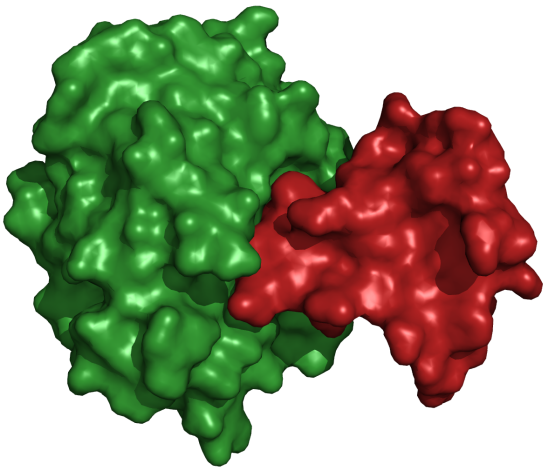
Technische Universität München
Fakultät für Physik
Lehrstuhl für Theoretische Biophysik (T38)

Flexible Docking Methods for Investigating Protein-Protein Interactions

Christina Eva Maria Schindler
Dipl.-Phys. (Univ.)



Technische Universität München
Fakultät für Physik
Lehrstuhl für Theoretische Biophysik (T38)

Flexible Docking Methods for Investigating Protein-Protein Interactions

Christina Eva Maria Schindler

Vollständiger Abdruck der von der Fakultät für Physik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktor der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender: Prof. Dr. Friedrich Simmel

Prüfer der Dissertation:

1. Prof. Dr. Martin Zacharias
2. Prof. Dr. Iris Antes
3. Prof. Dr. Alexandre Bonvin

Die Dissertation wurde am 24.08.2016 bei der Technischen Universität München eingereicht und durch die Fakultät für Physik am 02.12.2016 angenommen.

All models are wrong but some are useful.
George E. P. Box

What I cannot create, I do not understand.
Richard Feynman

Abstract

Protein-protein interactions are involved in virtually all important biological processes in the cell. Methods like X-ray crystallography or nuclear magnetic resonance spectroscopy have yielded atomic structural insights into many assemblies. However, experimental structure characterization is highly challenging and for a large fraction of complexes, atomic structural knowledge is lacking to date. Computational protein-protein docking methods aim to complement experimental studies by modeling the structure of protein-protein complexes from the structures of the individual constituents. In addition to elucidating unknown structures, designing accurate docking methods gives insights into the physical principles that govern non-covalent protein association and hence helps to understand the requirements for specific and nonspecific recognition in the cell. Due to the complexity of the problem, most docking programs keep the internal structure of the proteins rigid and explore only rotational and translational degrees of freedom. But in many cases, proteins undergo significant conformational changes upon binding and rigid-body docking may fail to correctly predict the native structure. Flexible docking methods aim to incorporate protein conformational flexibility, balancing the required level of detail with computational efficiency. This thesis presents several methodological advances that were implemented in the ATTRACT docking engine. The developed protocols expand ATTRACT's capabilities towards atomistic flexible docking and incorporating additional types of low-resolution experimental data. The protocols were tested on large benchmark sets and achieved significant improvements compared to previous approaches. This thesis also describes two applications of developed protocols in collaboration with experimental groups. In the first application, we studied peptide recognition in the chaperone cofactor ERdj5. In the second project, the full-length multi-domain structure of the flexible ISWI nucleosome remodeling enzyme was modeled. In the future, the developed protocols can be employed to study a broad variety of transient and flexible protein-protein complexes.

Zusammenfassung

Protein-Protein Wechselwirkungen sind an allen wichtigen zellulären Prozessen beteiligt. Durch Röntgenkristallographie und NMR Experimente konnte die Struktur von vielen Komplexen auf atomarer Skala aufgelöst werden. Allerdings stehen derartige Informationen aber für eine Vielzahl von Wechselwirkungen momentan nicht zur Verfügung. In Fällen, in denen die Struktur des Komplexes unbekannt ist aber experimentelle Strukturen für die einzelnen Komponenten vorhanden sind, kann der Bindungsmodus mit Protein-Protein Docking Computersimulationen vorhergesagt werden. Zudem bieten Dockingsimulationen und das Design von präzisen Dockingalgorithmen die Möglichkeit unser Verständnis von den physikalischen Kräften, die die Proteinassoziation bestimmen, zu testen und zu vertiefen. Das Protein-Protein Docking Problem ist hochkomplex, da es sich um Systeme mit vielen Atomen und dementsprechend vielen Freiheitsgraden handelt. Deshalb wird typischerweise in den meisten Dockingprogrammen die "Starre Körper" Näherung verwendet, die erlaubt den Bindungsprozess nur in Rotation und Translation des Schwerpunkts zu beschreiben. Allerdings verändern viele Proteine ihre Struktur, wenn sie an ihren Partner binden, und Docking in der "Starren Körper" Näherung kann in solchen Fällen oft die native Komplexstruktur nicht mehr korrekt vorhersagen. Flexible Dockingmethoden versuchen die Proteinflexibilität in der Strukturvorhersage in angemessenem Detail effizient zu berücksichtigen. In dieser Arbeit werden mehrere neue Erweiterungen des ATTRACT Dockingprogramms beschrieben, die es erlauben atomistische Flexibilität und zusätzliche experimentelle Daten während des Dockings einzubauen. Die entwickelten Protokolle wurden auf großen Benchmarks getestet und stellen klare Verbesserungen zu vorhandenen Methoden dar. Zudem wurden die neuen Protokolle in Kollaboration mit experimentellen Gruppen auf zwei spannende biologische Fragestellungen angewendet: 1. die Untersuchung des Peptiderkennungsmechanismus des Chaperoncofaktors ERdj5 im endoplasmatischen Retikulum, 2. die Modellierung der Struktur des Multidomänenproteins und Nukleosomremodelers ISWI. Die in dieser Arbeit vorgestellten Protokolle können in Zukunft zur Strukturvorhersage vieler kurzlebiger und flexibler Protein-Protein Komplexe verwendet werden.

Contents

Publications

Sjoerd J de Vries, Isaure Chauvot de Beauchêne, Christina EM Schindler, and Martin Zacharias. "Cryo-EM data is superior to contact or interface information in integrative modeling". *Biophysical Journal* 110.4 (2016), pp. 462–465. DOI: 10.1016/j.bpj.2015.12.038.

Sjoerd J de Vries, Christina EM Schindler, Isaure Chauvot de Beauchêne, and Martin Zacharias. "A web interface for easy flexible protein–protein docking with ATTRACT". *Biophysical Journal* 108.3 (2015), pp. 462–465. DOI: 10.1016/j.bpj.2014.12.015.

Nadine Harrer, Christina EM Schindler, Linda Bruetzel, Jan Lipfert, Martin Zacharias, and Felix Mueller-Planitz. "Integrative modeling of the full-length ISWI nucleosome remodeling enzyme using cross-linking/mass spectrometry and SAXS data" (2016), In preparation.

Johanna Ludwigsen, Sabrina Pfennig, Ashish K Singh, Christina EM Schindler, Nadine Harrer, Ignasi Forné, Martin Zacharias, and Felix Mueller-Planitz. "Concerted regulation of ISWI by an autoinhibitory domain and the H4 N-terminal tail". *eLife* 6 (2017), e21477. DOI: 10.7554/eLife.21477.

Alexander Sasse, Sjoerd J de Vries, Christina EM Schindler, Isaure Chauvot de Beauchêne, and Martin Zacharias. "Rapid design of knowledge-based scoring potentials for enrichment of near-native geometries in protein docking with the GRADSCOPT tool kit". *PLOS One* 12.1 (2017), e0170625. DOI: 10.1371/journal.pone.0170625.

Christina EM Schindler, Isaure Chauvot de Beauchêne, Sjoerd J de Vries, and Martin Zacharias. "Protein-protein and peptide-protein docking and refinement using ATTRACT in CAPRI". *Proteins: Structure, Function, and Bioinformatics* (2016), In press. DOI: 10.1002/prot.25196.

Christina EM Schindler, Sjoerd J de Vries, and Martin Zacharias. "Fully blind peptide-protein docking with pepATTRACT". *Structure* 23.8 (2015), pp. 1507–1515. DOI: 10.1016/j.str.2015.05.021.

Christina EM Schindler, Eva Stauffer, Thomas Mietzner, Klaus-Jürgen Schleifer, and Martin Zacharias. "Free energy calculations elucidate substrate binding and gating mechanism in herbicide target X" (2016), In preparation.

Christina EM Schindler, Sjoerd J de Vries, Alexander Sasse, and Martin Zacharias. "SAXS data alone can generate high-quality models of protein-protein complexes". *Structure* 24.8 (2016), pp. 1387–1397. DOI: 10.1016/j.str.2016.06.007.

Christina EM Schindler, Sjoerd J de Vries, and Martin Zacharias. "Development and implementation of a fully blind flexible peptide-protein docking protocol, pepATTRACT". *Bio-protocol* 6.11 (2016), e1831.

Christina EM Schindler, Sjoerd J de Vries, and Martin Zacharias. "iATTRACT: Simultaneous global and local interface optimization for protein–protein docking refinement". *Proteins: Structure, Function, and Bioinformatics* 83.2 (2015), pp. 248–258. DOI: 10.1002/prot.24728.

Christina EM Schindler and Martin Zacharias. "Application of the ATTRACT coarse-grained docking and atomistic refinement for predicting peptide-protein interactions". Ed. by Ora Schueler-Furman and Nir London. Methods in Molecular Biology. Springer Press, 2016. Chap. 7, In review.

Zhe Zhang, Christina EM Schindler, Oliver F Lange, and Martin Zacharias. "Application of enhanced sampling Monte Carlo methods for high-resolution protein-protein docking in Rosetta". *PLOS One* 10.6 (2015), e0125941. DOI: 10.1371/journal.pone.0125941.

1. Introduction

Proteins are truly the molecules of life governing all aspects and scales of cellular function. These molecular "machines"–linear polymers of 20 different building blocks–carry out most of the work in cells. Proteins catalyze a large array of different chemical reactions from breaking down nutrients, transport and motion to DNA replication and cell division. They also provide mechanical stability, act as messengers, storage facilities and cleaners, and defend the cell against external threats. In all these processes, they work in close association with other macromolecules such as nucleic acids, lipids, and of course other proteins. Proteins bind to each other to either jointly carry out a specific biological function, to regulate each other's activity or to pass on molecular signals. The number of interaction partners can vary from just two proteins associating to large assemblies of hundreds of protein components. The interactions are often transient and reversible and may depend on cellular conditions. Hence, protein-protein interactions (PPI) add an additional layer of complexity to cellular processes. This complexity greatly increases the versatility of the cell allowing it to respond to changing environments; e.g., changes in temperature or oxygen levels, and tightly control all its different biological functions. Furthermore, PPIs enable cells to communicate with each other. The ability to send and receive signals from different parts of the organism through interacting proteins is a prerequisite for multi-cellular life. The importance of these fine-tuned networks is further illustrated by their role in pathological disorders: aberrant interactions have been linked to severe diseases like Alzheimer's and cancer. In addition, many viruses hijack PPI networks to use them for their own ends. This therapeutic relevance and their abundance has put PPI targets into the focus of recent drug design efforts with several inhibitors advancing now to clinical trials [317].

Over the past decades, studies on proteins and PPIs using different methods from molecular biology, biochemistry and genetics have elucidated many of their functions and greatly increased our insights into the intricate processes that govern life. Still, a detailed understanding of how proteins work together can only be obtained by atomic-level knowledge of the three-dimensional (3D) structure of protein-protein complexes. Structural data can reveal the functional mechanisms like binding affinity and specificity, mechanical properties and conformational transitions in molecular detail. Only in this way, the biological role of the interaction can be completely

understood. Furthermore, knowledge of the complex structure can be used to specifically modulate the interaction either by rationally designing small molecules as inhibitors or by mutations. The aim of structural biologists is to provide this knowledge by experimentally characterizing protein-protein complexes at atomic precision.

X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy and recently cryo-EM experiments have resolved high-resolution 3D structures of many proteins and protein assemblies. The Protein Data Bank [40] contains over 100,000 structures of individual proteins and also several thousands of structures of two or more interacting proteins. Still, this is only a very small fraction of all known and putative interactions with estimates for the size of the human interactome ranging from 250,000 to 650,000 [429, 235]. Achieving sufficient structural coverage of this large interaction space experimentally may take several decades [148]. High-resolution studies of protein complexes are extremely challenging and laborious due to the strict requirements for pure and homogeneous samples, especially when it comes to characterizing large multimeric complexes. Moreover, many complexes–in particular those involved in signal transduction–exhibit weak binding which makes large-scale expression and purification even more demanding. On top of this, proteins are dynamic molecules which are subject to thermal motions and also complexes can display a range of different conformational states and intrinsic flexibility. The resulting conformational heterogeneity further complicates high resolution structure determination and in some cases even makes it impossible. However, on the bright side, even in such problematic cases, low-resolution structural information can often be obtained by other experimental techniques.

To complement experimental structure characterization, a variety of computational modeling approaches have been developed with the goal to shed light on protein-protein complexes. This thesis deals with the field of protein-protein docking. In many cases, structures of the individual protein partners are available or can be reliably modeled by homology. Docking methods aim to predict the 3D structure of the complex from the structures of the individual protein components. In other words, docking intends to expand structural knowledge for protein-protein interactions making use of the good coverage at the single protein level. The ATTRACT docking program [99] was developed in our group and can be used to model a variety of biomolecular complexes. However, the success of the approach is often limited to interactions that only exhibit small-scale conformational change upon binding. The aim of this thesis is to improve and extend the ATTRACT protein-protein docking approach by explicitly including atomistic protein flexibility and different types of low resolution experimental data.

The thesis is organized as follows. An introduction to physical and structural aspects of PPIs is given in Chapter 2, followed by a general overview on protein-

protein docking strategies and methodologies in Chapter 3. Chapter 4 presents the ATTRACT docking engine outlining its main features. The next chapters describe new methodological developments in ATTRACT. Chapter 5 introduces the flexible interface docking refinement approach iATTRACT. iATTRACT performs simultaneous optimization of global rigid-body and the local interface residue degrees of freedom of protein-protein complex geometries. In Chapter 6, ATTRACT is expanded towards ab-initio docking of highly flexible peptide-protein complexes. A new protocol for modeling interface loops on a given protein-protein complex geometry is tested in Chapter 7. Chapter 8 describes an integrative modeling approach in ATTRACT driven by small angle X-ray scattering data. ATTRACT's performance in rounds 28-36 of the blind prediction challenge CAPRI is discussed in Chapter 9. Chapter 10 describes an application of the pepATTRACT protocol (Chapter 6) to studying peptide binding to the co-chaperone ERdj5 that is involved in protein folding and quality control in the endoplasmic reticulum. Finally, in Chapter 11, different docking protocols in ATTRACT (including the one presented in Chapter 6) are applied to elucidate structural features of the ISWI nucleosome remodeling enzyme. Nucleosome remodelers are very flexible molecules and have been known to adopt different functional states in solution. These properties have made them elusive to traditional structural biology approaches. In order to deal with conformational heterogeneity and ambiguity, a new docking protocol driven by cross-linking/mass spectrometry data (ATTRACT-XL) data has been developed. This approach is also described in Chapter 11. The thesis concludes with a short perspective on the docking field, future challenges and expected developments (Chapter 12).

2. Protein-Protein Complexes

Protein-protein interactions are involved in virtually all biological processes in the cell. This chapter gives an overview on the physical principles and the architecture of protein-protein complexes. Differences between homomeric and heteromeric, and domain-domain and peptide-mediated interactions are discussed. Finally, methods for experimental structure characterization of protein-protein complexes at various resolutions are presented.

2.1. Introduction

Proteins are the "workhorse molecules" of the cell and they make great team players. Even though several proteins are already active in their monomeric form, most proteins interact with other proteins forming large assemblies in order to do their job. Protein-protein interactions (PPIs) are abundant in the cell and involved in important processes such as metabolism, transport, signal transduction, cell division and immune response. It was estimated that the number of PPIs in yeast exceeds the number of single proteins by a factor of 5 to 8 excluding self-assembly [162]. This allows proteins to contribute to different functions in the context of different complexes. The importance of PPIs is also underlined by their role in cellular malfunction; aberrant interactions can be traced to a variety of human diseases [379]. In these cases, the interaction between the proteins is either lost or the complex may form at an inadequate time or location. Viruses like the papilloma and the HI-virus produce proteins that bind to target proteins in their hosts [379, 117]. Current efforts in the field of cancer genomics have mapped out entire networks of modified PPI networks in tumors [211, 217]. Furthermore, dysfunctional interactions have been linked to bacterial infections and amyloid-related neurodegenerative diseases [379].

Large efforts have gone into characterizing PPIs. Interactomes for several organisms have been studied using high-throughput methods like two-hybrid assays and tandem affinity purification tagging [205, 268, 369, 428, 227]. Unfortunately, the fact that two proteins interact does not always give insights into how this interaction is actually established and how it can be potentially modulated. For this, structural

knowledge of the complex needs to be obtained. Since the structures of many individual proteins have been solved experimentally and are available through the Protein Data Bank [39, 40], the structure of the complex can in principle be inferred based on possible physical contacts between the proteins (see Chapter 3). In the following, I will present the main physical rules governing protein-protein association. Then the architecture of known complexes will be discussed. This chapter concludes with an overview on experimental methods for studying protein-protein complex structures at various resolutions.

2.2. Physical principles

Proteins are linear polymers consisting of 20 different amino acids (also termed residues) as building blocks. Most proteins fold into one or more three-dimensional specific conformations to carry out their biological function (Figure 2.1). The specific conformation is directly encoded in the amino acid sequence. Protein-protein complex formation; i.e., the non-covalent association of multiple protein chains, can be understood as a subset of the general protein folding problem (quaternary structure formation, Figure 2.1). Since no chemical bonds are formed between the interaction partners, the assembly is purely driven by physical forces and the binding process can be described in a statistical mechanics framework using an energy landscape concept. When describing the internal energy and the entropy of the system accurately, we can in principle determine the native structure for a protein-protein complex by locating the global minimum in the free energy landscape. This idea is the basis for many of the protein-protein docking programs discussed in Chapter 3. However, due to the large system size, structural inhomogeneity, ensemble properties and protein flexibility, it is in practice highly challenging to apply physical concepts and descriptions to protein-protein complex formation.

2.2.1. Thermodynamics

Binding of two or more proteins is a dynamic process. It is often assumed that the structures of the unbound proteins and the bound complex are in equilibrium (although this might not always be the case in vivo where proteins can be constantly synthesized and degraded). In thermodynamic equilibrium, the probability of the proteins to assemble depends on the free energy difference ΔG_{bind} between the bound and the unbound state

$$\frac{p_{\text{bound}}}{p_{\text{unbound}}} = e^{-\beta \Delta G_{\text{bind}}},$$

where β is the Boltzmann factor $\beta = \frac{1}{k_B T}$. The free energy difference can be decomposed in an enthalpic (energetic when neglecting volume changes) and an entropic contribution

$$\Delta G_{\text{bind}} = \Delta H - T\Delta S.$$

This can be further split up into terms resulting from changes in protein-protein, solvent-solvent (water) and interactions between solvent and proteins

$$\Delta G_{\text{bind}} = \underbrace{\Delta H_{pp} - T\Delta S_{pp}}_{\text{protein-protein}} + \underbrace{\Delta H_{ps} - T\Delta S_{ps}}_{\text{protein-solvent}} + \underbrace{\Delta H_{ss} - T\Delta S_{ss}}_{\text{solvent-solvent}}.$$

Hence, the equilibrium of complex formation depends not only on the interactions between the biological macromolecules but also on solvent conditions; e.g., salt concentration. In order for a complex to be stable, ΔG_{bind} needs to be negative resulting from a decrease in energy or an increase in entropy or both. In particular, loss of entropy upon binding, especially for highly flexible proteins and peptides, has to be compensated by favorable enthalpic changes.

2.2.2. Energy

The strength of PPIs is largely determined by the chemical properties of the amino acids at the interface both with respect to residue-residue interactions and with respect to residue-solvent interactions (desolvation properties). Amino acids consist of a backbone containing the amine and carboxylic functional groups and a side chain that is attached to the C_α carbon. Different functional groups are present as side chains in different amino acids and their atomic composition and chemical structure can vary widely (Figure 2.1). The electronic configuration or charge distribution on the amino acid determines its interaction characteristics. The electronic configuration can be obtained by solving the Schrödinger equation for the entire molecule. However, obtaining solutions for such large multi-body problems is very difficult. Even when using approximations, quantum mechanical calculations are too demanding to apply them to large systems like proteins that consist of thousands of atoms. We will limit ourselves here to discussing a few general properties of the electronic distribution of amino acids.

