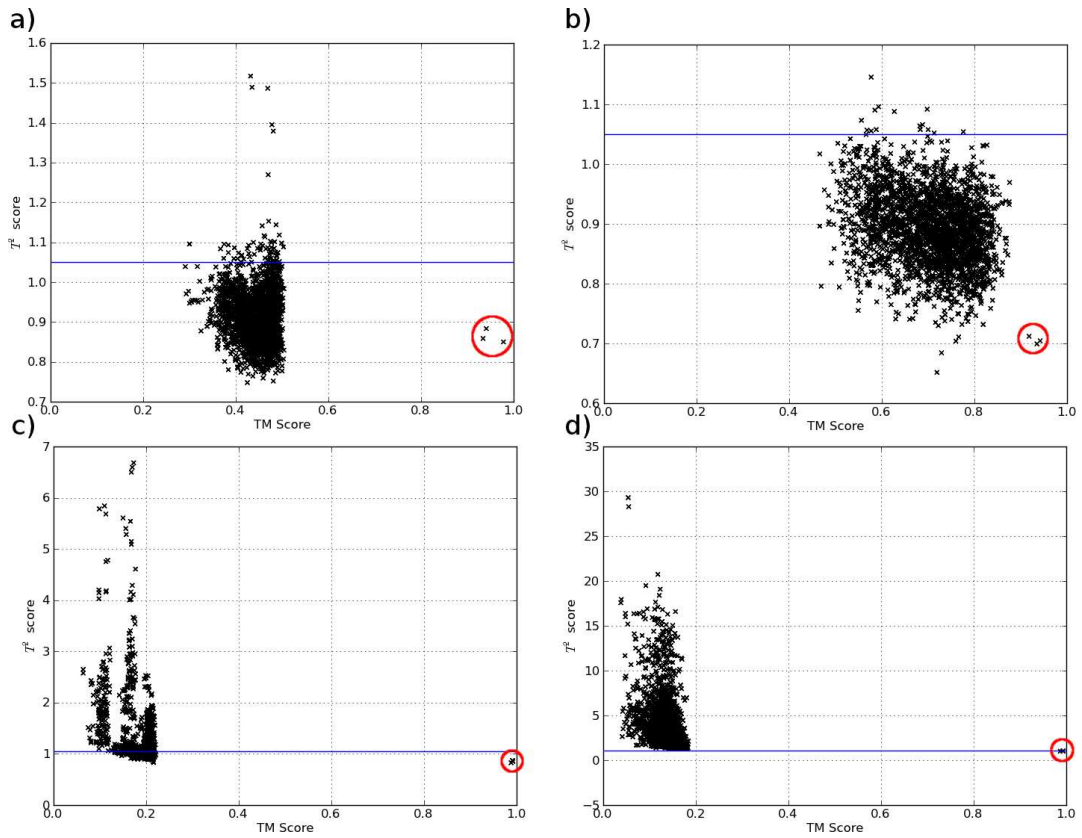


Fig. 4.6.: *Illustrative results for the quality assessment using single amino acid statistics of four decoy sets. The critical T^2 threshold was set to 1.05 (blue line). Relaxed native structures are encircled in red. a) Results for the bacillus cell fate determinant protein 1H4X: Although only low quality structures are present in the decoy set (TM-Score below 0.5), native and decoy structures cannot be differentiated by T^2 value. Most of the decoy structures exhibit per amino acid energies compatible with the native distributions. b) Results of the decoy set of Phl PII from timothy grass pollen 1BMW: The decoy set is higher in quality than the previous one of 1H4X. Although a tendency of the three native structures towards a lower T^2 deviation is visible, a classification by T^2 value is still impossible, as most of the bad quality decoy structures are considered native by our algorithm. c) Results of the decoy set screening of superantigen spe-h 1ET9: A tendency for the native structures towards lower T^2 values is visible. A significant amount of low quality decoys lies below the critical T^2 score. d) Results of the ligand binding domain of the EPHB2 receptor tyrosine kinase 1NUK: The low quality decoys are perfectly separated from the native structures. Within the population of 160 structures only very few decoy sets show this characteristic. a)+b)+c)+d) The native structures are below the T^2 threshold and therefore accepted for all four histograms.*

per-amino acid energy distributions are different between decoy sets and native experimental structures was true (see Fig. 4.4), the difference in these distributions was too small to differentiate between native and decoy structures. The proposed critical T^2 value lies above the T^2 value of many decoys: The decoy structures exhibit local amino acid energies compatible with the native amino acid energy distribution. Only for a few large proteins (see Fig. 4.6d) a good separation of decoy and native structures can be observed. Many optimization protocols, for example the Rosetta[85] and Modeller[76] toolkits, optimize the final prediction by amino acid energies. This could explain, why a large part of the amino acids are observed in states compatible with the native distribution. The results were similar for pairs of amino acids. We will therefore focus on the statistical test using triplet amino acid energies.

Statistical test for amino acid triplets

The statistical test using triplet amino acid energies calculates the energies for all triplet pairs of amino acids inside the protein. The energy of a triplet is the sum of three single amino acid energies:

$$e_{\text{triplet},i,j,k} = e_i + e_j + e_k \qquad . \tag{4.24}$$

The T^2 value is afterwards calculated against the 1540 means $\mu_{i,j,k}$ and standard deviations $\sigma_{i,j,k}$ of the triplet energy statistics of the experimental structures:

$$T^2 = \frac{1}{R_K^2(N_{\text{triplets}})} \sum_{i=0}^{N_{\text{triplets}}} \left(\frac{e_{i,j,k} - \mu_{i,j,k}}{\sigma_{i,j,k}} \right)^2 \qquad . \tag{4.25}$$

A T^2 score below the critical T^2 threshold of 1.05 denotes the classification as a native structure. Structures exhibiting T^2 scores exceeding 1.05 are classified as low quality models. The quality assessment plot for four representative proteins is shown in Fig. 4.7 using the same proteins, which could not be discriminated in the test considering only a single or pairs of amino acids, previously shown in Fig. 4.6. In the T^2 test incorporating triplet information, the separation is much more pronounced in three populations. Especially in the case of bacillus cell fate determinant protein 1H4X, the results are very favorable: While over 2500 structures resulted in false positive selections using single amino acid statistics (Fig. 4.6a), only five false positives remained using the triplet test (Fig. 4.7a).

We evaluated the True Negative Ratio (TNR = TN/(TN + FP)) for all structures of the test sets. Here TN is the number of correctly identified low quality decoys (True Negative) and FP the number of incorrectly identified low quality decoys (False Positive). Fig. 4.8 shows an explanation of these definitions. A structure with a TM-Score exceeding 0.94 was considered native; a structure with a TM-Score below 0.7 was considered a non-native decoy. We did not consider structures with a TM-Score in the range between 0.7 and 0.94 in the estimation of positive and negative results, as structures in this TM-Score range can already be considered native for the most part and only contain subtle local errors, which might lie within experimental resolution. A good quality assessment algorithm exhibits a very high True Negative Ratio (TNR), which indicates the fraction of structures correctly identified as low quality models. A TNR of 100% corresponds to the correct identification of all low quality models; a TNR of 0% corresponds to

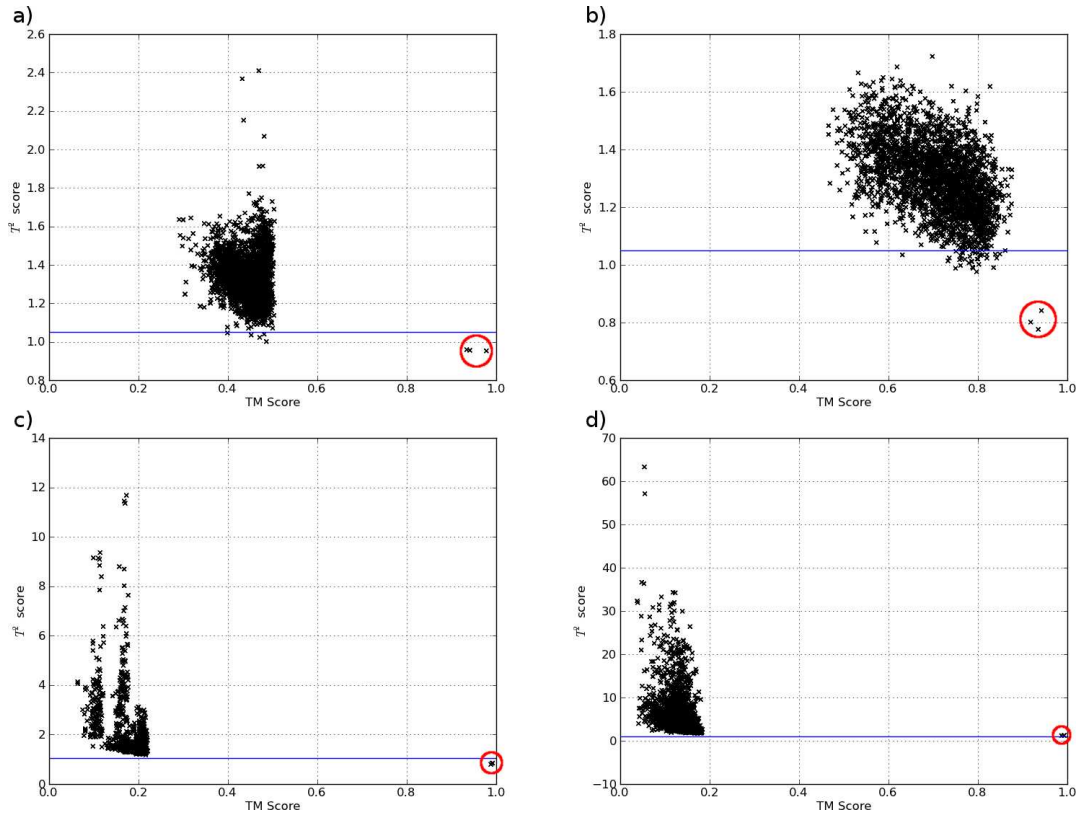


Fig. 4.7.: *Illustrative results for the quality assessment using triplet amino acid statistics of four decoy sets. The critical T^2 threshold was set to 1.05 (blue line). Relaxed native structures are encircled in red. Results are shown for the proteins also shown in Fig. 4.6. a) 1H4X: Although previously more than 2500 false positive results were present in the single amino acid statistical test, the new T^2 test correctly identifies the three native structures, while producing only 5 false positive structures. b) 1BMW: Only a single false positive result (TM-Score < 0.7) is observed for protein 1BMW. This is in stark contrast to the previous Fig. 4.6b, where no separation between decoys and native structures was visible at all. c) 1ET9: The false positive structures observed in the previous quality assessment of 1ET9 (Fig. 4.6c) are moved above the critical T^2 threshold. No false positive event is generated. d) 1NUK: Similar to the previous quality assessment in Fig. 4.6d, no false positive event is generated. The native conformations exhibit a slightly increased T^2 score by the new metric and are therefore rejected.*

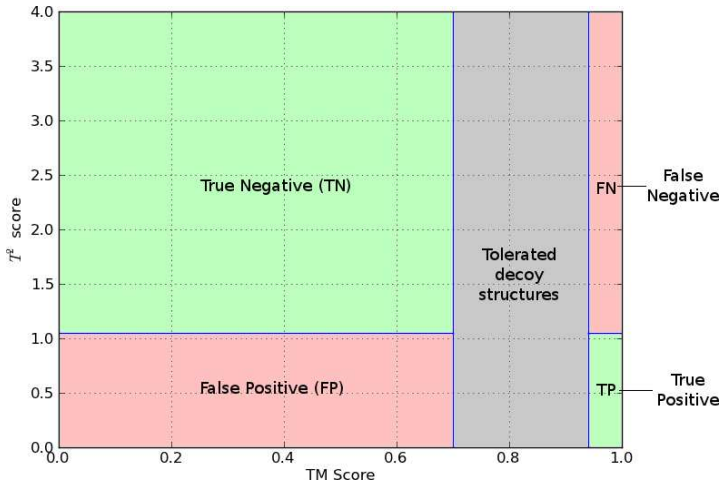


Fig. 4.8.: *Explanation of the variables TN, TP, FN, FP. A True Positive (TP) denotes a correctly identified native structure. A True Negative (TN) denotes a correctly identified low quality model. A False Negative (FN) denotes a misclassified native structure. A False Positive (FP) denotes a misclassified low quality model. Good statistical tests feature data points in the green regions and no data points in the red regions. The True Negative Ratio (TNR) is the ratio of the number of structures observed in the upper left region (TN) divided by the number of all low quality decoys, i.e. the structures in the lower left region (FP) and the upper left region (TN). It denotes the quality of the discrimination of a single decoy set.*

the identification of all low quality models as native models and would therefore indicate a bad statistical test.

In summary using the triplet amino acid quality assessment, native protein structures could be discriminated against low quality decoys. The average TNR for all 160 proteins is $96.0\% \pm 1.0\%$. The distribution of TNR scores for each decoy set (Fig. 4.10) shows that for most decoy sets a perfect TNR score of 100% could be achieved. For these decoy sets, all low quality decoys could be correctly identified and removed from the population. Only 11 decoy sets of the full population of 160 decoy sets feature a TNR of less than $85\,\%$. The distribution of the T^2 values of all structures is shown in Fig. 4.9. Only a small number of bad quality decoys fall below the critical T^2 score and are misclassified. In total 466143 models were investigated for their quality. Of these models, 4%, or 18517 structures, were found to be false positives, which may be worrisome at first. However: the total number of false positives in the 11 decoy populations with a TNR score below 85% already amounted to 13961, i.e. more than 75% of the false positives were generated by less than 7% of all decoy sets, suggesting a systematical error during the analysis of these decoy sets. This is also apparent in Fig. 4.10, as the 11 misclassified populations all exhibited TNR scores far lower than 85% with the smallest TNR score being 24%. Classifications for the eleven proteins, which could not be discriminated using the triplet algorithm, are shown in Tab. 4.1. Except for one protein (Carnobacteriocin B2 - PDB: 1CW5), all proteins, exhibiting a low TNR, bind to another molecule in their active conformation. The training set was based around globular, monomeric proteins. As DNA-binding or chaperone proteins were not explicitly included, it is therefore not surprising that a discrimination using the T^2 statistic will not be accurate for these protein classes. The two metal ion binding proteins (Whiting Parvalbumin - 1A75 and the phosphotransfer domain of Arcb - 1A0B) were not identified by the pruning script, which removed all metal-binding structures from the population prior to the decoy analysis. In summary 10 of these 11 structures should have been removed ini-

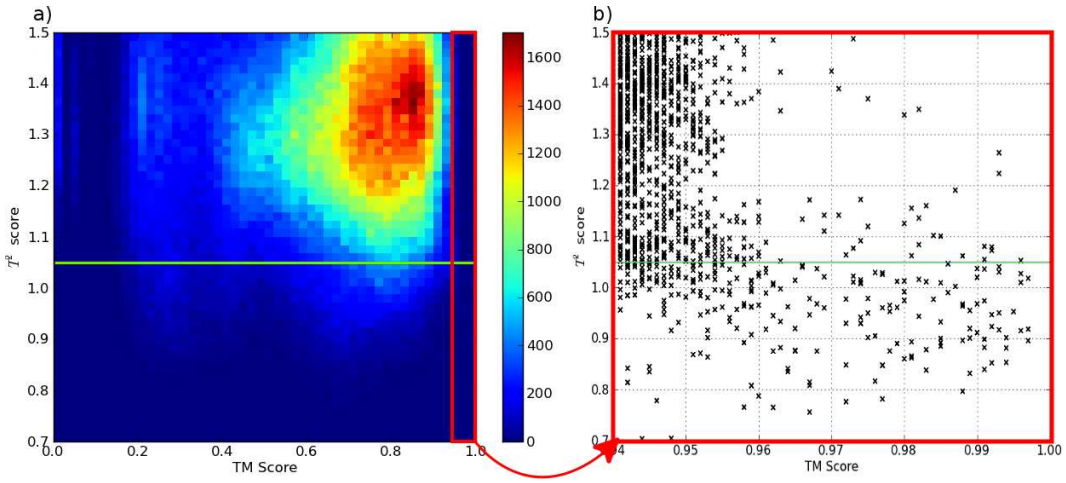


Fig. 4.9.: *Combined quality assessment results of all* 160 *decoy sets. a) Combined results of all quality assessment results for the low and high quality models. The bulk of bad decoy structures (TM-Score below* 0.7*) are situated above the critical* T^2 *value and are therefore recognized. As far more bad decoys were sampled than native structures, the native structures are not visible in this figure. Especially for native-like decoys (TM-Score between* 0.7 *and* 0.94*), structures are observed crossing the critical* T^2 *score. These structures can neither be considered native protein structures, nor bad, as many of the structures lie within experimental resolution or differ due to native protein flexibility. b) Quality assessment results for all high quality protein structures considered native (TM-Score above* 0.94*). For TM-Scores above* 0.95 *most of the protein structures are observed below the critical* T^2 *score and therefore correctly classified as being native. It is expected that many native-like structures lie above the critical* T^2 *threshold: Steric overlap can lead to very large non-physical energies, while only slightly changing the topology. a and b: The green bar denotes the critical* T^2 *below which a model is recognized as native.*

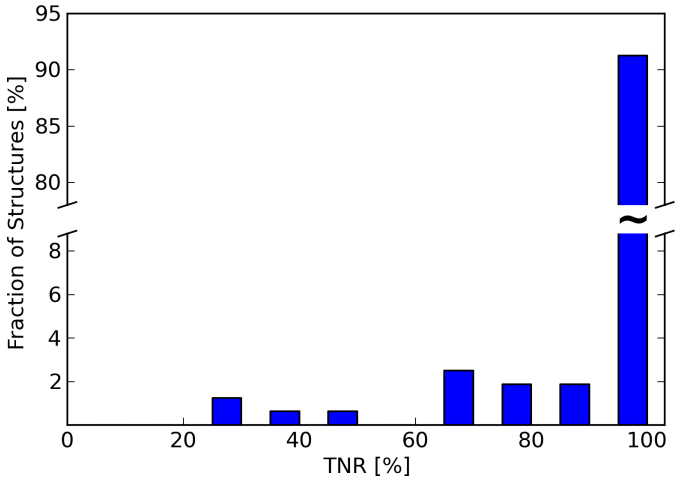


Fig. 4.10.: *True Negative Rates (TNR) for all* 160 *decoy sets. The* y *axis shows the percentage of decoy sets, for which a specific TNR score was achieved. Over* 90% *of the decoy quality assessments exhibit a very high True Negative Ratio (TNR). Only* 11 *decoy sets feature TNR values less than* 85%*.*

PDB	FP	TNR [%]	special properties
1FAF	2285	26	chaperone protein[86]
1ADR	2058	28	DNA binding[87]
1ICH	1817	34	aggregation[88]
1VGH	2583	46	heparin binding[89]
1B4Q	558	64	glutathione binding[90]
1ITP	751	65	chaperone[91]
1A0B	385	65	zinc ion binding[92]
1CW5	1553	66	large unstructured segments[93]
1MEK	488	75	quaternary complex[94]
1BQZ	1013	76	chaperone[95]
1A75	515	78	calcium ion binding

Tab. 4.1.: *Eleven structures with a True Negative Ratio below* 85%*. A large amount of false positive structures was observed for all these structures. Except for 1CW5, all proteins show propensities to bind other proteins or metal ions. As no multimer- or DNA binding- structures were considered in the testset, the discrimination using the* T^2 *test did not succeed. FP: False Positives, TNR: True Negative Ratio*

PDB	TP	FP	TN	FN	TNR	ACC	PDB	TP	FP	TN	FN	TNR	ACC
1JHJ	3	0	645	0	100.0	100.0	1C25	3	0	985	0	100.0	100.0
1IQ3	0	0	3333	0	100.0	100.0	1A29	0	0	2865	0	100.0	100.0
1BFI	0	4	2676	0	99.9	99.9	1B6F	2	0	747	0	100.0	100.0
1HQI	0	0	3183	0	100.0	100.0	1HBK	0	2	2440	0	99.9	99.9
1K3K	0	0	681	0	100.0	100.0	1CLH	0	0	1390	2	100.0	99.9
1I82	2	0	354	225	100.0	61.3	1BM8	2	0	1899	1	100.0	99.9
1GGL	3	0	794	7	100.0	99.1	1JW3	0	0	741	3	100.0	99.6
1CSK	1	41	2492	0	98.4	98.4	1CW5	0	1553	3010	0	66.0	66.0
2ALP	3	0	303	21	100.0	93.6	1QKF	0	67	2816	0	97.7	97.7
1G6Z	0	63	2946	0	97.9	97.9	1JT8	0	0	2544	0	100.0	100.0
1H5P	0	409	2544	0	86.1	86.1	1HDN	3	231	1484	0	86.5	86.6
1BCX	3	0	344	54	100.0	86.5	1DX8	0	11	2830	0	99.6	99.6
1RFS	3	0	1891	0	100.0	100.0	1I7K	2	1	930	0	99.9	99.9
1EWX	3	0	710	0	100.0	100.0	1ODD	1	40	1936	0	98.0	98.0
1BQZ	0	1013	3163	0	75.7	75.7	1KRS	0	6	2474	0	99.8	99.8
1G2R	0	16	2488	0	99.4	99.4	1A9V	0	0	1478	3	100.0	99.8
1QFT	0	0	550	3	100.0	99.5	1I2U	0	0	4803	0	100.0	100.0
1FJR	0	0	840	0	100.0	100.0	1DUJ	0	0	283	42	100.0	87.1
1H4X	2	5	2584	0	99.8	99.8	1I17	0	33	2832	0	98.8	98.8
1A7I	0	89	2464	0	96.5	96.5	1FHS	0	1	2515	0	100.0	100.0
1KQR	3	0	525	100	100.0	84.1	1E68	0	4	2791	0	99.9	99.9
1GMM	3	0	763	0	100.0	100.0	1FAF	0	2285	820	0	26.4	26.4
1A7H	3	5	2118	0	99.8	99.8	1NUK	0	0	2250	3	100.0	99.9
2NCM	3	0	1912	0	100.0	100.0	1VIB	0	0	4233	0	100.0	100.0
1ET9	3	0	2727	0	100.0	100.0	1GNY	0	0	626	23	100.0	96.5
1B2T	0	0	2979	0	100.0	100.0	1AEY	1	244	2414	0	90.8	90.8
2AFP	0	0	2210	0	100.0	100.0	1BOE	0	659	4134	0	86.3	86.3
1MUT	0	0	1295	3	100.0	99.8	1BOR	0	2	2437	0	99.9	99.9
1DBW	0	0	783	3	100.0	99.6	1CLF	0	3	2673	0	99.9	99.9
1ITP	0	751	1412	0	65.3	65.3	1AXJ	0	0	1512	3	100.0	99.8
1DG4	3	1	1998	0	99.9	100.0	1I8N	2	36	2109	1	98.3	98.3
1DAX	0	0	2512	0	100.0	100.0	1BMW	1	24	2331	0	99.0	99.0
1EHX	0	58	2242	0	97.5	97.5	1G28	3	0	2502	0	100.0	100.0
1HYK	0	0	4802	0	100.0	100.0	1DBY	3	16	1612	0	99.0	99.0
1G9P	0	234	4569	0	95.1	95.1	1FW9	3	0	436	53	100.0	89.2
1EM9	2	27	667	0	96.1	96.1	1JKZ	0	353	3860	0	91.6	91.6
1EWI	0	0	2573	0	100.0	100.0	1PFT	0	4	4757	0	99.9	99.9

PDB	TP	FP	TN	FN	TNR	ACC	PDB	TP	FP	TN	FN	TNR	ACC
1CMO	0	0	2290	0	100.0	100.0	1MEK	0	488	1436	0	74.6	74.6
1BVH	0	1	1146	4	99.9	99.6	1A0B	3	385	726	2	65.3	65.3
1G7E	1	1	1644	0	99.9	99.9	1HPW	0	2	2785	0	99.9	99.9
1JRM	0	16	2356	0	99.3	99.3	1B6E	0	0	1894	0	100.0	100.0
1ADR	0	2058	781	0	27.5	27.5	1AGY	3	0	304	25	100.0	92.5
1PIH	0	58	2518	0	97.7	97.7	1H6H	0	0	1212	0	100.0	100.0
1CL3	0	0	1275	1	100.0	99.9	1HKS	0	1	2981	0	100.0	100.0
1DNY	0	54	3361	0	98.4	98.4	1MKN	0	0	3684	0	100.0	100.0
1A44	3	0	831	0	100.0	100.0	1BS4	3	0	702	2	100.0	99.7
1BJX	0	0	2814	0	100.0	100.0	1I27	0	114	3374	0	96.7	96.7
1AAZ	1	130	2564	1	95.2	95.1	1VHR	3	4	493	3	99.2	98.6
1CTO	0	0	2504	0	100.0	100.0	1A75	2	515	1853	0	78.3	78.3
1MNL	0	5	2616	0	99.8	99.8	3CRD	0	0	2512	0	100.0	100.0
1H8U	0	0	1263	3	100.0	99.8	1LIH	1	9	665	0	98.7	98.7
1H75	0	247	2620	0	91.4	91.4	1F0Z	0	27	2712	0	99.0	99.0
1E5K	3	0	321	14	100.0	95.9	1B9G	0	27	3012	0	99.1	99.1
1AW0	3	182	2443	0	93.1	93.1	1E8R	0	115	2297	0	95.2	95.2
1B00	3	7	933	0	99.3	99.3	2PNA	0	0	2963	0	100.0	100.0
1IQQ	3	0	522	1	100.0	99.8	1XNA	3	0	891	0	100.0	100.0
1KXL	0	0	671	3	100.0	99.6	1A3K	2	0	686	1	100.0	99.9
1DT4	1	1	3134	1	100.0	99.9	1IJT	2	0	1215	1	100.0	99.9
3NCM	3	0	1868	0	100.0	100.0	1EUJ	3	0	504	22	100.0	95.8
1ICH	0	1817	934	0	34.0	34.0	1G12	7	3	356	7	99.2	97.3
1QLC	2	0	2226	0	100.0	100.0	1HX2	0	210	3204	0	93.8	93.8
1EHD	0	23	2618	0	99.1	99.1	1K5W	0	0	818	5	100.0	99.4
1VGH	0	2583	2220	0	46.2	46.2	1B9W	0	2	2227	0	99.9	99.9
1DZ7	0	0	2364	0	100.0	100.0	1IDY	0	13	3283	0	99.6	99.6
1AO3	3	0	106	132	100.0	45.2	1B75	0	0	3056	3	100.0	99.9
1MUP	0	0	640	3	100.0	99.5	1CDQ	0	24	2675	0	99.1	99.1
1GH8	0	0	2335	0	100.0	100.0	1K8M	0	0	2110	0	100.0	100.0
1B4Q	3	558	1009	0	64.4	64.5	1AHL	0	31	2608	0	98.8	98.8
1JGK	0	2	4798	0	100.0	100.0	1QMY	2	0	568	3	100.0	99.5
2FNB	0	0	2232	2	100.0	99.9	1AAL	0	1	2792	0	100.0	100.0
1QND	0	48	2419	0	98.1	98.1	1CQQ	3	0	407	8	100.0	98.1
1WIT	0	0	1616	3	100.0	99.8	1G1K	3	0	689	137	100.0	83.5
1JBI	3	0	2340	0	100.0	100.0	1AC0	0	0	2177	0	100.0	100.0
1FGY	1	0	2667	0	100.0	100.0	1EQK	0	0	2575	0	100.0	100.0
1B6B	1	0	633	2	100.0	99.7	1GPS	0	53	4382	0	98.8	98.8
1JH3	1	196	2416	0	92.5	92.5	1EWW	0	56	2390	0	97.7	97.7
1AAC	2	0	1027	3	100.0	99.7	1JF8	3	11	1056	0	99.0	99.0
1TIF	0	28	2444	0	98.9	98.9	1J8K	0	0	2496	0	100.0	100.0
1THX	3	35	1155	0	97.1	97.1	1YUF	0	84	4719	0	98.3	98.3
1COL	106	0	4	395	100.0	21.8	1BBN	0	1	958	1	99.9	99.8

Tab. 4.2.: *Results of the quality assessment of all* 160 *decoy structures using the triplet amino acid statistics. TP: True Positives, FP: False Positives, AP: Actual Positives within the population, PREC: Fraction of positives observed (set to* 100 *if AP=*0*), TN: True Negatives, FN: False Negatives, AN: Actual Negatives within the population, TNR: True Negative Ratio, TNR=TN/(TN+FP), ACC: Accuracy. ACC = (TP + TN)/(TP+TN+FP+FN)*

tially, prior to the testing benchmark. The initial selection process removed all proteins from the test and training sets, which were classified as membrane or multimeric proteins in the PDBML metadata available for every experimental protein structure in the RCSB database[3]. Our selection process did not detect structures, where the binding propensities to other proteins were only recorded in the manuscript usually deployed in conjunction with the experimental structure. This is very often the case for chaperone and DNA binding proteins, which are resolved in their

isolated form. Except for the 11 decoy sets, the triplet test could reliably remove the low quality decoys from the population of all models. The results of the quality assessment of all decoy sets are shown in Tab. 4.2.

4.4. Discussion

In this investigation we developed a method for absolute quality assessment of globular protein models. We first derived a N-dimensional statistical test, based on amino acid specific energies e_i and assembled statistics of the required energies for a representative high-resolution set of globular protein structures. Three test methods were developed using single amino acid energies and neighboring amino-acid pairs and triplets. While the single amino acid test was not sensitive enough to discriminate between native and decoy structures, the triplet amino acid test could deliver correct results for 149 of 160 structures with a average True Negative Ratio (TNR) of 96.0%. For most of the globular protein structures, native structures were observed very close to the critical 95% probability threshold of the T^2 test. Ten of the eleven proteins with false positives in the testset were shown to bind other proteins, DNA or other cofactors in their native conformation and should have therefore been removed from the testset.
The triplet test could reliably discriminate low quality protein models against the native protein structure. Theoretically an extension to four, five or more amino acids might increase the accuracy of the proposed method even further; practically not enough experimentally resolved PDB structures are available to prepare high-dimensional distributions of sufficient quality.
The success of fragment-based methods in protein structure prediction indicates that sequence fragments of size longer than three contain sufficient structural information to build protein models de-novo[96, 97]. These fragments are local in sequence, i.e. only three sequential amino acids are considered. Our method also considers non-local motifs ($3\,AA$) for protein quality assessment. This may indicate that also non-local motifs contain sufficient information to identify the correct tertiary fold. This structural information could therefore also be used in structure prediction methods akin to the Rosetta fragment assembly protocol and allow the prediction of non-local contacts.

5. Protein Structure Prediction

Biochemical machinery of all cellular life are made of proteins. Although the genome of several species has been completely sequenced[1, 98], a large gap exists between the number of about $80,000$ structurally resolved proteins[3] and that of millions of known protein-coding sequences[72]. This discrepancy results from the high cost of experimental methods for protein structure determination and the difficulty to prepare entire classes of proteins, such as transmembrane proteins, for analysis with experimental techniques. Structural insight is enormously helpful to analyze a protein's function and possibly to modulate its activity in pharmaceutical research. The ability to predict a protein's three-dimensional structure from the sequence alone promises to yield a wealth of biomedical information.
In this chapter, we apply and develop methods for protein structure prediction to various biological systems and investigate their aggregation. In section 5.1, we briefly discuss two methods for protein structure prediction. In section 5.2, we apply these methods to predict structures of hydrophobin proteins for functionalization as efficient and biocompatible coating of implants[99]. In section 5.3, we investigate how hydrophobins develop stable structures on air-water interfaces. In section 5.4, we investigate the mechanism of gas-vesicle formation in aqueous bacteria and provide the first nano-scale structural model of the main constituent of the gas-vesicle wall[100]. We conclude the chapter in section 5.5, where we present the development of a high-throughput technique for the structure prediction of peptides[101].

5.1. Knowledge-based protein structure prediction

The experimental determination of protein structures is a very difficult and resource-intensive task. The simulation of the folding process of a protein from an initial random-coil structure to the functional tertiary assembly would theoretically elucidate the protein's structure, practically the theoretical characterization of the folding process has been accomplished for a few, small proteins to date[6]. If only the tertiary structure of the protein in its native state is required, knowledge-based methods can sometimes build an adequate model. The quality of protein structure prediction methods is regularly assessed in the bi-annual CASP competition[2]. Methods used in this thesis include homology modeling[78] and fragment assembly[80].

Homology Modeling: Protein sequences of functional proteins are the product of millions of years of evolution. As such, many families of protein sequences exist, which encode similar 3D structures (similar ancestors), but differ in the primary sequence. This hereditary relationship is called *Homology*[102]. Homology modeling uses this relationship to infer structural information about the target protein, i.e. the protein of unknown structure, by analyzing the template

protein, i.e. a protein homologous to the target sequence with an experimentally resolved 3D - structure.
Homology modelling is usually separated into three stages:

- Template Search
  The number of experimentally known protein structures has surpassed $80,000$ structures in 2012[3]. Of these $80,000$ structures about $12,500$ are non-redundant (Data from the Vast database as of 2012, clustering at a p-value of 10^{-7}[103]). In the template search step, one uses a scoring function between these non-redundant sequences and the input sequence to find a homologous protein. Template search algorithms usually approximate alignment algorithms explained in the next step to construct scoring functions.

- Alignment
  For each template an optimal alignment is constructed, which is rated by a scoring function. The minimal requirement for a scoring function is that it needs to align the sequence with itself with a maximum score. Alignments of different amino acids with each other are often penalized by scores in a substitution matrix depending on the compatibility of their physical properties. One such choice for a substitution matrix is BLOSUM[104], which is derived from the naturally occurring mutations observed within a family of protein structures. Gaps in the alignment are allowed but also penalized by a gap-penalty. Two popular alignment algorithms are the Smith-Waterman[105] and Needleman-Wunsch[106] algorithms for local and global alignment respectively.

- Model construction
  Model information (backbone dihedrals, tertiary contacts) is extracted from the template structure and applied to the target chain. Protein segments, which could not be aligned to the template will be unstructured and need to be reconstructed, for example using Monte-Carlo or Molecular Dynamics simulations. Experience shows that about 40% sequence similarity is sufficient to build an adequate structure[73]. Modeller is one of the most well known programs implementing this strategy[76].

Fragment Based Modeling: When no homologous protein can be discovered, shorter alignments of part of the target sequence can yield structural fragments. Rosetta[80] uses fragments of fixed length of 3 and 9 amino acids and inserts them using a Monte-Carlo procedure. Fragments are selected not only according to their sequence similarity with the predicted sequence, but also based on the predicted secondary structure of the target sequence. Fragments similar in secondary structure to the prediction are selected with a higher probability. The final predicted structure is then optimized based on a physical or empirical scoring function. This approach was benchmarked in multiple occasions by predicting proteins without homologous templates during the CASP benchmark[96, 97].
Numerous other threading algorithms exist, which introduce longer alignments and thread them together to form a prediction. Notably Zhang et al.[107] were successful with this approach by using the constructed model structure to rate the initial alignment. If the model could show similarity with the alignments used to build it, a high score was assigned to the model.

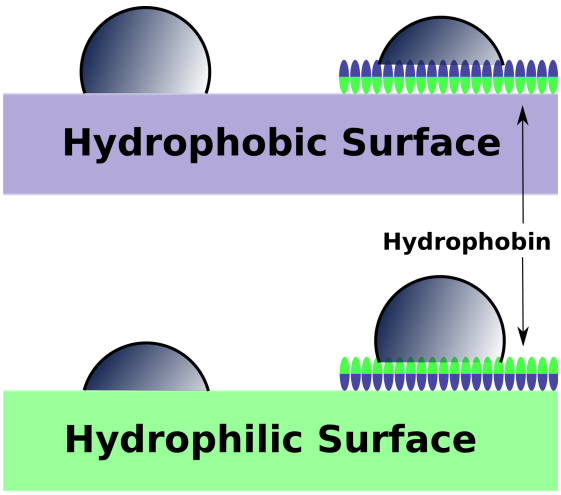


Fig. 5.1.: *Hydrophobins show the ability to reverse hydrophobicity of materials by coating. The scheme shows a high contact angle for water droplets on hydrophobic materials and a low contact angle for droplets on hydrophilic materials. Coating with hydrophobin reverses this relationship. This property can be used to convert the hydrophobicity of hydrophilic materials (for example implants) hydrophobic and therefore prevent biofilm formation.*

5.2. Design of genetically engineered variants of hydrophobin DewA

5.2.1. Motivation

The project presented in this section, describes an investigation to create a novel biocompatible coating material for cell adhesion of stem-cells, while preventing biofilm formation. Biocompatible surfaces possess a multitude of different applications in the medical sciences[108, 109]. They are topics of interest especially in the fields of tissue engineering and medical implants. If a material is used, which allows the attachment of stem cells and their further differentiation into bone stem cells (for example Mesenchymal stem cells - MSCs, which are present in the bone marrow), bone can grow around the implant and therefore stabilize it in the body[110, 111].
At the same time bacterial adhesion has to be prevented, as biofilm development and inflammation can lead to infection, which can be fatal[112]. In less severe cases it is often necessary to remove the implant. Functionalizing a surface to bind stem cells therefore must not facilitate bacterial adhesion[110].
In this project we investigated, whether we can modify DewA, a protein of the hydrophobin protein family, to exhibit both impaired biofilm growth and increased cell binding on hydrophobin covered surfaces compared to untreated implant surfaces. Unmodified hydrophobins have been shown to suppress the immune response[113] and are known to be biocompatible. Further they form hydrophobic-hydrophilic layers at air-water interfaces (Fig 5.1). Due to their hydrophobicity they were a likely target to inhibit biofilm formation. We therefore genetically modified hydrophobin DewA, for which high-throughput synthesis methods had been developed, to also enhance cell-binding. This project was carried out as a joint project with the experimental groups of Prof. Fischer and Prof. Schimmel from the KIT and Prof. Richter from Universitätsklinik Heidelberg funded by the Landesstiftung Baden-Württemberg[1]

[1]Parts of the following text have been published in Boeuf et al.[99] courtesy of Elsevier.

- X_{26-85}-C-X_{5-8}-C-C-X_{17-39}-C-X_{8-23}-C-X_{5-6}-C-C-X_{6-18}-C-X_{2-3} Class-I
- X_{17-67}-C-X_{9-10}-C-C-X_{11} -C-X_{16} -C-X_{6-9}-C-C-X_{10} -C-X_{3-7} Class-II[118].

Fig. 5.2.: *Hydrophobins can be grouped into two different classes by their characteristic Cysteine (C) pattern. An X_{i-j} in the above motif denotes a series of random amino acids (bar Cysteine) of a minimal length i up to a maximum length of j.*

This section is structured as follows: In section 5.2.2 I discuss the general properties of hydrophobin proteins and afterwards introduce the specifics of hydrophobin DewA in section 5.2.3. After presenting the used methods for model generation of DewA (section 5.2.4 and section 5.2.5) and the synthesized mutants (section 5.2.6), the models will be analyzed in section 5.2.6. The experimental binding efficiency of the proposed mutants is reported in section 5.2.7 and further set into context in the final discussion of the results in section 5.2.8.

5.2.2. Introduction - The family of hydrophobin proteins

The hydrophobin protein family consists of small (80 - 200 amino acid) amphipathic proteins structurally defined by a unique arrangement of 8 Cysteine residues forming 4 disulfide bonds[114]. Proteins are classified into the hydrophobin family, because of similar hydropathicity patterns around the characteristic Cysteine residues. Hydrophobins are fungal proteins with the ability to form amphipathic membranes by self-assembling at hydrophilic-hydrophobic interfaces. If these membranes are formed on a solid surface, they can invert the hydrophobicity of this surface. Janssen et al. could show that a Teflon surface can be made hydrophilic by the addition of a hydrophobin layer[115].
The assembly of class I hydrophobins is associated with the formation of amyloid fibrils[116] and assembled class I hydrophobins bind strongly to their supports, resisting harsh treatments such as boiling using water or detergents[117]. Two distinct hydropathicity patterns are recognized resulting in the classification of hydrophobins into two classes as shown in Fig. 5.2. Although both classes share the amphipathic nature of the proteins, some class-I hydrophobins are considered more stable, as they form distinct insoluble rodlets[119]. In contrast class-II hydrophobins are water soluble and adhere to the Cysteine spacing more strictly. Both families share the same Cysteine pairing:

$$C_1 - C_6, C_2 - C_5, C_3 - C_4, C_7 - C_8.$$

Various loops are attached to the hydrophilic base, a β-barrel fixed by the Cysteine residues prevalent throughout the complete hydrophobin family. In context of the structure determination of Class-I hydrophobin EAS[120] it was hypothesized that the rodlet formation is mediated by a rise in β-content of loop $C_3 - C_4$, which would result in an extended β-barrel. Further investigations[121] could show that not only is the loop unimportant in the formation of rodlets, but that the truncated mutant also exhibited more distinct rodlets, while retaining its amphipathic character.
Hydrophobins can be observed on the surface of fungal spores and tissues of fruiting bodies that have been exposed to air[122, 123]. Among other features a hydrophobin coating enables

the spore's dispersal in the air by creating an affinity with hydrophobic surfaces. Additionally it could be shown that the hydrophobin RodA from Aspergillus fumigatus is immunologically inert; i.e. it does not create an apparent immune-response by activating helper T-cells[113]. Due to their non-immunogenic nature, hydrophobins are a viable choice to change surface properties in medical applications[117].

5.2.3. Functionalization of DewA – Aspergillus Nidulans

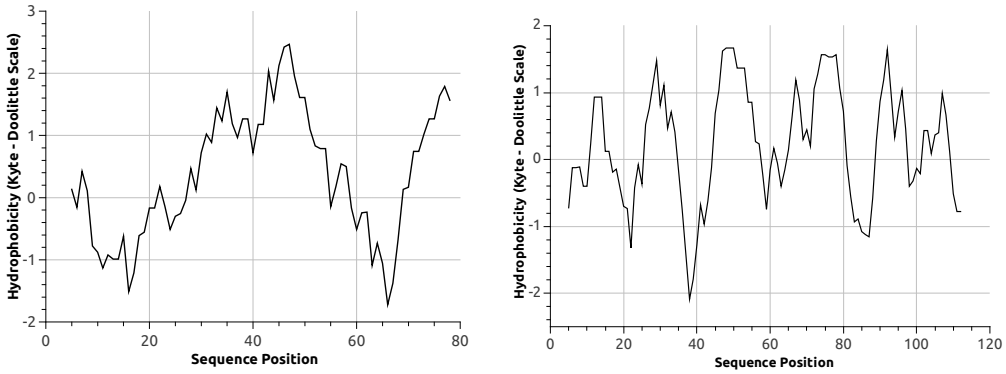
Here we focused on the hydrophobin DewA for which large-scale production has been achieved [124]. DewA is a class-I hydrophobin located in the sporewall of Aspergillus Nidulans[125]. For functionalization we wanted to integrate protein sequence motifs known to bind cells into appropriate places of the hydrophobin protein structure. We introduced two known binding motifs to increase cell-adhesion: the RGD (ARG GLY ASP) site of fibronectin[126, 127] and the laminin globular domain LG3[128].
In a previous study Janssen et al. fused an RGD peptide into the sequence of the SC3 hydrophobin from Schizophyllum commune, which promoted growth of fibroblasts on a hydrophobic solid[115]. Although the results were very promising the major downside of functionalizing SC3 is that currently SC3 cannot be mass-produced in E.Coli and has to be collected from the mushroom.
We used all-atom molecular modelling to predict suitable insertion sites for RGD or LG3 motives at surface-accessible sites in the engineered A. nidulans DewA molecule, and used these purified proteins to produce hydrophobin surfaces that enhance adhesion of human cells.

5.2.4. Homology Modeling of DewA

The Cysteine pattern of DewA matches with the class-I hydrophobin pattern:

$$X_{24}\text{-C-}X_{6}\text{-C-C-}X_{31}\text{-C-}X_{23}\text{-C-}X_{5}\text{-C-C-}X_{6}\text{-C-}X_{13}.$$

DewA exhibits a hydropathicity plot unlike most class-I hydrophobins (Fig. 5.3). In contrast to the class-I hydrophobin EAS, DewA does not form rodlets[125]
No experimental structure is available for DewA, which makes modelling a necessity. As no direct sequence similarity to a known protein could be observed, a motif conserving alignment (Fig. 5.4) with the only structurally resolved class-I hydrophobin EAS (PDB: 2FMC) is prepared to build a 3D-model using comparative modeling with Modeller[130].
Although the alignment has various big gaps, the model quality can still be adequate as it is dependent on the generation of the hydrophilic β-barrel: Most of the protein's structure is suspected to be in disordered loops that do not possess a distinct structure.
The alignment was used to build comparative models using the standard Modeller single alignment protocol[130] (Alignment in Fig. 5.4, Starting Model in Fig. 5.5 (a)). Although the models did not exhibit secondary structure elements the β-elements of the chain were aligned and present in the model. In addition the correct disulfide pairing is present in the model.



