Protein Structure Refinement with Adaptable Restraints

Inaugural dissertation

for the attainment of the title of doctor
in the Faculty of Mathematics and Natural Sciences
at the Heinrich Heine University Düsseldorf

presented by

André Wildberg
from Essen

Jülich, June 20, 2013

from the institute ICS-6
at the Forschungszentrum Jülich

Published by permission of the
Faculty of Mathematics and Natural Sciences at
Heinrich Heine University Düsseldorf

Supervisior: Jun. Prof. Dr. Gunnar Schröder
Co-supervisor: Prof. Dr. Dieter Willbold

Date of the oral examination: 13.05.2013

Declaration of Authorship

I, André Wildberg, declare that this thesis and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

mod·el/[mod-l]/*noun*

a simplified description, especially a mathematical one, of a system or process, to assist calculations and predictions.

cha·os/[keias]/*noun*

the property of a complex system whose behavior is so unpredictable as to appear random, owing great sensitivity to small changes in conditions.

oxford dictionary

Protein Structure Refinement with Adaptable Restraints

Abstract

Proteins and especially their structures play an essential role in computational structural biology. With accurate characterizations of proteins, processes like drug design and docking simulations can achieve high resolution results. Protein structures can be derived experimentally and computationally. Structures solved experimentally with X-ray crystallography are among the most detailed descriptions of proteins we can generate today. But since this is a complex process involving a lot of labor and time, structures are also solved computationally with empirical data. Here, the most successful method is homology modeling. Models produced with this method, though often very accurate, can carry severe structural errors. We focus on correcting these errors in a reliable, stable way, by refining the spacial arrangement of atoms.

The more detailed a model built with homology modeling methods actually is, the more the task of refining means to hold the correctly placed atoms and carefully introduce corrections just where needed. While starting models deviate more and more, confident refinement means allowing enough freedom to move atoms but still holding a substantial part of the protein in place.

Successful refinement approaches solve the task of knowing when to just hold an accurate model and when to move atoms more substantially automatically.

We achieve exactly that - holding the overall shape of a protein and still allowing needed movement of atoms - by introducing a novel deformable elastic network (DEN) inspired adaptive deformable position restraint (ADPt), either internally applied on a single copy approach or by introducing the use of coupled evolutionary related sequence data.

Proteinstrukturverfeinerung mit adaptierbaren Restraints

Zusammenfassung

Proteine und vor allem ihre Strukturen spielen eine wesentliche Rolle in der computergestützten Strukturbiologie. Mit genauen Beschreibungen von Proteinen können Prozesse wie Drug Design und Docking Simulationen eine hohe Auflösung erreichen.

Proteinstrukturen können experimentell oder, mit Hilfe von Computermodellen, empirisch abgeleitet werden. Strukturen, die experimentell mit Röntgenkristallographie erstellt wurden gehören zu den detailliertesten Beschreibungen von Proteinen, die heutzutage erzeugt werden können. Da dies aber ein komplexer Prozess ist, der generell viel Arbeit und Zeit benötigt, werden Strukturen auch empirisch gelöst, wobei hierbei die erfolgreichste Methode die Homologiemodellierung ist. Modelle, die mit dieser Methode erstellt wurden sind normalerweise sehr akkurat, können aber auch schwere strukturelle Fehler tragen. Wir konzentrieren uns auf die Korrektur dieser Fehler, bei der wir eine Verfeinerung der räumlichen Anordnung der Atome durchführen.

Je detaillierter ein Modell ist, desto mehr bedeutet Verfeinerung, die richtig platzierten Atome zu halten und Korrekturen nur dort einzubringen, wo sie wirklich nötig sind. Weichen die Startmodelle mehr und mehr ab, bedeutet zuverlässige Verfeinerung, den Atomen genug Optimierungsfreiheiten zu lassen, die Gesamtstruktur im Wesentlichen aber immer noch an Ort und Stelle zu halten.

Erfolgreiche Verfeinerungsansätze lösen die Aufgabe, zu wissen, wann das Protein grösstenteils nur gehalten werden muss und wann mehr Bewegungsfreiheit zugelassen werden darf, automatisch. Wir erreichen genau das - Halten der allgemeinen Form eines Proteins und gleichzeitig Zulassen von Bewegungen der Atome, die verbessert werden müssen - durch die Einführung eines neuartigen, vom elastischen verformbaren Netzwerk (DEN - deformable elastic network) inspirierten adaptiven verformbaren Positionsrestraints (ADPt - adaptiv deformable position restraints), entweder intern angewendet auf eine einzelne Proteinkopie oder durch Verwendung von gekoppelten, evolutionär verwandten Sequenzdaten.

Acknowledgements

First, I want to thank Gunnar Schröder, my advisor, for placing his trust in me and the work I did. Particularly, I am thankful for his perpetual patience and tremendous knowledge he holds.

I also want to thank the ICS-6 institute under the direction of Dieter Willbold at the Forschungszentrum Jülich for the great support and all people working there for the very nice working environment, especially our computational biology cluster (CBC). It was always very inspiring talking to them.

My special thanks go to my direct group members of our computational structural biology (CSB) group, in particular Benno Falkner, Zhe Wang and Kumaran Baskaran for the always very motivating energizing propelling discussions and for the very harmonic and funny office atmosphere and the inspiring and propulsive conversations.

Contents

List of Figures

List of Tables

Abbreviations

g	gram
k	kilo
J	Joule
K	Kelvin, temperature
°C	degree Celsius, temperature
Å	Ångström
J	Joule, work, energy
kJ	kilo Joule
mol	mole, substance elements
fs	femto second
ps	pico second
ns	nano second
μs	micro second
s	second
nm	nano meter
N	Newton, force
MD	molecular dynamics
RMSD	root mean square deviation
dRMSD	distance root mean square deviation
NMR	nuclear magnetic resonance
TBM	template-based modeling
TFM	template-free modeling
SSE	secondary structure elements

Physical Constants / Symbols

Avogadro Constant	N_A	=	$6.022\ 136\ 7 \times 10^{23}\ \mathrm{mol}^{-1}$
Gas constant	R	=	$8.314\ 462\ 1\ \mathrm{J \cdot K^{-1} \cdot mol^{-1}}$
Energy	J	=	$\mathrm{kg \cdot m^2 \cdot s^{-2}}$
Specific energy	J/kg	=	$\mathrm{m^2 \cdot s^{-2}}$
Temperature	K	=	$0\ \mathrm{K} = -273.15\ ^\circ\mathrm{C}$
Boltzman Constant	k_B	=	$\frac{\mathrm{R}}{N_A}$
Electric charge	C	=	$\mathrm{s \cdot A}$
Entropy	J/K	=	$\mathrm{m^2 \cdot kg \cdot s^{-2} \cdot K^{-1}}$
Specific entropy	J/(kg K)	=	$\mathrm{m^2 \cdot s^{-2} \cdot K^{-1}}$

Curiosity, my main encouragement.

Chapter 1

Introduction

1.1 Motivation

Highly accurate protein models are needed to understand the function of proteins, elucidate the molecular mechanisms they are using to fulfill this function, and to develop drugs to manipulate their behavior. For example the design of inhibitors to bind to receptor proteins requires protein models of high quality, which means the position of all individual atoms needs to be known precisely. However, oftentimes, only coarse or low-resolution models are available.

The determination of large proteins and protein complexes is one of the biggest challenges in structural biology. The intrinsic flexibility and heterogeneity however leads to generally lower resolution for such large molecular systems in crystallographic or cryo-electron microscopic experiments. Nevertheless the determination of protein complex structures is of high interest in the study of biological networks. Consequently, an increasing number of low resolution structures are determined and published.

Additionally and importantly, being able to build structures computationally can help to overcome the ever widening gap between experimentally solved structures and known sequences [1], as shown in Fig. 1.1. The rate with which new sequences are determined is orders of magnitude larger than the rate at which new structures are determined.

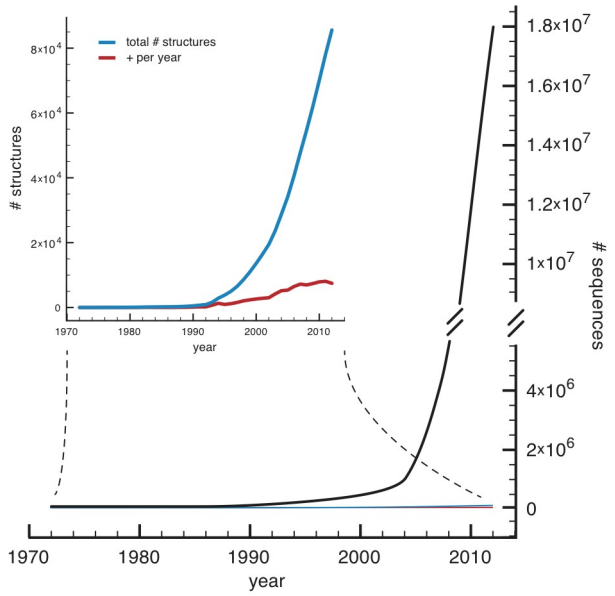

Figure 1.1: The gap between available sequence data and resolved structures. (top left inside: number of structures, main plot: approximate number of sequences). Structure data from RCSB Protein Data Bank, sequence data from non-redundant database (NCBI) with GenBank CDS translations, PDB, SwissProt, PIR and PRF, without ES of WGS (before 1999 are approximate values).

Homology modeling is a tool that uses the fact that structure is much more conserved than sequence, which means even remotely similar (homologous) sequences usually yield rather similar structures. Homology modeling can therefore be used, at least to some extent, to build protein models when a homologous structure is known.

Interestingly, the growth of information on individual domains is saturating [2], unlike the number of new multi domain structures. We are therefore getting closer to knowing all possible domain folds. This means we should soon be able to build homology models for all naturally existing proteins including those consisting of multiple domains.

The lower the sequence identity of the (homologous) template structure that is used to build the protein model, the more errors or structural differences need to be corrected by further optimization and structure refinement.

In this work, several methods were developed to refine low-quality protein structures and move them closer to the correct solution. Since there is no principal difference between the refinement of a homology model or the refinement of a protein model that has been built using low-resolution or sparse data from X-ray crystallography, single-particle Cryo-EM, or NMR spectroscopy, the methods developed in this work can be expected to be applicable in all these different scenarios.

Structure refinement is not an easy problem. The challenging task is to find the native structure given the sequence of amino acids and an approximate starting point. The search space, that is the conformational space, is enormously high-dimensional and, in particular when the problem is simplified, e.g., by reducing the degrees of freedom, the achievable accuracy is limited. Large differences between the approximate starting structure and the true solution are usually very difficult to correct.

The expected solution to this problem is to find the global energy minimum in the vast conformational space. Among the ensemble of energy minima one single basin is globally the lowest state. Finding it generally depends, however, on the shape of the energy landscape. For a simple energy landscape, as shown in Fig. 1.2 **A**, finding the global energy minimum is easy when following a path by straight-forward energy minimization. The direction of the reaction path is unimportant because there is only one basin. In the second, more complicated scenario (see Fig. 1.2 **B**), just one path is most probable. A dynamics simulation will find the global minimum, when it has enough time to sample sufficiently the relevant regions in conformational space which requires sufficient kinetic energy to surmount the barriers along the most probable paths. In the last example 1.2 **C**, searching the lowest energy within the very rugged landscape yields many paths and many barriers to overcome, maybe even insurmountable when sampling with an inappropriate method, with local minima which can be very near to the global minimum.

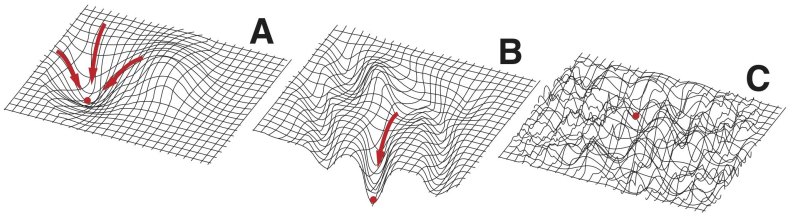

Figure 1.2: Different potential energy surfaces: **A** depicts a landscape with one obvious global basin. **B** shows a more complex surface. Searching the global minimum within **C** is only possible with

Finding the global free energy minimum is thought to eventually solve the problem of finding the native conformation of a folded protein [3, 4]. But since every protein has a different energy landscape, choosing the right method to achieve the required sampling is difficult because the nature of the energy landscapes and the height of the barriers cannot easily be predicted. The global energy minimum cannot be found if the sampling

method is unable to cross the existing energy barriers, and if the method is not optimized enough to sample sufficiently around very low energy basins.

This work focusses on protein structures but the findings obtained with these systems may also be relevant to RNA, DNA, and other macromolecules or polymers.

1.2 Proteins

Most of the mechanisms and processes that govern the cell and its function involve proteins.

Proteins can be considered as one of the major components of life as we know it. Made of amino acids and synthesized in every biological cell, they are responsible for almost all functions a cell has to master. They are the constituent parts of scaffolding elements, they are control elements of cell signal transduction, they are hormones, regulators, indicators and most simply the translated genetic code manifestation that every organism possesses.

Including RNA, proteins are the models that are transcribed within the genetic code. The ribosomal complex (made itself out of proteins and RNA) is the place at which proteins are synthesized. While a protein is synthesized it folds into a three-dimensional structure, and when finally completed, processed and eventually transported to its desired destination it will function for a designated time (controlled passively by its sequence stability and actively by other cell factors).

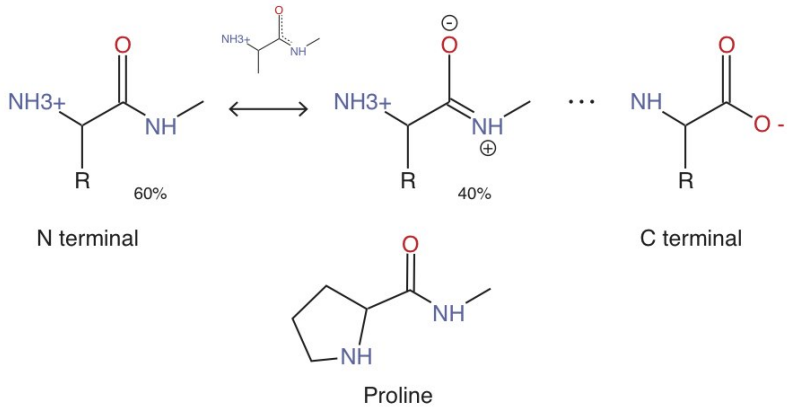

Figure 1.3: The amino acid resonance form within a peptide chain, including the N and C terminal ion form in water. Proline depicted because of its special ring form.

In fact, the three-dimensional shape of a protein (and its dynamics) defines its function and the shape is defined by its amino acid sequence. The amino acid sequence is determined by the sequence of codons in the DNA and its subsequent processing.

A well-defined protein structure is the result of a large number of amino acid interactions, and to a certain degree the structure is tolerant against amino acid substitutions. However, some interactions seem to be more important than others which is indicated by the fact that the corresponding amino acids are conserved between different species. Those amino acids can be expected to be important for the stability of the protein structure

In addition, amino acids share mutually similar properties, which leads to some level of degeneracy in the amino acid code. For example, all 20 amino acids can be divided into two groups: the hydrophobic (H) and the polar (P) group (see Fig. 1.4).

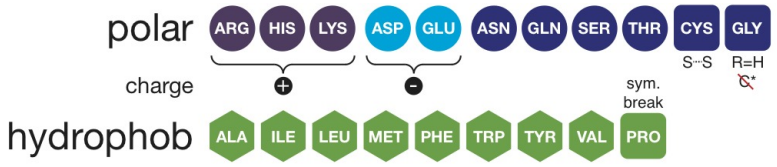

Figure 1.4: Amino acids and their properties. The polar amino acids that carry a positive charge are basic and those that carry a negative charge are acidic. Three of the amino acids have special functions that distinguishes them from the others: *Cystein*, forms disulfide bonds by changing into a cystin. *Glycein* has just one residual hydrogen atom, which leaves it without a chiral C_α atom. *Proline*, due to its 5-ring structure, introduces a change in the direction of the amino acid chain.

Hence a change in one of the amino acids will not automatically change the type of interaction between them and the other residues if it is still of the same group (HP) and not one of the special types (see Fig. 1.4). The exchange of amino acids with similar properties, such as similar charge, similar hydrophobicity or similar size is better tolerated than other exchanges.

This degeneration is the reason for the stability of the three-dimensional structure of proteins. Changing minor parts of a code segment on whichever level (DNA, RNA, amino acid) will most likely not destroy its function. And since the amino acid sequence of a protein defines its structure [5, 6] and therefore also its function [7–10], resilient protein structures are one of the cornerstones for highly developed (multi)cellular protein-based life.

Homology modeling methods exploit these facts because from the previous assumption it simply follows, that many inherited protein sequences must share almost the same fold, which, on the other hand also means, that the natively accessible conformational space of folded proteins is much smaller [11, 12] than its entire conformational space [13].

1.2.1 The structure of a protein

To understand molecular mechanisms, employed by proteins to perform their function, it is vital to know the structures of the molecules. Several methods exist that are able to determine the position of the atoms of the molecules, for example X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy and cryo-electron microscopy (Cryo-EM).

The topology of a protein, i.e. its general fold, is basically defined by the dihedral angles between the backbone atoms of the polypeptide chain (see Fig. 1.5).

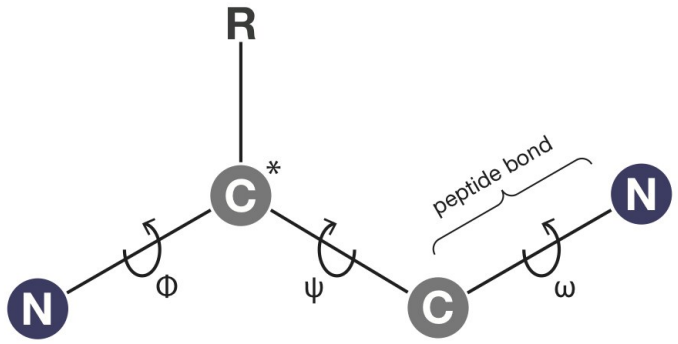

Figure 1.5: Atoms of the protein backbone (hydrogen atoms are not shown). The backbone dihedral angles ϕ ψ and ω are sufficient to define the fold of a protein and also report on the secondary structure. The '*'-sign marks the chiral C_α atom to which the different amino acid residues are attached.

Favorable patterns of hydrogen bonds in the protein backbone lead to formation of specific structural patterns such as α-helices and β-sheets, which is referred to as secondary structure. The three-dimensional arrangement of the secondary structure in a molecule is called the tertiary structure.

In some cases, none of the experimental structure determination methods can be applied. The reasons for this are manifold: X-ray crystallography requires protein crystals, which

are sometimes very difficult to obtain. An example is the difficulty to crystallize the majority of membrane proteins due to the unfolding of the proteins in water and the precipitation in polar environments. NMR cannot be used if the structures are too large, and Cryo-EM cannot be used if the structures are too small. And all techniques will yield limited resolution or quality if the molecule is highly flexible or heterogeneous. Or sometimes the protein of interest simply cannot be expressed in sufficient amounts necessary for these methods to be applicable.

Empirical structure modeling methods that do not deduce their knowledge entirely from experiments but make use of information about already resolved molecules can be applied instead. Ideally, computational methods could be used to build a protein model from nothing but the sequence "de novo" in the so called *ab initio* structure prediction. Unfortunately, this practice is generally not very successful, since it often produces structures severely misfolded or erroneous.

Both regimes of computational approaches are presented and explained in the following section, with a focus on the empirical part which is needed to understand the keystones of the refinement methods presented here.

1.2.2 Protein structure prediction

1.2.2.1 Ab initio folding

Various computational concepts for predicting the correct fold of a protein have been developed and were discussed and reviewed intensively [14–16]. They all try to predict the spatial arrangement of a protein just by knowing its sequence. These methods use thermodynamic principles and start from a random unfolded state or an extended chain of the model to subsequently find the native, lowest energy state in the vast conformational space. To simplify the search problem it is also very common and widely accepted to reduce the degrees of freedom and build coarse-grained models [17] which are then translated into all-atom models and further refined at later stages of the prediction procedure.

1.2.2.2 Homology modeling

Homology or comparative modeling takes advantage from the fact that a lot of protein folds share a similar sequence [18, 19]. Normally, the modeling process begins with a search for a homologous protein which structure has been solved experimentally. This structure is then used as a template to build a model for the new sequence. For the modeling to be reliable, the sequence identity, i.e. the number of identical amino acids at corresponding positions along the peptide chain must be higher than a minimum value, normally around (20 - 25) % [20–32]. Given a certain threshold of identity the expected correctness of the resulting models is depicted in Fig. 1.6.

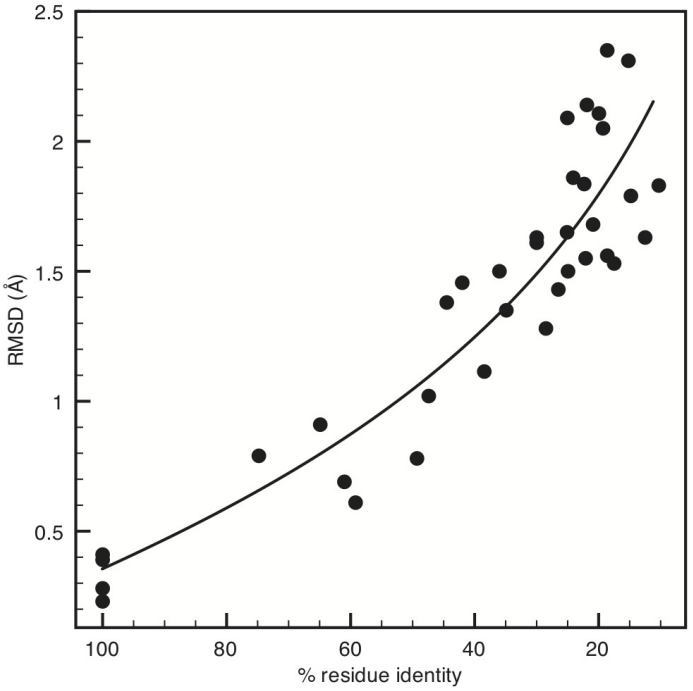

Figure 1.6: Correlation of the residue identity versus the RMSD in Å of homology models. Data from Chothia 1986 [18].

It shows the root mean square deviation (RMSD) of homologous structures with different sequence identities. Some models even share major structural similarities when the identity is very small such as around 10 %. By searching for similarity not in sequences but on a structural level it is possible to detect homologous structures for those models that would have never been identified by sequence comparison. For this it is necessary to have search engines that can search in data bases for structural elements [33].

The actual modeling step then requires to assign the target sequence to the template backbone. Since this process is normally not computationally expensive, comparative modeling is usually done in an all-atom representation.

Not every approach is evenly successful in generating high quality models from a given sequence. Due to the staggering vastness of the given problem, most methods that try to solve the problem de novo, just with physical approaches by exerting and iterating thermodynamic principles are not yet very reliable.

The quality of the homology model depends directly on the quality of the template. If the template is close, i.e. the sequence identity is high (such as 50% and better) and if the template structure has been solved to high resolution, the homology model can be of very high quality with an RMSD of 1 Å or better.

However, *ab initio* methods typically do not reach this level of accuracy. Reaching RMSD values of 4 to 2 Å is usually the exception, especially for larger proteins (> 150 residues).

As homology modeling often produces relatively good models it is common to just use the model as is, without refining it. In addition, attempts to refine a model can, and often did in the past, deteriorate parts of the model [34, 35], which made the practice of refining such models an unpopular and inaccessible task.

1.2.2.3 Refinement

Since homology models are often already close to the correct solution, one would think that some further optimization or refinement suffices to correct the small errors in the model. However, existing methods so far have not been able to consistently move the models closer to the correct solution. That is why until now the general consensus seems to be to leave homology models unrefined, without changing the protein backbone conformation [34].

With this work we show that consistent improvement can be achieved by carefully adjusting and augmenting modern molecular dynamics (MD) simulation techniques.

The starting structures for refinement discussed here are obtained by homology modeling techniques.

There is no strict definition of which model improvements should be considered a refinement and which is rather extensive remodeling or rebuilding. In this work, refinement means the improvement of models with an RMSD to the native structure of 4 to 1 Å [36].

Refinement in this work is performed with MD simulations [37], using a physics-based force field [38], which is extended by additional restraints to improve the sampling.

Three different restraining approaches were developed which are based on particle coupling mechanisms. The different coupling restraints all modify the original energy function and eventually smoothen the originally rugged energy landscape. The improved sampling on the one hand guides structures towards the global energy minimum and on the other hand reduce the risk of being trapped in local energy minima.

These three different MD setups were intensively tested and are described in Chapters 3, 4, and 5, respectively.

Chapter 2

Methods

2.1 Molecular dynamics simulations

Atoms move according to their interaction energies and the resulting forces between them. To simulate this motion one needs to describe these interactions in a way that allows calculating the forces in an efficient way. The collection of the interaction energy terms is called force field. It defines the energy landscape and contains all parameters which let the simulations behave like they should. Changing these parameters can have major consequences for the result of a simulation. Even tiny deviations can cause large differences and errors. The calculations for moving an atom based on the force field parameters are iterated in every single step of the simulation for each and every atom as it is described below in more detail.

2.1.1 Energy minimization

Prior to any productive molecular simulation, an energy minimization (EM) has to be done, which ensures that the system is in the nearest local energy minimum and removes atomic clashes and excessive bond strains. This further avoids large unnatural forces at the beginning of the simulation which could artificially destabilize the complete system.

The algorithm we use in our setups is the steepest descent approach. We define

$$\vec{r}_k, \text{position vector of all atoms, i.e. all 3N coordinates, at minimization step } k$$

$$h_0, \text{initial maximum displacement}$$

$$\vec{F}_k = -\nabla V, \text{the force vector on all atoms}$$

Then, the new position vector is calculated by

$$\vec{r}_{k+1} = \vec{r}_k + \frac{\vec{F}_k}{\max(|\vec{F}_k|)} h_k \tag{2.1}$$

$$if(V_{k+1} < V_k), \text{ new positions are accepted and } h_{k+1} = 1.2h_k$$

$$if(V_{k+1} \geq V_k), \text{ new positions are accepted and } h_k = 0.2h_k$$

Eq. 2.1 is repeated until a defined number of iterations has reached its threshold or when the maximum force is smaller than desired.

2.1.2 Force field

The high-dimensional energy landscape on which the atoms are moving is described by the force field. The forces F within a MD simulation are then obtained simply by the gradient of this energy landscape. The force on an atom i is given by

$$\vec{F}_i = -\nabla_i V \tag{2.2}$$

with V being the potential energy. All forces acting on i are conceptually divided within the force field into the bonded and non-bonded interaction part. The bonded interactions are summed over the covalent bonds (bon), the angles (ang), the proper (dih) and the

improper dihedrals (imp).

$$\begin{aligned} V_{bonded} = & \sum_{bon} l_d(\vec{d} - \vec{d_0})^2 + \\ & \sum_{ang} l_\theta(\theta - \theta_0)^2 + \\ & \sum_{dih} l_\psi(1 + \cos(n\psi - \sigma))^2 + \\ & \sum_{imp} l_\phi(\phi - \phi_0)^2 \end{aligned} \tag{2.3}$$

The non-bonded part below is normally computed pair-wise additive and describes the Lennard-Jones and the electrostatic interactions.

$$V_{non-bonded} = \sum_{non-bonded} = \overbrace{\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{d_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{d_{ij}}\right)^{6}\right]}^{Lennard-Jones} + \overbrace{\frac{q_i q_j}{d_l d_{ij}}}^{electrostatics} \tag{2.4}$$

with the charges q_i and q_j of atoms i and j and their distance d_{ij}.

With Eq. (2.3) and Eq. (2.4), we can calculate the complete force vector on all atoms

$$\vec{F} = \nabla V(\vec{r})_{bonded} + \nabla V(\vec{r})_{non-bonded} + \nabla V(\vec{r})_{restraints}, \tag{2.5}$$

where the forces from restraints are also added. The movement of the atoms is then calculated by numerically solving the Newton's equations of motion

$$m\,\frac{d^2\vec{r}}{dt^2} = \vec{F}(\vec{r}) \tag{2.6}$$

The total energy of the system is then

$$\mathcal{H} = T + V \tag{2.7}$$

or in other words,

$$E_{total} = \frac{m}{2}\,\dot{\vec{r}}^2 + \vec{F}(\vec{r}). \tag{2.8}$$

In order to simulate a canonical ensemble (NVT) as done in this work, the equation of motion must be extended by a heat bath term that regulates the systems temperature. The Nosé-Hoover temperature coupling method is used here for this purpose. In this

method the Hamiltonian is extended by a friction term to be correct for this type of ensemble. The friction force is then the product with the particle's velocity.

$$\frac{d^2\vec{r}_i}{dt^2} = \frac{\vec{F}_i}{m_i} - \frac{p_\xi}{Q}\frac{d\vec{r}_i}{dt} \tag{2.9}$$

where p_ξ is the momentum of the added friction parameter ξ. The constant Q is the mass parameter of the reservoir, which determines the strength of the coupling.

$$Q = \frac{{\tau_T}^2 T_0}{4\pi^2}$$

Several other coupling mechanisms that extend the system to a heat bath exist, but do not necessarily generate a correct canonical ensemble. One of them is the Berendsen temperature coupling. Because the Berendsen thermostat is suppressing the fluctuations of the kinetic energy, no correct canonical ensemble is produced. This fact for us was the reason to use the Nosé-Hoover temperature coupling within all simulations.