In general, electronic distributions from different atoms or molecules cannot overlap leading to a repulsion of non-bonded atoms at small separation distance (Pauli principle). In addition to this steric repulsion, there are several types of attractive interactions for amino acids (Figure 2.2). The electronic configurations from different amino acids can interact by induced dipole-dipole interactions (van der Waals interactions). Protein-protein interfaces are typically densely packed [86] to maximize

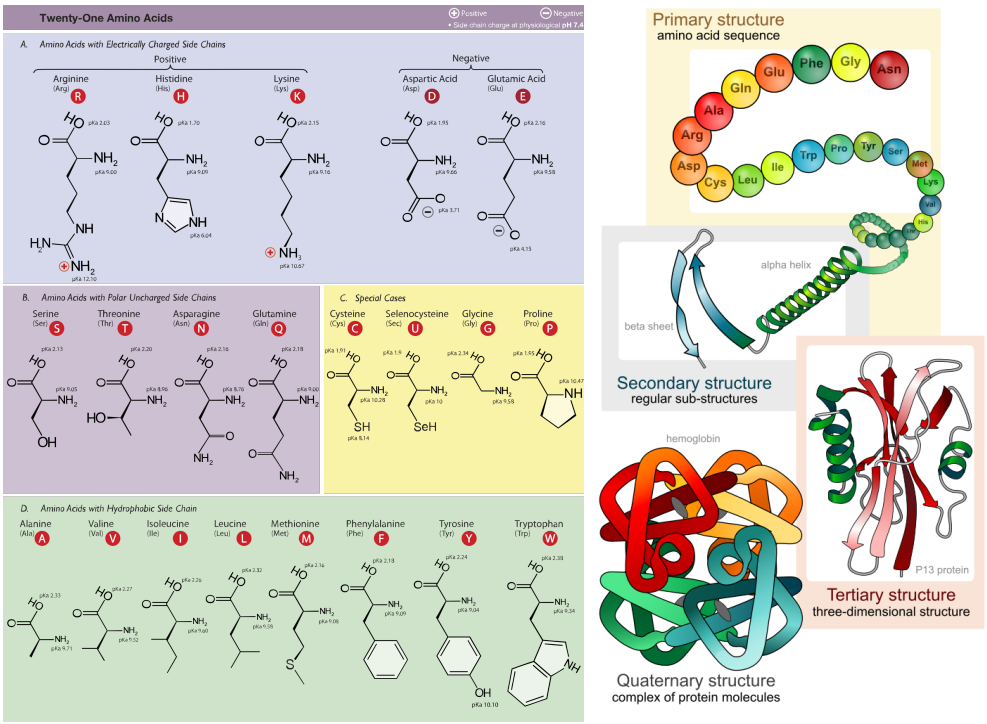
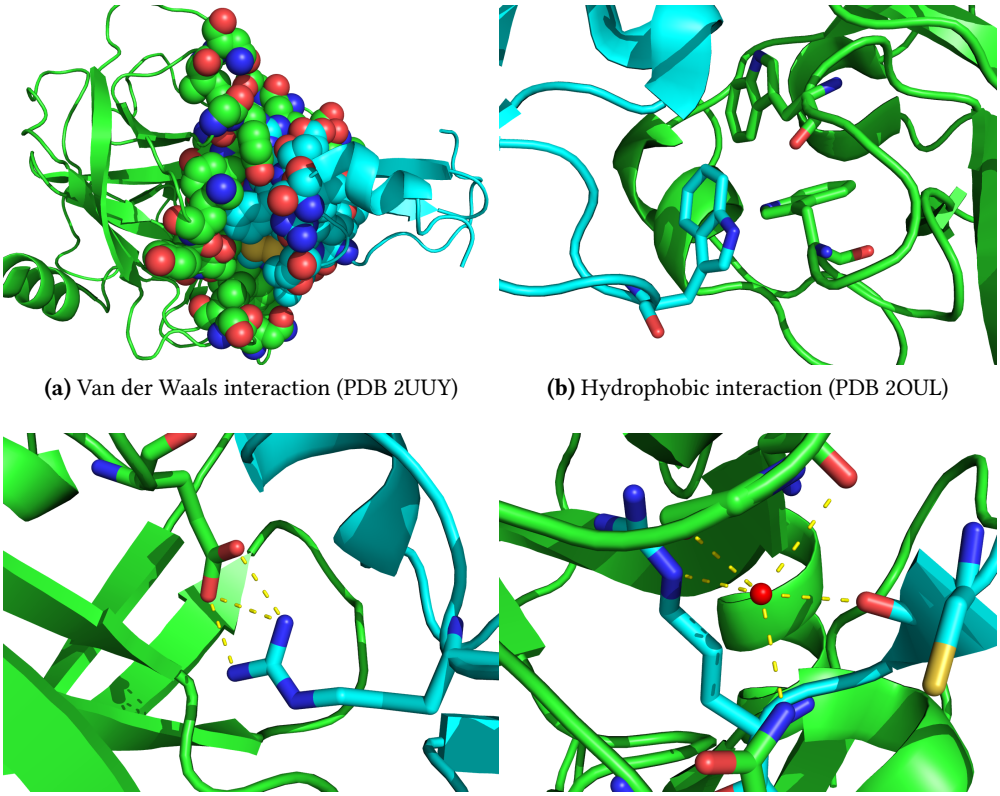


Figure 2.1. Twenty-one natural amino acids used as building blocks in proteins and main levels of protein structure. For each amino acid, its name, three-letter and one-letter code, its chemical properties and its chemical structure are shown. The figures were created by Dancojocari [CC BY-SA 3.0 (creativecommons.org/licenses/by-sa/3.0)] and by LadyofHats [Public domain], and obtained via Wikimedia Commons.

such favorable short-ranged van der Waals contacts. In many cases, the electronic distribution along the bonds is not uniform and the bond has a dipole moment (polar bond). In some molecules, the individual bond dipole moments do not add up to zero resulting in an overall dipole moment (water is an example of such a polar molecule). Amino acids contain polar OH and NH groups both in the backbone and the side chains. An interesting property of these polar groups is that they can form additional electrostatic interaction with each other (hydrogen bonds). Hydrogen bonds are directional and strong dipole-dipole interactions: the energy associated with hydrogen bond formation is in the range of a few kcal/mol (for reference, $k_B T \approx 0.5\,\mathrm{kcal/mol}$ at room temperature). Therefore, formation of hydrogen bonds between two protein partners also contributes to the stability of the complex. Furthermore, hydrogen bonding is related to the effects of the solvent on complex formation. Water molecules form strong hydrogen bond networks with each other and introducing a solute can potentially destroy these favorable interactions. Polar amino acids that can replace inter-solvent hydrogen bonding with hydrogen bonds between its side chain and water are therefore more easily solvated. In contrast, nonpolar residues tend to be excluded from the solvent to avoid disrupting the natural water-water hydrogen bonding network. In other words, such hydrophobic residues form favorable interactions at protein-protein interfaces (hydrophobic effect). Since water molecules can engage in multiple hydrogen bonds simultaneously, they have also been found at interfaces forming hydrogen bonds between residues on both partners (water-mediated hydrogen bonds). Another potential energetic contribution can come from stacking interactions. Some amino acids like phenylalanine and tryptophane contain aromatic rings. These special chemical structures result in a particular electronic configuration with delocalized electrons in p-orbitals that can engage in quadrupole interactions. This interactions are termed π-stacking interactions and have been frequently observed in nucleic acids, in protein structures and in protein-small molecule complexes. Finally, electrostatics also play an important role in complex formation. Several amino acids (histidine, lysine, arginine, glutamate, aspartate) can carry a net charge at physiological pH and therefore interact by Coulombic interactions (salt bridge/ionic bond). The total enthalpy changes can be determined by summing over the changes in all interaction types for all atoms in the system

$$\Delta H = \Delta E_{\text{vdW}} + \Delta E_{\text{Coulomb}} + \Delta E_{\text{hydrophobic}} + \Delta E_{\text{hbond}} + \Delta E_{\text{stacking}}.$$

The electronic configuration of the amino acids additionally depends on the local environment; i.e., the location of a particular amino acid within the protein and its neighboring residues. This can lead to changes in e.g. the protonation state of the



(a) Van der Waals interaction (PDB 2UUY)

(b) Hydrophobic interaction (PDB 2OUL)

(c) Salt bridge between charged residues (PDB 1A2K)

(d) Water-mediated hydrogen bonds (PDB 1PPE)

Figure 2.2. Interactions driving protein-protein complex formation. The protein partners are shown in green and cyan. Water molecules are drawn in red.

charged amino acids. The electronic properties of interface residues might potentially change during the association process and calculating the energies using the same electronic configurations for the individual proteins in bound and unbound state might not be accurate. These polarization effects could only be considered in a quantum mechanical description. However, such calculations are not feasible for large systems like proteins and protein-protein complexes. Hence, many approximations have to be made when evaluating the energy of protein-protein complexes in practice (see also Section 3.5.2).

2.2.3. Entropy

The energetic description of protein-protein interactions is already challenging, however, considering entropic effects is even more so. Conformational entropy and solvent entropy play a major role in protein-protein interactions. Upon protein-protein complex formation, conformational entropy is usually lost for the interface residues, whereas desolvation of hydrophobic residues typically increases the solvent entropy. However, in some cases, other non-interacting parts of the proteins can become more flexible [167] and in principle conformational entropy change can be either positive or negative. In principle, the entropy S of a system can be calculated using the Boltzmann formula

$$S = -k_B \sum_i p_i \ln p_i$$

where k_B is the Boltzmann constant, and the probability p_i for state i is proportional to $e^{-\beta U_i}$ with U_i the potential energy of configuration i. However, this calculation entails two problems: a) the possible conformations of the system have to be known and b) the energy U has to be accurately modeled.

In practice, entropy changes are usually decomposed into different contributions; e.g., side-chain conformational entropy, backbone conformational entropy or solvent entropy. In cases of a finite number of states with similar energy, the entropic change can be approximated by

$$\Delta S \approx -k_B \ln \left(\frac{n_2}{n_1} \right)$$

with n_1 and n_2 being the number of accessible conformations in the initial and final state. Such an approach can be used for estimating the side chain entropic changes by counting accessible rotamers. For degrees of freedom which are governed by a

harmonic potential, the change in entropy can be calculated by

$$\Delta S \approx -\frac{k_B}{2} \ln\left(\frac{k_2}{k_1}\right) = -k_B \ln\left(\frac{\nu_2}{\nu_1}\right)$$

where k_1 and k_2 are the force constants and ν_1 and ν_2 the frequencies connected to the motion. The frequencies can be obtained from normal mode calculations and from this, vibrational entropies for global backbone deformations can be calculated [55].

The main problem remains to determine the possible conformations of the proteins in the unbound and the bound state in the first place. Due to the large system size with in principle $3N$ degrees of freedom (N being the number of atoms), sampling the entire conformational landscape of proteins and protein-protein complexes is practically impossible. This is illustrated by several unsuccessful attempts to calculate conformational entropy changes by molecular dynamics simulations [167, 153, 191]. Similar problems arise for determining the solvent entropy changes.

2.3. Structural aspects

In order to better understand their biological function at the molecular level, many protein-protein complex structures have been resolved experimentally by methods like X-ray crystallography or NMR spectroscopy. These 3D structures have given important insights into the function of PPIs and the physical principles that govern protein association. Protein-protein complexes can be classified by their composition, the lifetime of the complex and the strength of the interaction (which directly relate to its binding free energy), and the characteristics of the interface. Homomeric complexes are formed from multiple identical or homologous chains, whereas heteromeric complexes contain at least two different subunits. Complexes can be distinguished on the basis of whether they are obligate or non-obligate. In an obligate complex, the individual constituents are not found as stable structures in the cell, whereas proteins from non-obligate complexes can exist independently. In terms of lifetime, complexes are classified as transient or permanent. Furthermore, weak and strong binders can be distinguished by measuring the binding affinity [292]. Interactions can further be classified based on the size and the degree of flexibility of the association partners (domain-domain interactions, peptidic interactions and interactions involving intrinsically disordered proteins). In the following, I will introduce some general interface properties and global structural features of protein-protein complexes and also discuss the role of protein flexibility. From this point on, I will only consider non-obligate, transient complexes.

2.3.1. General properties of protein-protein interfaces

Protein structures deposited in the Protein Data Bank (PDB, www.pdb.org) [39, 40] contain two types of information: the chemical identity of each atom and its 3D position. Hence, the structure of interfaces in protein-protein complexes can be characterized both by geometric and chemical properties. Typical geometric criteria are the number and type of atom-atom contacts (evaluated within a chosen cutoff distance) and the buried surface area; i.e., the surface patches that are part of the solvent accessible area (SASA) in the unbound proteins but not in the complex

$$\mathrm{BSA} = \mathrm{SASA}_A + \mathrm{SASA}_B - \mathrm{SASA}_{AB}.$$

The solvent accessible surface area can be calculated by rolling a particle with the radius of a water molecule ($r_{\mathrm{probe}} = 1.4\,\mathrm{\AA}$) over the surface and determining the border. Atoms are defined to be in the interface if they contribute to the BSA or if they are located within a certain distance of the partner molecule. The size of the interface as measured by BSA is widely-used for classifying protein-protein interfaces. The interface can also be analyzed in terms of its chemical nature. Amino acids can be either hydrophobic, polar, nonpolar or charged depending on the electronic side-chain properties (see also Section 2.2.2). Based on this, the interface composition in terms of e.g. non-polar, neutral polar and charged BSA can be assessed. Similarly, the contacts can be evaluated by the chemical properties of the residues. Another important property of protein interfaces is atomic packing. Protein-protein interfaces are closely packed similar to the interior of proteins [86] and in many cases form a single contiguous patch [62]. The high packing density and the connectivity of the interacting surfaces reflect the high degree of shape complementarity between the association partners (an example is shown in Figure 2.2 (a)). In general, amino acid composition at protein-protein interfaces is significantly different from the composition of the rest of the protein surface. Interfaces tend to be enriched in hydrophobic and aliphatic residues (phenylalanine, tyrosine, tryptophane, alanine, leucine, valine, methionine) relative to the protein surface and depleted in most charged residues (except arginine) [86]. Interfaces can be divided into a hydrophobic core region which is often very conserved and a more evolutionary variable, polar rim region [62, 170, 20]. Furthermore, not all interface residues are born equal: early on it was noted that certain residues contribute the major part of the interaction energy. These residues were termed “hot-spot” residues and often involve large amino-acids like tyrosine, arginine and tryptophane that anchor themselves into small pockets across the interface [76, 48, 104]. London et al. [276] proposed an extension to this concept by identifying so-called “hot segments”; i.e., short linear motifs in domain-domain interactions that dominate the binding energy. Hot segments can be found in approximately 50% of

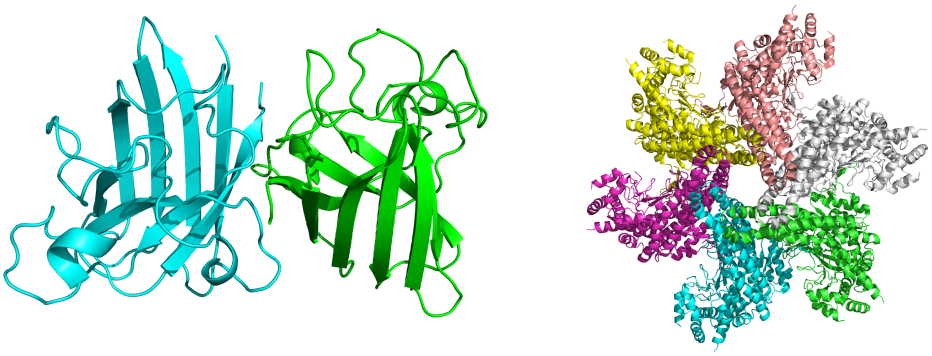
globular interactions [276] and probably make these complexes more amenable for modulation by small molecules [277].

2.3.2. Homomeric complexes

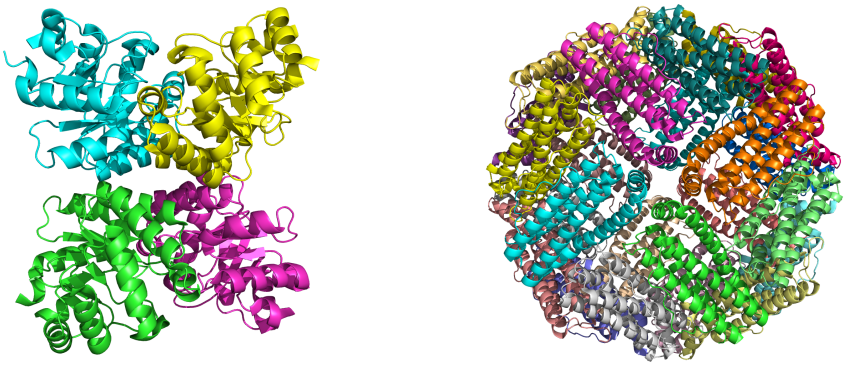
The majority of protein structures in the PDB are found with multiple copies in the asymmetric unit and the majority of protein-protein interactions falls into the category of homomeric assemblies. Most homooligomeric complexes of known structure display symmetry. The evolutionary origins of this symmetry are still under debate with proposed advantages in folding efficiency [34], reduced aggregation [158], potential of allosteric regulation [312] and adaptability [158, 312]. Interestingly, Baker and coworkers showed that symmetry can potentially arise as a feature of a sufficiently favorable interaction energy and thus might be an intrinsic property of the free energy landscape [10]. Based on symmetry operations, the overall architecture can be classified into a small number of groups. Examples of different symmetry groups are shown in Figure 2.3. Symmetric dimers (C_2 symmetry group) are the simplest homomeric assemblies. Due to the two-fold rotational symmetry, the proteins necessarily interact via identical surface patches (isologous interface). In contrast, higher-order cyclic multimers (C_n) are built from asymmetric interactions between distinct parts of the surface (heterologous interfaces) and form closed rings. Complexes of dihedral symmetry (D_n) are characterized by two orthogonal symmetry axes (e.g., a D_2 complex is a dimer of dimers). In addition, cubic, helical and asymmetric complexes have been observed [292]. Knowing the symmetry of a complex can greatly increase the accuracy of computational modeling (see Chapter 3). Note that even though homomeric complexes often display global symmetry, individual subunits can undergo local structural variations.

Bahadur et al. analyzed the interfaces of homodimers and found that the size of the interfaces as measured by BSA ranged from $1000\,\text{Å}$ to $14\,000\,\text{Å}$ with an average size of $\approx 3900\,\text{Å}$. The interface are composed to 35% of polar groups and 65% of nonpolar atoms. This makes homodimeric interfaces significantly more hydrophobic than the average protein surface with a clear enrichment in Leu, Ile, Val and Met. Still, hydrogen bonds are also frequently found at the interfaces with on average one hydrogen bond per $210\,\text{Å}$ BSA [20, 505].

When studying homomeric complexes by crystallographic experiments, distinguishing the biologically relevant interfaces from crystallographic interfaces is very challenging. In the PDB, the quaternary structure of a protein in vivo is usually determined by the authors and reported as the “biological unit”. However, this information is not available in all PDB entries and if present, the annotations might not always be reliable [494, 495]. A few computational methods have been devel-



(a) Dimeric (C_2), superoxide dismutase, PDB 3F7L. **(b)** Cyclic (C_6), glutamine synthetase, PDB 3O6X.



(c) Dihedral (D_2), aldolase, PDB 1UB3. **(d)** Cubic (Octahedral), apoferritin, PDB 4V1W.

Figure 2.3. Examples of symmetry found in homomeric protein complexes. For each example, the quaternary structure, the name of the protein and its PDB ID are shown.

oped for discriminating between biologically relevant interfaces and lattice contacts, however, the methods typically have an error rate of at least 10 % [116, 238]. In general, it would be desirable to independently determine the quaternary structure in solution and submission/annotation standards in the PDB should be adapted towards this goal in the future. The problems arising from uncertainties in quaternary state assignment have also featured prominently in the recent CASP-CAPRI blind prediction experiment [262].

2.3.3. Heteromeric complexes

Only a small fraction of all possible heteromeric assemblies have been characterized structurally to date [148], however, these structures already reveal a vast diversity in quaternary structure and interaction types. Heteromers can display symmetry (especially when composed of paralogous subunits), sometimes mixing different types of symmetry. Symmetry often occurs in larger heteromeric multimers with repeating subunits. In contrast, complexes composed only of very few components are often asymmetric [292]. When looking at the interface properties, it was found that heteromeric interfaces tend to be smaller than homomeric interfaces (average BSA $\approx$ 1.900 Å), although in both cases the distribution varies widely. Heteromeric interfaces contain a higher fraction of polar groups ($\approx$ 42%) compared to homomeric interfaces but are still enriched in hydrophobic residues compared to the average protein surface. On average one hydrogen bond is found per 190 Å BSA so per unit interface area heteromeric complexes form more hydrogen bonds than homomeric complexes [62, 505]. The observed differences between homomeric and heteromeric complexes most likely reflect the fact that homomeric complexes tend to form relatively long-lived assemblies whereas heteromeric complexes often associate and disassociate in a more dynamic manner.

2.3.4. Peptide-protein complexes

Protein-protein interactions can also be classified by the degree of flexibility present in the association partners. The majority of interactions involve binding of two or more ordered domains/proteins (as described above), but the importance of interactions mediated by flexible peptides or by intrinsically disordered protein (regions) has become more and more apparent [442]. Furthermore, even in domain-domain interactions, significant binding-induced conformational change can be observed (see also Section 3.6).

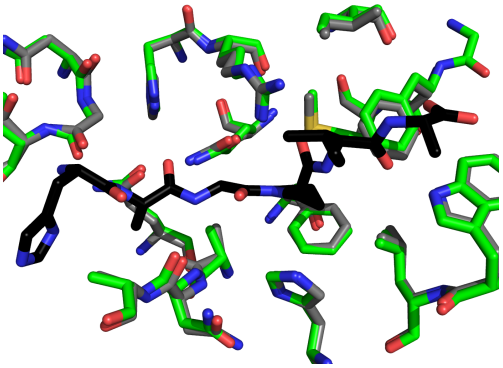
Peptide-mediated interactions are complexes between an ordered protein domain and a peptidic motif. The motif can be an isolated peptide but more often this refers to

short linear motifs from an intrinsically disordered protein region (IDR). Intrinsically disordered proteins (IDPs) do not adopt a fixed fold in solution and are often found as hub proteins in protein-protein interaction networks [487]. IDPs provide functional and evolutionary flexibility to interaction networks, since IDPs can be fine-tuned and adapted to a variety of interaction partners due to their structural flexibility and the possibility for regulation by post-translational modifications [202, 442]. A common feature of IDPs/IDRs is that they often interact via short recognition motifs [458]. Several complexes of such motifs bound to their partners have been characterized experimentally (e.g., in [159, 44, 32]) and currently several thousands of motifs have been identified [108]. However, the actual number of motifs might be above one million [441].

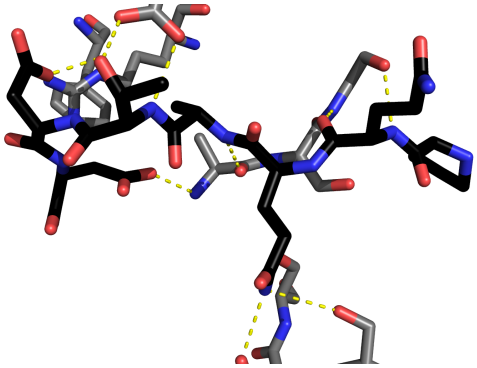
The structure of peptide-mediated interactions differs in some aspects from the structure of protein-protein complexes. A few typical features of peptide-protein complexes are shown in Figure 2.4. Since the motifs are typically very flexible in solution, the loss in peptide conformational entropy upon binding has to be efficiently compensated. London et al. conducted a systematic analysis of known peptide-protein complexes to unravel several peptide-protein binding strategies [275]. The protein usually undergoes only little conformational change upon peptide binding. This is in contrast to protein-protein interactions, where large conformational changes can be observed. The fact that the protein does not need to adopt its conformation reduces the overall entropic cost of binding. Since peptides can only form small interfaces due to their limited size, the interactions have to be highly optimized. Peptide-protein complexes therefore display higher packing density and form more hydrogen bonds per interface area than protein-protein complexes. A main contribution to this increase in hydrogen bonds is the increased ability of the flexible peptide to form hydrogen bonds between its main chain groups and the protein's side chains. The amino acid composition of the interface is in general similar to that of protein-protein complexes with an over-representation of leucine. Similar to protein-protein complexes, peptide-protein complexes also contain hot-spot residues (mainly phenylalanine, leucine, tryptophane, tyrosine and isoleucine). In the majority of complexes, these residues contribute more than 70% of the interaction energy [275].

2.4. Experimental methods

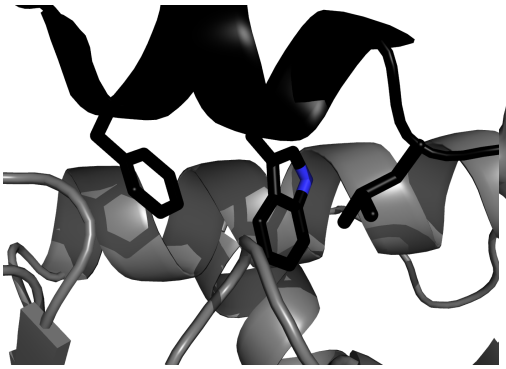
Experimental methods like X-ray crystallography and nuclear magnetic resonance spectroscopy have resolved high-resolution structures for many proteins and also many protein-protein complexes. As of April 2016, the Protein Data Bank (PDB, www.rcsb.org) [39, 40] contains roughly 110,000 protein structures, 13% of which



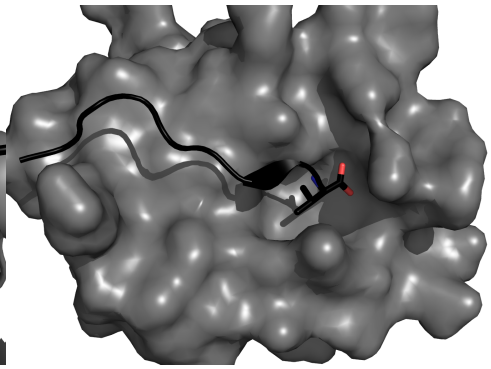
(a) Preordered protein interface, PDB 1AWR.



(b) Extensive hydrogen bonding, PDB 1CZY.



(c) Hydrophobic interactions, PDB 3DAB.



(d) Hot-spot residue anchoring the peptide, PDB 1MFG.

Figure 2.4. Interactions in peptide-protein complexes. The protein is shown in gray in its peptide-bound form and in green in the unbound form. The peptide is drawn in black. Hydrogen bonds are shown in yellow.

represent heteromeric protein-protein complexes from different organisms. Also when including template-based models, only a minority of all possible interaction types has been resolved [148]. Compared to the progress in high-resolution structural coverage achieved for individual proteins, progress for complexes is still limited by technical difficulties and the high costs of the methods. However, even in the absence of high-resolution data, structural insights can be gained from a variety of biophysical and biochemical experiments. This section introduces some of the most widely used techniques and explains which type of data can be obtained.