(a) EAS (b) DewA

Fig. 5.3.: *Hydropathicity patterns for class-I hydrophobins EAS and DewA. The frequent variation in local hydrophobicity of DewA differs from most class-I hydrophobins. The hydrophobicity shown in the plot is averaged over a window of 9 amino acids. The Kyte-Doolittle Scale was used for the y-Axis[129]. In this scale ARG has a value of -4.5 (hydrophilic) and Isoleucine has a value of $+4.5$ (high hydrophobicity). A large hydrophobic patch is observed in protein EAS. While also DewA features many hydrophobic sections, they are separated by short hydrophilic sections.*

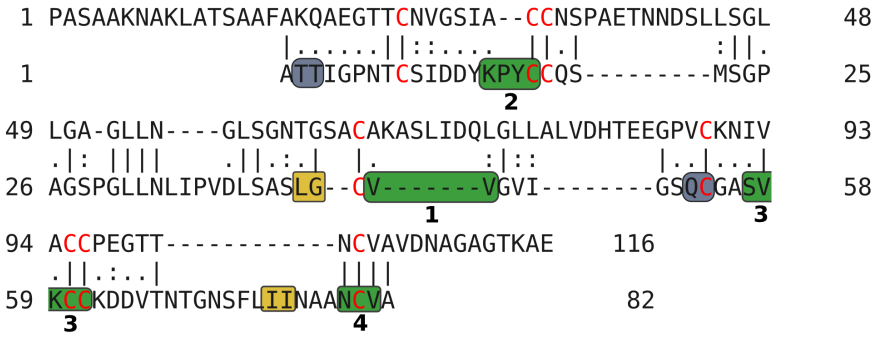
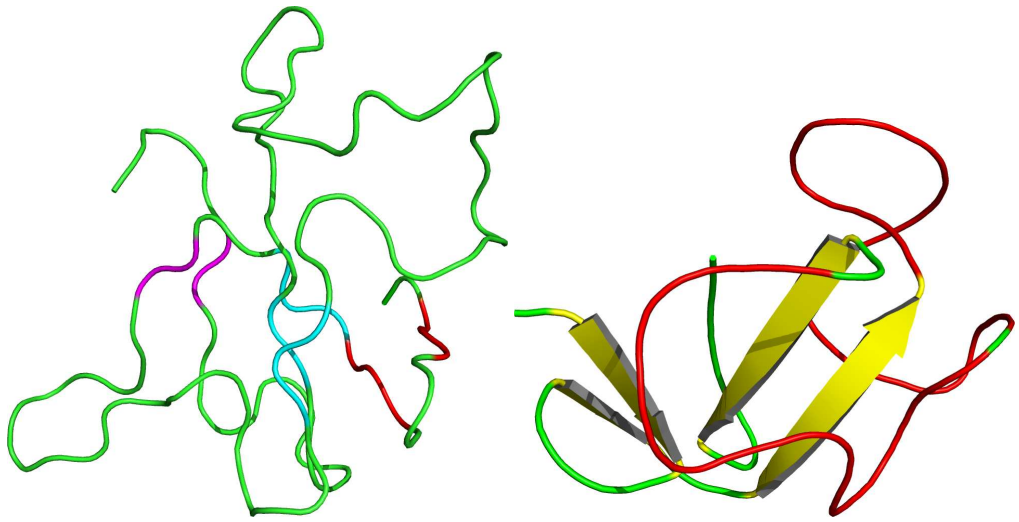


Fig. 5.4.: *Alignment between DewA and EAS (PDB: 2FMC). The location of the 8 Cysteine residues is conserved to permit the formation of a β-barrel in the molecular modeling simulations. Similar colored intervals represent paired β-strands in the template structure. One β-sheet is formed of four β-strands (green color). The numbers below the strands represent the bonding order: Strands with consecutive numbers are bound towards each other. β-sheets were classified according to the experimental structure[131].*



(a) starting Model, paired β-strands in the template are displayed in the same color

(b) refined Model, the two red loops are hypothesized to be unstructured in the globular structure

Fig. 5.5.: *Models obtained by comparative modeling using the alignment in Fig. 5.4. a) The resulting model is missing all secondary structure; the observed Cysteine-pairing corresponds to the one of Class-I hydrophobins. b) The refined model features a developed β-barrel. Unstructured loops in the first model a) are collapsed into a more compact structure.*

Pop. #	1	2	3	4	5
Temp. $[K]$	1176	676	388	223	128
Memb.	50				
Accepted	2407	2911	3142	3408	3345
Rejected	195265	272391	232733	270792	192078
Total	197672	275302	235875	274200	195423
Ratio	1.2%	1.1%	1.3%	1.2%	1.7%

Tab. 5.1.: *Statistics from the mEA relaxation simulation of hydrophobin DewA. The temperature values between the populations are geometrically scaled. The low acceptance probabilities are explained by the fact that the input structure was already near convergence and the large number of simulations.*

5.2.5. All-atom structural refinement of the DewA model

The local secondary structure is obtained by refinement simulations using the all-atom free-energy forcefield PFF02 in SIMONA with the multiple population evolutionary algorithm setup with the parametrization shown in Tab. 5.1. Probabilities to switch temperature were set to $1/3$. The RMSD and energy ranges sampled by this mEA configuration are shown in Fig. 5.6. While at high temperature only one cluster of structures can be observed, this cluster breaks down into substructures, which are separated by barriers at low temperatures. The lowest energy structure features a RMSD of 7 Å to the initial conformation.

In our simulations we observe a stabilization of the β-content. Three of the four proposed β-pairings (2: G28 - CYS32 3: ILE92-CYS96 4: ASP102-ALA105, compare Fig. 5.4) develop further and stay stable, while the fourth one (1: ALA66 - LEU74) is destroyed due to fluctuations (Fig. 5.5 (b)). This is in agreement with CD-spectra of EAS, which show a rise in β-content either due to the regularization of β-structure or the development of β-structure in the loops upon rodlet formation[120]. As the third β-sheet lies within the proposed rodlet binding interface of

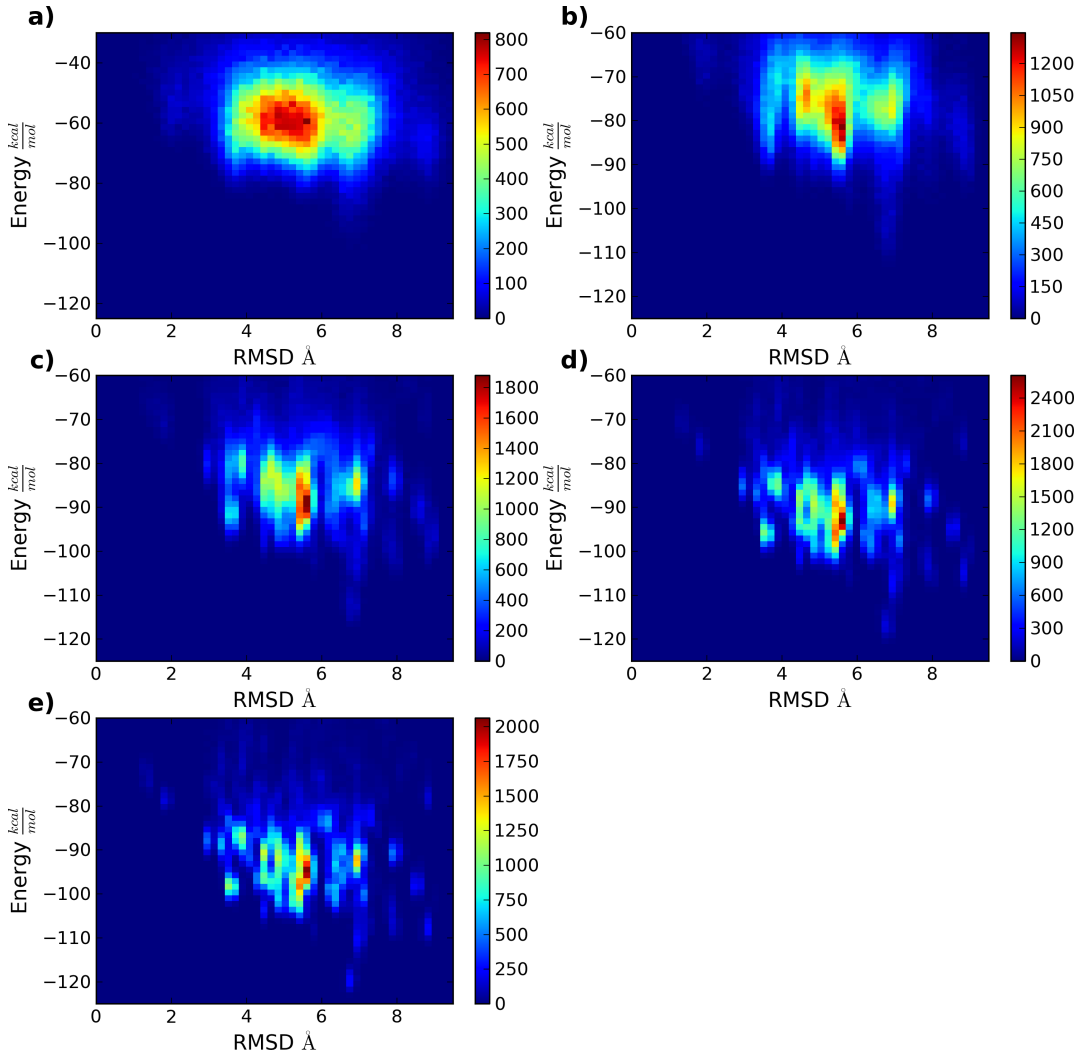


Fig. 5.6.: *Energy and RMSD statistics of the five mEA populations of protein DewA. RMSD is plotted against the initial structure. Temperatures: a)* $1176\,K$ *b)* $676\,K$ *c)* $388\,K$ *d)* $223\,K$ *e)* $128\,K$*. The lowest energy structure has a RMSD of* $7\,Å$ *to the initial structure. From high to low temperature it can be seen that tempering in multiple temperatures permits the protein to hop barriers by transitioning to different temperature populations: The single high-temperature cluster of structures seen in a) gradually separates into substructures separated by barriers, when moving to lower temperatures*

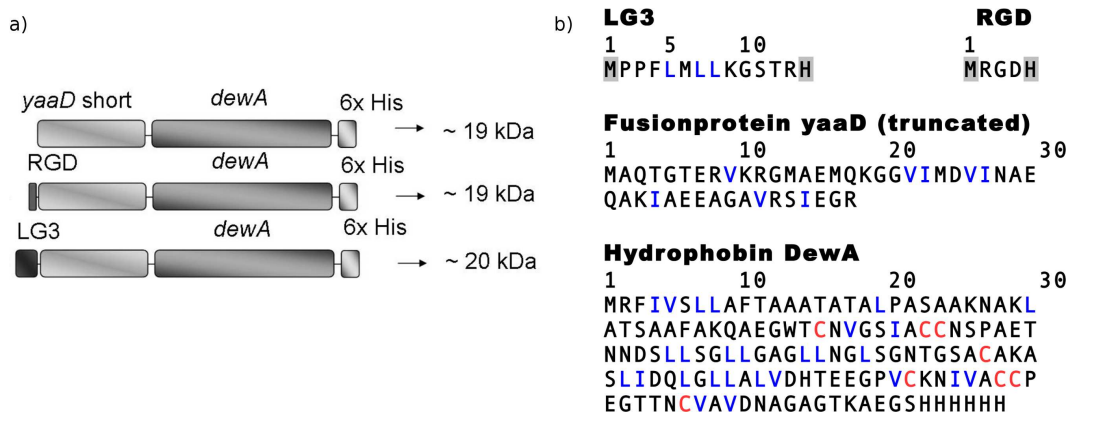


Fig. 5.7.: *Proposed sequences for increased cell-adhesion. Mass production of DewA required the fusion of a shortened variant of protein yaaD to the hydrophobin. LG3 and RGD sequences were attached to the N-Terminal of the fusion protein.*

hydrophobin DewA, it is possible that it is not stable in isolated structures.

5.2.6. Structure-based design of genetically modified DewA

In order to engineer cell-adhesive DewA surfaces, it was necessary to design genetically modified DewA variants. Among peptide sequences promoting cell adhesion, the RGD sequence (DewA-RGD) and a 12 amino acid long LG3 sequence (DewA-LG3) were selected (Fig. 5.7). For the large-scale expression of DewA in Escherichia Coli the protein yaaD[132] needed to be fused using a linker to protein DewA. These two modifications resulted in the sequences shown in Fig. 5.7.

To generate a model of the DewA-Fusion construct, the structure of the truncated yaaD protein was required. Fortunately, an experimentally resolved structure for the full yaaD protein exists[132].

As shown in Fig. 5.8 the shortened part of yaaD represents only a small fraction of the whole protein. In order to generate an unbiased ensemble of models, $20,000$ structures for each yaaD variant containing either the LG3 or RGD motif were prepared using the Rosetta 3 fragment assembly protocol and relaxed afterwards in SIMONA twice for $500,000$ steps from $200\,K$ to $50\,K$. Dihedral angles were perturbed randomly by drawing them from a Gaussian distribution with a width of 10° around the current angle. The complete DewA-fusion protein complex was then modeled comprising the existing models for DewA and yaaD with modeler (Fig. 5.10).

Model evaluation Analysis of the models shown in Fig. 5.10 demonstrated that the protein comprises two distinct domains (DewA-RGD: hydrophobin domain MET52-VAL176, fusion-interaction domain MET1-ARG51). The hydrophobin domain exhibited the same characteristic beta-barrel structure stabilized by cysteine-bridges as before (CYS99-CYS165, CYS102-CYS159, CYS103-CYS135, CYS166-CYS173 for DewA-RGD) and four β-sheets (ILE100-CYS103, CYS135-LYS137, ILE162-CYS165, CYS173-VAL176 for DewA-RGD), while the binding domains formed an unstructured conformational ensemble at the N-terminus of the

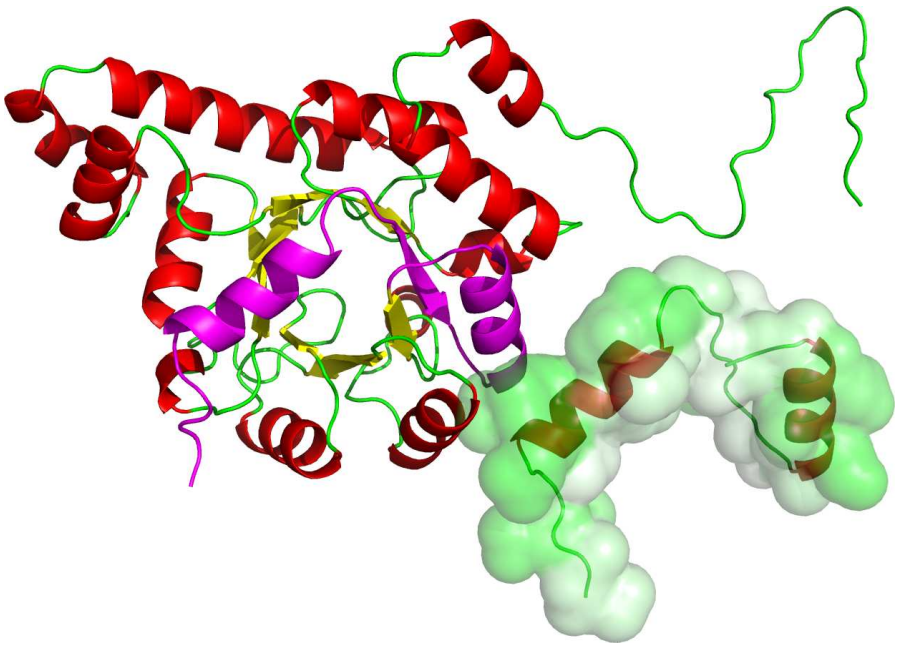
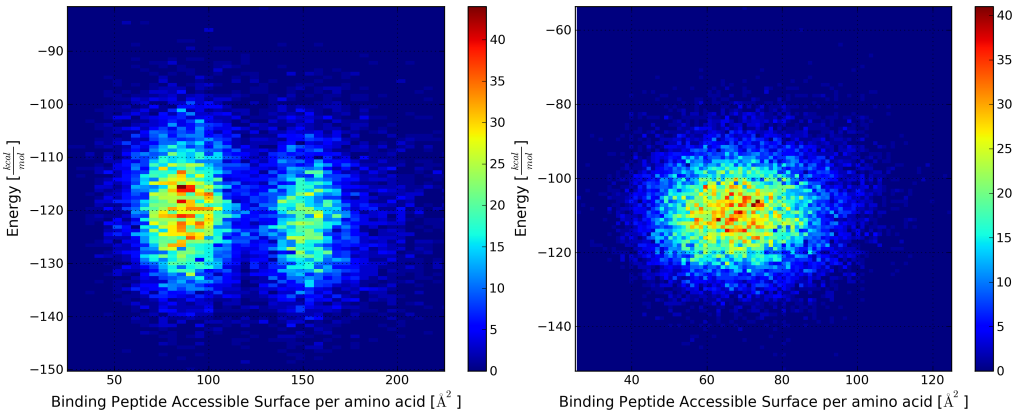
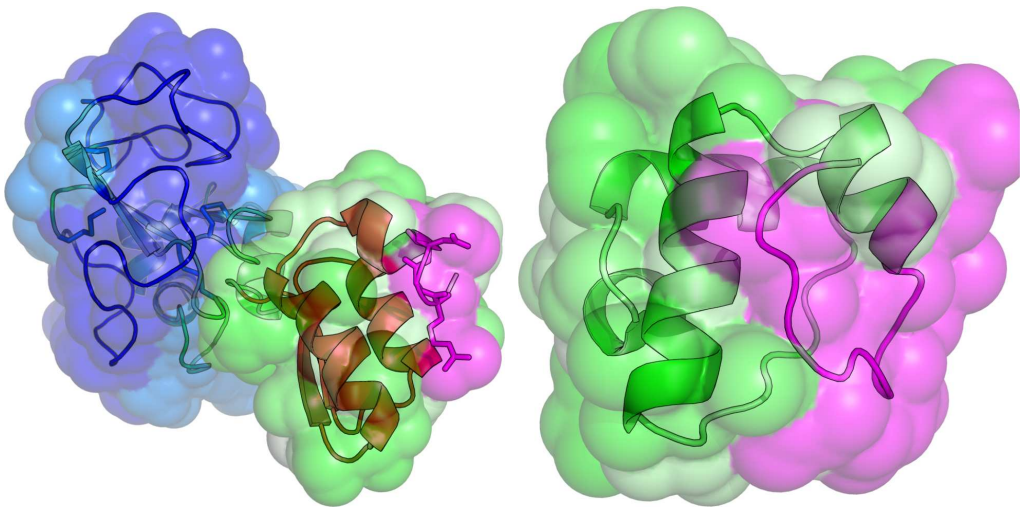


Fig. 5.8.: *Experimental structure of the full fusion protein yaaD. The helix bundle (shown in magenta in the full structure) is the only remaining part in the truncated yaaD variant. This part is also seen isolated in the lower right. As most structural features of the full protein are truncated, the remaining part will refold into a new conformation unlike the experimental structure. Molecular modeling has to therefore start ab-initio.*



(a) RGD (b) LG3

Fig. 5.9.: *Solvent Exposure of LG3 and RGD motifs in Fusion-Peptide constructs. Both sequence motifs feature significant solvent exposure per amino-acid. a) The RGD solvent motif features two preferred conformations. While the lowest energy structure features a solvent exposure of* $80\,\mathring{A}^2$ *per amino acid, a second cluster exists with a mean solvent exposure of* $150\,\mathring{A}^2$ *per amino acid. b) Only one cluster of structures was observed for the LG3 motif. Members of this cluster presented a mean accessible surface area of* $80\,\mathring{A}^2$.



(a) RGD (b) LG3

Fig. 5.10.: *Lowest energy structures of both fusion constructs of RGD and LG3. a) Cartoon and surface visualization of the lowest energy model of the full engineered protein variant fused with RGD. The model features two distinct domains: the hydrophobin domain MET52-VAL176 (light blue) and the hydrophobic loops (dark blue); and the fusion-interaction domain MET1-ARG51 (green) with the RGD (magenta) b) Cartoon and surface visualization of the lowest energy model of construct LG3 (magenta) fused with yaaD (MET1-ARG60, green). The DewA domain (MET61-VAL185) is not shown. Tight packing towards the fusion part of the protein can be explained by the big amount of hydrophobic amino acids present in the LG3 motif*

fusion-protein domain. Additionally, two large unstructured hydrophobic loops were identified, which give the hydrophobin its characteristic amphipathicity (SER105-SER133 and LYS137-PRO157 for DewA-RGD).

The resulting models were analyzed for the exposure of the cell-binding motifs RGD and LG3. Two populations of different solvent exposure of the RGD motif were identified among simulated ensembles seen in 5.9 a). Most structures in the ensemble exhibited tightly packed helical folds for the truncated fusion protein domain and no apparent secondary structure for the RGD motif. While the lowest energy structure of the complex featured a solvent exposure of 80 Å^2 per amino acid of the motif, the second cluster, with a mean solvent exposure of 150 Å^2 per amino acid, was separated by less than $1\ kcal/mol$. Models from both clusters showed the RGD motif to be exposed to the solvent. Fig. 5.9 b) illustrates that the lowest energy model for the RGD motif lies in plane with the hydrophobic amino acids of the hydrophobin. When estimating the partition between the two populations at $120\ \text{Å}^2$ surface area per amino acid, the population of structures with a mean accessible surface of about $80\ \text{Å}^2$ comprised 9100 members, while the population with a mean accessible surface area of about $150\ \text{Å}^2$ comprised 6000 members. The overall ratio was therefore roughly 2:3. The lowest energy structure was located in the $80\ \text{Å}^2$ population. The structures of the models with the lowest energies are shown in Fig. 5.10.

A mean solvent exposure of the LG3 motif of $80\ \text{Å}^2$ per amino acid was observed inside the single cluster of simulated models (Fig. 5.9 b). Compared to the RGD structure, the LG3 motif may tend to be less exposed to the solvent in the tertiary fold. The observed tight packing of the binding peptide to the fusion domain can be explained by the large number of hydrophobic amino acids in the LG3 sequence (Fig. 5.7); however, the low-energy ensemble contained many models in which a large fraction of LG3 is exposed to the solvent. Modelling showed

that both DewA fusion proteins may be able to enhance cell adhesion; therefore, these proteins were synthesized and purified for subsequent experimental testing by the groups Fischer and Richter.

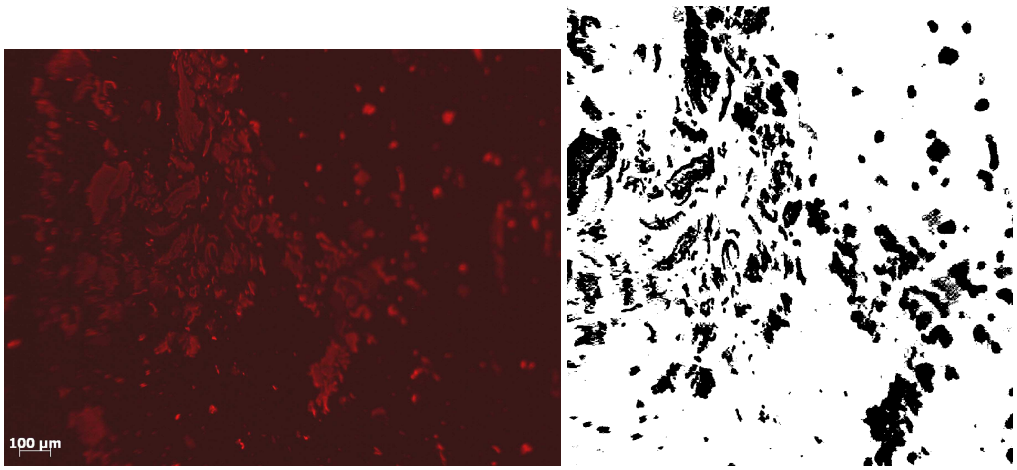
5.2.7. Experimental verification of Bacterial and Cell Adhesion

The proposed DewA mutants were inserted via primer ligation and expressed in transformed E.coli Rosetta (DE3) pLysS cells as reported in Boeuf et al.[99][2]. Cell-culture wells were coated with the unmodified hydrophobin with linker and RGD- and LG3 variants by exposing them to hydrophobin solution and incubation overnight. Cell-culture wells with fibronectin and BSA acting as positive and negative control were also prepared.
A medium containing multiple cell types (MSCs, chondrocytes, osteoblasts and fibroblasts) was applied to the cell culture wells incubated again and washed, after which the cells were counted. Results are shown in Fig. 5.12 a). MSC adhesion to DewA was significantly increased for both RGD and LG3 variants of the hydrophobin compared to the wild-type DewA (Fig. 5.12 a)). Although the adhesion was increased it remained significantly lower than the positive control fibronectin. More cells adhered to DewA-RGD than to DewA-LG3 at hydrophobin concentrations of $20\,\mu g/ml$ and upwards (Fig. 5.12 a). To mimic in-vivo conditions in patients with (bone-)implants, the above experiment was reproduced using titanium as a coating substrate. Adhesion was similar to the previous results. Bacterial adhesion was quantified by exposing the titanium discs for one hour with colonies of an S. aureus strain. After exposing the covered discs for one hour, they were washed, photographed and the percentage of the covered surface was estimated graphically.
Again the highest coverage was achieved with the positive-control fibronectin as shown in Fig. 5.12 b). Although also DewA, DewA-RGD and DewA-LG3 showed bacterial adhesion, the covered surface area was significantly lower than that of fibronectin, albeit being higher than for uncoated surfaces. It is interesting to note that the inclusion of the RGD motif did not increase bacterial adhesion compared to the unmodified DewA: Bacterial adhesion was not mediated by the RGD motif.

5.2.8. Discussion

In this section I have reported the successful structure based design of genetically modified DewA hydrophobins with enhanced cell-binding properties. After initial model generation a suitable site for mutation could be located. The integration of the RGD motif[126] and the laminin globular domain LG3[128] introduced binding sites for improved cell adhesion, which could be validated using adhesion assays. At the same time no increased bacterial adhesion was observed compared to unmodified DewA hydrophobins.

[2] As this work was a collaboration, it will only be presented in a brief fashion. Please review the cited manuscript for quantitative details. This experimental part of the work was conducted by AG Fischer (primer ligation and insertion), AG Schimmel (surface characterization using AFM) and AG Richter (cell adhesion assays)

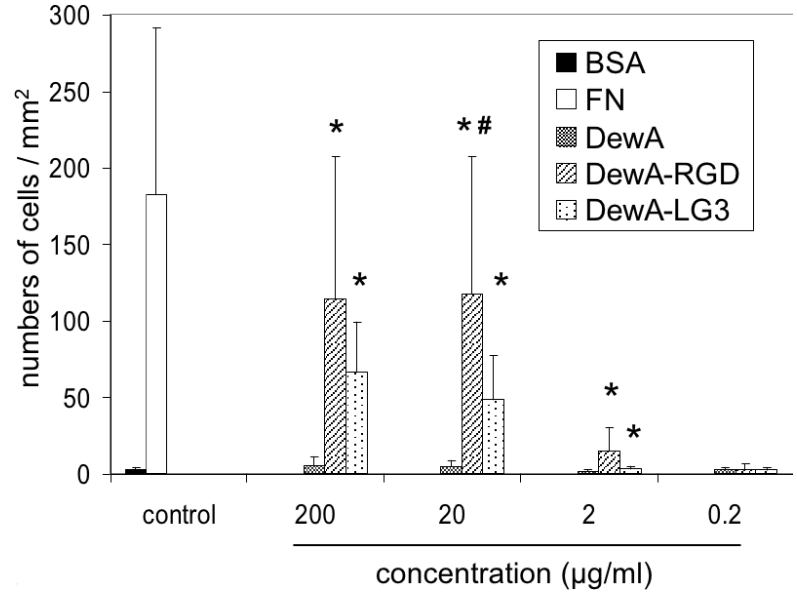


(a) Original Image (b) Masked used for counting

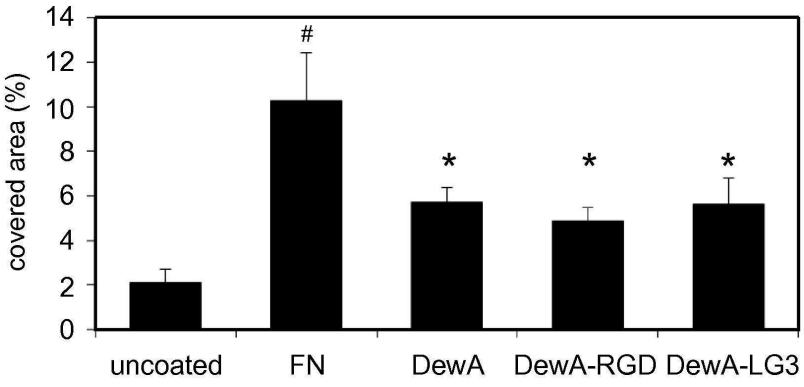
Fig. 5.11.: *Images used for the estimation of the DewA covered area used in the cell adhesion experiments. a) Fluorescence image of a DewA cell culture well. b) The in-focus part of the image shown in a) was extracted and converted into a black and white mask. By counting of all the black pixels in the image the coverage can be estimated to about* 20%.

One of the domains in the model of the fusion proteins corresponded to an intact amphipathic DewA domain with a large hydrophobic patch, which may be involved in impairing cell adhesion. Both RGD and LG3 motifs are part of the second domain, including the fusion peptide, and were exposed at least partially in many of the low-energy models. The RGD motif was identified as facing the same side as the hydrophobic loops, indicating that it is exposed to area accessible by cells, which is supported by the experimentally observed increased adhesion of cells to the engineered binding-peptide mutants.
On the basis of the proposed orientation (hydrophobic loop up) of the DewA-RGD variant, we estimated an area of $13\,nm^2$ per protein/RGD motif or $8 \cdot 10^{10}$ ligands per mm^2 for perfect packing, but the apparent surface density was lower. We have estimated the covered surface for a DewA-RGD covering used in the cell adhesion studies as about 20% of the total surface area of the cell culture wells. This was estimated by counting the fluorescent pixels on a sample image of about $1\,mm^2$ (Fig. 5.11).

Binding studies with similar surface densities of RGD were reported by Le Saux et al.[133]. Using a variety of mirror-polished and etched silicone materials, they investigated the influence of RGD ligand density and surface roughness for endothelial cell adhesion. For an observed RGD density of $6 \cdot 10^{11}$ ligands per mm^2 (compared to $8 \cdot 10^{10}$ ligands per mm^2 in this study), they report 700 endothelial cells per mm^2 for a mirror-polished surface, which fell to 300 cells per mm^2 for a silicone surface etched for 10 min. Adherent endothelial cells featured a mean cell surface of $400\,\mu m^2$. The cells used in this investigation featured a surface area of roughly $700 - 1000\,\mu m^2$ and therefore occupy roughly double the area. Our investigations were performed in cell culture wells of unknown roughness; however, the number of cells for the fibronectin positive control in this study is comparable to the etched silicone surface in the study by Le Saux et al.[133] (300 cells on silicone vs. 180 cells per mm^2 in this study), considering the increased surface area of the stem cells: The total area of 300 cells with a mean cell surface of $400\,\mu m^2$ (Le Saux et al.) is comparable to the total area of 180 cells with a mean surface area



(a) RGD



(b) LG3

Fig. 5.12.: *a) Cell adhesion on surfaces coated with DewA variants. For the quantification of cell adhesion, cells were allowed to adhere for 1 h. After fixation, the number of adherent cells per mm^2 was counted in three randomly selected photographic fields. MSCs were allowed to adhere to surfaces coated with BSA (2%), fibronectin ($10\mu g/ml$) and various concentrations of the hydrophobins DewA, DewA-RGD and DewA-LG3. *Significant difference in comparison to DewA at the same concentration and to fibronectin; #significant difference in comparison to DewA-LG3 at the same concentration ($p < 0.05$). Adhesion of S. aureus to titanium discs coated with fibronectin and DewA variants. The fraction of the surface of the disc covered with S. aureus was quantified using the measure function of ImageJ and is shown as a percentage. *Significantly lower than fibronectin, significantly higher than uncoated; # significantly higher than uncoated ($p < 0.05$). Image courtesy of Beouf et al.[99]*

of $800\,\mu m^2$ (Fibronectin positive control in this study).
About 40% less cells were observed on surfaces covered with the RGD-modified hydrophobin compared to the Fibronectin positive control. This efficiency of 60% of the RGD-modified hydrophobin construct in comparison to fibronectin can be explained by the unordered surface coverage of 20% of the hydrophobin (Fig. 5.11). Intuition would suggest that the incomplete coverage of the surface with DewA would reduce also the binding efficiency to 20% instead of the measured 60% compared to the fibronectin positive control due to the decreased amount of RGD binding motifs on the surface, however Le Saux et al.[133] could show that a decreased RGD ligand density can actually lead to increased endothelial cell binding. Our sample showed unordered ligand densities, which could explain the difference of 60% in binding efficiency, when comparing DewA and the Fibronectin positive control.
Furthermore, it should be noted that surface coating with DewA was heterogeneous. Whereas the A. nidulans hydrophobin RodA is able to form rodlets on the spore surface, DewA does not have this ability[134, 135]. The low adhesion on DewA-LG3 compared to DewA-RGD could result from a less exposed conformation of the cell binding motif as shown in the models or to differential adhesion potentials of cells to RGD and LG3.
Neither changes in proliferation of MSCs nor a change in differentiation potential of MSCs could be observed on DewA surfaces compared to uncoated cell culture wells (data in [99]). Further studies can include modular peptide sequences known to promote osteogenic differentiation[136]. As we could locate a sequence segment in the protein, where mutations will conserve the overall tertiary structure and function of the hydrophobin, introducing additional short sequences is possible.
Compared to untreated titanium disks a higher amount of bacteria was observed on titanium disks coated with DewA. Adhesion was still significantly lower than with fibronectin, a common wound fluid. Hydrophobin coated implants can therefore induce lower bacterial adhesion than fibronectin, which can immediately form after implantation. Binding motifs for cell adhesion did not increase the bacterial adhesion.

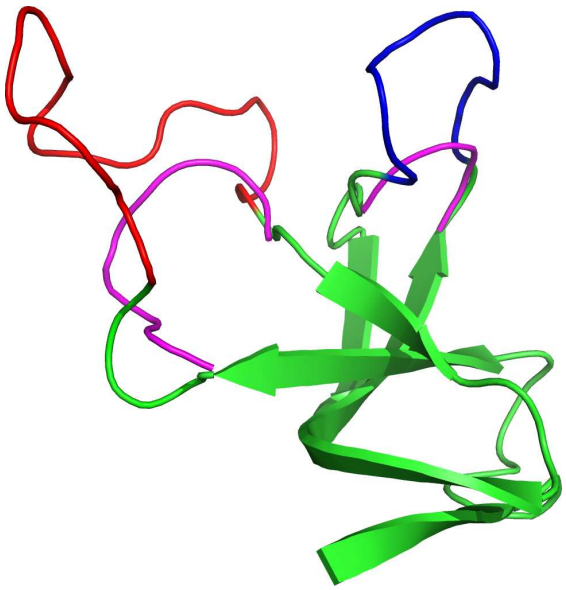


Fig. 5.13.: *Starting model of the docking simulations of protein EAS. The experimental structure comprises the green barrel in the center and the red and blue hydrophobic loops. Amino acids were removed in these loops to form the truncated mutant including the green barrel and the magenta loops.*

5.3. Modeling of Rodlet formation of hydrophobin EAS

5.3.1. Introduction

Class-I hydrophobins, presented in the previous section (section 5.2), form long, cylindrical rodlet structures at air-water interfaces. The hydrophobin EAS (PDB: 2FMC) is a hydrophobin protein, where such rodlet development was observed experimentally. The structural mechanism, these hydrophobins use to link into cylindrical rodlet structures, is currently unclear. Kwan et al. [131] indicated that the involvement of the flexible loops in hydrophobin EAS is unclear, but resolved a β-barrel as the main structural motif mediating aggregation into rodlets. A later publication even demonstrated that more defined rodlets could be observed once the flexible loops were truncated[137].
In this section, we want to elucidate the structural mechanism by which hydrophobin EAS develops rodlets, and investigate the role of the flexible loops in rodlet formation. Multiple simulations using mutants of protein EAS missing the flexible loops are carried out (Fig. 5.13), to assess, if a periodic structure is developed without the formation of additional β-content by the loops. In section 5.3.2, we assess, whether a β-barrel structure can be observed using a truncated EAS mutant modeled by means of homology modeling and protein-protein docking. We discuss the results in section 5.3.3.

5.3.2. Docking of a truncated mutant of protein EAS

Homology modeling was carried out with the standard Modeller protocol as used by Fiser et al.[76, 130]. The resulting ensemble does not present any variability. The structural ensemble created using homology modeling exhibits less structural variability, than the 20 experimental

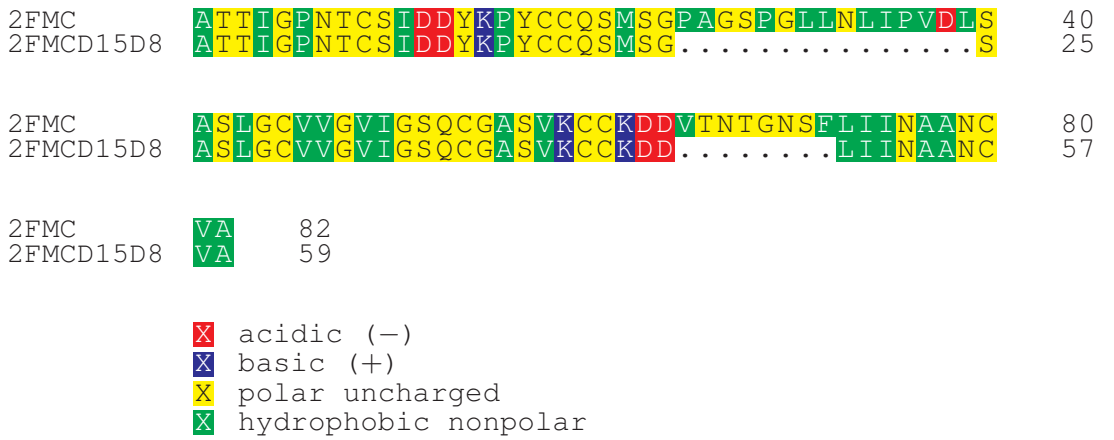


Fig. 5.14.: *Alignment of the two sequences of 2FMC. Especially the first loop contained multiple hydrophobic residues cut in the mutant.*

snapshots of the full structure of EAS[131]. The hydrophilic barrel structure remains intact, while the flexible loops lost their flexibility, due to truncation of most amino acids they contain (Fig. 5.13).

Next, we modeled 2000 complex structures using the FFT-based docking approach Zdock[138]. As rodlet covered surfaces show a defined uniform hydrophobicity along the rodlet surface, it can be assumed that rodlets do not include shear between the hydrophobin subunits: If a a hydrophobin would be twisted around the long axis of the rodlet, hydrophobicity of the rodlet would not be uniform. Therefore only structures with a shearing angle less than 40 degrees are included from the Zdock predictions for further analysis.

While the Zdock predictions might represent physical reality, high or low quality predictions cannot be easily discriminated. Therefore the resulting structures were offset by 10% of their center of mass distance and then relaxed in SIMONA on POEM@HOME[71] in a 25.000 step Metropolis Monte-Carlo simulation comprising 50% rigid body movement and 50% sidechain relaxation moves. During the simulations the temperature was annealed from $300\,K$ to $5\,K$ using a geometric cooling schedule (temperature units given in respect to the temperature scale of the used forcefield PFF02). Afterwards the energies of the structures and the RMSDs to the minimum energy structure were evaluated. Special care needed to be taken to evaluate the RMSD, as most programs provided false results, because they did not treat the two 2FMC subunits in the models equal. Therefore RMSDs were evaluated twice using the McLachlan algorithm[139] in the Profit program, one time for each mapping; the resulting RMSD is the smaller one of the two evaluations.

Results are shown in Fig. 5.15. Two highly symmetric low-energy conformations could be identified; the lowest energy structure at $-97kcal/mol$ is the primitive unit cell structure (Fig. 5.15c), which can be extended into periodical rodlets. Another low-energy structure is observed at $-94kcal/mol$ in an antisymmetric binding pose (Fig. 5.15b) and represents a possible dimer of two subunits, which cannot be easily periodically extended.

When aligning the lowest energy structure orthogonal to the hydrophobicity gradient, a periodical rodlet can be observed as displayed in Fig. 5.16. The hydrophobicity gradient is a likely orientation of the EAS subunits at air-water interfaces.

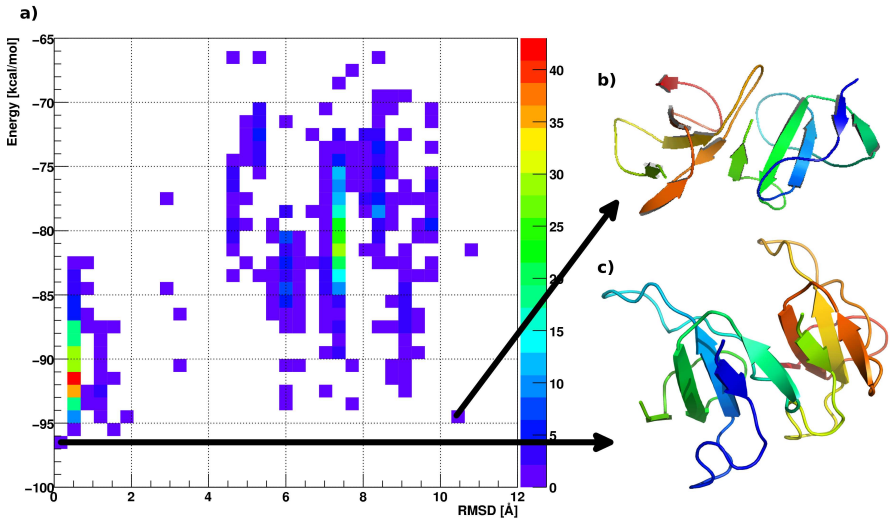


Fig. 5.15.: *Results of the docking approach of the truncated hydrophobin EAS mutant. a) Energy and RMSD values of the simulation of two docked EAS monomers. The RMSD is evaluated in comparison to the lowest energy structure. The two lowest energy structurally dissimilar models are observed at $-97kcal/mol$ and $-94kcal/mol$. It is striking that both binding poses involve the β-barrel interface and are highly symmetric b) The higher energy ($-94kcal/mol$) antisymmetric binding pose. This binding cannot be easily extended as the β interfaces on opposite point in different directions. c) The lowest energy ($-97kcal/mol$) symmetric binding pose. This binding pose can be extended into an extended rodlet structure.*

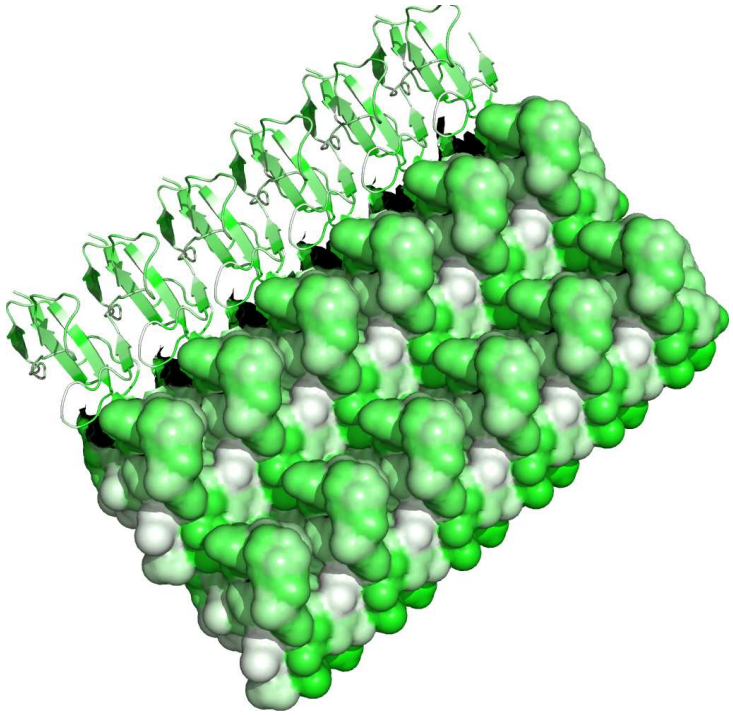
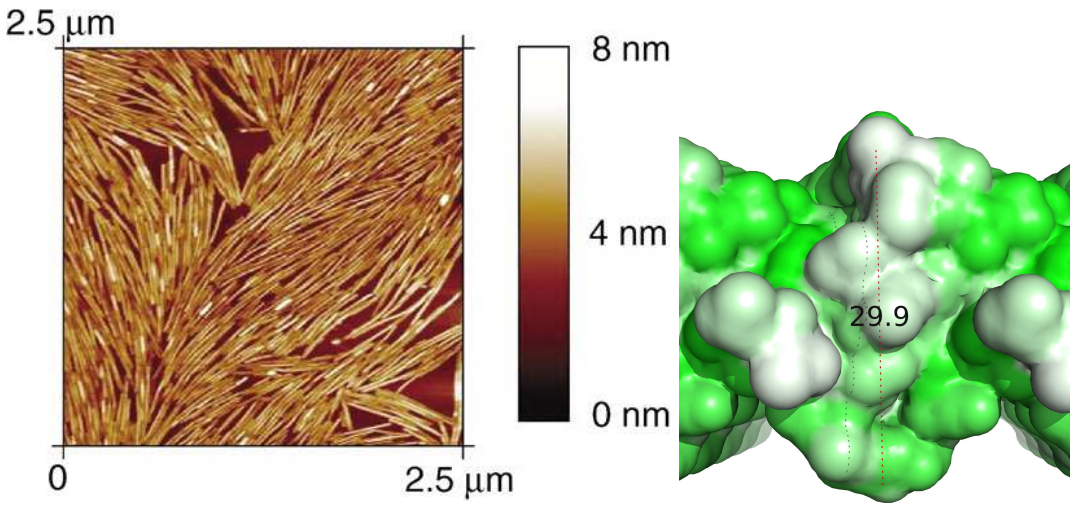


Fig. 5.16.: *Rodlet structure observed when assembling the lowest energy structure of the truncated 2FMC. The z-axis orthogonal to the rodlet plane is the axis of the hydrophobic gradient. White spots are hydrophobic, while green spots are hydrophilic. Exposed white spots can be seen on the top of the rodlets.*



(a) Rodlets of hydrophobin EAS on a surface

(b) Theoretical model of a single rodlet subunits

Fig. 5.17.: *Comparison of rodlet dimensions to experimental results. a) AFM image of EAS rodlets on surface by Mackay et al.[140]. The height of the rodlets can be estimated to roughly* $4-7\,nm$ *for the full EAS. Our truncated variant shows a height of* $3\,nm$*. The height difference can be explained by the truncated flexible loops. b) A simulated rodlet for comparison.*

5.3.3. Discussion

In this section, we predicted a structural model of rodlet formation of class-I hydrophobin EAS, which may explain rodlet assembly on air-water interfaces. Although earlier experimental studies hypothesized an involvement of the large flexible loops in rodlet formation[131], our simulations confirm the subsequent evidence[137] that a rodlet can be developed also by a structure without flexible loops.
The structural features of the proposed rodlet are compatible with experimental observations: Mackay et al. published AFM images of protein EAS on surfaces[140]. The height of the rodlets in these images can be roughly estimated to $5-7\,nm$. This is comparable to the height observed in the full rodlet image, measured to $3.4\,nm$. The difference in sizes can be attributed to the truncated loops, as the AFM image resolved the full EAS rodlet structure and the rodlet in Fig. 5.17 is built from the truncated mutants.