2.1.3 Canonical energy distribution

In a given canonical ensemble, a particle can obtain many different energy states. The probability for a certain state is given by the Boltzmann distribution.

$$P(i) = \frac{1}{Z}\, c_i\, e^{-\beta E_i} \tag{2.10}$$

where

$$\begin{aligned} e^{-\beta E_i}&, \text{Boltzmann factor} \\ \beta = \frac{1}{k_\mathrm{B} T}&, \text{inverse temperature} \\ Z = \sum_i e^{-\beta E_i}&, \text{partition function} \\ k_\mathrm{B}&, \text{Boltzmann constant} \\ c_i&, \text{degeneracy level of i.} \end{aligned}$$

Or, another possibility of writing it

$$\frac{N_i}{N} = \frac{1}{Z}\, c_i\, e^{-\beta E_i} \tag{2.11}$$

describes, which fraction of particles N_i occupy the energy state E_i.

2.1.4 Periodic boundary conditions

The simulations of protein presented here are performed in an explicit water environment. To reduce the overall computational cost, this water environment is chosen to be rather small.

A simulation box of finite size raises the question of how to treat the boundaries. Ideally, the water at the boundary behaves as if it was embedded in an infinite solvent environment. A simple solvent-vacuum boundary with a hard wall would clearly introduce severe artifacts. An elegant solution to this problem is provided by using periodic boundary conditions.

In this case atoms interact not only with other atoms within the box, but additionally with atoms in periodically extended simulation volume and thus with virtual copies of the original system. Furthermore, atoms that leave the box on one side re-enter the box on the other side, while keeping their momenta.

For small solvent environments the reduction of artifacts is significant compared to a simple hard wall boundary. The larger the simulation box, the more negligible are those errors introduced by the boundaries. But since simulation boxes should be kept as small as possible to reduce simulation overhead and reduce computational cost, periodic conditions are typically preferred.

However, precaution is required when simulating boxes with bonds (e.g. distance restraints) that are longer than half of the box size. This may introduce severe artifacts and simulation errors since the periodic images of the restraint atoms may interfere with itself. We came across this problem while producing ensembles in chapter 3. It was solved by expanding the box size, which was the main reason for dropping of the distance restraint approach and switching to an implemented code version in chapter 4 and 5.

This allowed at the same time to also benefit from the parallel domain decomposition approach, as outlined further below.

2.1.5 Numerical integration method

To simulate the motion of a particle on an energy landscape requires to solve Newton's equation of motion. For proteins described by a realistic force field, this can only be done numerically.

A common way of doing this is the leap frog method, which integrates the equations of motion by using the positions q at time t and the velocities at time $t - \frac{1}{2}\Delta t$. The update of the position and velocity is then calculated using the force $\vec{F}(q)$ at its position at time t.

$$\dot{q}(t + \frac{1}{2}\Delta t) = \dot{q}(t - \frac{1}{2}\Delta t) + \frac{\Delta t}{m}\vec{F}(t) \tag{2.12}$$

$$q(t + \Delta t) = q(t) + \delta t\dot{q}(t + \frac{1}{2}\Delta t) \tag{2.13}$$

2.1.6 Particle decomposition versus domain decomposition

The speed of a simulation is determined by its slowest component. In the single core or local memory approach, the slow part is the calculation of long range interactions. When distributing the system to cores on interconnected workstations, the network communication part becomes the limiting factor. In the worst case, when the system is not partitioned, a minimum of half of all atoms have to be communicated between all hosts or cores to maintain a stable setting. For N cores, the communication is made of $N \times N/2$ coordinates, which does not scale well at all. We had to use this setting for our first approach (chapter 3) because the long distance restraints used there were not dividable and required a non-decomposed system. All other systems presented here are optimized to work with the fast parallel domain decomposition approach. In those partitioned systems, only a subset of atoms has to be communicated, because the communication now only has to be accomplished with coarse domains and not with particle precision.

2.1.7 Molecular dynamics flow chart (real space/PME)

A general work flow of a simulation with all important steps is depicted. In this chart, the separation of the computational work load is split into the part of the real dynamics calculation and the solving of the PME grid summations. This procedure enables the possibility of ideal work load balancing and enhanced efficiency per computational unit.

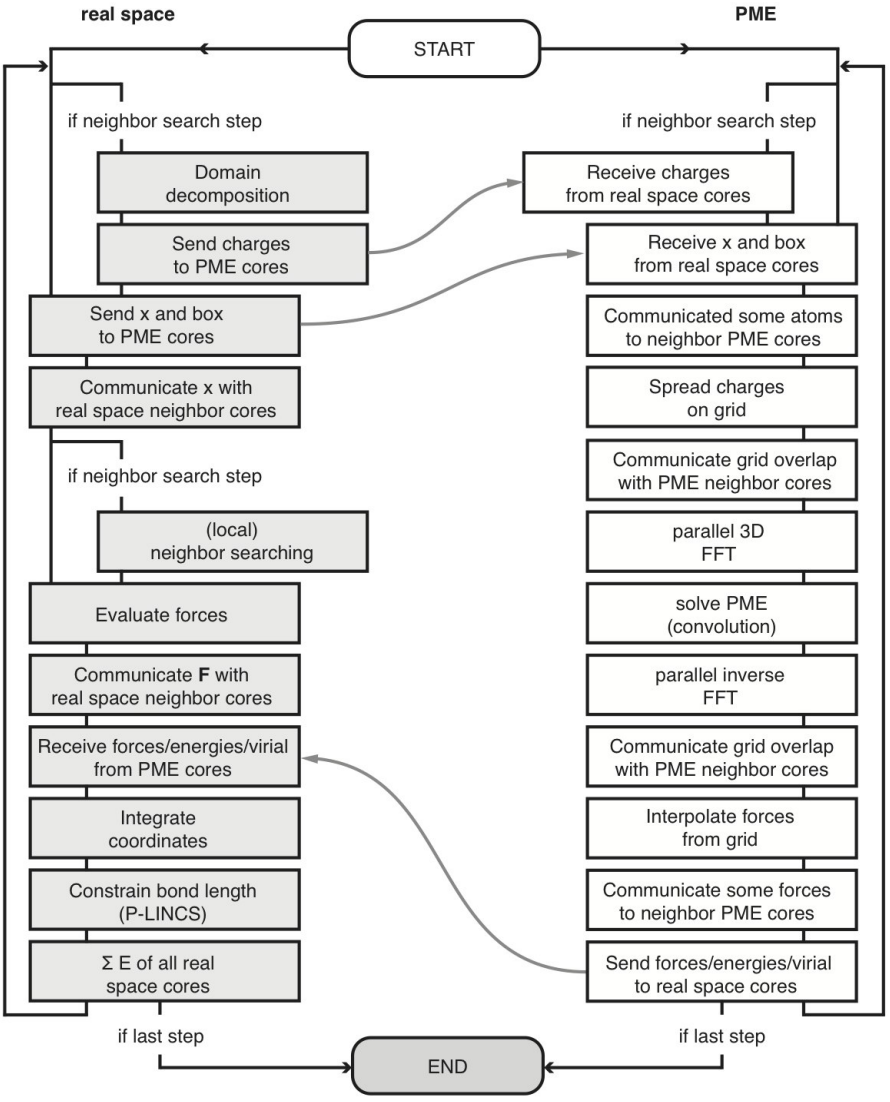

Figure 2.1: Molecular dynamics flow chart in the case, where real space calculations are separated from PME nodes. This is done, e.g. when using domain decomposition. Gray arrows show communications between the different types of nodes (see also [39]).

2.2 Monte Carlo simulations

In contrast to MD simulations, no forces are calculated in Monte Carlo (MC) simulations and its extension to canonical ensembles by Metropolis [40]. All particle movements are based on random numbers, hence no time steps are involved in any of the computations, and no time information is obtained. Despite this, MC simulations are used widely for molecular studies of all kind, but mostly with simplified systems with a reduced number of degrees of freedom.

Molecular dynamics systems with explicit solvent are generally too big for MC approaches because sampling can become very inefficient when also including all atom solvent effects into the calculations. The introduced Metropolis algorithm is a so called importance sampling approach, which follows the notion of excluding the huge space of improbable states and focus the sampling around a given problem, while still producing correct probabilities.

A trajectory of configurations is produced by comparing the energy of the previous state n with the energy of a new test state m (Markov chain). The test state m is accepted as the next step of the trajectory according to the transition probability $W_{n\to m}$, which depends on the energy difference of these two states.

The new state is accepted if the energy of the new state E_m is lower than the energy of the previous state E_n, or if a random number $x_i \in [0..1]$ is smaller than the factor $\exp(-(E_m - E_n)/kT)$.

While $Z = \sum_i e^{-\beta E_i}$ is generally inaccessible, the choice of the test states has to obey the detailed balance condition, which is

$$P_n(t)W_{n\to m} = P_m(t)W_{n\leftarrow m}, \tag{2.14}$$

which means that the ratio of the transition probabilities is equal to the ratio of the probabilities, as given by the Boltzmann factors.

We used MC simulations to verify the effects of coupled simulations in a very simplified system. Only one particle in the case of the approach in chapter 4 and two particles

in the case of the augmented approach in chapter 5 were simulated with a simple one-dimensional energy function that gave us a reasonable result on the basis of 5×10^6 steps.

Chapter 3

Ensemble-Restrained Structure Refinement

3.1 Introduction

One of the biggest challenges in structural biology is to solve the problem of how proteins attain their three-dimensional functional structure in a biological context, namely how proteins fold either in vitro or within a cell. After almost half a century of investigations and research, predicting the folding of proteins is still unsolved although many different approaches from different scientific disciplines have been proposed and progress has been made [41]. A broad range of methods exists to predict protein structures or refine protein structural model for which the best working ones are still the template-based methods, for example homology modeling.

In this thesis different methods have been developed to optimize homology models. The first of these methods is presented in this chapter and aims to refine a homology model by simulating the dynamics of several copies of the same starting model at the same time. The idea is to couple all these copies with restraints to keep the structures similar to each other. The resulting effect is similar to that of a particle swarm optimizer which is expected to smoothen the energy landscape and thereby facilitate the crossing of energy barriers.

Molecular dynamics simulations usually have a problem refining protein structures to high resolution, mostly because they easily become trapped in local minima and sample inefficiently low energy regions around the global minimum. Being able to add the capacity of overcoming energy barriers on the energy landscape to the accuracy needed to simulate around the native conformation will increase the probability of refining a given structure. Our approach extends the regular molecular dynamics approach and simulates a number of structures that are weakly coupled to each other simultaneously. They therefore have to move together within the simulation. This multiple copy-like optimization procedure narrows the obtained structural distribution while at the same time facilitates the crossing of energy barriers due to the introduced cooperativity between the structures.

This approach was tested in the international blind test CASP (Critical Assessment of methods of Protein Structure Prediction) [1, 23, 42–44]. CASP is an evaluation of the worldwide improvement of how powerful the prediction methods are and what is possible today. Since 1994, CASP takes place biannually and allows modelers to assess their methods on models that are solved experimentally but were not published yet. In this way, only the assessors know the answer to the given problem. Recently (CASP8, 2008), a specially arranged category dedicated to the refinement problem was introduced. It solely focusses on models submitted in the categories held earlier in the main template-based modeling (TBM) section and aims exclusively to improve the predicted structures. We are focussing only on this special refinement category.

Within the refinement category, the models selected for refinement are not dramatically wrong structures, because the models for the refinement section have already been filtered by the post modeling procedures of the corresponding server or production method.

The assessment itself begins when the first models targeted for refinement are posted on the workshops conference homepage (http://www.predictioncenter.org). This happens subsequently to the initial start of the model building event, which is divided into the template-based and template-free modeling categories. In addition, participating groups can be either humans who can use any method including manual modeling and inspection, or servers which run automatically without any human intervention. The best prediction is then selected for the refinement category.

Normally one to three targets are posted initially, given up to three weeks of process time. In total, 14 targets were posted. After the refinement of the model was finished the structures were uploaded back to the CASP server. 5 models could be submitted, ranking the submissions from 1 to 5, where the 5th would be the target which is thought to be the least refined one. Once all submissions were finished, the assessors compared the submissions with the experimental structures hold back and published their rankings on their website and on the conference meeting, held subsequently.

3.2 Methods

3.2.1 Restraints approach

In our approach for CASP (refinement category), eight structures were simulated simultaneously. The structures are weakly coupled to each other using restraints between corresponding C_α-atoms. For the molecular refinement we wanted to take advantage of the idea that simultaneously coupled simulations might benefit from the averaging moment the coupled component (another copy of the molecule in our case) introduces. Somewhat similar strategies were published before, but most of them focussed on *ab initio* modeling or refinement based mostly on MC approaches [45] or calculated averages from aligned models [46].

Since we simulate the coupled molecules all together in one box, under all-atom explicit solvent conditions, once the setup of the restraints is finished, the system is self maintained and needs no further intervention.

The complete system contains eight copies of the starting homology model. We restrained those C_α atoms which had a low degree of fluctuations within a short 1 ns pre run, since we just wanted to move atoms which were not in a low energy state. To allow just those atoms to move, the well placed low fluctuation atoms were stabilized through a scaffold provided by the other copies. To find the atoms for restraining, we used the following scheme:

1. After energy minimization and equilibration, perform a short 1 ns simulation of the targeted protein.

2. For all C_α atoms calculate the RMSF over all frames of the simulation, then average the values for each C_α atom.

3. Restrain those C_α atoms which fluctuate less than the calculated average over all fluctuations.

The motivation for this was to reduce the sampling in the well defined (mostly core) regions to keep regions that seemed to be acceptably positioned rather fixed and allow more freedom in the more variable (e.g. loop) regions.

The harmonic potential force constant between the distance restrained C_α atoms was 1000 kJ mol^{-1} nm^{-2}.

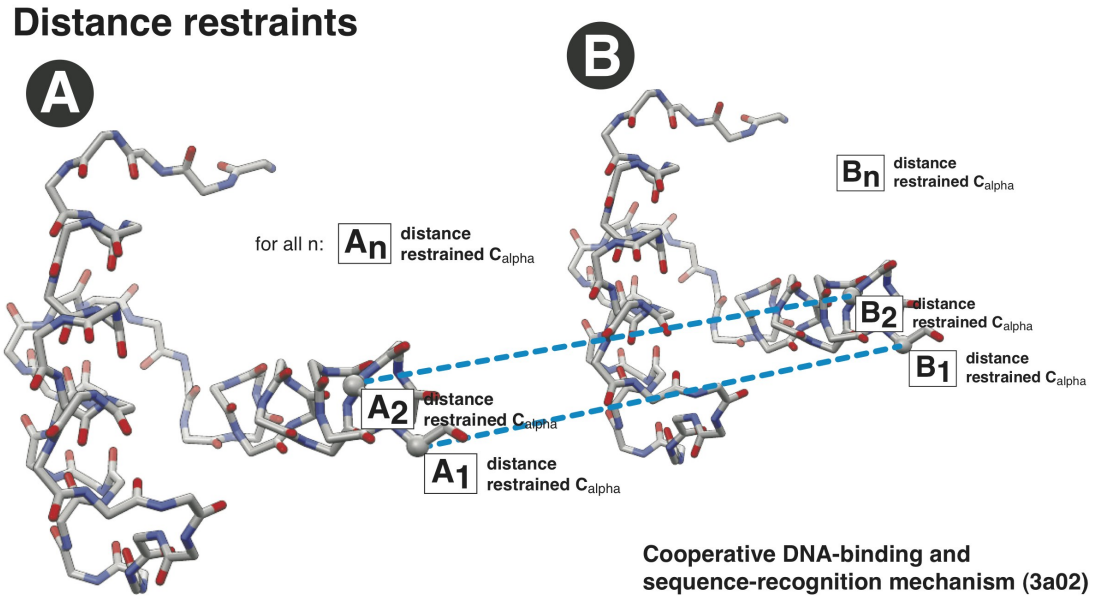

Figure 3.1: Distance restraints as used in our approach. Molecules **A** and **B** are just copies of each other. All C_αs are connected like **A1** with **B1** and **A2** with **B2**, where n is the number of residues. The connection strength can be controlled by the force constant of the harmonic potential between the atoms, represented by the dashed **blue line**.

In this way the coupled proteins were free to do any conformational motion as long as all restrained structures roughly follow this motion. The mechanism is depicted in Fig. 3.1, using the example of a simple two-copy system.

3.2.2 Simulation setup

3.2.2.1 Evaluation of restraints

First, an initial steepest descent energy minimization with a step size of 0.01 fs was carried out. For evaluation of the restraints we performed a 1 fs time step, 350 K temperature, NVT Nosé-Hoover, explicit tip3p solvent 1 ns simulated annealing MD simulation. The annealing temperatures were going from 350 to 320 and 300 to 280 K over 0, 300, 500, and finally 1000 ps, respectively. Then, the C_α atoms were investigated for their fluctuations and those which moved less than the average were distance restrained.

3.2.2.2 Production runs

Eventually, the system contained the box filled with solvent and the protein model copied to each corner of the simulation box. The distance restraints connected all C_α atoms intermolecular. That means no connections were made within a molecule. Just corresponding C_α atoms of the different copies were connected.

For the main refinement, 1000 8-copy-restrained simulated annealing runs of 100 ps length were performed starting at a temperature of 200 K and ending at 100 K. The MD simulations were done with Gromacs in version 4.0.7 [39, 47] using the AMBER03 force-field [48] and explicit solvent tip3p [49].

In the end, the simulation box was made of the solvent and the copies of the protein model. We took eight copies because this number makes up a symmetric box and a symmetric system of distance restraints (see Fig. 3.2).

3.3 Results

3.3.1 One-dimensional Monte Carlo simulations

The general assumption that coupling several copies of a molecule helps to guide our simulations to the global minimum and smoothen the energy landscape was tested by a

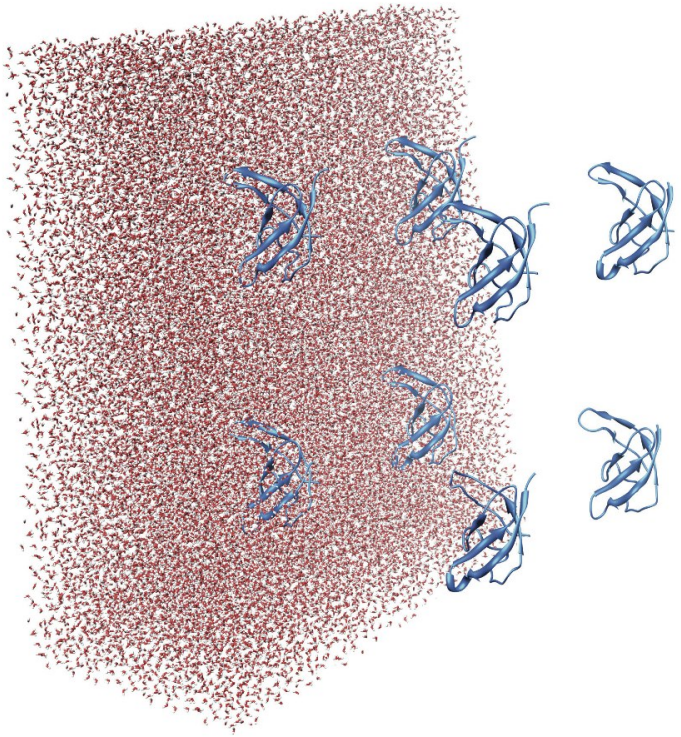

Figure 3.2: Composition of the simulation box. The solvent is truncated to show the protein arrangement. Embedded in the solvent are 8 copies of the protein, which are connected by distance restraints as shown in Fig. 3.1 (not shown here).

simple Monte Carlo simulation approach, where we coupled two particles and simulated them within a simplified energy landscape.

The results (see Fig. 3.3) showed that by increasing the coupling of the two particles, the probability for finding the particle in the global minimum was also increased. When we coupled the particles only weakly, the probability to also occupy the other local minima increased while decreasing the probabilities for finding it in the global minimum.

3.3.2 The structures

All refinement models of CASP9 are shown in Fig. 3.4. Some of the models were problematic and introduced difficulties while working with them. Especially model TR517 and TR614. Model TR517 had a poorly described region from residue 69 to 88 (no secondary structure present). This is a very long region in terms of a refinement approach. We remodeled this section with the program Coot (version 0.6.1) [50].

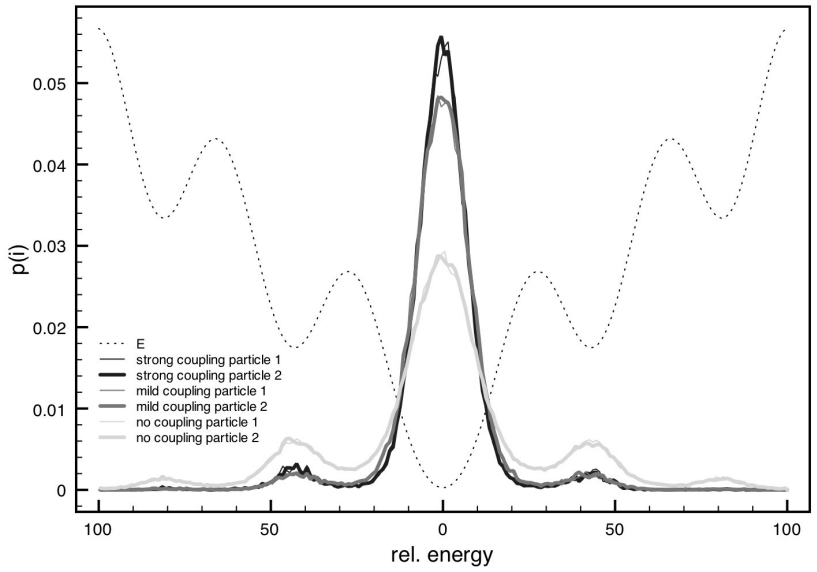

Figure 3.3: Monte Carlo simulation with coupled particles. The dotted line shows the plotted energy function with its local minima and one global minimum. The strongest coupling did perform best, appearing most often in the global energy minimum. By decreasing the coupling strength, the particles acted more and more like a single one, showing, that coupling confines a system closer to the global energy minimum.

Model TR614 was created with NMR spectroscopy methodology and two structures were given to refinement as both models were equally ranked by the assessors. Since the submission threshold of 5 models per target still took effect, deciding for structures to submit was a bit narrowed.

MODEL (*-id)	exp method	♯ atoms C_α	RMSD (Å)	GDT-TS	dRMSD (Å)
TR517	X-ray	159	6.932	0.8088	5.2611
TR530	X-ray	115	1.990	0.8594	1.3197
TR557	NMR	145	4.058	0.5441	3.1464
TR567	X-ray	145	3.435	0.7817	2.2811
TR568	X-ray	158	6.149	0.5490	4.4619
TR569	NMR	79	3.010	0.6552	2.2237
TR574	X-ray	126	3.583	0.6201	2.5333
TR576	X-ray	172	6.850	0.6431	4.7643
TR592	X-ray	144	1.257	0.9024	1.0310
TR594	X-ray	140	1.817	0.8661	1.2569
TR606	X-ray	169	4.850	0.7175	3.4613
TR614 a	X-ray	135	6.490	0.6963	5.1231
TR614 b	X-ray	135	4.199	0.6818	3.2299
TR622	X-ray	138	7.473	0.6680	4.0014
TR624	X-ray	81	5.189	0.5543	3.1542

Table 3.1: All CASP9 refinement models. TR517 was aligned and scored with 138 residues, for it had missing parts that were modeled poorly in region 65-88. * ♯ given by the CASP competition.

In all other cases, non of the information given by the assessors about the quality or the

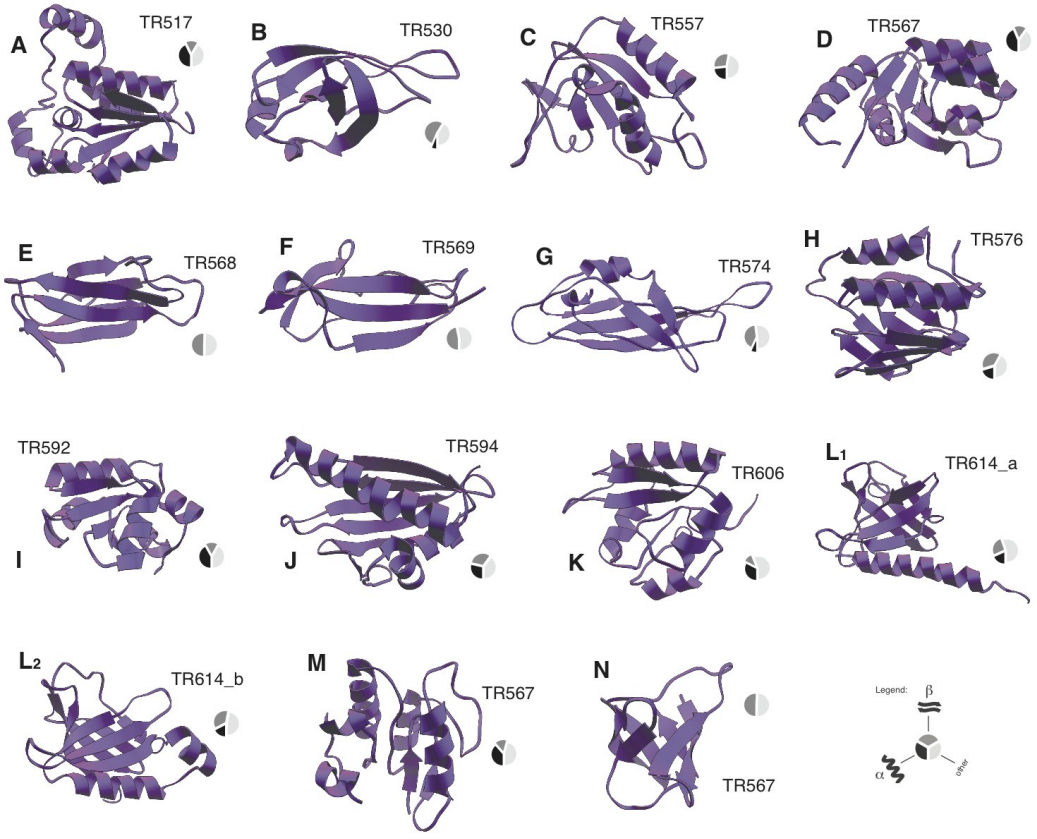

Figure 3.4: All models of the refinement section of CASP9. Also, the secondary structure content is highlighted and the legend in the right corner explains the colors.

region to focus on were taken into account. Interestingly, most of the times, the regions we identified as highly fluctuating coincided with the regions that were described on the CASP website as regions that needed special attention in the refinement.

3.3.3 Assessment of model quality

The evaluation of the CASP9 refinement results were based on several comparison metrics.

The generally and widely used measurement to assess the quality of a structure is the root mean square deviation (RMSD) method. It measures the positional distance of corresponding atoms after an ideal alignment of the two structures. Most of the times, the alignment of the models is done with a least squares quadratic (LSQ) fit. Though the resulting value of this method normally gives a good approximation of the quality of the model, minor deviations distributed throughout the complete structure can have a significant impact on the score.

One of the main methods for the assessments of structure quality in CASP is the global distance test (total score) (GDT-TS) [51]. It is based on RMSD but compared to this, it works iteratively while considering 4 different distance thresholds of (1, 2, 4 and 8) Å. Superpositions of the compared models are made and tested for the fraction of atoms that fall within one of the given thresholds. In the end, all percentiles are summed together and form the total score. Reasonable values of the score are within 0.2 and 1.0, where a value below 0.2 can not differentiate between arbitrary placed atoms anymore. A score of 1.0 means 100 % identity. Thus, higher means better. The main reason for a quality metric like this is the advantage of ignoring regions of the protein, that are far away from nativeness, since just atoms more distant than 8 Å are deteriorating the score significantly. That ensures good scores for models that are mostly correct, even when a loop is severely misfolded. Scored with a RMSD, those regions would have a higher impact on the scoring value and models with a considerably good core region have a higher (worse) rank, compared to a GDT value. Another version of the GDT score is the high accuracy (GDT-HA) implementation, in which the thresholds are lowered to (0.5, 1, 2, 4) Å, The slightly modified version of GDT-TS is more responsive for small changes and useful to compare very similar structures.

In the global distance calculation for side chains (GDC-SC) [52], a GDT-like evaluation procedure is used, but this time not by looking at the C_α atoms but a characteristic atom describing each side chain uniquely.

A combined all-atom per-residue score, which takes the side-chains into consideration is given by SphereGrinder. The "sphere" is constructed by a 6 Å distance around the C_α atoms, respectively. Then, a fit based on a RMSD of the sphere-region is performed, and eventually iterated through all available spheres of the structures. The final score is calculated by summing and averaging the fraction of all per residue scores that were within 2 Å compared against the native structure.

Finally, MolProbity was used to test the overall physical correctness of the submitted models. It tests side chain rotamers, steric clashes and Ramachandran scores [53].

3.3.4 Selecting structures from an ensemble of models

Our refinement method produced a large number of models. The task was to pick the best model from the ensemble of all models. To be able to rank the structures for separating the improved from deteriorated models, we used a combination of clustering and Ramachandran/H-bond scoring. To obtain the best five candidate models, first all models were clustered by using the Jarvis-Patrick clustering algorithm [54]. This method clusters the structures based on similarity. Structures were chosen from the largest cluster, where the fine grained score to pick within a cluster was done with a Ramachandran score and evaluation of H-bonds.