2.4.1. X-ray crystallography

Structural studies of proteins by X-ray crystallography date back to the 1950s and the vast majority of protein and nucleic acid structures in the PDB has been solved by this method. The molecular structure is inferred from the diffraction pattern of an X-ray beam scattered elastically at a protein crystal. X-ray crystallography can be applied to proteins of arbitrary size. Due to the repeated scattering in the crystal and the interference, the two-dimensional diffraction pattern recorded at different orientations of the crystal relative to the beam can be used build a three-dimensional model of the electron density by Fourier transform. Combined with knowledge of the sequence, the protein structure can then be reconstructed at resolutions typically higher than $3\,\text{Å}$. The major limitation in X-ray crystallography is the need to obtain diffracting protein crystals of sufficient size. Protein crystallization is very challenging, especially for (weak) complexes and membrane proteins. The use of free electron lasers might offer ways to overcome some of these limitations [64, 321, 286], but the levels of protein concentration and sample purity required for X-ray crystallography will remain the major bottleneck for studying protein-protein complexes.

2.4.2. NMR

Nuclear magnetic resonance spectroscopy experiments are often used to study the structures and dynamics of biomolecules. In contrast to X-ray crystallography, the molecules can be characterized in solution at near-physiological conditions and do not require protein crystallization. Still, relative large amounts of proteins and pure samples are needed. For structure determination by NMR, the key data are upper distance restraints derived from NOEs (Nuclear Overhauser Effect). NOEs arise from cross-relaxation due to dipole-dipole interactions. The inherent r^{-6} dependence of NOEs can be used to extract information on distances between hydrogen atoms. The distances are typically in a range of up to $6\,\text{Å}$. The main difficulty is to assign the cross-peaks in the spectra to a given pair of hydrogen atoms, for which additional

experiments need to be performed. For this reason, data collection can be very time-consuming. Furthermore, NOE assignment can often only be made in an ambiguous manner. Based on the assigned distances, structures can be calculated using specialized software [172, 405, 179]. For ab-initio structure determination by NMR, hundreds to thousands distances are needed already for small complexes making a large-scale application of NMR to resolve protein-protein complexes very difficult. Due to difficulties in assigning backbone resonances, structure determination by NMR is typically limiting to proteins/protein complexes smaller than $80\,\mathrm{kDa}$, although recent technological advances might make NMR studies of large assemblies more frequent [288].

If the structures of the individual subunits are known, less information is required to solve the structure of the complex. NMR experiments can provide a wide range of suitable low-resolution information. Interface residues can be identified based on chemical shift perturbation (CSP), NOEs, cross-saturation, hydrogen-deuterium exchange and solvent paramagnetic relaxation enhancements. Furthermore, the relative orientation of the association partners can be inferred from relaxation rates and residual dipole couplings (RDCs) [152]. These sparse data, which do not provide sufficient information for ab-initio structure determination, can then be combined with computational modeling to solve the structure of the protein-protein complex based on the structures of the individual partners (see also Section 3.7.2) [112, 144, 82]. Experimental NMR data are available through the BioMagResBank (Biological Magnetic Resonance Bank, www.bmrb.wisc.edu) [452].

2.4.3. cryo-EM

Instead of X-ray photons, electron scattering can be used to investigate biomolecular structures (electron microscopy). 2D scattering images are recorded for different orientations of the molecules and can then be assembled into a 3D reconstruction of the structure. The first electron microscope was already constructed in the 1930s [377], however, the wide-spread application to biological macromolecules was hindered by the requirement of a vacuum to avoid electron scattering by air molecules and the heavy radiation damage that often destroyed the sample. These issues were largely resolved by rapidly freezing the biological sample in vitreous ice and using this frozen sample for image collection (cryo-EM). Still, for many years, the resolution of images recorded by cryo-EM remained low ($> 7\,\mathrm{\AA}$). However, in the last three years, the field has made tremendous progress due to the development of direct-electron detectors and improved image processing procedures [22]. This fast advance has even been termed a "resolution revolution" [244], since 3D reconstructions of resolutions higher than $4\,\mathrm{\AA}$ have been obtained for many large protein assemblies (e.g., [57, 420,

476, 42]). This allows to build structures de novo into the cryo-EM density similar to the procedures used in X-ray crystallography. Even more exciting, cryo-EM studies can go beyond resolving a single, static structure allowing to study different functional states, flexibility and even dynamics of complexes [133]. Despite the great progress in the field, high-resolution structure determination of protein-protein complexes still faces several challenges. In order to obtain high resolution, the sample needs to be very pure and homogeneous and biochemical sample preparation will probably become the main bottleneck for studying protein-protein complexes by cryo-EM (Stark, personal communication). Also, currently high-resolution structures can only be obtained for large assemblies ($> 1\,\mathrm{MDa}$) due to difficulties in image alignment for smaller complexes. For smaller, unstable or flexible complexes, obtaining near-atomic resolution maps may remain challenging and time-consuming if not impossible [22]. However, if the structures of the individual constituents are known or can be modeled, low-resolution maps can still be used to guide the computational assembly of the complex [94, 459, 443, 333]. The majority of density maps stored in the EMDB (Electron Microscopy Data Bank, www.emdatabank.org) [433] are still of low resolution (60% $> 12\,\text{Å}$ as of May 2016) and even half of the maps deposited in 2015 had a resolution lower than $8\,\text{Å}$.

2.4.4. SAXS

Similar to NMR, in small-angle X-ray scattering (SAXS) experiments, the structure of biomolecules can be studied in solution at near-physiological conditions. In a SAXS experiment, the sample is illuminated by an X-ray beam and the X-ray photons scatter elastically on the sample. The diffuse scattering pattern (in contrast to scattering at a crystal, there are no sharp diffraction peaks) is recorded as a 2D image. Subsequently, the scattering image from the background solution (usually a buffer) is collected. Since the proteins are oriented randomly in solution, the scattering pattern represents the orientational average of the scattering on an individual particle. Hence, the 2D images can be radially averaged to obtain a 1D scattering intensity profile. The scattering of the protein solution is then subtracted from the scattering of the buffer and this difference scattering profile is the typical data obtained from a SAXS experiment. SAXS data contain low-resolution (10-50 Å) information on the overall macromolecular shape and several structural quantities can be extracted from the profile (e.g., radius of gyration, molecular mass, overall shape and maximum intraparticle distance). Data collection by SAXS experiments is fast (typically a few minutes on a well-equipped beamline) and sample preparation is relatively easy. Still, large amounts of purified proteins/protein complexes need to be obtained and measuring high-quality SAXS data is not always straightforward, since the data are

highly sensitive to the experimental technique and sample quality. Protein aggregation, poly-dispersity and poor background subtraction can dramatically affect the interpretability of the profiles [421]. Even in the case of high data quality, SAXS data of protein-protein complexes can still be challenging to interpret due to structural ensembles/flexibility of the complexes in solution and possibly incomplete complex formation. Still, SAXS has become more and more popular [160] and in the last years, SAXS experiments have been used to study several protein-protein complexes [367, 395, 9, 119, 115]. SAXS data can be also used for quaternary structure assignment [350] (see Section 2.3.2). Experimental SAXS data can be deposited in the SASBDB (Small Angle Scattering Biological Data Bank, www.sasbdb.org) [455] and the BIOISIS (www.bioisis.net) [198] databases.

2.4.5. Mass spectrometry

Over last couple of years, several mass spectrometric methods have been developed for obtaining low-resolution structural data of biomolecular complexes (typically 15-35 Å). The data can be combined with computational modeling to gain insights into the overall structure of the assembly [360] and such protocols have been applied successfully to a variety of molecular machines [250, 180, 72]. Mass spectrometry experiments require only very small amounts of protein and are in general not limited by protein size. Also the analysis is relatively fast. For studying protein-protein complexes, there are four main mass spectrometric methods: chemical cross-linking mass spectrometry (CX-MS or XL-MS), native mass spectrometry (analysis of intact assemblies), hydrogen/deuterium exchange mass spectrometry (H/DX-MS) and affinity-purification mass spectrometry (AP-MS) [165]. Native MS and AP-MS can yield information about subunit composition and connectivity and stoichiometry of the complex and subcomplexes. H/DX-MS can give insight into changes in solvent accessibility (similar to hydrogen/deuterium exchange combined with NMR experiments). In contrast, XL-MS can be used to extract information on residue distances within the complex which is the most useful data type for integrative modeling approaches [360]. In a XL-MS experiment, the protein complex is treated with a chemical agent that can form covalent bonds between adjacent amino acids. The complex is then split into peptides that are analyzed by mass spectrometry. Identification of cross-linked peptides/residues can give upper limits on the spatial proximity of these residues in the complex. Typical cross-linking agents yield upper distance limits of $\approx$ 30 Å between the C_α atoms of residue pairs. For a long time, the broad application of XL-MS to protein-protein complexes was hindered by the lack of suitable methods to specifically enrich and reliably identify cross-linked peptides from the large mixture generated by the digestion of protein assemblies [256]. However, recent advances in

instrumentation, cross-linking chemistry, and analysis software have helped XL-MS become a well established and versatile part of the structural biologist's toolbox [255]. Difficulties with using XL-MS data for structure determination are that they might potentially be incoherent (originating from multiple conformations) and can contain false positives. Also the cross-linked conformation might not necessarily be the biologically functional conformation. Therefore, XL-MS data should be combined and validated with other experimental data. XL-MS data are not as easily publicly available as other experimental data. As of May 2016, there are only two small databases: XLink-DB (brucelab.gs.washington.edu/xlinkdb) [518] and XLdb (manually curated from literature, no submission system) [213].

2.5. Conclusion and Outlook

Protein-protein interactions are involved in all aspects of cellular life. Due to their size and the large number of degrees of freedom, it is very challenging to apply physical concepts to these highly complex systems and understand their biological function and the driving forces of the interaction from first principles. Experimental methods like X-ray crystallography and nuclear magnetic resonance spectroscopy have provided atomistic insight into the structure of a large number of complexes. However, for the majority of the interactome, atomic structural data are lacking to date. Experimental structure determination for protein-protein complexes is very difficult, time-consuming and expensive. The major bottlenecks are sample preparation (expression and purification of complexes) and protein flexibility/conformational heterogeneity. Hence, obtaining high-resolution structures for hundreds of thousands possible complexes will not be feasible in the near future. Nevertheless, for many systems, low-resolution structural data from SAXS and XL-MS experiments are becoming more and more easily available.

3. Protein-Protein Docking

Protein-protein interactions are abundant in the cell, however, atomic structural data is only available for a small fraction of complexes. Computational protein-protein docking methods can complement experimental structure characterization by predicting the structure of protein-protein complexes from the structures of the individual constituents. This chapter gives an overview on the basic concepts and strategies in docking. Different docking programs are introduced.

3.1. Introduction

Most biological macromolecules exert their function in complexes. These complexes can consist of just two molecules but more often involve a multitude of biomolecular entities that form huge macromolecular machines like the ribosome or the nuclear pore complex. Atomic structural knowledge of these assemblies is necessary for better understanding their biological roles and hence the processes that govern life. However, only a small number of protein-protein complex structures has been characterized experimentally so far. In contrast, the structural coverage for individual proteins is a lot higher (an example for binary interactions is shown in Figure 3.1) [314]. Hence, being able to predict the three-dimensional structure of these assemblies has been a goal of theoretical modeling, ever since sufficient structural information on the building blocks; i.e., proteins and nucleic acids, became available. However, due to the size of biomolecular assemblies, progress in modeling and docking has been closely tied to progress in computing power and storage capacities. Still, already in 1978, Wodak and Janin performed the first protein-protein docking simulations [485] establishing many of the principles that guide docking methods today. This pioneering work already illustrated the major problem specific to docking—the extensive exploration of the degrees of freedoms when assembling macromolecular complexes—and the strategy for resolving this problem: simplifying the problem as much as possible while at the same time ensuring that these simplifications still allow sampling a near-native geometry and ranking it successfully. This chapter presents the different choices modelers face with regards to the complexity of the system representation, possible sampling algorithms and the degree of accuracy and level of

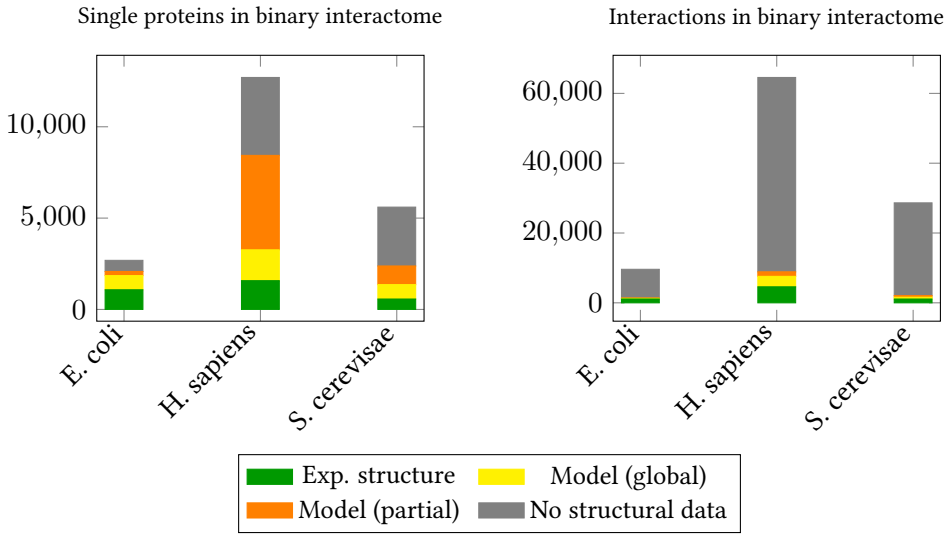


Figure 3.1. Structural coverage for single proteins and complexes in the binary interactome for several model organisms. The figure was created using data from [314].

detail of the scoring function. Current docking programs and their strategies are a result of an efficient combination of such choices. This chapter gives on overview on current docking methods and their underlying strategies. It also presents the criteria used to evaluate the performance of docking methods which have mostly been established by the blind prediction docking experiment CAPRI (Critical Assessment of PRediction of Interactions, www.ebi.ac.uk/msd-srv/capri). Since its initiation in 2001, the CAPRI challenge has taken place 36 times and covered more than 100 targets (as of April 2016). Regular evaluation meetings identify the limitations of current docking methods and provide a comparative assessment of diverse docking strategies (Section 3.8). The general challenges and bottlenecks for further development of docking methods will also be discussed in this chapter.

3.2. Terminology

The term protein-protein docking refers to any theoretical method that is capable of predicting the three-dimensional structure of protein-protein complexes from the structures of the individual constituents. A wide range of different docking programs have been developed to date covering a large spectrum of docking applications. This ranges from local, high-resolution docking to large-scale, often simplified/coarse-grained explorations of all possible positions and orientations of the binding part-

ners (global docking). In high-resolution docking, information about the interface is available for at least one partner and the main goal is to uncover the details of the interactions (local docking). High-resolution docking methods use a very detailed description of the interaction potentials and are therefore often computationally expensive. In contrast, global/ab-initio docking requires fast methods in order to screen many possible complex geometries. These methods often approximate parts of the precise interaction; representing all the atomic details is not required as long as the overall predictive power is retained. A specific terminology has been established in the docking field. For binary docking, the association partners are typically referred to as the receptor (typically the larger protein) and the ligand. For preliminary tests of docking algorithms, crystallized protein-protein complexes can be separated and the separated structures can be used to "redock" the complex. This problem is referred to as "bound-bound" docking, since the protein structures from the bound complex are used. Since all the residues are in the perfect position for association (as envisioned in the "lock-key" mechanism by Fischer [131]), bound-bound docking is considered an easy problem in the field [43]. Typically, docking algorithms should be tested in an "unbound-unbound" docking scenario. Unbound-unbound docking best represents the situation in a "real life" application where the free forms of the proteins or structures from complexes with other partners or homology models are used. Such a structure is referred to as the "unbound" form of the protein. Many methods that perform well in the bound-bound case fail in an unbound-unbound scenario. This bias towards bound-bound docking is a well-known problem in the docking field (see also Section 3.6).

3.3. General strategy

In general, protein-protein docking methods consist of three main components (Figure 3.2):

1. Choosing a protein representation together with the definition of the degrees of freedom that will be sampled.
2. Sampling/generating many possible models of the complex.
3. Classifying the generated models by a scoring function and identifying the best/near-native prediction.

Different levels of complexity can be found in different docking programs for each of these three components. The protein representation can range from a simplified, geometric surface representation via different levels of coarse-graining to detailed,

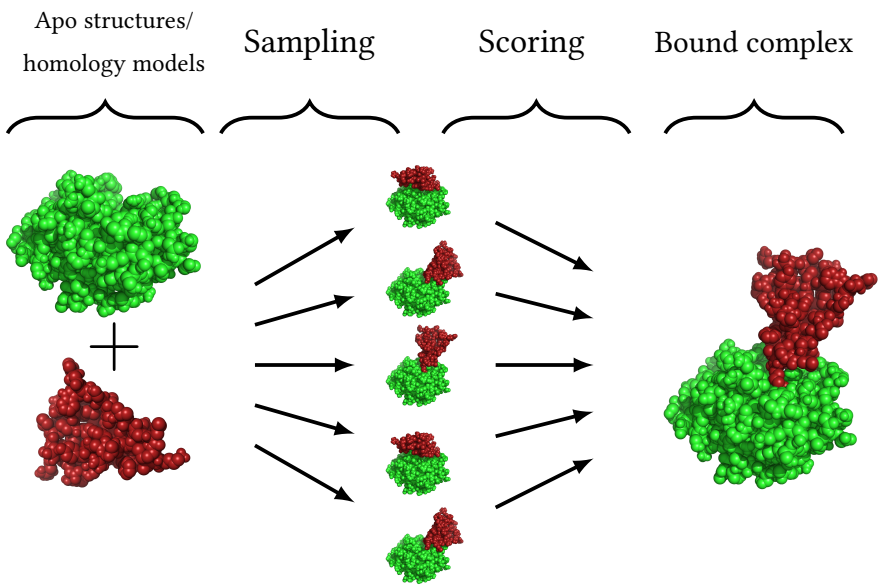


Figure 3.2. Overview of the docking process. First, a large number of possible models of the complex is generated (sampling). Then, these models are ranked to identify the near-native prediction (scoring).

atomistic models. In terms of degrees of freedoms, the different docking methods also vary: from considering only the six rigid-body degrees of freedom (for binary interactions) via including global motions of the proteins to sampling local conformational changes at the residue level. Similarly, docking scoring functions are diverse both in origin and complexity. Some of them were derived from statistical analysis or machine learning on crystallized protein-protein interfaces (knowledge-based potentials). Others are based on physical potentials using the system's free energy. Yet another type of scoring functions relies only on surface complementarity. Note that the level of detail of the protein representation, degrees of freedom and scoring function have to be adapted with respect to each other. In order to balance accuracy and sampling efficiency, docking methods typically carry out two phases:

1. a large-scale search using a low resolution protein representation, few degrees of freedom and a simple scoring function during which the overall geometry of the complex is exhaustively explored,

2. a refinement phase during which a subset of structures obtained from the low-resolution phase is optimized at higher resolution with more degrees of freedom and a more accurate and computationally demanding scoring function.

In practice, each of these two phases often consist of multiple sampling and scoring steps. A number of different docking methods have been developed to date (Table 3.1). Different approaches to initial sampling (large-scale search), refinement and scoring, and incorporating external information in docking will be discussed in the following.

3.4. Large-scale search

3.4.1. Exhaustive methods

Different docking programs employ different search algorithms to generate possible models of the complex. During the initial large-scale search, most methods consider the internal structure of the protein partners as rigid which limits the problem to sampling the three rotational and the three translational degrees of freedom. In an exhaustive enumeration scheme, the sampling over rotational and translational degrees of freedom is typically separated and Euler angles and center-of-mass translations are explored independently. For a given rotation of the molecule, the protein is moved relative to its partner (3D translational search). This process is then repeated for all possible rotations (at a specified level of discretization). In principle,

Table 3.1. Comparison of selected protein-protein docking methods.

Name	Method type[1]	Molecule types	Flexibility	Multi-body[2]	Symmetry	Refinement	Support for experimental data
ATTRACT	MIN, MC	Protein, DNA, RNA, small molecule	Normal modes, ensemble and domain docking	✓	✓	✓	Contact and interface data, cryo-EM, SAXS
ClusPro	FFT	Protein	-	-	✓	-	Contact data, SAXS
FiberDock	MC, MIN, R	Protein	Normal modes, side chain rotamers	-	-	✓	-
FTDock	FFT	Protein	-	-	-	-	-
GRAMM-X	FFT, MIN	Protein, small molecule	-	-	-	-	Interface data
HADDOCK	MD	Protein, DNA, RNA, small molecule	Full flexibility, ensemble and domain docking	✓	✓	✓	Contact and interface data, RDCs, cryo-EM, SAXS, IM-MS
Hex	FFT	Protein, DNA, small molecule	Ensemble docking	-	✓	✓	Interface data
IDOCK	FFT	Protein	-	-	✓	✓	Contact and interface data, cryo-EM, SAXS
PatchDock	G	Protein	Domain docking (FlexDock)	-	✓	-	-
pyDock	FFT	Protein	-	-	-	-	Contact and interface data, SAXS
RosettaDock	MC	Protein, small molecule	Ensemble docking, backrub, side chain rotamers	-	✓	✓	Contact and interface data
SwarmDock	PSO, MIN	Protein	Normal modes	-	-	-	Interface data
ZDOCK	FFT	Protein	-	-	✓	-	Contact and no-contact data

[1] Type of methods categorized into fast Fourier transform (FFT), geometric (G), Monte Carlo (MC), minimization (MIN), Particle Swarm Optimization (PSO) and molecular dynamics (MD). Refinement methods are marked additionally as R.

[2] Supports docking of more than two distinct (protein) bodies (heteromers).

the 6D rigi-body phase space can be sampled completely with such an exhaustive approach. For a typical protein size and a discretization of an Euler angle interval of 12 and 1.2 Å steps in the translational coordinates, a total of $\approx 10^{10}$ putative complex geometries have to be explored. In order to evaluate this huge number of possible conformations efficiently, two types of approaches have been developed: fast Fourier transform (FFT) correlation-based methods and shape matching algorithms.

Correlation methods FFT-based docking programs accelerate the sampling of the three translational degrees of freedom by performing a part of the calculation in Fourier space. Due to the high computational efficiency, the majority of currently used docking methods make use of an FFT correlation-based approach [144, 446, 69, 346, 348, 224, 149, 23, 237, 365, 326, 332, 267] and several web-servers based on FFT-based docking programs have been made available to the structural biology community [446, 287, 347, 83, 359]. I will describe the basic principle using a simple shape complementarity scoring scheme. The computation strategy is similar for different FFT-based search methods, however, they differ in terms of the potentials/scoring functions to describe the interaction between the proteins and in how these potentials are mapped to the grid. For both receptor and ligand protein, an energy grid is constructed with dimensions $N \times N \times N$ in Cartesian space. The respective protein is centered in its discretized grid. At each grid point (i, j, k), the scoring function

E is precalculated and stored. In our simple example, the protein grid $\mathbf{E}_{ijk}$ at voxel (i, j, k) with $1 \leqq i, j, k \geqq N$ is given by

$$\mathbf{E}_{ijk} = \begin{cases} 1 & : \text{on the surface of the protein} \\ \varrho & : \text{inside the molecule} \\ 0 & : \text{outside of the protein.} \end{cases}$$

For both receptor and ligand protein, this grid representation is created with ϱ being a large negative value ($\varrho \ll -1$) for the receptor protein and ϱ a small positive value ($0 \leqq \varrho \geqq 1$) for the ligand protein. A complex geometry for a translation (l, m, n) of the ligand protein relative to the receptor protein can now be evaluated based on shape complementarity by calculating the overlap product $C(l, m, n)$ between the two grids

$$C(l, m, n) = \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{k=1}^{N} \mathbf{E}^R_{i,j,k} \times \mathbf{E}^L_{i+l,j+m,k+n}.$$

Periodic boundary conditions are applied if the indices are greater than the dimension of the grid. For our simple protein representation, the overlap function adopts a favorable score (positive correlation) for surface contacts with a penalty for large overlap/penetration between the proteins (negative correlation). The overlap or correlation score has to be evaluated for all possible translations ($\forall l, m, n \in \{1, \ldots, N\}$); i.e. N^3 calculations have to be carried out. The entire calculation for $\mathbf{C} = \{C(l, m, n)\}$ can be carried out in one step according to the cross-correlation theorem by calculating the product of the Fourier transforms of the grids and transforming back into real space

$$\mathbf{C} = \mathfrak{F}^{-1}\left(\overline{\mathfrak{F}(\mathbf{E}^R)} \times \mathfrak{F}(\mathbf{E}^L)\right).$$

The FFT-based search in the translational degrees of freedom is then repeated for each rotation of the ligand protein. Instead of the simple shape complementarity score shown here, more elaborate scoring functions like pair-wise shape complementarity, electrostatic potentials, desolvation, hydrophobic complementarity and knowledge-based functions, can also be employed after mapping to the grid.