In conclusion, we could identify the mode of protein-protein aggregation, by attachment of the β-barrel subunits of hydrophobin monomers. The aggregation was possible in spite of the truncations introduced in the mutant protein. Our findings are compatible with experimental results obtained by Kwan et al.[137]

5.4. Structural model of the development of gas-vesicles in aqueous bacteria

5.4.1. Motivation

Protein-protein aggregation has been an important subject for study in the biological sciences, which gathered increasing momentum in the discovery of aggregation related diseases such as Alzheimer's[141, 142]. Because of the difficulties to study such assemblies with traditional experimental techniques, detailed structural information is only rarely available[143]. One example of a protein forming macroscopic aggregates is the gas-vesicle protein GvpA, which aggregates into a macroscopic helical, rib-shaped gas-vesicles. Gas-vesicles provide the lift for aquatic bacteria to enable flotation on natural water bodies by increasing the bacteria's buoyancy[144]. Recently new microscopic information about the GvpA protein and its assembly has been discovered by a combination of many different approaches[145, 146], but a complete structure for the protein and its assembly is not yet available. In this chapter, we present a first structural model of the gas-vesicle forming protein GvpA obtained by de novo modelling using the GvpA sequence of Haloferax mediterranei (Fig. 5.18).
In section 5.4.2, we first introduce known structural information of protein GvpA. Afterwards we present the methods used for de-novo prediction of GvpA and the analysis of the model and assessment of its stability in section 5.4.3. Subsequently, we briefly discuss the results regarding the various aspects of our investigation, such as the monomer model and its stability and dimerization in section 5.4.4 and compare the model's properties with experimental results in section 5.4.5. The model presented here is compatible with most of the currently available structural information on GvpA[3].

5.4.2. Introduction

Here we investigate the proteins forming the gas-vesicle wall of bacteria Haloarchaea. The gas-vesicle wall of Haloarchea is built solely of proteins with the $8\,kDa$ protein GvpA being the dominant species present in the wall[144]. Macroscopic imagery shows them as spindle- or cylinder-shaped structures up to $1\,\mu m$ in length and $200\,nm$ in diameter. To contain the gas, the inner gas-vesicle wall is more hydrophobic than the outer hydrophilic shell. The protein GvpC, which is also part of the vesicle wall, is attached to the outer shell, which stabilizes the gas-vesicle wall.
Gas enters the vesicles passively by diffusion of gas molecules dissolved in the cytoplasm. It is hypothesized that although water molecules might enter the structure, the hydrophobic, curved inner surface prevents condensation. Collapsed gas vesicles exhibit a rib-shaped structure with $5\,nm$ ribs running perpendicular to the long axis that are presumably formed by GvpA[147–149]. The amino acid sequence of GvpA is highly conserved among bacterial species (Fig. 5.18), whereas the sequences of GvpC are more divergent.
Although most of the macroscopic properties of gas-vesicles are known, there is still no

[3] Parts of this text were previously published in Strunk et al.[100]. I thank the publisher John Wiley and Sons and all the coauthors for the permission to republish it as part of my thesis.

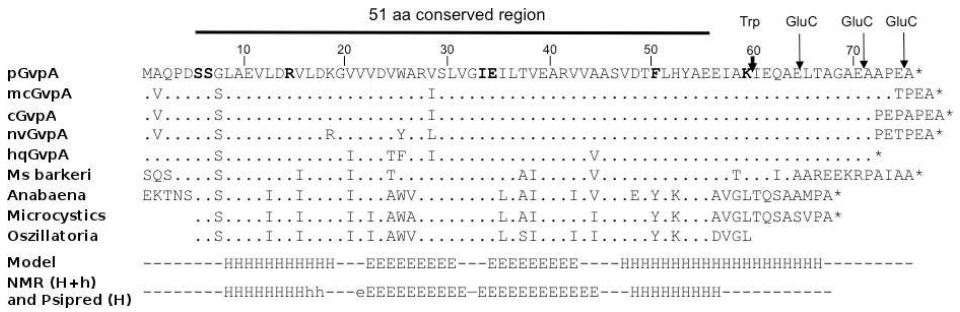


Fig. 5.18.: *Alignment of GvpA sequences and prediction of α-helices and β-sheets. The 51-amino-acid conserved core region is indicated by a bar and the sites accessible to peptidases (Trp: trypsin, GluC) are marked by arrows[145]. pGvpA and cGvpA derive from Hbt. salinarum, mcGvpA from Hfx. mediterranei, nvGvpA from Halorubrum vacuolatum and hqGvpA from Haloquadratum walsbyi. Ms, Methanosarcina. The model structure is given in the last line with (H) depicting α-helical and (E) depicting β-sheet regions. For comparison, the PSIPRED and NMR results for GvpA of Anabaena are added[146]. The differences between PSIPRED and NMR results are marked by small letters (h = α-helical, e = β-sheet, additionally observed in NMR, but not in PSIPRED).*

nano-scale structural model of the GvpA monomer or the formed vesicle. Previous solution NMR studies encountered difficulties to resolve the protein structure, since GvpA monomers aggregate in solution and dissolve only in 80% formic acid. Removal of the formic acid via dialysis results in amorphous GvpA precipitates[145].

Recently a solid-state NMR study suggested a coil-α-β-β-α-coil fold[146]. Furthermore FTIR spectra indicated antiparallel β-sheets in the GvpA structure, while X-ray diffraction and atomic force microscopy suggested that the β-strands of GvpA are tilted in the ribs at an angle of $54°$ to the axis of the rib[145, 148, 150, 151].

The exposure of peptide bonds of GvpA inside the gas vesicle structure was determined using proteolytic cleavage using trypsin among other proteases in the bacteria Anabaena flos-aquae and Hbt. salinarum[145]: It could be shown that a highly conserved (51 AA) part of the sequence of GvpA is not cleaved by trypsin in either case (Fig. 5.18). GvpA was accessible to trypsin as the C-terminal K60-I61 bond was cleaved in the case of Hbt. salinarum, whereas other possible trypsin cleavage sites were not affected. Cleavage at K60 resulted in collapsed gas vesicles[145]. Similar results were obtained, when cleaving with GluC endopeptidase. Also GluC endopeptidase could not cleave sites within the conserved segment of GvpA (Fig. 5.18); only the C-terminal portion could be cleaved, which is therefore presumed to lie at the outer surface of the gas-vesicle.

The single GvpA subunits inside the vesicle structures were shown to be non-equivalent in solid-state NMR studies of gas-vesicles of A. flos-aquae due to a small folding variation in alternating subunits[152]. The model derived implies that the β-sheet portion of GvpA achieves a hydrophobic surface, and that complementary charges and aromatic-aromatic interactions are present at the subunit interfaces.

```
mcGvpA  MVQPDSSSLAEVLDRVLDKGVVVDVWARISLVGIEILTVEARVVAASVDTFLHYAEEIAKIEQAELTAGAEAAPTPEA*
PSIPRED --------HHHHHHHH----EEEEEEEEEEE--EEEEEEEEEEEEEHHHHHHHHHHH---HHH--------------
DSSP    ---TT-S-HHHHHHHHHHTT-EEEEEEEEEEEEEHHHHHHHHHHHHHHHHHHHHHHHHHHHH-HHHHHHHHHH--H---
Stride  --TTTT--HHHHHHHHHTTEEEEEEEEEEEETTEEEEEEHHHHHHHHHHHHHHHHHHHHHH-HHHHHHHHHHHHHH-
sspro   -------HHHHHHHHHH--EEEEEEEEEEE-EEEEEEEEEEEEE-HHHHHHHHHHHHHHHHHHHH-----------

Model   --------HHHHHHHHHH---EEEEEEEEE---EEEEEEEEE----HHHHHHHHHHHHHHHHHHH-----------
```

Fig. 5.19.: *Helices (H) and β-sheets (E) in GvpA as determined by different model programs. Most notable is the agreement of the assignment of two beta-strands in three of the programs (PSIPRED, SAM-T02-Stride and SSPro). The single long helix as predicted by SAM-T02-DSSP is possible, but unlikely as the single β-strand would not be paired in the resulting structure.*

5.4.3. Methods

Model generation

Templates for template based modeling were obtained by querying the 3DJury[153] meta-server. Among the servers queried were PSIPRED[154] and two SAM-T02[155] secondary structure prediction servers. Furthermore the PHYRE fold recognition server[156] was used to identify the secondary structure in similar sequences. Finally a separate alignment using FUGUE[157] was created, which was translated into a three-dimensional model using the Modeller program. The GvpA protein was also modeled directly by the I-Tasser[158] and SAM-T08[159] web-based threading servers.

To enlarge the model diversity of the attained structures another modeling attempt using fragment based modeling was carried out. The de novo structure prediction scheme consisted of a model generation step and a refinement step. In the model generation step we used ROSETTA[160] to generate a set of decoys, which were clustered using MaxCluster in order to avoid rescoring the same model multiple times. Afterwards the decoys were ranked using the all-atom free-energy force field PFF02[9] as a scoring function. Each of the models had to be be relaxed in a short simulation to a close-by local minimum in the PFF02 force field, previously described in section 3.2.2.

Fragments were obtained using the Robetta server[161]. We assembled these fragments into mainchain models using the ROSETTA 2.3 software suite using the default ab initio protocol[161, 162]. Subsequently a short idealization simulation was carried out, which added the missing sidechain atoms and removed clashes in the ROSETTA force field. The resulting ensemble of models was clustered using MaxCluster with a 3DJury (MaxSub)[163] score threshold of 0.55 and a 3DJury pair threshold of 20. Each of the MaxCluster cluster centroids was relaxed in five independent simulations per model, using a simulated annealing protocol with a starting/end temperature of $700K/5K$, respectively, and a geometrical cooling protocol. We note that these temperatures must be scaled against the energy scale of the forcefield PFF02, which has not been calibrated. Each simulation comprised 1.5 million steps. In each step one randomly selected dihedral angle of the model was changed by a maximum of five degrees. Backbone/sidechain dihedral angles were selected for rotation with a ratio of 70%/30%[164].

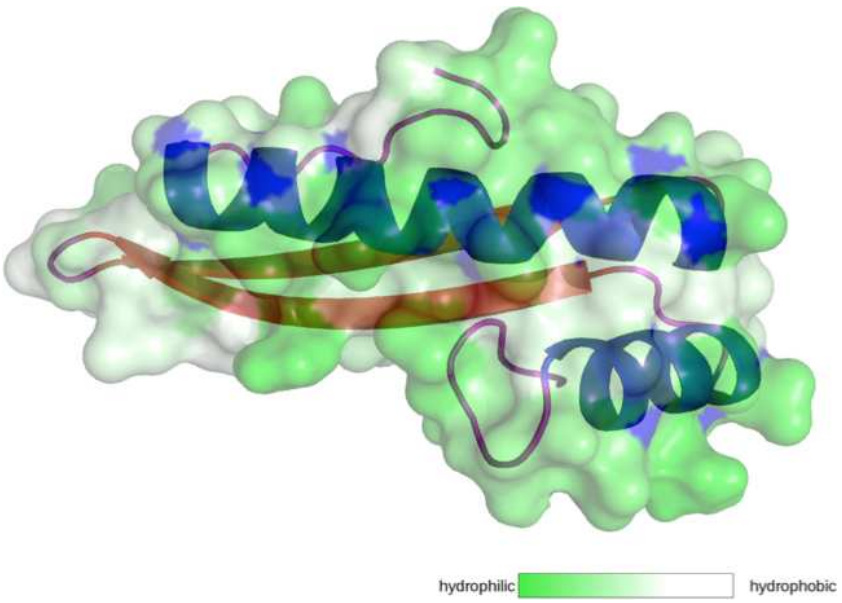


Fig. 5.20.: *Ribbon illustration of the lowest energy model of a GvpA monomer. A single central β-strand (red) is observed comprising of two bonded β-strands from positions V23 to L31 and E35 to V43. Furthermore there is a long helix (blue) from V48 to T67 and a smaller helix from amino acids L9 to K19. The accessible surface area is colored using the Eisenberg-Schwarz hydrophobicity scale encoded in the intensities of green and includes the two lysine residues that were observed in solid-state NMR results[146] to be solvent-accessible.*

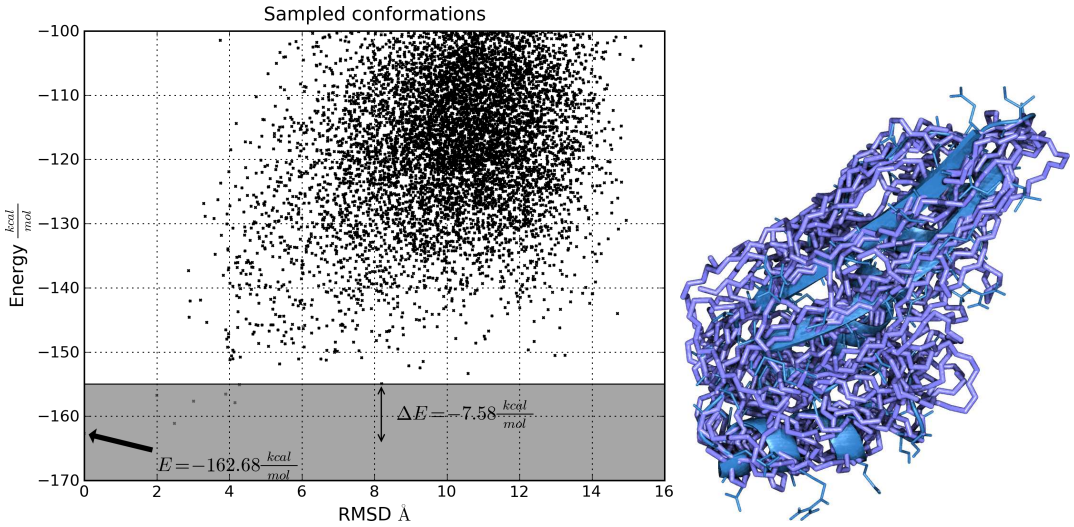
Model evaluation

To judge the stability of the model predicted for the monomer, we simulated the dynamics of this molecule with the NAMD molecular dynamics package[165] for $20\,ns$ at two different salt concentrations in reference to the halophilic organisms: (i) at $1\,M$ KCl and (ii) at $1\,M$ NaCl and $5\,M$ KCl in the CHARMM force field[166]. Trajectory analysis was performed with the GROMACS software suite[167]. We repeated the simulations five times for each salt concentration. A model of docked GvpA subunits was prepared using the ROSETTA protein-protein docking protocol[168] by the group of Prof. Dr. Kay Hamacher.

5.4.4. Results

Template-based modelling and secondary structure prediction

The amino acid sequence of mcGvpA derived from Hfx. mediterranei (see Fig. 5.18) was used to obtain a structural model. No alignments with significant prediction scores could be discovered using 3DJury. The FUGUE server delivered a single marginally significant result. While tertiary structure prediction using template-based modelling proved difficult, all secondary structure prediction methods resulted in similar secondary structure content (Fig. 5.19). Helical regions were predicted between L9-V16 and V48-I58, while a β-strand was predicted for V22-V32, both in good agreement with recent results from solid-state NMR[146]. PSIPRED's secondary structure prediction seems the most plausible, as it permits a pairing of two equal-length β-strands; again in correspondence with solid-state NMR[146] (see Fig. 5.18).



(a) Energy-RMSD values of the simulated structures

(b) Low energy ensemble

Fig. 5.21.: *a) RMSD versus energy of the relaxed* 1850 *centroid models scored with PFF02. The RMSD is measured with respect to the lowest energy structure. The lowest energy cluster of models (dark grey region) comprised similar models (see b)). The lowest energy model is separated by an energy gap of* $7.6 kcal/mol$ *from the first structurally dissimilar model, which has an RMSD of* $8\,$Å*. b) Structural ensemble of the ten lowest energy models. The cartoon plot in the middle represents the lowest energy structure.*

PHYRE did not detect homology to known proteins, but its secondary structure prediction element (using sspro) coincides with the PSIPRED secondary structure prediction.
In addition to the 3DJury results FUGUE was directly used as an alignment server. A marginally significant result (Z-score: -3.95) was obtained for the PDB-template 2JOI. A model was built using Modeller, 2JOI and the FUGUE alignment. In contrast to the previous hypothesis of two paired strands, the model shows four paired strands stabilized by two helices. Since a large fraction of the amino acids in this model remained unstructured, this approach was not pursued further.

De novo modelling

Since no models of promising quality emerged from 3DJury, separate FUGUE, I-Tasser and SAM-T08 modelling attempts, we resorted to a de novo prediction using the predicted secondary structure elements as constraints. In total $53,000$ mainchain models were created using ROSETTA, which resulted in 1850 centroid models after clustering. The lowest energy structure (Fig. 5.20) has a PFF02 energy of $-163 kcal/mol$ and is separated by a gap of about $8 kcal/mol$ to the next structure of different topology ($\Delta RMS = 8\,$Å - Fig. 5.21). Sampled models with an energy below $-155 kcal/mol$ fall into a structurally similar cluster with a $2\,$Å radius. An illustration of the ensemble of this cluster is shown in Fig. 5.21b. The model features two bonded β-sheet regions at V23-L31 and E35-V43, each of which is 9 AA long (Fig. 5.18). In addition there is one long α-helix (V48-T67) that is arranged almost in the same direction as the β-sheets but faces the other surface. These regions agree with the secondary structure predictions we used

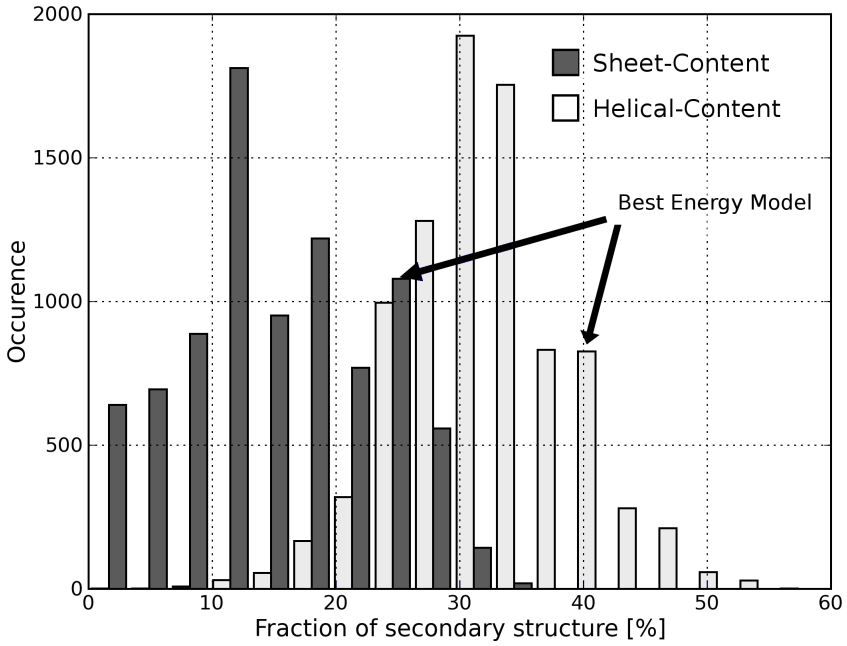


Fig. 5.22.: *Fraction of β-sheet and helical content in the models created by ROSETTA. The values of the lowest energy model are indicated by arrows. The models span a large interval of different secondary structure combinations. The selected prediction contains both high amounts of β and α topologies.*

to generate the models. Another α-helix is observed from amino acid L9 to amino acid K19. Eleven hydrogen bonds are located between the two strands of the β-sheet. The whole population of models created using ROSETTA has a large fraction of secondary structure elements (Fig. 5.22). Refining these models with PFF02 selected the lowest energy model among many competitors with similar secondary structure.

Stability analysis

The RMSD in the five simulated trajectories was observed to saturate around $4\,\text{Å}$ in both salt conditions (Fig. 5.23a,b), which is commensurate with the deviations observed in simulations started with high resolution crystal structures using the same forcefield[169].
These results indicate the overall stability of the proposed monomer structure. Especially the secondary structure elements proved to be stable as individual dihedral angles in α and β regions only showed small deviations from the starting structure.

Dimerization

Next we attempted to assemble the monomer models into dimers, generating $622,000$ decoys in total using the Rosetta dimerization simulations. Structures with the two lowest dimerization energies observed in the simulations were $-128.9 kcal/mol$ and $-128.7 kcal/mol$, which differed only by $0.09\,\text{Å}$ from each other (Fig. 5.24). The dimer model is nearly symmetric, i.e. the two subunits have a vanishing shearing angle towards each other. Furthermore, the dimer presents an *inner* concave surface with high hydrophobicity. We hypothesize that the surface faces towards the vesicle inner in agreement with results from solid-state NMR[146].

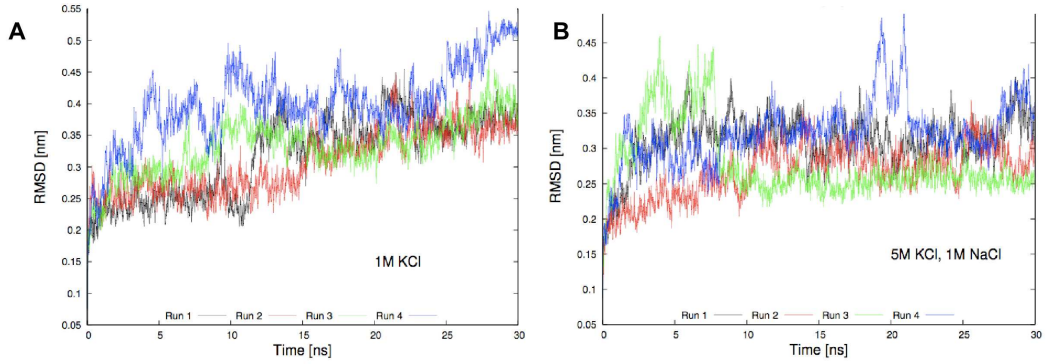


Fig. 5.23.: *RMSD over the course of time in the MD simulation for the two simulated high-salt concentrations (a and b). The reference structure is always the predicted GvpA monomer. We repeated the simulations five times each to show stability of the monomeric structure. The RMSDs lie within typical regions of fluctuations of proteins around an experimental structure.*

5.4.5. Discussion

In this section I have presented a first structural model for dimerization of the major gas-vesicle forming protein GvpA using ab-initio all-atom modeling. The model is consistent with previously determined properties, such as the existence of a hydrophobic inner gas-vesicle wall and the hydrophilic shell. The model features a hydrophobic-inner and hydrophilic-outer surface and contains two α-helices separated by two anti-parallel 9-AA long β-strands of $3.67\,nm$ length. Overall the structure contains 47% helical, 25% β-sheet and 28% of unknown structure. The antiparallel β-strands presumably form the inner surface of the gas vesicle wall. Nine out of ten residues in the beta-sheets are hydrophobic and point towards the gas-facing surface of the wall, whereas six out of eight amino acids, pointing to the other side of the β-sheet, are hydrophilic/charged. Helix I is also located at the gas-facing surface and forms the broader back of the triangular-shaped GvpA monomer (Fig. 5.24).
Our findings disagreed with previous investigations, which postulated that GVPA consists only of β-structures[144]. The presence of a significant amount of α-helices was confirmed by FTIR measurements performed by our experimental partners (data in Strunk et al.[100]). Apart from the presence of significant helical secondary structure elements, the FTIR measurements could also confirm connected anti-parallel β-sheets as observed in the proposed model. Our structural results and their assignment to secondary structure elements agrees very well with the NMR results obtained for cyanobacterial GvpA of Anabaena [146] (Fig. 5.18). The only differences occur in the C-terminal region, where the haloarchaeal and cyanobacterial GvpA sequences substantially differ.

The model of the GvpA dimer contains two antiparallel GvpA monomers of triangular shape that are connected by contacts between half of the antiparallel β-sheet region of monomer 1 and monomer 2 (Fig. 5.24). The relatively large extension (tip) formed by the second half of the anti-parallel β-strands and the β-turn should contact the next GvpA dimer located in the adjacent rib, with one monomer in contact with the rib on top and the other one with the rib below as indicated in Fig. 5.24B. It is striking how perfectly the dimers fit to each other when a single layer of GvpA is formed. The orientation of the antiparallel β-sheet relative

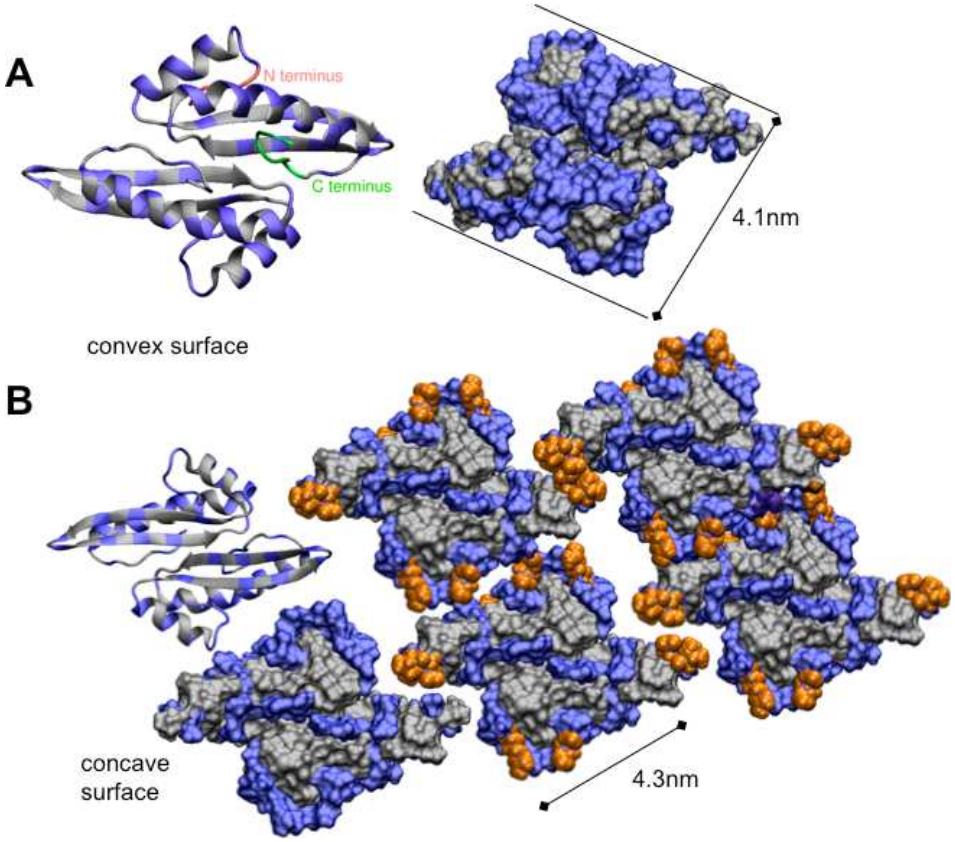


Fig. 5.24.: *Predicted dimerization mode. Blue resides are hydrophilic, Gray residues hydrophobic. Orange residues were mutated in a mutation experiment introduced in section 5.4.5 A) Two of the monomers shown in Fig. 5.20 are docked in an antisymmetrical mode in both a cartoon representation (left) and a surface representation (right). B) Concave side of the proposed vesicle wall. The hydrophobic wall facing outwards shows high hydrophobicity and is therefore implied to be facing towards the gas pocket. Two of the dimers shown in A) are arranged to form a* $4.3\,nm$ *wide rib. Figure courtesy of Strunk et al.[100].*

to the axis of the rib corresponds well with the $55°$ angle shown by X-ray diffraction studies of cyanobacterial gas vesicles [148]. Such periodicities are also observed by atomic force microscopy [150]. The unit cell of the GvpA monomer model measures $4.3\,nm$ (across the rib without the tip region) $\times 2.2$ nm (along the rib defined by the dimer structure) $\times 2.1\,nm$ (wall thickness) (Fig. 5.24). These dimensions were measured from the most distant atoms along the respective axes, effectively neglecting the unknown separation of unit cells. The measurements are close to the dimensions of the unit cell ($4.57 \times 1.15 \times 1.95\,nm$) known from fibre X-ray crystallography using gas vesicles from the cyanobacterium A. flos-aquae[148]. It should be noted that the length of the β-sheet ($3.67\,nm$) is significantly shorter compared with the width of the rib ($4.3\,nm$) and thus constitutes only part of it. However, the tip of an adjacent GvpA monomer (constituted by antiparallel β-sheet + β-turn) interacts with the β-sheet of a second GvpA monomer, which might result in the strong interactions between two adjacent ribs.

The structural model of GvpA deduced here defines putative contact sites of GvpA. The major contact sites besides the β-turn at the tip are helix I (near the bottom) and helix II (alongside the dimer). The contacts predicted by the model to be vital to the forming of the vesicle or monomer structure were deleted or substituted in mutagenesis experiments. In summary eight point mutants and three deletion mutants were produced and screened for their ability to produce gas-vesicles (quantitative details in Strunk et al.[100]). All mutations are shown in Fig. 5.24b.

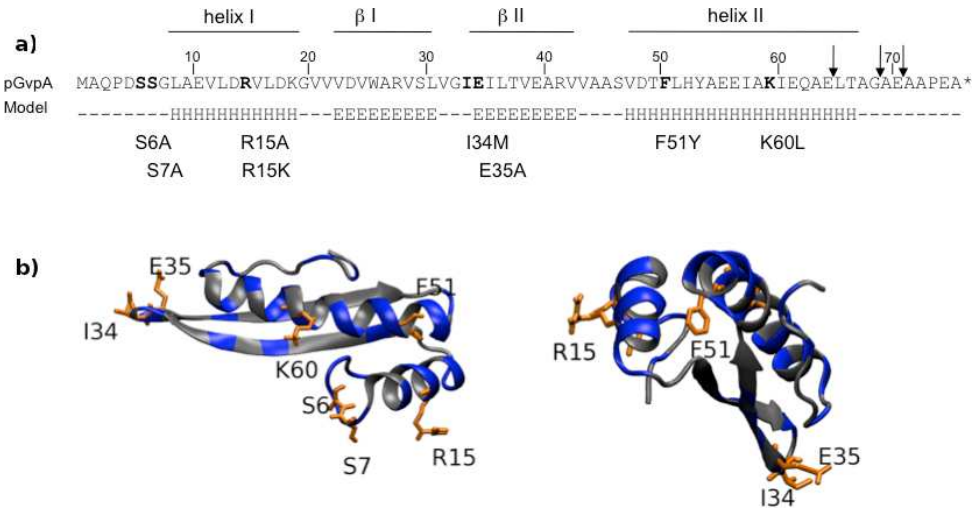


Fig. 5.25.: *Mutations assessing the validity of the GvpA model. a) Sequence and secondary structure for the predicted model of GvpA. Point mutations are indicated below the sequence. Truncations at the C-terminal are indicated by arrows. b) Locations of the single point mutations inside the structure. Mutated sidechains are indicated in orange. Blue indicates hydrophilic residues. Figure courtesy of Strunk et al.[100]*

The single point mutations I34M and E35A (location indicated in Fig. 5.25) changed the overall shape of the gas-vesicles to be long and thin or cylinder-shaped respectively. Similarly the two mutations (F51Y and K60L) in helix II, presumed to be in contact with the adjacent GvpA molecule of the same rib, change the form of the gas-vesicles, but keep them intact. The importance of the second helix was further assessed in the truncation mutants, as the Δ11 mutant (the truncation of the last 11 amino acids and thereby the last two amino acids of helix 2) inhibited all gas-vesicle formation. In contrast Δ5 and Δ7 mutants, which do not truncate the terminal helix, were still able to form gas vesicles. This is compatible with the fact that the C-Terminal of the GvpA family is not highly conserved and even missing in some species (Fig. 5.18). Another indicator for the the lack of importance of the C-terminal in the formation of gas-vesicles is that cleavage at the GluC endopeptidase cleavage sites contained in the C-terminal region left the gas-vesicles intact, while cleavage at the trypsin accessible site K60-I61, within the helix II region, led to the collapse of the vesicles[145].

The structural model shows ARG15 in the center of helix I to play an important role in the aggregation between two GvpA subunits of adjacent ribs. Mutations at this site (R15A and R15K) both led to the inhibition of gas-vesicle formation: ARG15 could be bound to the oppositely charged amino acids of helix I of the adjacent GvpA dimer. In summary the in vivo data on the GvpA mutants support the major structural features deduced from the model of GvpA.

5.5. High-throughput prediction of peptide structures

5.5.1. Motivation

The availability of effective antibiotics is one important contribution to improved health in the last century. As bacteria adapt to and resist current drugs, effectiveness of available antibiotics decreased for some diseases, e.g. tuberculosis[170, 171]. Since 1970 only three novel antibiotic drugs were released to market[172]. Antimicrobial, antibiotic and antifungal peptides[173, 174] present one possible alternative to complement current antibiotics. Recent investigations to develop novel antimicrobial and antibiotic drugs have therefore focused on the development of artificial peptides[175]. Short peptides are naturally involved in many important biological processes in the cell and therefore target many kinds of cells. For antimicrobial and antifungal applications, it is vital to design peptides, which can differentially target bacterial and eucariotic cells. Experimental large-scale screening endeavors introduce changes in the peptide sequence to discover new functional peptides[176–178], but the number of possible sequences is overwhelmingly large. For a peptide of length 20, $20^{20} = 10^{26}$ different possible combinations exist. Although the length of the peptides investigated in this study was limited to 16 amino acids, the number of possible peptide sequences is still too large to synthesize in a trial- and error manner, therefore requiring a method for directed, but also high-throughput peptide design. By predicting the structure of peptide proteins, this design process can be supported through structure-function analysis[179] and peptide-membrane interaction simulation[180], which would improve novel peptide development.
Here we present a method to generate structural ensembles of peptides in an automated manner. In section 5.5.2, we introduce the protocol used for structure prediction and explain the clustering method used to characterize representative low energy conformations. In section 5.5.3, we present the results of the de-novo predictions and analyze the low energy conformations. We conclude in section 5.5.4 and give a short overview of the success of the prediction protocol.[4]

5.5.2. Methods

Simulation Protocol

For a specific peptide sequence the initial random-coil structure was created using Profasi[181]. Bond-lengths were idealized using the standard Rosetta 3.0 idealization protocol[80]. Afterwards a SIMONA Metropolis Monte-Carlo simulation was started. In each Monte-Carlo step we randomly rotated one angular degree of freedom, i.e. the mainchain and sidechain dihedral angles. There are two move classes: In one move class, values for the angles were drawn from Gaussian distributions with a width of ten degrees around the original angles. In the other move class they were selected randomly from a distribution reflecting the naturally occurring distribution of phi- and psi-angles in the Ramachandran plot[182]. To model the Ramachandran plot we

[4] This work was published as part of Strunk et al.[101]. I thank the publisher Springer and all co-authors for the opportunity to publish it as part of my thesis.

used angles randomly drawn from equidistributions of two circles with radii of $45°$ at the centers ($-125°$, $135°$), for the right-handed helical region, and ($-70°$,$-35°$), for the β-sheet region of the Ramachandran plot.
Starting from a completely extended conformation, we performed 5000 simulations, each computing 1.5 million steps using a simulated annealing protocol with a geometrical cooling schedule ($T_{\text{start}} = 700\,K$, $T_{\text{end}} = 5\,K$). The lowest energy conformation was then selected as the predictive model of the experimental structure.

Clustering

The population of structures at the end of the simulations was clustered by RMSD to identify representative low-energy conformations using the following algorithm:

1. The current lowest energy structure is selected from the population.
2. All structures in the vicinity of the current lowest energy structure (RMSD $< 1.6\,$Å) are merged into a cluster.
3. The structures in this cluster are removed from the population.
4. The algorithm repeats from step 1, until the population is empty.

This procedure generates clusters around minimum energy conformations and always selects an energetically favorable centroid structure as a representative conformation of a cluster. Further analysis was based on the minimum energy conformations of all clusters. Only conformations within an energy threshold ($\Delta G < 2.5\,kcal/mol$) to the lowest energy structure were selected for further analysis.

5.5.3. Results

We validated the prediction protocol by investigating four peptides of different topologies. Among these are the one-turn helical fold of the GroES mobile loop 1EGS[183], the Tryptophan-zipper derived β-hairpin 1N0D[184] and two random coil-like folds, the RGD peptide isomer-A 1FUV[185] and the allostatin neuropeptide 2JQU[186] (four letter codes correspond to RCSB PDB ids[3]).
The overlay between the predicted and experimental structure of β-sheet 1N0D is shown in Fig. 5.26a. The predicted structure agrees with the experimental structure to a RMSD of $0.85\,$Å. The RMSD-energy distribution of the ensemble in Fig. 5.26b shows that the cluster with the next higher energy is separated by an energy gap of about $1\,kcal/mol$. It is notable that apart from these two clusters, no other low energy conformation of different topology was discovered. We note that the lowest energy structure features an energetically unfavorable hydrophobic patch (dark red spot in Fig. 5.26a) in agreement with the experiment.
The collapsed coil-fold of 1FUV could be predicted to a RMSD of $2.4\,$Å to the experimental conformation (Fig 5.26c). The cluster closest to the experimental structure has a RMSD of 1.6 Å and is separated by a large gap of $5\,kcal/mol$ to the lowest energy structure.
The fold of 1EGS (Fig. 5.27a) resembles a distorted β-sheet conformation. The stabilizing

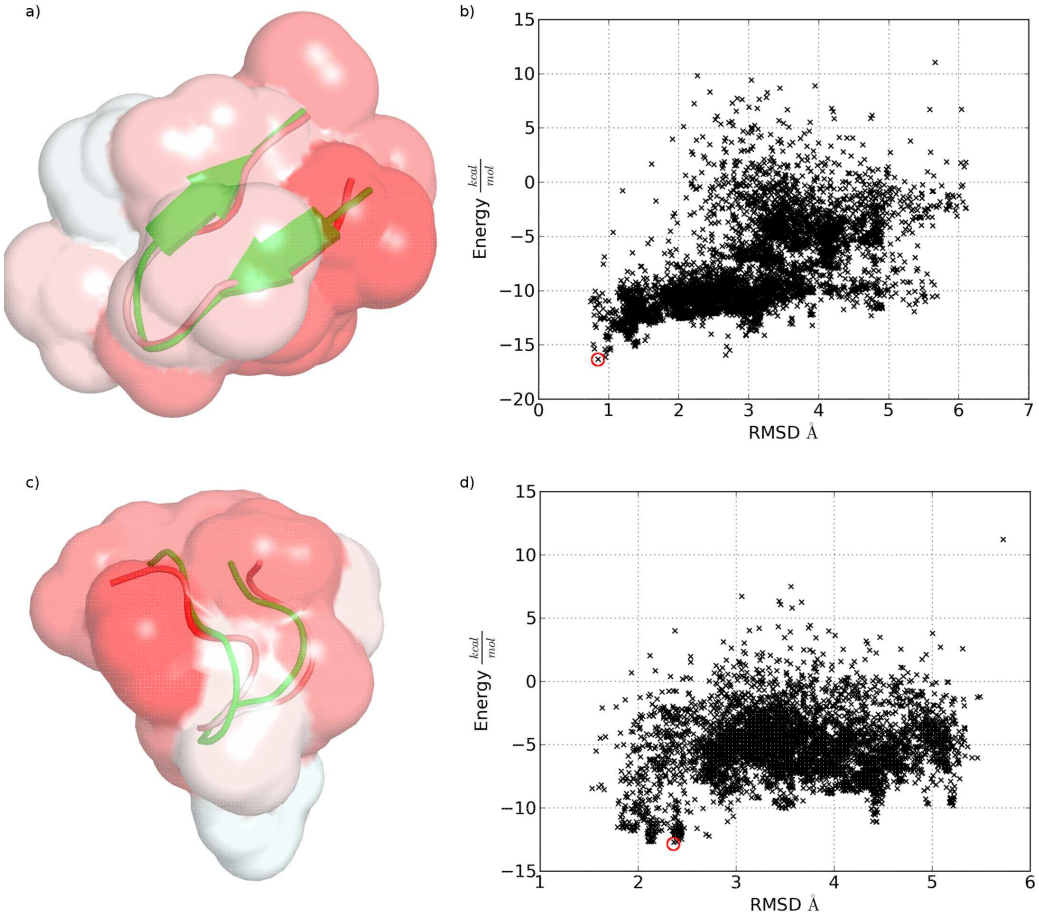


Fig. 5.26.: *a) Result of the predictions of peptide 1N0D (overlay). The predicted β-fold (green) agrees with the experimental structure (red-white) to within experimental resolution. b) RMSD-Energy plot of all simulated structures of 1N0D. The lowest energy structure has an all-atom RMSD of 0.85 Å to the experimental structure. (circle) c) Comparison of the predicted structure of 1FUV (green) with the experimental structure (red) with a RMSD of 2.4 Å. d) RMSD vs. Energy plot of all simulated conformations for 1FUV. Conformations closer in RMSD to the experimental structure were discovered; they were energetically disfavored. a and c: Dark red tones correspond to hydrophobic amino acids. Light red to white tones correspond to hydrophilic amino acids.*

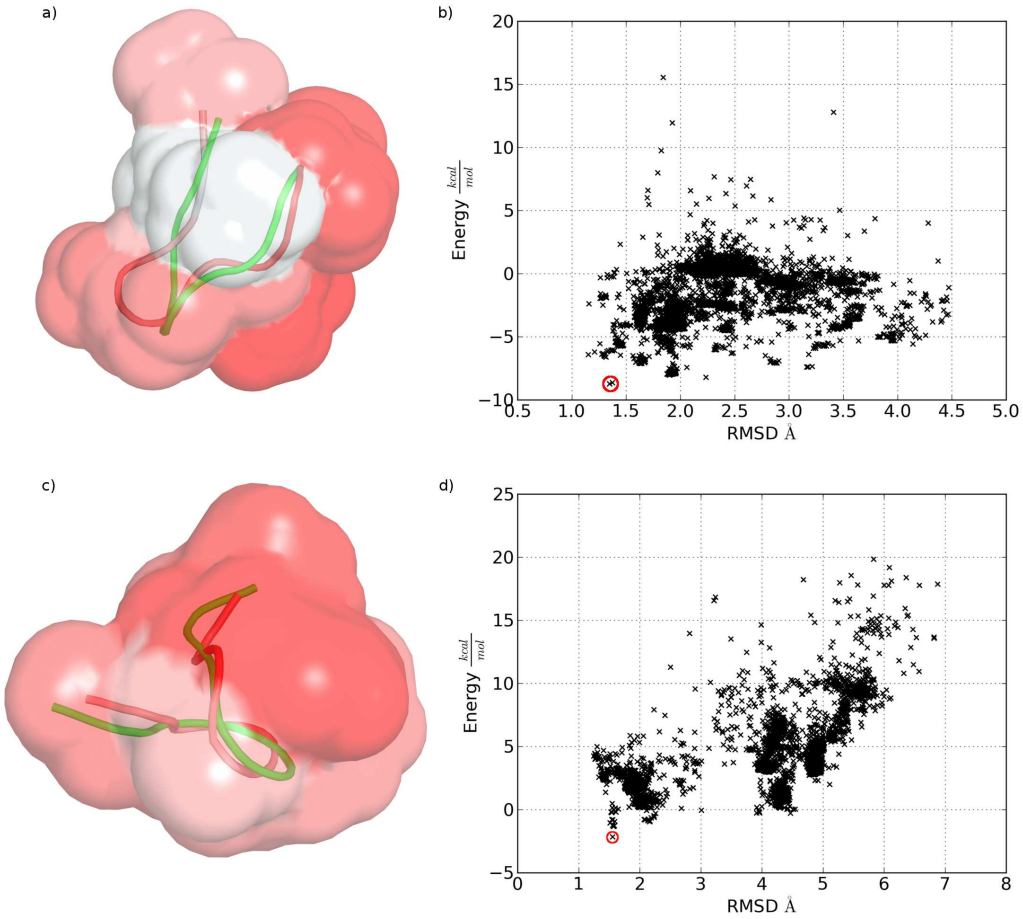


Fig. 5.27.: *a) Result of the predictions of peptide 1EGS. Similar to the experimental structure (red) the predicted structure shows a shift in the β-like fold. b) RMSD-Energy plot of all simulations of 1EGS. The lowest energy structure has an all-atom RMSD of 1.4 Å (circle) to the experimental structure. c) Comparison of the predicted structure (green) of 2JQU with the experimental structure (red). The helical-like fold of 2JQU was correctly identified. d) RMSD vs. Energy plot of all simulated conformations for 2JQU. The population comprises only few structures, which exhibit a smaller RMSD to the experimental conformation than the lowest energy structure at 1.5 Å (circle). a and c: Dark red tones correspond to hydrophobic amino acids. Light red to white corresponds to hydrophilic amino acids.*

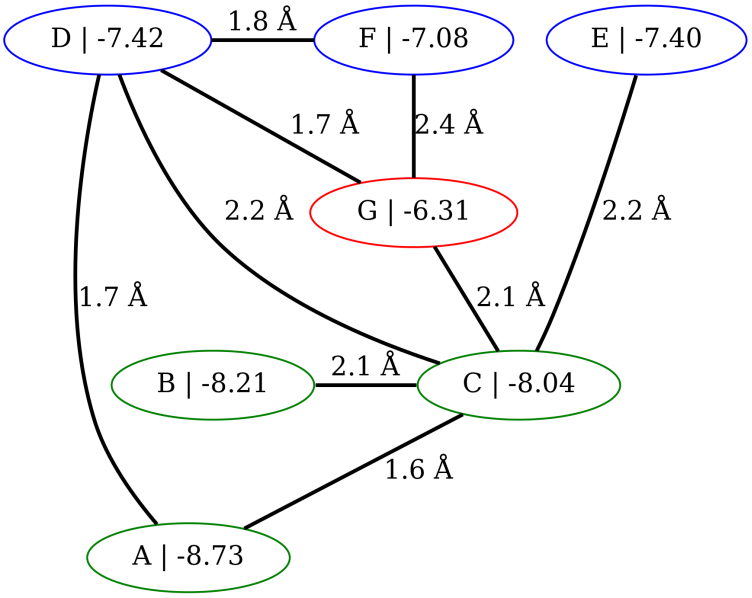


Fig. 5.28.: *Connectivity tree of the distinct low energy clusters of peptide 1EGS below an energy of* $6.3\,kcal/mol$*. The connections show the relative RMSDs between the connected structures, while the labels show the cluster ID and the respective energy in* $kcal/mol$*. IDs correspond to the plots in Fig. 5.29. Clusters with green ellipses are low in energy, clusters with red ellipses are high in energy, clusters with blue ellipses are intermediate.*

zipper-mechanism is not as pronounced as in peptide 1N0D. The lowest energy structure of 1EGS exhibited a RMSD of $1.4\,\text{Å}$ to the experimental structure, which is separated by about $3\,kcal/mol$ from the next unfolded conformation (Fig. 5.27c).
All sampled conformations of peptide 2JQU are displayed in Fig. 5.27d. An energy difference of $3\,kcal/mol$ separates the lowest energy structure and the next sampled unfolded conformation. There was no pronounced low energy conformation sampled with a RMSD bigger than $4.5\,\text{Å}$. This can be rationalized by the fact that the native helical conformation (Fig. 5.27c) is stabilized mostly by local interactions. The funnel towards the helical structure might therefore be so pronounced that no other structure of completely different topology was visited in the simulation.