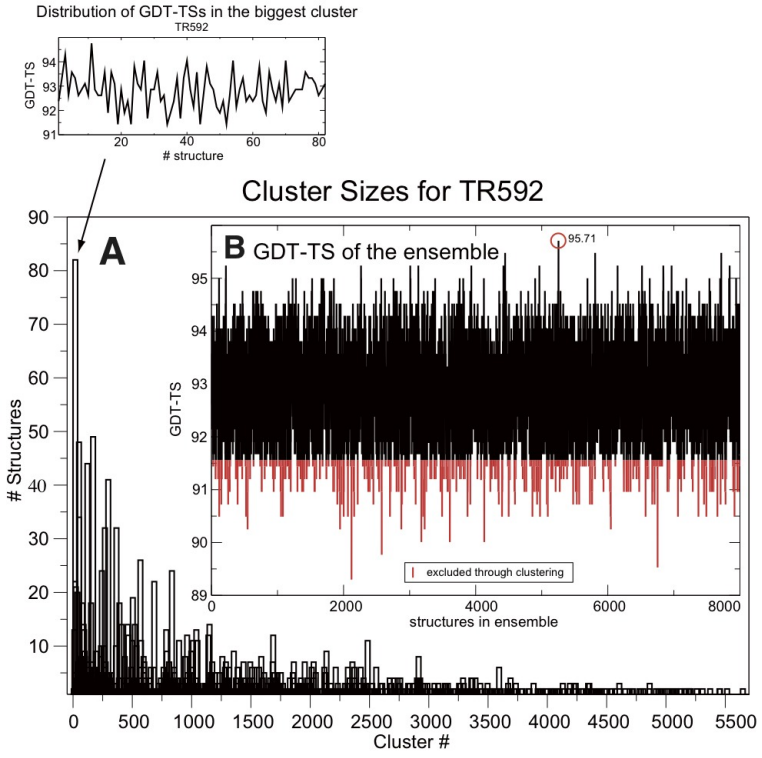

Figure 3.5: Ranking of models for CASP9. Performed with the Jarvis-Patrick [54] clustering method, which is based on a distance matrix, arranging very similar models into clusters.

Fig. 3.5 shows the clustering on the example of model TR592. On the right, **B**, the GDT-TS score of each frame of the simulation ranked against its crystal structure is plotted. The best frame is marked with a small circle. The red area shows the exclusion

of deteriorated frames through clustering. All clusters are depicted in **A**. The individual distribution of GDT scores within the cluster on top reflects the variance of this structural fraction of the ensemble.

The total ranking of the refinement section of CASP9 is depicted in Fig. 3.7, based on the ranking with methods outlined above.

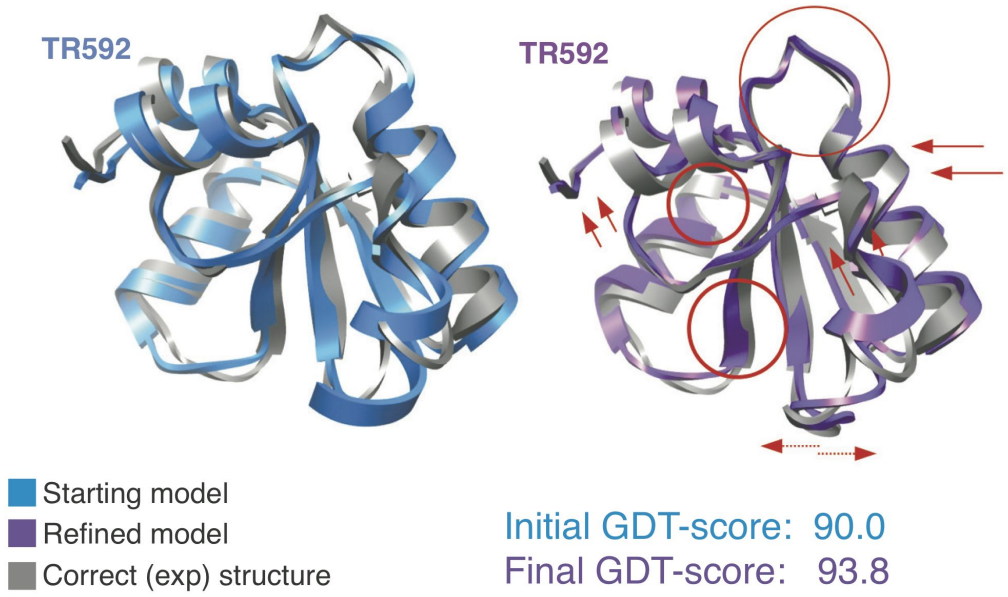

Figure 3.6: Result for model TR592. The model already started with a very high GDT-TS score. Still we were successful in refining the model, although improving a model close to the optimum is not easy because the chances of deteriorating parts of it are high. The starting model (blue) had a GDT-TS score of 90.0 compared to its crystal structure (gray). Our improved structure (purple) had a score of 93.8. The improved regions are highlighted in red on the right side.

Our molecular dynamics approach was able to improve seven of the 14 CASP9 refinement targets. For one of the most difficult targets TR592, Fig. 3.6 depicts a structure comparison in which the crystal structure (**gray**), the homology model (**light blue**) and additionally our improve model (**purple**) are superposed. The target can be considered difficult, for it was already in a near native state (RMSD 1.257 Å and GDT-TS 0.9024, compared to its crystal structure). Improving those very near native models can be very hard, because a significant amount of refinement has already been applied successfully during the modeling process. Here, the approach of holding in place what was already modeled very well and just refine what is misplaced by our restraints was somewhat successful. The encircled regions in Fig. 3.6 show the loop improvement, as well as the stabilizing effect for β-sheets and α-helical secondary structural elements. The arrows

point to a tendency to enhance a more compact representation. The falsely modeled α-helix in the lower region of the protein is about to destabilize and by that dissolves.

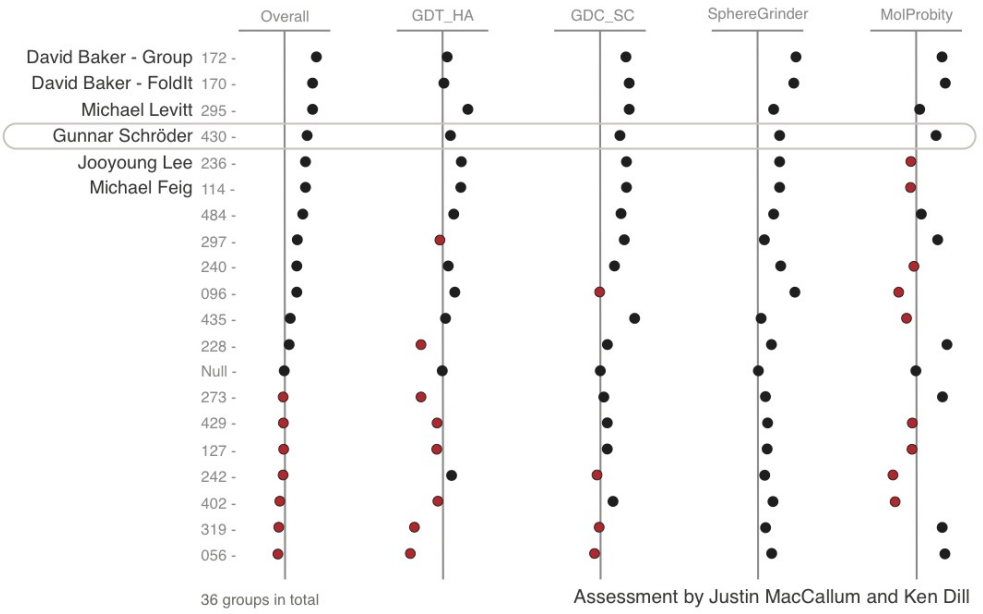

Figure 3.7: All scoring methods and results depicted. Listed are just the 20 first Groups. Assessment was done by Justin L. MacCallum and Ken A Dill.

The overall ranked result, considering all models submitted during the complete contest is depicted in Fig. 3.7. 3 groups which were already successful in various previous CASP rounds were on the first positions. The Baker group on position one used their software package Rosetta [26, 55, 56] to generate and score/rank their ensembles. The second group used a rather unconventional approach (also under the lead of the Baker group): the predicted models were placed into the computer game FoldIt [57]. Players from around the world were asked to participate in the refinement process by playing around with the structures in this computer game. To facilitate that, the FoldIt game makes use of an intuitive graphical user interface (GUI), which enabled untrained laymen to move side chain rotameres - and more - of protein models in an easy way. The goal of the game is to minimize the Rosetta score for the current model, calculated immediately during the interactive modeling process. The conformational changes introduced by a human player actually enhanced some models but also lead to some very false structures.

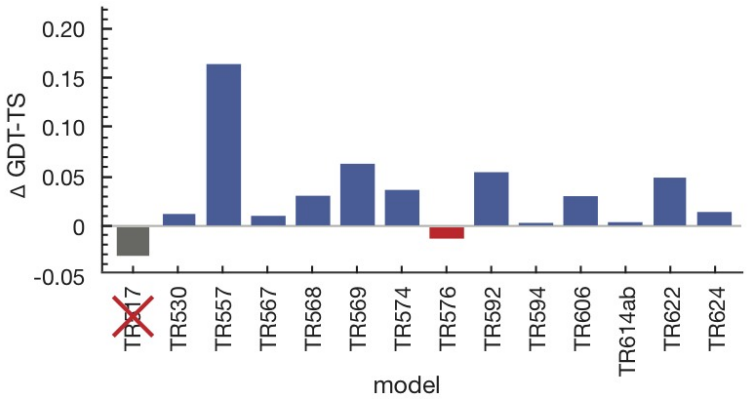

Figure 3.8: Refinement results when looked at the maximum GDT-TS change per model. The first model we did exclude, due to consistency reasons, since the model was stripped from a lot of atoms in a centered region, leading us to think that all generated structures were biased because of the missing interactions.

Finally, to give an overview of the capability of our CASP9 refinement, Fig. 3.8 shows the Δ GDT-TS when subtracting the initial start GDT-TS from the refinement result. All models except two showed improvement behavior.

3.4 Discussion of CASP9 results

The very short (100 ps) simulated annealing procedure we used was quite conservative since not only the achievable improvement but also the potential structure degradation was limited. A more aggressive annealing (longer simulation time and higher temperatures) has the potential for larger improvements but at the same time also higher chances of deteriorating the structures. Also the structure selection method needs to be improved, for example by using more structure validation procedures, as the clustering method performs only slightly better than choosing structures randomly. This was approached and partially solved subsequently for the two other approaches in chapter 4 and 5, where our compact score achieved better results in discriminating deteriorated structures from refined models.

Crystal contacts were proposed for several of the structures, namely for model TR517, TR567, TR576, and TR592. No knowledge of this was available at the time of the competition. Three of this targets missing essential crystal contacts, which were probably necessary for reliable refinement [58], also have been problematic in our simulations as they showed just limited refinement tendencies (TR567) or did not improve at all (TR576). Target TR517 was a specially difficult case and different compared to all

other refinement models. A significant number of residues from the mid-part of the model was removed so we were tempted to skip the model completely from all further investigations since a dominating part of interactions were missing in the simulations and hence no real but falsely biased statements whatsoever could be made by analyzing this data.

The starting model TR592 was too good to be targeted for backbone refinement, though we still were able to refine this model, and even score it, what possibly means, that this very careful approach was able to hold atoms of structures in place during a simulation, just crossing those energy barriers needed to simultaneously refine and improve the structure.

The general aspect of refinement underlined by the CASP competition is admittedly a bit biased towards finding only one best native model, which is normally not a realistic view of the energetic description of a native state. Clearly, there is just one global energy minimum, but this is not the only solution to the problem. The energies around a distinct native state can be thought of as a near native ensemble itself. On the other hand, with conformational differences that come along only slight total energy variations can be very substantial and may introduce huge structural changes.

Thus we instead did not focus on refinement for special purposes, as we wanted to develop a methodology that is applicable in most refinement scenarios. We therefore neglected all additional information for the refinement target given by CASP, and just focussed on the structures. This helped us to develop a workflow without or with less necessity for prior knowledge about the protein. Of course this can only be done when the systems inaccuracy is not harmful to the setup of the model straight away. If thinking of the abilities of creating realistic, nature mimicking molecular dynamics setups, those systems will always be devoid of details. So we argue, an all-atom simulation may be as unrealistic as a coarse grained simulation approach since it also misses a lot of potentially needed interactions. But it is as far as modern systems can get right now, and, most importantly, the approach modifies the system to become a self controlled self maintained all-atom simulation due to its coupled C_α atoms, that can gain detail from the simultaneous averaging.

Regarding the energies that drive the simulations, our models will be augmented with this cooperative effect, but indeed, a missing factor like an important crystal contact,

can have devastating consequences for the ability to synthesize a native conformation within a MD simulation. All motions of the system depend not only on the protein interactions but also highly on the protein-water interface. Therefore explicit solvent mitigates the absence of normally important antagonist effects on the targeted structure. Inevitably, time is a worsening factor in simulations with lacking components which are not systematical. The longer a simulation lasts, with a missing interaction on specific parts of the protein, the resulting conformation will be more and more incorrect over time.

CASP refinement for us was successful for some targets. Especially TR557, TR569, TR592 and TR622. The most impressive refinement was achieved with model TR557 with around 0.17 GDT-TS points and model TR592 with around 0.05 GDT-TS points. The former is remarkable not because of its sheer refinement but because it started from an already very accurate structure. This is a hint towards the capability of MD simulations to refine structures normally very susceptible to very small errors.

On total average, our CASP results were ranked slightly above the *null* refinement, which sounds small and limited but compared to the competitors this was among the best scoring results. Although we were not able to rank our five submitted models correctly, aligning the quality of the five structures per model with the correct label (1 to 5) is as hard as scoring an ensemble of near native structures. Hence, no group was able to put structures in the right order.

This leads to the question of the improvements since the last CASP refinement and the future perspective for upcoming events and the general proposition for refinement in the light of improving structure prediction and modeling.

Under the light of the different refinement approaches of all the teams that participated at CASP, all of them had their weaknesses and disadvantages but also strengths. MolProbity scores, for example, were among the most refined ones. So if model builders could integrate a refinement tailored to this score (for example by only allowing highly accurate side-chain rotameres within models), refinement itself will be more challenging because it already reached a maximum improvement. But while improving modeling methods sounds easy when demonstrating the possible goal to achieve, it might be very hard to actually work out and implement a strict rule into a refinement workflow, since every protein is different and needs special treatment.

Exactly this is the point that makes refining of proteins difficult and finding generalities over a wide spectrum of models really hard. Whenever a scheme or treatment works fine for one protein, it might completely crash another.

Chapter 4

Adaptive Position Restraints

The structure of a protein is important for understanding its function [6, 59–61]. Protein model precision is furthermore important when working with the structures, for example in computer aided simulations. Methods that use and work on structural data ideally start with the most detailed atomistic description available to be as realistic as one can possibly be to produce reliable results.

Predicting models de novo, without using an homologous template only by using the protein sequence via so called *ab initio* protein folding methods is attempted for decades [15, 45, 62–67] but unfortunately not yet ultimately successful (at least for bigger proteins) because its reliability of producing near native, high quality models is still limited. In general, models can be very wrong and hence not intended to be used in subsequent workflows yet. Template free modeling is still considered the "Holy Grail" of protein structure prediction.

Usually, very good models are produced by homology modeling [62, 64, 68, 69] methods, when the sequence identity of the target sequence and the sequence of the homologous structure is above about 30 %. This approach borrows the known structure and imposes it on the unknown model. Those models might still have erroneous regions, but the overall fold of the structure will most probably be correct.

If the identity between the compared sequences drops below 30 %, the fraction of incorrectly placed atoms rises quickly. Sources of error in homology modeling are highly flexible regions, for example not well described loops or unstructured coiled sites. But also conserved regions, mostly in the hydrophobic core of a protein can sometimes incorporate severe misfolded chains. Fortunately nowadays, thanks to the structure solving community, there is a huge number of structures available to pick high identity sequences from, albeit at the same time, the gap between known sequences and known structures gets bigger every day (see Fig. 1.1).

Starting with such a scenario, refinement methods aim to achieve a better atomic placement, closer to the native state. A number of methods exist that attempt to refine given targets without knowledge of the correct structure, but until now there is no simple straightforward way to accomplish that task consistently to a satisfying level. Methods range from Langevin Dynamics, Brownian Dynamics, Monte Carlo implementations of all kind, energy minimizations with knowledge based and physics based potentials, to, underestimated but increasingly used, molecular dynamics (MD) simulations [36, 70, 71] including simulated annealing, either free [36, 72] or with restrictions or enhancements which all try to extent the sampling to avoid being trapped in local energy basins. [73, 74]. Up to now it is still very unclear which properties do define the success or failure of a refinement method. With this work, we think we can shed some light onto the process of refinement, to pave the way for consistent, reliable improvement of homology models and contribute to deletion of the notion, that attempted refinement of a homology model usually deteriorates a structure more then it will improve it.

MD simulations are a very powerful tool to describe atomic motions with classical Newtonian mechanics [70, 71]. When the underlying force fields are exact enough [9, 75–78] it should in general be possible to mimic natural atomic behavior as a function of time. Thus, when simulating long enough, it is possible to observe real folding events. Regrettably, just fractions of the time normally needed to observe folding can be simulated today, at least when larger proteins or protein complexes with interesting and important biological functions are addressed. Additionally, when prolonging [72] simulations, more and more precise setups are needed [78] because every small detail (salt concentration, pH-value, aiding molecules etc.) can ultimately have a big impact on the sampled conformational folding space [79]. Refinement consequently can be very susceptible to small errors and lead to completely misfolded molecules, impossible to correct.

That means, despite the fact that refinement usually starts with a well defined near native state (or at least closer than a completely unfolded chain of amino acids), successfully reaching an elusive global minimum can be evenly difficult as folding an entire protein. Likewise, being trapped in a local energy minimum can also render refinement impossible.

To overcome unavoidable, with MD usually uncrossable energy barriers [80], we tested simulation setups augmented with a novel positional restraints, which may lower energy barriers and finally lead to substantially refined models, contributing like the other enhanced sampling methods mentioned above. Our approach can be described as locally adaptable position restraints method. Our adapted positional restraint procedure comprises permanently updated per-atom potential which, due to the updates, will adapt its local minimum as a function of time. Updates will be guided by the direction that the atom takes while it is moving within its potential. Since it is still influenced by all the other atoms in the simulation, the potential helps to stay within a reached minimum for at least up to one update interval until it is then again deformed iteratively to finally reach a more native-like, settled state for a maximum amount of atoms. This process is self directed and guided by the internal forces of the molecule.

To generate the structural ensemble, we performed three sets of MD simulations: First we carried out a fixed temperature restrained refinement simulation with all models. Then we cross checked the results with unrestrained, free simulations. Finally we tried to investigate the effect of a simulated annealing restraint simulation on the initial refinement approach, for which we hoped that it could widen up the possibly narrowed conformational space of the fixed temperature refinement setup.

In refinement simulations, the actual improvement, if at all, usually does not happen linearly, but will take a chaotic path, caused by the high dimensionality and complexity of the addressed problem [81–86]. Accordingly, simply picking the last structure from a trajectory will not automatically yield an improved structure [36]. To account for this, in addition to the refinement method, we developed an associated scoring function, which ideally would be able to filter the generated structures, separating the improved from the worsened structures. The scoring method assesses a combination of local and global compactness, so that regional- and molecule-wide distances reflect the quality of a model.

4.1 Methods

4.1.1 Restraints approach

We are introducing a deformable elastic network (DEN) [87] inspired positional restraint to MD systems, which is called adaptive deformable position restraint (ADPt). With this approach we are able to run a position restrained inspired simulation that is able to allow changes in all possible directions. Simultaneously it takes advantage of the notion, that simulations, coupled to an additional potential behave differently and alter the original energy function. Whether dynamics can now overcome energy barriers that would normally hinder simulations to improve, resulting in more efficient sampling, will be shown below.

An ADPt-enhanced simulation setup starts with applying position restraints to all C_α atoms with harmonic potentials of the form $e_{pos} = \sum_{i=1}^{3} m/2(x_i - X_i)^2$, with the energy e_{pos} and x being the position of the simulated atom and X is the location of the position restraint. All other atoms are not restrained in this approach, but the method would also allow all atom restraints. While a position restrained atom moves within the boundaries of its potential, after a predefined fixed update interval, we update the position restraint coordinate to the new location (see Fig. 4.1):

$$X_i^{\text{new}} = X_i^{\text{old}} + \kappa(x_i - X_i^{\text{old}}) \tag{4.1}$$

This update takes place at a fixed frequency, typically every 500 integration time steps.

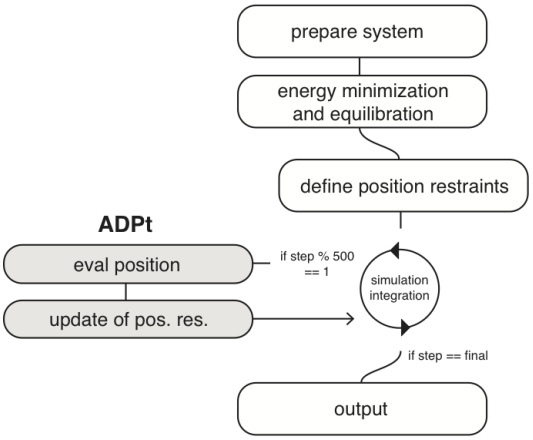

Figure 4.1: The adaptive deformable position restraint (ADPt) workflow schematically.

As a result we achieve, that atoms which are modeled very accurately near to their native position within their local energy minimum tend to stay where they are. Their movement within their potential is zero in average and hence the update will not change its position. In contrast, atoms which are far away from their native location tend to move away from their initial position, not affecting those atoms that are placed ideally, assuming they do not move too far away. With this approach we do have the advantage of, first, having position restrained atoms which do not leave their dedicated area because their movement is zero on average (for example core region of a protein), while at the same time, second, others freely move where ever the simulation force field directs them, both combined in one system.

4.2 Results

4.2.1 Monte Carlo simulation

To study the effect of the ADPt approach in which a particle feels a slowly adapted, updated harmonic potential, we used a MC simulation with one particle and a one-dimensional energy function.

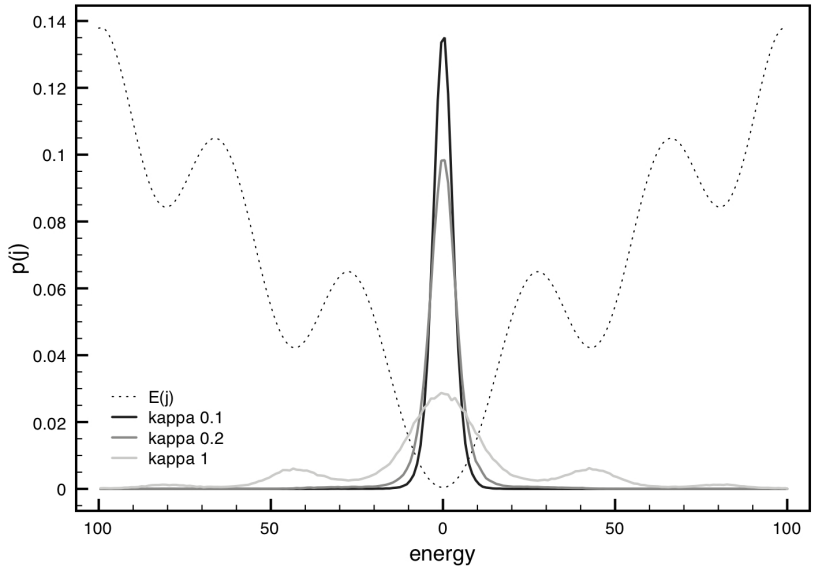

Figure 4.2: Monte Carlo simulation for a one-dimensional ADPt particle. κ is the rate of how fast the particle follows its potential. $E(j)$ is the energy of the particle at position j, $p(j)$ are the probabilities. In conclusion, the lower κ is set, the more the particle can be seen in the lowest energy state.

Simulations with three different values of κ, the update strength, 1, 0.2 and 0.1, were performed and the results are depicted in Fig. 4.2. The curves show impressively what effect this technique can have in a very simple setup. The black line for $\kappa = 0.1$ shows that the probabilities for being in the lowest energy state are highest of all settings. All higher κ-values show a wider probability distribution with more populated higher energy states. At last, for $\kappa = 1$, the particle behaves like an unrestrained normal single particle, with higher occupation rates of higher energy states.

4.2.2 Molecular dynamics setup

The complete ADPt approach is implemented in Gromacs 4.5.3 [39]. To be also suitable for large systems above about 5.0×10^5 atoms the implementation takes advantage of the domain decomposition functions within the main computation loop.

We performed canonical NVT ensemble explicit solvent simulations with the water model tip3p [49] in conjunction with the Amber 99SB-ILDN force field [75]. The duration of each simulation was 100 ns using an integration step size of 2 fs. For electrostatics calculations we used the Particle Mesh Ewald (PME) method [88, 89] with a temperature of 300 K controlled by the Nosé-Hoover temperature coupling [90, 91] (simulated annealing see below). For keeping the constraints and maintaining the overall structure over the decomposed computing cells we used P-LINCS [92, 93], the parallel implementation of the linear constraint solver unless mentioned otherwise. The restrained simulations (including the simulated annealing) position restraint force constant was 100 kJ mol^{-1} nm^{-2} (multiple different values have been tested, ranging from 10 to 1.0×10^6 kJ mol^{-1} nm^{-2}) with an ADPt adaptation parameter κ set to a value of 0.2 (in agreement with our Monte Carlo coupling simulations and various test runs) and an update frequency of 500 steps. κ is designed to allow adjustment of how fast the restraint potential follows its designated atom. For the free simulations we used the same setup but this time without any restraints. The simulated annealing simulations were started with 300 K, then heated until they reached 600 K after 0.6 ns. Then we cooled down the system to 400 K until reaching 1 ns, and afterwards finally slowly letting it reach 300 K after 100 ns. All simulation systems were standardly energy minimized with the steepest descent algorithm and then equilibrated for 100 ps.

We generated 100 ns of fixed temperature restrained simulation for all protein models and added another 100 ns of free, unrestrained simulations for cross checking and simulated annealing simulations for each model, respectively. Altogether, we produced 5.4 μs of simulation data. On our cluster a single simulation needed 2-6 days of core time per model, depending on the number of atoms involved, while using an average amount of 96 Intel® XEON® X5670 cpu cores per run.

4.2.3 Model selection and preparation

We selected 18 models from the refinement section of the international blind test meeting Critical Assessment of methods of protein Structure Prediction (CASP) [1] 8 (model id: 389 429 432 435 453 454 461 488) and CASP 9 (model id: 530 567 568 574 576 592 594 606 622 624) and excluded all NMR models (CASP8 model id: 462 464 469 476, CASP9 model id: 557 569 [614 no pdb.org data found]) for consistency reasons.

Some models were not fully described by their pdb data entries and we had some problems aligning a few templates with their targets. In detail, model TR389 had one more residue (LYS 135) than the crystal target, which in fact was not problematic for fitting procedures, since for all comparison calculations it was simply excluded. The crystal structure of target T0429 incorporated a gap from residues 56 to 72. Model T0435 also had a gap from residue 62 to 71 and missed 3 residues at the end. As all our measurement tools included a sequence alignment prior to the actual comparison, gaps were automatically skipped in both fitted structures.

Table 4.1 gives an overview of the models and their most important properties.

The model with the lowest initial RMSD (1.26 Å) was TR592, the one with the highest (7.47 Å) was TR622, both from CASP9. Very near native structures are included in the set (below 2 Å RMSD: TR432, TR453, TR461, TR530, TR592, TR594) as well as structures rather far away from their native state (above 6 Å RMSD: TR492, TR568, TR576, TR622).

The shape and size of the models are in the range from all helical (TR432, TR454) to all beta sheet (TR624), 69 residues as lowest (TR624) and 192 as highest (TR454) value (see Fig. 4.3).

MODEL (*-id)	♯ atoms C_α/all	RMSD (Å)	GDT-TS	dRMSD (Å)	all atom dRMSD (Å)
TR389	135/2136	2.638	0.8097	1.9643	4.3216
TR429	155/2510	6.796	0.4457	5.9684	7.0900
TR432	130/2159	1.646	0.9173	1.0505	2.1450
TR435	137/2199	2.153	0.8223	1.3952	2.4105
TR453	87 /1386	1.396	0.8879	1.0548	3.1475
TR454	192/2966	3.238	0.6406	2.7201	3.0578
TR461	157/2432	1.634	0.9029	1.1657	2.0900
TR488	95 /1471	2.109	0.8789	1.3405	1.9029
TR530	80 /1288	1.990	0.8594	1.3197	3.5878
TR567	142/2261	3.435	0.7817	2.2811	7.5075
TR568	97 /1541	6.149	0.5490	4.4619	4.9824
TR574	102/1515	3.583	0.6201	2.5333	2.8981
TR576	138/2197	6.851	0.6431	4.7643	5.1740
TR592	105/1635	1.257	0.9024	1.0310	1.3810
TR594	140/2258	1.818	0.8661	1.2569	1.8787
TR606	123/1894	4.850	0.7175	3.4614	4.1534
TR622	122/1996	7.474	0.6680	4.0014	3.5873
TR624	69 /1118	5.189	0.5543	3.1542	4.3264

Table 4.1: Initial model overview. All models were compared to the corresponding crystal structure with different distance metrics. * ♯ given by the assessors of CASP.