A drawback of FFT-based methods is that they need to employ a grid-based score. The docking results are hence not only sensitive to the choice of the potential but also to the discretization of the grid. Discretization can introduce large errors in the calculated score, especially when it comes to evaluating short-range interactions. To compensate for these discretization errors, the scoring function is typically softened.

But these inaccuracies often result in the production of false positive solutions; i.e., docking models that are far from the native structure but still have a good score. So false positive minima in the energy landscape are enhanced by the combined effect of softening the potentials and the remaining discretization error. In addition, as a result of the soft potentials, the resulting models often contain steric clashes and have to be optimized by further refinement stages (see Section 3.5). Even though the translational search is accelerated, the procedure still has to be carried out for all possible orientations of the partners. Several methods have tackled this problem by accelerating the search also in the rotational degrees of freedom [149, 364]. Note that the expansion in rotational degrees of freedom is usually only performed to a maximum shell radius and hence these methods cannot be applied for larger separations of the molecules. The strength but also the drawback of the FFT-based approaches is that inherently all possible geometries have to be sampled. However, a large fraction of these possible complex models are unphysical because they contain either large steric overlaps or only very few contacts. These geometries have an unfavorable score but are explored nevertheless in the FFT framework. Other drawbacks of FFT-based methods are that they are harder to adapt for multi-body docking problems and that experimental data can often not be directly included in the sampling process but have to be applied as a filter [492].

Geometric surface matching The second type of systematic/exhaustive search methods focuses on creating geometries that present a local shape complementarity between the partners leaving out some of the unphysical conformations generated by FFT-based approaches. The proteins are represented by molecular shapes (e.g., the Connolly surface [84]). The surfaces can be segmented into different parts; e.g., concave, convex and flat. Parts with high complementarity are then matched by an algorithm. Instead of explicit sampling, the six rigi-body degrees of freedom are only calculated when a binding orientation has been established based on a surface match. Since shape matching is local, many of the generated models contain steric clashes and the structures have to be relaxed in refinement phases. Several docking programs employ a geometric surface matching-based search including the small molecule docking program DOCK [416] and the protein-protein docking programs PatchDock [400], LZerD [467] and GAPDOCK [147]. Often geometric hashing algorithms are used to find local matches of shape descriptors [130] (e.g., surface patches in PatchDock and 3D Zernike descriptors in LZerD). Geometric surface matching methods have also been expanded towards modeling multi-component assemblies [123, 203, 252, 400] and including protein flexibility [400]. Similarly to FFT-based methods, experimental information is typically only applied as an a posteriori filter [403].

3.4.2. Guided methods

Similar to local shape matching, guided search algorithms do not explicitly explore all six rigi-body degrees of freedom but rather generate favorable complex geometries guided by an energy-like function. Exploration algorithms can be deterministic like multi-start energy minimization [506] and molecular dynamics (MD) [112, 283], or stochastic like Monte Carlo (MC) simulations [517, 516], Brownian dynamics [304] and genetic algorithms [147], or a combination of stochastic and deterministic sampling [310, 161, 1]. Typically, the proteins are represented in atomic detail with a force field where each atom is assigned van der Waals parameters and a (partial) charge. The interaction between the protein is then evaluated as the sum of all pair-wise atomic van der Waals and electrostatic interactions. Some methods also consider further energy terms like desolvation energy [128, 73] or hydrogen bond potentials [234]. The search uses at least the six rigi-body degrees of freedom and can also easily include other degrees of freedom to represent protein flexibility. The different terms in the interaction energy contribute differently according to the distance between the partners. At larger separation, electrostatic steering dominates the association. Desolvation becomes important when the partners get in contact, and van der Waals interactions dictate the scoring of closely packed interfaces. Due to the shape of the potential, van der Waals terms are very sensitive to slight misalignments/steric overlaps at the interface. This has to be considered when docking with unbound protein structures where conformational change at the interface could disfavor sampling near-native geometries.

Due to the rather detailed energy function, guided search methods are more computationally expensive than FFT and geometric matching methods. The calculations can be accelerated by either pre-calculating the energy function around the receptor protein on a grid or by employing a coarse-grained protein representation. The grid can be constructed for each of the energy terms (e.g., van der Waals and Coulomb interactions). The energy is then interpolated between the grid points (e.g., as in [127]). Interpolation errors for this classical grid-based energy evaluation increase significantly when the two proteins get closer and these inaccuracies might hamper the sampling process. Flexibility in the receptor protein may additionally modify the energy at the grid points and enlarge the approximation error.

Several docking programs employ coarse-grained protein representations in the initial large-scale search [506, 161, 422, 46]. Coarse-grained models group several heavy atoms into larger beads and hence reduce the number of particles that have to be considered in the pairwise energy calculations. In addition to accelerating the docking calculations, these reduced representations implicitly consider protein flexibility and are coupled to simplified force fields. These force fields often smoothen the

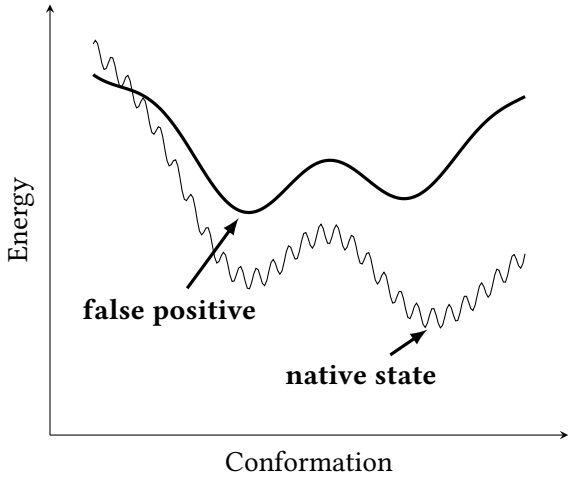


Figure 3.3. Schematic comparison of coarse-grained and atomistic potential energy landscapes. During the initial large-scale search, a low-resolution, soft scoring function is used that aims to retain the basic features of the energy landscape (thick line). During refinement, the resolution of the protein representation is increased and a more detailed search process is required to sample the details of the energy landscape.

energy landscape which again accelerates the sampling. However, on the downside, similar to grid acceleration, the coarse-grained representation might create new false positive minima in the energy landscape (Figure 3.3).

3.5. Refinement and Scoring

Most docking methods select a subset of generated models from the initial sampling and optimize and rerank these during further refinement stages. The reason for this is that the scoring functions and the resolution used in the initial stage are often not sufficient to reliably generate and distinguish near-native solutions from false-positive predictions. Typically, a few hundreds to thousands conformations are selected for refinement. Refinement can also be performed independently of docking methods; e.g., for models obtained from comparative modeling.

3.5.1. Increasing sampling resolution

A possibility to refine a prediction is to reproduce the search process at higher search resolution while using a more accurate scoring function. For discrete search pro-

cesses like FFT or MC docking, this corresponds to decreasing the step size in rotational and translational degrees of freedom. Typically, the vicinity of the previous docking models is explored in more detail (local search). This can also be done by switching to continuous sampling methods like energy minimization or MD simulations. Energy minimization can be performed in Cartesian coordinates of the protein atoms or in internal degrees of freedom (e.g., torsion angles). Cartesian minimizers are available through several molecular dynamics packages (e.g., GROMACS, AMBER, Charmm). Furthermore, the sampling resolution can be increased by considering additional degrees of freedom (see also Section 3.6). FiberDock, SwarmDock and ATTRACT include global backbone motions of the proteins through normal mode deformations [310, 297, 293]. ICM-DISCO and Rosetta sample local side chain conformations extensively using rotamer libraries [127, 161]. Sampling of side chain conformations has to be combined with adjustments of the relative positioning of the partners in order to improve the overall accuracy of the prediction [161] (see also Chapter 5). Molecular dynamics-based (MD) refinement allows to consider in principle full atomistic flexibility of the docking partners. However, MD simulations are limited by the restricted ability to overcome energy barriers in the short time scales accessible during the simulation. Therefore, MD refinement can often only yield slight improvements of the overall complex geometry [240]. It is possible to overcome the sampling limitations in MD by enhanced sampling methods such as simulated annealing or replica exchange [112, 216, 283, 284] or by coupling the atomistic simulations to a coarse-grained representation of the molecules [504].

3.5.2. Increasing representation resolution

During refinement and for the final ranking, a more detailed protein representation and scoring function are used. In most cases, the proteins are represented in atomistic detail. The scoring function can be either derived from a physics-based representation of the interaction or from training on known protein-protein interfaces (knowledge-based scoring functions).

Physical scoring functions Physical scoring functions aim to refine and rank docking models by calculating their total free energy. In principle, biomolecules should be described in a quantum-mechanical framework, however, this is computationally expensive and currently limited to representing systems of up to 100 atoms (see also Chapter 2). The majority of refinement methods are therefore based on a molecular mechanics force field description of the biomolecules (e.g., Rosetta [161], HADDOCK [112] and ATTRACT [393]). Molecular mechanics force fields have been used successfully in MD simulations to study the structure and dynamics

of biomolecules and can accurately reproduce thermodynamic and kinetic properties of biological systems. Force fields approximate the quantum mechanical energy landscape resulting from the distribution of electrons and nuclei in the molecule by potentials which only depend on the position of the atoms (Born-Oppenheimer approximation). The total potential V consists of several additive terms which represent chemical bonds and nonbonded interactions

$$
\begin{aligned}
V = & \underbrace{\sum_{i=1}^{N_{\text{bonds}}} \tfrac{1}{2} k_b^i \left(b_i - b_i^0\right)^2}_{\text{bond length}} + \underbrace{\sum_{i=1}^{N_{\text{angles}}} \tfrac{1}{2} k_\theta^i \left(\theta_i - \theta_i^0\right)^2}_{\text{bond angles}} \\
& + \underbrace{\sum_{i=1}^{N_{\text{torsion}}} \sum_{n=1}^{N_\tau} k_n \left(1 + \cos\left(n\tau_i + \delta_n^i\right)\right)}_{\text{dihedral angles}} \\
& + \underbrace{\sum_{i \neq j}^{N_{\text{pairs}}} \left[\epsilon_{ij} \left(\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6} \right) + \frac{q_i q_j}{4\pi\varepsilon r_{ij}} \right]}_{\text{nonbonded interactions}} .
\end{aligned}
$$

The first three terms are referred to as the bonded terms and sum over all bonds, angles and dihedral angles of the protein structure. Bond lengths and bond angles are controlled by harmonic potentials. Torsion angle potentials are represented by linear combinations of periodic functions. Non-bonded interactions are described by van der Waals and Coulomb potentials. Non-bonded interactions are also evaluated between proteins and solvent molecules. Solvation effects play a very important role for the association and stability of protein-protein complexes [265, 387, 5]. Water acts as a dielectric medium and favors association of hydrophobic residues due to entropic effects. Furthermore, water molecules are also found at the interface bridging the proteins via water-mediated hydrogen bonds [207]. A few groups have proposed methods to represent the solvent explicitly during docking [456, 222, 223, 334]. Most notably, van Dijk et al. docked the protein partners solvated by their first hydration shell using HADDOCK. They simulated desolvation in the encounter complex by expelling the initially present water molecules during association by a Monte Carlo procedure [456, 222, 223]. The HADDOCK program also routinely refines a few hundred docking models in a molecular dynamics simulation using explicit solvent (it2) [112, 97, 98]. Other methods consider water implicitly–if at all–either through knowledge-based potentials (see below) or by considering additional terms in the

scoring function such as

$$E = \alpha_1 E_{\text{bonded}} + \alpha_2 E_{\text{vdW}} + \alpha_3 E_{\text{ele}} + \alpha_4 E_{hb} + \alpha_5 E_{\text{BSA}} + \alpha_6 E_{\text{non-interacting}}$$

where E_{hb} is a hydrogen bonding potential, E_{BSA} accounts for buried (hydrophobic) surface area and $E_{\text{non-interacting}}$ considers the properties of the non-interacting protein surface. The weights of the different terms α_i can be chosen to improve the ranking of near-native docking models. Such scoring functions with optimal weights for each energy term are for example implemented in Rosetta [161], HADDOCK [112], ZRANK [349] and pyDock [73]. Atomic energy-like scoring functions always require a thorough sampling of side chain conformations and precise positioning of the protein partners.

Knowledge-based scoring functions With the increasing availability of 3D protein-protein complex structures in the PDB, knowledge-based scoring functions have become a powerful alternative to physics-based approaches. In many cases, it is very difficult and computationally expensive to accurately calculate the free energy for protein-protein binding (e.g., accounting for solvation effects, see Chapter 2 and above). Knowledge-based scoring functions circumvent this problem by extracting information from a set of known protein-protein interfaces to implicitly consider all these different effects.

In a statistical potential, the observed frequency of interface residue-residue contacts is compared to the expected contact frequency and over-or underrepresentation then is translated into a favorable; i.e., attractive or unfavorable; i.e., repulsive interaction potential. Expected contact frequencies are obtained by calculating the probability for random contacts of surface amino acids. In statistical mechanics, the N-body correlation function for a set of N particles is given by

$$g^{(N)}\left(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N\right) = \frac{p\left(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N\right)}{p_1(\mathbf{r}_1) p_2(\mathbf{r}_2) \ldots p_N(\mathbf{r}_N)}$$

where $p\left(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N\right)$ is the probability to find the N particles at $\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N$ and $p_i(\mathbf{r}_i)$ is the probability to find particle i in configuration $\mathbf{r}_i$. The N-body interaction potential (potential of mean force) can be calculated from g using the following formula

$$V^{(N)}\left(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N\right) = -k_B T \ln g^{(N)}\left(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N\right)$$

with k_B the Boltzmann constant and T the temperature of the system. For biological systems, typically a temperature of $T = 300\,\text{K}$ is used ($k_B T \approx 0.5\,\text{kcal/mol}$).

For deriving a statistical potential, a set of native 3D structures has to be supplied. In practice, the available data is not sufficient to derive the N-body potentials and therefore the evaluation is often limited to pairwise correlations and interactions. Hence, most knowledge-based potentials are based on contacts or distances between residues or atoms observed in experimental protein-protein complex structures (with a few recent exceptions [150, 388]). For residues A and B, a distance dependent potential $V_{AB}(r)$ can then be obtained by a Boltzmann inversion

$$V_{AB}(r) = -k_B T \ln \left(\frac{N_{\text{observed}}(A, B, r)}{N_{\text{expected}}(A, B, r)} \right)$$

with $N_{\text{observed}}(A, B, r)$ the number of observed contacts at distance r (within a distance interval) between residues A and B and $N_{\text{expected}}(A, B, r)$ the number of expected contacts if contacts were randomly distributed (no interactions between residues). The main difference between different knowledge-based potentials results from the calculation of the expected contact frequency; i.e., the determination of a reference state. Choosing the reference state is often the main hurdle in constructing statistical potentials. However, reference state-free scoring functions can also be derived [194].

Instead of a statistical analysis, scoring functions can also be directly trained on sets of decoy structures to distinguish near-native from non-native structures. For this different machine learning approaches like linear regression, neural networks and support vector machines can be used [242]. In the past, knowledge-based potentials have been applied successfully to several protein structure prediction prediction problems [309, 308, 514, 515, 489, 490, 103, 281, 521, 33, 509, 522, 412, 270, 388] and protein-ligand docking [154, 313, 195, 232, 125, 508, 163, 519]. Over the years, a large number of knowledge-based scoring functions for protein-protein docking have been developed as well [507, 506, 113, 56, 241, 11, 246, 523, 311, 460]. As mentioned above, most potentials are based on an analysis of contacts/pairwise distances. The potentials can be distinguished by the level of resolution used for the interface (atomic, residue based or larger parts of the interface), the type of interaction potentials used (distance-dependent or non-distance dependent), the number of atom/bead types used in the interaction potentials (which corresponds to the number of parameters for optimization of the potential) and the definition of the reference state. For example, the DFIRE (Distance-scale Finite Ideal Gas REference state) potential is an all-atom potential that represents proteins by 19 different atom types and was originally developed for protein structure prediction and stability analysis [522]. PISA (Protein Interactions Scored Atomically) also uses an atomic representation but trains a number of interpolated step potentials directly on a set of previously generated docking decoys to improve the ranking of near-native solutions [470]. Tobi used a linear programming approach to design side-chain based (coarse-grained) and atomistic

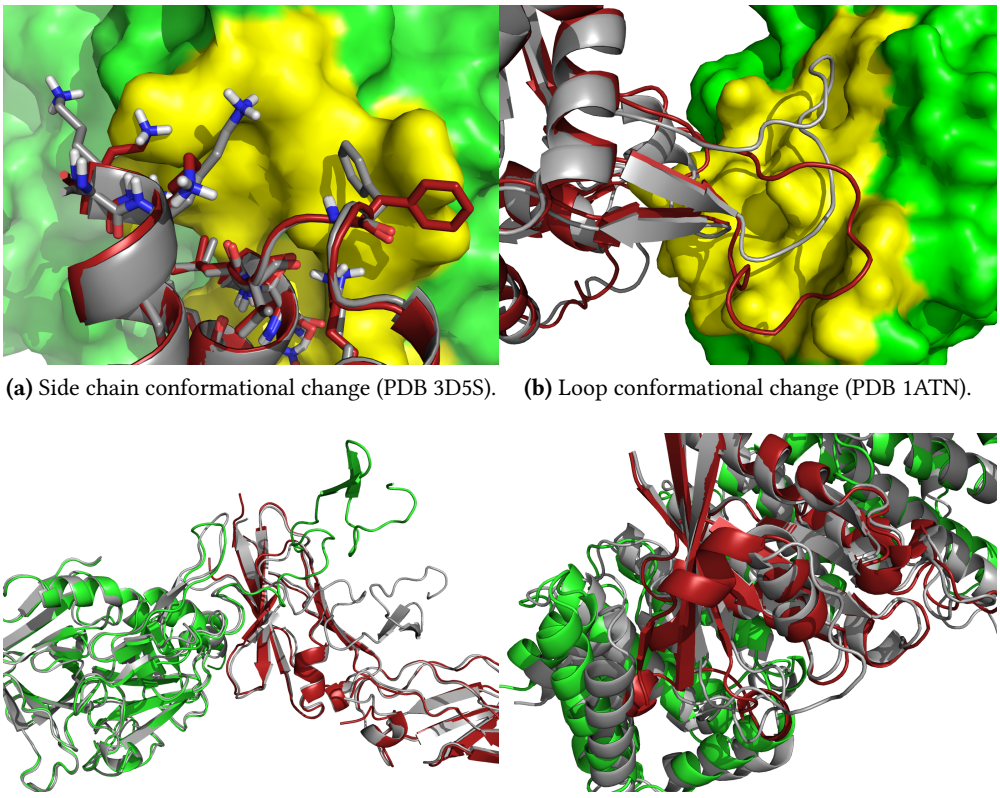
potentials for distinguishing a native complex from non-native decoys. The interactions between different atoms/residues were described by step potentials for which optimal distance cutoffs and depths of the potential were determined. Due to the simple shape of the potential (in comparison to Lennard-Jones type potentials), the resulting scoring functions were found to be less sensitive to conformational change when tested on unbound-unbound docking models [439]. Recently, a general tool box for designing scoring functions by machine learning on a set of decoy structures was developed in our lab [389].

3.6. Protein flexibility

In many experimentally known protein-protein complexes, the association partners undergo only little conformational change upon complex formation. These type of complexes can be efficiently treated by a rigid-body docking protocol combined with a flexible refinement step and rescoring as described above. However, in a large fraction of cases, the proteins undergo binding-induced conformational changes; i.e., the unbound forms of the proteins differ significantly from the bound form. These differences can range from small, local alterations (mostly, side chain rotamer changes or slight backbone changes at the interface) to large, global changes (loop and domain rearrangements, change in secondary structure or folding of disordered regions). Figure 3.4 illustrates some of these conformational changes upon complexation. Many rigid-body docking methods that perform well in the bound-bound case and in cases where little conformational change occurs, fail in dealing with increased protein flexibility. Hence, for the cases where the protein structures changes upon binding, it is necessary to consider conformational flexibility throughout the entire docking procedure. This is also desirable when using protein structures that have been obtained from comparative homology modeling. Depending on the degree of sequence identity, these models can deviate significantly from the native structure [124]. Furthermore, even in the case of high average target-template similarity, there may be regions that are less well aligned and that display structural inaccuracies which may affect the prediction. A range of different strategies have been developed to incorporate different types of protein flexibility in the docking process.

3.6.1. Ensemble docking

Rigid-body docking methods can be easily extended to include flexibility by representing the protein partners as an ensemble of structures. The structural ensemble can be obtained from experiments (e.g., from nuclear magnetic resonance spectroscopy or from multiple crystal structures) or from simulations (e.g., from molec-



(a) Side chain conformational change (PDB 3D5S). **(b)** Loop conformational change (PDB 1ATN).

(c) Global domain motion (PDB 1FAK). **(d)** Change in secondary structure (PDB 1IBR).

Figure 3.4. Illustration of types of binding-induced conformational changes. The unbound protein structures were superimposed on the bound complex. The unbound proteins are shown in red and green, the bound form in gray. The interface is highlighted in yellow.

ular dynamics, elastic network calculations or homology modeling based on different template structures). In the simplest possible setup, the docking procedure is repeated for each ensemble member independently, although switching between ensemble members can also be performed on-the-fly (Zhang, unpublished data). Ensemble docking mimics a conformational-selection-type process and selecting a structure that is more similar to the bound state can significantly improve docking accuracy. Obviously, ensemble docking is computationally more expensive. It may also increase the number of false positive solutions due to the large number of protein conformations (many of which deviate from the bound form). Ensemble docking has already been used successfully in the field of small-molecule docking [445, 196] and several protein-protein docking methods have adopted this strategy as well [65, 29, 112]. Gray and coworkers tested the effect of including computational and NMR ensembles in a flexible backbone protein-protein docking approach in Rosetta [65]. They recently also explored the possibility of generating an ensemble closer to the bound form based on the unbound protein structure with several tools in Rosetta, but found rather small improvements (Gray, personal communication). Zhang and Zacharias explored ensemble generation by molecular dynamics simulations and Rosetta tools and found similar results (Zhang, unpublished data). Hence, the main obstacle to successful ensemble docking is to generate a suitable structural ensemble and reliable strategies towards this goal have yet to be determined.

3.6.2. Loop modeling

Many loop modeling methods have been developed in the protein structure prediction field [134, 229, 290, 423, 254, 427, 4, 323, 185, 92, 335] and can in principle also be applied to modeling interface loop flexibility during protein-protein docking. Bastard et al. assessed a multi-copy/mean-field approach to include different loop conformations during docking with ATTRACT [29]. During docking, the energy of each loop copy is evaluated and a weight/probability is assigned to each copy according to the Boltzmann distribution. The higher the weight the stronger the interactions of a given loop copy and eventually the loop copy with the highest weight completely drives the docking minimization. Hence, the most suitable loop conformation is selected on-the-fly. This approach can also be applied to multiple loops simultaneously [29]. Note that in the ATTRACT approach the loop conformations have to be modeled prior to docking. Another strategy would be to model the loops a posteriori (see also Chapter 7). For CAPRI target 20 [206], Wang et al. used such a sequential modeling procedure in Rosetta to predict a protein-protein complex with flexible loops at the interface [478]. Flexible loops were identified prior to docking by comparison to orthologous complex structures. Initial models of the complex were generated

without flexible loops. The loops were then rebuilt onto the lowest energy docking models and resulting models were filtered with respect to additional constraints on the biological function of the complex. By combining available functional and structural information, it was possible to accurately model this extremely challenging target [478]. This CAPRI target demonstrated the importance of accurately identifying flexible regions prior to docking. The docking success achieved by the Baker lab most likely results from this accurate flexibility representation in combination with reducing the sampling space by available experimental information [478].