Low energy conformations of 1EGS

The low energy conformations of 1EGS were analysed using the clustering scheme presented in section 5.5.2. Seven distinct low energy conformations were identified within a range from $-8.73\,kcal/mol$ to $-6.31\,kcal/mol$. The connectivity of the clusters is shown in Fig. 5.28. Two clusters were considered adjacent, if the RMSD of their centroids was smaller than $2.5\,\text{Å}$. The minimum energy conformation is connected to two distinct conformations with an RMSD of $1.7\,\text{Å}$ and $1.6\,\text{Å}$, respectively. Structure C (Fig. 5.29) is the cluster with the most connections and acts like a ”travel hub“ among the low energy conformations. This may be understood, by analyzing the conformation of structure C. Fig. 5.29 C shows that structure C features a relaxed β conformation, which resembles the experimental conformation (Fig. 5.29 H), the broken β conformation (Fig. 5.29 B) and the partially helical coiled conformation (Fig. 5.29 D). The

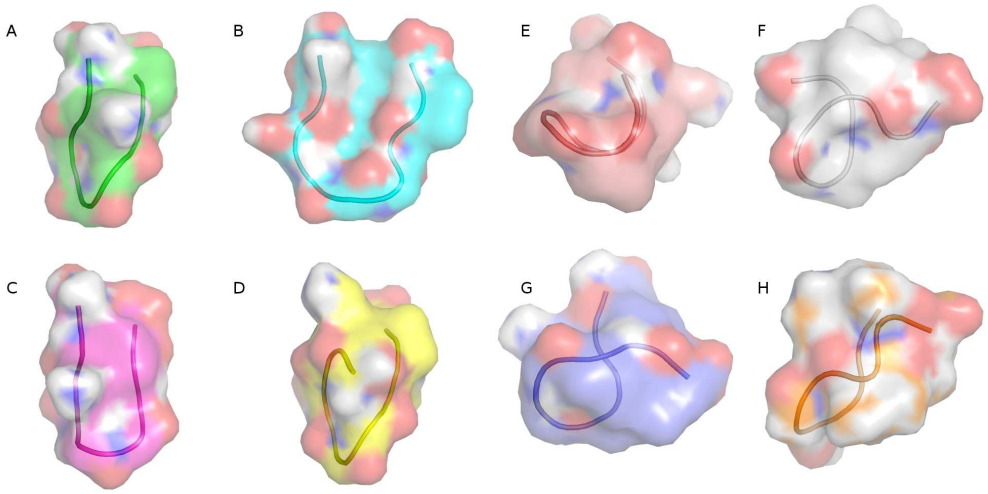


Fig. 5.29.: *Low energy conformations of peptide 1EGS. Energies are sorted in increasing order from A to G. The structure H is the experimental conformation shown for comparison. Conformations C and G can be converted to most of the other conformations as only a slight shearing move is necessary to transform them into either coiled or β-conformations. The connectivity tree for these structures is shown in Fig. 5.28.*

higher energy structures included a warped β-sheet (Fig. 5.29 E) and a coiled helical turn similar to structure (C) in Fig. 5.29 F.
Traversing the low energy landscape of protein 1EGS can therefore be represented by a shearing motion for most of the peptide structures, which transforms many of the low energy conformations into one-another either by increasing the shear, leading to a coiled structure, or reducing the shear, leading to the correct β-fold.

5.5.4. Discussion

In this investigation we have developed a technique to predict peptide structures de-novo, i.e. based on the sequence information alone, using a massively parallel simulation scheme. We sample the peptide's conformational space using Monte-Carlo simulations in the free-energy forcefield PFF02[9] on the volunteer computing network POEM@HOME[71]. We could identify the native conformation of peptides of different topologies in a completely automated manner that allows for the high-throughput screening of large peptide databases for their structural features, enabling the rapid prototyping needed for novel peptide design. Due to the short peptide length, homology modeling and fragment based modeling attempts will not work, as even single amino acid mutations can have a large impact on the function and therefore the structure of a peptide[187]. We could generate structural ensembles for peptides of very different structure, including collapsed folds without apparent secondary structure. The simulations did not only elucidate the biologically active structure of the peptides, but could also characterize their low energy ensemble, which might be involved during the folding process of the peptide. This method may enable further analysis to establish structure function relationships for peptide libraries and thereby help optimize their biological activity. Due to the automated nature of this prediction scheme, many different sequences can be structurally characterized in a short time allowing the integration of this method in experimental peptide-design investigations.

6. Protein-Ligand Interactions

Many processes in cells are regulated by the association of proteins with other proteins and small-molecule ligands[188, 189]. The manipulation of these processes presents an opportunity for the development of novel drugs. The experimental characterization and directed targeting of these interfaces is a complicated and costly process and requires the analysis of millions of ligands[190, 191]. In-silico protein-ligand screening may present a viable alternative to alleviate some of these costs[192].
In this chapter, we discuss simulation protocols that may contribute to the development of novel therapeutic agents. In section 6.1, we first present a method for identifying specific hotspots as possible therapeutic targets for structure-based drug design. In section 6.2, we report the first applications of SIMONA to protein small-ligand docking. Our group has developed FlexScreen[193], a high throughput receptor ligand docking program, for the last decade. For technical reasons this implementation has now reached the limits of its efficiency. Implementation in the general purpose simulation program SIMONA would make it possible to rapidly develop highly efficient protocols for in-silico screening using all the tools and force fields available within SIMONA.[1]

6.1. Computational Alanine Screening

6.1.1. Motivation

Many different biological signaling processes are mediated by protein-protein interactions. Understanding of protein-protein interactions gives insights into a wealth of biological processes and may even enable their manipulation. Protein-protein interfaces are emerging as novel drug targets[195]. In comparison to interactions between proteins and small-molecule ligands, protein-protein interfaces are more extended. While targeting these extended interfaces with antibodies has been successful[190], smaller ligands are more desirable, because of their lower cost, ease of handling and better bioavailability. Due to the often large area of protein-protein interaction sites, it is difficult to modify the attachment site to affect binding[196]. In spite of these difficulties, high-throughput screening studies were successful targeting protein-protein interfaces with small-molecule ligands[197–199]. Designed binding partners, which are derived from the natural ligand, were used to inhibit protein-protein binding and served as a model for

[1] The two protein-protein docking simulations of 1MFG and 1ILP were previously published in Meliciani et al.[194]. The third docking simulation and the cascaded strategy was published in Strunk et al.[100]. I thank all of the co-authors and especially the publishers American Institute of Physics (Meliciani et al.) and John Wiley and Sons (Strunk et al.) for their permission to publish these results again as part of my thesis. In the investigation presented in this chapter, I was responsible for the brute-force protein-protein and protein-ligand docking simulations without prior knowledge of the docking interface and the forcefield implementations. Alexander Biewer parametrized the FlexScreen forcefield in SIMONA and did most of the work on the simulation cascade.

the development of small-molecule analogs[200–202].
Alanine exchange screening is a widely used method to discover and characterize binding hotspots, i.e. key interactions between the two interacting proteins. The amino acids located at these hotspots provide most of the binding energy supporting a stable complex. After identification of the binding interface, charged, polar or bulky amino acids in its vicinity are mutated in alanine screening experiments to the non-polar alanine residue and the binding affinity is measured[203–205]. After identification of the binding hotspots, ligands can be designed to target the most important amino acids in the interface and compete with the natural binding partner.
Experimental alanine screening is a time-consuming process, as it requires creation of large mutagenesis assays. Simulation of the properties of the binding interface can reduce the experimental work involved by estimating the differences in interaction energy upon mutation and therefore guide the experiment. Additional insight about the the mode of binding can be obtained by analyzing the energies contributing most to the binding mechanism[2].
In section 6.1.2, we will introduce the method for computational alanine screening and the validation protocol. In section 6.1.3, we present the mutational protocol. The results of our simulations are presented in section 6.1.4. In section 6.1.5, we discuss our findings in the context of experimental screening investigations.

6.1.2. Introduction

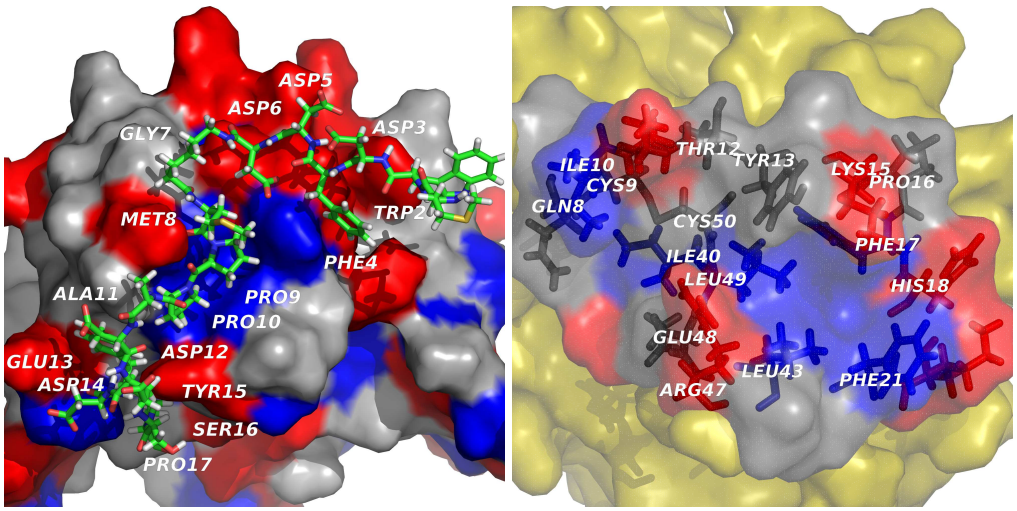
Many computational methods for computational alanine screening have been reported in the literature[206–210]. The bulk of these methods are based on Molecular Dynamics simulations, the first of which investigated a complex of oncoprotein Mdm2 bound to a peptide from tumor suppressor protein p53[211]. Ideally such simulations should estimate the difference in the free energy of binding $\Delta\Delta G$ upon mutation, but estimation of this quantity is notoriously difficult in Molecular Dynamics simulations and their derivatives. For this reason many sequence- and knowledge-based methods have been developed to provide rapid and computationally efficient predictors for the affinity change.
Here we investigate a Metropolis Monte-Carlo-based protocol to study protein-protein interactions and their modification under mutation and validate the approach on two pharmaceutically relevant complexes. We calculate the binding energy of the wild-type complex from the experimental structure and subsequently mutate each of the amino acids in the vicinity of the binding interface to alanine. We then compute the binding energy and thus obtain an estimate of the enthalpic contribution to the free energy change. Hot-spots are being identified as the sequence positions, where a significant energy change was induced by alanine substitution. We rationalize the energy changes by decomposition of the total binding energy into the most important physical contributions.

[2]Parts of this section have been published in Meliciani et al.[194]. I thank the publisher AIP and all the authors for their permission to publish them as part of this dissertation. In the presented project Dr. Irene Meliciani investigated the specific systems and their properties and performed the subsequent analysis. I was responsible for the development of the simulation and evaluation scripts and their implementation on BOINC. Dr. Konstantin Klenin took care of the mutagenesis scripts and adapted the simulation parameters.

First we screen the influence of mutations for the complex of the chemokine Interleukin-8 (CXCL8) and a N-terminal peptide mimic of its cognate receptor CXCR1[212]. The family of chemokine proteins direct the movement of cells carrying receptors for chemokines towards sites of inflammation. Due to the large number of cells carrying chemokine receptors, many chemokines are involved in clinical conditions like chronic inflammations, allergies, autoimmune diseases and even cancer[213, 214]. We further tested our approach by investigating the ERBIN/ERBB2 complex[215, 216] and identify the most important interaction hot spots in agreement with experimental data[215, 217].

6.1.3. Methods

To predict the change in interaction energy between receptor and ligand upon mutation, we first anneal the compound structure to a near minimum in the selected forcefield PFF02 (see section 3.2.2) using long simulated annealing simulations. From this simulation we obtain the internal free-energy of the complex before mutation G_C, which comprises the inter- and intra-molecular contributions to the energy, as well as an approximation to the solvent interactions, including the solvent entropy. It is important to note that this estimate contains no contribution of the backbone entropy of either binding partner. After moving both binding partners apart, we calculate their internal free-energies G_R and G_L in the unbound state. We define the energy of binding as $\Delta G_B = G_C - (G_R + G_L)$.
We then repeat the simulations, after mutating a single interface residue to alanine to obtain the energy after mutation G_B^a. The change $\Delta\Delta G$ in binding energy is then $\Delta\Delta G = \Delta G_B^a - \Delta G_B$. Neglecting entropic effects in computing the differential binding energy is justified if the change in the free-energy of binding is dominated by the enthalpic effect. In the simulations reported below, we have investigated only mutations of the peptide-ligand, which implies that the majority of the missing entropic correction is likely to arise from contributions to the entropy of the unbound peptide. The overall interaction energy ΔG is therefore overestimated using this protocol.
We modeled the mutation of a single residue to alanine by removal of all sidechain atoms except for the β-carbon atom C_β. Hydrogen is then placed into the now open bond positions of the β-carbon. For all relaxation and annealing simulations, moves for the dihedral-angles are drawn from an equidistributed interval with a maximum change of 5°. Rigid-body transformations (rotations and translations) and dihedral angle rotations are selected at a ratio of 1:2. The molecule is rotated by a random angle, drawn from an equidistribution with maximal rotation of 5° around a random axis, or displaced along a random axis by a maximal radius of $1\,\text{Å}$ also drawn from an equidistribution.
In the relaxation simulations, we annealed the native and mutated structures for $50,000$ steps starting at $300\,K$ and ending at $2\,K$ using a geometric cooling schedule. The low final temperature is selected to differentiate between low-lying metastable conformations. Unless otherwise specified only sidechain dihedral angles and rigid-body motions were sampled in this investigation. For more complex systems, in particular when backbone reorganization is expected to play a significant role, more complex simulation protocols may be required[68, 218–220]. The



(a) Receptor-peptide present (b) Only the ligand is shown

Fig. 6.1.: *Structure of Interleukin-8 (CXCL8) a) Interleukin-8 (surface representation) in complex with a peptide derived from its native receptor CXCR1 (drawn in stick representation). Red surfaces correspond to charged amino acids, blue surfaces correspond to hydrophobic amino acids. The receptor-peptide fits into the rift on the surface of Interleukin-8. b) Isolated Interleukin-8. Only amino acids in contact with the CXCR1 receptor peptide (not drawn) are colored according to a). Additionally yellow amino acids are not in contact with CXCR1. Large hydrophobic patches are covered by CXCR1 in the bound state. Many charged amino acids are adjacent to the rift (red).*

simulations were carried out on the POEM@HOME network[71, 221], allowing us to simulate all possible mutations of a single ligand in 1000 independent simulations.

6.1.4. Results

The NMR structure of Interleukin-8 (ligand) bound with a peptide derived from its native receptor CXCR1 contains the non-natural amino acid ACA (PDB code: 1ILP)[212]. For parametrization in PFF02 this non-natural amino acid was mutated to the naturally occurring glycine. The NMR structure of the receptor-ligand complex is displayed in Fig. 6.1. The figure shows the close fit between ligand and receptor: The receptor-peptide fills a rift inside the surface of the ligand structure. The N-terminal part (MET1-TRP2) of the peptide shows a high degree of disorder in the NMR ensemble. As it also exhibits no contacts to the binding partner, it will presumably not partake strongly in the stabilization of the binding.

Prior to the alanine screen, the native complex is annealed into its pocket to establish the reference energy in the forcefield PFF02. In this simulation, mainchain dihedral angles were perturbed. The obtained energy-optimized conformation differs by 2.25 Å RMSD from the NMR structure. Convergence of this simulation was verified by long simulated annealing simulations ($500,000$ steps). After moving the receptor away from the ligand structure the energy was recalculated.

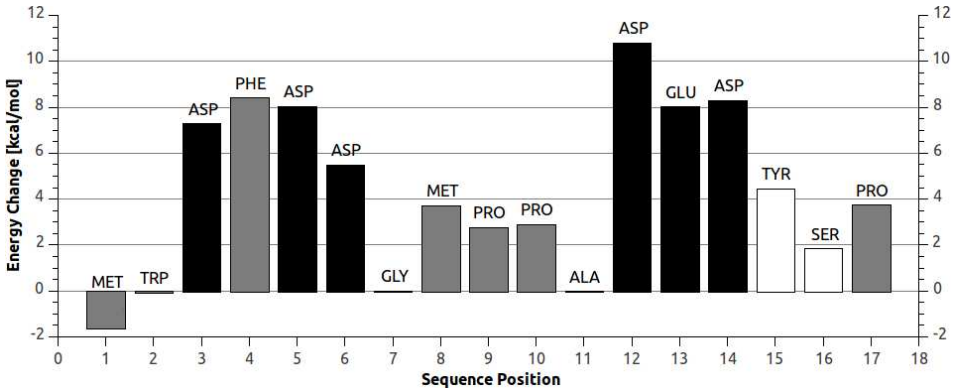


Fig. 6.2.: *Interaction energy difference for all alanine mutations of the receptor peptide CXCR1. The bar color indicates the sidechain properties: charged sidechains (black), nonpolar sidechains (gray), polar sidechains (white).*

Mutational analysis of CXCR1

Each of the 17 receptor peptide amino acids was mutated to alanine. The resulting estimates of the changes in the free energy of binding are shown in Tab. 6.1. Mutations in the N-terminal part (MET1,TRP2) did not contribute significantly to the stabilization of the complex as expected, as their alanine mutations had little effect on the interaction energy. The mutations of most non-polar and polar sidechains (GLY7, MET8, PRO9, PRO10, TYR15, SER16, and PRO17) also exhibited only a slight energy change upon mutation. The only exception was the non-polar PHE4 (see Fig. 6.2). Mutation of charged amino acids had the most severe effect on the interaction energy (ASP3, PHE4, ASP5, ASP6, ASP12, GLU13, and ASP14), with the largest impact being the ASP12 hotspot.

The NMR structure (Fig. 6.1, right panel) features a large number of charged or polar residues in the vicinity of the binding site explaining the large impact of mutation of charged residues. Only a small change in solvation energy is observed for most polar amino acids since the polar residues are exposed in the complex and the isolated molecules. To better understand these results the energy contributions of the interactions were analyzed independently (solvation, main-chain and sidechain- electrostatics, torsion potential and Lennard-Jones interactions). Pair-wise interactions were assigned to both of the interaction partners with half interaction strength. We found that most of the interaction energy can be assigned to the sidechain electrostatic interactions of the charged amino acids. Contributions to the backbone electrostatics energies are almost unchanged, because the mainchain conformation is unchanged and intermolecular contacts are mediated by sidechain atoms. Apart from the electrostatics hot-spots the amino acid PHE4 has a large contribution towards the binding energy, due to the large change in solvation energy upon its mutation.

In summary the dominant changes in interaction energy can be attributed to hydrophobic effects for most polar and hydrophobic amino acids and to electrostatic contributions for charged amino acids. Except for PHE4 the important energy differences arise from electrostatic interactions (see Tab. 6.1). Solvation energies are favorable for PHE4 in comparison to the mutated ALA4, since PHE4 covers the hydrophobic pocket of the ligand (PHE17, PHE21 and LEU47; IDs are relative to the ligand) better than ALA4 and also contributes a destabilizing interaction

Res.	Amino	Solvation	Electrostatics		Backbone	Lennard	Total
No.	acid		Sidechain	Mainchain	Torsion	Jones	
1	Met	0,90	−2,82	−0,30	0,09	0,48	−1,65
2	Trp	0,85	−1,21	−0,29	0,09	0,50	−0,06
3	Asp	0,90	6,14	−0,30	0,09	0,48	7,31
4	Phe	7,79	−0,05	−0,24	0,09	0,84	8,43
5	Asp	1,02	6,95	−0,54	0,03	0,59	8,05
6	Asp	−1,01	6,11	−0,30	0,09	0,62	5,51
7	Gly	−0,52	0,29	−0,30	0,09	0,45	0,01
8	Met	3,07	0,29	−0,30	0,09	0,59	3,74
9	Pro	2,09	0,25	−0,31	0,09	0,67	2,79
10	Pro	2,07	0,20	−0,23	0,09	0,78	2,91
12	Asp	1,09	9,42	−0,30	0,09	0,53	10,83
13	Glu	1,29	6,23	−0,30	0,09	0,72	8,03
14	Asp	0,90	7,14	−0,30	0,09	0,48	8,31
15	Tyr	4,15	−0,07	−0,30	0,09	0,59	4,46
16	Ser	1,28	0,29	−0,30	0,09	0,50	1,86
17	Pro	2,99	0,25	−0,26	0,09	0,69	3,76

Tab. 6.1.: *Result of the alanine substitutions of the individual amino acids in the receptor peptide derived from CXCR1. Energy contributions are decomposed into solvation, electrostatics, torsion and Lennard-Jones potentials. Numeric values are in $kcal/mol$. The large changes in interaction energy can almost exclusively be attributed to changes in solvation and sidechain-electrostatics contributions. Darker values present hotspots in the peptide. Light-Dark values are significant contributions. Residue ALA11 is not listed.*

in the unbound state due to its own exposed hydrophobic surface.
Experimentally the residues PRO9, TYR15, PRO16, ASP12, GLU13, and ASP14 were identified as hotspots by Skelton et al.[212], while another study also attributed a crucial role to ASP3[222, 223]. The data in Tab. 6.1 demonstrate that these hot-spots were also identified by this study. No false negative was reported. Additionally a stabilizing contribution was observed for PHE4, ASP5, ASP6, MET8, PRO10, and SER17.

Mutational analysis of the ERBIN/ERBB2 complex

The second system studied here is the complex of a 15 amino acid C-terminal peptide of human receptor tyrosine kinase ERBB2 bound to the PDZ domain of the ERBIN (ERBb2 Interacting proteiN) protein. The ERBB2 C-terminal domain contains two recognition motifs. It was simulated as a benchmark system as it is widely studied by mutagenesis and functional studies in cellular environments[215, 216, 224]. One important hotspot identified in these studies was a critical tyrosine residue (TYR2 in the following investigation) in a recognition motif of ERBB2, where substitution resulted into the misdirection of the mutant protein. A similar study found that even phosphorylation of the tyrosine residue inhibits binding between the two domains[225].
Similar to the previous analysis of Interleukin-8, the simulations started using the experimental structure of the ERBIN-PDZ domain bound to the C-terminal tail of the ERBB2 receptor (PDB

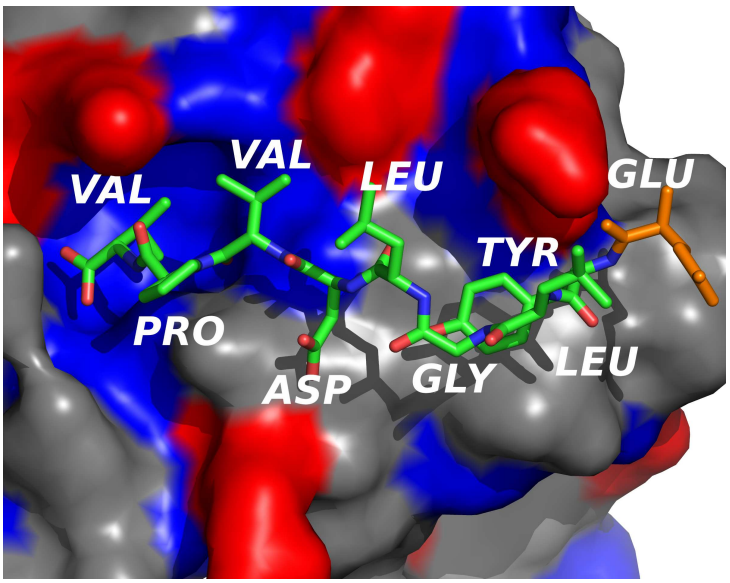


Fig. 6.3.: *Structure of the ERBIN PDZ domain bound to the C-terminal tail of the ERBB2 Receptor (PDB: 1MFG). The ERBB2 receptor peptide (stick representation) docks into a burrow of charged and hydrophobic amino acids on the surface of ERBIN (surface representation). The sidechain of TYR2 (left from the orange peptide N-terminal) is located below the mainchain adjacent to the surface. Red surfaces correspond to charged amino acids, blue surfaces correspond to hydrophobic amino acids.*

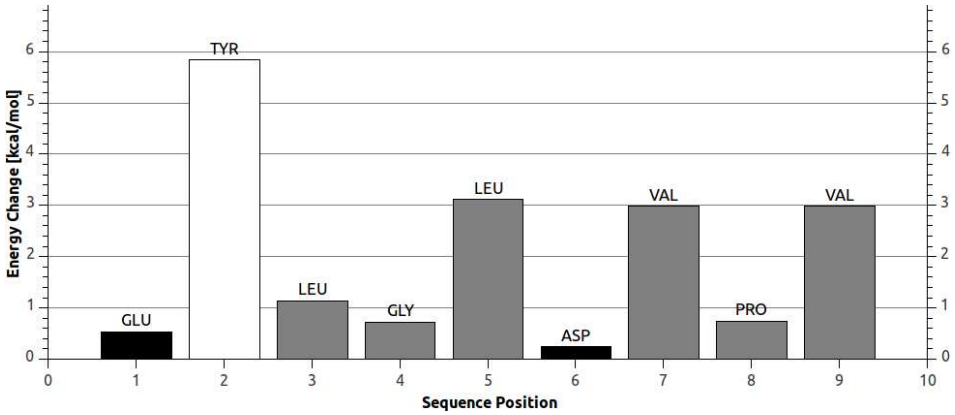


Fig. 6.4.: *Result of the computational alanine screen of the ERBIN receptor peptide. Especially mutation of TYR2 resulted in a large change in interaction energy. Color coding: charged sidechains (black), nonpolar sidechains (gray), polar sidechains (white).*

Res. No.	Amino acid	Solvation	Electrostatics		Backbone Torsion	Lennard Jones	Total
			Sidechain	Mainchain			
1	Glu	0.53	−0.11	−0.05	0.05	0.12	0.54
2	Tyr	5.42	−0.23	−0.05	0.05	0.66	5.85
3	Leu	0.77	0.1	−0.06	0.05	0.28	1.14
4	Gly	0.53	0.1	−0.05	0.05	0.1	0.73
5	Leu	3.21	−0.36	−0.05	0.05	0.28	3.13
6	Asp	−0.46	0.44	−0.05	0.05	0.27	0.25
7	Val	2.72	0.1	−0.05	0.05	0.17	2.99
8	Pro	0.54	0.13	−0.12	0.05	0.15	0.75
9	Val	2.42	0.1	−0.06	0.05	0.48	2.99

Tab. 6.2.: *Energy contributions (in kcal/mol) of alanine substitutions of individual amino acids in the N-terminal peptide of ERBIN. The total energy is split into contributions by solvation effects, electrostatics (sidechain and mainchain), backbone torsion and Lennard-Jones potential. Large changes in interaction energy can be attributed to solvation effects in case of polar and nonpolar residues. Compared to the Interleukin-8 screen no significant changes in binding energy can be attributed to electrostatic effects. Darker colors indicate stronger interactions.*

Code: 1MFG, shown in Fig. 6.3) in agreement with the experimental observations. Only the alanine mutation of TYR2 could show a significant change in binding energy of $5.85\,kcal/mol$, mostly due to a rise in solvation energy (Fig. 6.4 and Tab. 6.2). The increase in solvation energy relates to the burial of a hydrophobic pocket (SER1296, GLY1301, ASN1304, and PRO1305, IDs relate to ERBB2) by TYR2. Furthermore TYR2 develops a hydrogen bond with ASN1304 (ERBB2) in the experimental structure. Mutation of TYR2 to alanine exposes the pocket and therefore increases the solvation energy.
In contrast to the previous investigation of Interleukin-8 most hotspots can be attributed to changes in solvation energy. Less significant solvation contributions were observed for LEU5, VAL7 and VAL9. The only significant hotspot observed was TYR2 (Tab. 6.2).

6.1.5. Discussion

The understanding of key interactions in protein-protein association is vital to allow design of small-molecule ligands binding to a protein-protein interface. Large databases of protein-protein interaction hotspots were accumulated by experimental studies[10, 226]. These were used to develop and benchmark computational methods for prediction of interaction hotspots[227–230] or to provide insight into the common means of protein association, like densely packed regions of conserved polar residues[231–234].
Previous approaches to predict the importance of hotspot residues included the physical and knowledge-based neural network approach *KFC*, which had been trained to recognize features of interfaces from existing structural data[229, 235] and an approach trained on data based on the sequence environment, environmental trace and the accessible surface area; information obtainable from sequence data alone[230]. Although these bioinformatics-based approaches are far less computationally demanding than the one presented here, they do not always offer insight into the structural mechanisms underlying binding energy changes. Qualitative insight into such

Method	True Positive	False Positive	True Negative	False Negative	Precision	Recall
PFF02	8	10	8	0	44%	100%
FoldX	3	5	15	3	38%	50%
Robetta	2	2	17	5	50%	29%
KFC	3	1	18	4	75%	43%

Tab. 6.3.: *Performance of our method (PFF02) in comparison to three other approaches FoldX, Robetta, and KFC. Precision is the number of true positives divided by the number of predicted hot spots; recall is the number of true positives divided by the number of experimental hot spots.*

effects could be gained by multiple structure alignments of conserved residues in the vicinity of the binding interface[231, 232, 236].

In comparison to the knowledge-based approaches, the energy-based approach presented here permits a quantitative analysis of the binding interface. Energy-based approaches can further be divided into empirical bioinformatics-based approaches, such as functional matrices[237] and molecular mechanics based approaches using physical forcefields such as MM-GBSA, a generalized Born surface area (GBSA) based approach, which estimates the free-energy of binding by calculating gas-phase energies, solvation free energies, and entropic contributions for the free proteins and the complex[162, 206, 207, 211, 236, 238–240]. In contrast to MM-GBSA, the approach presented here is less computationally expensive. We approximate the enthalpic contributions to the free-energy of a protein complex in the bound and unbound states using the PFF02 forcefield, which also models part of the solvent entropy. In contrast to MM-GBSA, backbone entropy contributions are not considered in our approach.

The accuracy of the approach studied here is dependent on the quality of the underlying forcefield. To compare it with other similar approaches, we recalculated the energies given in the results section 6.1.4 with the ROBETTA server (and its corresponding potential)[74, 241], the FoldX server[227, 242] and the KFC server[229, 235]. A mutation was considered indicative of a hotspot if a binding energy change of more than $1\ kcal/mol$ was observed (Tab. 6.3). Only our method predicted all seven hot spots of the IL-8 receptor peptide, while ROBETTA, FOLDX, and KFC predicted one, two, and three hot spots, respectively. Additionally six false positive hotspots were identified, while ROBETTA, FOLDX, and KFC yielded only one, two, and one false positive. Three of the four methods considered PHE4 of the IL-8 receptor peptide to be a hotspot. The involvement of PHE4 in binding could be underestimated by current experimental methods.

With the exception of KFC all methods identified the TYR2 hotspot in the ERBB2/ERBIN complex. PFF02 generates three more false-positive structures with minor energy contributions compared to TYR2. Precision and recall statistics are shown in Tab. 6.3. Precision is the fraction of true positives divided by the sum of true and false positives. Recall is the fraction of true positives divided by the number of experimentally determined hotspots[229]. The precision was 44% for our method, comparable to the one of ROBETTA (50%) and FOLDX (38%). While KFC had the best precision (75%) in the Interleukin-8 system, it did not identify the hotspot in the ERBB2/ERBIN complex.

In summary PFF02 yields more false positives than the other methods; no false negative was

reported. Most of the false positives are related to an overestimation of the electrostatic contribution. The overestimation of the electrostatic contributions has since been addressed by the development of a new generalized Born electrostatics and implicit solvent model. Preliminary results are available in Strunk et al.[38].

6.2. De-Novo Protein-Protein and Protein-Ligand Docking

6.2.1. Motivation

The development of SIMONA aimed to replace the three in-house programs for protein folding simulations (POEM)[9, 61], for in silico screening (FlexScreen)[193] and for the modeling of thin-film materials (DEPOSIT). FlexScreen performs high throughput small-molecule protein docking simulations to discover and optimize novel small-molecule ligands. In contrast to other in-silico docking tools FlexScreen is based on a biophysical model of the interactions and incorporates very efficient sampling techniques to model structural changes in the receptor and the ligand during the binding process[243]. The field of in silico-screening has increasingly recognized the importance of receptor flexibility in the binding process, which was simply too costly to sample when these methods were introduced two decades ago[244]. With the incorporation of receptor flexibility many of the components required for in-silico screening now overlap with the techniques required to model protein conformational change. For this reason, we implemented the FlexScreen forcefields[193] in SIMONA and establish and validate its novel docking protocols for various benchmark systems.
In section 6.2.3 we validated a very simple brute-force docking protocol for the streptavidin biotin complex. Using the enhanced capabilities of SIMONA it is possible to apply the same type of protocol not only to protein small-molecule docking, but also to protein-protein docking, which was not possible with FlexScreen. We also present a cascaded docking strategy, which was merged into SIMONA in the scope of the diploma thesis of Alexander Biewer[245]. In section 6.2.4 initial results for protein-protein docking simulations are reported for three test cases: ERBIN-PDZ domain bound to the C-Terminal peptide derived of its receptor ERBB2 (PDB: 1MFG), the complex of Interleukin-8 with its native receptor peptide CXCR1 (1ILP) and the fire ant venom homodimer (2YGU). The two systems 1MFG and 1ILP were discussed in the previous section concerning in-silico alanine screening (section 6.1).

6.2.2. Methods

The docking simulations are initiated from the experimental structure. Prior to the simulation, the ligand is offset multiple times from the receptor by a random translation and rotation. Each structure is simulated using Metropolis Monte-Carlo including rigid-body (rotations and translations) and sidechain degrees of freedom. The quantitative details, such as stepcount, temperatures and forcefield used for the simulations are located in the respective results sections.

6.2.3. Results of the Protein-Ligand docking benchmarks

Protein-Ligand Docking was benchmarked using the streptavidin-biotin complex (PDB-ID: 1STP), one of the most common examples for protein-ligand docking, as its binding energy is one of the highest measured for noncovalent binding systems[246]. It is therefore one of

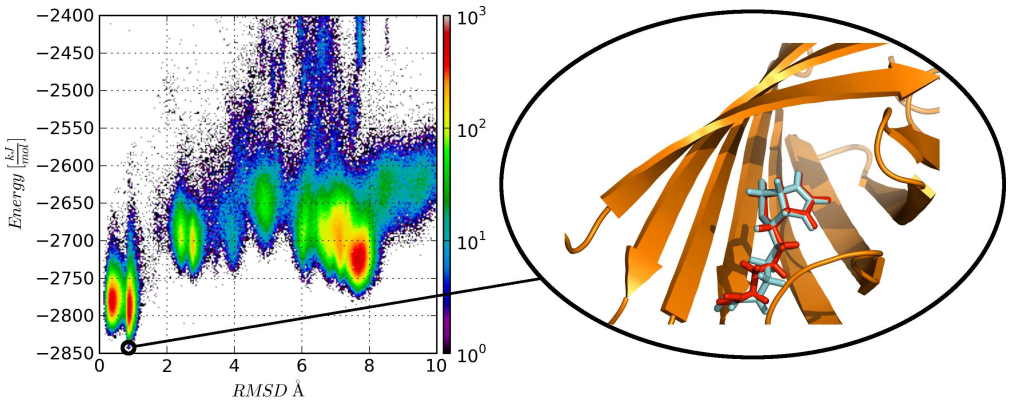


Fig. 6.5.: *Results of the protein-ligand docking of a streptavidin-biotin complex done in SIMONA. Displayed are* $660,720$ *structures with an energy below* $-2400\,kJ/mol$ *and a RMSD below* $10\,\text{Å}$*. The correct binding pose was discovered at* $0.88\,\text{Å}$ *to the native structure.*

Stage	No. of poses	No. of steps	Forcefields
0	96	1000	E_{LJ}
1	96	5000	$E_{LJ} + E_C + E_{HB} + E_{ISE}$
2	16	30000	$E_{LJ} + E_C + E_{HB} + E_{ISE}$
3	8	75000	$E_{LJ} + E_C + E_{HB} + E_{ISE}$

Tab. 6.4.: *Configuration of the cascaded docking approach. The first stage only comprises a short relaxation simulation inside a Lennard Jones potential to remove steric overlap in the initial structure. Further stages comprise extended simulation times and incorporate the full FlexScreen potential[193]. Only the last eight remaining lowest energy structures are relaxed for the full runtime of* $111,000$ *MC steps.*

the most popular benchmark systems for protein-ligand docking. We relaxed randomly placed ligand conformation at constant temperature $T = 300\,K$ for $300,000$ Monte-Carlo steps. The maximum translational displacement was set to $2.0\,\text{Å}$. The ligand was completely flexible in dihedral space, but we kept the receptor backbone dihedrals fixed and allowed the rotation of 29 sidechain dihedrals around the binding pocket. All energies were evaluated in the FlexScreen forcefield[193].

The ligand in the 1STP docking simulations was docked into the correct docking pose. The low-energy subset of the results of this brute-force sampling approach, including the $660,720$ structures with an energy below $-2400\,kJ/mol$ (absolute energy) and an all atom RMSD below $10\,\text{Å}$ is shown in Fig. 6.5. The structure of lowest energy has a RMSD of $0.88\,\text{Å}$ to the native structure and therefore models the correct docking pose to experimental resolution.

The results presented here should be understood as a proof-of-concept. While elucidating the mode of binding for protein-protein interactions is possible with the presented protocol, efficient drug design requires protocols that are far less computing intensive.

FlexScreen[10] implemented a four-stage cascaded docking approach using the prior knowledge of the docking pocket.

In this protocol, the ligand is placed into the docking pocket, displaced randomly by maximum displacement of $2\,\text{Å}$ and randomly perturbed $10,000$ times by rotations around single bonds, which are identified by an automated procedure. These structures enter stage 0 of the cascade as starting conformations. The idea of the cascades is to perform many short simulations on a large possible set of ligand poses and then to select the lowest energy poses at the end of

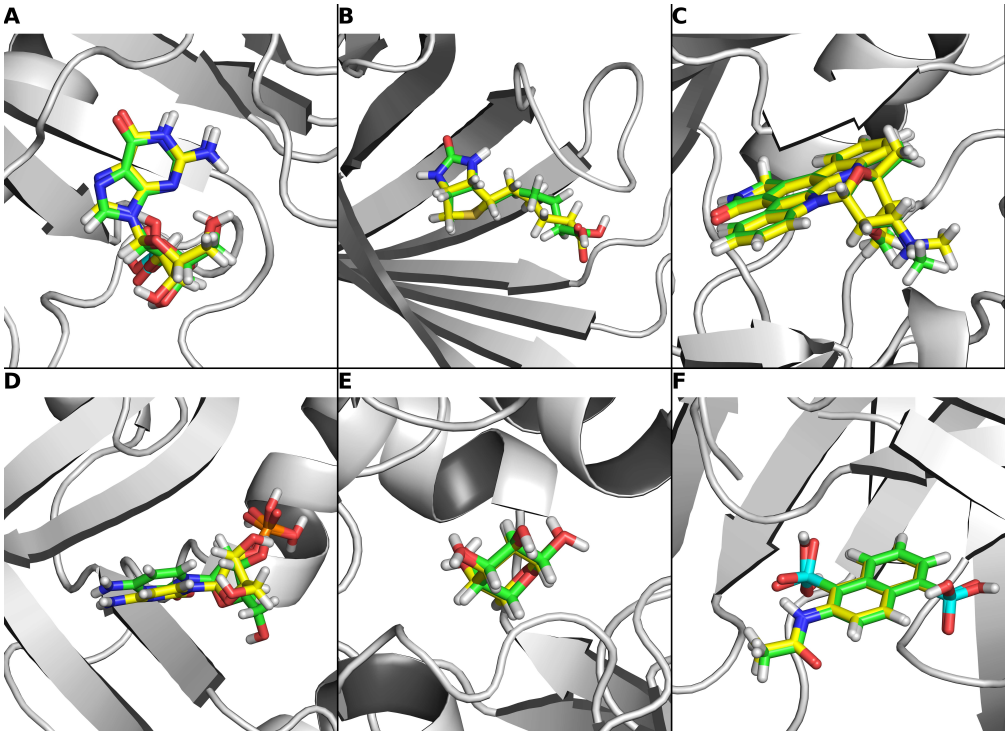
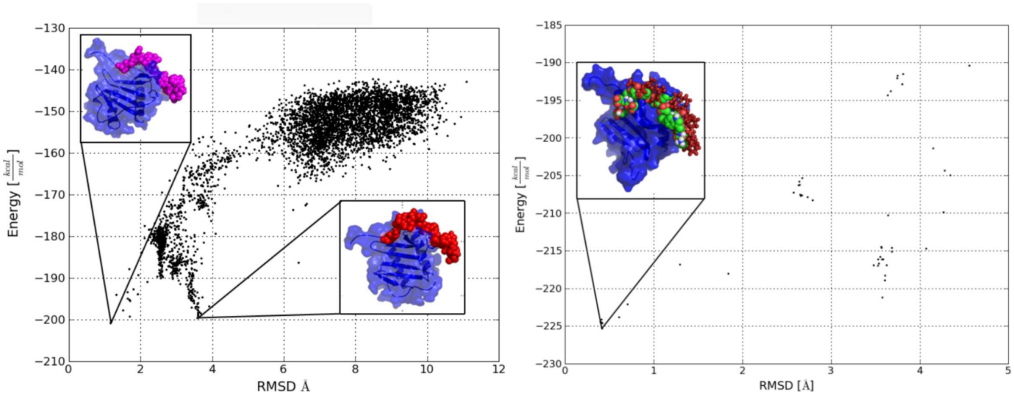


Fig. 6.6.: *Illustrative results of the cascaded docking strategy used for the efficient sampling of six different drug targets. The images show the experimental reference structure in green and the prediction in yellow. Following are the respective PDB IDs of the benchmark complexes and the RMSDs toward the respective experimental structure: A 1RNT,* $1.45\,\text{Å}$ *B: 1STP,* $0.95\,\text{Å}$ *C: 1XBC,* $0.38\,\text{Å}$ *D: 1ROB,* $1.45\,\text{Å}$ *E: 1ABE,* $0.31\,\text{Å}$ *F: 1C5C,* $1.34\,\text{Å}$.

the stage. These poses are passed to the next stage and subjected to longer simulations until only a few conformations are left in the last stage. As in the standard FlexScreen protocol, individual simulations use the stochastic tunneling method. Although single simulations of this approach still simulate for a maximum of $111,000$ Monte-Carlo steps, this is only done for 8 of 96 replica. Fig. 6.6 shows six benchmark simulations conducted with a fraction of the CPU resources needed in the benchmark simulations presented here. The configuration of the simulations is shown in Tab. 6.4. All six predicted protein-ligand complexes could be predicted to experimental resolution.