4.2.4 Assessment of model quality

To assess the quantity and quality of the refinement approach we compared and applied several measurement methods, since deeper understanding of the refinement processes demanded for a slightly wider variety of methods.

Normally, structures of the same type, the same reference frame and almost the same subset of comparable atoms, but with a (moderate) different spatial arrangement of atoms, can be compared and assessed reasonably well with the root mean square deviation (RMSD) method. With a RMSD, usually all C_α atoms of two models are superimposed so that the root of the mean squares of positional differences of the atoms give a value of the overall similarity. The lower the value the more similar is the spatial arrangement of the two sets of atoms that are being compared. Since this method is well established, the value normally gives a good impression of the quality of a model. The disadvantage of this method is the unavoidable necessity of a superposition of the models. This is normally done via a least square quadratic fit (LSQ) of the two structures. Though this approach is now widespread and routinely used, it is not very trivial to implement. Another shortcoming is that all atoms are weighted equally. Ideally, flexible regions of a protein model (e.g. loops or N/C-terminal regions) should be weighted less because they can impair an otherwise perfect matching structure score. By this,

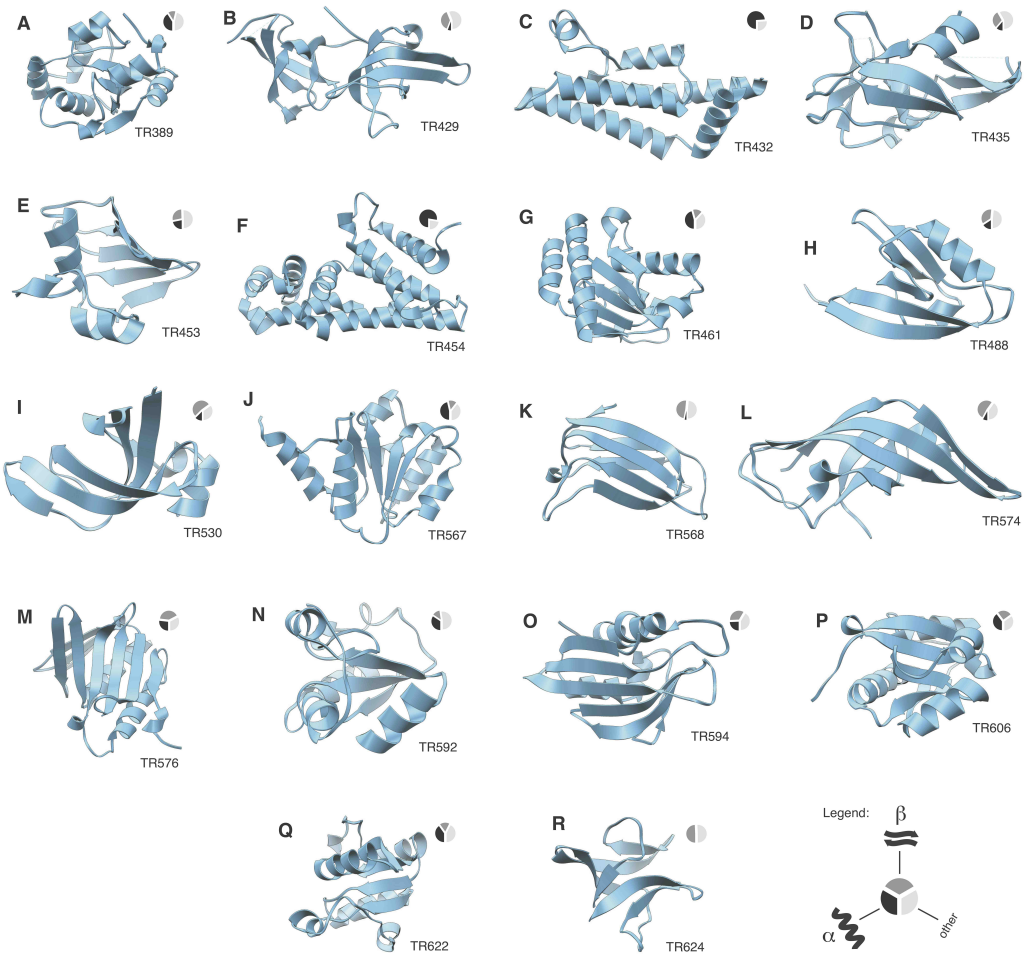

Figure 4.3: Overview of models used in this refinement test. A-H: CASP8, I-R: CASP9. Also included is a secondary structure icon, showing the percentage of α-helix or β-sheet content of a model, respectively.

unfortunately, a lot of information is hidden or simply lost within the single-value score of the RMSD (which we address in Fig. 4.9, where we use a ranged all atom version of our distance RMSD function).

Consequently, our main quality measure is a distance based RMSD (dRMSD) (see Fig. 4.4), which measures internal distance differences of the two compared models. Unless otherwise mentioned we just use the backbone C_α - C_α distances to compute the score.

$$\text{dRMSD} = \sqrt{\frac{1}{n^2 - (\frac{n^2+n}{2})} \sum_{i=1}^{n} \sum_{j=i+1}^{n} \left(d_{ij}^{(t)} - d_{ij}^{(m)} \right)^2} \tag{4.2}$$

with $d_{ij}^{(t)}$ and $d_{ij}^{(m)}$, which are the distances between atom i and j in the target structure and the model, respectively.

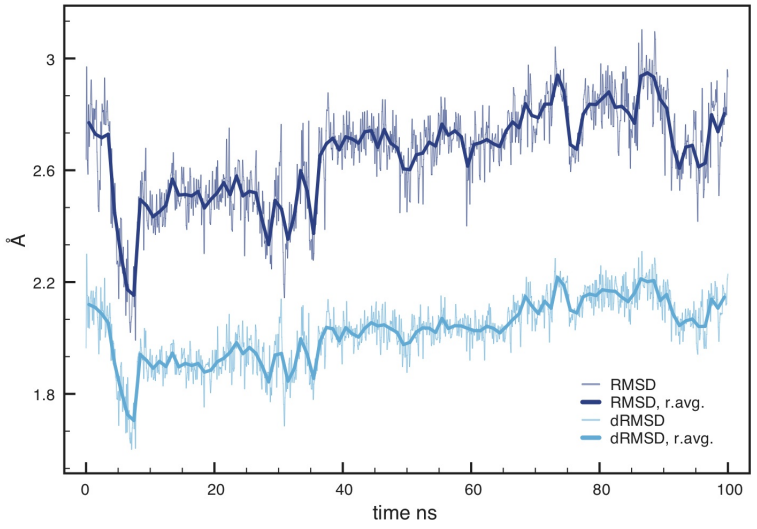

Figure 4.4: Comparison of dRMSD and RMSD. The plot shows the unfitted and unnormalized results of the RMSD and the dRMSD calculations applied to a protein model. The overall trace of the values are similar, the details are different, as this is a complete different value: dRMSD compares distance differences and not positional differences like RMSD.

Structures evaluated with the dRMSD method do not need to be aligned since we also do a sequence alignment prior to the distance difference calculation. Accordingly it can be used to show distance specific clustered data of the investigated model, where clustered here means focussing just on specific ranges of distances within the model to depict structural changes that happen during simulations.

Another similarity measure we use is the Global Distance Test (GDT) [51]. This measure tries to find the largest group of atoms that can be aligned with an RMSD of better than a certain threshold.

This measure therefore focusses on regions in the protein that form the correct core and completely ignores outliers. It is an iterative method which takes a small number of consecutive C_α atoms and executes a least-squares (LSQ) fit with those regions, measuring four predefined contributing maximum thresholds and looping through that process until the end of the chain is reached. The GDT Total Score (GDT-TS), which we used, takes (1, 2, 4, and 8) Å as those RMSD thresholds. Unfortunately, this sort of structure quality measurement is not ideal when trying to correlate it with potential energy values from simulations or qualitative rankings like our compact score, because it is intrinsically unstable regarding the score cutoffs. Since the score disregards parts of very misplaced structural elements, the real quality of a structure is not reflected in the absolute score, or at least not as accurate as a direct RMSD method would be.

4.2.5 Scoring method

Generally, the correct structure is not available to determine the quality of a model. It is therefore necessary to have a scoring that, given an ensemble of models, quantifies which model is better, i.e. which model is closer to the true answer.

This scoring could be using energies, structural properties or a combination of multiple variables of the model and the resulting scoring measure should be closely correlated to the real quality of the model. A template-free scoring should ideally be stable, meaning it can score all kinds of proteins, no matter how they are composed, generated or treated, without loss of generality. Unfortunately, no scoring method can be successful in all cases. There will always be tradeoffs specific scoring approaches have to face. For example, taking just the energies of all-atom explicit solvent simulations will usually not yield a good scoring measure (except regional coincidently) since it will typically be dominated by large fluctuations arising from the multitude of interactions from particles within the simulation (mostly water). To create a very low level and highly general scoring function we combined two compactness scores, a local and a global one. The local compactness score is focussed on local distances of atoms being within a certain range within of the model.

The compactness score is calculated as averages over atomic C_α-C_α-distances d_i by

$$\text{if } d_i <= 6.7 \text{ Å }, \mathrm{E}_{\text{local}} = \frac{1}{n}\sum_{i=1}^{n} d_i \tag{4.3}$$

$$\mathrm{E}_{\text{global}} = \frac{1}{N}\sum_{i=1}^{N} d_i \tag{4.4}$$

$$\mathrm{E}_{\text{compact}} = \frac{\mathrm{E}_{\text{local}} \times \mathrm{E}_{\text{global}}}{2} \tag{4.5}$$

where N is the number of all C_α atoms, and $n =$ is the fraction of distances that obey the local distance criterion.

The global score reveals a mean overall model compactness. The resulting compact score (Eq. 4.5) is composed of the local score (Eq. 4.3), weighted by the global compactness (Eq. 4.4).

MODEL	♯ residues	♯ decoys	min / max RMSD
1ctf	68	630	1.319 / 9.071
1r69	63	675	0.876 / 8.311
1sn3	65	659	1.310 / 9.134
2cro	65	674	0.806 / 8.311
3icb	75	653	0.945 / 9.391
4pti	58	688	1.414 / 9.265
4rxn	54	678	1.356 / 8.140

Table 4.2: Overview of 4-state decoy set by Park and Levitt [94]. The list shows the model identifier, the number of residues of the decoy, the number of structures within a set and the minimum and maximum RMSD values of the decoys in the set compared to the native model.

The score as it is used here is an unbiased, geometric measure, which does not need to be trained or parameterized in any way.

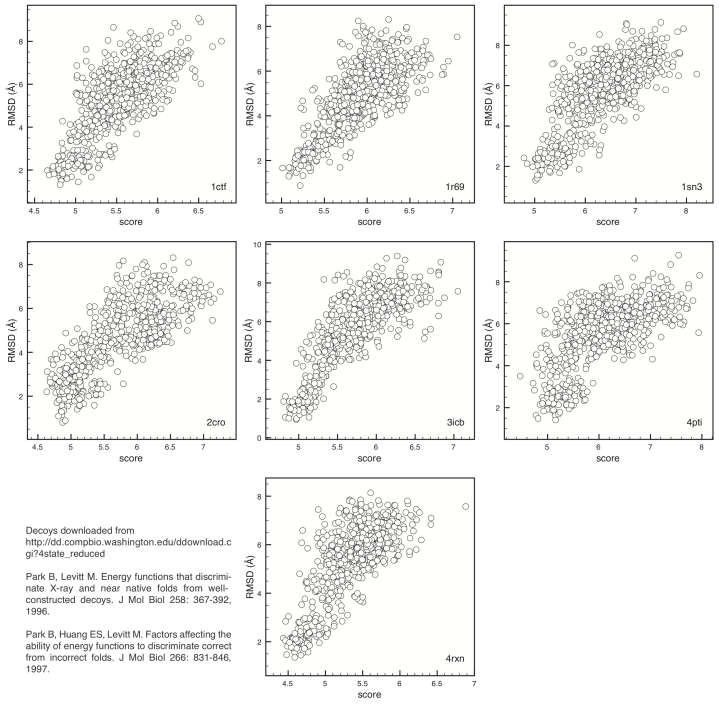

Figure 4.5: Assessment of the compactness score, which was used to rank all models in the ensemble of decoys. Each circle represents a decoy structure [94, 95]. The high correlation of the score to the RMSD shows the ability of the score to depict the real quality of the investigated decoys. Pearson correlation coefficients of each set: $r_{1ctf} = 0.69$, $r_{1r69} = 0.67$, $r_{1sn3} = 0.71$, $r_{2cro} = 0.76$, $r_{3icb} = 0.77$, $r_{4pti} = 0.64$, $r_{4rxn} = 0.70$.

To neutrally assess the quality of the score, we took the Park/Levitt decoy set [94] (4-state reduced), including all seven models (see table 4.2) and measured them with our scoring function.

The result of the test can be seen in Fig. 4.5. It clearly shows, that it can discriminate most near native structures from the worse consistently throughout the complete set of

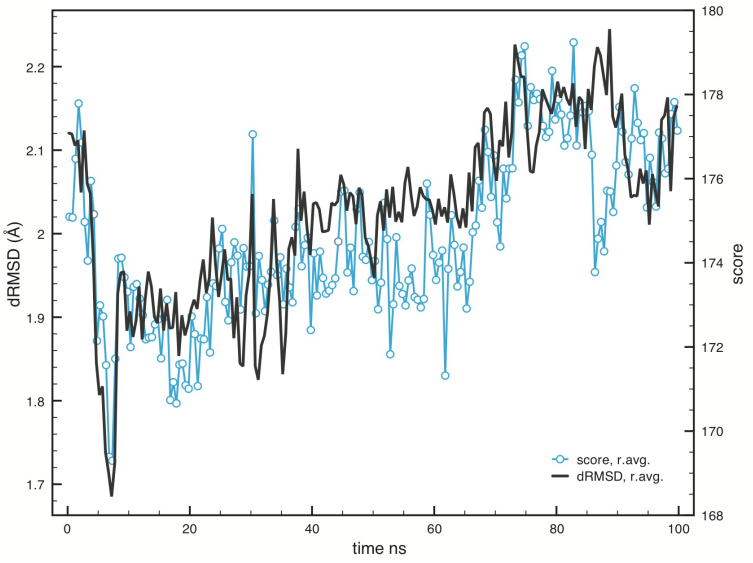

Figure 4.6: Scoring of model TR389 in a fixed temperature ADPt MD simulation. Clearly observable is the nearness of the score (blue) to the distance of the model to the crystal structure (dRMSD, black) in this case.

decoys.

How close our scoring function can be in comparison to the dRMSD quality measurement of a model trajectory (TR398 of our set) can be seen in Fig. 4.6. The dRMSD curve resembles very closely in the curve of the compact score values. So the score really guides to the most native structure of the ensemble. This is a very accurate example (r= 0.7) which has to be considered as an ideal case and may not be generalized.

4.2.6 Simulation results

The ADPt approach has been applied to 18 models from CASP8 [1, 96–100] and CASP9 [58, 101] as described in the method section. From all 18 models in this test set, only model TR622 could not be refined with our approach, even for the annealing simulation approach. Two simulations even produced only improved structures, meaning at each time in the refinement trajectory the model was better than the starting model.

To explore, investigate and assess the generated ensembles, we evaluated simulation properties like potential energies, coulombic and Lennard-Jones short range and 1-4 protein-protein interaction energies, Ramachandran energies, amount of secondary structure elements (SSE) (α-helices, β-sheets), model size (residues), best refinement frame, score correlations, frame improvement, starting dRMSD and sampled conformational

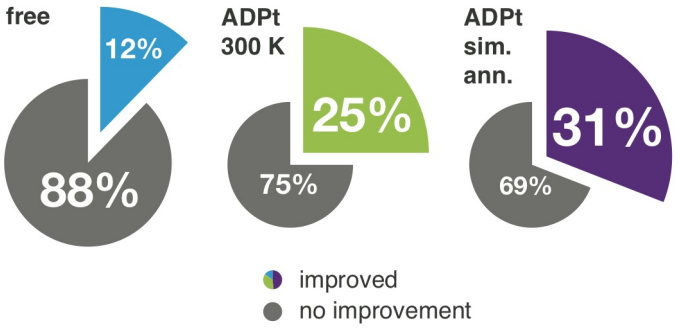

Figure 4.7: Comparison of the total improvement of the three different simulation setups when considering all frames produced. The improvement capability increases with the ADPt (middle) - and the simulated annealing ADPt (right) approach.

space, for which best agreements towards the real quality of the structures we only found in our compact score.

Overall, standard unrestrained MD simulations produced 12 % improved frames, fixed temperature ADPt restrained simulations improved 25 % frames and simulated annealing ADPt simulations were able to further increase that by 6 %, ending up with 31 % of refinement (see Fig. 4.7).

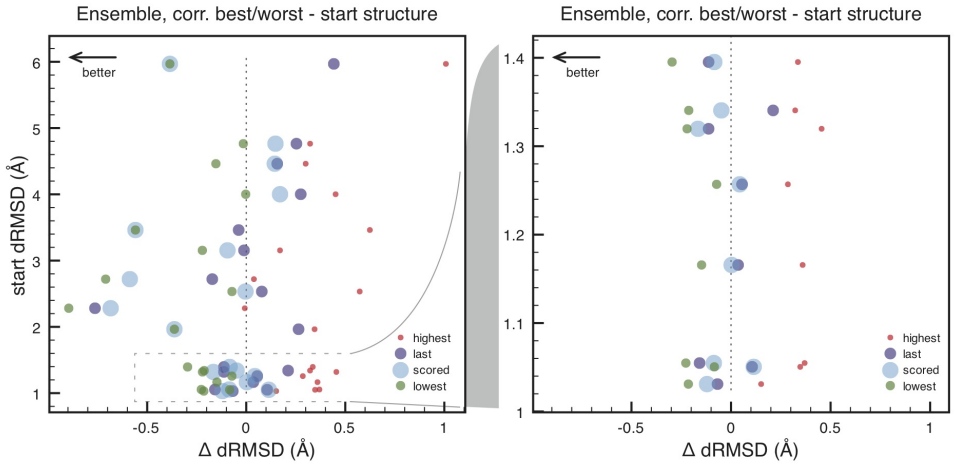

Figure 4.8: Correlation of the Δ dRMSD of the best (green), scored (light blue), last [meaning the last structure of a simulation as a function of time] (purple), and worst (red) structures and their starting dRMSD of the 300 K simulation, respectively. Horizontally, each dot with the same start dRMSD belongs to the same ensemble. The dotted vertical *null* line means no change, i.e. a *null* refinement. Each dot to the left of the *null* line is a refined model, each dot to the right is a deteriorated model. Models with starting dRMSDs between 1.8 and 3.8 Å (TR606, TR454, TR567) were refined the most, indicating that neither the farther away, nor the very close models from or to the native structure showed the best improvement.
Nevertheless this could suggest that models of intermediate quality, with a certain degree of correctness can be improved until a limit, defined by the incorporated level of detail, either by the limitations of the force field or by the sampled conformational landscape, is reached.

Focussed on the properties of the models from the fixed temperature ADPt runs, did the starting quality of the homology model have any impact on the ability to refine it (see Fig. 4.8)?

Those models in our set with a starting dRMSD value below 1.4 Å showed constant improvement of 0.05 to 0.4 Å dRMSD for the best model produced. The scoring function was able to pick a model out of the better lower third part of the ensemble, which was almost all the time better than simply picking the last produced structure from the trajectory. No trend was observable within this range of starting quality. Above 1.4 Å, beginning with the first model slightly above 1.9 Å, but below 4 Å, first of all the sampled space and the amount of improvement doubled, ranging from 0.1 to 0.9 Å. Here, the scoring function also was better than simply picking the last structure, except for one example. Two out of six models in this range were perfectly scored, meaning that the best structure produced was actually picked. Above 4 Å, the models showed very little overall improvement, with 0.4 Å improvement for TR429 as the best performance, which was also picked by the scoring function. All scored models were better than the last structure of the corresponding runs.

Models in the range of below 4 Å and above 1.9 Å dRMSD starting quality were refined the most. Below 1.4 Å and above 4 Å dRMSD, there were limitations, possibly founded in the lack of fine grained near native sampling for the first and too much error introducing models with no guiding path to near nativeness for the latter. A balanced structural error, combined with a fundamental overall correctness seems to be helpful when trying to refine a comparative model. To sum up, refinement performance in our simulation is lower either, when the model is too far away from nativeness (above 6 Å RMSD) or too close in the very near native region (below 1.4 Å RMSD). Expansion of the runtime from 100 ns to 200 ns can possibly lead to better performance for the models more distant from the native state, as can be seen in Fig. 4.18, and the simulated annealing ADPt results in Fig. 4.13.

With eight of the models (TR389, TR429, TR530, TR568, TR576, TR594, TR622 and TR624), the simulated annealing simulations were able to enhance the refinement significantly. The free runs in most cases just expanded the sampling space without increasing the chances to visit near native structures.

Whenever we expanded the conformational space and sampled structures being more distant from the native state, either with simulated annealing (just TR568) or with free runs (all but TR568), in 12 of 18 cases we were able to distinguish them with our scoring function, discriminating those as deteriorated high energy structures (see Fig. 4.13, **A**, **B**, **E**, **H**, **I**, **J** partly, **K**, **L**, **M**, **N**, **O**, **P**).

A detailed look at the refined models with our all-atom dRMSD assessment score should give a closer insight into which of the internal distances were actually affected by our refinement. The question of whether a general deterioration arises from the overall shape impairment or a drift in short-range interactions of a structure needs to be answered, as the outcome could give rise to the possible strengths and weaknesses of MD refinement or MD simulations at all. Unfortunately there is no general trend observable when measuring the models in finer ranges explicitly, which is also good on the one hand, because that means there is no systematic error or bias within our MD refinement approach. Nevertheless, the examples we studied below will illustrate that MD simulations are not limited in their refinement spectrum.

The detailed measurement was done with the all-atom version of the dRMSD method, in which multiple chosen subsets of intra-distances were taken into the distance difference calculation. We divided all intra-distances of the crystal structure into ranges of 0-4, 4-8, 8-12, 12-16, 16-20 and 20-max Å and compared those with the corresponding set of distances of the model. Each comparison yields a score, not observable in a general C_α RMSD or dRMSD, GDT-TS or GDT-HA (high accuracy).

The most prominent result of that score is shown in Fig. 4.9, which shows that our simulations are able to maintain the overall shape of a protein and just refine local interactions (see Fig. 4.9, **B**, light blue and green line) as well as refine the global shape and keeping local interactions stable (see Fig. 4.9, **A**, red line). Compared to the free simulations, this is a unique feature of our ADPt setup.

In summary, the simulated annealing approach resulted in an extra improvement of 8 models compared to the ADPt setup without simulated annealing. Additionally, drastically reduction of noise (lowered dRMSD variance as a function of time per model) during a simulation is observed in the fixed temperature and the simulated annealing restrained setup.

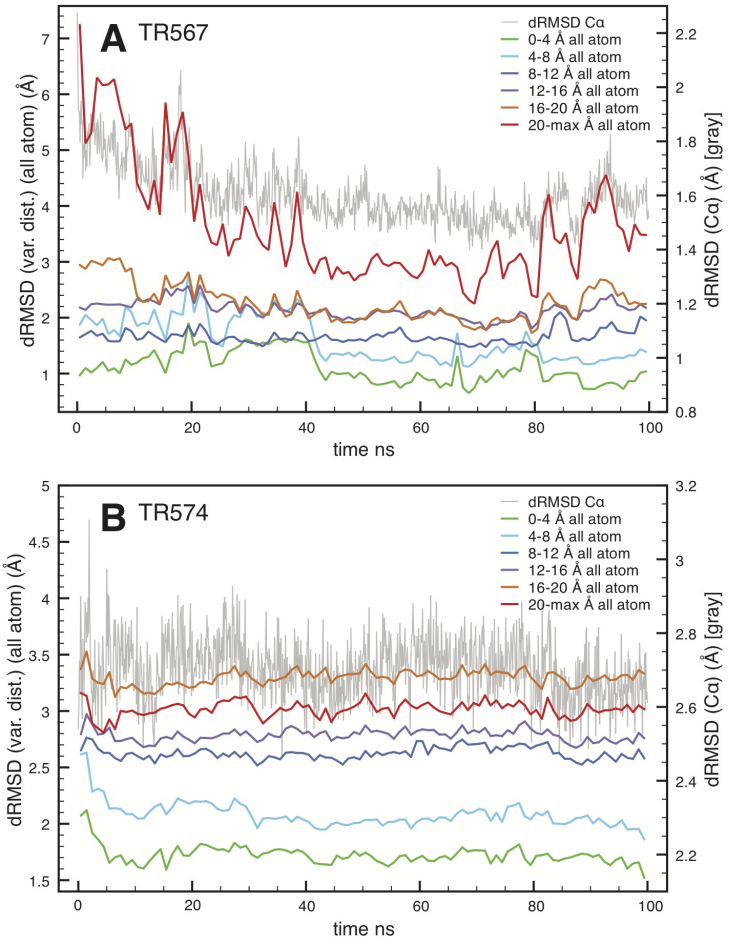

Figure 4.9: Refinement results analyzed for different internal distances. The dRMSD is plotted for C_α atoms of each frame against its crystal structure (right y-axis). The other colored lines show a windowed all-atom dRMSD for a specific distance. For example for 0-4 Å(green line), we first evaluate the distance between two atoms and then take them into the calculation only when the distance is at least 0 and at most 4 Å. This is done for all listed distances (left y-axis). By this it is possible to depict the quality of a refinement and to detect which distance-regime of the model could be improved most during a simulation.
A TR567: The red line shows that the refinement happened almost exclusively at the longer distances, indicating that the overall shape of the model was corrected. It started with an all-atom dRMSD of 7.2418 Å and followed a continuous drop until it reached 1.7198 Å after about 70 ns simulation time (Not observable in this plot; curves are running averages). Then it deteriorated a bit, which is also observable in the overall dRMSD. Also noteworthy is the fact, that the improvement in the long distances did not cause a decline of quality of all other distance ranges. **B** TR574: In contrast to the overall shape refinement of model TR567, we can see in this example, that the small impact to the refinement for this model came from the less distant atoms from 0 to 8 Å (green and light blue line), indicating a better side chain packing.

Consequently, in contrast to the restrained simulations, free simulations behaved much more chaotic per time, since it was simply producing worsened structures, in a complete undirected manner with low chances of yielding improved structures. The chance of deteriorating the model simply was very high. All in all, there was only one free simulation (TR568), that did not produce the worst structures from a combined ensemble of all runs for a model.

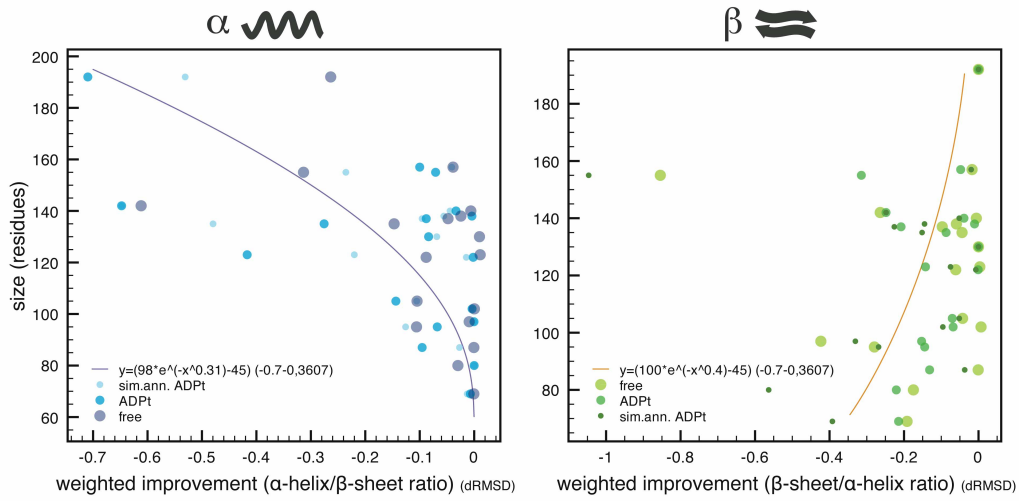

Figure 4.10: Improvement, weighted with the amount of α-helices/β-sheet rate, against the size of the model. It is observable here that the improvement of a model is related to the α-helix content and its size (residues). The bigger a model and the more α-helices it has, the better was the refinement. On the other hand, there also exists a (weaker but also observable) relation between the improvement, the size and the amount of β-sheets. The smaller the model, together with increased amounts of β-sheet content, the better was the refinement. An exception to that rule was one outlier (TR429) with 155 residues, which was large and showed a good refinement result, although it had more β-sheet than α-helix content. An explanation for that behavior could be its dimeric shape, though its difficult to explain, what exact impact it had on the refinement.

We found, that simulations were most successful when either models were around 120 to 200 residues large containing mainly α-helices (see Fig. 4.10) or smaller and containing more β-sheet content. This means big α-helical and small β-sheet models were affected the most in our refinements. This effect could either point to a force field bias towards stabilizing α-helices. Since this issue was addressed in the past [75] this can be neglected. By naive juxtaposition of helical and sheet secondary structure elements one can come to the conclusion, that the sheet feature is, most of the times, much more fragile because it involves loop motives and introduces a plane area which is difficult to predict because the exact twist of a β-sheet involves many long range interactions.