3.6.3. Multi-body domain docking

Several methods have adopted a ”divide-and-conquer” strategy to deal with large-scale domain rearrangements. The flexible protein is partitioned into rigid-body domain and connectivity restraints between the domains are employed during the docking. The hinge regions can be predicted; e.g., using an elastic network model [122]. This allows to include global conformational changes already in the initial rigi-body docking stage. MolFit splits the proteins into domains and sequentially docks them with a multi-stage two-body docking protocol [236]. A similar sequential strategy is adopted by the FlexDock method that performs docking with PatchDock on rigid domain pairs [400]. Rosetta uses a fold-tree representation to define flexible regions between centers of rigid molecules and hence enables domain rearrangement during the rigid-body sampling stage [477]. Karaca et al. developed a flexible multi-domain docking protocol in the HADDOCK software [218]. Again, the proteins were split into rigid domains with restraints imposed on the linkers. The domains were simultaneously docked in a rigid-body stage and then the resulting models were subjected to a semi-flexible refinement in internal coordinates and a fully flexible refinement in water to improve backbone and side chain conformations. For a benchmark of 11 protein-protein complexes, the authors found that this approach was capable of modeling domain rearrangements as large as 19.5Å. They also identified indicators that might allow to predict the extent of protein flexibility in order to choose the appropriate docking strategy [218].

3.6.4. Collective mode deformations

Global conformational changes in the protein partners have also been successfully modeled based on normal mode analysis in Elastic Network Models (ENM). The low-frequency eigenvectors can capture domain opening-closing motions and large loop rearrangements. Several studies have investigated whether the bound form of the protein is dynamically accessible from the unbound form through deformations in

low-frequency eigenmodes [109, 310, 342, 21, 226, 440, 24]. ENMs provide a simple strategy to calculate these soft modes. The protein is represented by distance-dependent harmonic springs between the atoms and this approximation yields a quite accurate description of the protein mobility [182]. ENM-based normal modes have been previously used to detect flexible regions and hinges in proteins [122, 430] and also to generate sets of conformations for ensemble docking [376]. Instead of using discrete conformations derived from deformations along the eigenmodes, the soft collective normal modes can also be used as additional degrees of freedom during docking. Normal modes have been used in FiberDock [294, 293], Eigenhex [466] and SwarmDock approach [310]. In the ATTRACT program, mode deformations have also been applied in the initial large-scale docking stage using a coarse-grained protein representation. In cases where the proteins undergo conformational changes that are captured by the normal mode deformations, the approach can improve the sampling and the scoring of near-native docking geometries and also enrich the pool of solutions that are close to the native state [297].

3.6.5. Folding and docking

Folding upon binding is a problem that extends beyond the field of docking as it requires to include protein structure prediction into the docking procedure. Only a few attempts have been made so far to tackle this problem. Most notably, Baker and coworkers developed a fold-and-dock protocol for predicting the structure of symmetric homo-oligomers from extended conformations of the individual monomers [93]. In a recent joint CASP-CAPRI experiment (CAPRI round 30), structures of homo-oligomers had to be both modeled and docked. The results of this round demonstrated the importance of accurate monomer modeling for the success of the quaternary structure prediction [262]. Simultaneously predicting the tertiary and the quaternary protein structure will also be important for modeling complexes involving intrinsically disordered protein regions.

3.7. Including external information during docking

For many complexes, external information is available about the binding mode. In contrast to ab-initio docking where no a priori knowledge about the structure of the complex is assumed and all possible binding of proteins are explored, including external information can significantly reduce the sampling space and improve the scoring of near-native geometries. External information can be either derived from biological experiments, bioinformatic predictions or from similarity of the binding mode.

We call methods that use information from experiments and bioinformatic analysis ”integrative modeling” approaches and methods that use structural similarity to known complexes ”template-based” docking methods.

3.7.1. Template-based docking

Since the 1980s, structures of individual proteins have been successfully modeled by similarity/homology (comparative modeling). If the sequence of a protein with unknown structure is sufficiently similar to the sequence of another protein with known structure, the structure can be modeled by assuming a similar fold based on the alignment of the target to the template sequence. The critical aspect of such a comparative modeling approach is the availability of a suitable template. Due to the rapid growth of the PDB and the limited structural scope of the protein universe, template-based modeling of individual proteins has become the dominant approach for building protein structures and is more reliable and efficient than ab-initio structure prediction.

Similarly, protein-protein complexes can also be modeled based on known complex structures. An advantage of these methods is that they are capable of including binding-induced conformational changes, provided that a suitable template is available. The following steps need to be carried out:

1. finding one or more appropriate templates,
2. aligning the target sequence with the templates,
3. building an initial model for the target by copying the structural segments from the aligned regions,
4. refining the structures by replacing the sidechains and constructing missing loops, insertions and termini.

The template search and the alignment of the sequences may be applied to each partner separately, but a template for the entire assembly also needs to be obtained. The alignment can be generated based on sequence, sequence profile, or a combination of the sequence and structure feature information. The quality of template-based models strongly depends on the accuracy of the template identification. Structures of protein-protein complexes are still relatively rare in the PDB, since complexes are generally more difficult to crystallize than individual proteins. While estimates for the number of protein interaction types vary [8, 148], it is clear that currently the Protein Data Bank only encompasses less than 50% of all interaction types and it may take probably decades to achieve sufficient coverage [148]. This lack of suitable

templates has generally hindered the broad application of template-based modeling to the problem of structure prediction for protein-protein interactions.

In contrast to threading-based modeling approaches [316, 314, 143], template-based docking uses a structural alignment of the individual docking partners onto an interface template library. Unlike the whole protein-protein interaction space, the interface structural space appears to be more limited and even chains with different folds often have similar interfaces. Structural alignments can be performed either based on the global fold or for interface fragments. The complex models are then constructed from the templates by superimposing the monomers on the selected interaction template. These tentative geometries are then evaluated by scoring functions that measure both the structural similarity between model and complex template and physiochemical properties of the resulting interface. Since the initial template selection is based on structural alignment, these methods have the potential to detect distant or even non-homologous templates. Vakser and coworkers even showed that in principle suitable structural templates are available for all complexes where the structures of the individual association partners are known. However, the accuracy of template-based docking is still low ($\approx$ 23% when at least one of the chains has no homologous template with more than 40% sequence identity with the target) and it remains unclear how to exploit interface similarity to model the global complex structure. Furthermore, the crude initial geometries derived by template-based docking need to be optimized. The quality of the complex models depends strongly on the quality of the templates and template-based docking typically succeeds for cases where templates with a sequence identity of $>$ 40% are available [245, 418, 472]. For low-quality templates, efficient full-length complex structure refinement methods are still lacking to date [432]. Several template-based docking methods have been developed during the last 10 years. Kundrotas et al. were among the first to use a template-based docking approach based on an interface library. This library was extracted from the DOCKGROUND database of $\approx$ 12,000 complexes. Docking models were then built by partial or full alignment of the partners to the template and scored by the TM-scores of this alignment [418]. The protein-protein interaction database PrePPI stores information on experimental and predicted protein-protein interactions. PrePPI generates models of the complex by querying the PDB for structural neighbors of the individual proteins and structurally aligning the docking partners to complexes involving the detected structural neighbors. The models are then assessed by a Bayesian network to determine the probability that these proteins interact [511]. PRISM [451] uses an interface database derived from non-redundant structures from the PDB and aligns the target protein surfaces to the interface templates using MultiProt [411]. If complementary partners of a template interface are similar to surface regions of the two target proteins, it is

assumed that the two proteins can interact through these motifs. The chains are superimposed on the interface template and scored by their similarity to the template and steric overlap between the partners. These initial models are then refined by the FiberDock method [293]. Similarly, ISEARCH and iWrap use special domain-domain interaction libraries to scan the surfaces of the association partners for interaction sites and construct the models of the bound complex [173, 189]. Recently, Xue et al. proposed a new template-based docking method (CA-CA docking). They derived interfacial residue restraints by similarity to homologous complexes and used these to drive a standard docking run in HADDOCK [496]. This approach combines a protocol used successfully in integrative modeling/data-driven docking with information derived from the interface templates and can therefore be considered intermediate between template-based docking and free docking. Compared to a simple alignment of the partners as used in most template-based docking methods, this protocol allows for larger sampling flexibility. Interestingly, Xue et al. found that CA-CA docking generates more accurate docking models than true interface-driven docking and refinement of the bound-unbound superposition [496].

3.7.2. Integrative modeling

For many complexes, biochemical and biophysical experiments have provided a wide range of low resolution structural information (see also Chapter 2). This limited amount of data can be combined with docking to yield more accurate models of protein-protein interactions. Similarly, bioinformatic predictions (e.g.; interface predictions or co-evolution data) can be incorporated [95, 331]. During docking, the scoring function is typically supplemented by a pseudo-energy term that accounts for the available data. In the following, we will focus on discussing methods that include experimental data. We can distinguish between experiments that give information about the overall shape of the biomolecular complex, about the interface and local contacts, and about orientation and symmetry. Considering experimental data during docking is highly challenging, since the data may be sparse, noisy, ambiguous and even incoherent due to conformational heterogeneity [402]. The uncertainty in the data affects the accuracy of the prediction, however, reliable standards for assessing model error have yet to be established.

In general, an integrative modeling protocol consists of the following steps:

1. Collecting available experimental data,

2. Choosing an appropriate representation for the docking partners and the experimental data,

3. Sampling possible complex models,

4. Analyzing and validating the models.

Integrative modeling approaches have been applied successfully to elucidate the structure of large molecular machines like the 26S proteasome [250], the nuclear pore complex [7, 386, 414], the Salmonella typhimurium Type III secretion system [279] and the Mediator complex [366]. Many docking programs have been expanded over the years to include different kinds of experimental information and to cope with the inherent uncertainty of the data [392, 112, 248, 371, 107, 491, 209, 94, 459, 378, 398]. Some of them are specialized towards one type of experimental data, but many allow combining data from multiple types of experiments. The Rosetta program works by exhaustive Monte Carlo sampling considering the underlying stereochemistry and physical principles of proteins and can also use experimental restraints from NMR [248, 291], SAXS [371] or cryo-EM [107, 480] to improve the sampling process. This approach was applied successfully to several biological problems [279, 510, 371, 409, 408]. HADDOCK (High Ambiguity Driven DOCKing) was one of the first docking programs that specialized in dealing with a large variety of experimental input [112]. Currently, $\approx$ 130 complex structures calculated with HADDOCK have been deposited in the PDB. HADDOCK uses ambiguous interaction restraints (AIRs) in order to include information about putative interface residues. Such information can be obtained from site-directed mutagenesis, NMR spectroscopy (CSP, RDCs, NOEs), cross-linking/mass spectrometry [178] or H/D exchange. HADDOCK distinguishes between *active* residues; i.e., residues that are known to form contacts, and *passive* residues; i.e., residues that might be at the interface. For every active residue A, an AIR restraint is defined between this residue and all active and passive residues on the partner molecule by calculating the effective distance d_A^{eff}

$$d_A^{\text{eff}} = \left(\sum_{i=1}^{N_A} \sum_{j=1}^{N^B_{\text{active+passive}}} \frac{1}{d_{ij}^6} \right)^{-\frac{1}{6}}$$

with N_A the number of atoms in the active residue and $N^B_{\text{active+passive}}$ the number of atoms in the active and passive residues on the partner molecule. An upper limit of typically 2 Å for the effective distance is enforced by a flat bottom harmonic potential that becomes linear beyond a given cut-off distance. For effective distances shorter than the given limit, no penalty is applied. Due to the inverse summation, the shortest distances d_{ij} essentially dominate the effective distance d_A^{eff} and typically a residue-residue distance of 4-5 Å is sufficient to fulfill the restraint. AIR restraints enforce active residues to make contact but allow ambiguity with respect to the partner

residue. Furthermore, in order to deal with false positives, 50% of the AIR restraints are randomly removed during docking. HADDOCK also supports orientational restraints in the form of RDC or diffusion anisotropy restraints. Symmetry restraints can be imposed as well. In the last years, protocols including shape data (SAXS and cryoEM) have also been tested [459, 219]. However, these protocols are not yet able to systematically deal with a high degree of ambiguity and noise in the same way as it is done for interface data.

The Integrative Modeling Platform (IMP) has been used repeatedly to study the structure of individual proteins and large assemblies [6, 31, 250, 465, 126]. IMP is a software platform that was designed to facilitate writing integrative modeling applications. IMP can use a variety of representations for the biological macromolecules ranging from atomistic scale to highly coarse-grained beads. This allows to model different parts of the system at different resolutions depending on the available information. A range of restraints for experimental data are available including SAXS profiles [399], proteomics data [251], EM images [465], FRET [403], cross-linking [352] and NMR data [386]. To consider experimental uncertainty, the restraints are often formulated using a Bayesian approach [50]. IMP also provides a large collection of sampling algorithms and analysis tools [378].

IDOCK [403] is an integrative modeling approach based on the PatchDock docking method [400] that can filter docking models by a large number of experimental data. IDOCK can handle SAXS profiles, 3D cryo-EM densities, 2D cryo-EM class averages, residue-type content at the interface and cross-linking/mass spectrometry data by making use of functions implemented in IMP [378]. The authors also tested a combination of different experimental data and found that combining global shape data with local contact data was especially beneficial [403].

3.8. Assessment of docking methods

Very early in its development, the docking community decided to organize a common evaluation procedure. As a result, in 2001, the CAPRI blind prediction experiment (www.ebi.ac.uk/msd-srv/capri) was initiated as a community-wide experience. In close dialogue with the community, the CAPRI management team has designed and established consensus assessment criteria for classifying the accuracy of docking methods. Besides the CAPRI challenge, docking benchmarks have been created for testing protocols on a common basis.

3.8.1. The CAPRI blind prediction experiment

In the CAPRI challenge, participating groups can test their docking algorithms on not-yet published experimental structures of biomolecular complexes. Structural biologists can bring a newly-solved structure from X-ray crystallography, NMR or high-resolution cryo-EM to the CAPRI organizing committee. This structure can then be offered to the docking community as a target. The structural biologists need to wait until the round closes before releasing their structure to the public. Predictors are given typically between one and four weeks to submit 10 models for the target (depending on the timeline for publishing the target and its difficulty). If possible, unbound structures are supplied, but in recent rounds the participating groups often had to model at least one of the structures by homology. In addition to the Predictor round, there is also a scoring challenge. For this predictor groups can upload up to 100 models. All uploaded models are supplied to the groups participating in the scoring challenge. The idea of this separate scoring round is that good predictions may be overlooked during scoring and other scoring methods might be able to detect them; scorers can therefore take advantage of all the models sampled to evaluate their methods. This corresponds to the generally accepted concept of separating the docking problem in a sampling and a scoring problem.

Between 2001 and 2016, 36 rounds of CAPRI have been run with 107 targets. About 25 groups participated in the latest CAPRI rounds. Over the years, the targets have become more and more diverse including e.g., predicting peptide-protein complexes, interfacial water molecules, binding free energies and the effect of mutations. The difficulty of the targets has also increased, most recent targets presented a high degree of protein flexibility. In general, the targets proposed during CAPRI have stimulated the field and accelerated inclusion of experimental data and design of new protocols (e.g., a range of peptide-protein docking protocols were created in response to several peptide-protein targets in CAPRI between 2013 and 2016). The failures for certain targets in CAPRI also highlight the challenges faced by the docking commu-

nity, namely binding-induced conformational change and its prediction based on the unbound structure, and accurate (homology) modeling of the docking partners. Indeed, in the last couple of rounds, there were several targets for which none of the participating groups submitted any model close to the native form. Targets and results for the ATTRACT docking engine from the most recent CAPRI evaluation are discussed in Chapter 9.

3.8.2. Docking benchmarks

In addition to the evaluation of docking methods provided by CAPRI, docking benchmarks have been published for protein-protein [306, 199, 200, 473, 146, 228], peptide-protein [275, 253], protein-DNA [457] and protein-RNA docking [28, 193]. Parts of Weng's protein-protein docking benchmark have been supplemented by binding affinity data [221, 473]. A small benchmark for multi-body docking has also been proposed [220]. The benchmarks contain non-redundant complex structures found in the PDB. The structures of the complex were determined by X-ray crystallography at high resolution. In many benchmarks, the unbound structures for the components have also been collected from the PDB allowing to create realistic test scenarios. In some benchmarks, the unbound structures have been (homology) modeled to achieve greater coverage ; e.g., in DOCKGROUND where input models with different deviations from the bound structure have been systematically created [228] and in the large-scale benchmark PPI4DOCK that contains more than 1,000 test cases (Guerois, unpublished data). Many benchmarks classify the complexes by their function/type and the degree of docking difficulty. Docking difficulty is determined by the amount of binding-induced conformational change. Benchmarks are useful for developing docking methods and assessing their performance for a large variety of systems and difficulty levels.

3.8.3. Assessment criteria

According to the standards introduced by CAPRI, the quality of a protein-protein docking model is assessed by evaluating its interface root-mean-square-deviation (IRMSD), ligand root-mean-square-deviation (LRMSD) and fraction of native contacts (fnat) with respect to the native complex structure. The interface is defined by all residues that have heavy atoms within 10 Å of any heavy atom of the partner. The docking model is fitted onto the interface of the native complex and the deviation of the backbone interface atoms is then evaluated (IRMSD). For calculating LRMSD, the docking model is fitted onto the receptor protein of the native structure and the RMSD between the backbone atoms of the ligand protein are evaluated. For

Table 3.2. CAPRI quality measure used to evaluate protein-protein docking models.

Quality	Criteria
High accuracy ($\star\star\star$)	(IRMSD $\leqq 1$ Å or LRMSD $\leqq 1$ Å) and fnat $\geqq 0.5$
Medium accuracy ($\star\star$)	(IRMSD $\leqq 2$ Å or LRMSD $\leqq 5$ Å) and fnat $\geqq 0.3$
Acceptable ($\star$)	(IRMSD $\leqq 4$ Å or LRMSD $\leqq 10$ Å) and fnat $\geqq 0.1$

Table 3.3. CAPRI quality measure used to evaluate peptide-protein docking models.

Quality	Criteria
High accuracy ($\star\star\star$)	(IRMSD$\leqq 0.5$ Å or LRMSD $\leqq 1$ Å) and fnat $\geqq 0.8$
Medium accuracy ($\star\star$)	(IRMSD $\leqq 1$ Å or LRMSD $\leqq 2$ Å) and fnat $\geqq 0.5$
Acceptable ($\star$)	(IRMSD $\leqq 4$ Å or LRMSD $\leqq 4$ Å) and fnat ≥ 0.2

calculating fnat, the contacts are extracted from the native complex. Two residues are in contact if any of their heavy atoms are within 5 Å of each other. Then the contacts on the docking model are extracted and the fraction of native contacts that is correctly reproduced by the model is evaluated. Based on these criteria, CAPRI classifies docking models as of high ($\star\star\star$), medium ($\star\star$) and acceptable ($\star$) quality. The quality criteria are summarized in Table 3.2.

For peptide-protein complexes, tighter criteria have been introduced in the 2016 CAPRI evaluation to reflect the smaller size of the peptide and the interface. The interface is defined by all residues where the C_β atom is within 8 Å from any C_β atom on the partner. Contacting residues are defined as having heavy atoms within 4 Å distance. The overall quality classification has also been tightened as summarized in Table 3.3.

3.9. Conclusion and Outlook

The Protein Data Bank contains a large number of three-dimensional structures of isolated proteins but rather few complexes. Computational docking methods can help to fill this gap by predicting the 3D structure of biomolecular complexes from the structures of the individual constituents. As documented by the community-wide

blind prediction experiment CAPRI, protein-protein docking methods have become faster and more versatile and are now able to give insight into a wide range of biological systems. Benefiting from enormous progress in computer technology and algorithms, including the usage of GPUs, docking experiments can now be carried out on large scales in reasonable time frames and genome-wide investigations of biomolecular interactions are coming within reach. A major trend is the use of ever-increasing amounts of data: be it by training new scoring functions on known structures of protein-protein complexes, by modeling the complexes themselves from experimental complex structures, or by including experimental data that could be even extracted automatically from the literature via text mining [17]. But docking methods still face three large challenges. The first one concerns the modeling of binding-induced conformational change and (protein) flexibility. As assessed by CAPRI, accurate predictions can be achieved for complexes that only present small-scale conformational change between bound and unbound form. However, approximately 35 % of all complexes found in docking benchmark 5 or among the CAPRI targets present larger-amplitude conformational change upon complexation. Not considering flexibility affects the scoring of near-native models. In the worst case, the docking protocols might even fail to sample a near-native geometry due to clashes. Unfortunately, it is very difficult to predict when and what type of conformational change will occur, since induced-fit processes can only be observed in an encounter complex close to the native state. The second challenge refers to modeling multi-component assemblies. Many complexes are composed of more than two entities and fast methods are needed in order to deal with these large combinatorial sampling problems. Only few methods can simultaneously dock multiple structures to date. The third challenge relates to including experimental data in docking. Dealing with new types of data, the inherent experimental uncertainty and possible incoherence between multiple data sources will become necessary.

4. The ATTRACT Docking Engine

The ATTRACT software suite is a large collection of programs and tools for modeling biomolecular complexes. ATTRACT has been successfully applied to a range of interesting biological problems and also regularly participates in the blind docking challenge CAPRI. This chapter describes the most common features and characteristics of the ATTRACT docking program. Parts of this chapter have been previously published in [394] and [99].

4.1. Introduction

Virtually all cellular processes involve the interaction of biomolecules. Three-dimensional (3D) structures of interacting complexes are essential for understanding their biological function, regulation and potential modulation. However, experimental structure determination is highly challenging and has so far only succeeded in elucidating a small fraction of complexes. Furthermore, it may not be feasible for all low-affinity/transient interactions. Therefore, reliable predictions of protein-protein complexes and protein-nucleic acid complexes are in high demand and over the last 15 years a range of different docking programs have been developed (Chapter 3).

The ATTRACT docking engine [506, 296, 99] can perform structural modeling for a large variety of biomolecular interactions. It has been developed for over a decade in the Zacharias lab and as of April 2016 comprises a core program with approximately 20,000 lines of code and over 80 tools for input preparation, data processing and analysis. ATTRACT has been applied to protein-protein, protein-DNA [406], protein-RNA[407, 67] and protein-small molecule complexes [298] and successfully predicted targets in various rounds of the blind protein-protein docking challenge CAPRI [299, 101, 260] (rank 2 in 2016 CAPRI evaluation) (Lensink, personal communication, see also Chapter 9). ATTRACT distinguishes itself from other docking programs by its coarse-grained force field, the possible use of protein flexibility throughout the docking search, and the simultaneous docking of any number of (protein) bodies (Figure 4.1). In addition to predicting complexes ab-initio, ATTRACT can also include a variety of experimental data in the docking process. It was expanded to fitting molecules in low-resolution cryo-EM density [94, 96] and supports

incorporating information obtained from e.g. NMR, cross-linking/mass spectrometry and mutational experiments. Recently, an integrative modeling approach based on small-angle X-ray scattering data was developed (Chapter 8) [392]. Flexible interface refinement of ATTRACT-generated rigid-body models can be performed by the iATTRACT protocol [393] (Chapter 5). A part of the functionality in ATTRACT has been made easily accessible through web-interfaces [99]. In this chapter, I will describe the ATTRACT docking engine, its characteristics and parts of its basic modeling capabilities in more detail. I will present the different protein representations and interaction potentials, ab-initio rigid-body docking, different ways of including (protein) flexibility and the ATTRACT web interfaces.