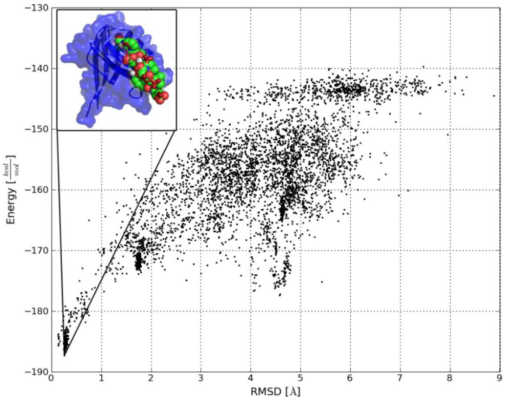
6.2.4. Results of the Protein-Protein docking benchmark

1ILP and 1MFG The protein-protein docking simulations for Interleukin-8 with its native receptor peptide CXCR1 (1ILP) and the complex of the ERBIN-PDZ domain bound to the C-Terminal peptide derived of its receptor ERBB2 (1MFG) were carried out using the PFF02 forcefield. For each complex we generated $5,000$ random structures by a center-of-mass translation of $15\,\text{Å}$ away from the docking site and a random rigid-body rotation. All simulations started from conformations, where ligand and receptor were not in contact.
This initial conformation was then relaxed in 5000 center-of-mass moves comprising random rigid-body displacements of $0.5\,\text{Å}$ and random rigid-body rotations of $3°$ of the molecule. Additionally sidechain dihedral angles were perturbed by a maximal angle of $5°$ drawn equidistributed. Side-chain and rigid-body moves were selected with a fraction of 2:1 respectively. In



(a) 1ILP short simulations

(b) 1ILP further relaxation

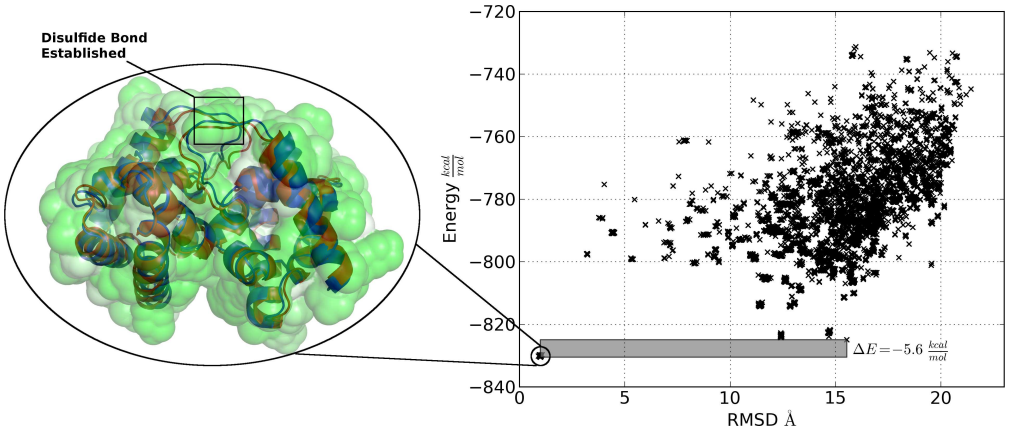


(c) 1MFG

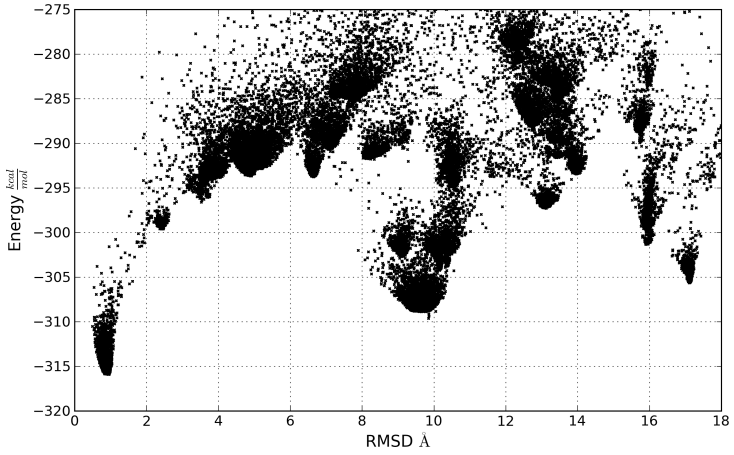
Fig. 6.7.: *Results of docking simulations for 1ILP and 1MFG: a) Plotting the energy vs. RMSD of the relaxed conformations for 1ILP yields two candidate conformations with nearly identical energy. The chemokine is shown in blue, while the competing peptide conformations are shown in red and magenta. b) Further relaxation of the energetically lowest* 60 *conformations uniquely identified the native model to within* $0.5\,\AA$*. The chemokine is shown in blue, the native model in red and the docked model as spheres (colors: N - blue, C - green, O -red, H-white). c) For 1MFG the native model was identified in the first set of relaxation simulations. (Colors as in b).*

the case of 1MFP the 60 lowest energy structures were relaxed for another $100,000$ steps to achieve convergence. The simulations were conducted using simulated annealing with geometrical temperature scaling from 700 to $50\,K$.

A near-native conformation was selected in the $5,000$ step simulations for 1MFG (Fig. 6.7c). The lowest energy conformation had a RMSD of below 0.5 Å to the native conformation. Two competing low energy conformations were observed in the docking simulations of 1ILP at 1.3 Å and 3.8 Å RMSD distance to the experimental conformation, separated by less than $1\,kcal/mol$ (see Fig. 6.7a). We concluded that the length of the short Monte-Carlo annealing simulation was too short to achieve convergence. The 60 lowest energy structures observed in the first simulation were therefore relaxed another time for $100,000$ steps. The resulting ensemble is displayed in Fig. 6.7 b). Upon convergence, the lowest energy conformation had a RMSD of 0.5 Å in comparison to the experimental structure. Structures similar to the incorrect prediction observed in the short relaxations remained at a RMSD of 3.6 Å to the experimental conformation.



(a) Docking results, annealed conformations



(b) Docking results, snapshots

Fig. 6.8.: *Energy/RMSD of all results for the docking of protein 2YGU. a) The native conformation was discovered to experimental resolution with a RMSD of* $1.0\,\text{Å}$*. It is offset to the next low energy structure of different topology by an energy difference* $5.6\ kcal/mol$*. Due to the strong energetical funnel towards the correctly docked structure, no structures near the correct docked pose were discovered. b) Energy/RMSD snapshots taken during the simulations. The funnel is clearly visible for structures below* $4.0\,\text{Å}$*. To speed up the sampling in the snapshot simulations, internal energies were not evaluated leading to an offset in the energy scales, when comparing the two figures.*

2YGU We studied the dimerization of the fire ant venom allergen (PDB: 2YGU)[247]. The protein occurs as a homodimer in its native state and comprises 125 amino acids per chain. The crystal structure was resolved to $2.6\,\text{Å}$ and the complex is stabilized via a disulfide bond that links the CYS21 of both chains. In forming such a complex, the question arises, whether the native conformation of the complex is a unique free energy minimum in the absence of the disulfide bridge, or selected via conformational selection from a multitude of competing structures[248, 249].

The simulations replicated the setup of the protein-protein docking benchmark of 1ILP and 1MFG with modifications by generating $200,000$ starting structures with an offset of 15 Å in a random direction and rotating them about an arbitrary axis by a random angle. The simulations comprised $50,000$ steps to achieve convergence. Rigid-Body translations were drawn from a uniform distribution with a maximum displacement of $1.4\,\text{Å}$. In contrast to the previous simulation a biased rigid-body translation was carried out, whenever the two proteins were further away than $15\,\text{Å}$. In the event the two structures moved too far away from each other, the ligand was pulled towards the receptor structure by a maximum displacement of $2.0\,\text{Å}$.

The lowest energy conformation of 2YGU lies within experimental resolution at a RMSD of $1.0\,\text{Å}$ to the experimental structure and is offset by an energy gap of $5.6\,kcal/mol$ from the next lowest energy conformation, which has a RMSD of $15.5\,\text{Å}$ to the experimental conformation. The intermolecular disulfide bridge (not included in the forcefield) can only form correctly in the cluster of lowest energy structures shown in the left panel of Fig. 6.8. The forcefield PFF02 favored complex conformations which envelop the docking interface near the actual docking site of 2YGU. The lowest energy structure selects the native conformation in the absence of a potential modeling the disulfide bridge, because most of the surface of the docking interface is covered. It is interesting to note that there are few competing low-energy structures featuring a low RMSD around the lowest energy structure. This can be attributed to a sharp funnel of the energy towards the correctly docked pose. Due to the long relaxation simulations, intermediate structures inside the funnel are not observed. To further test this hypothesis RMSD/energy snapshots during the simulations are shown in Fig. 6.8b. A sharp funnel can be observed for conformations near the correct docking pose (RMSD < $4\,\text{Å}$).

6.2.5. Discussion

In this section, we presented a protein-ligand docking approach and benchmarked it, by predicting the complex of streptavidin-biotin to experimental resolution. We further presented a method suitable for the high-throughput screening of protein-ligand complexes required to assist experimental drug design investigations.

The presented protocol was further used without modifications for protein-protein docking to predict the docking pose of two protein-protein complexes with different binding partners and one homodimer. While the simulation of 1ILP required further manual intervention to achieve convergence and used information about the binding interface, the later simulations of 1MFG and 2YGU were simulated without knowledge about the binding interface and converged immediately due to exhaustive sampling. As SIMONA includes both, efficient protocols to study protein conformational change and protein-ligand docking, later studies will include

protein-ligand docking studies with a higher degree of receptor flexibility to also enable the study of induced-fit docking events.

7. Simulations of nano-scale systems

The Monte-Carlo simulation techniques applied to biological systems in the previous chapter can be applied to many different systems on the nano-scale currently studied in the material sciences. Morphology simulations using the techniques introduced in chapter 3 are possible for systems at energies, where no covalent bond breaking occurs. If it is known prior to the simulation that no covalent bond breakage will occur, many nano-scale systems can be parametrized and studied with semi-classical Monte-Carlo approaches. Above these energies simulations of changes in covalent bonding require hybrid QM/MM schemes, such as the Car-Parinello method[250, 251], Density Functional Theory[252] or Hartree-Fock based quantum chemistry methods[253]. These techniques are very resource intensive and only allow the simulation of either short timescales or small conformational change. Previous Monte-Carlo investigations report polymer crystallization simulations[254], the morphology of multi-component metallic glasses[255] or self-organization of magnetic nanoparticles and its phase transitions[256], to name just a few. We have implemented a generic methodology to parametrize these systems and simulate them in SIMONA. We first applied it to simulate morphologies of organic materials.
The emergence of printable (organic) electronics sparked an increased interest in the research of organic materials. In comparison to traditional semiconductors, like silicon, organic electronics do not require clean-room handling and multi-step lithography. Often the required materials are flexible and biodegradable opening new application areas, such as the usage as flexible e-paper. Many promising applications of organic electronic components like OLED, OFET and organic CMOS were developed[257–260]. Presently organic materials have deficiencies, such as lower mobilities and a higher resistance than their inorganic counterparts. To optimize the electronic structure of organic materials, morphologies of the material have to be modeled. This chapter therefore focuses on two applications, simulation of morphologies of amorphous pentacene and carbon nanotube sorting.
In section 7.1, we simulate morphologies of amorphous pentacene, a material, which rivals silicon in its electronic properties due to its high mobility and low resistance. The generation of morphologies of amorphous pentacene is the first step to enable electronic calculations of this material. In section 7.2, we focus on the specific sorting of carbon nanotubes. The electronic properties of carbon nanotubes are very well known and very attractive for usage in both, mechanical and electronical, applications[261]. The usage of carbon nanotubes however requires a homogeneous carbon nanotube bulk. We therefore introduce a polymer, which selectively binds nanotubes with specific chiral angles to allow their sorting.

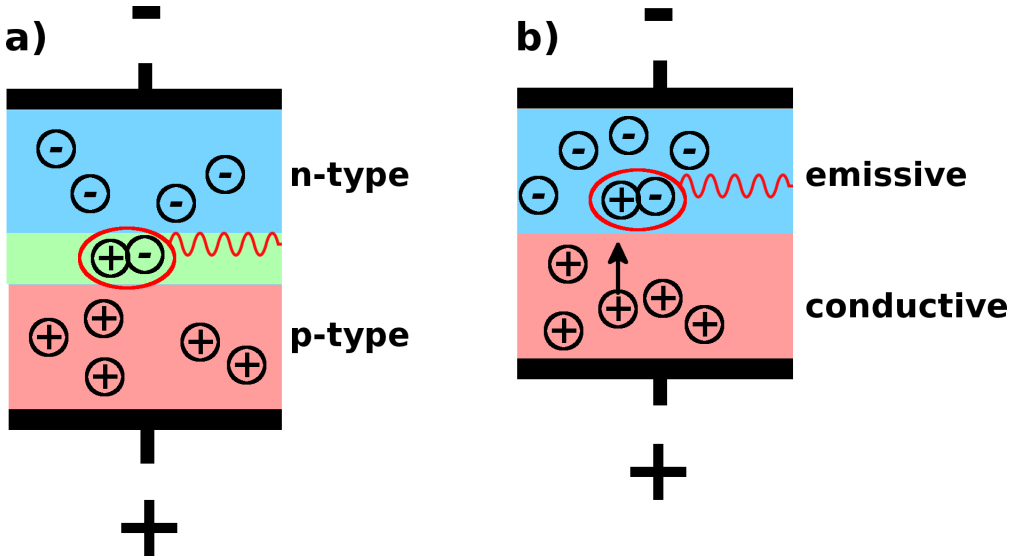
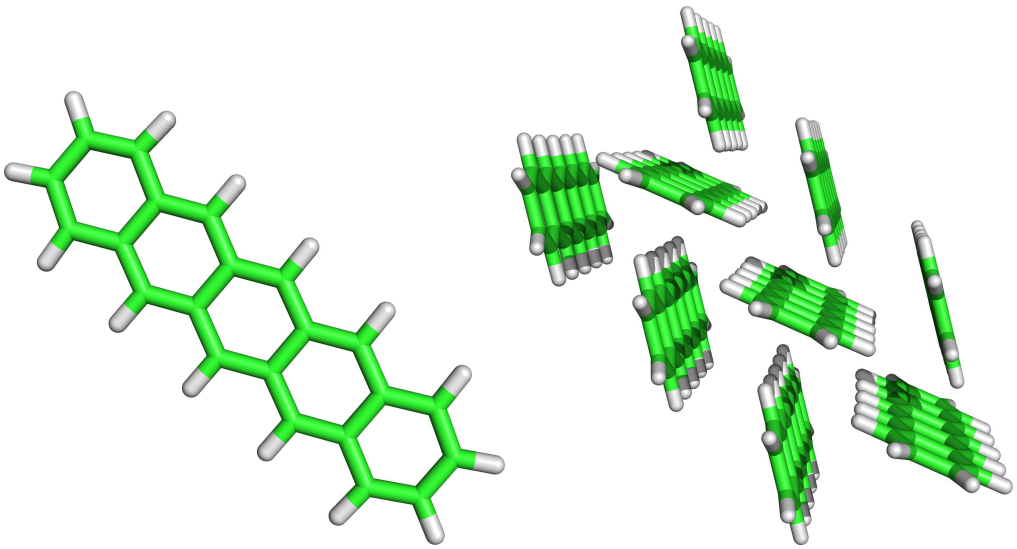


Fig. 7.1.: *Differences between semiconductor LED and organic LED. a) A pn-junction of a semiconductor LED. When applying a voltage, holes and electrons travel to the junction between the p- and n-doped semiconductors. They recombine emitting a photon. b) Holes inside the conductive layer diffuse into the emissive layer. Electrons and holes recombine in the emissive layer and emit a photon.*

7.1. Morphology simulations of amorphous pentacene

7.1.1. Motivation

Organic components were developed for many devices, which are traditionally constructed from inorganic semiconductors like OLED, OFET and organic CMOS[257–260]. In spite of their attractive properties like low cost, application using printing techniques and very often biodegradability, they cannot replace their inorganic counterparts completely, due to their often lower mobility, higher resistance and shorter life period.
The light-emitting layer of traditional light emitting diodes (LEDs) consists of a np-junction between an anode and a cathode (see Fig. 7.1a). If a forward voltage is applied between anode and cathode, electrons are relocating to the p-doped side and holes are moving towards the n-doped side. Recombination in the np-junction then allows photon emission for direct-gap semiconductors with a frequency according to the band-gap of the semiconductor[262].
OLEDs replace the n and p-type semiconductors by a conductive and an emissive layer (see Fig. 7.1b). When applying a voltage, holes are created in the conductive layer, which can diffuse into the emissive layer due to their higher mobility (compared to electrons in the emissive layer). Recombination of electrons from the cathode with holes from the conductive layer inside the emissive layer then generates photons. The gap between the HOMO and LUMO orbitals of the organic material, forming the emissive layer, define the wavelength of the emitted light. The emissive layer of an OLED is formed of amorphous arrangements of small organic molecules or polymers. For the design of OLEDs a p-type semiconductor with a high mobility is required in the emissive layer. Pentacene ($C_{22}H_{14}$) is an interesting candidate for the electroluminescent layer of OLED[263]. Pentacene is a planar molecule consisting of five benzene rings as displayed in Fig. 7.2a. Composites of pentacene rival mobilities of amorphous silicone, with mobilities of approximately $5\,cm^2V^{-1}s^{-1}$ for polycrystalline pentacene films. Interest in pen-



(a) Single pentacene Molecule (b) Herringbone-stacked layer

Fig. 7.2.: *Isolated and crystallized pentacene. a) Isolated pentacene. Pentacene consists of five benzene rings. Green atoms represent carbon, white atoms represent hydrogen. b) Layer of crystallized pentacene. Pentacene crystallizes in a face-on-edge herringbone conformation.*

tacene grew especially after the discovery that bulk and thin-film pentacene is a p-type organic semiconductor and therefore compatible with the usage in the emissive layer of OLED[264]. To model electronic transport in such materials, the morphology of the pentacene layer needs to be known.

Upon crystallization bulk pentacene stacks in herringbone conformation with a layer spacing of 14.1 Å. Apart from that three thin-film conformations could be identified with spacings slightly bigger than the bulk[265–267]. The native crystal conformation can be observed in Fig. 7.2b. In this section, we investigate initial pentacene cluster nucleation. The simulations are preparatory for a project including compound Monte-Carlo moves, moving complete clusters of pentacene as a whole to allow for crystal formation.[1] We present the simulation parameters and the analysis method of the nucleated clusters in section 7.1.2 and analyze the obtained cluster geometry in section 7.1.4. In conclusion we show the shortcomings of this simulation strategy in the scope of crystal formation, which will be published as part of Schönauer[269].

7.1.2. Methods

The cluster growth simulations started from a random assembly of molecules. As input, we used the pentacene morphology from the Cambridge Crystallographic Data Centre (CIF ID 2012157)[270] and assigned partial charges of pentacene using DFT[271]. Lennard-Jones parameters were included from the standard FlexScreen parametrization[10]. The partial charges were used for a vacuum electrostatics potential evaluated on the GPU (see section 3.2.1 for

[1] Results in this project were published as part of Strunk et al.[38]. Pentacene was parametrized initially in the scope of the bachelor thesis of Paul-Jakob Kleine[268]. The mentioned cluster move scheme is still being researched as part of a bachelor thesis of Benedikt Schönauer. I helped supervise both of the theses and developed the cluster move code together with Benedikt Schönauer[269]. I also wrote the analysis code explained in the methods section.

the GPU implementation). Starting from this setup, we ran $400,000$ simulations with $500,000$ steps each with a constant temperature of $300.0\,K$ on the distributed POEM@HOME architecture[71].
The resulting cluster topologies were analyzed for their size and order using the following algorithm:

- We generate the power diagram (described in Klenin et al.[63]) and calculate its dual Delaunay tesselation.
- We iterate over all vertex neighbor pairs (corresponding to atoms in the pentacene). If two vertices are close to each other (two atoms closer than $4\,\text{Å}$) they are assigned to the same cluster.
- For each pentacene molecule inside a cluster the tensor of inertia is calculated and transformed into its principal axis system. The axis of smallest inertia then corresponds to the long axis of the pentacene.
- Starting from a random pentacene molecule, we iterate over all members in the cluster and calculate the relative angle between two pentacene molecules. If the relative orientations of two pentacene molecules are similar (angle $\theta < 0.8$) they were counted as in-plane and therefore ordered. Otherwise they were counted as facing in different directions.
- The ordered pentacene molecules inside the current stack are counted and accumulated.

7.1.3. Results

In the following text, we will separate between *total* and *ordered* cluster sizes. *Total* cluster sizes count all pentacene members within a cluster, even unordered ones. *Ordered* cluster sizes count only pentacene molecules within a cluster, which point into the same long direction (estimated using the algorithm in section 7.1.2). This differentiation is made to elucidate, whether or not clusters attained local order. Pentacene crystallizes in a herringbone conformation (see Fig. 7.2b); all pentacene molecules are aligned along their long axis and would be identified inside an ordered layer by our algorithm. The histogram of all cluster sizes encountered during the simulation is shown in Fig. 7.3. The *total* cluster sizes show a near-exponential decay over five orders of magnitude. Only a small amount of clusters with *total* sizes above 50 are formed. The distribution of *ordered* cluster-sizes trails the distribution of *total* cluster sizes. The probability to find an *ordered* cluster size is negligible for ordered clusters of a size of 20 and above. A significant amount of all *ordered* clusters below size 20 has attained local order. Characteristic examples of clusters observed during the simulation are shown in Fig. 7.4. Simulations comprised clusters with pentacene molecules ordered in a herringbone formation, as seen in Fig. 7.4a. Furthermore single layers of π-stacked clusters were also observed (Fig. 7.4b). Many clusters featured amorphous pentacene arrangements unable to attain higher local order (Fig. 7.4 c). Bigger cluster sizes then consisted mostly of smaller clusters as seen in 7.4 a-c, merged into higher superclusters (Fig. 7.4 d and e).

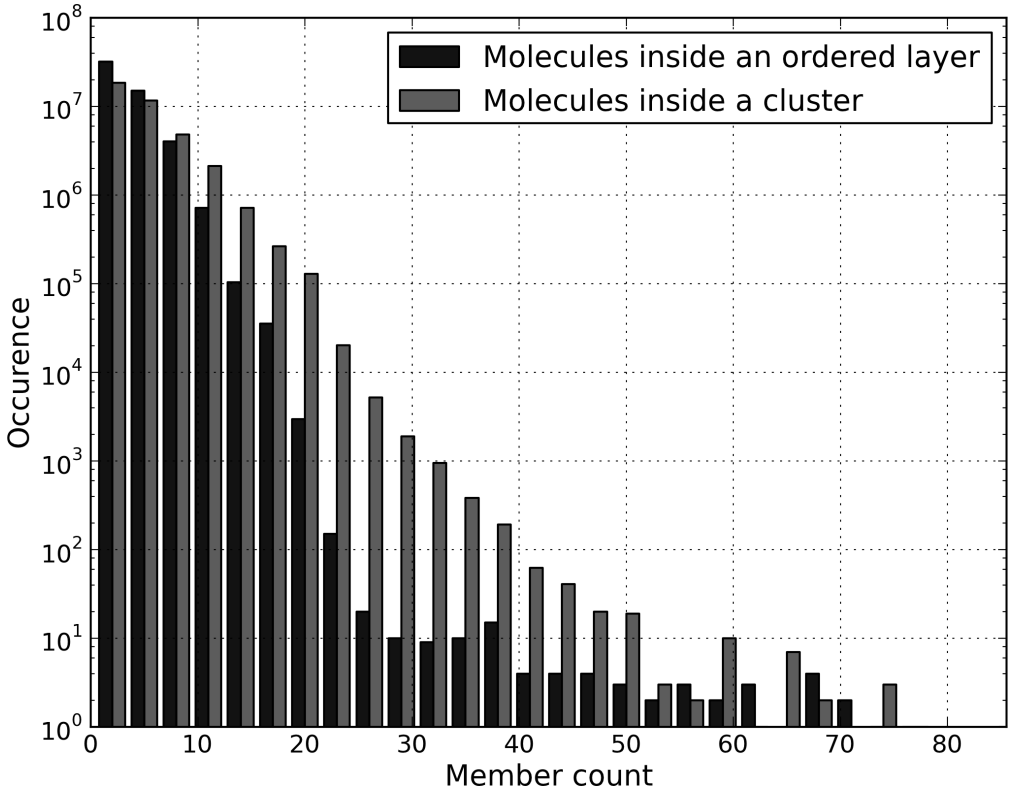


Fig. 7.3.: *Histogram of cluster sizes and number of pentacene molecules in ordered layers. For small counts, both distributions decrease exponentially. Although pentacene forms a crystal, the Monte-Carlo simulation freezes and crystals stop growing. As only rigid body moves of single pentacene molecules are allowed, it gets exponentially harder to move a complete cluster.*

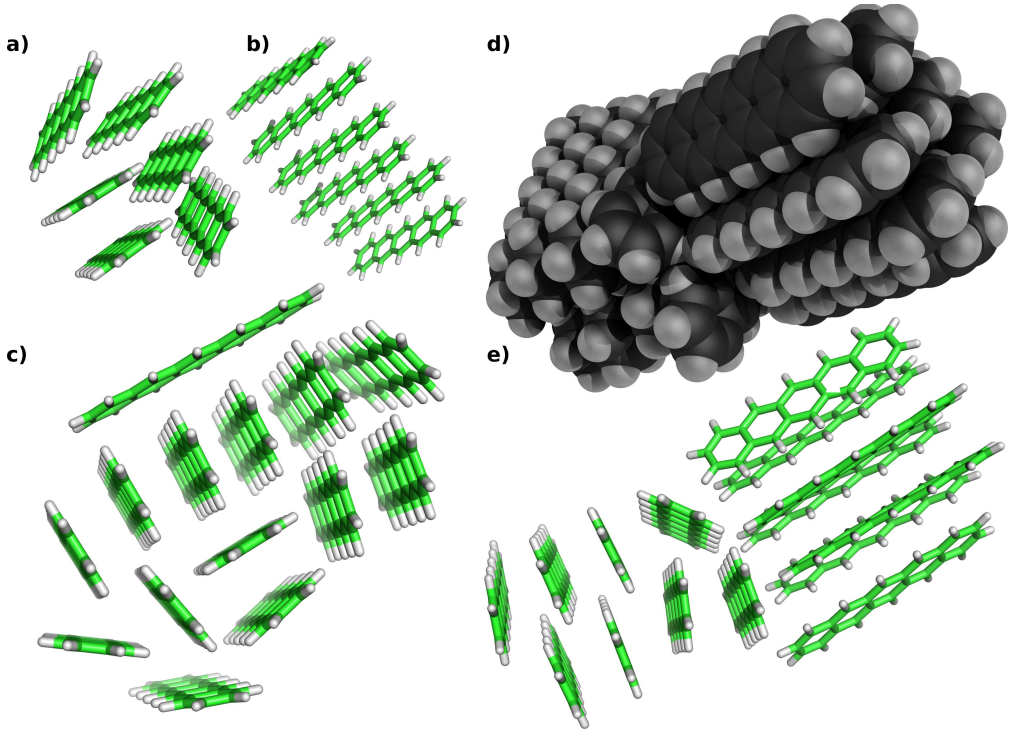


Fig. 7.4.: *Various examples characteristic for the cluster topologies observed in the simulations of pentacene clustering. a) Initial stage of building a herringbone conformation. b) Direct Face-on-Face π stacking conformation. c) Unordered pentacene bundle with pentacene ordered in z-direction. d) and e) Two pentacene cluster directions merged into a single cluster, space-fill, and licorice visualizations.*

7.1.4. Discussion

We observed pentacene clusters exhibiting pentacene binding in a herringbone formation akin to the native crystal structure, which demonstrates the correct parametrization of pentacene in SIMONA. The further growth of these clusters will allow future electronic structure calculations based on the morphology and thereby allow the optimization of the electronic properties of OLED.

Significant numbers of ordered clusters with cluster sizes beyond 20 could not be produced. This is due to a freezing effect in our Monte-Carlo simulations: As we only perturb single molecules inside a cluster, acceptance rates for diffusion processes of a complete cluster are very low. Molecules inside a cluster rarely break off making cluster extension by single pentacene molecules unlikely once most molecules have migrated to a cluster. Although better results might be obtained by using more advanced simulation techniques like parallel tempering[42–44] or multiple try Monte-Carlo[69], the main simulation flaw impeding further cluster growth is the usage of single subunit moves.

Simply choosing to move one cluster at a time would break detailed balance and therefore skew evaluation of observables. Multiple cluster move schemes have therefore been developed[272–274]. We implemented the simulation scheme of Whitelam et al.[275], previously tested on single atoms and applied it to pentacene clusters: Every Monte-Carlo step a random particle of the simulation is selected, the seed particle i (in our case one of the pentacene molecules).

In this method, a pseudocluster C, which is a subcluster of a cluster observed in the simulation, is selected and an acceptance probability for moving this pseudocluster in the ensemble is calculated. For this single molecule, a transformation is suggested from state μ to state ν. For every neighbor of the seed particle a link probability $p_{ij}(\mu \to \nu)$ is calculated to include the particle j also in the subcluster C. The transformation $\mu \to \nu$ can be interpreted as a translation of particles i and j. The link-fail probability is then $q_{ij} = 1 - p_{ij}$. Both are recorded and the particle is accepted to the pseudocluster with the link-probability p_{ij}. For a specific i this is done for all other particles. Whenever a particle j is accepted to the pseudocluster all link probabilities p_{jk}, of pseudoparticles not yet in the cluster are also tested.

Using the still unspecified link and link-fail probabilities p_{ij} and q_{ij}, the probability to generate a specific configuration C is:

$$W_{\text{gen}}(\mu \to \nu) = P_{\text{Seed}}(\mu) \sum_{R}^{C} \prod_{[i,j] \notin R} q_{i,j}(\mu \to \nu) \prod_{[i,j] \in R} p_{ij}(\mu \to \nu) \quad . \tag{7.1}$$

Here R denotes a link-representation of C. For a specific pseudocluster C, multiple link-orders called representations R exist. The first product in the equation comprises of all pairs i, j, which were tested, but rejected. The second product comprises all accepted i, j. For big pseudoclusters, the evaluation of the whole sum is not feasible as the enumeration of all representations grows approximately with the factorial of cluster members.

If W_{gen} is the probability of generating a move and W_{acc} is the acceptance criterion, the transition probability is $W = W_{\text{gen}} \cdot W_{\text{acc}}$. To derive an acceptance probability, which maintains the detailed balance criterion (Eq. 7.2) the exact probability of move generation W_{gen} has to be

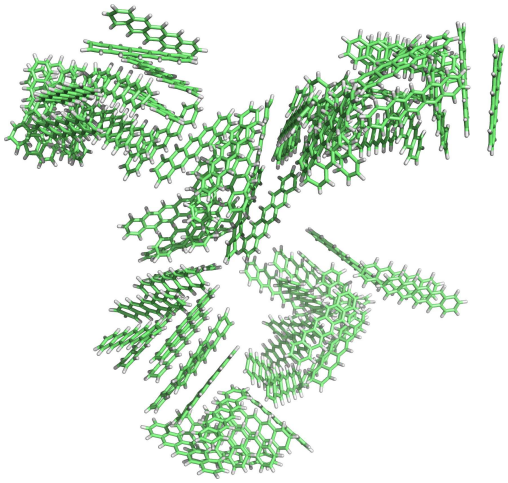


Fig. 7.5.: *Final configuration of a* $400,000$ *step simulation of pentacene using cluster moves started from* 100 *isolated pentacene molecules; one of two clusters observed in the simulation is shown. Due to the mobility of the clusters during the simulation, clusters can merge more frequently and form bigger assemblies.*

known.

$$\rho(\mu)\, W(\mu \to \nu) = \rho(\nu)\, W(\nu \to \mu) \qquad . \tag{7.2}$$

Another possibility to converge to an equilibrium state is to fulfill super-detailed balance:

$$\rho(\mu)\, W(\mu \to \nu | R) = \rho(\nu)\, W(\nu \to \mu | R) \qquad . \tag{7.3}$$

Eq. 7.3 fulfills the weaker detailed balance criterion even for a specific representation R of the pseudocluster C. In this case $W_{\text{gen}}(\mu \to \nu)$ is obtained as the product of accepted and rejected probabilities of links in the cluster (Eq. 7.4):

$$W_{\text{gen}}(\mu \to \nu | R) = P_{\text{Seed}}(\mu) \prod_{[i,j] \notin R} q_{i,j}(\mu \to \nu) \prod_{[i,j] \in R} p_{ij}(\mu \to \nu) \qquad . \tag{7.4}$$

With this definition a possible choice for W_{acc} is given in Eq. 7.5:

$$W_{\text{acc}} = \min \left[1, \frac{P_{\text{seed}}(\nu)}{P_{\text{seed}}(\mu)}\, e^{-\beta\,(E_\nu - E_\mu)} \times \frac{\prod\limits_{ij \notin R} q_{ij}(\nu \to \mu) \prod\limits_{ij \in R} p_{ij}(\nu \to \mu)}{\prod\limits_{ij \notin R} q_{ij}(\mu \to \nu) \prod\limits_{ij \in R} p_{ij}(\mu \to \nu)} \right] \qquad . \tag{7.5}$$

A specific link probability p_{ij} based on the binding energies between two subunits can now be proposed. This approach uses the interaction energy between the two particles i and j to propose a link (Eq. 7.6):

$$p_{ij}(\mu \to \nu) = \max(0, 1 - e^{\beta\,(E_c(i,j) - E_I(i,j))}) \qquad . \tag{7.6}$$

The algorithm starts by choosing a seed particle with even probability $P_{\text{Seed}}(\mu) = P_{\text{Seed}}(\mu) = 1/N$. Then from this seed particle, other particles are selected, for which $p_{ij}(\mu \to \nu)$ is significantly bigger than 0; these are particles, which are in close vicinity in the limit of strongly interacting particles. Once the algorithm has tried all particles within a cluster, the move is carried out and back probabilities $\nu \to \mu$ are calculated. All p_{ij} values and q_{ij} values are recording

during this process; in the end W_{acc} is evaluated and the new state is accepted or rejected. Various further optimizations to this algorithm can be found in Whitelam and Geissler[275].
Initial results of simulations carried out using this algorithm are promising. Fig. 7.5 shows a typical simulation starting from 100 isolated pentacene molecules for $400,000$ steps of Cluster-moves interleaved with single pentacene rigid-body translations and rotations. Only two clusters of amorphous pentacene exist in this simulation, which still show remarkable mobility. Quantitative results analyzing the effect of cluster moves on cluster formation and dissociation will be reported in the bachelor thesis of Benedikt Schönauer[269].

7.2. Dispersion of Single-Walled Carbon Nanotubes by chiral index

7.2.1. Motivation

Due to their interesting properties and emerging applications, single-walled carbon nanotubes (SWNT) have been the focus of research attention. While nanotube bulk production methods have been established[276, 277], the attained bulk of nanotubes is often heterogeneous with nanotubes of many different chiral angles and diameters. As physical properties of nanotubes depend on diameter, chiral index and the length of an individual nanotube, it is important to separate the bulk selectively into samples with specific properties[276–282]. This was an experimental collaboration between the group Mayor of the INT. Experimental details and parts of this text can be found in Lemasson et al.[283].
In section 7.2.2, we will characterize the main properties of nanotubes and the polymers used for dispersion in the simulations. The results of the dispersion simulations are found in section 7.2.3. Section 7.2.4 discusses the dispersive properties of the polymers observed in the simulations in the context of the experimental results.

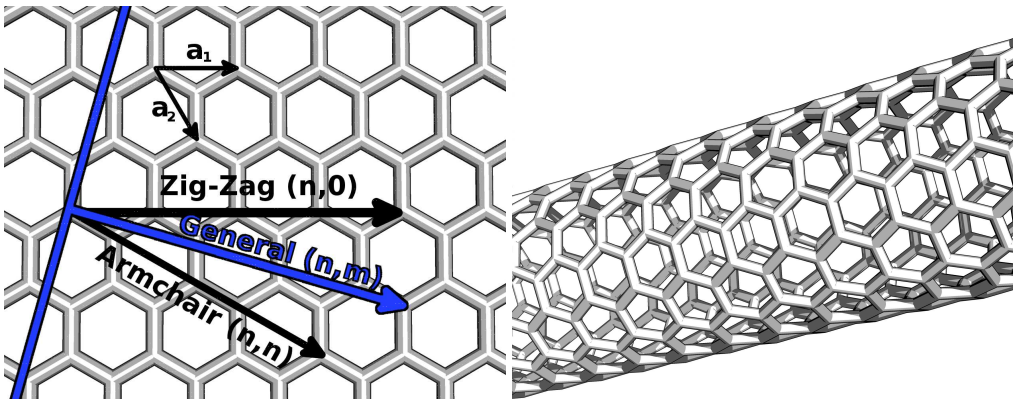
7.2.2. Introduction

Carbon nanotubes are rolled up sheets of graphene. Graphene is a monoatomic hexagonal lattice of carbon molecules as shown in Fig. 7.6a. Carbon atoms inside nanotubes are bound in sp^2 configurations leaving one electron per carbon in the π-orbital. As seen in Fig. 7.6b a nanotube can be characterized solely by its two chiral indices (n, m). Depending on the chiral index, a nanotube can be either metallic or semiconducting. As a rule of thumb, nanotubes are semiconducting with a vanishing band-gap, if $n - m$ is a multiple of 3. They are metallic, if $n = m$ and moderately semiconducting otherwise. There are exceptions to this rule especially for the case of large chiral indices. Apart from their extraordinary strength and stiffness, many of those configurations also possess unique electronic characteristic like their high electric current density[261].
Due to their attractive properties they have been the focus of research attention for some time. Nanotubes interact by interactions of the π-electrons. For morphology simulations, which keep the overall bond structure of the nanotube intact, it is therefore sufficient to parametrize the nanotube in a Van-der-Waals forcefield[51]. Using the definitions in Fig. 7.6b, one can easily obtain the chiral angle Θ and the diameter d of the nanotube from the chiral indices as in Eq. 7.7 and Eq. 7.8:

$$\Theta = \arctan\left(\sqrt{3}\,\frac{m}{m+2\,n}\right) \quad , \tag{7.7}$$

$$d = \frac{\sqrt{3}\,a}{\pi}\,\sqrt{n^2 + n\,m + m^2} \quad . \tag{7.8}$$

Before sorting of the nanotubes can be accomplished they need to be debundled, for example by sonication-assisted dispersion in a suitable solvent. The attained dispersion can then be stabilized by adding water soluble polymers[284] or surfactants in water or by polymers in organic



(a) Graphene (b) (7,7) nanotube

Fig. 7.6.: *Graphene sheet and carbon nanotube. a) A flat graphene sheet. a_1 and a_2 are the lattice vectors of the unit cell. A nanotube is defined solely by its chiral indices (n, m). The tube is rolled along the blue axis orthogonal to the lattice vector. Two specific nanotube configurations are drawn: an armchair tube (n,n) and a zig-zag tube (n,0) b) A rolled up (7,7) armchair nanotube for comparison.*

solvents[278–282, 285].

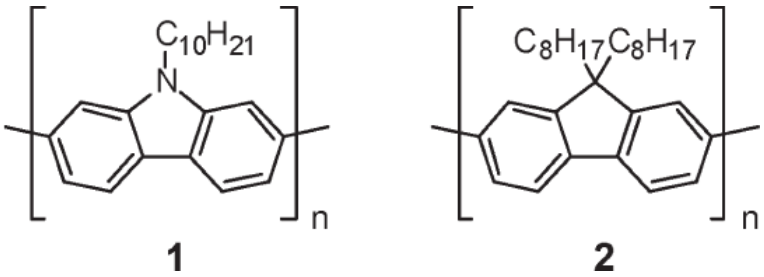
Polymers containing 9,9 dialkyl-2,7-fluorene subunits were found to be unexpectedly selective for semiconducting tubes with differences in chiral indices $(n - m) = 1$ or 2. Further research focused on the fine tuning of the aryl subunits to increase specificity towards the binding of single nanotube configurations[278–282], but the mechanisms remain unclear. These polymers were shown to only slightly alter the nanotube's electronic properties upon binding[280, 281, 286].

Here we present Monte-Carlo simulations using polymers of a fluorene decamer and additionally a carbazole decamer and show their selectivity towards different chiral indices exemplary for (10,2) and (7,6) nanotubes. Using models of these polymers and two nanotubes known to be dispersed by only one polymer, we could show that differences in the $\pi - \pi$ stacking interaction are the reason for the unexpected selectivity. We observed energetically favored complexes between the (10,2) nanotube and the carbazole decamer and between the (7,6) nanotube and the fluorene decamer, respectively. These results agree with experimental findings[283], where multiple other polymer - nanotube combinations were analysed with respect to their selectivity.

The simulated polymers are shown in Fig. 7.7. The carbazole decamer (polymer 1) has a planar N-bridging atom, whereas the fluorene decamer (polymer 2) has a tetragonal C-bridging atom (Fig. 7.7) leading to different steric Van-der-Waals configurations of the polymers. Furthermore polymer 1 has only a single alkyl sidechain, while the fluorene polymer possesses 2 (also Fig. 7.7).

For our binding studies, we simulated single-walled nanotubes with the chiral indices (10,2) and (7,6), as they both have comparable diameters (Eq. 7.9), but were each preferentially dispersed by only one of the two polymers. The forcefield used in these simulations was comprised only of a standard 6-12 Lennard-Jones potential as the system was uncharged. By applying Eq. 7.7 and 7.8 to the n, m pairs we obtain the nanotube diameters $d_{n,m}$ and their chiral angles $\theta_{n,m}$:

$$d_{7,6} = 8.95\,\text{Å} \quad \Theta_{7,6} = 27.457^\circ \quad d_{10,2} = 8.84\,\text{Å} \quad \Theta_{10,2} = 8.95^\circ \qquad . \tag{7.9}$$



1 2

Fig. 7.7.: *Single subunits of the two simulated polymers. 1. Structure of the carbazole polymer. In comparison to the fluorene polymer, the carbazole one incorporates only a single alkyl sidechain per subunit. The nitrogen atom leads to a planar bond of the sidechain, while the fluorene polymer has a tetragonal binding carbon atom. 2. Structure of the fluorene polymer.*

7.2.3. Results

Starting from conformations with well-separated structures of both polymer decamers and nanotubes, we conducted atomistic basin-hopping simulations.
All four decamer - nanotube combinations were initially optimized by four independent basin hopping simulations. The lowest energy conformation was then subjected to long simulated annealing simulations to obtain an estimate for the energy minimum[164, 287–289].
Each single hopping step comprised a Monte-Carlo annealing running initially for $N_{\text{Steps}} = 10,000$ steps as shown in Eq. 7.10:

$$N_{\text{Steps}} = \sqrt{100 \cdot N_{\text{iter}}} \cdot 1000 \qquad . \tag{7.10}$$

The annealing process employed a geometrical cooling schedule starting from a random high temperature (between $1000\,K$ and $3000\,K$) to an end temperature of $1000\,K$ to allow for the hopping over barriers. As we did not gauge the forcefield the temperature unit K is only an arbitrary unit in this case and represents the forcefield gauge. We immediately observed that the fluorene decamer, representative of polymer 2, showed no tendency to wrap around the (10,2) nanotube, lying essentially flat on the tube. This qualitative result was obtained in all four independent simulations. In contrast, some wrapping of the fluorene compound was observed for the (7,6) tube, as illustrated in the bottom part of Fig. 7.8. The energies of this complex were significantly lower than for the (10,2) tube (red symbols in Fig. 7.9). Close alignment of the rings of the polymer and the tube was observed for the carbazole decamer. Here we observed the complexes formed with the (10,2) tube (Fig. 7.8) to be energetically favored compared to the ones formed with the (7,6) tubes (Fig. 7.9). Because the energy model incorporates a short-range term for the $\pi - \pi$ interactions between the aromatic carbons and the polymer models and the nanotubes, the binding energy is dominated by the number of aromatic C-C contacts in the complex.
The number of such contacts is constrained by the geometry of the molecule, which has only one free backbone dihedral per unit and steric repulsion between the sidechain atoms and the nanotube. The results of the fluorene-(10,2) simulations indicate that only the *trivial* solution (all

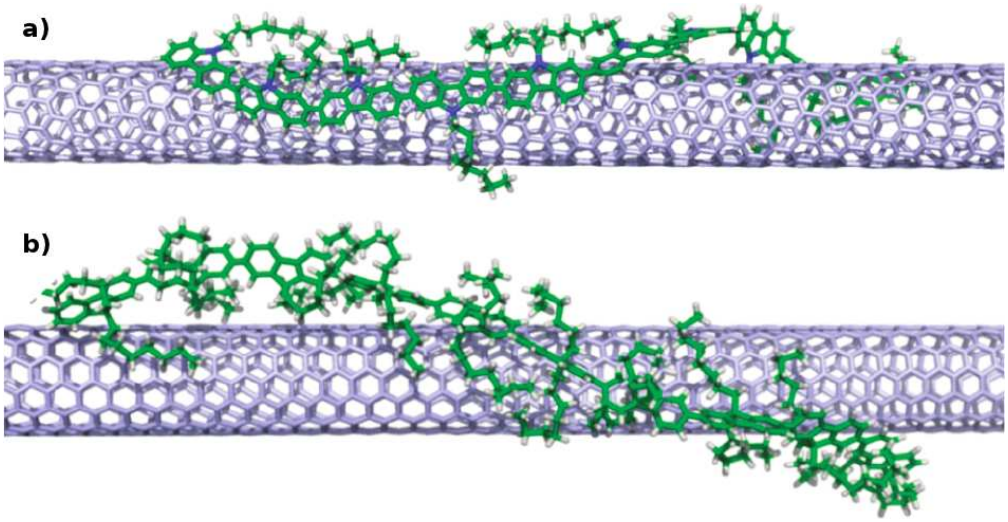


Fig. 7.8.: *Energy-minimized structures of the complex of the two nanotubes with the simulated decamers. a) The carbazole decamer attached to a* (10,2) *nanotube. The flat alignment of the polymer to the tube is stabilized by* $\pi - \pi$ *stacking. b) Wrapping of the fluorene decamer to the* (7,6) *tube. Compared to the carbazole decamer this stacking is not as uniform. Only small segments are* π*-stacked. The sidechains seem to be more involved in the binding of the polymer.*

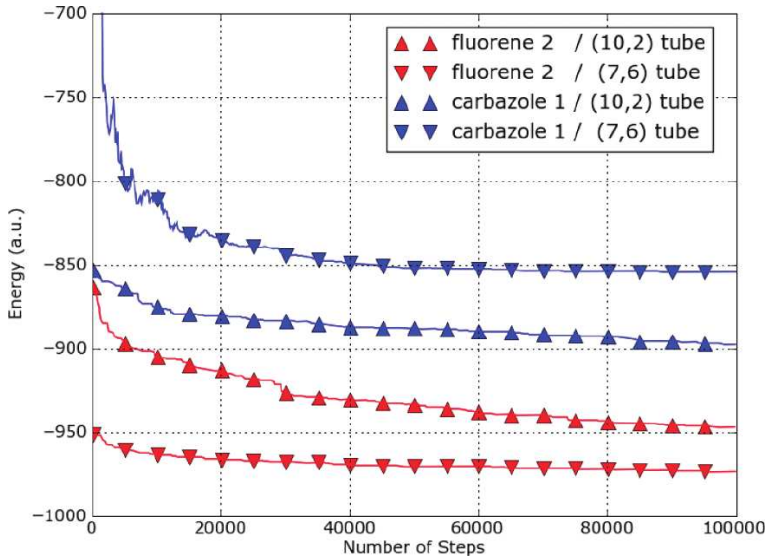


Fig. 7.9.: *Energy evolution of the final simulated annealing basin-hopping simulations. Carbazole (blue) favors binding to the* (10,2) *tube (made apparent by the lower energy in the wrapping simulations). The simulations of the fluorene polymer (red) show an inverted binding propensity: the* (7,6) *tube features a lower energy bound to fluorene compared to the* (10,2) *system. Shown binding energy units are arbitrary as solvation effects were not taken into account.*

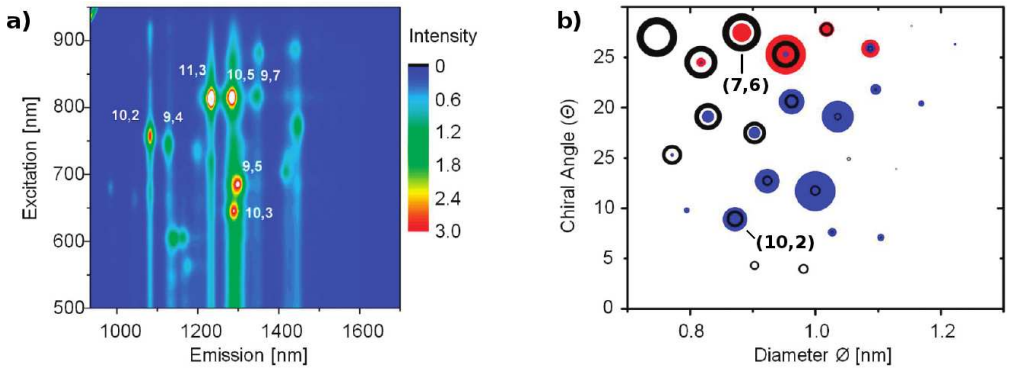


Fig. 7.10.: *Fractions of nanotube content observed using Photoluminescence Excitation. a) PLE image of nanotubes dispersed with polymer 1 in toluene. b) Merged intensities transformed into the diameter/chiral angle plane. Blue: PLE signal observed in solution dispersed by polymer 1, Red: PLE signal observed in solution dispersed by polymer 2. Black: Control dispersed by Na-cholate in D_2O. Images courtesy of Lemasson et. al[283].*

backbone dihedrals straight) is possible for this system. When the radius of the tube changes or the sidechain constraints simplify by removing one sidechain, other solutions for the backbone dihedrals become feasible. The chiral index of the tube will then determine how many favorable aromatic contacts become possible, which obviously depends on many nontrivial geometrical constraints. The increased tendency of the carbazole decamer to wrap around both types of tubes thus results from an increased $\pi - \pi$ interaction made possible by the reduced steric requirements of a single alkyl chain at the bridging atom.