While the total amount of secondary structure elements of a structure during a simulation is usually not correlated with the RMSD of the model, the simulation should aim for reaching the secondary structure composition of the crystal structure. Fig. 4.11 shows, that we indeed achieve a more native-like distribution of secondary structure elements. The starting homology model secondary structure distribution was shifted, so that it missed essential parts of β-sheets and a few α-helices. In the end, the free simulation produces too much β-sheet content while not being able to accumulate enough α-helical content. The simulated annealing simulation had the same trend but was at

MODEL (*-id)	♯ atoms C_α	initial RMSD (Å)	end RMSD (Å)	initial GDT-TS	end GDT-TS	initial dRMSD (Å)	end dRMSD (Å)
TR389	135	2.638	2.934	0.8097	0.7537	1.9643	2.22852
TR429	155	6.796	7.258	0.4457	0.4293	5.9684	6.41063
TR432	130	1.646	1.877	0.9173	0.8962	1.0505	1.15654
TR435	137	2.153	**1.991**	0.8223	**0.8781**	1.3952	**1.28272**
TR453	87	1.396	**1.344**	0.8879	**0.9109**	1.0548	**0.89694**
TR454	192	3.238	3.287	0.6406	0.6406	2.7201	**2.54871**
TR461	157	1.634	1.771	0.9029	0.8965	1.1657	1.20115
TR488	95	2.109	2.404	0.8789	0.8763	1.3405	1.55187
TR530	80	1.990	**1.640**	0.8594	**0.8813**	1.3197	**1.20707**
TR567	142	3.435	**2.230**	0.7817	**0.7993**	2.2811	**1.5175**
TR568	97	6.149	6.625	0.5490	**0.5773**	4.4619	4.61816
TR574	102	3.583	**3.391**	0.6201	**0.6446**	2.5333	2.61181
TR576	138	6.851	7.677	0.6431	0.5688	4.7643	5.01773
TR592	105	1.257	1.263	0.9024	0.9000	1.0310	**0.96313**
TR594	140	1.818	1.997	0.8661	0.8500	1.2569	1.31224
TR606	123	4.850	**4.594**	0.7175	**0.7459**	3.4614	**3.42305**
TR622	122	7.474	7.689	0.6680	0.6373	4.0014	4.27649
TR624	69	5.189	5.296	0.5543	**0.5761**	3.1542	**3.14294**

Table 4.3: Comparison of crystal structures with the last structure from the refinement trajectory. This last structure is usually not the best structure. * ♯ given by the assessors of CASP.

least able to add some more α-helical content. The fixed temperature ADPt simulation was able to approach the native secondary structure content the best (see Fig. 4.11, **C**). For some models even simulated annealing was able to expand the sampling space and to lead to better RMSD values. A careful look regarding the native secondary structure distribution should be made.

As already mentioned and depicted in Fig. 4.12, TR429 (light gray rings) is a model hard to improve in two ways. First, it is composed of two intra model domains with a bridge-like interconnection. This link is very flexible and introduces difficulties to refine the model. Here, modeling the entire periodic crystal structure including crystal contacts could help to maintain the overall shape. However, in addition to that problem, one of the cores of this model is poorly modeled and therefore also very difficult to refine because it gives the simulation no real guidance, at least not with simulation times used here. So for the three different setups (ADPt, sim. ann., free), three different improvement results are observable. Expanding the conformational space did have a huge effect in the most efficient form in case of the simulated annealing setup. Because all three systems behaved quite differently, we can imply, that the underlying energy landscape of this model is in the starting region of 6.8 Å more like a flat golf course, far away from a possible direct refinement funnel path to the global minimum.

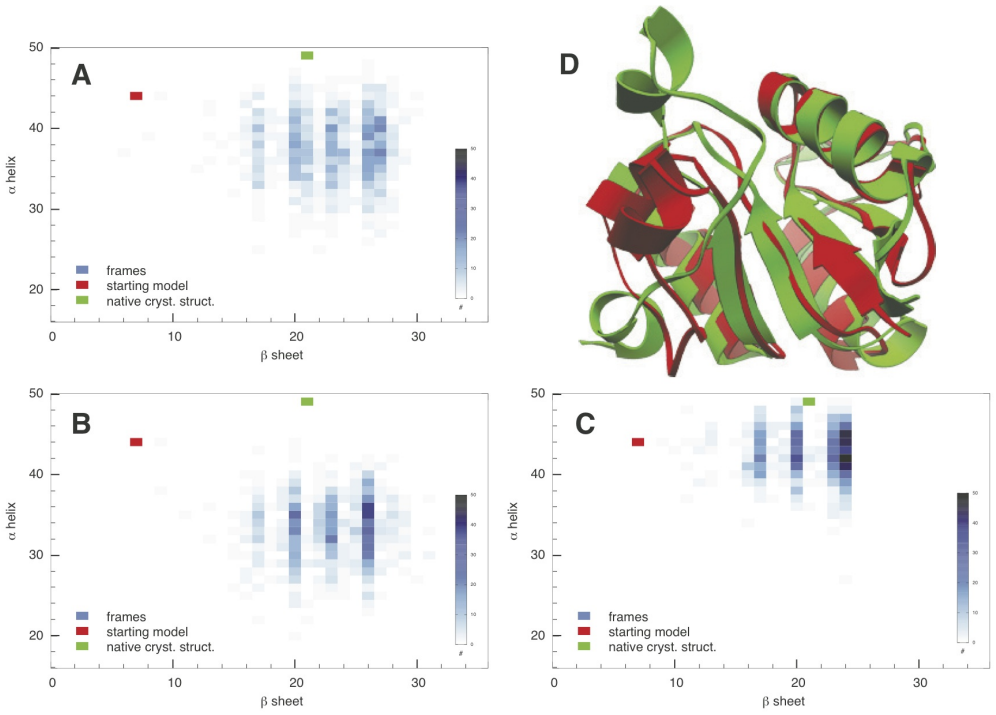

Figure 4.11: Distribution of secondary structure during the MD simulation. The darker the color the more structures from the refinement trajectory fall into this range of α-helix and β-sheet content. Red and green squares mark the starting and native composition of secondary structure of that model, respectively. The goal should be to reach the area of the green square with the native amount of α-helices and β-sheets. **A** Simulated annealing of TR622. **B** Free MD of TR622. **C** ADPt fixed temperature MD. **D** Crystal structure **green** and homology model **red** of model TR622, color scheme is the same as in the secondary structure plots.

When comparing this to the results of model TR567 (4.12, dark gray ring), we see that all setups reach the same 1.4 Å (also see Fig. 4.13, **J**) area, but are unable to refine further. By taking the energy landscape analogon, here we are not starting at an outer flat region of of the landscape but instead well in the funnel region with a direct path to a lower energy minimum. Further improvement to the global native minimum would include reshaping specific sites within the model, temporary dwelling the whole protein or at least some parts of it. Indication for this could be a decrease of the scoring value, although the dRMSD is increasing (v-shape of the scored ensemble in all 3 setups). The consistency of this behavior in all 3 cases, with almost the same minimum points to a very stable, dominant path, not to the global minimum, but towards the basin at 1.4 Å.

When taking a closer look at the scoring of the ensembles generated in **D**, **E**, **G**, **Q** and **R** (see Fig. 4.13) all simulations were less or even not successful in refining. Interestingly all those models except **R** have very good starting models. Difficulties with scoring were observed when structural changes within the frames were not sufficient, mostly when the structures become more and more native. So enhancing the scoring with additional

MODEL (*-id)	♯ atoms C_α	initial RMSD (Å)	scored RMSD (Å)	initial GDT-TS	scored GDT-TS	initial dRMSD (Å)	scored dRMSD (Å)
TR389	135	2.638	**2.076**	0.8097	**0.8116**	1.9643	**1.60163**
TR429	155	6.796	7.130	0.4457	0.4384	5.9684	**5.58303**
TR432	130	1.646	1.869	0.9173	0.8962	1.0505	1.16438
TR435	137	2.153	**2.147**	0.8223	**0.8719**	1.3952	**1.31124**
TR453	87	1.396	1.429	0.8879	**0.8966**	1.0548	**0.96851**
TR454	192	3.238	**2.966**	0.6406	**0.6562**	2.7201	**2.13244**
TR461	157	1.634	1.647	0.9029	**0.9156**	1.1657	1.16861
TR488	95	2.109	2.155	0.8789	**0.8947**	1.3405	**1.2918**
TR530	80	1.990	**1.733**	0.8594	**0.8813**	1.3197	**1.15396**
TR567	142	3.435	**2.172**	0.7817	**0.8063**	2.2811	**1.59672**
TR568	97	6.149	6.425	0.5490	**0.5722**	4.4619	4.606
TR574	102	3.583	**3.462**	0.6201	**0.6765**	2.5333	**2.52954**
TR576	138	6.851	7.468	0.6431	0.5888	4.7643	4.91183
TR592	105	1.257	**1.221**	0.9024	0.9048	1.0310	**0.91054**
TR594	140	1.818	1.910	0.8661	0.8446	1.2569	1.30044
TR606	123	4.850	**4.485**	0.7175	**0.7297**	3.4614	**2.90214**
TR622	122	7.474	7.679	0.6680	**0.6803**	4.0014	4.172
TR624	69	5.189	**5.164**	0.5543	**0.5833**	3.1542	**3.05959**

Table 4.4: Comparison between crystal structures and refined models that were selected by our scoring function. * ♯ given by the assessors of CASP.

complementary information, such as energy functions can be expected to enable better resolution when structures are closer to the global energy minimum.

Apart from the good results with our compact score, a major advantage of the position restraints used here is the fact, that the position restraint energies do reveal the quality of the ensemble (see Fig. 4.14). The higher the position restraint energies, the further away is the starting structure from its native state.

In the three examples, structural superpositions of different models from our set does give a good insight into how structures are really refined and what regions where affected (see Fig. 4.15, 4.16 and 4.17). All examples truly show the advantages and capabilities of the ADPt approach, since models are refined in one part of the model while others are kept stable where needed.

Finally, our results include a recommendation for the general length of free and restrained classical MD refinement setups. Limited improvement was observed in the beginning of a simulation within 20 ns, followed by a overall decrease in quality in the region of about 20 to 60 ns of simulation time. After that, frames began to be better again, and by the time simulations reached 90 ns, the frame quality raised and included more improved than deteriorated structures 4.18. So our suggestion is to set simulation times at least

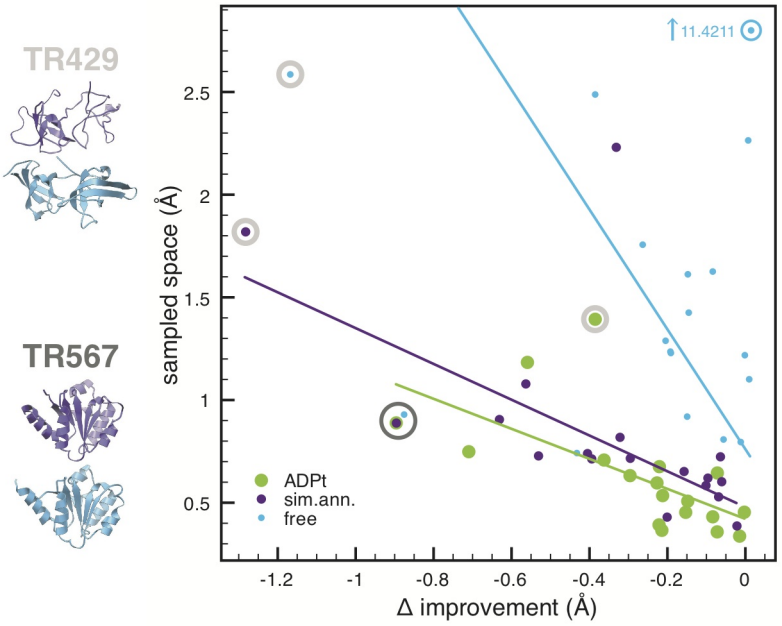

Figure 4.12: Improvement of the models of each setup compared to the sampled space. Models at the left are in purple, the homology model, and light blue, the corresponding crystal structure. An important finding regarding the general ADPt setup is the fact that, when we expand the conformational space, we improve models more. Compared to the free runs, chances here are also higher to expand the conformational space without generating more improved models (light blue line). The more horizontal the line in this plot, the more models were refined without expanding the conformational space too much, hence following a more direct, quicker and computationally less demanding simulation path that leads to improved structures faster.
Models TR429 (light gray rings) and TR567 (dark gray ring) are picked to point out the different improvements (see text) and to show the different improvement behavior.

to 200 ns to be able to profit from our observed improvement in the last 10 % of a simulation, and maybe observe even more improvement afterwards.

This of course does not guarantee convergence, but is rather a guidance for future molecular dynamics simulation implementations.

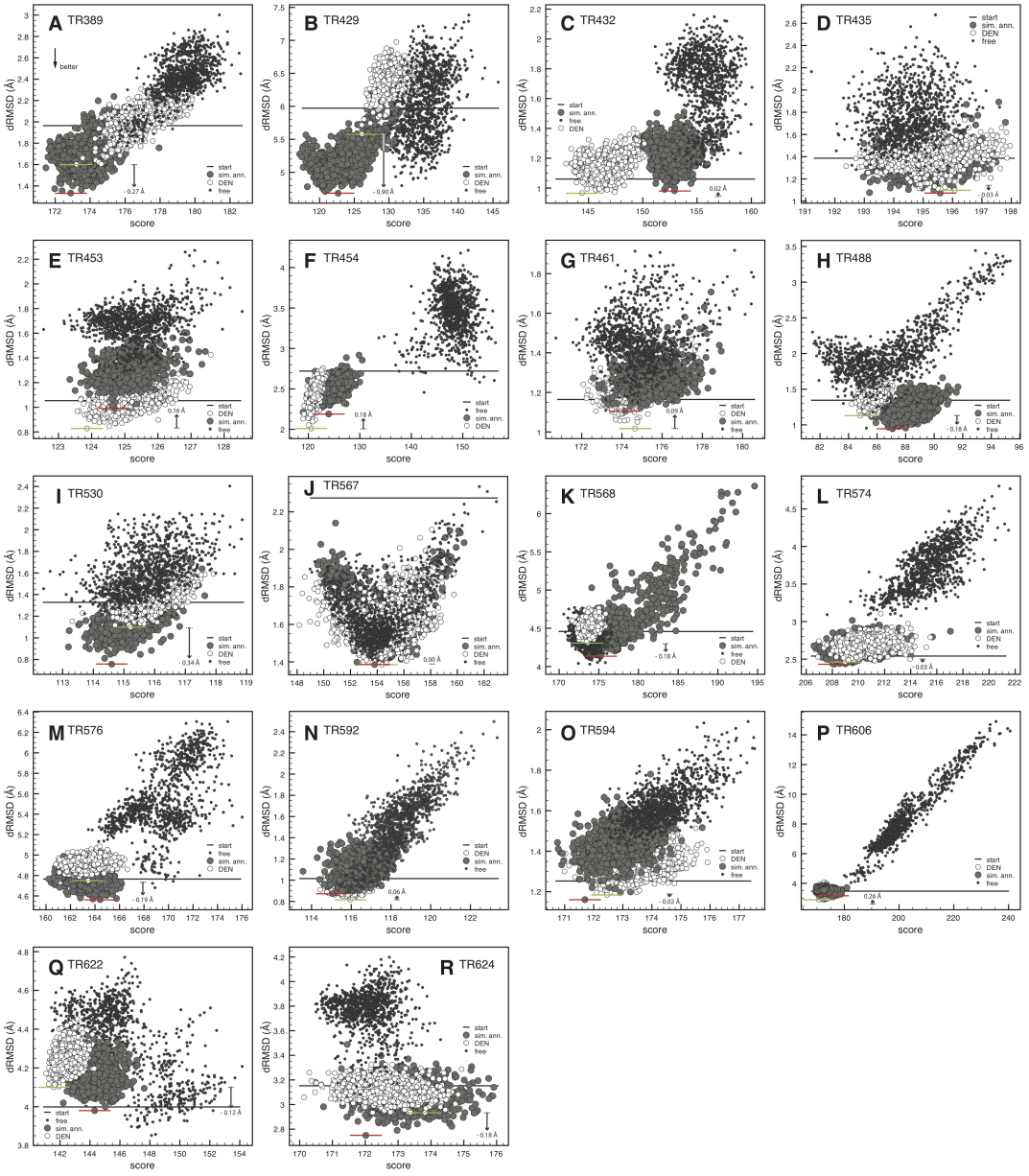

Figure 4.13: All simulations evaluated by the compact score. Also shown is the increased sampling with simulated annealing and free simulations. Green line: Best structure in the ADPt run. Red line: Best structure in the simulated annealing run. With a simulated annealing approach it is possible to expand the sampling space. It is more likely to overcome energy barriers without the negative effect of just deteriorating the structure, which happens in the free runs. Also, the scoring needs conformational changes to successfully discriminate misshaped structures. With a simulated annealing we could enforce more of those conformational differences than a normal run with 300 K may create. Success or failure of this method depends on the energy landscape and the energy barriers.

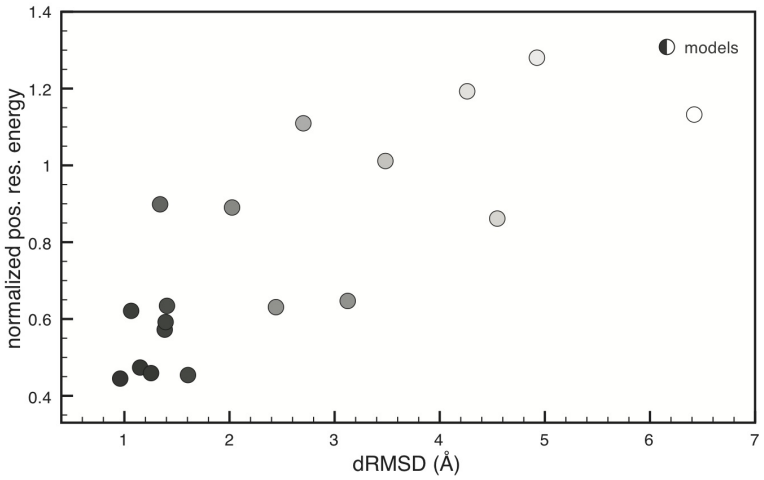

Figure 4.14: Correlation of position restraint energies against the dRMSD per model. Each dot represents the mean dRMSD and the mean position restraint energy value of each model, respectively. Structures which are closer to the native state have lower energies than those being further away from their native structure.

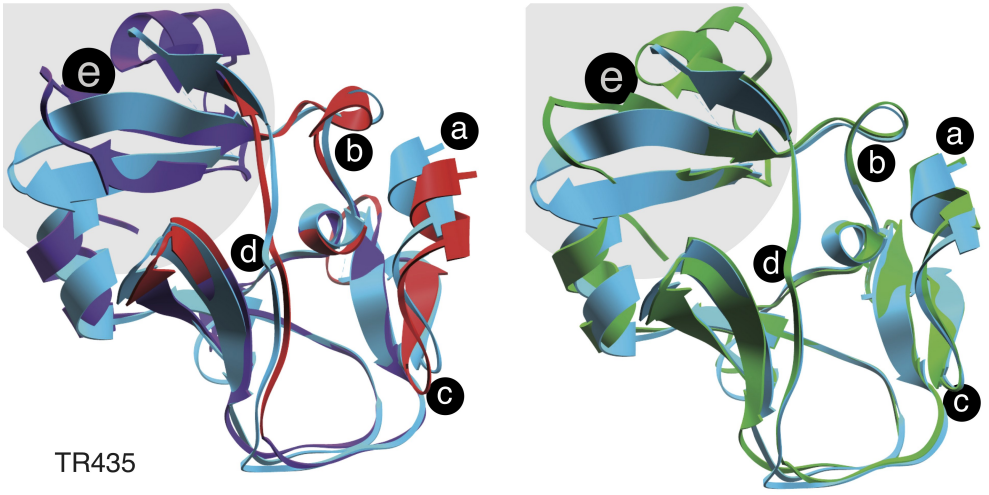

Figure 4.15: Starting model TR435. Crystal structure (blue), homology model (purple/red) (GDT-TS: 0.8223, dRMSD: 1.40 Å), refined model (green) (GDT-TS: 0.8719, dRMSD: 1.31 Å). The homology model was already very good, just with local problems in region **e** and an improvable side-chain packing **a, c**. The refined model still has problems in region **e** but shows tendencies to resolve the misplaced helix. Within this model, region **e** was problematic because it needed some major refolding and melting of a falsely modeled helix. Apart from those regions **a** to **d** reflect a rather accurate refinement. The helix in **a** to **c** is placed more to the core of the protein, resulting in a more dense part, where also regions **b** and **d** profit from, resulting in a very accurate hit of the native trace.

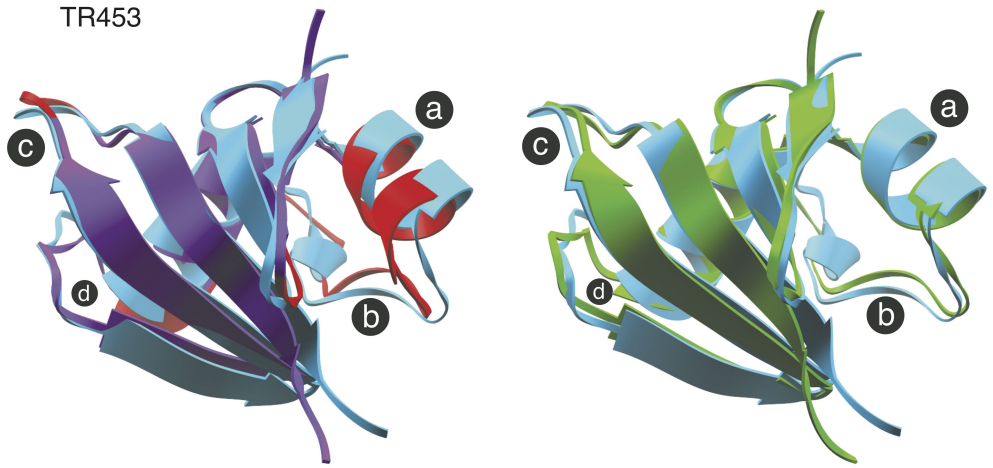

Figure 4.16: Starting model TR453. Crystal structure (blue), homology model (purple/red) (GDT-TS: 0.8879, dRMSD: 1.05 Å), refined model (green) (GDT-TS: 0.8966, dRMSD: 0.97 Å). This is a very accurate model overall, with local problems and a reversed packing problem in region **a**. Here, it is actually better to loosen the rigid compact packing. With this, the model revealed its major refinement in regions **a** (corrected helical structure placement), **b** and **c** (both coil structures agree better with the native trace), and a very good placing of the background α-helical structure in site **d**.

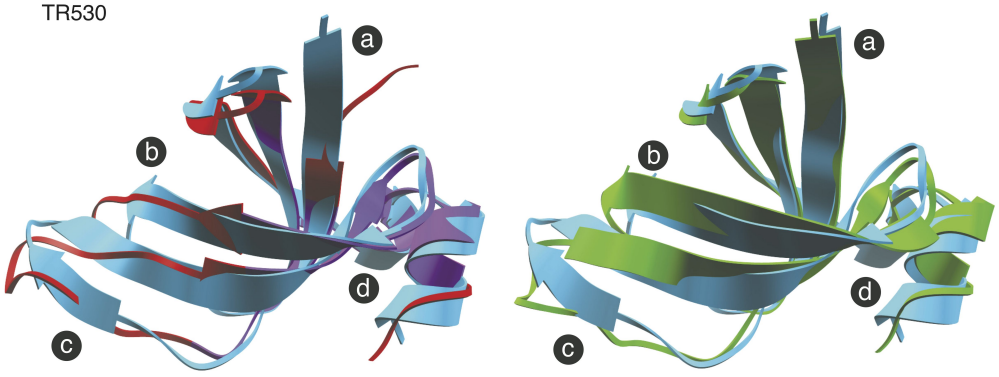

Figure 4.17: Starting model TR530. Crystal structure (blue), homology model (purple/red) (GDT-TS: 0.8594, dRMSD: 1.32 Å), refined model (green) (GDT-TS: 0.8813, dRMSD: 1.15 Å). Simulations improved both β-sheet contents in region **a** and **b**, where a major amount of secondary structure elements were added to the structure and, impressively, kinked the terminal region **a** up to its native contact. The small missing β-sheet in **c** was a bit fluctuating during the simulations and could possibly be stabilized by a crystal contact. Region **d** did almost not move, indicating a stable state of the modeled helix.

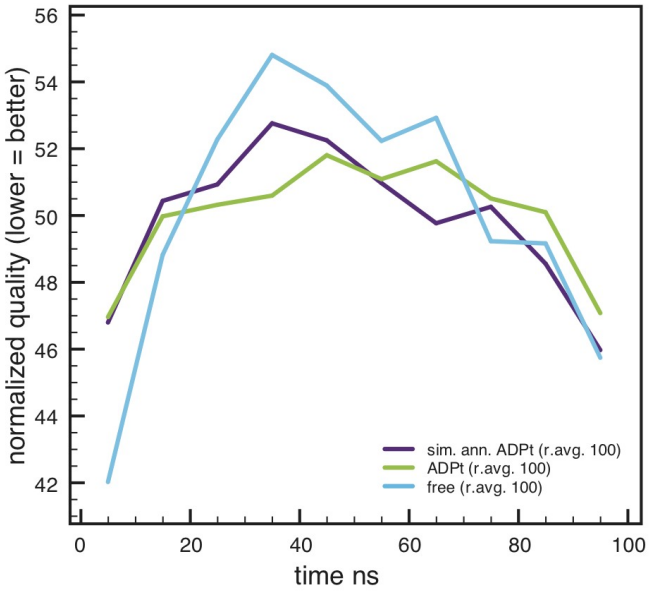

Figure 4.18: Histogram of improved frames (0 is best, 100 is worst) for all simulation sets, running averages for all models. For clarification, each line is not comparable directly, it is just a measure, when refinement runs improve or deteriorate a model, meaning that no absolute numbers are plotted here, but normalized representations. The plot shows the deterioration tendencies of the quality of the structures during the middle of the 100 ns simulations. Structures had a better measured quality (better RMSD against the native structure) in the beginning and the very end of our simulations.

4.3 Discussion

Generalization of MD simulation results is often difficult because of the large statistical fluctuations. To reveal a general trend requires starting a large number of similar simulations with random initial velocities. The number of simulations that can be performed is clearly limited by available computational resources and is a trade-off between reliability of the observed trend and computational cost.

We observe three categories of refinement problems: The first category is the refinement of models which have an RMSD to their crystal structure of about and above 5 Å. The models can be very correct locally but carry severe misfolds in some regions. Or the structure has an overall low quality, with a medium error all over the whole model. The former is seen more often in homology models, while the latter may be more improbable, because once the homology towards the similar sequences breaks or is fractional, no minor overall error is put into the homology model. When homology vanishes, the transition from good models to massively misfolded proteins is quite steep. Within a cut off range of about above 10 Å, models can still be refined, but due to the errors which one starts with it will take longer and may even fail at all to refine, while destroying more than correcting.

The second category is the refinement of models that are in the range of below five and above roughly 1.4 Å RMSD towards their crystal structure. This may be the area into which most of the homology models fall. Our results show consistent success in refining models in this range.

The last category is the refinement of models with a RMSD below 1.4 Å. It turns out, that this region is very hard to address with either simulations and scoring because we saw a limited success in correlating those scoring values with the real structural dRMSD towards the crystal structure. A possible solution could be to enhance sampling even more (apart from simulated annealing) to generate more near native models and/or extend the scoring measure with additional terms that allow for discrimination of the very dense similar native like structures.

Qualities of above 10 Å RMSD are not considered to be mainly addressed by our MD refinement any more, because their local error can be so large that the methodology may need different considerations and the problem is then to refold significant parts of

the model. Fortunately, our simulations can detect the quality quantity of the targeted model quite well (see Fig. 4.14).

These findings may help to categorize and rate targeted refinement attempts and aid to avoid unnecessary considerations concerning timescales and system setups.

Within our approach, the energy function is not original anymore since we introduced a self-adapting position restraint into the simulation. However, when the simulation converges and the atoms do not move much anymore, our ADPts relax slowly towards to the atom positions. In this way the restraint energy slowly converges to zero and the structure rests in an almost static part of the energy landscape, which is particularly true in the case of minimization. That means when minimizing a structure, the local minimia of the energy landscape are not altered by the adaptable position restraints.

On one hand one might argue, that a slightly changed behavior of the energy function is undesired since the notion is plainly always to be more and more precise and mimick a natural unaltered forcefield with highly accurate parameters. But there will never exist the "one and only best and optimal" force field in simulations, best for all simulation conditions. For example classical force fields do not describe the polarizability of atoms, and cannot treat the hydrogen binding equilibrium of protonatable groups depending on the pH-value. So perturbing the interactions with our approach to overcome local energy barriers is not destroying force field parameters and preventing correct results. Despite the fact, that MD is able to describe proteins in a realistic way and do the right integration for moving atoms correctly, physics based force fields are not designed for correlating its output with the so called (X-ray crystallographic) native state of a molecule [102]. That means, when a crystal contact or even the whole crystal environment is required to keep a loop in place, a simple free MD simulation will not yield any structure closer to a native state, other than by chance [103]. In this scenario, the simulation will first try to escape its locally deformed, non natural state and subsequently relax to its biological structure. This process alone could need more time than 100 ns, which is a normal efficient simulation length today, pushing the minimal simulation times necessary for senseful refinement to approximately 200 ns and longer. That also means, that unrecovered errors within the model due to the crystallization process are not obvious or even detectable, though high variance sites within the molecule (atomic position fluctuations) are a clue but not in the least a general hint towards misplaced regions, since

intrinsically flexible loops would have the same signature. It can be stated that some protein models cannot be (efficiently) refined without the presence of a crystal contact environment or an enhanced sampling method, given that the target which is aimed for is a X-ray crystallography model (own work). Furthermore, models which seem to be wrong locally compared with their crystal structure may be right when compared with their biologically active (non-crystal) counterpart [103].