4.2. Protein representation

4.2.1. ATTRACT coarse-grained force field

Coarse-grained protein representations have the advantage of being coupled to a simplified, smoothened energy landscape that contains fewer docking energy minima and therefore allow for much rapid and fully converged energy minimization compared with an atomic resolution representation. The empirical, coarse-grained protein representation in ATTRACT is intermediate between a residue-level and full atomistic description. It represents each amino acid of a protein by up to four pseudo atoms (Figure 4.2). The protein main chain is represented by two pseudo atoms per residue (located at the backbone nitrogen and backbone oxygen atoms, respectively). Small amino acid side chains (Ala, Asp, Asn, Cys, Ile, Leu, Pro, Ser, Thr, Val) are represented by one pseudo atom (geometric mean of side chain heavy atoms). Larger and more flexible side chains (Arg, Gln, Glu, His, Lys, Met, Phe, Trp, Tyr) are represented by two pseudo atoms to better account for their shape and the dual chemical character of some side chains. Effective interactions between pseudo-atoms A and B are described by soft distance-dependent Lennard-Jones (LJ)-type potentials of the following form (Figure 4.2)

$$V_{AB}(r_{ij}) = \begin{cases} \epsilon_{AB}\left[\left(\frac{\sigma_{AB}}{r_{ij}}\right)^8 - \left(\frac{\sigma_{AB}}{r_{ij}}\right)^6\right] & \text{for attractive pairs} \\ -\epsilon_{AB}\left[\left(\frac{\sigma_{AB}}{r_{ij}}\right)^8 - \left(\frac{\sigma_{AB}}{r_{ij}}\right)^6\right] & \text{for repulsive pairs if } r_{ij} > r_{\min} \\ 2e_{\min} + \epsilon_{AB}\left[\left(\frac{\sigma_{AB}}{r_{ij}}\right)^8 - \left(\frac{\sigma_{AB}}{r_{ij}}\right)^6\right] & \text{for repulsive pairs if } r_{ij} \leqq r_{\min}. \end{cases}$$

where σ_{AB} and ϵ_{AB} are effective pairwise radii and bonding energies for attractive or repulsive LJ pairs. At the distance $r_{\min} = \sqrt{\frac{4}{3}}\sigma_{AB}$ between two pseudo

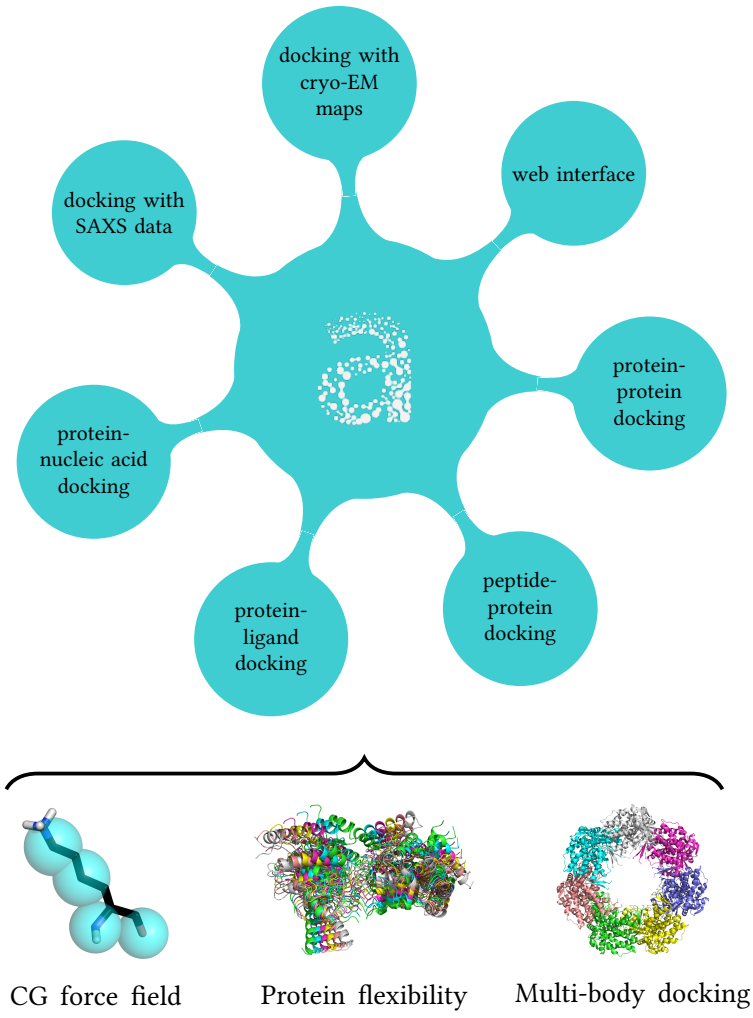


Figure 4.1. The ATTRACT docking engine, its functionality and its main features.

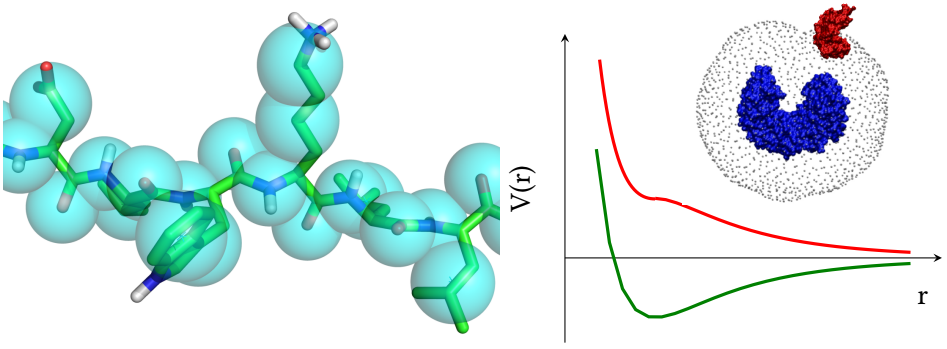


Figure 4.2. The ATTRACT force field. (a) The ATTRACT coarse-grained force field represents each amino acid by up to 4 pseudo-atoms (2 for the backbone and 1-2 for the sidechain). (b) Interactions between pseudo-atoms are described by attractive (green) and repulsive (red) Lennard-Jones-type potentials.

atoms, the attractive LJ-type potential has the energy $e_{\text{min}} \approx -0.1\epsilon_{AB}$. In contrast to the original force field [506], this form allows for purely repulsive interacting pseudo-atom pairs. For each pseudo-atom pair, attractive and repulsive LJ parameters were initially derived from a statistical analysis of contact probabilities at known protein-protein interfaces and then iteratively optimized by minimizing the root-mean-square-deviation of near-native docking minima and comparing the scoring of near-native minima with many high-scoring decoy complexes [129]. The LJ-type interaction potentials were parametrized to reflect both surface complementarity and physico-chemical properties of protein-protein interfaces. The LJ interactions also implicitly include solvation effects by favoring association of hydrophobic residues at the interface.

In addition to the LJ-type potentials, a Coulomb-type term accounts for electrostatic interactions between charged side chains (Lys, Arg, His, Glu, Asp)

$$V(r_{ij}) = \frac{q_i q_j}{\varepsilon(r_{ij}) r_{ij}}.$$

The Coulomb potential is damped by a distance dependent dielectric constant $\varepsilon(r_{ij}) = 15 r_{ij}$ in order to mimic Debye-Hückel screening by the solvent molecules. ATTRACT also provides coarse-grained models for nucleic acids and their interactions with proteins [406, 407].

4.2.2. OPLS atomistic force field

In addition to the coarse-grained ATTRACT force field presented above, an atomistic force field is also available in ATTRACT. The force field uses a united-atom representation for nonpolar hydrogen and represents all other atoms, including all polar hydrogens, explicitly. The force field employs standard Lennard-Jones (LJ) potentials to describe van der Waals interactions and a Coulomb-type term to represent charge-charge and dipolar interactions between atom types i and j

$$V(r_{ij}) = 4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] + \frac{q_i q_j}{\varepsilon r_{ij}}$$

where σ_{ij} and ϵ_{ij} are given by the combination rules $\sigma_{ij} = \frac{\sigma_i+\sigma_j}{2}$ and $\epsilon_{ij} = \sqrt{\epsilon_i \epsilon_j}$. The parameters σ_i, ϵ_i and q_i are based on the OPLS force field [212] and were derived from a modified version of the parallhdg5.3.pro parameter file [271]. The dielectric constant ϵ is typically set to 10 to account for solvent screening effects.

4.2.3. Grid-based energy calculation

Docking calculations can be accelerated by precalculating the potential energy on a grid [100]. This fast energy evaluation in combination with the coarse-grained docking approach makes it possible to scan hundreds of thousands configurations in the initial docking stage of the ATTRACT program. The grid represents the complete spatially discretized interaction potential around one of the protein partners (which is typically the larger partner and also referred to as the "receptor protein"). At each grid point, energy and forces are stored. The potentials are evaluated at run time by trilinear interpolation between the values of neighboring voxels. In contrast to previous grid-based approaches, ATTRACT still calculates short-range interactions explicitly (for distances $r_{ij} < d_P$ with d_P being the distance cutoff or plateau distance) using a stored neighbor list of atoms at each grid point and only uses the interpolation to evaluate the long-range interactions ($r_{ij} > d_P$)

$$V_{\text{total}}(r_i) = \underbrace{\sum_{r_{ij}<d_P} V(r_{ij})}_{\text{explicit; short-range}} + \underbrace{V(r_i) - V(d_P)}_{\text{precalculated; long-range}} \quad \text{for atom } i.$$

This reduces the interpolation error significantly [100] and also allows to include moderate flexibility of the receptor protein during docking. Recently, this energy calculation routine was ported to the GPU yielding a speed-up of up to 100 [121] compared to the current implementation in ATTRACT [100]. Grid-acceleration is

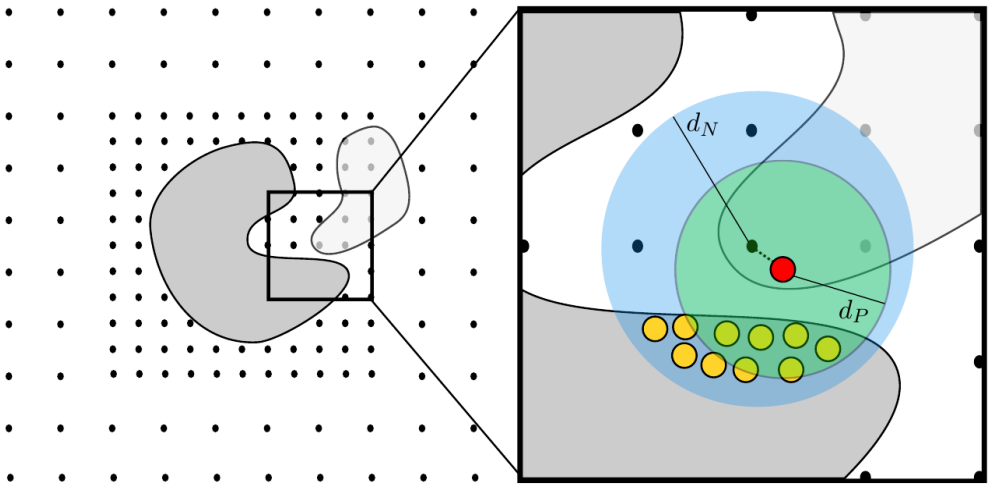


Figure 4.3. Grid-based energy calculation in ATTRACT. Short range interactions are calculated explicitly, long-range interactions are calculated implicitly by interpolation from the grid. Inset: Illustration of the energy calculation for one atom (red). Interactions with receptor atoms within the plateau distance d_p (yellow) are evaluated explicitly. A list of neighboring receptor atoms within a neighbor distance d_N is taken from the nearest grid point (dashed line). Long-range interaction energies are interpolated from the nearest grid points. The figure was taken from [121].

compatible with multi-body docking, ensemble docking (for small conformational changes within the ensemble) and the use of normal mode deformations during docking.

4.3. Standard docking protocol

Figure 4.4 shows a flowchart of a typical docking run in ATTRACT. For simplicity, a protein-protein docking run is described, however, most of this is also valid e.g. for protein-nucleic acids docking. As an input, the user needs to supply the atomic coordinates for the two proteins of interest (PDB files). The docking run consists of the following steps:

1. **Structure processing and protein representation.** The structures of the proteins are converted into the ATTRACT format with the ATTRACT tools `aareduce` and `reduce`. `aareduce` uses PDB2PQR [111, 110] to rebuild missing side chain heavy atoms and if necesssary adds hydrogens to the structure of the biomolecule.

2. **Starting positions.** Starting positions for the docking minimization are generated by placing the ligand protein at various positions and orientations around the receptor protein. The position and orientation of the ligand protein is defined by its center-of-mass translational (x, y, z) and rotational (ϕ, θ, ψ) (Euler angles) degrees of freedom.

3. **ATTRACT rigid-body docking.** Each of the starting positions is optimized during a potential energy minimization in the six rigid-body degrees of freedom using a quasi-Newtonian minimizer (default settings). Alternatively, ATTRACT can also perform a Monte Carlo search in the translational and orientational degrees of freedom. During the rigid-body docking, the proteins are represented by the coarse-grained ATTRACT force field.

4. **Ranking and data processing.** The generated docking models are ranked according to their ATTRACT score evaluated within a short cutoff. Redundant docking models (within 0.05 Å from a better scored model) are discarded (ATTRACT tool `deredundant`).

5. **Refinement.** Typically the top-ranked 200 models are selected for atomistic refinement with the iATTRACT protocol [393] (Chapter 5).

In the end, models for the bound complex are obtained. ATTRACT usually generates hundreds of thousands of complex models in a single docking run. A typical docking run is executed in a few hours on a standard Desktop PC.

The ATTRACT standard docking protocol has been tested on 226 protein complexes from protein-protein docking benchmark 5.0 [473]. The protocol yielded a success rate of 48% among the top-ranked 100 models and an overall success rate of 96% when evaluating by the CAPRI one-star criterion (Schindler, unpublished data). For docking cases in which both partners undergo no or only very little conformational change upon binding, this protocol already yields very accurate results placing near-native predictions often among the top-ranked 50 models. However, in many cases, the conformation of the proteins in the bound complex differ significantly from their unbound structures in solution and this flexibility has to be taken into account during docking.

4.4. Protein flexibility

Several different possibilities are available to model flexibility throughout the different docking stages in ATTRACT: normal mode deformations, ensemble docking and domain docking. Docking with soft harmonic mode deformations and ensemble

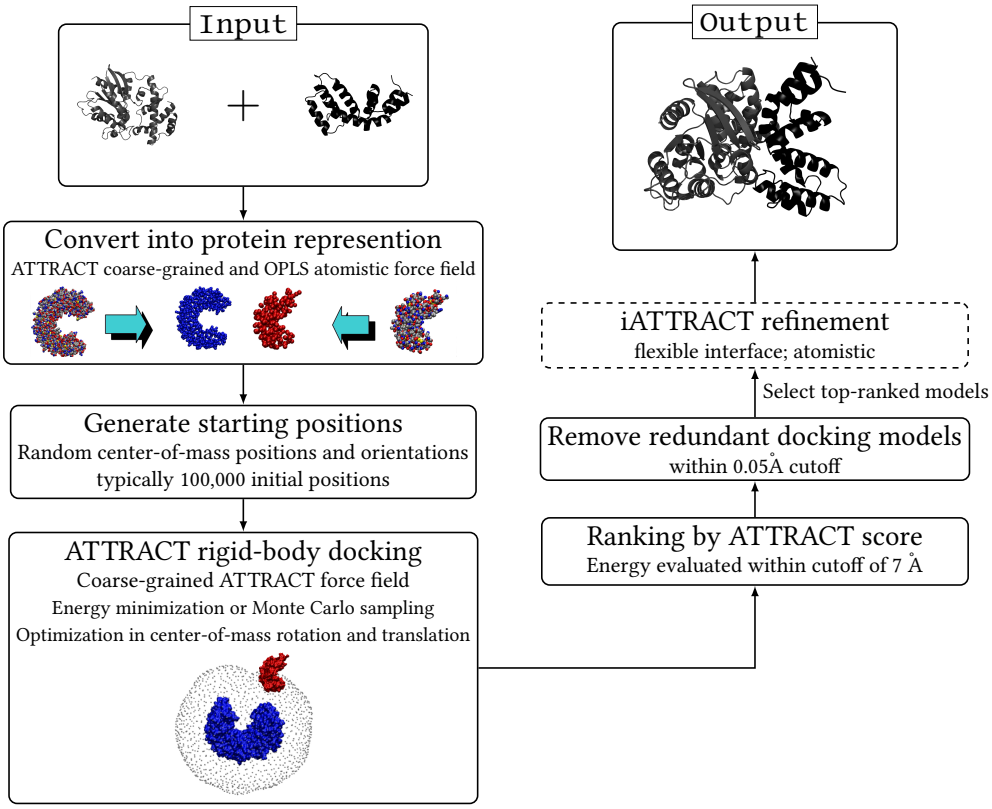


Figure 4.4. Overview of a typical docking run with ATTRACT.

docking are available through the ATTRACT Easy web-interface. Multi-body docking scripts can be generated with the ATTRACT full web interface (see Section 4.5).

4.4.1. Soft harmonic modes

Global backbone flexibility (e.g. domain-domain motion) can be included by energy minimization along the directions of precalculated soft normal modes for each partner structure [296, 297]. The soft normal modes corresponded to eigenvectors of the protein calculated using an approximate normal-mode analysis related to an elastic network model (ENM) as described by Hinsen [182]. In an elastic network model, the experimental structure serves as a reference (energy minimum) structure and the mobility of a residue or protein segment depends on the local density and number of short range contacts. The calculation of modes based on ENMs is computationally inexpensive and global protein flexibility can be included in the docking process at moderate additional computational cost. The normal modes are calculated with respect to the C_α atoms of the protein based on a pair-wise dependent energy-function

$$V(\mathbf{R}_1, \ldots, \mathbf{R}_N) = \sum_{i,j} k_{ij} \left(|\mathbf{R}_{ij}| - |\mathbf{R}^0_{ij}| \right)^2 (\mathbf{R}_i - \mathbf{R}_j)$$

where $|\mathbf{R}^0_{ij}| = |\mathbf{R}^0_i - \mathbf{R}^0_j|$ is the equilibrium distance of the C_α pair derived from the experimental/input structure and the distance-dependent force constant k_{ij} is given by

$$k_{ij}(|\mathbf{R}_{ij}|) = \lambda \exp\left(-\frac{|\mathbf{R}_{ij}|^2}{\sigma^2} \right).$$

The elastic network model hence has the parameters λ (coupling strength) and σ (coupling cutoff) which are set by default to 1.2 and 5 Å in ATTRACT (ATTRACT tools `modes.py` and `modes-aa.py`). Harmonic modes with respect to the above energy function can be obtained by diagonalization of the second derivative matrix of the energy function [182]. Each residue is treated as a rigid-body during normal mode deformations. ATTRACT typically uses deformations in the lowest 5 normal modes as additional degrees of freedoms (in addition to center-of-mass rotation and translation). The normal-mode deformations mimic an induced-fit process upon binding. Normal modes can also be calculated for nucleic acids (Setny, unpublished data).

4.4.2. Ensemble docking

For modeling a conformational selection-type binding process, a set of multiple rigid conformations (multi-model PDB file) can be supplied for each partner allowing to choose the most likely conformation during docking (ensemble docking). Recently,

the ensemble docking approach in ATTRACT was expanded towards a replica exchange docking scheme where ensemble conformations can be exchanged based on a Metropolis criterion (Zhang, unpublished data).

4.4.3. Multi-body docking and domain docking

The ATTRACT program can dock up to 100 biomolecules simultaneously. In contrast, the widely-used HADDOCK docking program is limited to a maximum of six individual docking partners [220]. This multi-body docking property can also be used to dock individual protein domains in order to represent domain-domain motions, an approach that was successfully employed earlier by Karaca et al. [218]. The fragment-docking approach for modeling protein-RNA complexes [67] also makes extensive use of this feature.

4.5. ATTRACT web interfaces

Many docking programs can now be easily accessed via web interfaces and servers [83, 401, 446, 285, 98, 287, 444, 347]. With the large user base acquired by many of these web services, docking has been established as a valuable tool for the structural biology community. Docking web services further offer the possibility to systematically compare published methods and hence promote reproducible research.

The ATTRACT docking engine can tackle a large variety of docking problems, due to an extensive set of features and options. ATTRACT is implemented as a suite of command line tools and options that can be combined at will. Therefore, ATTRACT is typically invoked via a custom, hand-written shell script. While this approach is very flexible, it limits the accessibility of ATTRACT to expert users only. However, with the Spyder framework (used in similar web servers [98, 95]), it is possible to generate docking protocols automatically, based on a set of parameters that can be edited in a web browser.

We developed five web interfaces for setting up docking runs in ATTRACT: the ATTRACT Easy web interface, the pepATTRACT web interface, the ATTRACT standard web interface, the ATTRACT full web interface and the upload web interface. The protocols that can be generated allow for many different docking applications and expose a large fraction of the functionality present in ATTRACT. The ATTRACT web interfaces are not web servers: they return docking protocols (shell scripts) for execution on a local machine (the user has to install the ATTRACT program). Alternatively, these docking scripts can be submitted to the servers at the Mobyle platform (de Vries, personal communication).

4.5.1. ATTRACT Easy web interface

The ATTRACT Easy web interface (www.attract.ph.tum.de) provides a convenient way to set up an ab-initio two-body protein-protein docking protocol and hence a user-friendly, general-purpose entry point for protein-protein docking with ATTRACT. On the one hand, it is sufficient to provide just a PDB file for both protein partners. On the other hand, a number of options are available (but not required) to customize the protocol. For example, the web interface offers several possibilities to include protein flexibility in the docking search. If an induced fit model of binding is hypothesized, the "harmonic modes" option can be enabled, selecting collective modes that will be calculated from an elastic network model [296]. The protein will then be deformed along these modes during the docking. Alternatively, an ensemble of multiple rigid conformations can be provided as a multi-model PDB file, allowing the most likely conformation to be selected during the docking. The initial rigid-body docking search may be followed by a flexible refinement using the iATTRACT protocol [393] (Chapter 5). Finally, for benchmarking purposes, the docking results can be assessed against a user-supplied reference structure with the same statistics as used in CAPRI (IRMSD, LRMSD and fnat).

4.5.2. pepATTRACT web interface

Peptide-protein docking protocols [391] (Chapter 6) can be easily generated with the pepATTRACT web interface (www.attract.ph.tum.de/peptide.html). The web interface helps the user set up a script that performs the rigid-body sampling stage and the flexible interface refinement starting from the structure of the unbound protein and the peptide sequence. It also provides the option to specify residues for ambiguous interaction restraints [322, 112] to include experimental information and restrict the search for the peptide binding site to a portion of the protein's surface. Furthermore, conformational change on the protein side can be included by providing multiple protein structures (ensemble docking). This option is also useful if the protein structure has been derived from template-based homology modeling or derived from NMR experiments. Snapshots from MD simulations could also be used as an ensemble in docking. Instead of uploading a PDB file containing a single protein structure to the web interface, the user can upload a multi-model PDB file and specify the number of conformers. Like the ATTRACT Easy web interface, the pepATTRACT web interface provides benchmarking options. The usage of the pepATTRACT web interface is illustrated in Figure 4.5. The docking script generated by the web interface provides an easy entry point for non-expert users into fast peptide-protein docking in ATTRACT.