7.2.4. Discussion

We could show selective binding of two polymers containing either fluorene or carbazole subunits towards nanotubes of different chiral angles. These results represent the first step towards further sorting of nanotubes into bulks homogeneous in chiral angle and diameter. We were able to pinpoint this selective interaction to be based on the aromatic stacking of the polymer mainchain with the nanotube. Geometric constraints allow the polymers to only bind a specific nanotube selectively by optimizing the stacking interactions.
Our findings agree with experimental results. Photoluminescence excitation spectra were used to determine the fractions of nanotubes in solutions treated with Polymer 1 or 2 after sonication and density gradient centrifugation[283]. Due to the characteristic excitation and emission spectra of nanotubes, it is possible to convert the PLE image into an image normalized in chiral angle Θ and diameter d (Fig. 7.10). The excitations observed in solutions of Toluene with polymers 1 or 2 clearly disperse the nanotube also energetically favored in the simulations (Fig. 7.10b)[283].

8. Summary

Proteins govern many biological processes of the human body. Among countless other functions these nanomachines catalyze complex biochemical reactions, regulate ion concentrations or mediate the immune response. Mutation and misfolding of proteins are implied in many diseases, such as Alzheimer's disease and cancer. Most proteins fold spontaneously into a unique tertiary structure, encoded by their amino acid sequence. Although complete genomes of protein-coding sequences were compiled, only a small fraction of experimentally resolved protein structures are currently available. Of the $80,000$ resolved proteins only $12,500$ are structurally dissimilar. The huge gap between millions of known protein-coding sequences and the available protein structures shows the difficulty involved in resolving a protein structure experimentally.

Methods for theoretical prediction of protein structures produce mixed results. While models with reasonable accuracy can be expected for proteins with at least 40% sequence similarity to a protein with a known structure, no reliable method exists to predict structures with no or very distant homologs. This state of the art led to mistrust in the biological and medical community towards theoretical protein structure predictions and resulted in the pruning of the PDB database of all theoretical predictions. If a reliable method to assess the quality of these structures would exist, theoretical predictions could be introduced as a new source of information for biological and medical research and provide insight into many biological processes and diseases, where we lack understanding.

It is now well established that beyond single structures, the dynamics of proteins and their association is very important to understand their function. Some of these properties are accessible via Molecular Dynamics simulations, but severe limitations remain. Due to the timescales involved, Molecular Dynamics simulations are frequently not able to simulate large conformational change or de-novo folding in spite of recent technological advances.

In this thesis, I have therefore explored Monte-Carlo simulations as a tool to investigate structure and thermodynamic properties of biomolecular and nanoscale systems. Monte-Carlo simulations do not suffer from the timescale problem, as they characterize the thermodynamic ensemble of the system and avoid sampling processes on very short timescales that are often irrelevant for the large-scale properties of the underlying thermodynamic ensemble. We have implemented the novel generic Monte-Carlo simulation program, SIMONA, because previously no efficient general-purpose Monte-Carlo code has been available.
We have optimized this program package on multiple architectures and parallelized it using a variety of optimization algorithms. The resulting program, SIMONA, is available for free for

academic use under http://www.int.kit.edu/nanosim/simona.php .

I have then applied the simulation methodology to various problems, including absolute quality assessment of protein models, several applications requiring protein structure prediction, protein-protein and protein-ligand docking and the simulation of morphologies in organic nanostructures. Most of the results obtained during the course of these simulations have been experimentally verified showing the validity and generality of our approach. In the following, I provide an overview of the insights gained in the various projects.

Absolute Quality Assessment of Protein Structures

Protein structures are frequently used to explain biological processes on the nano-scale. Although only a limited set of experimental protein structures of medical and biological relevance is available, the acceptance of theoretical predictions of protein models in the life-sciences is low. Using methods like homology modeling, very often a prediction within experimental resolution can be made for highly homologous sequences. In the absence of homology, there are isolated cases where protein models close to the native conformation were constructed. However in the grey area of intermediate degrees of homology, little is known about the quality of protein models, in particular those generated from fully automated servers. Development of methods for absolute quality assessment of protein structures would therefore go a long way to increase the acceptance of theoretical models in life-science research.
We therefore developed a protocol for absolute quality assessment, based on the concept of marginal stability of proteins. We hypothesized that every amino acid must contribute an optimal energy contribution towards the global protein structure in its biologically active state. We collected statistics for these energy contributions for a set of high-resolution protein structures and derived a N-dimensional statistical test, which assesses the quality of a protein model by comparing against these statistics. We found that the energy statistics of amino acids in their folded state differ from those of low quality protein models. By introducing energy statistics of triplets of amino acids, we could increase the specificity of our methods and reject 93% of the low quality protein models for 160 proteins tested. The remaining 7% of the protein models were found to be either oligomeric, not globular, or bound to cofactors, all classes of proteins, where the initial hypothesis was bound to fail. Given the present state of the art it is important to develop methods that reject false positives with high certainty, even if these work only for a specific subclass of proteins. There are bioinformatics-based approaches that can be used to predict with high reliability, whether, for a given protein sequence, a model belongs to one of the classes that our present algorithm cannot discriminate well. In combination of these techniques we hope that our approach will serve as a prototype for further development of quality assessment protocols and help the further acceptance for those models which are evaluated positively.

Design of genetically engineered variants of hydrophobin DewA

Adhesion and proliferation of biofilms on implants frequently requires the replacement of the implant. Attributable to the increasing life-expectancies, many current patients will face the

problem of implant replacement. Especially for patients of old age, implant surgeries can be life-threatening. It is therefore important to design implants, which can remain in the body for more than two decades. One method, which would allow longer usage times of implants in the human body is the coating of the implant with a material suppressing the formation of biofilms, while allowing for the development of new tissue cells. Hydrophobin proteins have been considered as one promising candidate for implant coatings, because hydrophobin aggregates form hydrophobic-hydrophilic surfaces at air-water interfaces and can turn a hydrophilic (implant-) surface hydrophobic and less susceptible to the formation of biofilms. One drawback discovered in previous studies was the low degree of cell-adhesion on hydrophobin coatings.
We therefore functionalized the class-I hydrophobin DewA to allow cell adhesion on DewA covered surfaces. As no experimentally resolved structure of hydrophobin DewA was available, we predicted a model using homology modeling on the basis of a distant homolog and identified a solvent exposed area, which could be modified without impairing the overall hydrophobin protein structure. We inserted two previously identified cell-binding peptide motifs into the sequence and modeled the fusion protein. Protein expression was then carried out by the group of Prof. Fischer (KIT) and the resulting surface was characterized by the group of Prof. Schimmel (KIT). Cell binding assays performed by the group of Prof. Richter (UIC Heidelberg) could show significantly increased binding propensities of mesenchymal stem cells, required for bone growth compared to the unmodified hydrophobin. Remarkably, at the same time, we did not observe any increase in the formation of biofilms. Apart from applications for implant coating, hydrophobins allow many other uses for industrial application areas in the field of medical engineering. One area of application, which we are currently investigating with an industrial partner is their application in dialysis machines to allow the filtering of blood cells, while preventing the accumulation of life-threatening biofilms. Future development will focus not only on the adhesion, but especially the specificity of the cells bound to the hydrophobin.

Rodlet development of Class-I hydrophobins

Even better surfaces could be engineered if it were possible to exploit the subclass of hydrophobin that form stable monolayer rodlets on the implant surface. The mechanism and requirements for rodlet formation are presently not well understood. We therefore investigated possible reasons, why class-I hydrophobins form stable amyloid structures upon aggregation on surfaces. It was hypothesized that the flexible loops, prevalent throughout the family of class-I hydrophobins, take a considerable part in rodlet formation. However, subsequent publications could show indications that a rodlet might even develop upon truncation of the flexible loops.
Using a truncated hydrophobin model, obtained from the experimental structure of the class-I hydrophobin EAS by means of homology modeling, we docked subunits of the hydrophobin together and discovered a periodically extensible dimer with defined hydrophobicity. This structure is compatible with the previous experimental results, suggesting that even a truncated mutant can form a rodlet, and may serve as the basis for further structure characterization of the rodlets and the mechanism of their formation.

Structural model of the gas-vesicles in aqueous bacteria

In collaboration with the groups of Prof. Pfeifer and Prof. Hamacher (Darmstadt University), we could develop the first nano-scale structural model explaining the development of gas-vesicles in aqueous bacteria. Protein-protein aggregation has been a thoroughly studied topic, since the discovery of protein-aggregation related diseases, such as Alzheimer's disease. The experimental study of protein aggregation is difficult with traditional structure determination methods, such as X-ray crystallography or NMR methods, as these proteins are frequently insoluble or aggregate into amorphous structures, when subjected to the requirements of the structure determination experiments. In this project, we studied gas-vesicles, macroscopic protein aggregates, which allow bacteria to swim afloat natural water bodies.

Prior to our investigation, no structural model for the monomer of protein GvpA, the main constituent of the gas-vesicle wall, was available. Using fragment-based modeling techniques we were able to not only generate a model compatible with prior solid-state NMR and ATR-FTIR measurements, but also to provide insight into the quaternary structure of the gas vesicle and its rib-like superstructure. We explain the gas retention by identifying the inner hydrophobic gas-vesicle wall and predict various contact sites important for the formation of the vesicle, which were validated in mutagenesis experiments. When introducing protease proteins only sites shown to be accessible in our model were cleaved. Remarkably, cleavage of the evolutionary non-conserved C-terminal did not impair gas-vesicle formation at all in accordance with our model.

High-throughput prediction of peptide structures

In the last four decades only four new classes of antibiotics achieved FDA approval. Antimicrobial, antifungal and antibiotical peptides are vigorously investigated as possible alternatives to complement current antibiotic drugs and therefore alleviate the increasing problems stemming from resistance of bacteria against existing antibiotics. In contrast to small molecule drugs in-silico design of peptides is complicated by the lack of structural information for novel peptide sequences. Although the conformational space of peptides is far smaller than the one of large proteins, knowledge-based methods frequently fail to produce correct structure predictions, as even single point mutations often affect peptide structure.

We therefore developed a high-throughput peptide structure prediction method and benchmarked it on the distributed volunteer computing architecture POEM@HOME by de-novo prediction of four peptides of different secondary structure. The results of the four peptide structure models are within experimental resolution. We further investigated the low energy conformations and elucidated possible shearing conformations, one of the peptides may switch between, near its native conformation. Only a small fraction of the resources of POEM@HOME is required to screen these peptides. This methodology will be further pursued in a new project of the BMBF biotechnology 2020+ initiative aimed at the rational design of peptides for protein immobilization on magnetic nanoparticles and other technologically relevant surfaces.

Computational Alanine Screening

A multitude of biological processes are mediated by protein-protein interaction and their mal-

function is implied in many diseases. Understanding protein-protein interaction gives insight into the underlying biological processes and might even offer avenues for their manipulation. One problem impeding the progress in the exploitation of protein-protein interfaces as drug targets is their large area in comparison to the compact docking pockets targeted by many small molecule inhibitors. Although the inhibition of protein-protein binding was possible with antibodies, the bioavailability of these costly molecules is low. Development of smaller inhibitors is therefore desirable due to better bioavailability, lower price and ease of handling. The targeted design of these small-molecule ligands is aided by the identification of interaction hotspots, which are most important for the stabilization of the protein-protein interface. One of the most widely used experimental methods for the identification of these hotspots is alanine screening, where each individual amino acid in the interface must be mutated.
To reduce the cost and effort involved in these experiments, we implemented an in-silico alanine screening protocol and benchmarked it using two important chemokine systems. Chemokines are small proteins, which guide cells of the immune system towards sites of infection. As many cells contain binding receptors for chemokine proteins, chemokines are implied in many serious conditions, such as chronic inflammations or allergies, but also fatal diseases such as autoimmune disorders or cancer. Using the alanine screening protocol, we predict the hotspots of the ERBIN/ERBB2 complex and Interleukin-8 complexed with its native receptor CXCR1. Our method was the only of the tested methods to predict all seven hot spots of the Interleukin-8 receptor peptide. We could also correctly identify the interaction hotspot in the ERBIN/ERBB2 complex. We did not observe a single false-negative hotspot, as validated by subsequent experimental alanine screens that could focus on the reduced set of amino acids identified by the computational screen. Our method therefore reduces the work involved for a complete mutagenesis assay of a protein interface significantly.

De-novo protein-protein and protein-ligand docking

The study of protein-protein interactions presented in the previous section is only possible, if an experimental structure of the bound mode is available. When only the monomer structures are known, it is not immediately evident, which amino acids constitute the binding interface. We therefore predicted the binding pose of three protein-protein complexes. We extended the protocol to small-molecule protein docking, integrating the functionality of the FlexScreen in-silico screening approach into SIMONA. In addition to brute-force docking simulations, we developed a cascaded docking strategy for protein-ligand docking and verified it by docking six protein-ligand complexes to within experimental resolution. This project extended the application area of SIMONA to also efficiently tackle problems of protein-ligand docking.

Morphology simulations of pentacene clusters

The electronic properties of amorphous organic materials are widely investigated in the context of the development of organic light emitting diodes and materials for organic photovoltaics. Prior to a characterization of the electronic properties realistic models for the physical morphology of the amorphous material have to be generated. We have implemented forcefields to describe such materials and specifically parametrized pentacene as one widely studied mate-

rial for the emissive layer in OLED and simulated self organization in clusters. The simulated clusters feature herringbone sub-structures as observed in the pentacene crystals, but extended amorphous superclusters of those topologies are also observed. To speed these simulations we implemented a cluster move scheme, which maintains superdetailed-balance. Using the novel approach the clusters showed faster relaxation rates, which helps us to better describe extended clusters.

Dispersion of Single-Walled Carbon Nanotubes by chiral index

Carbon nanotubes are widely investigated as an attractive nanomaterial for many applications in electronics and mechanics. As their properties depend on the specific type of nanotubes as defined by their chiral indices, nanotubes need to be separated and dispersed, before they can be used in specific industrial applications. Presently few low-cost separation techniques exist, but recent investigations could show that some polymers selectively bind towards specific nanotube subpopulations. We were able to explain this selectivity by analyzing the preferential binding mode for two different polymers towards two exemplary nanotubes of different chiral indices. This finding is remarkable as both nanotubes exhibit similar diameters. The proposed binding preferences are compatible with photoluminescence maps. The dominant binding mode between nanotubes and polymers can be attributed to aromatic stacking, which, due to geometrical constraints of the polymers, can only be optimal for either one of the polymers. The results pose a first step to enable the targeted design of polymers dispersing nanotubes of specific chiral indices.

A. Additional data of the absolute quality assessment methods

A.1. Proof of equation 4.8

We want to show:

$$I_N := \int_0^{R_\kappa} R^N \, e^{-\frac{R^2}{2}} \, dR \tag{A.1}$$

$$= \begin{cases} (N-1)!!\, I_0 - \sum_{i=1}^{\frac{N}{2}} \frac{(N-1)!!}{(2i-1)!!} f_{2i-1}\Big|_0^{R_\kappa} & N \quad \text{even} \\ (N-1)!!\, I_1 - \sum_{i=1}^{\frac{N-1}{2}} \frac{(N-1)!!}{(2i)!!} f_{2i}\Big|_0^{R_\kappa} & N \quad \text{odd} \end{cases} \tag{A.2}$$

$$\text{with: } f_N = R^N \, e^{-\frac{R^2}{2}} \quad . \tag{A.3}$$

Proof. By differentiating f_i, Eq. A.4 is obtained:

$$\frac{d}{d\,R} R^{N-1} \, e^{-\frac{R^2}{2}} = \left((N-1)\, R^{N-2} - R^N \right) e^{-\frac{R^2}{2}} \tag{A.4}$$

$$\text{which gives: } f_N(R) = (N-1)\, f_{N-2}(R) - \frac{d}{d\,R} f_{N-1}(R) \quad . \tag{A.5}$$

This can now be inserted into the definition of I_N to obtain Eq. A.7:

$$I_N = \int_0^{R_\kappa} (N-1)\, f_{N-2}(R)\, dR - \frac{d}{d\,R} f_{N-1}(R)\, dR \tag{A.6}$$

$$= (N-1) \int_0^{R_\kappa} f_{N-2}(R)\, dR - f_{N-1}\Big|_0^{R_\kappa} \tag{A.7}$$

By observing Eq. A.7 an ansatz for the case of even N can be guessed (Eq. A.8):

$$\int_0^{R_\kappa} f_N(R)\, dR = (N-1)!! \int_0^{R_\kappa} f_0(R)\, dR - \sum_{i=1}^{\frac{N}{2}} \frac{(N-1)!!}{(2i-1)!!} f_{2i-1}\Big|_0^{R_\kappa} \quad . \tag{A.8}$$

By inserting the definition of Eq. A.1, one obtains:

$$I_N = (N-1)!!\, I_0 - \sum_{i=1}^{\frac{N}{2}} \frac{(N-1)!!}{(2i-1)!!} f_{(2i-1)}\Big|_0^{R_\kappa} \quad . \tag{A.9}$$

This can be proven for even N using complete induction starting from $N = 2$. The case of

$N = 2$ is trivial, when comparing with Eq. A.7. The inductive step ($(N-2) \rightarrow N$) will now be shown starting from Eq. A.7:

$$I_N = (N-1) I_{N-2} - f_{N-1}\Big|_0^{R_\kappa} \quad . \tag{A.10}$$

Eq. A.2 can now be inserted for I_{N-2}:

$$I_N = (N-1)\left[(N-3)!!\, I_0 - \sum_{i=1}^{\frac{N-2}{2}} \frac{(N-3)!!}{(2i-1)!!} f_{(2i-1)}\Big|_0^{R_\kappa}\right] - f_{N-1}\Big|_0^{R_\kappa} \quad , \tag{A.11}$$

$$= (N-1)!!\, I_0 - \sum_{i=1}^{\frac{N-2}{2}} \frac{(N-1)!!}{(2i-1)!!} f_{(2i-1)}\Big|_0^{R_\kappa} - f_{N-1}\Big|_0^{R_\kappa} \quad , \tag{A.12}$$

$$= (N-1)!!\, I_0 - \sum_{i=1}^{\frac{N}{2}} \frac{(N-1)!!}{(2i-1)!!} f_{(2i-1)}\Big|_0^{R_\kappa} \quad . \tag{A.13}$$

This concludes the proof for even N. For odd N Eq. A.8 has to be modified as in Eq. A.14:

$$I_N = (N-1)!!\, I_1 - \sum_{i=1}^{\frac{N-1}{2}} \frac{(N-1)!!}{(2i)!!} f_{(2i),R_\kappa} \quad . \tag{A.14}$$

The case $N = 3$ is also trivial (compare with Eq. A.7). In this case for the inductive step ($(N-2) \rightarrow N$) one obtains:

$$I_N = (N-1) I_{N-2} - f_{N-1,R_\kappa} \tag{A.15}$$

$$= (N-1)\left[(N-3)!!\, I_1 - \sum_{i=1}^{\frac{N-3}{2}} \frac{(N-3)!!}{(2i)!!} f_{(2i)}\Big|_0^{R_\kappa}\right] - f_{N-1}\Big|_0^{R_\kappa} \quad , \tag{A.16}$$

$$= (N-1)!!\, I_1 - \sum_{i=1}^{\frac{N-3}{2}} \frac{(N-1)!!}{(2i)!!} f_{(2i)}\Big|_0^{R_\kappa} - f_{N-1}\Big|_0^{R_\kappa} \quad , \tag{A.17}$$

$$= (N-1)!!\, I_1 - \sum_{i=1}^{\frac{N-1}{2}} \frac{(N-1)!!}{(2i)!!} f_{(2i)}\Big|_0^{R_\kappa} \quad . \tag{A.18}$$

□

A.2. Set used to train the absolute quality assessment method

The following list contains all PBD codes used in the training set used to compile the distributions of the per-amino-acid energies in the absolute quality assessment protocol:

1JF4, 1BKR, 1JF8, 1JHJ, 1JL1, 1JNI, 1JO0, 1JOV, 1BX7, 1JYK, 1K4N, 1K5C, 1K7C, 1K7J, 1KMT, 1A62, 1KMV, 1KNG, 1KOE, 1KQ6, 1KQR, 1KT6, 1L3K, 1LC0, 1C1K, 1LMI, 1LNI, 1LS1, 1LTZ, 1LWB, 1LYQ, 1M1Q, 1M40, 1M4L, 1M9Z, 1MC2, 1MF7, 1MFW, 1MJ4, 1C75, 1MJ5, 1MK0, 1MQO, 1N08, 1C7K, 1N8V, 1NG6, 1CC8, 1NNF, 1NNX, 1NQJ, 1NU0, 1NWA, 1NWZ, 1NYC,

1NZJ, 1O4Y, 1O7I, 1OI7, 1OK0, 1OOH, 1OQJ, 1P3C, 1PJX, 1CY5, 1PMH, 1PZ4, 1Q0R, 1QDD, 1QG8, 1QGI, 1QTW, 1CZP, 1QWY, 1R5L, 1R9L, 1AH7, 1RG8, 1RJU, 1D4O, 1RKI, 1ROC, 1RTQ, 1RTT, 1RUT, 1RV9, 1RW1, 1RYL, 1RYO, 1RYQ, 1DD9, 1S29, 1S9U, 1SAU, 1SFS, 1SVS, 1SZH, 1T3Y, 1T8K, 1TCA, 1TJX, 1TP6, 1TQG, 1TT8, 1TU9, 1TUA, 1TUK, 1U84, 1UCD, 1UCS, 1UI0, 1UJP, 1UKF, 1DS1, 1UNQ, 1UOY, 1US0, 1UTE, 1UYL, 1DY5, 1V2B, 1VBW, 1VCC, 1VE4, 1E29, 1VKK, 1VLY, 1VYI, 1E5K, 1W0H, 1W0N, 1W4S, 1W66, 1WC2, 1WCW, 1WHZ, 1WNA, 1WPA, 1AHO, 1WVH, 1EAQ, 1X0T, 1X6O, 1X6Z, 1X8Q, 1X91, 1XBI, 1XDN, 1EB6, 1XDZ, 1XGK, 1XMK, 1XMT, 1XOD, 1XQO, 1ECA, 1Y8A, 1Y9L, 1YD0, 1YE8, 1YN3, 1Z2N, 1Z2U, 1Z3X, 1Z67, 1Z6M, 1EJG, 1ZCE, 1ZDY, 1ZGK, 1ZI8, 1ZK5, 1ZMM, 1ZZK, 1ELK, 2A6Z, 2ABS, 2AP3, 2AYD, 2B97, 2BBR, 2BKM, 2BL8, 2BOG, 2BRF, 2BWF, 2C71, 2CB8, 2CBZ, 2CCQ, 2CCW, 2CG7, 2CKK, 2CMP, 2CNQ, 2CS7, 2CVE, 2CWS, 2CXA, 2CXY, 2CYG, 2CYJ, 1EW4, 2D3D, 2DDX, 2DHO, 2DSX, 1EYH, 2DT8, 2DXA, 2E3H, 2E3N, 2END, 2ENG, 2ERF, 2ERL, 2EW0, 2F23, 2F46, 2F60, 1F94, 1F9V, 1FK5, 1ARB, 1FT5, 1FYE, 1G2R, 1G66, 1ATG, 1G8A, 1GBS, 1GCI, 1GK7, 1GMX, 1GP0, 1GPP, 1GU2, 1B3A, 1GVD, 1GWM, 1H4A, 1H97, 1HDO, 1HXI, 1I27, 1I2T, 1I5G, 1I71, 1IE9, 1IFR, 1IN4, 1IO0, 1IQZ, 1J0P, 1J3A, 1J77

A.3. Statistics of single amino acid energies

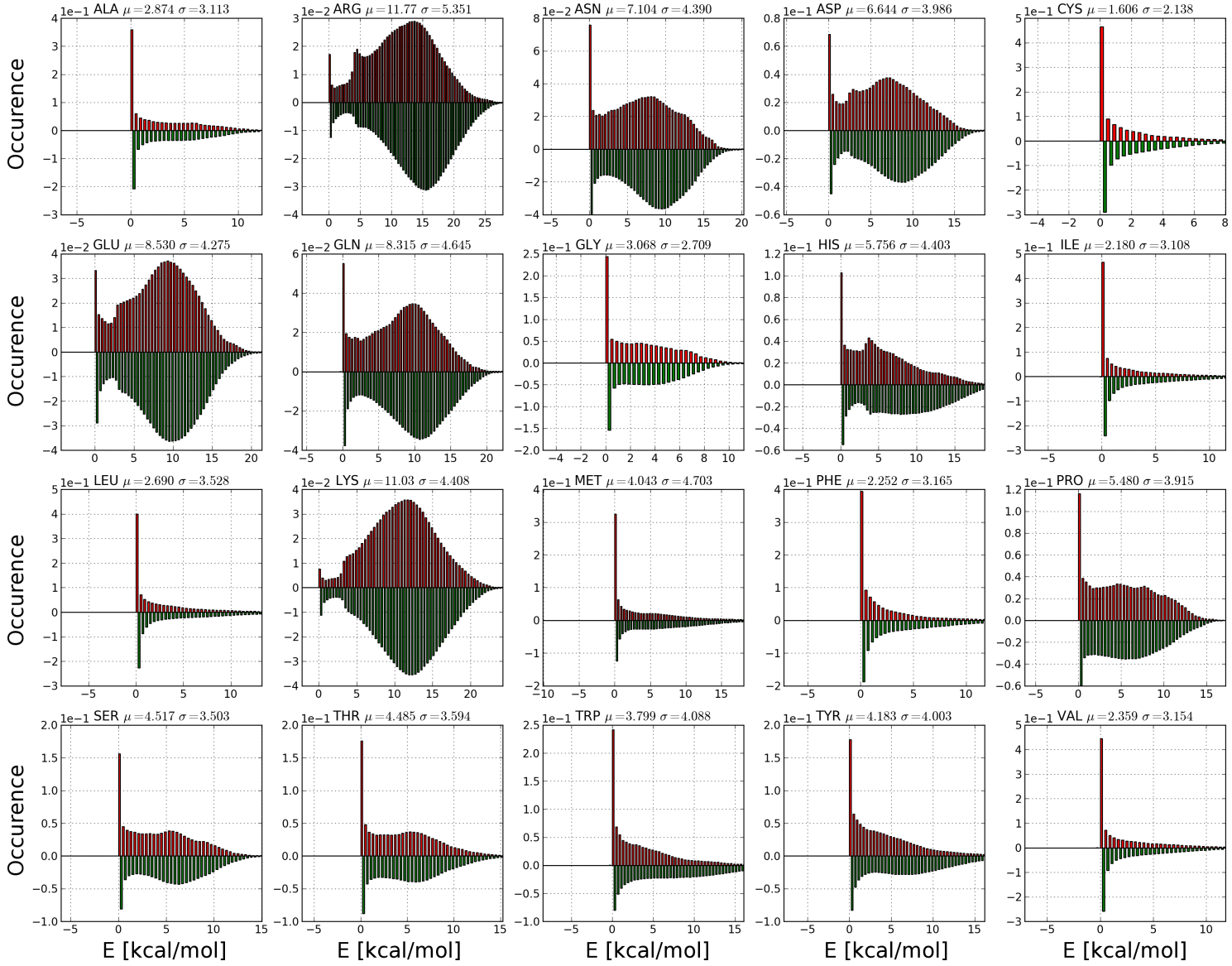


Fig. A.1.: *Statistics for the amino-acid specific PFF02 contributions of all 20 amino acids. Positive occurences relate to energy distributions the set of native protein structures (red); negative occurences relate to the decoy energy distributions. The decoy statistics feature a bigger solvent exposure for hydrophobic amino acids. This is most apparent, when comparing the values for completely occluded amino acids. For PHE, ILE, LEU, VAL, MET and HIS nearly two times more amino acids are completely occluded in the native structures in comparison to the decoy structures.*

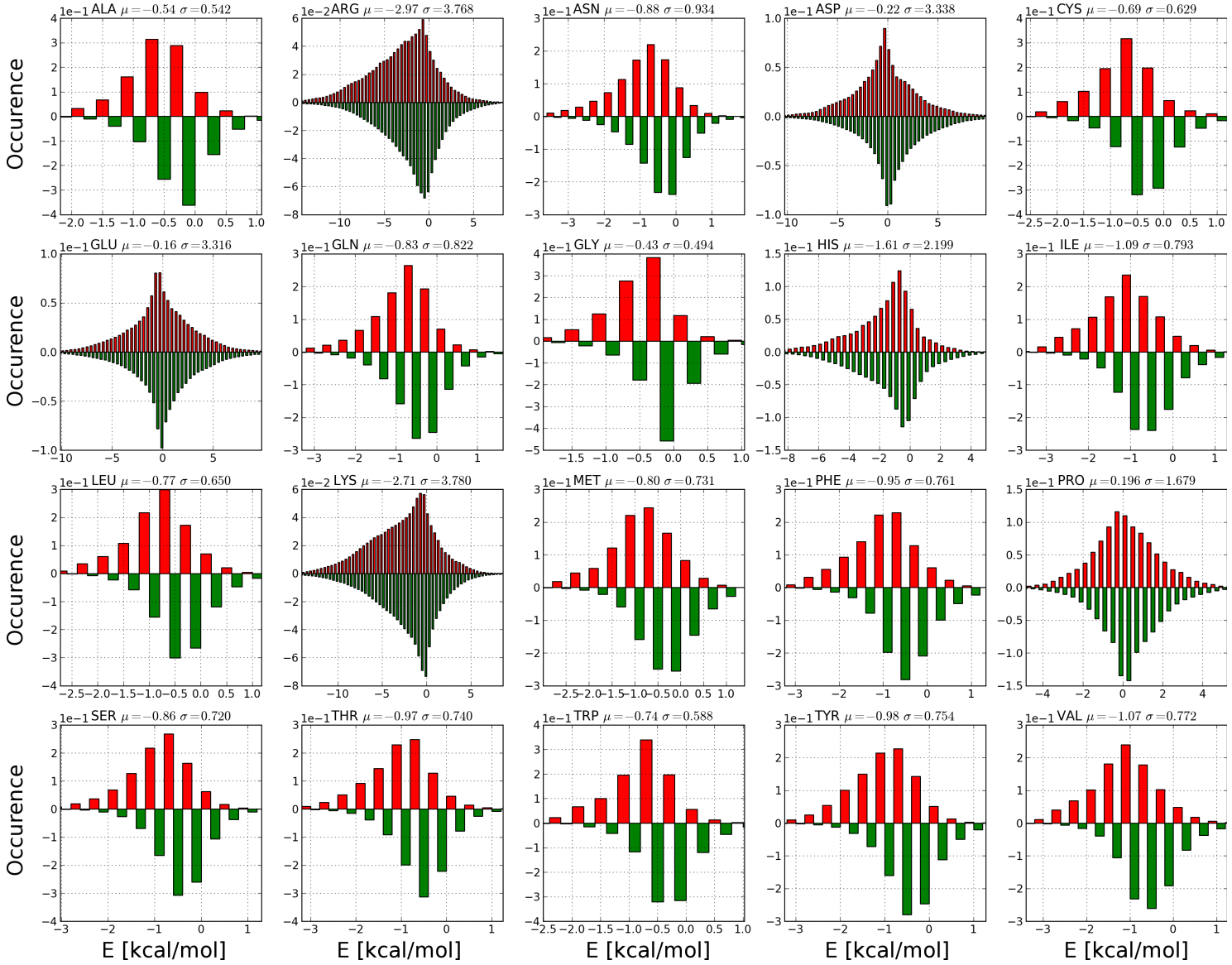


Fig. A.2.: *Statistics for the amino-acid specific PFF02 contributions of all 20 amino acids. Positive occurences relate to energy distributions the set of native protein structures (red); negative occurences relate to the decoy energy distributions. Charged amino acids (ARG, ASP, GLU, HIS, LYS) cover a large energy range due to their electrostatic contribution. Only slight deviations between native structures (red) and decoy structures (green) can be observed.*

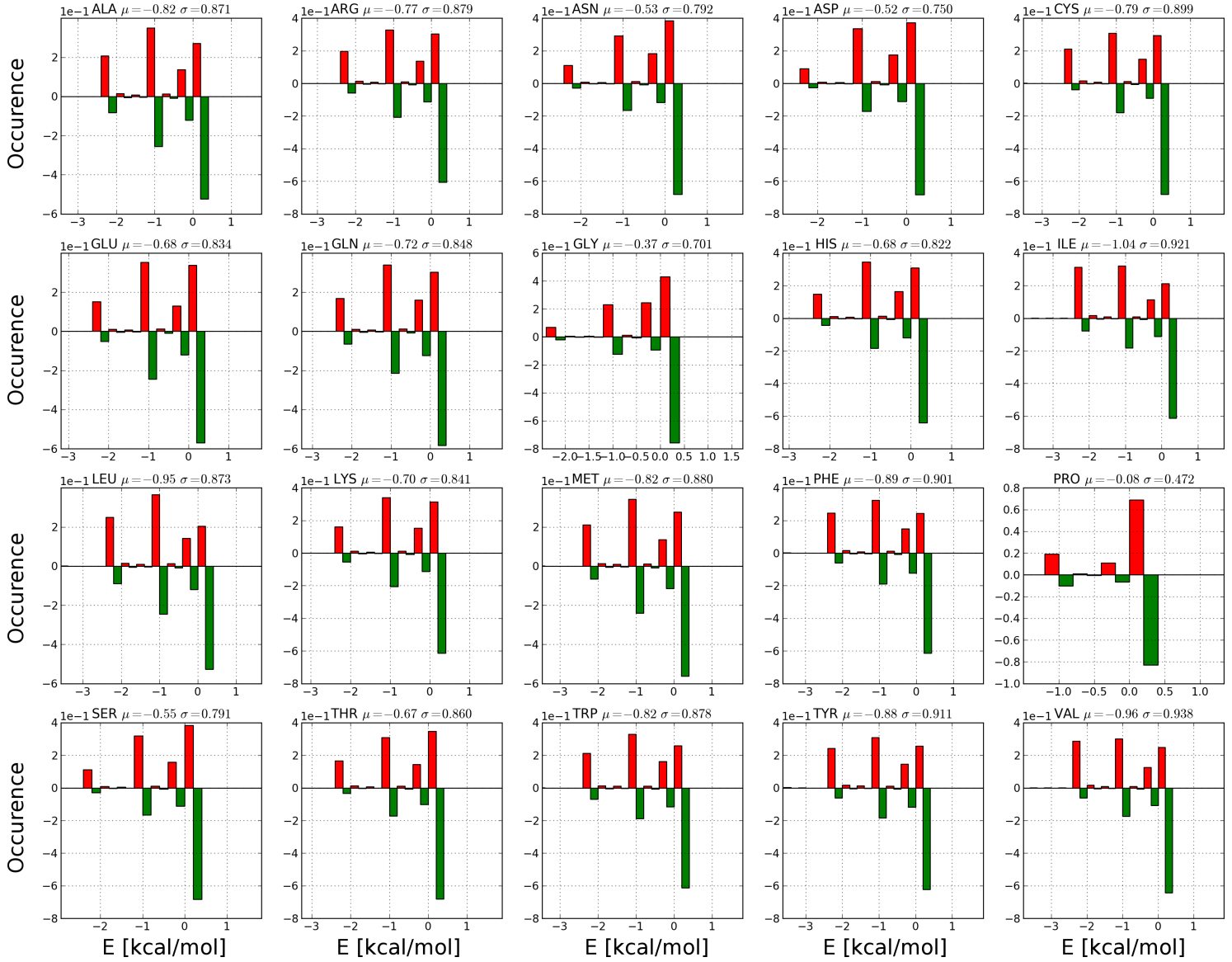


Fig. A.3.: *Statistics for the amino-acid specific main-chain hydrogen bonding contributions of all 20 amino acids. Positive occurences relate to energy distributions the set of native protein structures (red); negative occurences relate to the decoy energy distributions. All amino acids except for proline are more likely to develop two main-chain hydrogen bonds in the native structures than in the lower quality decoy sets.*

Bibliography

[1] J. Craig Venter et al. “The Sequence of the Human Genome”. In: *Science* 291.5507 (2001), pp. 1304–1351. URL: http://www.sciencemag.org/content/291/5507/1304.abstract.

[2] John Moult et al. “Critical assessment of methods of protein structure prediction—Round VIII”. In: *Proteins: Structure, Function, and Bioinformatics* 77.S9 (2009), pp. 1–4. ISSN: 1097-0134. DOI: 10.1002/prot.22589. URL: http://dx.doi.org/10.1002/prot.22589.

[3] Helen M. Berman et al. “The Protein Data Bank”. In: *Nucleic Acids Research* 28.1 (2000), pp. 235–242. URL: http://nar.oxfordjournals.org/content/28/1/235.abstract.

[4] Helen M. Berman et al. “Outcome of a Workshop on Archiving Structural Models of Biological Macromolecules”. In: *Structure* 14.8 (Aug. 2006), pp. 1211–1217. ISSN: 0969-2126. DOI: 10.1016/j.str.2006.06.005. URL: http://www.sciencedirect.com/science/article/pii/S0969212606002875.

[5] Vincent B. Chen et al. “MolProbity: all-atom structure validation for macromolecular crystallography”. In: *Acta Crystallographica Section D* 66.1 (2010), pp. 12–21. URL: http://dx.doi.org/10.1107/S0907444909042073.

[6] David E. Shaw et al. “Atomic-Level Characterization of the Structural Dynamics of Proteins”. In: *Science* 330.6002 (2010), pp. 341–346. URL: http://www.sciencemag.org/content/330/6002/341.abstract.

[7] Mikhail Efremov et al. “Probing Glass Transition of Ultrathin Polymer Films at a Time Scale of Seconds Using Fast Differential Scanning Calorimetry”. In: *Macromolecules* 37.12 (May 2004), pp. 4607–4616. DOI: 10.1021/ma035909r. URL: http://dx.doi.org/10.1021/ma035909r.

[8] T. Herges and W. Wenzel. “An all-atom force field for tertiary structure prediction of helical proteins.” In: *Biophys J* 87.5 (Nov. 2004). 15507688, pp. 3100–3109. ISSN: 0006-3495. URL: http://www.hubmed.org/display.cgi?uids=15507688.

[9] Abhinav Verma and Wolfgang Wenzel. “A free-energy approach for all-atom protein simulation.” In: *Biophys J* 96.9 (May 2009). 19413955, pp. 3483–3494. ISSN: 1542-0086. URL: http://www.hubmed.org/display.cgi?uids=19413955.

[10] T.B. Fischer et al. “The binding interface database (BID): a compilation of amino acid hot spots in protein interfaces.” In: *Bioinformatics* 19.11 (July 2003). 12874065, pp. 1453–1454. ISSN: 1367-4803. URL: http://www.hubmed.org/display.cgi?uids=12874065.

[11] Fangqing Xie et al. “Multilevel Atomic-Scale Transistors Based on Metallic Quantum Point Contacts”. In: *Advanced Materials* 22.18 (2010), pp. 2033–2036. ISSN: 1521-4095. DOI: 10.1002/adma.200902953. URL: http://dx.doi.org/10.1002/adma.200902953.

[12] Carl-Ivar Branden and John Tooze. *Introduction to Protein Structure*. Garland Publishing, Jan. 1999. ISBN: 0815323050. URL: http://www.amazon.ca/exec/obidos/redirect?tag=citeulike09-20&path=ASIN/0815323050.

[13] Craig D. Livingstone and Geoffrey J. Barton. “Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation”. In: *Computer applications in the biosciences : CABIOS* 9.6 (1993), pp. 745–756. URL: http://bioinformatics.oxfordjournals.org/content/9/6/745.abstract.

[14] Linus Pauling and Robert B. Corey. “The Pleated Sheet, A New Layer Configuration of Polypeptide Chains”. In: *Proceedings of the National Academy of Sciences* 37.5 (May 1951), pp. 251 –256. URL: http://www.pnas.org/content/37/5/251.short.

[15] Linus Pauling, Robert B. Corey, and H. R. Branson. “The structure of proteins: Two hydrogen-bonded helical configurations of the polypeptide chain”. In: *Proceedings of the National Academy of Sciences* 37.4 (Apr. 1951), pp. 205 –211. URL: http://www.pnas.org/content/37/4/205.short.

[16] C.B. Anfinsen. “Principles that govern the folding of protein chains.” In: *Science* 181.4096 (July 1973). 4124164, pp. 223–230. ISSN: 0036-8075. URL: http://www.hubmed.org/display.cgi?uids=4124164.

[17] Cyrus Levinthal. “Are there pathways for protein folding?” In: *Journal de Chimie Physique et de Physico-Chimie Biologique.* 65 (1968), pp. 44–45.

[18] J N Onuchic et al. “Toward an outline of the topography of a realistic protein-folding funnel”. In: *Proceedings of the National Academy of Sciences* 92.8 (Apr. 1995), pp. 3626 –3630. URL: http://www.pnas.org/content/92/8/3626.abstract.

[19] Vishwas R. Agashe, M. C. R. Shastry, and Jayant B. Udgaonkar. “Initial hydrophobic collapse in the folding of barstar”. In: *Nature* 377.6551 (Oct. 1995), pp. 754–757. DOI: 10.1038/377754a0. URL: http://dx.doi.org/10.1038/377754a0.

[20] Peter S. Kim and Robert L. Baldwin. “Specific Intermediates in the Folding Reactions of Small Proteins and the Mechanism of Protein Folding”. In: *Annual Review of Biochemistry* 51.1 (1982), pp. 459–489. URL: http://www.annualreviews.org/doi/abs/10.1146/annurev.bi.51.070182.002331.

[21] Timo Strunk. “Protein Structure Prediction using Free-Energy Methods and Volunteer Distributed Computing Networks”. Diploma Thesis. Dortmund: TU Dortmund, 2008.

[22] D. Eisenberg et al. “Analysis of membrane and surface protein sequences with the hydrophobic moment plot.” In: *Journal of molecular biology* 179.1 (Oct. 1984), pp. 125–42.

[23] C. Nick Pace. “Measuring and increasing protein stability”. In: *Trends in Biotechnology* 8.0 (1990), pp. 93–98. ISSN: 0167-7799. DOI: 10.1016/0167-7799(90)90146-O. URL: http://www.sciencedirect.com/science/article/pii/016777990901460.