Being able to find native-like structures within a populated ensemble plays a major role while assessing methods for structure refinement. Results of our scoring method reveal, that the function works for at least 12 models as it is able to exclude the majority of non native structures. The calculation of the compact score as it was used here depends only on geometric distances. Further modulating it with knowledge-based approaches could improve and enhance its success rate even more. Residue propensities and secondary structure motives could increase the sensitivity for ranking near native structures more accurately (like mentioned earlier). Additionally, effects from solvation, electrostatic interactions and H-bonds can be assessed to enhance the existing function with the ability to successfully score models with high resolution (approx. below 1.4 Å).

In the past, enhanced sampling methods like replica exchange MD were successful in refinement [74, 104], because it was able to sample in otherwise unexplored ensemble regions and thereby increasing the probability of sampling in the vicinity of the global energy minimum. The method developed here is trying to take advantage of two major properties: Local enhancement of correctly folded regions, to finally keep atomistic arrangements that are of low energy, and stable non fluctuating high kinetics sampling, combined in one simulation. Local enhancement and peak fluctuation mediating is achieved by applying a positional restraint. Overcoming energy barriers is enforced by simulated annealing and adaptive adjustment of the positional restraint coordinates. The overall effect of our method is self adjustment and local settling of kinetic and potential energies. Maximum refinement should be reached in general, when all local fluctuations are minimized by relocation and when the model sits in a globally settled state. This is reflected in Fig. 4.14, where we show that with our simulations we can correlate the globally normalized RMSD over all frames of a trajectory with our special position restraint energies.

In general, we were successful in refining structures to a certain degree, structure refinement was successful although limited by an improvable sampling when aiming for qualities below 1.4 Å RMSD, but expanding the conformational space could improve the success rate of refinement attempts immensely like simulated annealing simulations plainly showed. With non optimal sampling, scoring is pointless and with bad scorings, sampling is of no use.

When comparing to crystal structures, the refinement performance could possibly be improved by modeling the realistic crystal environment and include the crystal contacts, which for example would limit the conformational changes in the contact zones of the units.

Moreover, we think that force fields are accurate enough [76] for approaches like ours, where methods enhance the sampling of MD simulations, since most of them change the dynamics significantly (even replica exchange methods fall into this category, since most of the force fields are not parametrized to correctly represent and handle temperatures above e.g. 373 K) and the added terms do introduce effects not observed in unmodified normal simulations [105].

A possible guideline for simulation result assessment can be a measurement, where secondary structure content, when correlated with the size of a protein shows propensities which are utilizable to choose appropriate system settings to match their needs for additional improvement.

Finally, our compact score was helpful to enhance the quality of the results and may possibly be further enhanced by empirical functions in future implementations. Extending the simulation time from 100 to 200 ns, or even longer could enhance results even more.

Chapter 5

Coupling Structures with Multiple Sequences

The success of comparative modeling stems from the fact, that homologous sequences share a similar structure with an accuracy depending on the degree of identity that the sequences incorporate. Several studies investigated the reasons, implications and possibilities that go together with this fact [106, 107], with the aid of simplified two-state lattice models to deal with the tremendous size the sequence and structural spaces.

The limited (postulated) number of folds [11, 12], the robustness of the amino acid code [108, 109], the limited sequence space that nature favors [110] and the fact that similar sequences can share a very similar structure [111] narrows down the search spaces which would have to be explored when randomly trying out each possible sequence or conformation in the hunt for finding the correct native fold. But homologous sequences can only be used directly for homology modeling when information about the homologous structure is available. The sequences with unknown structure are unfortunately unused for refinement within molecular dynamics simulations, yet there should be a way of exploiting the tremendously vast space of known sequences, because they can contribute to homology model refinement with their evolutionary filtered sequence information (see Fig. 1.1).

In this study, we focus on the refinement of structures derived from homology modeling methods which are enhanced, stabilized and averaged with homologous sequences, having a sequence identity in the range of 50 to 80 % (Tab. 5.2).

Our approach shares similarities with methods that try to solve multi-dimensional energy landscape problems.

For example in Particle Swarm Optimization (PSO) [112, 113], a large set of agents is applied to find an optimal solution for a search problem. It exploits the fact that many individuals which can share information between each other, are better in solving a problem, than a single individual can ever be.

In our case, to reuse this specific example, we replace the agents by homolog proteins and the sharing of information by our special interconnected ADPt restraints. A step further would be, to not use a communicating system of agents alone and let it self adjust, but to incorporate additional information on top of the plain agent, that adds and contributes to the targeted problem. Again, in our case, this would translate to the evolutionary mutated sequence.

In protein *ab initio* folding as well as in refinement approaches, homologous sequences on the one hand, share common properties with a targeted fold to not distract and disturb or eventually break and destroy the system, and on the other hand introduce a perturbing force to overcome the local trap problem.

The general impact of an average over homologous sequences that is used to diversify the energy landscape was investigated in the past on the example of *ab initio* folding and structure prediction by using 2D lattice models and small proteins with Monte Carlo (MC) simulations [114–116]. The approach compared aligned structures of multiple simulations to penalize moves that lead to structural diversity.

Coarse-grained (one bead per amino acid) 2D lattice models were also used to investigate the effect of mutations of a protein sequence on the example of 4-12 bead models [117]. Here, one structure at a time was sampled to subsequently couple energies by using MC simulations.

Another approach connected the considerations that govern the random energy model (REM) with the aspects of homologous sequences [118] and the free energy minimum of the native fold of a protein.

In the SWARM-MD approach [46], in vacuo MD simulations of one 64 residue protein in an united atom force field were investigated. To smoothen the rugged energy landscape, a root-mean-square dihedral angle difference based measurement was introduced to keep the energies of an added potential energy function low. Here, the same molecule was simulated multiple times, however no information on homologous sequences are considered.

The SWARM-MD approach was recently [119] implemented into the Amber molecular dynamics package, and tested by using 3 test models, a 11-mer alanine, a 17-mer polypeptide and the globular 20-mer TRP-cage. The implementation uses the same proposed additional potential of the original SWARM-MD approach, but takes just 20 simulation replicas for building up the swarm of conformations.

Besides *ab initio* folding of proteins, with important examples shown above, the refinement of comparative models remains to be a challenging problem. Since both, accurate fine tuned sampling within small energy differences and overcoming relatively large energy barriers is needed at the same time to improve protein models, the task of refining a polypeptide within the boundaries of limited detail simulation environments and finite computer resources is a balancing act.

Our approach, broadly described as simultaneous coupling of homologous molecules derived from similar sequences, wants to accomplish that task, yet it is kept as simple as possible. To work properly, it only needs the homology model and additional similar sequences as input. Restraint force constants and the global coupling term κ can be adjusted. Everything else is self-managed and integrated into the Gromacs MD suite (version 4.5.3) (see Fig. A.1). With this it takes advantage of the speed improvements of the fast parallel domain decomposition code implementations. The sequence augmented system setup combines and connects all added sequences into one simulation system and simulates the homology models simultaneously. All calculations needed happen in-place and are parallel, while no superpositioning algorithms and no global restrictions are needed or imposed on the whole setup. The system is comprised of one multi molecule MD simulation box with the homology models and explicit solvent including the ADPt restraints.

The impact of homologous sequences on refinement simulations, namely the improvement of the quality of the targeted structure, is based on the modification of the underlying energy function. The differences between homologous sequences are not random. All sequences are the result of natural selection during evolution, e.g. for optimized protein function.

Proteins evolve because of naturally occurring mutations. A mutation can cause amino acid changes, deletions and additions. During the folding process of a mutated protein, the new structure is either equally stable, unstable, more stable than its unmutated predecessor. Given enough time, this process introduces a sieving of valid and working mutated protein sequences. The original function was either enhanced or shifted. All dysfunctional mutations are not maintained because they simply do not fold correctly. In a very basic procaryotic cell this can be fatal and lead to cell death. Even in multi cellular organisms, this mutations can be fatal, though other cells can buffer the dysfunction of a few. If the function was shifted, a new protein class can evolve. In this way, random mutations probe for sites in the sequence whose change do not break the structure of the protein. This ultimately evolves proteins, who share the same fold but possess different sequences.

Exactly this homologs can provide the energetic perturbance needed for the refinement of structures.

Unlike other methods, that try to optimize the speed of the search for free energy minima in the vast conformational space by simplifying the complexity of the model itself by reducing its degrees of freedom (coarse graining) [120] [17], we narrow down the space to search in. By coupling our models simultaneously, the observed conformational space during the simulation is extremely reduced (though theoretically free to move everywhere). This is needed to introduce small conformational changes without destroying the rest of the structure. And this is where reliable refinement can occur. The coupled sequences can be thought of opening a sub space in the vast conformational room. In this sub space, the copies serve as the energetic boundaries (manifested in the force field). Still, given enough simulation time to change, the explored conformational space will be huge but with contributions not only from one structure with his own free energy landscape but from seven. All these energies add together, forming an averaged landscape, which hopefully lowers energetic barriers along the conformational path during

the simulation.

To study the advantage of our new method over standard MD simulations, and to separate the effect of the adaptive position restraints (ADPt) from the effect of the coupling of homologous structures, three different refinement protocols were performed:

A Simulations with ADPt restraints averaged over eight models with different, evolutionary related sequences.

B Simulations with ADPt restraints averaged over eight models with identical sequence.

C Free simulations without any restraints or coupling.

5.1 Methods

5.1.1 Selection of Test Cases

We selected five models out of the set of 31 available from the 2-4 Å in the homology model benchmark set by A. Sali (Badretdinov decoy set). Those five models were selected to represent "normal" protein properties that cover a wide range in aspects of size, SSEs and globularity. Also, we chose a protein model that was among the largest of that set, with a large number of amino acids.

MODEL (RCSB-ids)	♯ atoms C_α/all	RMSD (Å)	dRMSD (Å)
1dvrA-1ak2	220/3452	2.790	2.250
1hdn-1ptf	87 /1297	2.150	1.712
1lpt-1mzl	93 /1240	3.887	2.572
1pod-1poa	118/1730	2.347	1.860
1utrA-1utg	70 /1116	3.002	2.509

Table 5.1: Overview of all test proteins. The model name consists of two identifiers: The first identifier denotes the PDB ID of the template structure which was used to build the homology model. target where the structures was borrowed from, and the second identifier is the PDB ID of the target structure, which was not used for the refinement. The RMSD and dRMSD value measured between the starting model and the corresponding target (crystal) structure is shown.

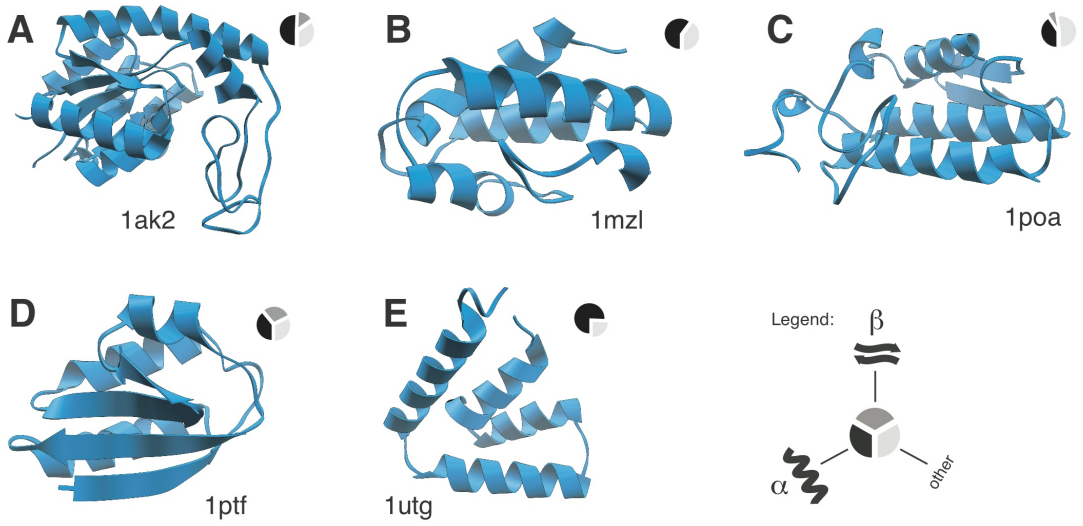

Figure 5.1: Crystal structures of all models used. The legend on the right shows the secondary structure composition of the models.

5.1.2 Sequence Selection for Coupling

We used psi-blast [121] to search for homologous sequences and decided to choose sequences with an average identity of 60.5 % towards the other included sequences and an average identity towards the target sequence of 61.8 %. For the search the blast package 2.2.23, build Mar 8 2012 14:49:45 was used. We used the updated "nr" database (including all non-redundant GenBank CDS translations, pdb, swissProt PIR and PRF, excluding environmental samples from WGS projects) with 11,205,216 sequence entries for our search.

The sequences were chosen such that the corresponding structures can be expected to be highly similar to the target structure. This is achieved by choosing sequences with identities to the target sequence of higher than 50 %.

Also, the similarities between the chosen homolog sequences should be considered. They must not be too high, because their influence ought to be as independent as possible to introduce as much meaningful variation as possible.

5.1.3 Model building for coupling

The modeling process starts with building an initial homology model that will be the starting structure for refinement and also provides the template structure for building

MODEL	id	identity (to target) %	identity (to all) % min avg max
1dvrA-1ak2	1	70	55 62.3 71
	2	68	56 64.2 71
	3	57	54 58.5 61
	4	53	49 55.7 59
	5	65	49 58.2 63
	6	61	57 58.4 61
	7	66	56 64.3 76
1hdn-1ptf	1	68	56 68.8 82
	2	66	59 68.6 82
	3	64	53 62.1 71
	4	51	53 57.2 61
	5	59	57 64.4 71
	6	71	53 64.2 73
	7	54	53 59.5 62
1lpt-1mz	1	58	42 54.1 68
	2	56	41 48.8 62
	3	55	47 55.1 63
	4	52	41 45.6 50
	5	67	47 55.5 68
	6	76	41 56.2 62
	7	63	45 53.2 63
1pod-1poa	1	77	57 65.4 73
	2	69	59 66.0 73
	3	63	56 66.8 84
	4	74	52 60.6 72
	5	59	52 62.5 73
	6	61	52 67.6 84
	7	77	57 63.7 73
1utrA-1utg	1	58	54 63.6 76
	2	52	43 50.7 59
	3	54	52 62.7 77
	4	55	53 62.7 77
	5	53	43 60.6 76
	6	57	49 65.3 90
	7	55	43 62.7 90

Table 5.2: For each test case (MODEL), seven homology models (with ids 1-7) were generated in addition to the homology model that was built for the target sequence. All these eight homology models were built using the same template structure. The sequence identity of each homology model to the corresponding target sequence is shown in the third column. The last columns show the minimum, average, and maximum sequence identity of each homology model to all other seven models used for one test case.

models for all other sequences.

Now, the added sequences simply serve as an extension of the simulation setup on the level of the evolutionary perturbed sequence information and any improvement in the targeted structure purely arises from the added sequences, when compared to the similar sequence approach.

All modeling steps are performed with Modeller [20, 24, 122] version 9.7.

5.1.4 Sequence Coupling

The simulation box consists of eight homology models that are placed far away from each other, such that there is no direct interaction between them. One of the models has the target sequence, the other seven models have sequences that are similar to the target sequence, with an identity of better than 50 %.

Each C_α atom i in each model j is restrained by an adaptable harmonic position restraint at position $\vec{p}_{ij}$. The initial coordinates of the restraints are identical to the C_α coordinates, $\vec{x}_{ij}(0)$, in the starting model.

After a specified update interval, typically 500 integration time steps, the displacement vector $\vec{v}$ of C_α atoms i in model number j is given by

$$\vec{v}_{ij} = \vec{x}_{ij}(t+1) - \vec{x}_{ij}(t), \tag{5.1}$$

where $\vec{x}_{ij}(t)$ is the position of C_α atom i at update step t.

The average displacement vector $\overline{\vec{v}}_i(t+1)$ of each position restraint is obtained by averaging over all corresponding C_α atoms:

$$\overline{\vec{v}}_i = \frac{1}{8}\sum_{j=1}^{8} \vec{v}_{ij}\,. \tag{5.2}$$

This average displacement vector is then added to all positions of corresponding restraints,

$$\vec{p}_{ij} = \vec{v}_i + \vec{x}_{ij}\,. \tag{5.3}$$

Our approach can deal with homologous sequences that do have different amounts of residues compared to the targeted model, so it is allowed to have gaps and insertions. In case of any occurring insertion or gap that has no corresponding partner no connection or interaction is made, it is just left free.

5.1.5 Molecular dynamics setup

All three different setups **A**, **B**, and **C** (list 5) used the same MD parameters and the same force field, except for the restraints. The free simulations were not restrained or

modified in any way.

After a steepest descent energy minimization and a short 100 ps position restraint equilibration of the system, a 300 K fixed temperature Nosé-Hoover NVT simulation with tip3p water and the Amber 99SB-ILDN force field with a time step of 2 fs was performed. For the restraints, we used $\kappa = 0.5$ and a force constant of 100 kJ/mol^{-1}/nm^{-2} (not for the free runs). All electrostatic long range interactions were calculated with PME and the constraints were controlled by the P-LINCS approach. On average, 32 - 120 cores, distributed on several nodes, were used.

5.1.6 Model Quality Assessment/Compactness Score

The compactness of protein models is a unique feature of nativeness [79, 123].

Our compact scoring function is a combination of a local and a global evaluation of compactness. It is used as defined in 4.5.

5.2 Results

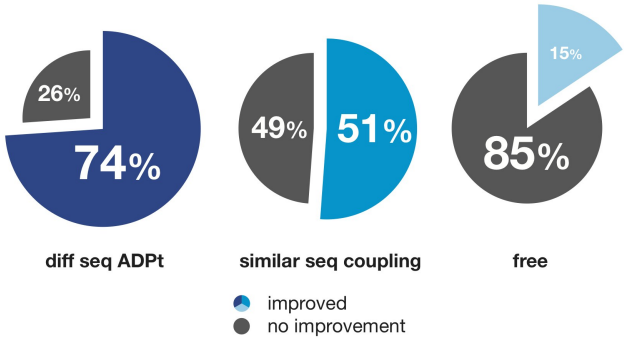

Figure 5.2: All simulation results. Showing the improvement percentages of all frames from all runs for each method in *blue* and the deteriorated frames in *gray*.

The complete impact in improvement of the three approaches are resumed in Fig. 5.2, where the coupling method augmented with different sequences on the left shows the best result by far. Nearly $\frac{3}{4}th$ of all produced frames were refined structures. Compared to the similar sequence coupling approach this is $\frac{1}{4}th$ more. Although the similar sequence approach was able to refine only 51 % of the runs, compared to the free runs, which only refined 15 % in total, this is also remarkable.

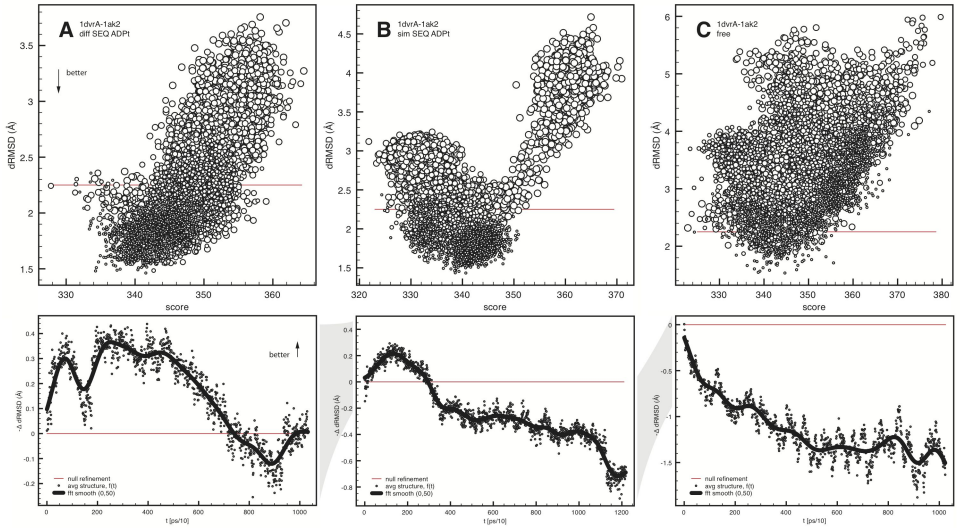

Figure 5.3: Averaged simulation results for model 1dvrA-1ak2. The upper 3 plots show our compact score results applied on the generated ensembles for the sequence augmented, the ADPt and the free simulations. The lower plots show the averaged simulation results of all simulations per model and approach, averaged per frame. In this case, our compact scoring works best with the sequence augmented approach **A**. The free simulations **C** show a comparatively unordered distribution, what means it is more difficult to rank structures properly. The simulation results are best for **A**, can rarely improve in case of the similar sequence approach **B** and show no trend of improvement for the free runs **C**.

Figures 5.3 - 5.7 show the individual simulation results, averaged over all frames produced for one target, respectively. The plots **A** - coupling different sequences, **B** - coupling similar sequences and **C** - free simulation runs, on top always show the compact scoring (different symbol sizes correspond to different simulations) and the red line in all cases marks the *null* refinement, namely the starting quality of the homology model. Here, the pure dRMSD is plotted against the scoring energies, so a lower value in the left corner is the ideal case of a refined and very well scored structure.

The smaller rectangular plot below **A**, **B** and **C** directly shows the averaged refinement $-\Delta$ dRMSD in Å of every approach, so every dot above the red line means improvement. For *1dvrA-1ak2* the scoring and the refinement was successful. As can be seen in Fig. 5.11 the model size is relatively large (220 residues) and the refinement problem is rather complex because approximately $\frac{1}{3}rd$ of the structure basically needs a remodeling while the rest is build quite accurately. Despite the drop of the quality of the structure at the end of the simulation, improvement adopted very fast, more or less directly after starting the simulation. This is true for all simulations and models we tested, and can be considered as a direct effect of our sequence coupling augmentation approach. In comparison to **B** and **C**, **A** gave the best results by far. The compact score directly reflects the quality of the three approaches. The energies of the worsened models is higher than those of

the better models throughout the different approaches. The overall correlation values for the scorings are $r = 0.54$ for **A**, $r = 0.11$ for **B** (the best $r = 0.87$ was lowered by the other scoring values which were partly negative. It is interesting to see, that the scoring had problems in the lower energy regions, just like proposed in the discussion of chapter 4), and $r = 0.50$ for the free approach **C**.

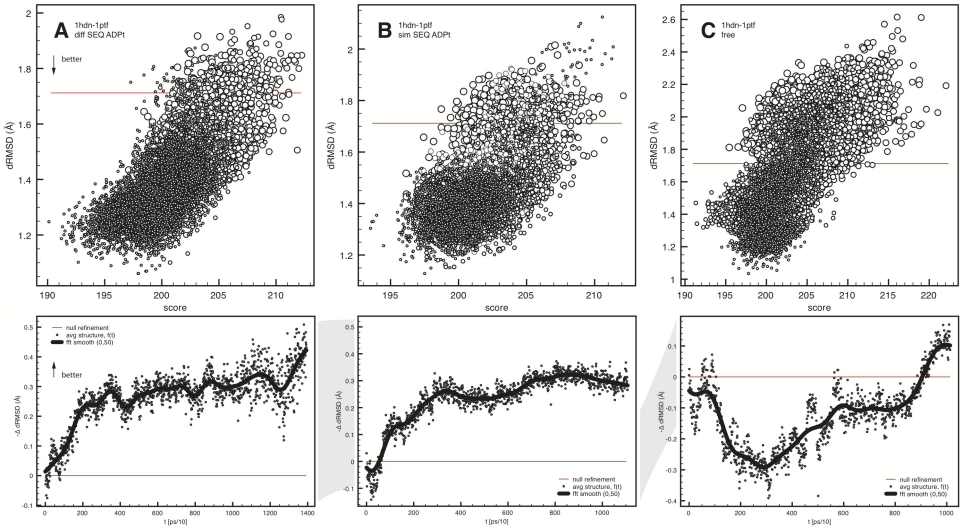

Figure 5.4: Averaged simulation results for model 1hdn-1ptf. The upper 3 plots show our compact score results applied on the generated ensembles for the sequence augmented, the ADPt and the free simulations. The lower plots show the averaged simulation results of all simulations per model and approach, averaged per frame. Our compact scoring worked good in all 3 cases **A**, **B** and **C**. The simulation refinement was again best for the sequence augmented simulations **A**. 0.1 Å worse was **B** and the free simulations **C** just degraded the frames, although slight improvement occurred at the end of the simulation.

The simulation and scoring results of model *1hdn-1ptf* are very promising. All compact scores are high and point very accurately to improved models. The correlation coefficient values are $r = 0.59$ for **A**, $r = 0.57$ for **B** and $r = 0.62$ for **C**.
The averaged simulation results in this case show the two coupled approaches relative close together, slightly more improvement for **A** at the end of the runs. The free runs show a tendency to refinement but compared to the two other runs here the variance of the dRMSD per time is much higher.

A more difficult target was *1lpt-1mzl*. The scoring did not yield any additional information about the model quality in none of the three approaches and the refinement was very limited for **A** and **B**. During the end of the two runs a tiny trend towards better structures are recognizable, but overall the coupling just aided to maintain the given starting model, which indeed is also a very good observation. The correlation coefficients

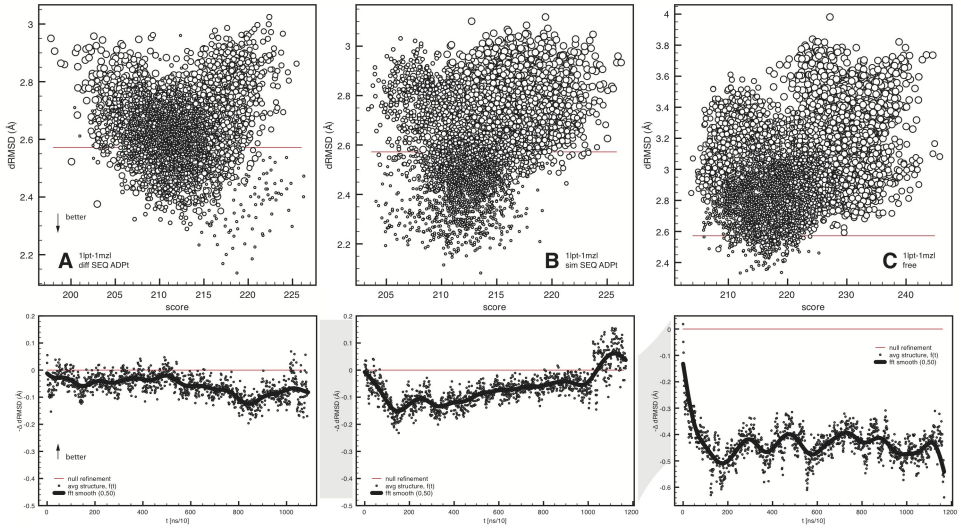

Figure 5.5: Averaged simulation results for model 1lpt-1mzl. The upper plots show our compact score results applied on the generated ensembles for the sequence augmented, the ADPt and the free simulations. The lower plots show the averaged simulation results of all simulations per model and approach, averaged per frame. Here, no real good scoring was observable for all cases, and the simulation results just show, that **A** and **B** are both able to keep the structure in its current energy minimum. In contrast, the free simulation **C** deteriorates the structure quite fast.

for the scoring are $r = -0.15$, $r = -0.03$, $r = 0.17$ for **A**, **B** and **C**, respectively.

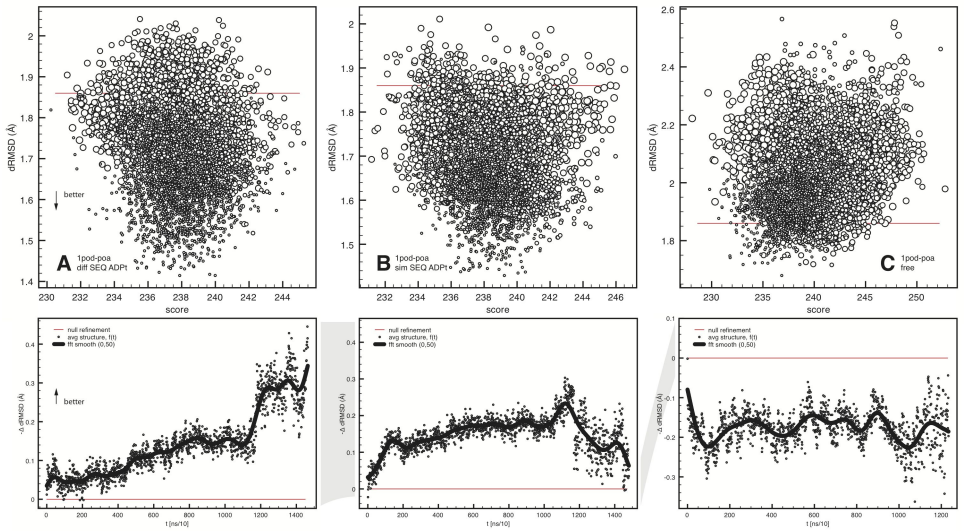

Figure 5.6: Averaged simulation results for model 1pod-1poa. The upper 3 plots show our compact score results applied on the generated ensembles for the sequence augmented, the ADPt and the free simulations. The lower plots show the averaged simulation results of all simulations per model and approach, averaged per frame. Scoring this target was difficult, since no approach showed significant tendencies to work well. The simulation results instead were remarkably good for **A**, a bit worse for **B** and again quite bad for the free run **C**.