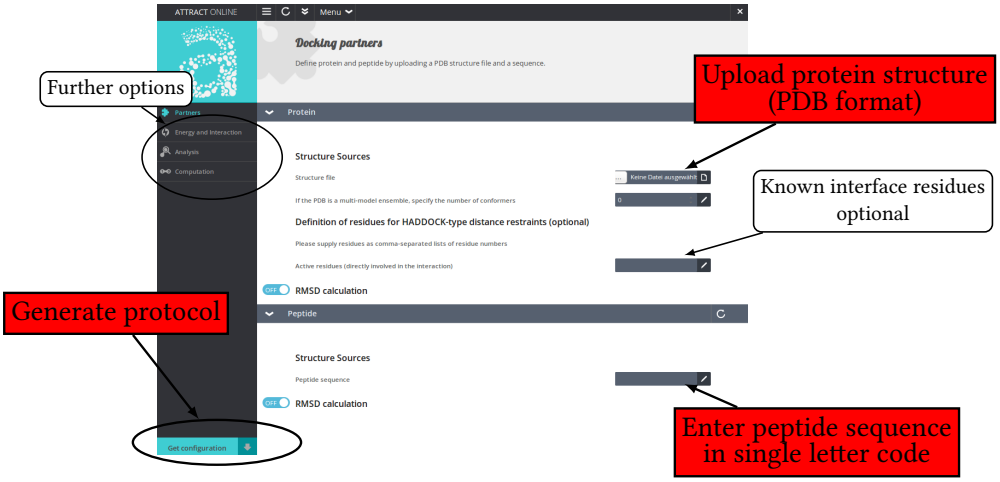


Figure 4.5. Instructions for generating a peptide-protein docking protocol with the pepATTRACT web interface. The web interface is available at www.attract.ph.tum.de/peptide.html. Required input is highlighted in red.

4.5.3. ATTRACT standard web interface

The ATTRACT standard web interface (www.attract.ph.tum.de/services/ATTRACT/standard.html) is an extension of the ATTRACT Easy web interface [99]. It can be used to set up two-body protein-protein, peptide-protein, protein-DNA and protein-RNA docking scripts for ATTRACT. In addition to the options available in the Easy web interface, the standard web interface supports including experimental information on interface residues via specifying active and passive residues for ambiguous interaction restraints [322, 112] for both docking partners. Furthermore, information on contacts can be supplied in CNS/HADDOCK tbl format in order to generate distance restraints during docking ("Harmonic distance restraints file", section "Sampling"). Instead of using the ATTRACT coarse-grained force field, the OPLS force field can be selected for rigid-body docking (in general, this is not recommended). The standard web interface can be used for advanced peptide-protein docking with a peptide ensemble specified by the user.

4.5.4. ATTRACT full web interface

The ATTRACT full web interface exposes the majority of the functionality present in ATTRACT and is suitable for advanced users. It is available at www.attract.ph.tum.de/services/ATTRACT/full.html. Most importantly, the full web interface allows the

user to set up multi-body docking scripts (currently up to 10 docking partners). In contrast to the Easy and the standard web interface, users have to provide more input parameters. In general, it is therefore recommended to set up a docking script with one of the other web interfaces and only modify certain parameters with the full web interface. The full web interface offers a large variety of options to customize the docking script. It allows to design the sampling process by specifying the number of sampling iterations and settings for each iteration (section "Iterations"). The user can change the number of minimization steps ($v_{\max}$) and also enable Monte Carlo sampling instead of energy minimization. In the "Sampling" section, the user can choose different setups for generating the initial starting positions ("Docking search") and modify parameters for iATTRACT refinement. Parameters for experimental restraints (e.g., ambiguous interaction restraints) can be customized in the "Energy and Interaction" section. For sampling without a physiochemical force field (only soft repulsion between the docking partners), atom density grids can be specified. The full web interface also supports symmetry restraints (e.g., for docking of homo-multimers, cf. Section 2.3.2).

4.5.5. The upload web interface

The upload web interface can be used to upload previously generated embedded parameter files and modify them. It can also be used to convert parameter models (e.g., convert an ATTRACT Easy model to a full ATTRACT model). The upload web interface is available at www.attract.ph.tum.de/services/ATTRACT/upload.html.

4.6. Conclusion and Outlook

The ATTRACT docking engine can model a wide range of biomolecular interactions. The program has evolved over the years and many new functions and protocols have been added. ATTRACT can now harness the computing capabilities of GPUs to further accelerate docking calculations [121]. We will integrate the GPU-accelerated version into the main branch of ATTRACT and add support for modified amino acids, ions, and cofactors during docking. Then a new ATTRACT version (ATTRACT 2.0) will be released (scheduled December 2016). The release will contain all published docking protocols [393, 391, 99], new tools for developing scoring functions [389], example scripts [96, 392, 67] and web-interfaces to make ATTRACT docking calculations available to the scientific community. In the future, this GPU-accelerated ATTRACT version can be used to offer docking computation services (web servers).

5. iATTRACT: Simultaneous Global and Local Optimization for Protein-Protein Docking Refinement

A major caveat for docking success is accounting for protein flexibility. Especially, interface residues undergo significant conformational changes upon binding. This limits the performance of methods that keep partner structures rigid or allow only limited flexibility. This chapter describes a new docking refinement approach, interface-ATTRACT (iATTRACT). iATTRACT combines simultaneous full interface flexibility (both backbone and side chain flexibility) and rigid-body optimizations during docking energy minimization. The performance of the approach was systematically assessed on a large protein-protein docking benchmark, starting from an enriched decoy set of rigidly docked protein-protein complexes. Large improvements in sampling and slight but significant improvements in scoring/discrimination of near-native docking solutions were observed. Improvements in the fraction of native contacts were especially favorable, yielding increases of up to 70%. The work presented in this chapter has been previously published [393].

5.1. Introduction

Protein-protein interactions (PPIs) play an essential role in most biological events within a cell. The interacting macromolecules carry out cellular processes such as DNA synthesis, gene expression, signal transduction, and immune responses. At the same time, aberrant PPIs are at the heart of pathological processes, such as Alzheimer's disease and cancer. Atomic-level structural knowledge of protein-protein complexes is vital for understanding their biological function, for predicting the effect of mutations [233] and for rationally designing new PPI-based therapeutic agents [483]. For the latter, it is particularly important to determine which protein residues form the physical contacts between the macromolecules, i.e., to identify the native residue contacts.

Experimental structure determination methods based on nuclear magnetic resonance (NMR) spectroscopy or X-ray crystallography have been used successfully to characterize many proteins and also protein-protein complexes. However, the number of possible PPIs far exceeds the number of known proteins [429] and experimental characterization of all PPIs is currently not feasible. Computational prediction methods like protein-protein docking aim to use available experimental or homology-modeled structures of individual constituents to predict the structure of protein-protein complexes. Hence, docking can complement experimental information and is an extremely valuable tool for structural biologists.

Protein-protein docking methods perform satisfactorily, especially in cases with little or no structural changes in the protein-protein interface upon complex formation. Progress in the protein-protein docking field in recent years has been monitored and documented by the community-wide blind docking challenge CAPRI (Critical Assessment of Predicted Interactions) [257, 259, 260]. Still, predicting the correct three-dimensional structure of protein complexes and predicting the native residue contacts remain an enormous challenge. Solutions of acceptable or even medium quality according to standards used in the CAPRI challenge often only retrieve one third or less of the interacting residues. This is insufficient for using such complexes, for example in drug design applications. Major limitations of current docking algorithms are the rigid-body assumption as well as simplistic and inaccurate force fields [51, 12, 502, 302, 192]. Typically, docking algorithms perform broad sampling of protein complex conformations while keeping the internal structure of the protein partners rigid. However, most proteins undergo conformational change upon binding and thus, sampling of the bound conformation is often not possible using rigid unbound protein partners. Additionally, this sampling failure impairs the scoring to distinguish near-native candidate structures with correct residue contacts from decoys which display few, if any, correct physical contacts. To tackle these problems, refinement protocols optimize a subset of rigid-body candidate structures using a computationally more demanding, detailed energy function and including a greater portion of protein flexibility. Ideally, refinement protocols should establish more native contacts between the protein partners, increase interface complementarity and yield improved structural agreement with respect to the native complex structure. In addition, the energetic evaluation or scoring of refined complexes should increase the specificity for native or native-like docking geometries.

A variety of different docking refinement approaches have been developed to date. In the ATTRACT docking approach, low-frequency normal modes of partner proteins can be employed during systematic search and refinement stage to account for global changes [296, 297]. RosettaDock's refinement stage uses successive rigid-body moves, backbone perturbations, rotamer selection and full side chain repacking in a

Monte Carlo minimization scheme [65, 66]. The FiberDock method models backbone flexibility through a combination of normal mode deformations and includes side chain flexibility by selecting appropriate states from a rotamer library. Side chain repacking and backbone remodeling are employed iteratively with additional Monte Carlo minimization in rigid-body space in each cycle [293]. The EigenHex algorithm also uses a combination of simple elastic network normal modes but individually selects them for each pose [466]. SwarmDock employs up to 40 normal modes derived from an all-atom elastic network model to represent backbone conformational change with a focus on the interface region [310]; the authors propose to combine this with side chain rebuilding using SCWRL [239]. In ICM-DISCO, the refinement focuses on the interface side chains which are treated as fully flexible [127]. The HADDOCK program uses a semi-flexible refinement stage in torsion angle space of both interface backbone and side chains followed by a full Cartesian dynamics refinement in explicit solvent comparable to a molecular dynamics simulation [97]. Most of these algorithms perform well with regards to optimizing interface residue conformations and resolving steric clashes in order to improve the scoring. However, they are less successful for improving the sampling of the overall protein complex geometry and moving the protein partners significantly from the starting geometry.

Here, we present a new approach, iATTRACT (interface-ATTRACT) that improves sampling in the refinement stage by combining full atomistic flexibility of interface residues with global translational and orientational motions of the protein partners. Hence, conformational space is simultaneously explored in global and local degrees of freedom in a potential energy minimization. To control conformational changes at the interface, we employ a structure-based potential with the unbound protein structure as reference, whereas a classical molecular mechanics force field is used to model the intermolecular physical interactions. We demonstrate the efficiency of the approach by refining rigid-body docking candidates of a large number of protein complexes.

5.2. Methods

The iATTRACT method is an all-atom refinement procedure that includes full Cartesian flexibility of the interface residues and combines this with rigid-body optimization in an all-at-once approach. It consists of an interface selection step, the generation of an intramolecular structure-based protein force field for interface residues, followed by energy minimization in overall translation, rotation and local atom movement at the interface.

5.2.1. Interface selection

The interface was selected for each initial docking start geometry; i.e., the residues which were in contact in the input structure were treated as flexible during the refinement. Both side chain and backbone atoms were taken into account. The interface residues were selected automatically by accounting for contacts between the protein partners of the input structure within a cutoff range of $d_{\text{cut}} = 3.0\,\text{Å}$. If necessary, the cutoff was increased to ensure that each protein partner had at least 8 flexible interface residues. On the other hand, if the total number of flexible atoms N_{flex} was larger than 333, the cutoff was decreased successively by steps of $0.5\,\text{Å}$ and the number of selected interface residues was decreased accordingly. This adjustable scheme accounts for the different sizes of the protein partners and the respective interface, and the varying quality of the candidate structures. The maximum total number of flexible atoms was 333, which is only a small fraction of the atoms in a typical protein complex.

5.2.2. Intra-protein force field

A structure-based force field for each unbound protein structure was generated and applied to the flexible interface atoms throughout the refinement process. This structure-based force field contains harmonic potentials for controlling bond lengths and bond angles as well as a double-quadratic potential to represent steric repulsion between non-bonded atoms. The force field representation allows motions that result in changes of dihedral torsion angles and does not include any attractive non-bonded interactions within the protein.

$$V_{\text{rest}} = \underbrace{\sum_{i=1}^{N} \tfrac{1}{2} k_r \left(r_{ii+1} - r^0_{ii+1}\right)^2}_{\text{bonds}} + \underbrace{\sum_{i=1}^{N} \tfrac{1}{2} k_\theta \left(r_{ii+2} - r^0_{ii+2}\right)^2}_{\text{angles}} + \underbrace{\sum_{i,j} \tfrac{1}{2} k_{\text{rep}} \left(r^2_{ij} - r^2_{\min}\right)^2}_{\text{excluded volume}} \quad \text{for } r_{ij} < r_{\min}$$

The parameters r^0_{ii+1} for bond length and r^0_{ii+2} for bond angles were extracted directly from the unbound protein structure. The parameters k_r, k_θ, k_{rep} were set to $1000\,\text{kcal/mol/Å}^2$, $100\,\text{kcal/mol/Å}^2$ and $1\,\text{kcal/mol/Å}^4$ respectively. This allows for distance fluctuations in the order of $10^{-2}\,\text{Å}$ for nearest neighbors and $10^{-1}\,\text{Å}$ for second nearest neighbors at room temperature. $r_{\min}$ was set to $3.5\,\text{Å}$ which corresponds roughly to twice the van der Waals radius of a carbon atom and thus approx-

imately represents the steric exclusion between heavy atoms. The form of the soft repulsive potential allows for partial atom-atom overlap (of atoms within one protein partner but not between atoms of separate partners, see below). Hydrogen atoms were not subjected to steric repulsion. Different parameters for k_r and k_θ were tested and the model showed only slight dependence on the values of the force constants (data not shown). For each atom, the bonds to its nearest covalently bound neighbors, the angles to its second nearest covalently bound neighbors and the steric repulsion to all neighboring atoms within a cutoff of $5\,\text{Å}$ were considered. If the number of neighboring atoms was less than 30, the cutoff was increased and neighbors were determined within the increased cutoff range to account for steric repulsion. During refinement, atoms perform only small-scale rearrangements and their movement is restricted by the presence of the neighboring atoms. If side chains are less densely packed, they are more likely to rearrange on a larger scale, this is implicitly taken into account by the variable neighbor list. Calculating non-bonded interactions within a cutoff range balances computational load with a physically accurate representation.

5.2.3. Combining full interface flexibility and global translational and orientational motion

We use Cartesian degrees of freedoms for the flexible interface atoms and calculate the displacements in a reference frame attached to the protein's center of geometry; i.e., relative to the template structure of the protein. The force on a single atom is the sum of the respective forces resulting from both the inter-protein force field and the intra-protein restraints and the displacements are directly derived from these forces. This flexibility mechanism, which we will also refer to as imodes (interface modes), was implemented as a separate option in ATTRACT. An efficient variable metric minimizer [296, 297] then allows simultaneous large-scale translational and rotational optimization (of the smaller ligand protein partner) combined with small-scale, local, relative movements of interface atoms. A single energy minimization step consists of the following parts. First, the pairwise non-bonded interactions and their derivatives between all the atoms from the protein partners are calculated in the global reference frame. These forces are rotated in the local protein-associated coordinate frame. Then the penalty potentials of the intra-protein force field and their derivatives for the flexible interface residues are computed in the protein reference frame and these forces are added to the atom forces derived from the protein-protein interaction. Local deformations of the atoms were derived directly from the atomistic forces. The total force and the torque are applied to the ligand protein's center of geometry position and orientation. For each structure, 2500 minimization steps were performed.

5.2.4. Evaluation of docking solutions

The quality of the refinement solutions was assessed by interface root mean square deviation (IRMSD) and fraction of native contacts (fnat, see Chapter 3). Stars were awarded in a fashion similar to the CAPRI criteria [303] to grade the quality of the predictions. The criteria are summarized in Table 5.1. We refer to structures of CAPRI one star quality or better as (one star) quality structures. To investigate the contribution of the interface flexibility on the sampling, we removed the interface mode deformations by superimposing the unbound protein structures on the generated iATTRACT models. Then, we compared IRMSD and fnat with and without interface residue movements. To compare the similarity of the iATTRACT models to the bound complex structure at the interface, we superimposed the individual protein structures of the iATTRACT models onto the bound structures and calculated a heavy atom $\mathrm{IRMSD}_{\mathrm{heavy}}$ with backbone and side chain atoms including all residues which were within a cutoff of $3.0\,\mathrm{\AA}$ from the protein partner in the bound crystal structure. The same was done for the unbound protein structures superimposed on the bound complex structure. These values were compared to check whether the proteins sampled conformations closer to the bound form. As a score we used the intermolecular energies of the complexes derived from the OPLS parameters. To see how the generated structures score with respect to their IRMSD, we clustered the structures by pair-wise backbone ligand-RMSD with a cutoff of $7.0\,\mathrm{\AA}$ and a minimum cluster size of 4 and calculated the percentage of models of a certain CAPRI quality in the top N clusters. Within each cluster, the cluster members were sorted by their intermolecular energy. The clusters were then ranked by the average of the intermolecular energies of the top 4 members. A cluster is designated as of one star CAPRI quality if any of the top 4 members is of one star CAPRI quality. Finally, to fix possible deformations in the side chain structure due to the simplistic nature of the intra-protein force field, each refined solution was subjected to an energy minimization (5500 steps) with the Amber 12 program [61] using the parm10 force field [188] with an implicit Generalized Born solvation model. Significance of observed improvements was tested by calculating the p-value using the Wilcoxon signed-rank test.

5.2.5. Dataset and structure preparation

Refinement was performed on 166 protein complexes from benchmark 4.0 [200]. Docking cases were classified according to RMSD between bound and unbound forms as “rigid-body”, “medium” and “hard” cases. Note that even the “rigid-body” cases can involve side chain rearrangements upon complex formation and for the “medium”

Table 5.1. CAPRI quality measure used to evaluate the refined docking models.

Quality	IRMSD[Å]	fnat
High accuracy ($\star\star\star$)	$\leqq 1.0$	$\geqq 0.5$
Medium accuracy ($\star\star$)	$\leqq 2.0$	$\geqq 0.3$
Acceptable ($\star$)	≤ 4.0	$\geqq 0.1$ and < 0.3

and "hard" cases typically both backbone as well as side chains differ significantly in unbound and bound partner structures. A few cases where the protein partners undergo large global conformational changes were excluded from the dataset (PDB accession codes: 2NZ8, 1F6M, 1DE4, 1FAK, 1H1V, 1IRA, 1Y64). Additionally, the homodimer/homotetramer 1N2C was excluded since RMSD evaluation on solutions was considered to be too complex. The final dataset contained 119 rigid-body, 28 medium and 19 hard docking cases.

The protein structures were downloaded from the PDB. If necessary residues were renumbered in the unbound structures to match the bound forms, parts in the unbound form that are not present in the bound form were removed (and vice versa), and point mutations were introduced to resolve minor differences in the protein sequences. The structures were then converted into the OPLS atom type description with the ATTRACT tool `aa-reduce`. Missing hydrogens were built with PDB2PQR [111, 110] and protonation states were determined by PropKa [266]. For histidine protonation states the bound structure was used as a reference to ensure that unbound and bound structures contained the same atoms. Disulfide bridges were also determined according to the bound structure based on a cutoff criterion. This was necessary for easy evaluation of the refinement results against the bound crystal structure.

5.2.6. Rigid docking for start complex generation and inter-protein force field

The refinement input structures were generated by first performing an atomistic ab-initio rigid-body docking with ATTRACT employing the OPLS force field. The intermolecular energy function consisted of pair-wise van der Waals energy terms and full electrostatics within a $50\,\text{Å}$ distance cutoff (see Chapter 4). The parameters for these non-bonded interactions are OPLS non-bonded parameters [212] derived from a modified version of the parallhdg5.3.pro parameter file [271]. The dielectric con-

stant ϵ was set to 10. The evaluation of the interaction potentials was accelerated using a pre-calculated grid [100]. The receptor protein was kept fixed. The energy minimization utilizes standard ATTRACT routines as described previously [296, 297]. For each starting structure, 1000 minimization steps were performed. The best 200 structures (ranked by interface root mean square deviation) were subjected to a reminimization in rigid-body space with exact energy calculation using OPLS parameters within a $50\,\text{Å}$ distance cutoff. 2500 minimization steps were used. Convergence was assured by reminimizing all structures again, which yielded no further change (data not shown). After ranking again by interface root mean square deviation (IRMSD), the top 200 structures for each case were selected as input for the refinement procedure. This resulted in a test set of 33,200 structures. Of these, less than 10% (3,092) were CAPRI one star quality structures.

5.3. Results

The new refinement method iATTRACT, (interface-ATTRACT), aims to enhance sampling of protein complex geometries by combining interface residue movements with rigid-body optimizations of the protein partners. A flow chart of the methodology is shown in Figure 5.1. Initially, flexible interface residues were selected on both protein partners from the contacts present in the input structure. A structure-based force field was constructed to control the induced deformations of the flexible interface region. This force field is optimally adapted to each individual protein partner and prevents deviations from the experimental structure which often occur in molecular mechanics force fields upon optimization. Subsequently, for each starting docking geometry, a simultaneous energy minimization in rotational and orientational degrees of freedom of the ligand and the Cartesian degrees of freedom associated with the flexible residues was performed using the ATTRACT minimization routine [296, 297]. The electrostatic and van der Waals interactions between the protein partners were calculated with parameters from a molecular mechanics force field [212, 271].

5.3.1. Overall performance

The iATTRACT flexible docking refinement approach was systematically evaluated on a large benchmark set of known protein-protein complexes [200]. We wanted to compare the sampling capacity of iATTRACT to that of ATTRACT rigid-body docking. Since it was not the aim of this study to investigate the scoring of rigid-body candidates, we chose the input structures by IRMSD rank which yielded an enriched decoy pool. This allows us to study the effects of interface flexibility in the refinement process of structures within a range of initial deviations of approximately

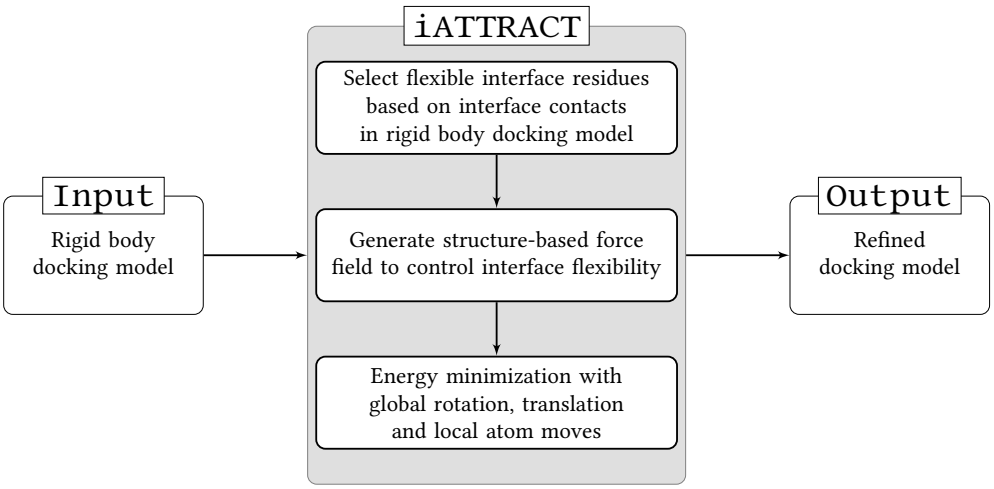


Figure 5.1. The iATTRACT refinement method. iATTRACT aims to enhance sampling of protein complex geometries by combining interface residue movements with rigid-body optimizations of the protein partners.

15 Å IRMSD. This corresponds to the common scenario (in practice) that a binding region on two protein partners is approximately known; e.g., based on additional experimental data.

We first analyzed the combined effect of sampling and scoring improvements during iATTRACT refinement in a blind prediction manner. We ranked the structures before and after refinement by their intermolecular energies based on the OPLS parameters. Overall, refinement improved the success rate for one star quality structures in the top 5 yielding 10% more benchmark cases with at least one acceptable structure within the top 5 ranked structures (Figure A.4). A solution with lower IRMSD in the top 5 comparing refined and unrefined docked structures was found for 65% of the benchmark cases (Table A.2). This represents a significant improvement ($p < 0.001$). However, no overall improvement was found for the top ranked structure (Table A.2). To smoothen the fluctuations of individual energy terms in the atomistic force field scoring function, we clustered the solutions and ranked the clusters by the average energy of the top 4 ranking members. The results are displayed in Figure 5.2. After refinement, we obtained a top ranking near-native cluster in 40% of the 154 successful docking cases and the top 5 clusters contain a near-native cluster for 81% of the successful docking cases. Also for two and three star clusters, there is an improvement compared to the cluster ranking before refinement.

Table 5.2. Docking results for flexible refinement. The results for the initial structures (rigid-body docking) are shown in brackets. For each docking case we report the IRMSD and fnat values for the best sampled structure ranked by IRMSD.