[24] Gerhard Vogt and Patrick Argos. “Protein thermal stability: hydrogen bonds or internal packing?” In: *Folding and Design* 2, Supplement 1.0 (June 1997), S40–S46. ISSN: 1359-0278. DOI: 10.1016/S1359-0278(97)00062-X. URL: http://www.sciencedirect.com/science/article/pii/S135902789700062X.

[25] Darin M. Taverna and Richard A. Goldstein. “Why are proteins marginally stable?” In: *Proteins: Structure, Function, and Bioinformatics* 46.1 (2002), pp. 105–109. ISSN: 1097-0134. DOI: 10.1002/prot.10016. URL: http://dx.doi.org/10.1002/prot.10016.

[26] Nobuhiko Tokuriki et al. “How Protein Stability and New Functions Trade Off”. In: *PLoS Comput Biol* 4.2 (Feb. 2008), e1000002. DOI: 10.1371/journal.pcbi.1000002. URL: http://dx.plos.org/10.1371%2Fjournal.pcbi.1000002.

[27] Raquel Godoy-Ruiz et al. “Relation Between Protein Stability, Evolution and Structure, as Probed by Carboxylic Acid Mutations”. In: *Journal of Molecular Biology* 336.2 (Feb. 2004), pp. 313–318. ISSN: 0022-2836. DOI: 10.1016/j.jmb.2003.12.048. URL: http://www.sciencedirect.com/science/article/pii/S0022283603015651.

[28] Jon A Kenniston et al. “Effects of local protein stability and the geometric position of the substrate degradation tag on the efficiency of ClpXP denaturation and degradation”. In: *Journal of Structural Biology* 146.1–2 (Apr. 2004), pp. 130–140. ISSN: 1047-8477. DOI: 10.1016/j.jsb.2003.10.023. URL: http://www.sciencedirect.com/science/article/pii/S1047847703002375.

[29] Neil P. King et al. “Computational Design of Self-Assembling Protein Nanomaterials with Atomic Level Accuracy”. In: *Science* 336.6085 (2012), pp. 1171–1174. URL: http://www.sciencemag.org/content/336/6085/1171.abstract.

[30] Ron O. Dror et al. “Biomolecular Simulation: A Computational Microscope for Molecular Biology”. In: *Annual Review of Biophysics* 41.1 (2012), pp. 429–452. URL: http://www.annualreviews.org/doi/abs/10.1146/annurev-biophys-042910-155245.

[31] Guy G Dodson, David P Lane, and Chandra S Verma. “Molecular simulations of protein dynamics: new windows on mechanisms in biology”. In: *EMBO Rep* 9.2 (Feb. 2008), pp. 144–150. ISSN: 1469-221X. DOI: 10.1038/sj.embor.7401160. URL: http://dx.doi.org/10.1038/sj.embor.7401160.

[32] K. Binder. “Monte-Carlo Methods”. In: *Mathematical Tools for Physicists*. Wiley-VCH Verlag GmbH & Co. KGaA, 2006, pp. 249–280. ISBN: 9783527607778. URL: http://dx.doi.org/10.1002/3527607773.ch9.

[33] Berk Hess et al. “GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation”. In: *Journal of Chemical Theory and Computation* 4.3 (Feb. 2008), pp. 435–447. ISSN: 1549-9618. DOI: 10.1021/ct700301q. URL: http://dx.doi.org/10.1021/ct700301q.

[34] David A. Case et al. “The Amber biomolecular simulation programs”. In: *Journal of Computational Chemistry* 26.16 (2005), pp. 1668–1688. ISSN: 1096-987X. DOI: 10.1002/jcc.20290. URL: http://dx.doi.org/10.1002/jcc.20290.

[35] B. R. Brooks et al. “CHARMM: The biomolecular simulation program”. In: *Journal of Computational Chemistry* 30.10 (2009), pp. 1545–1614. ISSN: 1096-987X. DOI: 10.1002/jcc.21287. URL: http://dx.doi.org/10.1002/jcc.21287.

[36] Loup Verlet. “Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules”. In: *Phys. Rev.* 159.1 (July 1967), 98–103. DOI: 10.1103/PhysRev.159.98. URL: http://link.aps.org/doi/10.1103/PhysRev.159.98.

[37] Arthur Voter. “INTRODUCTION TO THE KINETIC MONTE CARLO METHOD”. In: *Radiation Effects in Solids*. Ed. by Kurt Sickafus, Eugene Kotomin, and Blas Uberuaga. Vol. 235. Springer Netherlands, 2007, pp. 1–23. ISBN: 978-1-4020-5293-4. URL: http://dx.doi.org/10.1007/978-1-4020-5295-8_1.

[38] T. Strunk et al. “SIMONA 1.0: An efficient and versatile framework for stochastic simulations of molecular and nanoscale systems”. In: *Journal of Computational Chemistry* (2012), n/a. ISSN: 1096-987X. DOI: 10.1002/jcc.23089. URL: http://dx.doi.org/10.1002/jcc.23089.

[39] Nicholas Metropolis et al. “Equation of State Calculations by Fast Computing Machines”. In: *The Journal of Chemical Physics* 21.6 (June 1953), pp. 1087–1092. URL: http://dx.doi.org/10.1063/1.1699114.

[40] Fred Glover. “Tabu Search—Part I”. In: *ORSA Journal on Computing* 1.3 (1989), pp. 190–206. URL: http://joc.journal.informs.org/content/1/3/190.abstract.

[41] Ting Wang and Xiaolong Zhang. “A case study of 3D protein structure prediction with genetic algorithm and Tabu search”. English. In: *Wuhan University Journal of Natural Sciences* 16.2 (2011), pp. 125–129. ISSN: 1007-1202. DOI: 10.1007/s11859-011-0723-1. URL: http://dx.doi.org/10.1007/s11859-011-0723-1.

[42] Charles Geyer. “Practical Markov Chain Monte Carlo”. In: *Statistical Science* 7.4 (1992), pp. 473–483. ISSN: 08834237. URL: http://dx.doi.org/10.2307/2246094.

[43] Robert H. Swendsen and Jian-Sheng Wang. “Replica Monte Carlo Simulation of Spin-Glasses”. In: *Phys. Rev. Lett.* 57.21 (Nov. 1986), 2607–2609. DOI: 10.1103/PhysRevLett.57.2607. URL: http://link.aps.org/doi/10.1103/PhysRevLett.57.2607.

[44] Koji Hukushima and Koji Nemoto. “Exchange Monte Carlo Method and Application to Spin Glass Simulations”. In: *Journal of the Physical Society of Japan* 65.6 (1996). Copyright (C) 1996 The Physical Society of Japan, p. 1604.

[45] Ulrich H.E. Hansmann. “Parallel tempering algorithm for conformational studies of biological molecules”. In: *Chemical Physics Letters* 281.1–3 (Dec. 1997), pp. 140–150. ISSN: 0009-2614. DOI: 10.1016/S0009-2614(97)01198-6. URL: http://www.sciencedirect.com/science/article/pii/S0009261497011986.

[46] Viktor Hornak et al. “Comparison of multiple Amber force fields and development of improved protein backbone parameters”. In: *Proteins: Structure, Function, and Bioinformatics* 65.3 (2006), pp. 712–725. ISSN: 1097-0134. DOI: 10.1002/prot.21123. URL: http://dx.doi.org/10.1002/prot.21123.

[47] Thomas E. Cheatham, Piotr Cieplak, and Peter A. Kollman. “A Modified Version of the Cornell et al. Force Field with Improved Sugar Pucker Phases and Helical Repeat”. In: *Journal of Biomolecular Structure and Dynamics* 16.4 (1999), pp. 845–862. ISSN: 0739-1102. DOI: 10.1080/07391102.1999.10508297. URL: http://dx.doi.org/10.1080/07391102.1999.10508297.

[48] Junmei Wang, Piotr Cieplak, and Peter A. Kollman. “How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules?” In: *Journal of Computational Chemistry* 21.12 (2000), pp. 1049–1074. ISSN: 1096-987X. DOI: 10.1002/1096-987X(200009)21:12<1049::AID-JCC3>3.0.CO;2-F. URL: http://dx.doi.org/10.1002/1096-987X(200009)21:12<1049::AID-JCC3>3.0.CO;2-F.

[49] William L. Jorgensen, David S. Maxwell, and Julian Tirado-Rives. “Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids”. In: *Journal of the American Chemical Society* 118.45 (Jan. 1996), pp. 11225–11236. ISSN: 0002-7863. DOI: 10.1021/ja9621760. URL: http://dx.doi.org/10.1021/ja9621760.

[50] Kresten Lindorff-Larsen et al. “Improved side-chain torsion potentials for the Amber ff99SB protein force field”. In: *Proteins: Structure, Function, and Bioinformatics* 78.8 (2010), pp. 1950–1958. ISSN: 1097-0134. DOI: 10.1002/prot.22711. URL: http://dx.doi.org/10.1002/prot.22711.

[51] J. E. Jones. “On the Determination of Molecular Fields. II. From the Equation of State of a Gas”. In: *Proceedings of the Royal Society of London. Series A* 106.738 (Oct. 1924), pp. 463 –477. URL: http://rspa.royalsocietypublishing.org/content/106/738/463.short.

[52] Charles de Coulomb. “Premier Memoire sur l’Electricite et le Magnetisme”. In: *Histoire de l’Academie Royale des Sciences* (1785), pp. 569–577.

[53] P.O. Jääskeläinen et al. “OpenCL-based design methodology for application-specific processors”. In: *Embedded Computer Systems (SAMOS), 2010 International Conference on*. July 2010, pp. 223–230. DOI: 10.1109/ICSAMOS.2010.5642061.

[54] Josh Barnes and Piet Hut. “A hierarchical O(N log N) force-calculation algorithm”. In: *Nature* 324.6096 (Dec. 1986), pp. 446–449. DOI: 10.1038/324446a0. URL: http://dx.doi.org/10.1038/324446a0.

[55] L. Greengard and V. Rokhlin. “A Fast Algorithm for Particle Simulations”. In: *Journal of Computational Physics* 135.2 (Aug. 1997), pp. 280–292. ISSN: 0021-9991. DOI: 10.1006/jcph.1997.5706. URL: http://www.sciencedirect.com/science/article/pii/S0021999197957065.

[56] Erich Elsen et al. “N-Body simulation on GPUs”. In: *Proceedings of the 2006 ACM/IEEE conference on Supercomputing*. Tampa, Florida: ACM, 2006, p. 188. ISBN: 0-7695-2700-0.

[57] Timo Strunk, Moritz Wolf, and Wolfgang Wenzel. “Development and evaluation of a GPU-optimized N-body term for the simulation of biomolecules”. In: *Computational Methods in Science and Engineering: Proceedings of the Workshop SimLabsKIT, November 29 - 30, 2010, Karlsruhe, Germany*. Ed. by I. Kondov et al. Karlsruhe, Germany: KIT Scientific Publishing, Karlsruhe, 2011, pp. 35–46. URL: http://digbib.ubka.uni-karlsruhe.de/volltexte/1000023323.

[58] Martin Burtscher and Keshav Pingali. “Tree-Based Barnes Hut n-Body Algorithm”. In: *GPU Computing Gems Emerald Edition*. Chapter 6. Morgan Kaufmann, 2011, p. 75.

[59] R. Tsuchiyama et al. “The OpenCL Programming Book”. In: *Group* (2011).

[60] D. Eisenberg and A.D. McLachlan. “Solvation energy in protein folding and binding.” In: *Nature* 319.6050 (Jan. 1986). 3945310, pp. 199–203. ISSN: 0028-0836. URL: http://www.hubmed.org/display.cgi?uids=3945310.

[61] T Herges. “Entwicklung eines Kraftfelds zur Strukturvorhersage von Helixproteinen”. PhD thesis. Universität Dortmund, 2003.

[62] Amedeo Caflisch and Martin Karplus. “Acid and Thermal Denaturation of Barnase Investigated by Molecular Dynamics Simulations”. In: *Journal of Molecular Biology* 252.5 (Oct. 1995), pp. 672–708. ISSN: 0022-2836. DOI: 10.1006/jmbi.1995.0528. URL: http://www.sciencedirect.com/science/article/pii/S0022283685705281.

[63] Konstantin V. Klenin et al. “Derivatives of molecular surface area and volume: Simple and exact analytical formulas”. In: *Journal of Computational Chemistry* 32.12 (2011), pp. 2647–2653. ISSN: 1096-987X. DOI: 10.1002/jcc.21844. URL: http://dx.doi.org/10.1002/jcc.21844.

[64] F. Fogolari, A. Brigo, and H. Molinari. “The Poisson–Boltzmann equation for biomolecular electrostatics: a tool for structural biology”. In: *Journal of Molecular Recognition* 15.6 (2002), pp. 377–392. ISSN: 1099-1352. DOI: 10.1002/jmr.577. URL: http://dx.doi.org/10.1002/jmr.577.

[65] W. Clark Still et al. “Semianalytical treatment of solvation for molecular mechanics and dynamics”. In: *J. Am. Chem. Soc.* 112.16 (1990), pp. 6127–6129. ISSN: 0002-7863. DOI: 10.1021/ja00172a038. URL: http://dx.doi.org/10.1021/ja00172a038.

[66] Junmei Wang et al. “Development and testing of a general amber force field.” In: *J Comput Chem* 25.9 (July 2004). 15116359, pp. 1157–1174. ISSN: 0192-8651. URL: http://www.hubmed.org/display.cgi?uids=15116359.

[67] Jerry Tsai et al. “An improved protein decoy set for testing energy functions for protein structure prediction.” In: *Proteins* 53.1 (Oct. 2003). 12945051, pp. 76–87. ISSN: 1097-0134. URL: http://www.hubmed.org/display.cgi?uids=12945051.

[68] Abhinav Verma et al. “All-atom de novo protein folding with a scalable evolutionary algorithm.” In: *J Comput Chem* 28.16 (Dec. 2007). 17486550, pp. 2552–2558. ISSN: 0192-8651. URL: http://www.hubmed.org/display.cgi?uids=17486550.

[69] Silvia Pandolfi, Francesco Bartolucci, and Nial Friel. “A generalization of the Multiple-try Metropolis algorithm for Bayesian estimation and model selection”. In: *Journal of Machine Learning Research - Proceedings Track* (2010), pp. 581–588.

[70] Juha Nieminen and Joel Yliluoma. *Function Parser library for C++*. 2011. URL: http://warp.povusers.org/FunctionParser/ (visited on 09/26/2012).

[71] Timo Strunk et al. “Benchmarking the POEM@HOME Network for Protein Structure Prediction”. In: Westminster, London, UK: University of Westminster, CEUR-WS, 2011.

[72] Michele Magrane and UniProt Consortium. “UniProt Knowledgebase: a hub of integrated protein data”. In: *Database* 2011 (Jan. 2011). URL: http://database.oxfordjournals.org/content/2011/bar009.abstract.

[73] Narayanan Eswar et al. “Comparative Protein Structure Modeling Using Modeller”. In: *Current Protocols in Bioinformatics*. John Wiley & Sons, Inc., 2002. ISBN: 9780471250951. URL: http://dx.doi.org/10.1002/0471250953.bi0506s15.

[74] Dylan Chivian et al. “Automated prediction of CASP-5 structures using the Robetta server.” In: *Proteins* 53 Suppl 6 (2003). 14579342, pp. 524–533. ISSN: 1097-0134. URL: http://www.hubmed.org/display.cgi?uids=14579342.

[75] Ambrish Roy and Yang Zhang. “Protein Structure Prediction”. In: *eLS*. John Wiley & Sons, Ltd, 2001. ISBN: 9780470015902. URL: http://dx.doi.org/10.1002/9780470015902.a0003031.pub2.

[76] Andrej Šali and Tom L. Blundell. “Comparative Protein Modelling by Satisfaction of Spatial Restraints”. In: *Journal of Molecular Biology* 234.3 (Dec. 1993), pp. 779–815. ISSN: 0022-2836. DOI: 10.1006/jmbi.1993.1626. URL: http://www.sciencedirect.com/science/article/pii/S0022283683716268.

[77] Min-yi Shen and Andrej Sali. “Statistical potential for assessment and prediction of protein structures”. In: *Protein Science* 15.11 (2006), pp. 2507–2524. ISSN: 1469-896X. DOI: 10.1110/ps.062416606. URL: http://dx.doi.org/10.1110/ps.062416606.

[78] David Eramian et al. “A composite score for predicting errors in protein structure models”. In: *Protein Science* 15.7 (2006), pp. 1653–1666. ISSN: 1469-896X. DOI: 10.1110/ps.062095806. URL: http://dx.doi.org/10.1110/ps.062095806.

[79] T. Strunk and W. Wenzel. “Absolute Quality Assessment of Protein Models”. In preparation. 2013.

[80] Andrew Leaver-Fay et al. “Rosetta3: An Object-Oriented Software Suite for the Simulation and Design of Macromolecules”. In: *Computer Methods, Part C*. Vol. Volume 487. Academic Press, 2011, pp. 545–574. ISBN: 0076-6879. URL: http://www.sciencedirect.com/science/article/pii/B9780123812704000196.

[81] Yang Zhang and Jeffrey Skolnick. “Scoring function for automated assessment of protein structure template quality”. In: *Proteins: Structure, Function, and Bioinformatics* 57.4 (2004), pp. 702–710. ISSN: 1097-0134. DOI: 10.1002/prot.20264. URL: http://dx.doi.org/10.1002/prot.20264.

[82] Guoli Wang and Roland L. Dunbrack. “PISCES: a protein sequence culling server”. In: *Bioinformatics* 19.12 (2003), pp. 1589–1591. URL: http://bioinformatics.oxfordjournals.org/content/19/12/1589.abstract.

[83] Y. Zhang and J. Skolnick. “TM-align: a protein structure alignment algorithm based on the TM-score”. In: *Nucleic acids research* 33.7 (2005), pp. 2302 –2309.

[84] R. Rajgaria, S. R. McAllister, and C. A. Floudas. “A novel high resolution Cα-Cα distance dependent force field based on a high quality decoy set”. In: *Proteins: Structure, Function, and Bioinformatics* 65.3 (2006), pp. 726–741. ISSN: 1097-0134. DOI: 10.1002/prot.21149. URL: http://dx.doi.org/10.1002/prot.21149.

[85] D. Baker and A. Sali. “Protein structure prediction and structural genomics”. In: *Science* 294.5540 (2001), pp. 93 –96.

[86] Mark V. Berjanskii et al. “NMR Structure of the N-terminal J Domain of Murine Polyomavirus T Antigens”. In: *Journal of Biological Chemistry* 275.46 (2000), pp. 36094–36103. URL: http://www.jbc.org/content/275/46/36094.abstract.

[87] P. Sevilla-Sierra, G. Otting, and K. Wüthrich. “Determination of the Nuclear Magnetic Resonance Structure of the DNA-binding Domain of the P22 c2 Repressor (1 to 76) in Solution and Comparison with the DNA-binding Domain of the 434 Repressor”. In: *Journal of Molecular Biology* 235.3 (Jan. 1994), pp. 1003–1020. ISSN: 0022-2836. DOI: 10.1006/jmbi.1994.1053. URL: http://www.sciencedirect.com/science/article/pii/S0022283684710539.

[88] Steven F Sukits et al. “Solution structure of the tumor necrosis factor receptor-1 death domain”. In: *Journal of Molecular Biology* 310.4 (July 2001), pp. 895–906. ISSN: 0022-2836. DOI: 10.1006/jmbi.2001.4790. URL: http://www.sciencedirect.com/science/article/pii/S0022283601947904.

[89] Wayne J Fairbrother et al. “Solution structure of the heparin-binding domain of vascular endothelial growth factor”. In: *Structure* 6.5 (May 1998), pp. 637–648. ISSN: 0969-2126. DOI: 10.1016/S0969-2126(98)00065-3. URL: http://www.sciencedirect.com/science/article/pii/S0969212698000653.

[90] Yanwu Yang et al. “Reactivity of the Human Thioltransferase (Glutaredoxin) C7S, C25S, C78S, C82S Mutant and NMR Solution Structure of Its Glutathionyl Mixed Disulfide Intermediate Reflect Catalytic Specificity†,‡”. In: *Biochemistry* 37.49 (Nov. 1998), pp. 17145–17156. ISSN: 0006-2960. DOI: 10.1021/bi9806504. URL: http://dx.doi.org/10.1021/bi9806504.

[91] Hiroaki Sasakawa et al. “Structure of POIA1, a homologous protein to the propeptide of subtilisin: implication for protein foldability and the function as an intramolecular chaperone”. In: *Journal of Molecular Biology* 317.1 (Mar. 2002), pp. 159–167. ISSN: 0022-2836. DOI: 10.1006/jmbi.2002.5412. URL: http://www.sciencedirect.com/science/article/pii/S0022283602954124.

[92] Masato Kato et al. “Insights into Multistep Phosphorelay from the Crystal Structure of the C-Terminal HPt Domain of ArcB”. In: *Cell* 88.5 (Mar. 1997), pp. 717–723. ISSN: 0092-8674. DOI: 10.1016/S0092-8674(00)81914-5. URL: http://www.sciencedirect.com/science/article/pii/S0092867400819145.

[93] Yunjun Wang et al. “Solution Structure of Carnobacteriocin B2 and Implications for Structure-Activity Relationships among Type IIa Bacteriocins from Lactic Acid Bacteria†,‡”. In: *Biochemistry* 38.47 (Nov. 1999), pp. 15438–15447. ISSN: 0006-2960. DOI: 10.1021/bi991351x. URL: http://dx.doi.org/10.1021/bi991351x.

[94] Johan Kemmink et al. “Structure Determination of the N-Terminal Thioredoxin-like Domain of Protein Disulfide Isomerase Using Multidimensional Heteronuclear 13C/15N NMR Spectroscopy†”. In: *Biochemistry* 35.24 (Jan. 1996), pp. 7684–7691. ISSN: 0006-2960. DOI: 10.1021/bi960335m. URL: http://dx.doi.org/10.1021/bi960335m.

[95] Kai Huang, John M. Flanagan, and James H. Prestegard. “The influence of C-terminal extension on the structure of the “J-domain” in E. coli DnaJ”. In: *Protein Science* 8.1 (1999), pp. 203–214. ISSN: 1469-896X. DOI: 10.1110/ps.8.1.203. URL: http://dx.doi.org/10.1110/ps.8.1.203.

[96] Philip Bradley et al. “Rosetta predictions in CASP5: Successes, failures, and prospects for complete automation”. In: *Proteins: Structure, Function, and Bioinformatics* 53.S6 (2003), pp. 457–468. ISSN: 1097-0134. DOI: 10.1002/prot.10552. URL: http://dx.doi.org/10.1002/prot.10552.

[97] Dylan Chivian et al. “Prediction of CASP6 structures using automated robetta protocols”. In: *Proteins: Structure, Function, and Bioinformatics* 61.S7 (2005), pp. 157–166. ISSN: 1097-0134. DOI: 10.1002/prot.20733. URL: http://dx.doi.org/10.1002/prot.20733.

[98] Mark D. Adams et al. “The Genome Sequence of Drosophila melanogaster”. In: *Science* 287.5461 (2000), pp. 2185–2195. URL: http://www.sciencemag.org/content/287/5461/2185.abstract.

[99] Stephane Boeuf et al. “Engineering hydrophobin DewA to generate surfaces that enhance adhesion of human but not bacterial cells”. In: *Acta Biomaterialia* 8.3 (Mar. 2012), pp. 1037–1047. ISSN: 1742-7061. DOI: 10.1016/j.actbio.2011.11.022. URL: http://www.sciencedirect.com/science/article/pii/S1742706111005253.

[100] Timo Strunk et al. “Structural model of the gas vesicle protein GvpA and analysis of GvpA mutants in vivo”. In: *Molecular Microbiology* 81.1 (2011), pp. 56–68. ISSN: 1365-2958. DOI: 10.1111/j.1365-2958.2011.07669.x. URL: http://dx.doi.org/10.1111/j.1365-2958.2011.07669.x.

[101] Timo Strunk, Moritz Wolf, and Wolfgang Wenzel. “Peptide structure prediction using distributed volunteer computing networks”. In: *Journal of Mathematical Chemistry* 50.2 (Feb. 2012), pp. 421–428. ISSN: 0259-9791. DOI: 10.1007/s10910-011-9937-x. URL: http://dx.doi.org/10.1007/s10910-011-9937-x.

[102] C. Chothia and A.M. Lesk. “The relationship between the divergence of sequence and structure in proteins”. In: *EMBO J.* 5 (1986), 823–826.

[103] Jean-Francois Gibrat, Thomas Madej, and Stephen H Bryant. “Surprising similarities in structure comparison”. In: *Current Opinion in Structural Biology* 6.3 (June 1996), pp. 377–385. ISSN: 0959-440X. DOI: 10.1016/S0959-440X(96)80058-3. URL: http://www.sciencedirect.com/science/article/pii/S0959440X96800583.

[104] S. Henikoff and J. G. Henikoff. “Amino acid substitution matrices from protein blocks”. In: *Proceedings of the National Academy of Sciences* 89.22 (1992), pp. 10915–10919. URL: http://www.pnas.org/content/89/22/10915.abstract.

[105] T.F. Smith and M.S. Waterman. “Identification of common molecular subsequences”. In: *Journal of Molecular Biology* 147.1 (Mar. 1981), pp. 195–197. ISSN: 0022-2836. DOI: 10.1016/0022-2836(81)90087-5. URL: http://www.sciencedirect.com/science/article/pii/0022283681900875.

[106] Saul B. Needleman and Christian D. Wunsch. “A general method applicable to the search for similarities in the amino acid sequence of two proteins”. In: *Journal of Molecular Biology* 48.3 (Mar. 1970), pp. 443–453. ISSN: 0022-2836. DOI: 10.1016/0022-2836(70)90057-4. URL: http://www.sciencedirect.com/science/article/pii/0022283670900574.

[107] Yang Zhang. “I-TASSER server for protein 3D structure prediction”. In: *BMC Bioinformatics* 9.1 (2008), p. 40. URL: http://www.biomedcentral.com/1471-2105/9/40.

[108] Muhammad Sabir, Xiaoxue Xu, and Li Li. “A review on biodegradable polymeric materials for bone tissue engineering applications”. In: *Journal of Materials Science* 44.21 (Nov. 2009), pp. 5713–5724. ISSN: 0022-2461. DOI: 10.1007/s10853-009-3770-7. URL: http://dx.doi.org/10.1007/s10853-009-3770-7.

[109] S Ramakrishna et al. “Biomedical applications of polymer-composite materials: a review”. In: *Composites Science and Technology* 61.9 (July 2001), pp. 1189–1224. ISSN: 0266-3538. DOI: 10.1016/S0266-3538(00)00241-4. URL: http://www.sciencedirect.com/science/article/pii/S0266353800002414.

[110] Jurgen Groll et al. “Novel surface coatings modulating eukaryotic cell adhesion and preventing implant infection.” In: *Int J Artif Organs* 32.9 (2009). 19882552, pp. 655–662. ISSN: 0391-3988. URL: http://www.hubmed.org/display.cgi?uids=19882552.

[111] Stephane Boeuf and Wiltrud Richter. “Chondrogenesis of mesenchymal stem cells: role of tissue source and inducing factors.” In: *Stem Cell Res Ther* 1.4 (2010). 20959030, p. 31. ISSN: 1757-6512. URL: http://www.hubmed.org/display.cgi?uids=20959030.

[112] Andrej Trampuz and Andreas F. Widmer. “Infections associated with orthopedic implants.” In: *Curr Opin Infect Dis* 19.4 (2006). 16804382, pp. 349–356. ISSN: 0951-7375. URL: http://www.hubmed.org/display.cgi?uids=16804382.

[113] Vishukumar Aimanianda et al. “Surface hydrophobin prevents immune recognition of airborne fungal spores.” In: *Nature* 460.7259 (2009). 19713928, pp. 1117–1121. ISSN: 1476-4687. URL: http://www.hubmed.org/display.cgi?uids=19713928.

[114] J G Wessels. “Hydrophobins: proteins that change the nature of the fungal surface.” In: *Adv Microb Physiol* 38 (1997), pp. 1–45.

[115] M.I. Janssen et al. “Coating with genetic engineered hydrophobin promotes growth of fibroblasts on a hydrophobic solid.” In: *Biomaterials* 23.24 (Dec. 2002). 12361625, pp. 4847–4854. ISSN: 0142-9612. URL: http://www.hubmed.org/display.cgi?uids=12361625.

[116] Martijn F.B.G. Gebbink et al. “Amyloids–a functional coat for microorganisms.” In: *Nat Rev Microbiol* 3.4 (Apr. 2005). 15806095, pp. 333–341. ISSN: 1740-1526. URL: http://www.hubmed.org/display.cgi?uids=15806095.

[117] K. Scholtmeijer, J.G. Wessels, and H.A. Wosten. “Fungal hydrophobins in medical and technical applications.” In: *Appl Microbiol Biotechnol* 56.1-2 (July 2001). 11499914, pp. 1–8. ISSN: 0175-7598. URL: http://www.hubmed.org/display.cgi?uids=11499914.

[118] Han Wösten and Joseph Wessels. “Hydrophobins, from molecular structure to multiple functions in fungal development”. In: *Mycoscience* 38.3 (Oct. 1997), pp. 363–374. DOI: {10.1007/BF02464099}. URL: http://dx.doi.org/10.1007/BF02464099.

[119] Margaret Sunde et al. “Structural analysis of hydrophobins.” In: *Micron* 39.7 (Oct. 2008), pp. 773–84.

[120] A H Y Kwan et al. “Structural basis for rodlet assembly in fungal hydrophobins.” In: *Proc Natl Acad Sci U S A* 103.10 (Mar. 2006), pp. 3621–6.

[121] Ann H. Kwan et al. “The Cys3-Cys4 Loop of the Hydrophobin EAS Is Not Required for Rodlet Formation and Surface Activity”. In: *Journal of Molecular Biology* 382.3 (2008), pp. 708 –720. ISSN: 0022-2836. DOI: DOI:10.1016/j.jmb.2008.07.034. URL: http://www.sciencedirect.com/science/article/B6WK7-4T1Y41K-3/2/fadf321e7ecf2f693b1187c977302376.

[122] Marie-Anne Van Wetter, Han A. B. Wösten, and Joseph G. H. Wessels. “SC3 and SC4 hydrophobins have distinct roles in formation of aerial structures in dikaryons of Schizophyllum commune”. In: *Molecular Microbiology* 36.1 (2000), pp. 201–210. ISSN: 1365-2958. DOI: 10.1046/j.1365-2958.2000.01848.x. URL: http://dx.doi.org/10.1046/j.1365-2958.2000.01848.x.

[123] Han A. B. Wösten. “HYDROPHOBINS: Multipurpose Proteins”. In: *Annu. Rev. Microbiol.* 55.1 (2001), pp. 625–646. ISSN: 0066-4227. DOI: 10.1146/annurev.micro.55.1.625. URL: http://dx.doi.org/10.1146/annurev.micro.55.1.625.

[124] Wendel Wohlleben et al. “Recombinantly produced hydrophobins from fungal analogues as highly surface-active performance proteins.” In: *Eur Biophys J* 39.3 (Feb. 2010). 19290518, pp. 457–468. ISSN: 1432-1017. URL: http://www.hubmed.org/display.cgi?uids=19290518.

[125] M A Stringer and W E Timberlake. “dewA encodes a fungal hydrophobin component of the Aspergillus spore wall.” In: *Mol Microbiol* 16.1 (Apr. 1995), pp. 33–44.

[126] E. Ruoslahti and M.D. Pierschbacher. “New perspectives in cell adhesion: RGD and integrins.” In: *Science* 238.4826 (Oct. 1987). 2821619, pp. 491–497. ISSN: 0036-8075. URL: http://www.hubmed.org/display.cgi?uids=2821619.

[127] Ulrich Hersel, Claudia Dahmen, and Horst Kessler. “RGD modified polymers: biomaterials for stimulated cell adhesion and beyond.” In: *Biomaterials* 24.24 (Nov. 2003). 12922151, pp. 4385–4415. ISSN: 0142-9612. URL: http://www.hubmed.org/display.cgi?uids=12922151.

[128] Jin-Man Kim, Won Ho Park, and Byung-Moo Min. “The PPFLMLLKGSTR motif in globular domain 3 of the human laminin-5 alpha3 chain is crucial for integrin alpha3beta1 binding and cell adhesion.” In: *Exp Cell Res* 304.1 (Mar. 2005). 15707596, pp. 317–327. ISSN: 0014-4827. URL: http://www.hubmed.org/display.cgi?uids=15707596.

[129] Jack Kyte and Russell F. Doolittle. “A simple method for displaying the hydropathic character of a protein”. In: *Journal of Molecular Biology* 157.1 (May 1982), pp. 105–132. ISSN: 0022-2836. DOI: 10.1016/0022-2836(82)90515-0. URL: http://www.sciencedirect.com/science/article/pii/0022283682905150.

[130] András Fiser et al. “Modeller: Generation and Refinement of Homology-Based Protein Structure Models”. In: *Macromolecular Crystallography, Part D*. Vol. Volume 374. Academic Press, 2003, pp. 461–491. ISBN: 0076-6879. URL: http://www.sciencedirect.com/science/article/pii/S0076687903740208.

[131] A.H.Y. Kwan et al. “Structural basis for rodlet assembly in fungal hydrophobins.” In: *Proc Natl Acad Sci U S A* 103.10 (Mar. 2006). 16537446, pp. 3621–3626. ISSN: 0027-8424. URL: http://www.hubmed.org/display.cgi?uids=16537446.

[132] Marco Strohmeier et al. “Structure of a bacterial pyridoxal 5′-phosphate synthase complex”. In: *Proceedings of the National Academy of Sciences* 103.51 (2006), pp. 19284–19289. URL: http://www.pnas.org/content/103/51/19284.abstract.

[133] Guillaume Le Saux et al. “The Relative Importance of Topography and RGD Ligand Density for Endothelial Cell Adhesion”. In: *PLoS ONE* 6.7 (July 2011), e21869. DOI: 10.1371/journal.pone.0021869. URL: http://dx.doi.org/10.1371%2Fjournal.pone.0021869.

[134] M.A. Stringer et al. “Rodletless, a new Aspergillus developmental mutant induced by directed gene inactivation.” In: *Genes Dev* 5.7 (July 1991). 2065971, pp. 1161–1171. ISSN: 0890-9369. URL: http://www.hubmed.org/display.cgi?uids=2065971.

[135] M.A. Stringer and W.E. Timberlake. “dewA encodes a fungal hydrophobin component of the Aspergillus spore wall.” In: *Mol Microbiol* 16.1 (Apr. 1995). 7651135, pp. 33–44. ISSN: 0950-382X. URL: http://www.hubmed.org/display.cgi?uids=7651135.

[136] Jae Sam Lee, Jae Sung Lee, and William L. Murphy. “Modular peptides promote human mesenchymal stem cell differentiation on biomaterial surfaces.” In: *Acta Biomater* 6.1 (Jan. 2010). 19665062, pp. 21–28. ISSN: 1878-7568. URL: http://www.hubmed.org/display.cgi?uids=19665062.

[137] Ann H. Kwan et al. “The Cys3-Cys4 Loop of the Hydrophobin EAS Is Not Required for Rodlet Formation and Surface Activity”. In: *Journal of Molecular Biology* 382.3 (Oct. 2008), pp. 708–720. ISSN: 0022-2836. DOI: 10.1016/j.jmb.2008.07.034. URL: http://www.sciencedirect.com/science/article/pii/S0022283608008760.

[138] Rong Chen, Li Li, and Zhiping Weng. “ZDOCK: An initial-stage protein-docking algorithm”. In: *Proteins: Structure, Function, and Bioinformatics* 52.1 (2003), pp. 80–87. ISSN: 1097-0134. DOI: 10.1002/prot.10389. URL: http://dx.doi.org/10.1002/prot.10389.

[139] A. McLachlan. “Rapid comparison of protein structures”. In: *Acta Crystallographica Section A* 38.6 (1982), pp. 871–873. URL: http://dx.doi.org/10.1107/S0567739482001806.

[140] Joel P Mackay et al. “The Hydrophobin EAS Is Largely Unstructured in Solution and Functions by Forming Amyloid-Like Structures”. In: *Structure* 9.2 (Feb. 2001), pp. 83–91. ISSN: 0969-2126. DOI: 10.1016/S0969-2126(00)00559-1. URL: http://www.sciencedirect.com/science/article/pii/S0969212600005591.

[141] Adriano Aguzzi and Tracy O’Connor. “Protein aggregation diseases: pathogenicity and therapeutic perspectives”. In: *Nat Rev Drug Discov* 9.3 (Mar. 2010), pp. 237–248. ISSN: 1474-1776. DOI: 10.1038/nrd3050. URL: http://dx.doi.org/10.1038/nrd3050.

[142] Christopher A. Ross and Michelle A. Poirier. “Protein aggregation and neurodegenerative disease”. In: *Nature Medicine* 10.7 (2004). DOI: 10.1038/nm1066.

[143] Jingxian Liu and Jianxing Song. “Insights into Protein Aggregation by NMR Characterization of Insoluble SH3 Mutants Solubilized in Salt-Free Water”. In: *PLoS ONE* 4.11 (Nov. 2009), e7805. DOI: 10.1371/journal.pone.0007805. URL: http://dx.doi.org/10.1371%2Fjournal.pone.0007805.

[144] A.E. Walsby. “Gas vesicles.” In: *Microbiol Rev* 58.1 (Mar. 1994). 8177173, pp. 94–9144. ISSN: 0146-0749. URL: http://www.hubmed.org/display.cgi?uids=8177173.

[145] Marina Belenky, Rebecca Meyers, and Judith Herzfeld. “Subunit structure of gas vesicles: a MALDI-TOF mass spectrometry study.” In: *Biophys J* 86.1 Pt 1 (Jan. 2004). 14695294, pp. 499–505. ISSN: 0006-3495. URL: http://www.hubmed.org/display.cgi?uids=14695294.

[146] Astrid C. Sivertsen et al. “Solid-state NMR characterization of gas vesicle structure.” In: *Biophys J* 99.6 (2010). 20858439, pp. 1932–1939. ISSN: 1542-0086. URL: http://www.hubmed.org/display.cgi?uids=20858439.

[147] Walther Stoeckenius and Wolf H Kunau. “FURTHER CHARACTERIZATION OF PARTICULATE FRACTIONS FROM LYSED CELL ENVELOPES OF HALOBACTERIUM HALOBIUM AND ISOLATION OF GAS VACUOLE MEMBRANES”. In: *The Journal of Cell Biology* 38.2 (1968), pp. 337–357. URL: http://jcb.rupress.org/content/38/2/337.abstract.

[148] A.E. Blaurock and A.E. Walsby. “Crystalline structure of the gas vesicle wall from Anabaena flos-aquae”. In: *Journal of Molecular Biology* 105.2 (Aug. 1976), pp. 183–199. ISSN: 0022-2836. DOI: 10.1016/0022-2836(76)90106-6. URL: http://www.sciencedirect.com/science/article/pii/0022283676901066.

[149] A.E. Blaurock and W. Wober. “Structure of the wall of Halobacterium halobium gas vesicles.” In: *J Mol Biol* 106.3 (1976). 978738, pp. 871–878. ISSN: 0022-2836. URL: http://www.hubmed.org/display.cgi?uids=978738.

[150] T. McMaster, M. Miles, and A. Walsby. “Direct observation of protein secondary structure in gas vesicles by atomic force microscopy”. In: *Biophysical Journal* 70 (May 1996), pp. 2432–2436. DOI: 10.1016/S0006-3495(96)79813-2.

[151] P.K. Hayes, A.E. Walsby, and J.E. Walker. “Complete amino acid sequence of cyanobacterial gas-vesicle protein indicates a 70-residue molecule that corresponds in size to the crystallographic unit cell.” In: *Biochem J* 236.1 (May 1986). 3098234, pp. 31–36. ISSN: 0264-6021. URL: http://www.hubmed.org/display.cgi?uids=3098234.

[152] Astrid C. Sivertsen et al. “Solid-state NMR evidence for inequivalent GvpA subunits in gas vesicles.” In: *J Mol Biol* 387.4 (Apr. 2009). 19232353, pp. 1032–1039. ISSN: 1089-8638. URL: http://www.hubmed.org/display.cgi?uids=19232353.

[153] Krzysztof Ginalski et al. “3D-Jury: a simple approach to improve protein structure predictions.” In: *Bioinformatics* 19.8 (May 2003). 12761065, pp. 1015–1018. ISSN: 1367-4803. URL: http://www.hubmed.org/display.cgi?uids=12761065.

[154] Kevin Bryson et al. “Protein structure prediction servers at University College London.” In: *Nucleic Acids Res* 33.Web Server issue (July 2005). 15980489, pp. 36–38. ISSN: 1362-4962. URL: http://www.hubmed.org/display.cgi?uids=15980489.

[155] Kevin Karplus et al. “Combining local-structure, fold-recognition, and new fold methods for protein structure prediction.” In: *Proteins* 53 Suppl 6 (2003). 14579338, pp. 491–496. ISSN: 1097-0134. URL: http://www.hubmed.org/display.cgi?uids=14579338.

[156] Lawrence A. Kelley and Michael J.E. Sternberg. “Protein structure prediction on the Web: a case study using the Phyre server.” In: *Nat Protoc* 4.3 (2009). 19247286, pp. 363–371. ISSN: 1750-2799. URL: http://www.hubmed.org/display.cgi?uids=19247286.

[157] J. Shi, T.L. Blundell, and K. Mizuguchi. “FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties.” In: *J Mol Biol* 310.1 (June 2001). 11419950, pp. 243–257. ISSN: 0022-2836. URL: http://www.hubmed.org/display.cgi?uids=11419950.

[158] Yuji Zhang et al. “Network motif-based identification of transcription factor-target gene relationships by integrating multi-source biological data.” In: *BMC Bioinformatics* 9 (2008). 18426580, p. 203. ISSN: 1471-2105. URL: http://www.hubmed.org/display.cgi?uids=18426580.

[159] Sol Katzman et al. “PREDICT-2ND: a tool for generalized protein local structure prediction.” In: *Bioinformatics* 24.21 (Nov. 2008). 18757875, pp. 2453–2459. ISSN: 1367-4811. URL: http://www.hubmed.org/display.cgi?uids=18757875.

[160] Carol A. Rohl et al. “Protein structure prediction using Rosetta.” In: *Methods Enzymol* 383 (2004). 15063647, pp. 66–93. ISSN: 0076-6879. URL: http://www.hubmed.org/display.cgi?uids=15063647.

[161] Richard Bonneau et al. “De novo prediction of three-dimensional structures for major protein families.” In: *J Mol Biol* 322.1 (2002). 12215415, pp. 65–78. ISSN: 0022-2836. URL: http://www.hubmed.org/display.cgi?uids=12215415.

[162] Thomas Simonson, Georgios Archontis, and Martin Karplus. “Free energy simulations come of age: protein-ligand recognition.” In: *Acc Chem Res* 35.6 (June 2002). 12069628, pp. 430–437. ISSN: 0001-4842. URL: http://www.hubmed.org/display.cgi?uids=12069628.

[163] N. Siew et al. “MaxSub: an automated measure for the assessment of protein structure prediction quality.” In: *Bioinformatics* 16.9 (2000). 11108700, pp. 776–785. ISSN: 1367-4803. URL: http://www.hubmed.org/display.cgi?uids=11108700.

[164] A. Verma et al. “Basin hopping simulations for all-atom protein folding.” In: *J Chem Phys* 124.4 (Jan. 2006). 16460193, p. 044515. ISSN: 0021-9606. URL: http://www.hubmed.org/display.cgi?uids=16460193.

[165] James C. Phillips et al. “Scalable molecular dynamics with NAMD.” In: *J Comput Chem* 26.16 (Dec. 2005). 16222654, pp. 1781–1802. ISSN: 0192-8651. URL: http://www.hubmed.org/display.cgi?uids=16222654.

[166] Alexander D. Mackerell, Michael Feig, and Charles L. Brooks. “Extending the treatment of backbone energetics in protein force fields: limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations.” In: *J Comput Chem* 25.11 (2004). 15185334, pp. 1400–1415. ISSN: 0192-8651. URL: http://www.hubmed.org/display.cgi?uids=15185334.

[167] David Van Der Spoel et al. “GROMACS: fast, flexible, and free.” In: *J Comput Chem* 26.16 (Dec. 2005). 16211538, pp. 1701–1718. ISSN: 0192-8651. URL: http://www.hubmed.org/display.cgi?uids=16211538.

[168] Chu Wang, Philip Bradley, and David Baker. “Protein-protein docking with backbone flexibility.” In: *J Mol Biol* 373.2 (Oct. 2007). 17825317, pp. 503–519. ISSN: 0022-2836. URL: http://www.hubmed.org/display.cgi?uids=17825317.

[169] Manuel Rueda et al. “A consensus view of protein dynamics”. In: *Proceedings of the National Academy of Sciences* 104.3 (2007), pp. 796–801. URL: http://www.pnas.org/content/104/3/796.abstract.

[170] Dan I. Andersson and Diarmaid Hughes. “Antibiotic resistance and its cost: is it possible to reverse resistance?” In: *Nat Rev Micro* 8.4 (Apr. 2010), pp. 260–271. ISSN: 1740-1526. DOI: 10.1038/nrmicro2319. URL: http://dx.doi.org/10.1038/nrmicro2319.

[171] Julian Davies and Dorothy Davies. “Origins and Evolution of Antibiotic Resistance”. In: *Microbiology and Molecular Biology Reviews* 74.3 (Sept. 2010), pp. 417–433. URL: http://mmbr.asm.org/content/74/3/417.abstract.

[172] J. H. Powers. “Antimicrobial drug development – the past, the present, and the future”. In: *Clinical Microbiology and Infection* 10 (2004), pp. 23–31. ISSN: 1469-0691. DOI: 10.1111/j.1465-0691.2004.1007.x. URL: http://dx.doi.org/10.1111/j.1465-0691.2004.1007.x.