Model *1pod-1poa* was similarly unsuccessful in its scoring with $r = -0.23$, $r = -0.15$, $r = 0.063$ for the three approaches **A**, **B** and **C**. But here, the refinement was very stable

and showed an increase of refinement for **A** in the last 2 ns of the simulation, whereas **B** decreased during that time. The free simulation resembled all other free simulations for all models and resulted in a high variance deterioration of the model.

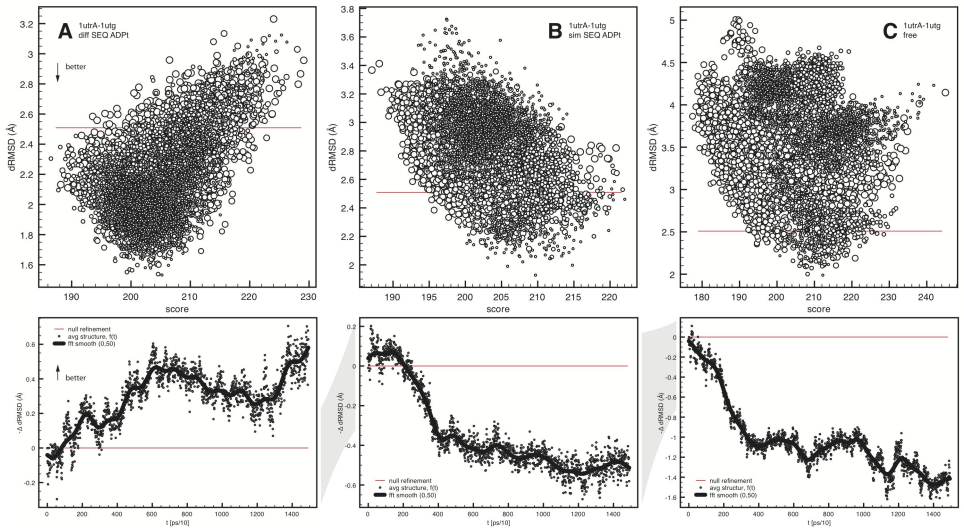

Figure 5.7: Averaged simulation results for model 1utrA-1utg. The upper plots show our compact score results applied on the generated ensembles for the sequence augmented, the ADPt and the free simulations. The lower plots show the averaged simulation results of all simulations per model and approach, averaged per frame. For this target the compact scoring again worked very good for the sequence augmented approach **A**. For the other 2 examples it did not produce helpful results. The simulation outcomes were again best for **A**. Here, the maximum refinement achieved was a remarkable 1 Å dRMSD. **B** did not refine significantly and **C** just performed good in worsening the structure.

The last model of our set *1utrA-1utg* showed a very good refinement capability and, in case of the different sequence approach **A** also the scoring was successful with a correlation coefficient of $r = 0.53$. Neither the similar sequence coupling nor the free approach was able to drive the structures any closer to the native state. Also the scoring correlations were undirected with values of $r = -0.48$ and $r = -0.33$ for **B** and **C**, respectively.

All simulations with the different sequence coupling approach were either successful in refining the given model or, in case of model *1lpt-1mzl*, were able to hold the structure at the *null* refinement starting level. Additionally, three of the models could be scored with our compact scoring. In contrast, the free simulations were altogether unsuccessful. For given protein model simulations it is not only important to gain insight into the single simulation performance. Equally important is the possibility to rank the model in relation to other models. For this, Fig. 5.8 lists the averaged compact score energies

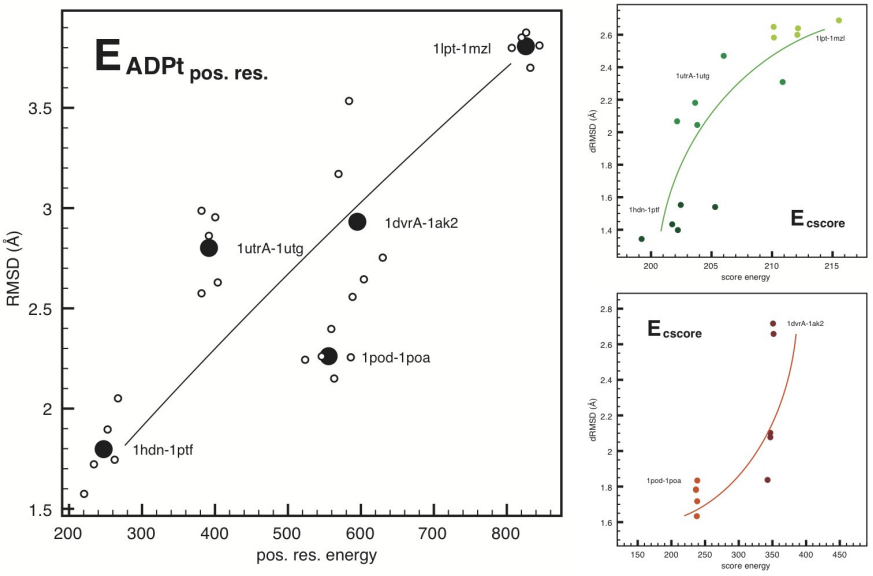

Figure 5.8: Position restraint energies from the simulations and compact score energies. The position restraint energies (left, *black*) can be correlated to the overall starting RMSD in Å of the homology model. This is also true for all averaged compact score energies (right, *green* and *red*).

and the averaged special ADPt position restraint energies from the coupled simulations. The position restraint energies are able to roughly rank the models into their problem set difficulty, which is the global RMSD in Å of the combined averaged simulation set of all runs belonging to one model.

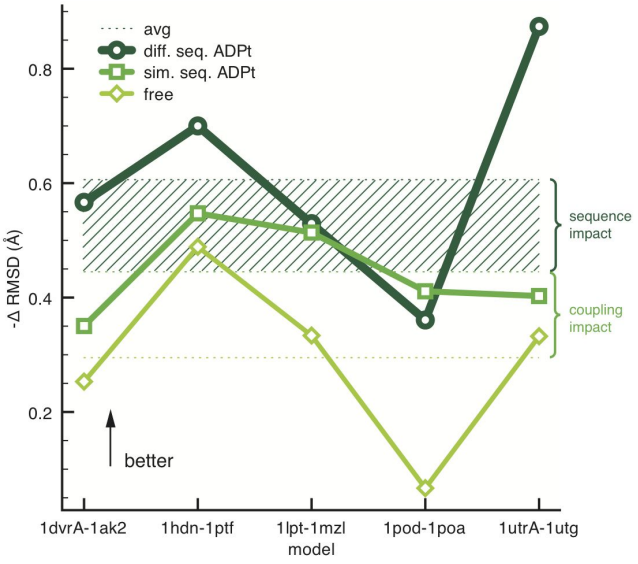

Figure 5.9: Averaged maximum peak simulation results relative to the method used. The dotted lines are the averaged results and the shaded areas mark the difference in gained improvement per approach.

The same can be achieved, to some extent, with the global averaged compact score

energies (right of 5.8), though the spread of data points is slightly higher here than in the ADPt energies.

Sorting out the maximum refinement performance depicted in Fig. 5.9 acts as a monitor to point out the capability of the augmentation of simulations with sequence information. The green shaded area is the difference of the averaged results of all best structures generated within all simulations compared to the coupling approach with just the same model. From the free simulations view, the augmentation that was achieved with coupling in peak performance on average was 0.15Å in the same sequence case (middle green), and another 0.15 Å resulting in 0.3 Å in the different sequence case (dark green).

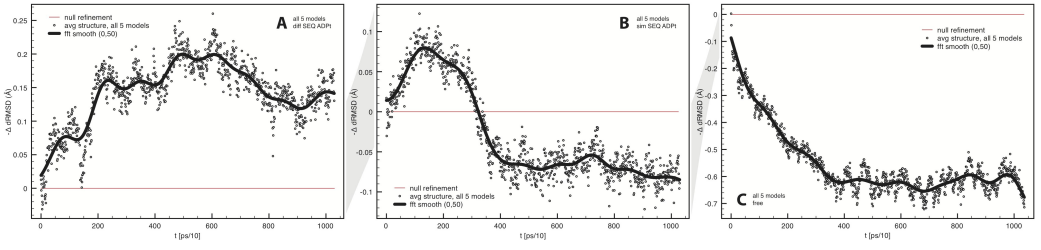

Figure 5.10: All simulation results averaged (higher means better). These plots depict the averages of 75 simulation runs in total. This means that each of the 3 plots includes 25 runs, totaling in ≈ 250 ns simulations for A, B and C, respectively. Including all runs, 1.13 μs simulations were produced. **A** shows that ADPt simulations augmented with seven different homolog sequences produce stable and reliable refinement results compared to **B**, the simple coupling of 8 identical sequences and the **C** free runs. The refinement visible in **A** is directly caused by the homolog sequence coupling, because similar sequences coupled, shown in **B** do not hold this effect. But still, compared to free simulations **C**, coupling the same sequence also has a stabilizing effect, though not as efficient as in **A**.

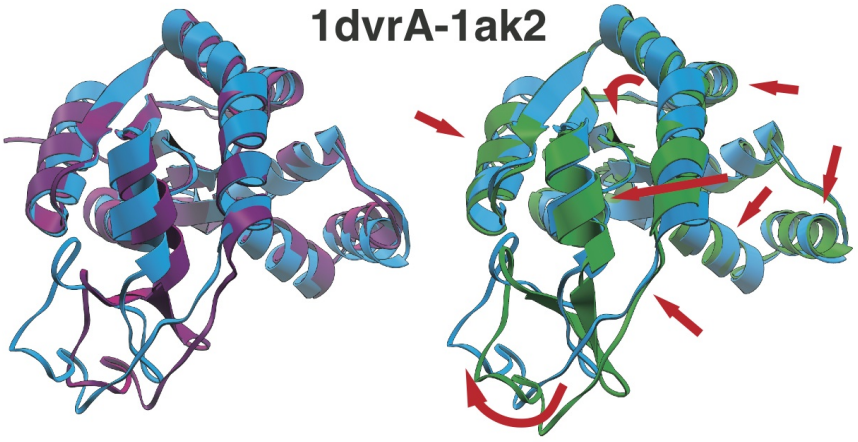

Figure 5.11: Model improvement 1ak2. The crystal structure in **blue**, the homology model in **purple** and the refined model in **green**. The red arrows point out sites with major improvements.

The results and the improvement of a single model can be taken to claim success on a specific set of structures. But the averaged results over all structure-frames generated in each of the three test scenarios (see Fig. 5.10) give a feeling about how far the approach could affect the different model qualities in the total overview. In case **A**, the different sequence coupling result showed that the augmentation of the simulation was indeed successful, for it was able to generate a significant better refinement compared to **B** and to **C**, the simple coupling and the free simulations, respectively.

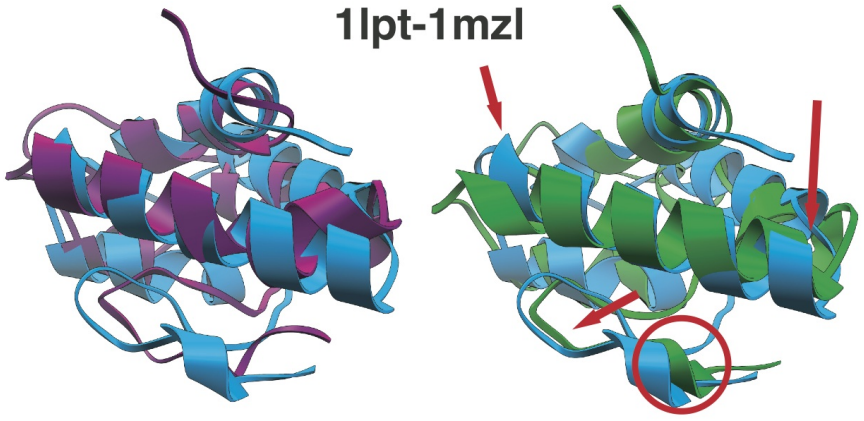

Figure 5.12: Model improvement 1mzl. The crystal structure in **blue**, the homology model in **purple** and the refined model in **green**. The red arrows point out sites with major improvements.

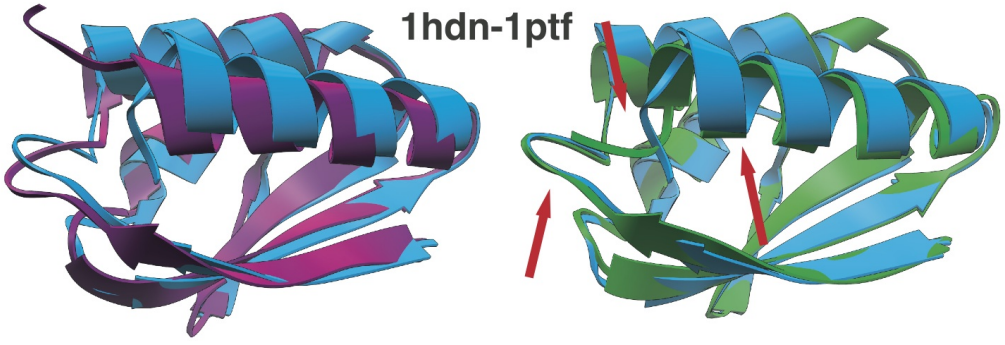

Figure 5.13: Model improvement 1ptf. The crystal structure in **blue**, the homology model in **purple** and the refined model in **green**. The red arrows point out sites with major improvements.

What refinement in the light of energies from compact scoring or ADPt really means can be seen in the structural overlay of models beginning with Fig. 5.11. It shows the crystal structures in blue, the homology starting model in purple and the resulting refined structure from the different sequence coupling in green. For all of the models, those figures show the profound structural improvement possible with this methodology.

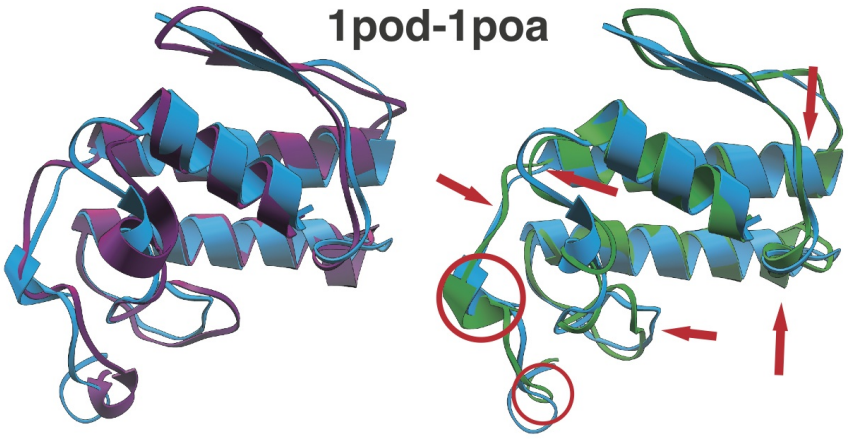

Figure 5.14: Model improvement 1poa. The crystal structure in **blue**, the homology model in **purple** and the refined model in **green**. The red arrows point out sites with major improvements.

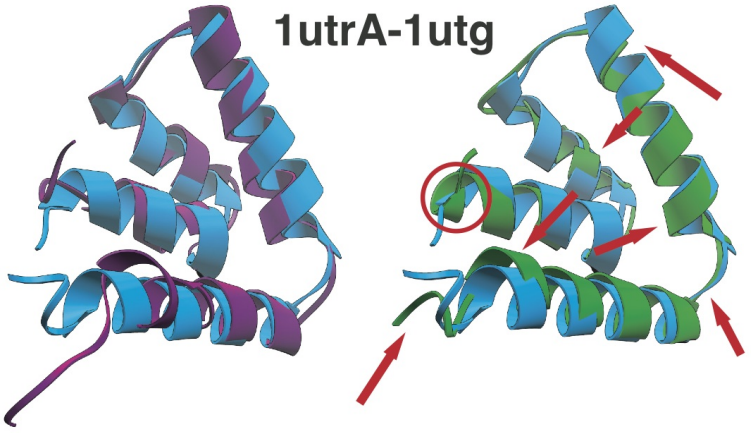

Figure 5.15: Model improvement 1utg. The crystal structure in **blue**, the homology model in **purple** and the refined model in **green**. The red arrows point out sites with major improvements.

The red arrows point to prominent sites within the structures which were refined. For model *1ak2*, the refinement happens consistently on the more correct helical part, moving the helices into the right positions, and interestingly, pushing the whole poorly build lower part closer to the crystal structure representation.

The smaller model *1mzl* was compacted by a lesser degree but still observable and marked by the red arrows.

An interesting fact that is true for all models is that the simulations were able to more or less precisely reproduce the secondary structure, but for some, the details in extended loop structures are a bit off like in model 5.12 where, for example, the terminal end

sticks out or, for model 5.14, where the upper long loop-like structure intermitted by a small β-sheet is not modeled very precisely.

The opposite is true for models 5.13 and 5.15. Due to their relatively small size and compact shape, it seems that they are also very precisely modeled in the loop regions.

5.3 Discussion

Refining a protein structure with classical MD simulations means searching for a deeper minimum on a multidimensional energy landscape which is a difficult problem.

With our sequence coupling approach, refinement often happened in the very beginning of the simulations, indicating that the added homologous sequences impose an additional force on the model which directly pulls atoms over energy barriers, thereby speeding up the improvement. After 4 - 5 ns simulation time all simulations have either reached a more native state (4/5 simulations) or stayed stable at a *null* refined state (1/5 simulations). It is interesting, that even the systems coupled with the same sequence show improvement in the very beginning (2 ns) of the simulations, but then decrease by on average about 0.1 Å dRMSD (see Fig. 5.10). The free simulations decrease in quality significantly faster and reach a plateau after 4 ns between 0.6 and 0.7 Å dRMSD deterioration. Simulations coupled to homologous sequences show no convergence at the end of our simulations at approximately 10 ns. It is possible that longer simulations can improve models even more. The long term stability needed will then be a direct effect of the strength of the implied propensities of the homolog sequences and therefore the properties of the energy landscape. The observable improvement of the models can be regarded as an effect of the added homolog sequences. Our biggest model (220 residues) was a special case to refine as it was modeled very accurately in a major part of the system and needed a significant improvement in the region of residue 129 to 167 (17.3 % of the whole model), which contributed to the deterioration with 0.87 Å dRMSD (2.79 Å instead of 1.92 Å when measured without that region). The task of refining that model was therefore to keep a large region in place and reshape a smaller area. Doing this with homolog sequences means, broadly, that its necessary to introduce less modifications in a very large part of the protein and apply more modifications to the problematic region. Since the modeled homologous structure is related to the homologous sequences taken

for modeling but also and in the same way, to the homologous sequences coupled within our simulations, the steps necessary to improve a desired region should be taken care of automatically because the homolog modeling process is just a depiction of the homolog structure space.

Normally, none of the energies produced by an all-atom explicit solvent simulation can be used unmodified to rank the quality or nativeness of a produced model to find the global energy minimum. In contrast, the restraint energies and the energies of our compactness score are able to identify the regime in which the refinement takes place. Broad averages of mean energies do give an insight to the overall quality of a model, making it possible to reveal the average distance to the native state and rank the difficulty of the refinement problem (see Fig. 5.8).

Which and how many sequences have to be coupled to achieve the best and maximum refinement is debatable. The effect of different sequences on the same setup has not been investigated in this work, but possible scenarios can be drawn very quickly. Several degrees of identity between the model and the added homolog sequences can be tested successively. Also, sequences with specific mutations in predefined domains or regions of the homology model can be selected.

Homology modeling approaches become more accurate when more homolog sequences with resolved structural information exist. With this, the refinement problem would decrease qualitatively, but probably remain equally difficult, because "closer to the native state" does not mean "easier to refine". Ultimately, using coupled different sequences in MD simulation for refinement makes sense when and if there exist more sequences than resolved structures to a given problem (a scenario which will be true for a long time), because solved structures improve the homology modeling process directly and do not need to be incorporated into a refinement simulation via a homolog sequence. In this way, no more coupling is needed to achieve a spacial improvement, since a significant rectification which originated from an evolutionary source would have already been imposed on the model. It may be questionable, if more homolog information in form of the sequences coupled to a simulation will yield more improvement. On the other hand, coupled simulations can still modify the underlying force field in a way, that can make the simulations more stable and better adjusted to a given problem, for example for a docking simulation aiming at drug design.

Here, the sequence selection process is performed by simply aiming for a minimal sequence identity via sequence comparison methods that still gives reasonable structural similarity to the targeted homology model. Various other methods for finding models that are more distantly related but still share a dominant fraction of the structure are existing [124]. Those methods and furthermore incorporating information about the evolutionary distance of the focussed sequences will have an impact on the stability and efficiency of the augmentation of our coupling approach.

It is possible to further improve this method by picking a structure produced by our augmented simulations and put those into the refinement process again, either with the previously used sequences or with new homologs. With this strategy it would be possible to use more evolutionary sequence data without bloating the simulation box at the same time, especially when dealing with proteins bigger than 200 residues. The bigger the box, the more time is wasted by unnecessarily simulating solvent molecules. By taking the same model and refine it again, more regions of the model could be improved, for it is possible to surmount energy barriers step by step with each iteration of the simulations.

Given the short simulation time of 10 ns, the impact of the coupled sequences, similar or different, is clearly observable. The information perturbation/disturbing effect when using homolog sequences within MD simulations was carried out on the basis of 5 comparative models, placing 8 sequences within the box and couple them via our integrated algorithm. The different naturally occurring sequences additionally do have a huge impact on the ability to refine a given structure.

With sequence identity of around 60 % towards the homology model and themselves, respectively, an increase of approximately 25 % of improvement was accomplished.

Bringing evolutionary information into MD simulations to refine homology models should work well if these setups meet one prerequisite: The targeted model should not include any crystal contacts, since those interactions cannot be compensated by anything else then the contact itself. In all other cases, smoothing the energy landscape to overcome barriers will be enhanced by introducing more data to the simulations [125, 126].

Chapter 6

Discussion

In this thesis, refinement of proteins derived from homology modeling methods was attempted.

The concluding result is that all simulation approaches were able to refine the given problem sets above the *null* refinement level. Improvement of the quality of selecting native-like structures from large ensembles has been achieved with the compact scoring function, which in general is able to allow picking of improved structures out of the best $\frac{1}{3}rd$ of a large dense ensemble, but sometimes also points exactly to the most native structures of the ensemble.

General practice, once a model was derived, was, and mostly still is, to not alter it afterwards because chances are high that modifications would just worsen the model instead of improving it. That was also the result of the last two CASP meetings (9 and 10), broadly. Constant reliable refinement, if at all, was just achievable on a very small level. Though at CASP10, an approach which for each model used informations given by the assessors, ranked best upon all contestants. They used a molecular dynamics simulation approach and added restraints or constraints to simply hold the unmarked regions to enable refinement just at the specified sites. Our approaches do not use any given guiding information since we can not rely on those to be available all the time, for example in real applications using sequences with unknown structures. So our methods will stay unbiased, neutral and universally applicable. Although we did not want to be dependent on external information about the improvable sites within a

model, we have means to elucidate regions within the protein which may need particular attention. During our simulations, residues which do not converge to a certain position with a positional variance close to the mean variance of the converged residues signal us that those sites are possible candidates for refinement. To know where to address the improvement in a model is a very important part in any refinement process.

Generally, without prior investigational positional variance check simulations like in our case, it is not known which parts of a protein are modeled close to the native representation or very badly, far away from their ideal placement. This is one of the main fundamental problems for a refinement approach. By looking at the mean square fluctuations of the complete C_α atom set of a molecular simulation ensemble, we do have a tool at hand that could tell us the overall quality of a targeted homology model. High fluctuations point to a dissatisfied model, which has most probably a higher RMSD than models with a lower overall fluctuation. Even more light is shed on the general quality of a model when we look at the averaged ADPt energies. The average value calculated from the whole ensemble gives us a very good correlation to its real quality (see Fig. 4.14). With this value at hand, it would be possible to extent the method with an estimator, to assess the refinements made. In that sense, if the differences between the initial and the refined model are greater than the estimator predicts, it may be possible that the refinement modified too much.

In the past, refinement was tried with many computational methods available, for example Monte Carlo approaches, all kinds of molecular dynamics sampling approaches, in each case with more or less success. As stated above by results of the recent CASP meetings, non of the methods have been overwhelmingly successful, yet. All of our refinement approaches used molecular dynamics simulations to generate an ensemble of structures. Since ensembles generated by free molecular dynamics simulations tend to follow unstructured, chaotic paths, restraints were added to help structures to stay in shape. Generally, when this is done with classical positional restraints, the improvement will be limited by the strength of the force constant used for the restraints. If chosen to low, the difference between a free run and a restrained run is marginal. On the other hand, if the force constant is too high, not enough movement is allowed, so the structural changes will be low. Hence, a good approach is to choose, by whatever criterion, which atoms have to be hold in place because their position is almost native and which have to be more or less free to achieve some improvement. Our ADPt position restraints, either

in the single copy or in the evolutionary sequence enhanced approach, aim to allow free movement in low quality sites and lesser, more restrictive movements of atoms in high quality regions of the protein.

In the first realization step, the system for the CASP9 refinement approach consisted of eight copies of the same protein, with distance restraints between all corresponding C_α atoms. The short simulations were repeated 1000 times with different starting seeds to ensure variations in the initial conditions. The resulting 8000 structures were clustered with the Jarvis-Patrick clustering [54]. We picked structures from the most populated cluster by performing a mixture of H-bond and Ramachandran score. The result of the first approach showed, that a well adjusted coupled simulation with low temperatures, even with very short simulations is capable to refine the given structures substantially.

In the ADPt system, only one model was simulated at a time. It did not use any distance restraints but pure adaptable position restraints, which were implemented in Gromacs to profit from the simulation speed achievable only without distance restraints. Protein models profited from the changed dynamics and enhanced the generation of improved structures. We also tried to investigate the effect of a moderate simulated annealing treatment, which indeed raised the probability of finding improved structures within our ensemble. Overall, the ensembles of the single copy ADPt method contained 25 % improved structures, enhanced through the simulated annealing this raised to 31 %. The free untreated simulations just improved 12 % of all structures of the ensemble.

To filter improved from worsened structures, the compact score helped to directly pick structures from a very large ensemble. This was a drastic improvement compared to the clustering filter of the CASP method.

The third approach, in which we used the coupled sequences to improve the related homology models, profited from all insights of the previous accomplishments. Due to the bigger systems, without a fast implementation, simulation times would have been too long. The Gromacs ADPt implementation made this approach feasible. The sequences used were not picked with regard to their biological evolutional relation. It was just assured that the sequences did share some degree of similarity and were not too similar. It is conceivable that a closer relation between the sequences or a more systematic selection of them might have an impact on the improvement capability. Another point is the amount of sequences per refinement simulation. We added seven sequences to

our setups, but this was a value derived from our previous distance restraint method, which took eight models because of symmetry reasons, a consideration not important any more in this case. We can think of a quality criterion, that could define the amount of added sequences. The better the quality of the structure supposedly is, the more sequences one might have to add (or the higher κ could be set). Altogether, 74 % of the structures produced with the sequence augmentation were improved models, whereas a coupling of the same structure (resembling the CASP approach with the improved ADPt methodology) resulted in 51 % improvement. Again, the free simulations improved just 15 % of all structures.

The most effective and successful approach among all was the incorporation of sequence information into the simulations. Here we could show that the incorporated sequences had a tremendous effect on the quality of the produced frames in the first 10 ns of our simulations. Even after just 2 ns of simulation time, almost all models improved already. Compared to the free runs that were not able to improve within the given time, this is a remarkable success.

What were the problems? Finding a structure in its global minimum is difficult and most often just a simplification of the given problem. There is not just only one native state that is the solution for the problem, but a collection, itself an ensemble of possible outcomes and solutions. Those ensembles around the global minimum represent a narrowed set of microstates, all with the same possibility of being almost equally probable. For the refinement of a given structure that may have several implications. Below a certain threshold, defined by the complexity and shape of the given problem, it is not only difficult but elusive to refine further, because we are not searching for one single structure but for an ensemble of conformations. The problem is to know when to finish the search. When is the problem solved, when does refining not make any more sense, because within the refined ensemble, all structures are active and exist in a cell and solve the given biological task equally well. So the sampled space can still be optimized. Though the structure may be deteriorated, for example at higher temperatures, chances are higher to also generate enough native like structures. But those would be very difficult to separate from the very similar wrong models. By sharpening the scoring function on the fine grained level, the sieving of good structures may be more reliable. This could be achieved by including or combining the existing function with other available scores, like a Ramachandran or H-bond score, or an accessible surface approach.

I suggest, that our methods featured here, especially the sequence augmented approach, will be very helpful in improving homology models, even though the generated homology models may improve over time due to better templates that will be accessible. For a long foreseeable time, the availability of sequences will be larger than the amount of solved structures. Consequently there will always be room for improvement by using the information that is contained within the primary structure of proteins.

Appendix A

Gromacs implementation

Gromacs (modified), version 4.5.3.

We wanted the modification of the gromacs default code to be as non-invasive as possible to ensure that the program remains to be fast and easy to maintain.

Apart from minor changes throughout a couple of files, the main changes were done in the decomposition setup/maintaining routine file *domdec.c*, the main simulation routine *md.c* and the communication initiation file *domdec_network.c* (see figure A.1).

Our modified version of gromacs can be used for any purpose, either for normal simulations with or without normal position restraints, for our single copy ADPt simulations or the multi copy sequence augmented ADPt simulations.

Parameters for our special purpose simulations can be set in the .mdp file (main grompp input parameter file). The connections for the sequence augmented simulations are defined by placing the special connection file *disres.con* in the working directory of the simulation. Each row of this file connects the interacting, corresponding atoms. So the sum of the rows of this file match the total number of residues, and the column elements are the individually connected atoms. The force constants needed for the coupling are set in the standard "posre.itp" file which is also used to introduce the interacting atoms itself.

As far as we could test it, our modifications had no negative effect on the speed of the program, neither in default simulations, nor in our special setups.

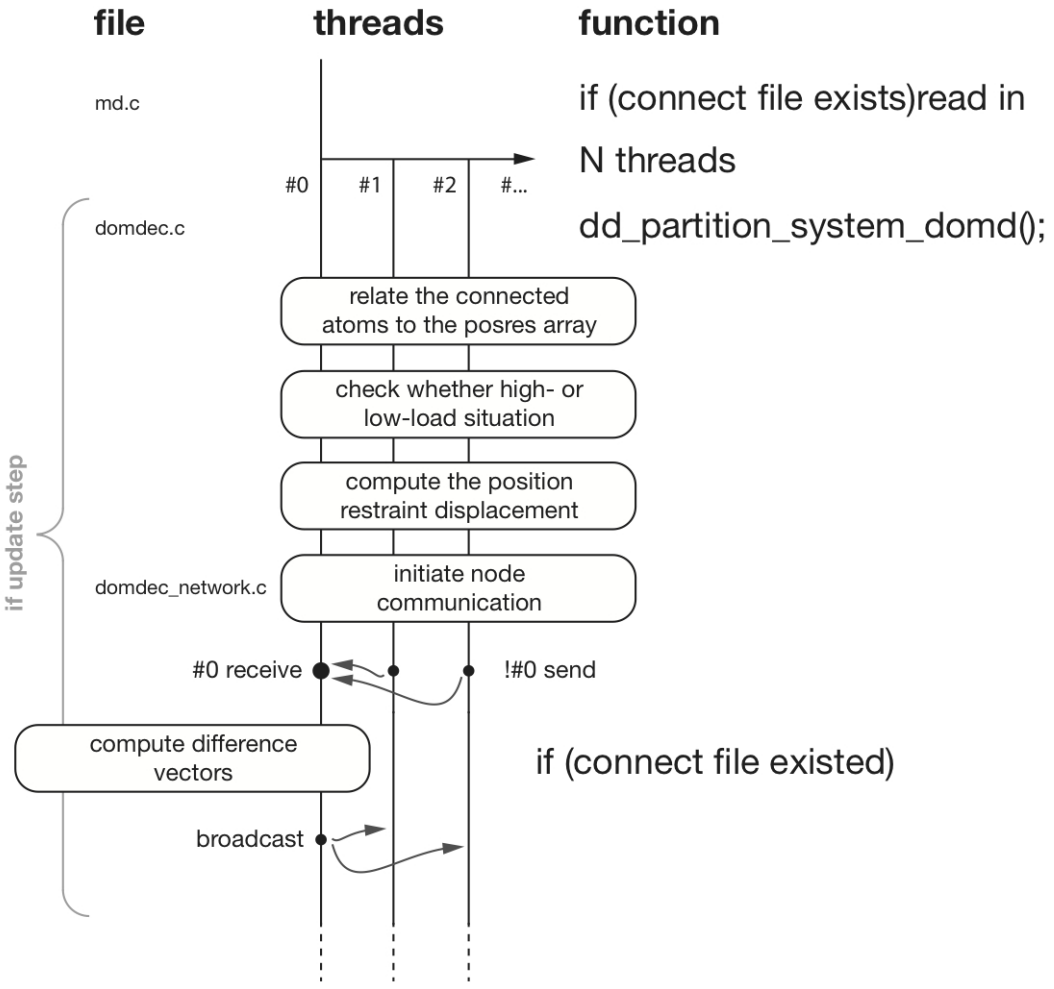

Figure A.1: Schematic simplified gromacs implementation for the adaptive restraints.

```
if(dd->ga2la->lal){

    int myi_mod;
    for(myi=0;myi<mydfmax_atm;myi++){

        myi_mod = myi % dd->ga2la->mod;

        if( dd->ga2la->lal[myi_mod].ga < mydfmax_atm &&
dd->ga2la->lal[myi_mod].ga >= 0 && dd->ga2la->lal[myi_mod].cell == 0 &&
dd->gatindex[dd->ga2la->lal[myi_mod].la] == myi ){

            top_global->molblock[...].posres_xA[...][XX]+=(state_local->x[...][XX]
                    -top_global->molblock[...].posres_xA[...][XX])*ir->den_kappa;
            top_global->molblock[...].posres_xA[...][YY]+=(state_local->x[...][YY]
                    -top_global->molblock[...].posres_xA[...][YY])*ir->den_kappa;
            top_global->molblock[...].posres_xA[...][ZZ]+=(state_local->x[...][ZZ]
                    -top_global->molblock[...].posres_xA[...][ZZ])*ir->den_kappa;
        }

    }

}
```

Listing A.1: high load update

```
if(dd->ga2la->laa){

    for(myi=0;myi<mydfmax_atm;myi++){

        if(dd->ga2la->laa[myi].cell==0){

            top_global->molblock[...].posres_xA[...][XX]+=(state_local->x[...][XX]
                    -top_global->molblock[...].posres_xA[...][XX])*ir->den_kappa;
            top_global->molblock[...].posres_xA[...][YY]+=(state_local->x[...][YY]
                    -top_global->molblock[...].posres_xA[...][YY])*ir->den_kappa;
            top_global->molblock[...].posres_xA[...][ZZ]+=(state_local->x[...][ZZ]
                    -top_global->molblock[...].posres_xA[...][ZZ])*ir->den_kappa;
        }

    }

}
```

Listing A.2: low load update

```
for(myi=0;myi<mynr_proteins;myi++){

    if(MASTER(cr)){

        for(ii=1;ii<dd->nnodes;ii++){

            dd_recv(dd,3*top_global->molblock[myi].nposres_xA,&pos_coords[
(ii * mynr_proteins * mydfmaxii_atm * 3) + (myi * mydfmaxii_atm * 3) ],ii,123);
```

```
9           }
10
11      }
12
13      else{
14
15          dd_send(dd,3*top_global->molblock[myi].nposres_xA,top_global->molblock[myi]
16                .posres_xA,123);
17
18      }
19
20  }
```

Listing A.3: communication, send - receive

```
1   for(myi=0;myi<mynr_proteins;myi++){
2
3       for(myh=0;myh<top_global->molblock[myi].nposres_xA;myh++){
4
5           dd_bcast(dd,sizeof(top_global->molblock[myi].posres_xA[myh]),top_global->
6                 molblock[myi].posres_xA[myh]);
7
8       }
9
10  }
```

Listing A.4: communication, final broadcast

Bibliography

[1] Kryshtafovych, A., Krysko, O., Daniluk, P., Dmytriv, Z. & Fidelis, K. Protein structure prediction center in CASP8. *Proteins* **77**, 5–9 (2009).

[2] Levitt, M. Nature of the protein universe. *PNAS* **106**, 11079–11084 (2009).

[3] Epstein, C. J., Goldberger, R. F. & Anfinsen, C. B. The Genetic Control of Tertiary Protein Structure: Studies with Model Systems. *Cold Spring Harbor Symposia on Quantitative Biology* **28**, 439–449 (1963).

[4] Anfinsen, C. B. Principles that Govern the Folding of Protein Chains. *Science* **181**, 223–230 (1973).

[5] Guzzo, A. V. The influence of Amino-Acid Sequence on Protein Structure. *Biophysical Journal* **5**, 809–822 (1965).

[6] Eyes wide open. *Nature Chemical Biology* **5**, 773 (2009).

[7] Karplus, M. & McCammon, J. A. Molecular dynamics simulations of biomolecules. *Nature* **9**, 646–652 (2002).

[8] Bolognesi, M., Smith, J. L., Bolognesi, M. & Smith, J. L. Proteins: ever larger, stranger and more dynamic. *Current Opinion in Structural Biology* 690–692 (2008).

[9] Klepeis, J. L., Lindorff-Larsen, K., Dror, R. O. & Shaw, D. E. Long-timescale molecular dynamics simulations of protein structure and function. *Current Opinion in Structural Biology* 120–127 (2009).

[10] Travaglini-Allocatelli, C., Ivarsson, Y., Jemth, P. & Gianni, S. Folding and stability of globular proteins and implications for function. *Current Opinion in Structural Biology* 3–7 (2009).

[11] Orengo, C. A. & Thornton, J. M. Protein Families and their Evolution - a Structural Perspective. *Annual Review of Biochemistry* **74**, 867–900 (2005).

[12] Chothia, C. One thousand families for the molecular biologist. *Nature* **357**, 543–544 (1992).

[13] Alva, V., Remmert, M., Biegert, A., Lupas, A. N. & Söding, J. A galaxy of folds. *Protein Science* **19**, 124–130 (2010).

[14] Lesk, A. M., Conte, L. L. & Hubbard, T. J. P. Assessment of Novel Fold Targets in CASP4: Predictions of Three-Dimensional Structures, Secondary Structures, and Interresidue Contacts. *Proteins* **45**, 98–118 (2001).

[15] Hardin, C., Pogorelov, T. V. & Luthey-Schulten, Z. Ab initio protein structure prediction. *Current Opinion in Structural Biology* **12**, 176–181 (2002).

[16] Jothi, A. Principles, Challenges and Advances in ab initio Protein Structure Prediction. *Protein and Peptide Letters* **19**, 1194–1204 (2012).

[17] Tozzini, V. Coarse-grained models for proteins. *Current Opinion in Structural Biology* **15**, 144–150 (2005).

[18] Chothia, C. & Lesk, A. M. The relation between the divergence of sequence and structure in proteins. *EMBO Journal* **5**, 823–826 (1986).

[19] Mosimann, S., Meleshko, R. & James, M. N. G. A Critical Assessment of Comparative Molecular Modeling of Tertiary Structures of Proteins *. *Proteins* **23**, 301–317 (1995).

[20] Sali, A. & Blundell, T. L. Comparative Protein Modelling by Satisfaction of Spatial Restraints. *Journal of Molecular Biology* **234**, 779–815 (1993).

[21] Li, L., Darden, T., Foley, C., Hiskey, R. & Pedersen, L. Homology modeling and molecular dynamics simulation of human prothrombin fragment 1. *Protein Science* **52**, 2341–2348 (1995).

[22] Sanchez, R. & Sali, A. Evaluation of Comparative Protein Structure Modeling by MODELLER-3. *Proteins* **1**, 50–58 (1997).

[23] Moult, J., Fidelis, K., Zemla, A. & Hubbard, T. Critical Assessment of Methods of Protein Structure Prediction (CASP): Round IV. *Proteins* **5**, 2–7 (2001).

[24] Mart, M. A., Stuart, A. C., S, R., Melo, F. & Sali, A. Comparative Protein Structure Modeling of Genes and Genomes. *Annu. Rev. Biophys. Biomol. Struct.* 291–325 (2000).

[25] Eswar, N., Eramian, D., Webb, B., Shen, M.-y. & Sali, A. *Protein Structure Modeling with Modeller*, vol. 426, chap. 8, 145–159 (Totowa, 2006).

[26] Misura, K. M. S. & Baker, D. Progress and Challenges in High-Resolution Refinement of Protein Structure Models. *Proteins* **59**, 15–29 (2005).

[27] Melo, F. & Sali, A. Fold assessment for comparative protein structure modeling. *Protein Science* **16**, 2412–2426 (2007).

[28] Xiang, Z. Advances in Homology Protein Structure Modeling. *Current Protein and Peptide Science* **7**, 217–227 (2006).

[29] Krieger, E. *et al.* Improving physical realism, stereochemistry, and side-chain accuracy in homology modeling: Four approaches that performed well in CASP8. *Proteins* **77**, 114–122 (2009).

[30] Bordoli, L. *et al.* Protein structure homology modeling using SWISS-MODEL workspace. *Nature Protocols* **4**, 1–13 (2009).

[31] Kaufmann, K. W., Lemmon, G. H., Deluca, S. L., Sheehan, J. H. & Meiler, J. Practically Useful: What the ROSETTA Protein Modeling Suite Can Do for You. *Biochemistry* **49**, 2987–2998 (2010).

[32] Misura, K. M. S., Chivian, D., Rohl, C. A., Kim, D. E. & Baker, D. Physically realistic homology models built with ROSETTA can be more accurate than their templates. *PNAS* **103**, 5361–5366 (2006).

[33] Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. SCOP: A Structural Classification of Proteins Database for the Investigation of Sequences and Structures. *Journal of Molecular Biology* **247**, 536–540 (1995).

[34] Schonbrun, J., Wedemeyer, W. J. & Baker, D. Protein structure prediction in 2002. *Current Opinion in Structural Biology* **12**, 348–354 (2002).

[35] Han, R. *et al.* An efficient conformational sampling method for homology modeling. *Proteins* **71**, 175–188 (2008).

[36] Fan, H. & Mark, A. E. Refinement of homology-based protein structures by molecular dynamics simulation techniques. *Protein Science* **13**, 211–220 (2003).

[37] Lee, M. R., Tsai, J., Baker, D. & Kollman, P. A. Molecular dynamics in the endgame of protein structure prediction. *Journal of Molecular Biology* **313**, 417–430 (2001).

[38] Lin, M. S. & Head-Gordon, T. Reliable Protein Structure Refinement Using a Physical Energy Function. *Journal of Computational Chemistry* **32**, 709–717 (2010).

[39] Hess, B., Kutzner, C., Spoel, D. v. d. & Lindahl, E. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *Journal of Chemical Theory and Computation* **4**, 435–447 (2008).

[40] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics* **21**, 1087–1092 (1953).

[41] Kryshtafovych, A., Fidelis, K. & Moult, J. CASP9 results compared to those of previous CASP experiments. *Proteins* **79**, 196–207 (2011).

[42] Moult, J. The curent state of the art in protein structure prediction. *Current Opinion in Biotechnology* **7**, 422–427 (1996).

[43] Moult, J. Predicting protein three-dimensional structure. *Current Opinion in Biotechnology* **10**, 583–588 (1999).

[44] Kryshtafovych, A. & Fidelis, K. Protein structure prediction and model quality assessment. *Drug Discov Today* **14**, 386–393 (2010).

[45] Lu, H. & Skolnick, J. Application of Statistical Potentials to Protein Struture Refinement from Low Resolution ab initio Models. *Biopolymers* **70**, 575–584 (2003).

[46] Huber, T. & Gunsteren, W. F. V. SWARM-MD: Searching Conformational Space by Cooperative Molecular Dynamics. *Journal of Physical Chemistry A* **5639**, 5937–5943 (1998).

[47] Spoel, D. v. d., Lindahl, E., Hess, B. & Groenhof, G. GROMACS: Fast, Flexible, and Free. *Journal of Computational Chemistry* **26**, 1701–1718 (2005).

[48] Duan, Y. *et al.* A Point-Charge Force Field for Molecular Mechanics Simulations of Proteins Based on Condensed-Phase Quantum Mechanical Calculations. *Journal of Computational Chemistry* **24**, 1999–2012 (2003).

[49] Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *Journal of Chemical Physics* **79**, 926–935 (1983).

[50] Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and Development of Coot. *Acta Crystallographica Section D - Biological Crystallography* **66**, 486–501 (2010).

[51] Zemla, A. LGA: a method for finding 3D similarities in protein structures. *Nucleic Acids Research* **31**, 3370–3374 (2003).

[52] Keedy, D. A. *et al.* The other 90% of the protein: Assessment beyond the Calphas for CASP8 template-based and high-accuracy models. *Proteins* **77**, 29–49 (2009).

[53] Chen, V. B. *et al.* MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallographica* **Section D**, 12–21 (2010).

[54] Jarvis, R. A. & Patrick, E. A. Clustering Using a Similarity Measure Based on Shared Near Neighbors. *IEEE Transactions on Computers* **C-22**, 1025–1034 (1973).

[55] Ramelot, T. A. *et al.* Improving NMR Protein Structure Quality by Rosetta Refinement: A Molecular Replacement Study. *Proteins* **75**, 147–167 (2009).

[56] Raman, S. *et al.* Structure prediction for CASP8 with all-atom refinement using Rosetta. *Proteins* **77**, 89–99 (2009).

[57] Cooper, S. *et al.* Predicting protein structures with a multiplayer online game. *Nature* **466**, 756–760 (2010).

[58] MacCallum, J. L. *et al.* Assessment of protein structure refinement in CASP9. *Proteins* **79**, 74–90 (2011).

[59] Laskowski, R. A., Watson, J. D. & Thornton, J. M. From protein structure to biochemical function? *Journal of Structural and Functional Genomics* **4**, 167–177 (2003).

[60] Zhang, C. & Kim, S.-H. Overview of structural genomics: from structure to function. *Current Opinion in Chemical Biology* **7**, 28–32 (2003).

[61] Karplus, M. & Kuriyan, J. Molecular dynamics and protein function. *PNAS* **102**, 6679–6685 (2005).

[62] Jones, D. T. Progress in protein structure prediction. *Current Opinion in Structural Biology* **7**, 377–387 (1997).

[63] Bonneau, R., Tsai, J., Ruczinski, I. & Baker, D. Functional Inferences from Blind ab Initio Protein Structure Predictions. *Journal of Structural Biology* **134**, 186 –190 (2001).

[64] Sternberg, M. J. E., Bates, P. A., Kelley, L. A. & MacCallum, R. M. Progress in protein structure prediction: assessment of CASP3. *Current Opinion in Structural Biology* **9**, 368–373 (1999).

[65] Bonneau, R. & Baker, D. Ab Iinitio Protein Structure Prediction: Progress and Prospects. *Annu. Rev. Biophys. Biomol. Struct.* **30**, 173–189 (2001).

[66] Xu, D. & Zhang, Y. Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins* **80**, 1715–1735 (2012).

[67] Ding, F., Tsao, D., Nie, H. & Dokholyan, N. V. Ab Initio Folding of Proteins with All-Atom Discrete Molecular Dynamics. *Structure* **16**, 1010–1018 (2008).

[68] Browne, W. J. *et al.* A Possible Three-dimensional Structure of Bovine a-Lactalbumin based on that of Hens Egg-White Lysozyme. *Journal of Molecular Biology* **42**, 65–86 (1969).

[69] Zhou, H. *et al.* Analysis of TASSER-based CASP7 protein structure prediction results. *Proteins* **69**, 90–97 (2007).

[70] Chen, J. & Brooks, C. L. Can Molecular Dynamics Simulations Provide High-Resolution Refinement of Protein Structure? *Proteins* **67**, 922–930 (2007).

[71] Van Gunsteren, W. F. & Berendsen, H. J. C. Computer Simulation of Molecular Dynamics: Methodology, Applications, and Perspectives in Chemistry. *Angewandte Chemie International Edition* **29**, 992–1023 (1990).

[72] Raval, A., Piana, S., Eastwood, M. P., Dror, R. O. & Shaw, D. E. Refinement of protein structure homology models via long, all-atom molecular dynamics simulations. *Proteins* **online**, online (2012).

[73] Zhu, J., Fan, H., Periole, X., Honig, B. & Mark, A. E. Refining homology models by combining replica-exchange molecular dynamics and statistical potentials. *Proteins* **72**, 1171–1188 (2008).

[74] Fan, H., Periole, X. & Mark, A. E. Mimicking the action of folding chaperones by Hamiltonian replica-exchange molecular dynamics simulations: Application in the refinement of de novo models. *Proteins* **80**, 1744–1754 (2012).

[75] Lindorff-Larsen, K. *et al.* Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins* **78**, 1950–1958 (2010).

[76] Beauchamp, K. A., Lin, Y.-s., Das, R. & Pande, V. S. Are Protein Force Fields Getting Better? A Systematic Benchmark on 524 Diverse NMR Measurements. *Journal of Chemical Theory and Computation* 1–19 (2012).

[77] Lindorff-Larsen, K. *et al.* Systematic Validation of Protein Force Fields against Experimental Data. *PLoS ONE* **7**, 1–6 (2012).

[78] Piana, S., Lindorff-Larsen, K. & Shaw, D. E. How Robust Are Protein Folding Simulations with Respect to Force Field Parameterization? *Biophysical Journal* **100**, L47–L49 (2011).

[79] Dill, K. A. Polymer principles and protein folding. *Protein Science* **8**, 1166–1180 (1999).

[80] Bitetti-Putzer, R., Dinner, A. R., Yang, W. & Karplus, M. Conformational sampling via a self-regulating effective energy surface. *The Journal of Chemical Physics* **124**, 174901–1,174901–15 (2006).

[81] Levinthal, C. Mossbauer Spectroscopy in Biological Systems: Proceedings of a meeting held at Allerton House, Monticello, Illinois. *University of Illinois Press 1969* 22–24 (1969).

[82] Zwanzig, R., Szabo, A. & Bagchi, B. Levinthal' s paradox. *PNAS* **89**, 20–22 (1992).

[83] Karplus, M. The Levinthal paradox: Yesterday and today. *Folding and Design* **2 Suppl**, 69–75 (1997).

[84] Dill, K. A. & Chan, H. S. From Levinthal to pathways to funnels. *Nature Structural Biology* **4**, 10–19 (1997).

[85] Dill, K. A., Ozkan, S. B., Weikl, T. R., Chodera, J. D. & Voelz, V. A. The protein folding problem: when will it be solved. *Current Opinion in Structural Biology* **17**, 342–346 (2007).

[86] Brown, W. M., Martin, S., Pollock, S. N., Coutsias, E. A. & Watson, J.-P. Algorithmic dimensionality reduction for molecular structure analysis. *The Journal of Chemical Physics* **129**, 1–13 (2008).

[87] Schroeder, G. F., Brunger, A. T. & Levitt, M. Combining Efficient Conformational Sampling with a Deformable Elastic Network Model Facilitates Structure Refinement at Low Resolution. *Structure* **15**, 1–12 (2007).

[88] Darden, T., York, D. & Pedersen, L. Particle mesh Ewald: An N log(N) method for Ewald sums in large systems. *Journal of Chemical Physics* **98**, 10089–10092 (1993).

[89] Essmann, U., Perera, L. & Berkowitz, M. L. A smooth particle mesh Ewald method. *Journal of Chemical Physics* **103**, 8577–8593 (1995).

[90] Nose, S. A unified formulation of the constant temperature molecular dynamics methods. *Journal of Chemical Physics* **81**, 511–519 (1984).

[91] Hoover, W. G. Canonical dynamics: Equilibrium phase-space distributions. *Physical Review A* **31**, 1695–1697 (1985).

[92] Hess, B., Bekker, H., Berendsen, H. J. C. & Fraaije, J. G. E. M. LINCS: A Linear Constraint Solver for Molecular Simulations. *Journal of Computational Chemistry* **18**, 1463–1472 (1997).

[93] Hess, B. P-LINCS: A Parallel Linear Constraint Solver for Molecular Simulation. *Journal of Chemical Theory and Computation* **4**, 116–122 (2008).

[94] Park, B. H. & Levitt, M. Energy Functions that Discriminate X-ray and Near-native Folds from Well-constructed Decoys. *Journal of Molecular Biology* **258**, 367–392 (1996).

[95] Park, B. H., Huang, E. S. & Levitt, M. Factors Affecting the Ability of Energy Functions to Discriminate Correct from Incorrect Folds. *Journal of Molecular Biology* **266**, 831–846 (1997).

[96] Kryshtafovych, A., Fidelis, K. & Moult, J. CASP8 results in context of previous experiments. *Proteins* **77**, 217–228 (2009).

[97] MacCallum, J. L. *et al.* Assessment of the protein-structure refinement category in CASP8. *Proteins* **77**, 66–80 (2009).

[98] Moult, J., Fidelis, K., Kryshtafovych, A., Rost, B. & Tramontano, A. Critical assessment of methods of protein structure prediction - Round VIII. *Proteins* **77**, 1–4 (2009).

[99] Cozzetto, D. *et al.* Evaluation of template-based models in CASP8 with standard measures. *Proteins* **77**, 18–28 (2009).

[100] Cozzetto, D., Kryshtafovych, A. & Tramontano, A. Evaluation of CASP8 model quality predictions. *Proteins* **77**, 157–166 (2009).

[101] Moult, J., Fidelis, K., Kryshtafovych, A. & Tramontano, A. Critical assessment of methods of protein structure prediction (CASP) - Round IX. *Proteins* **79**, 1–5 (2011).

[102] Krieger, E., Darden, T., Nabuurs, S. B., Finkelstein, A. & Vriend, G. Making optimal use of empirical energy functions: force-field parameterization in crystal space. *Proteins* **57**, 678–683 (2004).

[103] Tyka, M. D. *et al.* Alternate States of Proteins Revealed by Detailed Energy Landscape Mapping. *Journal of Molecular Biology* **405**, 607–618 (2011).

[104] Kannan, S. & Zacharias, M. Application of biasing-potential replica-exchange simulations for loop modeling and refinement of proteins in explicit solvent. *Proteins* **78**, 2809–2819 (2010).

[105] Jagielska, A., Wroblewska, L. & Skolnick, J. Protein model refinement using an optimized physics-based all-atom force field. *PNAS* **105**, 82688273 (2008).

[106] Xia, Y. & Levitt, M. Funnel-Like Organization in Sequence Space Determines the Distributions of Protein Stability and Folding Rate Preferred by Evolution. *Proteins* **55**, 107–114 (2004).

[107] Xia, Y. & Levitt, M. Simulating protein evolution in sequence and structure space. *Current Opinion in Structural Biology* **14**, 202–207 (2004).

[108] Davis, I. W., Arendall, W. B., Richardson, D. C., Richardson, J. S. & Carolina, N. The Backrub Motion: How Protein Backbone Shrugs When a Sidechain Dances. *Structure* **14**, 265–274 (2006).

[109] Del Sol, A., Fujihashi, H., Amoros, D. & Nussinov, R. Residue centrality, functionally important residues, and active site shape: Analysis of enzyme and non-enzyme families. *Protein Science* **15**, 2120–2128 (2006).

[110] Dryden, D. T. F., Thomson, A. R. & White, J. H. How much of protein sequence space has been explored by life on Earth? *Journal of the Royal Society Interface* **5**, 953–956 (2008).

[111] Jacobson, M. & Sali, A. Comparative Protein Structure Modeling and its Applications to Drug Discovery. *Annual Reports in Medicinal Chemistry* **39**, 259–276 (2004).

[112] Kennedy, J. & Russell, E. Particle Swarm Optimization. *IEEE* 1942–1948 (1995).

[113] Eberhart, R. C. & Shi, Y. Particle Swarm Optimization: Developments, Application and Resources. *Evolutionary Computation* **1**, 81 – 86 (2001).

[114] Keasar, C. & Elber, R. Homology as a Tool in Optimization Problems: Structure Determination of 2D Heteropolymers. *Journal of Physical Chemistry* **99**, 11550–11556 (1995).

[115] Keasar, C., Elber, R. & Skolnick, J. Simultaneous and coupled energy optimization of homologous proteins: a new tool for structure prediction. *Current Biology* **2**, 247–259 (1997).

[116] Keasar, C., Tobi, D., Elber, R. & Skolnick, J. Coupling the folding of homologous proteins. *PNAS* **95**, 5880–5883 (1998).

[117] Nanda, V. & Degrado, W. F. Automated Use of Mutagenesis Data in Structure Prediction. *Proteins* **59**, 454 – 466 (2005).

[118] Finkelstein, A. V. 3D Protein Folds: Homologs Against Errors - a Simple Estimate Based on the Random Energy Model. *Physical Review Letters* **80**, 4823–4825 (1998).

[119] Bruce, N. J. & Bryce, R. A. Ab Initio Protein Folding Using a Cooperative Swarm of Molecular Dynamics Trajectories. *Journal of Chemical Theory and Computation* **6**, 1925–1930 (2010).

[120] Levitt, M. A simplified representation of protein conformations for rapid simulation of protein folding. *Journal of Molecular Biology* **104**, 59–107 (1976).

[121] Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**, 3389–3402 (1997).

[122] Eswar, N. *et al.* Comparative Protein Structure Modeling Using Modeller. *Current Protocols in Bioinformatics* 1–30 (2006).

[123] Chan, H. S. & Dill, K. A. Polymer Principles in Protein Structure and Stability. *Annual Review of Biophysics and Biophysical Chemistry* **20**, 447–490 (1991).

[124] Yu, L., White, J. V. & Smith, T. E. A homology identification method that combines protein sequence and structure information. *Protein Science* **7**, 2499–2510 (1998).

[125] Villa, A., Fan, H., Wassenaar, T. & Mark, A. E. How Sensitive Are Nanosecond Molecular Dynamics Simulations of Proteins to Changes in the Force Field? *Journal of Physical Chemistry B* **111**, 6015–6025 (2007).

[126] Thompson, J. & Baker, D. Incorporation of Evolutionary Information Into Rosetta Comparative Modeling. *Proteins* **79**, 2380–2388 (2011).