[173] Kim A. Brogden. “Antimicrobial peptides: pore formers or metabolic inhibitors in bacteria?” In: *Nat Rev Micro* 3.3 (Mar. 2005), pp. 238–250. ISSN: 1740-1526. DOI: 10.1038/nrmicro1098. URL: http://dx.doi.org/10.1038/nrmicro1098.

[174] Michael R Yeaman and Nannette Y Yount. “Mechanisms of Antimicrobial Peptide Action and Resistance”. In: *Pharmacological Reviews* 55.1 (2003), pp. 27–55. URL: http://pharmrev.aspetjournals.org/content/55/1/27.abstract.

[175] Y. Jerold Gordon, Eric G. Romanowski, and Alison M. McDermott. “A Review of Antimicrobial Peptides and Their Therapeutic Potential as Anti-Infective Drugs”. In: *Current Eye Research* 30.7 (Jan. 2005), pp. 505–515. ISSN: 0271-3683. DOI: 10.1080/02713680590968637. URL: http://dx.doi.org/10.1080/02713680590968637.

[176] Kai Hilpert, Andrea Giuliani, and Andrea C. Rinaldi. “Antimicrobial Peptides”. In: vol. 618. Methods in Molecular Biology. Humana Press, 2010, pp. 125–133. ISBN: 978-1-60761-594-1. URL: http://dx.doi.org/10.1007/978-1-60761-594-1_9.

[177] Kai Hilpert et al. “Peptide-Based Drug Design”. In: vol. 494. Methods in Molecular Biology. Humana Press, 2008, pp. 127–159. ISBN: 978-1-59745-419-3. URL: http://dx.doi.org/10.1007/978-1-59745-419-3_8.

[178] Jianhui Xiao et al. “Efficient Screening of a Novel Antimicrobial Peptide from Jatropha curcas by Cell Membrane Affinity Chromatography”. In: *J. Agric. Food Chem.* 59.4 (Jan. 2011), pp. 1145–1151. ISSN: 0021-8561. DOI: 10.1021/jf103876b. URL: http://dx.doi.org/10.1021/jf103876b.

[179] HÅvard Jenssen et al. “QSAR modeling and computer-aided design of antimicrobial peptides”. In: *Journal of Peptide Science* 14.1 (2008), pp. 110–114. ISSN: 1099-1387. DOI: 10.1002/psc.908. URL: http://dx.doi.org/10.1002/psc.908.

[180] Benjamin A. Hall, Alan P. Chetwynd, and Mark S.P. Sansom. “Exploring Peptide-Membrane Interactions with Coarse-Grained MD Simulations”. In: *Biophysical journal* 100.8 (Apr. 2011), pp. 1940–1948. ISSN: 0006-3495. URL: http://linkinghub.elsevier.com/retrieve/pii/S000634951100261X.

[181] Anders Irbäck and Sandipan Mohanty. “PROFASI: A Monte Carlo simulation package for protein folding and aggregation”. In: *Journal of Computational Chemistry* 27.13 (2006), pp. 1548–1555. ISSN: 1096-987X. DOI: 10.1002/jcc.20452. URL: http://dx.doi.org/10.1002/jcc.20452.

[182] G.N. Ramachandran, C. Ramakrishnan, and V. Sasisekharan. “Stereochemistry of polypeptide chain configurations”. In: *Journal of Molecular Biology* 7.1 (July 1963), pp. 95–99. ISSN: 0022-2836. DOI: 10.1016/S0022-2836(63)80023-6. URL: http://www.sciencedirect.com/science/article/pii/S0022283663800236.

[183] S J Landry et al. “Interplay of structure and disorder in cochaperonin mobile loops”. In: *Proceedings of the National Academy of Sciences* 93.21 (1996), pp. 11622–11627. eprint: http://www.pnas.org/content/93/21/11622.full.pdf+html. URL: http://www.pnas.org/content/93/21/11622.abstract.

[184] Stephen J. Russell et al. “Stability of Cyclic Hairpins: Asymmetric Contributions from Side Chains of a Hydrogen-Bonded Cross-Strand Residue Pair”. In: *Journal of the American Chemical Society* 125.2 (2003). PMID: 12517150, pp. 388–395. DOI: 10.1021/ja028075l. eprint: http://pubs.acs.org/doi/pdf/10.1021/ja028075l. URL: http://pubs.acs.org/doi/abs/10.1021/ja028075l.

[185] Nuria Assa-Munt et al. “Solution Structures and Integrin Binding Activities of an RGD Peptide with Two Isomers†”. In: *Biochemistry* 40.8 (Feb. 2001), pp. 2373–2378. ISSN: 0006-2960. DOI: doi:10.1021/bi002101f. URL: http://dx.doi.org/10.1021/bi002101f.

[186] Monimoy Banerjee et al. “Probing the conformation and dynamics of allatostatin neuropeptides: A structural model for functional differences”. In: *Peptides* 29.3 (2008), pp. 375 –385. ISSN: 0196-9781. DOI: DOI:10.1016/j.peptides.2007.11.016. URL: http://www.sciencedirect.com/science/article/pii/S0196978107004652.

[187] Monali V. Sawai et al. “Impact of single-residue mutations on the structure and function of ovispirin/novispirin antimicrobial peptides”. In: *Protein Engineering* 15.3 (Mar. 2002), pp. 225 –232. URL: http://peds.oxfordjournals.org/content/15/3/225.abstract.

[188] D. Reichmann et al. “The modular architecture of protein–protein binding interfaces”. In: *Proceedings of the National Academy of Sciences of the United States of America* 102.1 (2005), pp. 57–62. URL: http://www.pnas.org/content/102/1/57.abstract.

[189] Takashi Ito et al. “Toward a protein–protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins”. In: *Proceedings of the National Academy of Sciences* 97.3 (2000), pp. 1143–1147. URL: http://www.pnas.org/content/97/3/1143.abstract.

[190] K.S. Midelfort et al. “Substantial energetic improvement with minimal structural perturbation in a high affinity mutant antibody.” In: *J Mol Biol* 343.3 (Oct. 2004). 15465055, pp. 685–701. ISSN: 0022-2836. URL: http://www.hubmed.org/display.cgi?uids=15465055.

[191] Michael M. Mysinger et al. “Structure-based ligand discovery for the protein–protein interface of chemokine receptor CXCR4”. In: *Proceedings of the National Academy of Sciences* (2012). URL: http://www.pnas.org/content/early/2012/03/14/1120431109.abstract.

[192] Bohdan Waszkowycz, David E. Clark, and Emanuela Gancia. “Outstanding challenges in protein–ligand docking and structure-based virtual screening”. In: *Wiley Interdisciplinary Reviews: Computational Molecular Science* 1.2 (2011), pp. 229–259. ISSN: 1759-0884. DOI: 10.1002/wcms.18. URL: http://dx.doi.org/10.1002/wcms.18.

[193] Bernhard Fischer, Kaori Fukuzawa, and Wolfgang Wenzel. “Receptor-specific scoring functions derived from quantum chemical models improve affinity estimates for in-silico drug discovery”. In: *Proteins: Structure, Function, and Bioinformatics* 70.4 (2008), pp. 1264–1273. ISSN: 1097-0134. DOI: 10.1002/prot.21607. URL: http://dx.doi.org/10.1002/prot.21607.

[194] Irene Meliciani et al. “Probing hot spots on protein-protein interfaces with all-atom free-energy simulation”. In: *The Journal of Chemical Physics* 131.3 (July 2009), pp. 034114–11. URL: http://dx.doi.org/10.1063/1.3177008.

[195] James A. Wells and Christopher L. McClendon. “Reaching for high-hanging fruit in drug discovery at protein-protein interfaces.” In: *Nature* 450.7172 (Dec. 2007). 18075579, pp. 1001–1009. ISSN: 1476-4687. URL: http://www.hubmed.org/display.cgi?uids=18075579.

[196] Nelly Andrusier et al. “Principles of flexible protein-protein docking.” In: *Proteins* 73.2 (Nov. 2008). 18655061, pp. 271–289. ISSN: 1097-0134. URL: http://www.hubmed.org/display.cgi?uids=18655061.

[197] Lyubomir T. Vassilev et al. “In vivo activation of the p53 pathway by small-molecule antagonists of MDM2.” In: *Science* 303.5659 (Feb. 2004). 14704432, pp. 844–848. ISSN: 1095-9203. URL: http://www.hubmed.org/display.cgi?uids=14704432.

[198] Bruce L. Grasberger et al. “Discovery and cocrystal structure of benzodiazepinedione HDM2 antagonists that activate p53 in cells.” In: *J Med Chem* 48.4 (Feb. 2005). 15715460, pp. 909–912. ISSN: 0022-2623. URL: http://www.hubmed.org/display.cgi?uids=15715460.

[199] Peter W. White et al. “Inhibition of human papillomavirus DNA replication by small molecule antagonists of the E1-E2 protein interaction.” In: *J Biol Chem* 278.29 (July 2003). 12730224, pp. 26765–26772. ISSN: 0021-9258. URL: http://www.hubmed.org/display.cgi?uids=12730224.

[200] W. Takasaki et al. “Structure-based design and characterization of exocyclic peptidomimetics that inhibit TNF alpha binding to its receptor.” In: *Nat Biotechnol* 15.12 (Nov. 1997). 9359109, pp. 1266–1270. ISSN: 1087-0156. URL: http://www.hubmed.org/display.cgi?uids=9359109.

[201] Hang Yin et al. “Terphenyl-Based Bak BH3 alpha-helical proteomimetics as low-molecular-weight antagonists of Bcl-xL.” In: *J Am Chem Soc* 127.29 (July 2005). 16028929, pp. 10191–10196. ISSN: 0002-7863. URL: http://www.hubmed.org/display.cgi?uids=16028929.

[202] Jack D. Sadowsky et al. “Exploration of backbone space in foldamers containing alpha- and beta-amino acid residues: developing protease-resistant oligomers that bind tightly to the BH3-recognition cleft of Bcl-xL.” In: *Chembiochem* 8.8 (May 2007). 17503422, pp. 903–916. ISSN: 1439-4227. URL: http://www.hubmed.org/display.cgi?uids=17503422.

[203] B.C. Cunningham and J.A. Wells. “Rational design of receptor-specific variants of human growth hormone.” In: *Proc Natl Acad Sci U S A* 88.8 (Apr. 1991). 2014261, pp. 3407–3411. ISSN: 0027-8424. URL: http://www.hubmed.org/display.cgi?uids=2014261.

[204] Jennifer GRODBERG, Kerry L. DAVIS, and Arthur J. SYTKOWSKI. “Alanine scanning mutagenesis of human erythropoietin identifies four amino acids which are critical for biological activity”. In: *European Journal of Biochemistry* 218.2 (1993), pp. 597–601. ISSN: 1432-1033. DOI: 10.1111/j.1432-1033.1993.tb18413.x. URL: http://dx.doi.org/10.1111/j.1432-1033.1993.tb18413.x.

[205] V. Magdolen et al. “Systematic mutational analysis of the receptor-binding region of the human urokinase-type plasminogen activator.” In: *Eur J Biochem* 237.3 (May 1996). 8647121, pp. 743–751. ISSN: 0014-2956. URL: http://www.hubmed.org/display.cgi?uids=8647121.

[206] Shuanghong Huo, Irina Massova, and Peter A. Kollman. “Computational alanine scanning of the 1:1 human growth hormone-receptor complex.” In: *J Comput Chem* 23.1 (Jan. 2002). 11913381, pp. 15–27. ISSN: 0192-8651. URL: http://www.hubmed.org/display.cgi?uids=11913381.

[207] Irina S. Moreira, Pedro A. Fernandes, and Maria J. Ramos. “Unraveling the importance of protein-protein interaction: application of a computational alanine-scanning mutagenesis to the study of the IgG1 streptococcal protein G (C2 fragment) complex.” In: *J Phys Chem B* 110.22 (June 2006). 16771349, pp. 10962–10969. ISSN: 1520-6106. URL: http://www.hubmed.org/display.cgi?uids=16771349.

[208] J.W. Pitera and P.A. Kollman. “Exhaustive mutagenesis in silico: multicoordinate free energy calculations on proteins and peptides.” In: *Proteins* 41.3 (Nov. 2000). 11025549, pp. 385–397. ISSN: 0887-3585. URL: http://www.hubmed.org/display.cgi?uids=11025549.

[209] J.C. Burnett et al. “Computationally accessible method for estimating free energy changes resulting from site-specific mutations of biomolecules: systematic model building and structural/hydropathic analysis of deoxy and oxy hemoglobins.” In: *Proteins* 42.3 (Feb. 2001). 11151007, pp. 355–377. ISSN: 0887-3585. URL: http://www.hubmed.org/display.cgi?uids=11151007.

[210] Qizhi Cui et al. “Molecular Dynamics-Solvated Interaction Energy Studies of Protein-Protein Interactions: The MP1-p14 Scaffolding Complex”. In: *Journal of Molecular Biology* 379.4 (June 2008), pp. 787–802. ISSN: 0022-2836. DOI: 10.1016/j.jmb.2008.04.035. URL: http://www.sciencedirect.com/science/article/pii/S0022283608004671.

[211] Irina Massova and Peter A. Kollman. “Computational Alanine Scanning To Probe Protein-Protein Interactions: A Novel Approach To Evaluate Binding Free Energies”. In: *J. Am. Chem. Soc.* 121.36 (1999), pp. 8133–8143. ISSN: 0002-7863. DOI: 10.1021/ja990935j. URL: http://dx.doi.org/10.1021/ja990935j.

[212] N.J. Skelton et al. “Structure of a CXC chemokine-receptor fragment in complex with interleukin-8.” In: *Structure* 7.2 (Feb. 1999). 10368283, pp. 157–168. ISSN: 0969-2126. URL: http://www.hubmed.org/display.cgi?uids=10368283.

[213] Timothy N.C. Wells et al. “Chemokine blockers–therapeutics in the making?” In: *Trends Pharmacol Sci* 27.1 (Jan. 2006). 16310864, pp. 41–47. ISSN: 0165-6147. URL: http://www.hubmed.org/display.cgi?uids=16310864.

[214] Samantha J. Allen, Susan E. Crown, and Tracy M. Handel. “Chemokine: receptor structure, interactions, and antagonism.” In: *Annu Rev Immunol* 25 (2007). 17291188, pp. 787–820. ISSN: 0732-0582. URL: http://www.hubmed.org/display.cgi?uids=17291188.

[215] Christian Dillon et al. “Basolateral targeting of ERBB2 is dependent on a novel bipartite juxtamembrane sorting signal but independent of the C-terminal ERBIN-binding domain.” In: *Mol Cell Biol* 22.18 (2002). 12192053, pp. 6553–6563. ISSN: 0270-7306. URL: http://www.hubmed.org/display.cgi?uids=12192053.

[216] Fanny Jaulin-Bastard et al. “The ERBB2/HER2 Receptor Differentially Interacts with ERBIN and PICK1 PSD-95/DLG/ZO-1 Domain Proteins”. In: *Journal of Biological Chemistry* 276.18 (2001), pp. 15256–15263. URL: http://www.jbc.org/content/276/18/15256.abstract.

[217] Ying-Xin Fan, Lily Wong, and Gibbes R. Johnson. “EGFR kinase possesses a broad specificity for ErbB phosphorylation sites, and ligand increases catalytic-centre activity without affecting substrate binding affinity.” In: *Biochem J* 392.Pt 3 (Dec. 2005). 16122376, pp. 417–423. ISSN: 1470-8728. URL: http://www.hubmed.org/display.cgi?uids=16122376.

[218] Paul N. Mortenson and David J. Wales. “Energy landscapes, global optimization and dynamics of the polyalanine Ac(ala)[sub 8]NHMe”. In: *The Journal of Chemical Physics* 114.14 (Apr. 2001), pp. 6443–6454. URL: http://dx.doi.org/10.1063/1.1343486.

[219] Joanne M. Carr and David J. Wales. “Global optimization and folding pathways of selected alpha-helical proteins.” In: *J Chem Phys* 123.23 (Dec. 2005). 16392943, p. 234901. ISSN: 0021-9606. URL: http://www.hubmed.org/display.cgi?uids=16392943.

[220] T. Herges, A. Schug, and W. Wenzel. “Exploration of the free-energy surface of a three-helix peptide with stochastic optimization methods”. In: *International Journal of Quantum Chemistry* 99.5 (2004), pp. 854–863. ISSN: 1097-461X. DOI: 10.1002/qua.20052. URL: http://dx.doi.org/10.1002/qua.20052.

[221] David P. Anderson. “Boinc: A system for public-resource computing and storage”. In: *5th IEEE/ACM International Workshop on Grid Computing*. 2004, 4–10.

[222] M. Baggiolini, B. Dewald, and B. Moser. “Human chemokines: an update.” In: *Annu Rev Immunol* 15 (1997). 9143704, pp. 675–705. ISSN: 0732-0582. URL: http://www.hubmed.org/display.cgi?uids=9143704.

[223] M. Baggiolini and B. Moser. “Blocking chemokine receptors.” In: *J Exp Med* 186.8 (Oct. 1997). 9379143, pp. 1189–1191. ISSN: 0022-1007. URL: http://www.hubmed.org/display.cgi?uids=9379143.

[224] J.P. Borg et al. “ERBIN: a basolateral PDZ protein that interacts with the mammalian ERBB2/HER2 receptor.” In: *Nat Cell Biol* 2.7 (July 2000). 10878805, pp. 407–414. ISSN: 1465-7392. URL: http://www.hubmed.org/display.cgi?uids=10878805.

[225] Gabriel Birrane, Judy Chung, and John A.A. Ladias. “Novel mode of ligand recognition by the Erbin PDZ domain.” In: *J Biol Chem* 278.3 (Jan. 2003). 12444095, pp. 1399–1402. ISSN: 0021-9258. URL: http://www.hubmed.org/display.cgi?uids=12444095.

[226] K.S. Thorn and A.A. Bogan. “ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions.” In: *Bioinformatics* 17.3 (Mar. 2001). 11294795, pp. 284–285. ISSN: 1367-4803. URL: http://www.hubmed.org/display.cgi?uids=11294795.

[227] Raphael Guerois, Jens Erik Nielsen, and Luis Serrano. “Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations.” In: *J Mol Biol* 320.2 (July 2002). 12079393, pp. 369–387. ISSN: 0022-2836. URL: http://www.hubmed.org/display.cgi?uids=12079393.

[228] Tanja Kortemme and David Baker. “A simple physical model for binding energy hot spots in protein-protein complexes.” In: *Proc Natl Acad Sci U S A* 99.22 (Oct. 2002). 12381794, pp. 14116–14121. ISSN: 0027-8424. URL: http://www.hubmed.org/display.cgi?uids=12381794.

[229] Steven J. Darnell, David Page, and Julie C. Mitchell. “An automated decision-tree approach to predicting protein interaction hot spots.” In: *Proteins* 68.4 (2007). 17554779, pp. 813–823. ISSN: 1097-0134. URL: http://www.hubmed.org/display.cgi?uids=17554779.

[230] Yanay Ofran, Venkatesh Mysore, and Burkhard Rost. “Prediction of DNA-binding residues from sequence.” In: *Bioinformatics* 23.13 (July 2007). 17646316, pp. 347–353. ISSN: 1367-4811. URL: http://www.hubmed.org/display.cgi?uids=17646316.

[231] Buyong Ma et al. “Protein-protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces.” In: *Proc Natl Acad Sci U S A* 100.10 (May 2003). 12730379, pp. 5772–5777. ISSN: 0027-8424. URL: http://www.hubmed.org/display.cgi?uids=12730379.

[232] Ozlem Keskin, Buyong Ma, and Ruth Nussinov. “Hot regions in protein–protein interactions: the organization and contribution of structurally conserved hot spot residues.” In: *J Mol Biol* 345.5 (Feb. 2005). 15644221, pp. 1281–1294. ISSN: 0022-2836. URL: http://www.hubmed.org/display.cgi?uids=15644221.

[233] A.A. Bogan and K.S. Thorn. “Anatomy of hot spots in protein interfaces.” In: *J Mol Biol* 280.1 (July 1998). 9653027, pp. 1–9. ISSN: 0022-2836. URL: http://www.hubmed.org/display.cgi?uids=9653027.

[234] Z. Hu et al. “Conservation of polar residues as hot spots at protein interfaces.” In: *Proteins* 39.4 (June 2000). 10813815, pp. 331–342. ISSN: 0887-3585. URL: http://www.hubmed.org/display.cgi?uids=10813815.

[235] Steven J. Darnell, Laura LeGault, and Julie C. Mitchell. “KFC Server: interactive forecasting of protein interaction hot spots.” In: *Nucleic Acids Res* 36.Web Server issue (July 2008). 18539611, pp. 265–269. ISSN: 1362-4962. URL: http://www.hubmed.org/display.cgi?uids=18539611.

[236] Osman N. Yogurtcu et al. “Restricted mobility of conserved residues in protein-protein interfaces in molecular simulations.” In: *Biophys J* 94.9 (May 2008). 18227135, pp. 3475–3485. ISSN: 1542-0086. URL: http://www.hubmed.org/display.cgi?uids=18227135.

[237] Yovani Marrero-Ponce et al. “Protein linear indices of the ’macromolecular pseudograph alpha-carbon atom adjacency matrix’ in bioinformatics. Part 1: prediction of protein stability effects of a complete set of alanine substitutions in Arc repressor.” In: *Bioorg Med Chem* 13.8 (Apr. 2005). 15781410, pp. 3003–3015. ISSN: 0968-0896. URL: http://www.hubmed.org/display.cgi?uids=15781410.

[238] Holger Gohlke, Christina Kiel, and David A. Case. “Insights into protein-protein binding by binding free energy calculation and free energy decomposition for the Ras-Raf and Ras-RalGDS complexes.” In: *J Mol Biol* 330.4 (July 2003). 12850155, pp. 891–913. ISSN: 0022-2836. URL: http://www.hubmed.org/display.cgi?uids=12850155.

[239] Alexander Benedix et al. “Predicting free energy changes using structural ensembles.” In: *Nat Methods* 6.1 (Jan. 2009). 19116609, pp. 3–4. ISSN: 1548-7105. URL: http://www.hubmed.org/display.cgi?uids=19116609.

[240] Deepa Rajamani et al. “Anchor residues in protein-protein interactions.” In: *Proc Natl Acad Sci U S A* 101.31 (2004). 15269345, pp. 11287–11292. ISSN: 0027-8424. URL: http://www.hubmed.org/display.cgi?uids=15269345.

[241] K.T. Simons et al. “Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions.” In: *J Mol Biol* 268.1 (Apr. 1997). 9149153, pp. 209–225. ISSN: 0022-2836. URL: http://www.hubmed.org/display.cgi?uids=9149153.

[242] Joost Schymkowitz et al. “The FoldX web server: an online force field.” In: *Nucleic Acids Res* 33.Web Server issue (July 2005). 15980494, pp. 382–388. ISSN: 1362-4962. URL: http://www.hubmed.org/display.cgi?uids=15980494.

[243] Johannes Flick, Frank Tristram, and Wolfgang Wenzel. “Modeling loop backbone flexibility in receptor-ligand docking simulations”. In: *Journal of Computational Chemistry* (2012), n/a–n/a. ISSN: 1096-987X. DOI: 10.1002/jcc.23087. URL: http://dx.doi.org/10.1002/jcc.23087.

[244] David S. Goodsell and Arthur J. Olson. “Automated docking of substrates to proteins by simulated annealing”. In: *Proteins: Structure, Function, and Bioinformatics* 8.3 (1990), pp. 195–202. ISSN: 1097-0134. DOI: 10.1002/prot.340080302. URL: http://dx.doi.org/10.1002/prot.340080302.

[245] Alexander Biewer. “Untersuchung der Rezeptor-Liganden Bindung an einem Homologiemodell des Aryl-hydrocarbon-Rezeptors”. PhD thesis. Universität Karlsruhe, 2012.

[246] P. C. Weber et al. “Structural origins of high-affinity biotin binding to streptavidin”. In: *Science* 243.4887 (1989), pp. 85–88. URL: http://www.sciencemag.org/content/243/4887/85.abstract.

[247] Aline S. Borer et al. “Crystal Structure of Sol i 2: A Major Allergen from Fire Ant Venom”. In: *Journal of Molecular Biology* 415.4 (Jan. 2012), pp. 635–648. ISSN: 0022-2836. DOI: 10.1016/j.jmb.2011.10.009. URL: http://www.sciencedirect.com/science/article/pii/S0022283611011338.

[248] Oliver F. Lange et al. “Recognition Dynamics Up to Microseconds Revealed from an RDC-Derived Ubiquitin Ensemble in Solution”. In: *Science* 320.5882 (2008), pp. 1471–1475. URL: http://www.sciencemag.org/content/320/5882/1471.abstract.

[249] Tomasz Wlodarski and Bojan Zagrovic. “Conformational selection and induced fit mechanism underlie specificity in noncovalent interactions with ubiquitin”. In: *Proceedings of the National Academy of Sciences* 106.46 (2009), pp. 19346–19351. URL: http://www.pnas.org/content/106/46/19346.abstract.

[250] R. Car and M. Parrinello. “Unified Approach for Molecular Dynamics and Density-Functional Theory”. In: *Phys. Rev. Lett.* 55.22 (Nov. 1985), 2471–2474. DOI: 10.1103/PhysRevLett.55.2471. URL: http://link.aps.org/doi/10.1103/PhysRevLett.55.2471.

[251] Elizabeth M. Lupton and Irmgard Frank. *Probing the mechanical strength of chemical bonds by stretching single molecules.* English. Published: Wagner, Siegfried (ed.) et al., High performance computing in science and engineering, Garching/Munich 2007. Transactions of the third joint HLRB and KONWIHR status and result workshop, December 3–4, 2007, Leibniz Supercomputing Centre, Garching/Munich, Germany. Berlin: Springer. 165-172 (2008). 2008.

[252] Martin K. Beyer. “The mechanical strength of a covalent bond calculated by density functional theory”. In: *The Journal of Chemical Physics* 112.17 (2000), pp. 7307–7312. URL: http://dx.doi.org/10.1063/1.481330.

[253] James B. Robinson and Peter J. Knowles. “Breaking multiple covalent bonds with Hartree-Fock-based quantum chemistry: Quasi-Variational Coupled Cluster theory with perturbative treatment of triple excitations”. In: *Physical Chemistry Chemical Physics* 14.19 (2012), pp. 6729–6732. ISSN: 1463-9076. URL: http://dx.doi.org/10.1039/C2CP40698E.

[254] Wenbing Hu et al. “Polymer crystallization under nano-confinement of droplets studied by molecular simulations”. In: *Faraday Discussions* 143 (2009), pp. 129–141. ISSN: 1359-6640. URL: http://dx.doi.org/10.1039/B901378D.

[255] Y. Q. Cheng, E. Ma, and H. W. Sheng. “Atomic Level Structure in Multicomponent Bulk Metallic Glass”. In: *Phys. Rev. Lett.* 102.24 (June 2009), p. 245501. DOI: 10.1103/PhysRevLett.102.245501. URL: http://link.aps.org/doi/10.1103/PhysRevLett.102.245501.

[256] J. Richardi, M. P. Pileni, and J.-J. Weis. “Self-organization of magnetic nanoparticles: A Monte Carlo study”. In: *Phys. Rev. E* 77.6 (June 2008), p. 061510. DOI: 10.1103/PhysRevE.77.061510. URL: http://link.aps.org/doi/10.1103/PhysRevE.77.061510.

[257] Wolfgang Brütting, Stefan Berleb, and Anton G. Mückl. “Device physics of organic light-emitting diodes based on molecular materials”. In: *Organic Electronics* 2.1 (Mar. 2001), pp. 1–36. ISSN: 1566-1199. DOI: 10.1016/S1566-1199(01)00009-X. URL: http://www.sciencedirect.com/science/article/pii/S156611990100009X.

[258] Andreas Fuchs et al. “Molecular origin of differences in hole and electron mobility in amorphous Alq3-a multiscale simulation study”. In: *Phys. Chem. Chem. Phys.* 14.12 (2012), pp. 4259–4270. ISSN: 1463-9076. URL: http://dx.doi.org/10.1039/C2CP23489K.

[259] Michele Muccini. “A bright future for organic field-effect transistors”. In: *Nat Mater* 5.8 (Aug. 2006), pp. 605–613. ISSN: 1476-1122. DOI: 10.1038/nmat1699. URL: http://dx.doi.org/10.1038/nmat1699.

[260] Andrzej Dzwilewski, Piotr Matyba, and Ludvig Edman. “Facile Fabrication of Efficient Organic CMOS Circuits”. In: *J. Phys. Chem. B* 114.1 (Nov. 2009), pp. 135–140. ISSN: 1520-6106. DOI: 10.1021/jp909216a. URL: http://dx.doi.org/10.1021/jp909216a.

[261] Francois Leonard. “The Physics of Carbon Nanotube Devices - Introduction”. In: *The Physics of Carbon Nanotube Devices*. Norwich, NY: William Andrew Publishing, 2009, pp. 1–26. ISBN: 978-0-8155-1573-9. URL: http://www.sciencedirect.com/science/article/pii/B9780815515739500045.

[262] Charles Kittel. *Einführung in die Festkörperphysik*. Oldenbourg. ISBN: 3486577239. URL: http://www.amazon.de/exec/obidos/redirect?tag=citeulike01-21&path=ASIN/3486577239.

[263] Mason A. Wolak et al. “Functionalized Pentacene Derivatives for Use as Red Emitters in Organic Light-Emitting Diodes”. In: *J. Phys. Chem. B* 108.18 (Apr. 2004), pp. 5492–5499. ISSN: 1520-6106. DOI: 10.1021/jp036199u. URL: http://dx.doi.org/10.1021/jp036199u.

[264] N Karl. “Charge carrier transport in organic semiconductors”. In: *Proceedings of the Yamada Conference LVI. The Fourth International Symposium on Crystalline Organic Metals, Superconductors and Ferromagnets (ISCOM 2001).* 133–134.0 (Mar. 2003), pp. 649–657. ISSN: 0379-6779. DOI: 10.1016/S0379-6779(02)00398-3. URL: http://www.sciencedirect.com/science/article/pii/S0379677902003983.

[265] Daniel Holmes et al. “On the Nature of Nonplanarity in the [N]Phenylenes”. In: *Chemistry – A European Journal* 5.11 (1999), pp. 3399–3412. ISSN: 1521-3765. DOI: 10.1002/(SICI)1521-3765(19991105)5:11<3399::AID-CHEM3399>3.0.CO;2-V. URL: http://dx.doi.org/10.1002/(SICI)1521-3765(19991105)5:11<3399::AID-CHEM3399>3.0.CO;2-V.

[266] Christine C. Mattheus et al. “Polymorphism in pentacene”. In: *Acta Crystallographica Section C* 57.8 (2001), pp. 939–941. URL: http://dx.doi.org/10.1107/S010827010100703X.

[267] Sandra E. Fritz et al. “Structural Characterization of a Pentacene Monolayer on an Amorphous SiO2 Substrate with Grazing Incidence X-ray Diffraction”. In: *J. Am. Chem. Soc.* 126.13 (Mar. 2004), pp. 4084–4085. ISSN: 0002-7863. DOI: 10.1021/ja049726b. URL: http://dx.doi.org/10.1021/ja049726b.

[268] Paul-Jakob Kleine. “Morphologie-Simulationen von SMOLED-Materialien mit Monte-Carlo”. PhD thesis. Universität Karlsruhe, 2011.

[269] Benedikt Matthias Schönauer. “Implementierung eines Cluster-Move-Algorithmus in SIMONA - in preparation”. PhD thesis. Karlsruhe Institute of Technology, 2012.

[270] R. B. Campbell, J. M. Robertson, and J. Trotter. “The crystal and molecular structure of pentacene”. In: *Acta Crystallographica* 14.7 (1961), pp. 705–711. URL: http://dx.doi.org/10.1107/S0365110X61002163.

[271] JJP Stewart. “MOSOL, MOPAC for solid-state physics”. In: 5 (1985), pp. 62–63.

[272] S. W. Davis et al. “Cluster-based Monte Carlo simulation of ferrofluids”. In: *Phys. Rev. E* 59.2 (Feb. 1999), 2424–2428. DOI: 10.1103/PhysRevE.59.2424. URL: http://link.aps.org/doi/10.1103/PhysRevE.59.2424.

[273] N. G. Almarza and E. Lomba. “Cluster algorithm to perform parallel Monte Carlo simulation of atomistic systems”. In: *The Journal of Chemical Physics* 127.8 (Aug. 2007), pp. 084116–6. URL: http://dx.doi.org/10.1063/1.2759924.

[274] A. Bhattacharyay and Alessandro Troisi. “Self-assembly of sparsely distributed molecules: An efficient cluster algorithm”. In: *Chemical Physics Letters* 458.1–3 (June 2008), pp. 210–213. ISSN: 0009-2614. DOI: 10.1016/j.cplett.2008.04.052. URL: http://www.sciencedirect.com/science/article/pii/S0009261408005010.

[275] Stephen Whitelam and Phillip L. Geissler. “Avoiding unphysical kinetic traps in Monte Carlo simulations of strongly attractive particles”. In: *The Journal of Chemical Physics* 127.15 (Oct. 2007), pp. 154101–19. URL: http://dx.doi.org/10.1063/1.2790421.

[276] P. M. Ajayan. “Nanotubes from Carbon”. In: *Chem. Rev.* 99.7 (1999), pp. 1787–1800. ISSN: 0009-2665. DOI: 10.1021/cr970102g. URL: http://dx.doi.org/10.1021/cr970102g.

[277] Justin Opatkiewicz, Melburne C. LeMieux, and Zhenan Bao. “Nanotubes on Display: How Carbon Nanotubes Can Be Integrated into Electronic Displays”. In: *ACS Nano* 4.6 (2010), pp. 2975–2978. ISSN: 1936-0851. DOI: 10.1021/nn101092d. URL: http://dx.doi.org/10.1021/nn101092d.

[278] Holger F. Bettinger. “Carbon Nanotubes-Basic Concepts and Physical Properties. By S. Reich, C. Thomsen, J. Maultzsch.” In: *ChemPhysChem* 5.12 (2004), pp. 1914–1915. ISSN: 1439-7641. DOI: 10.1002/cphc.200400387. URL: http://dx.doi.org/10.1002/cphc.200400387.

[279] David Mann. “Synthesis of carbon nanotubes”. In: *Carbon Nanotubes*. 0. CRC Press, May 2006, pp. 19–49. ISBN: 978-0-8493-2748-3. URL: http://dx.doi.org/10.1201/9781420004212.ch2.

[280] Richard Martel. “Sorting Carbon Nanotubes for Electronics”. In: *ACS Nano* 2.11 (2008), pp. 2195–2199. ISSN: 1936-0851. DOI: 10.1021/nn800723u. URL: http://dx.doi.org/10.1021/nn800723u.

[281] Adrian Nish et al. “Highly selective dispersion of single-walled carbon nanotubes using aromatic polymers”. In: *Nat Nano* 2.10 (Oct. 2007), pp. 640–646. ISSN: 1748-3387. DOI: 10.1038/nnano.2007.290. URL: http://dx.doi.org/10.1038/nnano.2007.290.

[282] Jeong-Yuan Hwang et al. “Polymer Structure and Solvent Effects on the Selective Dispersion of Single-Walled Carbon Nanotubes”. In: *J. Am. Chem. Soc.* 130.11 (2008), pp. 3543–3553. ISSN: 0002-7863. DOI: 10.1021/ja0777640. URL: http://dx.doi.org/10.1021/ja0777640.

[283] Fabien A. Lemasson et al. “Selective Dispersion of Single-Walled Carbon Nanotubes with Specific Chiral Indices by Poly(N-decyl-2,7-carbazole)”. In: *J. Am. Chem. Soc.* 133.4 (Dec. 2010), pp. 652–655. ISSN: 0002-7863. DOI: 10.1021/ja105722u. URL: http://dx.doi.org/10.1021/ja105722u.

[284] Fuming Chen et al. “Toward the Extraction of Single Species of Single-Walled Carbon Nanotubes Using Fluorene-Based Polymers”. In: *Nano Lett.* 7.10 (2007), pp. 3013–3017. ISSN: 1530-6984. DOI: 10.1021/nl071349o. URL: http://dx.doi.org/10.1021/nl071349o.

[285] Fuyong Cheng et al. “Soluble, Discrete Supramolecular Complexes of Single-Walled Carbon Nanotubes with Fluorene-Based Conjugated Polymers”. In: *Macromolecules* 41.7 (2008), pp. 2304–2308. ISSN: 0024-9297. DOI: 10.1021/ma702567y. URL: http://dx.doi.org/10.1021/ma702567y.

[286] Valerie C. Moore et al. “Individually Suspended Single-Walled Carbon Nanotubes in Various Surfactants”. In: *Nano Lett.* 3.10 (2003), pp. 1379–1382. ISSN: 1530-6984. DOI: 10.1021/nl034524j. URL: http://dx.doi.org/10.1021/nl034524j.

[287] Tetyana V. Bogdan, David J. Wales, and Florent Calvo. “Equilibrium thermodynamics from basin-sampling”. In: *The Journal of Chemical Physics* 124.4 (Jan. 2006), pp. 044102–13. URL: http://dx.doi.org/10.1063/1.2148958.

[288] Charusita Chakravarty et al. “Effects of three-body (Axilrod-Teller) forces on the classical and quantum behavior of rare-gas trimers”. In: *Physical Review E* 56.1 (July 1997), pp. 363–377. URL: http://link.aps.org/doi/10.1103/PhysRevE.56.363.

[289] Alexander Schug et al. “Comparison of Stochastic Optimization Methods for All-Atom Folding of the Trp-Cage Protein”. In: *ChemPhysChem* 6.12 (2005), pp. 2640–2646. ISSN: 1439-7641. DOI: 10.1002/cphc.200500213. URL: http://dx.doi.org/10.1002/cphc.200500213.

Acknowledgements

First and foremost I would like to thank the Carl-Zeiss Stiftung for making the work in this thesis possible by funding my project over the course of three years.
I would like to convey my sincere thanks to Prof. Dr. Wolfgang Wenzel for guiding me throughout the course of my thesis and allowing me to present my research on countless national and international conferences and initiating most of the experimental collaborations presented as part of this thesis. I also thank Prof. Dr. Ulrich Nienhaus and his group for interesting discussions and of course for refereeing my thesis.
I thank all the co-authors of my papers for the fruitful collaborations and the opportunity to present the work as part of my thesis: AG Richter (UIC Heidelberg), AG Fischer (KIT) and AG Schimmel (KIT) for their work on the genetically engineered hydrophobin, AG Hamacher and AG Pfeifer (Darmstadt University) for their work on gas-vesicle formation, AG Schmitz (KIT) for their work on computational alanine screening and AG Mayor (KIT) and AG Kappes (KIT) for their work on nanotube wrapping. Of these groups I especially thank Prof. Dr. Kay Hamacher for his support and motivation on the gas-vesicle formation project.
All of this work would not have been possible without the support of my group: Moritz Wolf not only supported me with coffee infusions every morning, but also helped me debug countless lines of code. We are a great team and could solve most problems with ease. SIMONA is the program it is today, because of him. The good working atmosphere in the group helped to efficiently tackle most problems. I especially thank Julia Setzler and Nana Heilmann for their support, while finalizing my thesis and Frank Tristram, Dr. Konstantin Klenin, Priya Anand, Simon Widmaier, Angela Poschlad, Dr. Velimir Meded, Carolin Seith, Matthias Ernst and Tobias Neumann, who supported me in many ways during the preparation of my thesis. Of course, also the many previous group members supported me with guidance and their insight gained during their work. Therefore my gratitude goes to Dr. Robert Maul, Dr. Horacio Pérez-Sánchez, Dr. Irene Meliciani and Alexander Biewer. I would also like to thank the whole group of Dr. Alexander Schug for the good collaboration between our groups.
The effort put into the development of SIMONA by all the developers cannot be quantified. I would especially like to thank Benedikt Schönauer, Felix Ehrler and Paul-Jakob Kleine for the contributions they made during their bachelor theses.
I wish to thank my entire family for their support during the writing of this thesis. They supported me in all my decisions and understood the time constraints this thesis was created under. Finally I would like to thank my girlfriend Dr. Marina Albrizio for her continuous encouragement and patience.

Publications

[1] Probing hot spots on protein-protein interfaces with all-atom free-energy simulation
Irene Meliciani, Konstantin Klenin, Timo Strunk, Katja Schmitz and Wolfgang Wenzel
J. Chem. Phys. 131, 034114 (2009)
Awarded with the 2009 JCP Editor's choice award

[2] Selective dispersion of single walled carbon nanotubes with specific chiral indices by poly(N-decyl-2,7-carbazole)
Fabien Lemasson, Timo Strunk, Peter Gerstel, Frank Hennrich, Sergei Lebedkin, Christopher Barner-Kowollik, Wolfgang Wenzel, Manfred Kappes, Marcel Mayor
J. Am. Chem. Soc. 133, no. 4 (December 20, 2010): 652–655.

[3] Structural Model of the Gas Vesicle Protein GvpA and Analysis of GvpA Mutants in vivo,
Timo Strunk, Kay Hamacher, Franziska Hoffgaard, Harald Engelhardt, Martina Zillig, Karin Faist, Wolfgang Wenzel and Felicitas Pfeifer
Molecular Microbiology 81, no. 1 (2011): 56-68.

[4] Derivatives of molecular surface area and volume: Simple and exact analytical formulas
Konstantin V. Klenin, Frank Tristram, Timo Strunk, Wolfgang Wenzel
Journal of Computational Chemistry 32, no. 12 (2011): 2647-2653.

[5] Engineering of the hydrophobin DewA to generate surfaces enhancing cell adhesion
Stephane Boeuf, Tanja Throm, Timo Strunk, Leonie Mühlberg, Marc Hoffmann, Wolfgang Wenzel, Reinhard Fischer, Wiltrud Richter
Acta Biomaterialia 8, no. 3 (March 2012): 1037–1047.

[6] Peptide structure prediction using distributed volunteer computing networks
Timo Strunk, Moritz Wolf, Wolfgang Wenzel
Journal of Mathematical Chemistry 50, no. 2 (February 1, 2012): 421–428.

[7] SIMONA 1.0: An efficient and versatile framework for stochastic simulations of molecular and nanoscale systems
Timo Strunk, Moritz Wolf, Martin Brieg, Konstantin Klenin, Alexander Biewer, Frank Tristram, Matthias Ernst, Paul-Jakob Kleine, Nana Heilmann, Ivan Kondov, Wolfgang Wenzel
Journal of Computational Chemistry (2012): preprint

Proceedings

[1] Free-energy based all-atom protein folding using worldwide distributed computational resources. *Timo Strunk, Srinivasa M. Gopal, Irene Meliciani, Konstantin Klenin, Wolfgang Wenzel*

From Computational Biophysics to Systems Biology (CBSB08) : Proc. of NIC Symp. 2008, Jülich, May 19-21, 2008, John von Neumann Institute for Computing, 2008 S.381-84

[2] Development and Evaluation of a GPU-optimized N-Body term for the simulation of biomolecules

Timo Strunk, Moritz Wolf, Wolfgang Wenzel

Proceedings of SimLab@KIT Workshop, Karlsruhe KIT Campus South, November 29/ 30, 2010

[3] High Throughput peptide structure prediction with distributed volunteer computing networks

Timo Strunk, Wolfgang Wenzel

CMMSE 2011 : Proceedings of the 11th International Conference on Mathematical Methods in Science and Engineering, Alicante 26-30 June 2011: ISBN: 978-84-614-6167-7

[4] Benchmarking the POEM@HOME Network for Protein Structure Prediction

Timo Strunk, Priya Anand, Martin Brieg, Moritz Wolf, Konstantin Klenin, Irene Meliciani, Frank Tristram, Ivan Kondov, Wolfgang Wenzel

IWSG life 2011, 08.-10.06.11 Westminster UK

Invited Talks

[1]Rationale Optimierung von Zelladhäsion auf nanoskalig strukturierten Hydrophobin-Oberflächen

Timo Strunk, Wolfgang Wenzel

Forschungstag der Landesstiftung Baden-Württemberg, Landesstiftung Baden-Württemberg, Heidelberg 29.06.11

Talks

[2]Free-energy based all-atom protein folding using worldwide distributed computational resources. *Timo Strunk, Srinivasa M. Gopal, Irene Meliciani, Konstantin Klenin, Wolfgang Wenzel*

Integration of Nanotechnology-Related Life Science Research in Europe: Frontiers Bio-Nano Winterschool 2008, Zermatt, CH, March 9-13, 2008

[3]Rational optimisation of cell adhesion, differentiation and growth on nanoscale structured hydrophobin surfaces.

Timo Strunk, Wolfgang Wenzel

Doktorandenkolloquium der Landesstiftung Baden-Württemberg, Sonnenbühl, 10.-11.09. 2009

[4]High-Throughput Proteinstrukturvorhersage auf hybriden und verteilten Höchstleistungs-Architekturen

Timo Strunk, Ivan Kondov, Konstantin Klenin, Wolfgang Wenzel

High Performance Computing in Science and Engineering, Stuttgart, 27.Januar 2010

[5]Development and Evaluation of a GPU-optimized N-Body term for the simulation of biomolecules

Timo Strunk, Moritz Wolf, Wolfgang Wenzel

Proceedings of SimLab@KIT Workshop, Karlsruhe KIT Campus South, November 29/ 30, 2010

[6]Analysis of amino-acid specific energy contributions to native conformations in high-resolution protein structures

Timo Strunk, Moritz Wolf, Wolfgang Wenzel

COMPUTER SIMULATION AND THEORY OF MACROMOLECULES 2011, Hünfeld, April 15-16, 2011

[7]High Throughput peptide structure prediction with distributed volunteer computing networks

Timo Strunk, Wolfgang Wenzel

CMMSE 2011 : 11th International Conference on Mathematical Methods in Science and Engineering, Alicante 26-30 June 2011