	All Structures			Best structure				All Structures			Best structure		
PDB	⋆ ⋆ ⋆	⋆⋆	⋆	CAPRI	IRMSD	fnat	PDB	⋆ ⋆ ⋆	⋆⋆	⋆	CAPRI	IRMSD	fnat
1A2K	0 (0)	0 (0)	22 (11)	* (*)	2.1 (2.7)	0.74 (0.38)	1NCA	3 (1)	0 (0)	16 (9)	*** (***)	0.5 (0.4)	0.85 (0.75)
1ACB	0 (0)	0 (0)	17 (8)	* (*)	2.5 (3.0)	0.46 (0.19)	1NSN	0 (0)	6 (2)	47 (28)	** (**)	1.3 (1.5)	0.75 (0.38)
1AHW	1 (0)	23 (19)	37 (39)	*** (**)	1.0 (1.4)	0.75 (0.55)	1NW9	0 (0)	0 (0)	16 (13)	* (*)	2.7 (3.0)	0.61 (0.39)
1AK4	0 (0)	0 (0)	10 (8)	* (*)	3.0 (3.3)	0.63 (0.39)	1OC0	0 (0)	0 (2)	60 (38)	* (**)	2.1 (1.6)	0.64 (0.45)
1AKJ	0 (0)	0 (0)	18 (15)	* (*)	2.3 (2.4)	0.58 (0.32)	1OFU	0 (0)	0 (0)	7 (5)	* (*)	2.5 (3.3)	0.56 (0.24)
1ATN	0 (0)	0 (0)	3 (0)	* (-)	3.6 (4.1)	0.45 (0.27)	1OPH	0 (0)	0 (0)	16 (5)	* (*)	2.5 (3.4)	0.84 (0.45)
1AVX	0 (0)	4 (1)	23 (7)	** (**)	1.6 (1.8)	0.76 (0.51)	1OYV	0 (0)	0 (0)	39 (14)	* (*)	2.1 (2.5)	0.52 (0.31)
1AY7	0 (0)	0 (0)	4 (1)	* (*)	2.3 (3.2)	0.64 (0.36)	1PPE	5 (0)	4 (1)	73 (43)	*** (**)	0.6 (2.0)	0.94 (0.52)
1AZS	0 (0)	8 (0)	40 (36)	** (*)	1.4 (2.6)	0.88 (0.48)	1PVH	1 (0)	0 (1)	18 (23)	*** (**)	1.0 (1.6)	0.97 (0.57)
1B6C	0 (0)	0 (0)	11 (4)	* (*)	2.3 (3.5)	0.56 (0.25)	1PXV	0 (0)	0 (0)	8 (0)	* (-)	3.6 (4.4)	0.42 (0.14)
1BGX	0 (0)	0 (0)	0 (0)	- (-)	4.9 (5.6)	0.13 (0.11)	1QA9	0 (0)	19 (15)	50 (34)	** (**)	1.5 (1.6)	0.79 (0.68)
1BJ1	2 (0)	2 (0)	14 (10)	*** (*)	0.8 (2.2)	0.95 (0.34)	1QFW	0 (7)	15 (0)	38 (33)	** (***)	1.1 (0.8)	0.92 (0.69)
1BKD	0 (0)	0 (0)	0 (0)	- (-)	4.5 (5.4)	0.19 (0.12)	1R0R	0 (0)	1 (1)	31 (18)	** (**)	1.8 (1.7)	0.67 (0.31)
1BUH	4 (0)	15 (16)	34 (20)	*** (**)	0.8 (1.6)	0.79 (0.49)	1R6Q	0 (0)	0 (0)	32 (21)	* (*)	2.1 (2.1)	0.66 (0.42)
1BVK	0 (0)	0 (0)	27 (24)	* (*)	2.3 (2.4)	0.59 (0.38)	1R8S	0 (0)	0 (0)	0 (0)	- (-)	5.0 (5.9)	0.18 (0.10)
1BVN	0 (0)	0 (0)	26 (6)	* (*)	2.4 (3.0)	0.47 (0.22)	1RLB	0 (0)	0 (0)	21 (19)	* (*)	2.2 (2.7)	0.62 (0.42)
1CGI	0 (0)	3 (0)	18 (8)	** (*)	2.0 (3.3)	0.59 (0.25)	1RV6	0 (0)	2 (2)	15 (13)	** (**)	1.8 (1.6)	0.69 (0.48)
1CLV	0 (0)	0 (0)	32 (9)	* (*)	2.9 (3.0)	0.35 (0.20)	1S1Q	0 (0)	6 (5)	40 (19)	** (**)	1.3 (1.4)	0.76 (0.57)
1D6R	0 (0)	2 (0)	41 (26)	** (*)	1.3 (2.2)	0.64 (0.33)	1SBB	0 (0)	6 (0)	6 (8)	** (*)	1.1 (2.2)	0.82 (0.45)
1DFJ	0 (0)	0 (0)	20 (0)	* (-)	2.2 (4.2)	0.47 (0.19)	1SYX	0 (0)	0 (0)	70 (46)	* (*)	2.3 (2.2)	0.88 (0.51)
1DQJ	0 (0)	3 (0)	23 (15)	** (*)	1.4 (2.3)	0.64 (0.45)	1T6B	0 (0)	0 (0)	8 (5)	* (*)	3.0 (2.6)	0.38 (0.28)
1E4K	0 (0)	0 (0)	0 (0)	- (-)	4.1 (4.4)	0.43 (0.25)	1TMQ	0 (0)	0 (0)	0 (0)	- (-)	6.0 (6.3)	0.14 (0.12)
1E6E	0 (0)	22 (14)	61 (39)	** (**)	1.1 (1.5)	0.89 (0.65)	1UDI	0 (0)	1 (0)	27 (10)	** (*)	1.5 (3.2)	0.68 (0.19)
1E6J	0 (0)	3 (0)	26 (21)	** (*)	1.3 (2.1)	0.87 (0.45)	1US7	0 (0)	2 (0)	40 (31)	** (*)	1.6 (2.5)	0.82 (0.64)
1E96	0 (0)	6 (5)	33 (23)	** (**)	1.3 (1.4)	0.87 (0.58)	1VFB	0 (0)	4 (2)	16 (13)	** (**)	1.5 (1.4)	0.81 (0.55)
1EAW	0 (0)	4 (0)	41 (11)	** (*)	1.5 (2.1)	0.69 (0.46)	1WDW	0 (0)	9 (0)	15 (7)	** (*)	1.7 (2.5)	0.62 (0.34)
1EER	0 (0)	0 (0)	8 (1)	* (*)	3.0 (3.8)	0.44 (0.20)	1WEJ	6 (4)	3 (3)	22 (18)	*** (***)	0.8 (1.0)	0.91 (0.65)
1EFN	0 (0)	4 (0)	60 (42)	** (*)	1.8 (2.1)	0.81 (0.49)	1WQ1	0 (0)	0 (0)	16 (5)	* (*)	2.5 (2.8)	0.36 (0.27)
1EWY	0 (0)	5 (6)	25 (16)	** (**)	1.7 (1.5)	0.75 (0.59)	1XD3	0 (0)	0 (0)	73 (69)	* (*)	2.1 (2.5)	0.68 (0.49)
1EZU	0 (0)	0 (0)	0 (0)	- (-)	4.8 (5.6)	0.17 (0.09)	1XQS	0 (0)	0 (0)	55 (25)	* (*)	2.0 (2.9)	0.72 (0.37)
1F34	0 (0)	0 (0)	0 (0)	- (-)	8.4 (9.4)	0.05 (0.05)	1XU1	0 (0)	1 (0)	19 (13)	** (*)	1.8 (2.8)	0.64 (0.39)
1F51	0 (0)	0 (0)	3 (1)	* (*)	2.2 (2.3)	0.65 (0.49)	1YVB	0 (3)	20 (24)	46 (31)	** (***)	1.1 (0.9)	0.86 (0.59)
1FC2	0 (0)	0 (0)	11 (4)	* (*)	2.7 (3.0)	0.65 (0.29)	1Z0K	2 (0)	3 (7)	40 (42)	*** (**)	1.0 (1.1)	0.88 (0.60)
1FCC	0 (0)	0 (0)	18 (15)	* (*)	2.0 (2.5)	0.69 (0.53)	1Z5Y	0 (0)	0 (0)	40 (19)	* (*)	2.1 (2.7)	0.59 (0.34)
1FFW	0 (0)	0 (0)	88 (82)	* (*)	2.0 (2.2)	0.79 (0.45)	1ZHH	0 (0)	0 (0)	1 (1)	* (*)	3.8 (3.9)	0.49 (0.29)
1FLE	0 (0)	0 (0)	40 (16)	* (*)	2.6 (3.2)	0.60 (0.23)	1ZHI	0 (0)	1 (7)	21 (17)	** (**)	2.0 (2.0)	0.68 (0.60)
1FQ1	0 (0)	0 (0)	1 (1)	* (*)	3.5 (3.6)	0.46 (0.26)	1ZLI	0 (0)	0 (0)	4 (0)	* (-)	3.8 (4.3)	0.46 (0.32)
1FQJ	0 (0)	1 (0)	30 (20)	** (*)	1.4 (2.3)	0.63 (0.44)	1ZM4	0 (0)	0 (0)	39 (21)	* (*)	2.5 (2.7)	0.60 (0.60)
1FSK	0 (2)	4 (2)	17 (7)	** (***)	1.3 (0.8)	0.90 (0.77)	2A5T	0 (0)	0 (0)	13 (16)	* (*)	2.9 (3.3)	0.39 (0.25)
1GCQ	0 (2)	9 (2)	33 (23)	** (***)	1.0 (0.9)	0.96 (0.67)	2A9K	0 (0)	1 (1)	14 (7)	** (**)	1.9 (1.7)	0.70 (0.30)
1GHQ	0 (3)	0 (0)	35 (28)	* (***)	2.7 (0.9)	0.82 (0.71)	2ABZ	0 (0)	4 (0)	20 (12)	** (*)	1.4 (2.2)	0.64 (0.30)
1GL1	0 (0)	0 (0)	58 (25)	* (*)	2.4 (2.5)	0.62 (0.38)	2AJF	0 (0)	2 (0)	21 (16)	** (*)	1.9 (2.3)	0.58 (0.36)
1GLA	0 (0)	1 (0)	12 (9)	** (*)	1.9 (1.8)	0.56 (0.30)	2AYO	0 (0)	3 (0)	34 (20)	** (*)	1.9 (2.6)	0.46 (0.30)
1GP2	0 (0)	0 (0)	24 (13)	* (*)	2.1 (2.5)	0.37 (0.27)	2B42	0 (0)	0 (0)	0 (0)	- (-)	4.1 (4.1)	0.26 (0.17)
1GPW	0 (0)	6 (0)	46 (24)	** (*)	1.1 (2.0)	0.78 (0.35)	2B4J	0 (0)	0 (0)	19 (10)	* (*)	2.9 (2.5)	0.68 (0.37)
1GRN	0 (0)	1 (6)	48 (31)	** (**)	1.8 (1.6)	0.68 (0.47)	2BTF	0 (0)	2 (0)	33 (17)	** (*)	1.1 (2.7)	0.69 (0.36)
1GXD	0 (0)	2 (0)	38 (31)	** (*)	1.7 (2.5)	0.67 (0.39)	2C0L	0 (0)	0 (0)	10 (11)	* (*)	3.0 (3.0)	0.57 (0.28)
1H9D	0 (0)	0 (0)	8 (2)	* (*)	3.0 (3.5)	0.32 (0.23)	2CFH	0 (0)	2 (0)	20 (9)	** (*)	1.8 (2.0)	0.59 (0.34)
1HCF	0 (0)	1 (0)	43 (31)	** (*)	1.4 (2.8)	0.75 (0.42)	2FD6	0 (0)	1 (0)	13 (10)	** (*)	1.8 (2.1)	0.76 (0.52)
1HE1	0 (0)	0 (0)	37 (19)	* (*)	2.1 (2.8)	0.66 (0.34)	2FJU	0 (0)	1 (1)	33 (30)	** (**)	1.9 (1.8)	0.81 (0.49)
1HE8	0 (0)	2 (2)	12 (8)	** (**)	1.6 (1.1)	0.84 (0.70)	2G77	0 (0)	2 (0)	27 (11)	** (*)	1.9 (2.6)	0.53 (0.28)
1HIA	0 (0)	0 (0)	16 (3)	* (*)	3.1 (3.8)	0.42 (0.22)	2H7V	0 (0)	0 (0)	7 (6)	* (*)	2.8 (2.8)	0.61 (0.33)
1I2M	0 (0)	0 (0)	32 (9)	* (*)	2.7 (2.9)	0.46 (0.26)	2HLE	0 (0)	5 (0)	25 (19)	** (*)	1.7 (2.3)	0.55 (0.32)
1I4D	0 (0)	0 (0)	16 (12)	* (*)	2.1 (2.8)	0.49 (0.30)	2HMI	0 (0)	0 (0)	1 (0)	* (-)	3.1 (4.2)	0.45 (0.25)
1I9R	0 (0)	4 (2)	16 (17)	** (**)	1.6 (1.6)	0.87 (0.85)	2HQS	0 (0)	0 (0)	14 (10)	* (*)	2.3 (2.5)	0.52 (0.37)
1IB1	0 (0)	0 (0)	3 (0)	* (-)	3.6 (5.1)	0.31 (0.26)	2HRK	0 (0)	2 (0)	37 (20)	** (*)	1.7 (2.3)	0.87 (0.44)
1IBR	0 (0)	0 (0)	0 (0)	- (-)	5.3 (6.7)	0.10 (0.08)	2I25	0 (0)	0 (0)	15 (3)	* (*)	2.1 (3.7)	0.53 (0.23)
1IJK	0 (0)	7 (0)	41 (45)	** (*)	1.4 (2.2)	0.83 (0.47)	2I9B	0 (0)	0 (0)	0 (0)	- (-)	4.3 (4.7)	0.33 (0.22)
1IQD	9 (0)	10 (4)	31 (24)	*** (**)	1.0 (1.5)	0.74 (0.48)	2IDO	0 (0)	0 (0)	8 (0)	* (-)	3.3 (4.0)	0.46 (0.28)
1J2J	0 (0)	4 (3)	47 (35)	** (**)	1.4 (1.4)	0.85 (0.47)	2J0T	0 (0)	0 (0)	0 (0)	- (-)	7.2 (7.1)	0.20 (0.08)
1JIW	0 (0)	0 (0)	15 (11)	* (*)	2.6 (2.8)	0.77 (0.49)	2J7P	0 (0)	0 (0)	5 (1)	* (*)	3.2 (3.6)	0.35 (0.16)
1JK9	0 (0)	0 (0)	10 (6)	* (*)	2.9 (3.1)	0.88 (0.47)	2JEL	3 (0)	1 (1)	18 (12)	*** (**)	0.7 (1.2)	0.85 (0.48)
1JMO	0 (0)	0 (0)	2 (0)	* (-)	3.9 (4.5)	0.31 (0.12)	2MTA	1 (0)	3 (3)	38 (14)	*** (**)	0.9 (1.0)	0.87 (0.50)
1JPS	3 (0)	34 (33)	70 (48)	*** (**)	0.6 (1.6)	0.84 (0.53)	2O3B	0 (0)	0 (0)	8 (2)	* (*)	3.6 (3.9)	0.52 (0.31)
1JTG	0 (0)	2 (2)	9 (6)	** (**)	1.0 (1.1)	0.68 (0.40)	2O8V	0 (0)	0 (3)	12 (7)	* (**)	2.8 (1.7)	0.82 (0.59)
1JWH	0 (0)	2 (0)	35 (15)	** (*)	1.6 (2.5)	0.69 (0.42)	2OOB	0 (0)	1 (1)	31 (35)	** (**)	1.7 (2.0)	0.73 (0.47)
1JZD	0 (0)	0 (0)	1 (4)	* (*)	3.3 (3.9)	0.37 (0.23)	2OOR	0 (0)	0 (0)	2 (1)	* (*)	3.5 (3.8)	0.32 (0.22)
1K4C	13 (0)	2 (1)	40 (40)	*** (**)	0.6 (1.9)	0.90 (0.46)	2OT3	0 (0)	0 (0)	0 (0)	- (-)	4.3 (4.7)	0.23 (0.16)
1K5D	0 (0)	0 (0)	13 (7)	* (*)	2.8 (3.4)	0.29 (0.14)	2OUL	9 (4)	0 (0)	27 (10)	*** (***)	0.7 (0.7)	0.86 (0.72)
1K74	0 (0)	17 (8)	50 (33)	** (**)	1.7 (1.8)	0.72 (0.55)	2OZA	0 (0)	0 (0)	10 (8)	* (*)	2.5 (3.2)	0.51 (0.31)
1KAC	0 (0)	6 (0)	41 (31)	** (*)	1.5 (2.0)	0.67 (0.45)	2PCC	0 (0)	1 (9)	58 (37)	** (**)	1.4 (1.5)	0.83 (0.54)
1KKL	0 (0)	0 (0)	6 (4)	* (*)	3.2 (3.8)	0.53 (0.35)	2SIC	0 (0)	0 (0)	2 (2)	* (*)	3.5 (3.7)	0.26 (0.17)
1KLU	0 (0)	1 (2)	6 (7)	** (**)	1.9 (1.8)	0.65 (0.43)	2SNI	0 (0)	6 (2)	32 (13)	** (**)	1.6 (1.7)	0.67 (0.37)
1KTZ	0 (9)	29 (26)	112 (98)	** (***)	1.1 (0.6)	0.91 (0.68)	2UUY	3 (0)	3 (0)	72 (43)	*** (*)	0.8 (2.4)	0.88 (0.38)
1KXP	0 (0)	5 (0)	4 (6)	** (*)	1.3 (2.8)	0.68 (0.32)	2VDB	1 (0)	1 (1)	14 (11)	*** (**)	0.9 (1.4)	0.85 (0.62)
1KXQ	0 (0)	14 (1)	34 (23)	** (**)	1.0 (1.7)	0.83 (0.56)	2VIS	0 (0)	6 (2)	28 (26)	** (**)	1.5 (1.6)	0.74 (0.48)
1LFD	0 (0)	0 (0)	47 (38)	* (*)	2.4 (2.9)	0.63 (0.50)	2Z0E	0 (0)	0 (0)	2 (0)	* (-)	3.1 (4.3)	0.33 (0.15)
1M10	0 (0)	0 (0)	14 (5)	* (*)	2.9 (3.7)	0.35 (0.21)	3BP8	0 (0)	4 (5)	31 (15)	** (**)	1.5 (1.3)	0.79 (0.58)
1MAH	3 (0)	13 (4)	25 (25)	*** (**)	0.9 (1.9)	0.71 (0.36)	3CPH	0 (0)	0 (0)	22 (15)	* (*)	2.9 (2.6)	0.49 (0.28)
1ML0	0 (0)	14 (0)	48 (25)	** (*)	1.2 (2.4)	0.75 (0.43)	3D5S	0 (0)	23 (3)	95 (73)	** (**)	1.1 (1.9)	0.88 (0.54)
1MLC	0 (0)	0 (0)	6 (4)	* (*)	2.2 (2.9)	0.77 (0.32)	3SGQ	1 (0)	4 (1)	34 (27)	*** (**)	0.7 (1.3)	0.91 (0.59)
1MQ8	0 (0)	0 (0)	9 (9)	* (*)	3.0 (3.6)	0.62 (0.42)	4CPA	0 (0)	6 (5)	45 (24)	** (**)	1.2 (1.6)	0.86 (0.49)
1N8O	0 (0)	0 (0)	8 (4)	* (*)	2.3 (2.4)	0.66 (0.25)	7CEI	0 (0)	29 (5)	99 (78)	** (**)	1.1 (1.6)	0.85 (0.57)

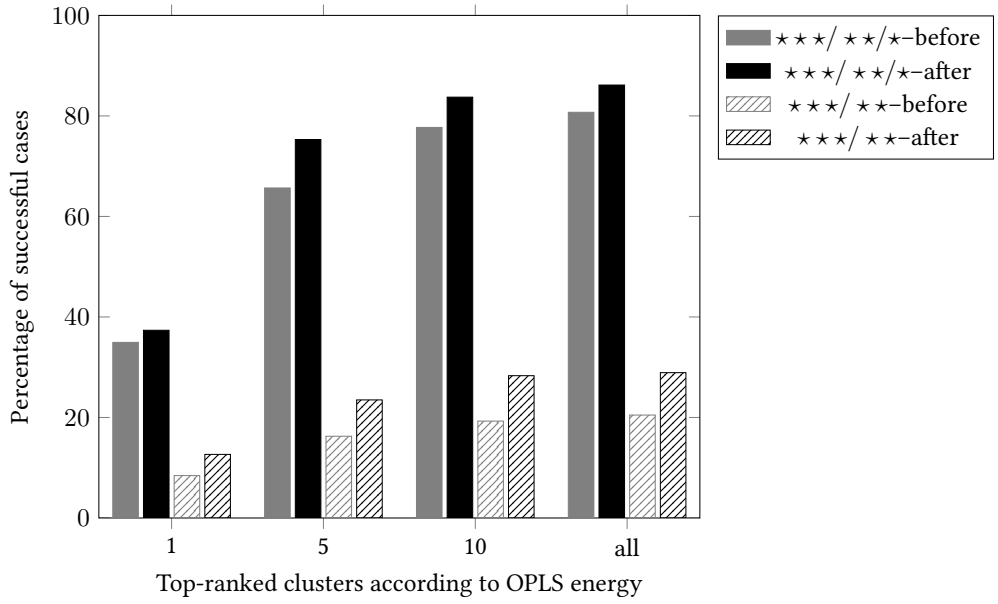


Figure 5.2. Percentage of successful refinement cases yielding acceptable (or better) CAPRI quality for the top scoring clusters. Clusters of docking solutions were generated based on pairwise RMSD (see Methods for details). The clusters were ranked by the average of the intermolecular energies derived from the OPLS based energy function of the top 4 scoring members. A cluster is designated as of one star CAPRI quality or better if any of the top 4 members is of one star CAPRI quality or better. The evaluation of the scoring was done both for rigid-body and iATTRACT refinement results.

We then aimed to elucidate the individual contributions of sampling and scoring in iATTRACT. First, we examined improvements in sampling; i.e., whether the refinement procedure yields protein docking geometries that are in closer agreement with the corresponding native complexes compared to the initial rigidly docked complexes. This analysis was performed on all structures independent of their energetic ranks. Second, improvement in scoring was investigated by comparing these ranks before and after iATTRACT refinement.

5.3.2. Sampling performance

Figure 5.3 shows that significant improvements in sampling were achieved, comparing initial and final IRMSD and fnat values before and after refinement. Most important, the improvements in the predicted native contacts are extremely favorable (up to 70%) and match our goal to improve sampling of structures with a high level of correct residue interactions. iATTRACT not only optimizes interface side chain conformations but also allows for large-scale refinement, as illustrated by the docking trajectories shown in Figure 5.4 where poses improved by up to $5\,\text{Å}$ in IRMSD. The average change in IRMSD for all structures was $\Delta\text{IRMSD} = i{-}0.293\,\text{Å}$ and the average change in fnat was $\Delta\text{fnat} = 0.071$. A limitation of the analysis to one star structures or better yields $\Delta\text{fnat} = 0.189$ and $\Delta\text{IRMSD} = -0.151\,\text{Å}$. The average increase of roughly 20% in fnat for quality structures is extremely encouraging. In nearly all cases, this resulted in an improvement (or no change) in the CAPRI quality of the best model (Table 5.2). At the same time, even for "non-quality" structures, improvements in fnat of on average 0.058 were observed. These observations underline the overall improvement in terms of predicting native contacts by the iATTRACT protocol. In fact, a considerable number of these structures were refined to one-star quality or better: iATTRACT refinement leads to a large increase (more than 50%) in the pool of CAPRI-quality refined structures. Most of these structures were already close to one-star quality, but as many as 10 % of these "new" quality structures were generated from structures with an initial IRMSD $> 6\,\text{Å}$ (see also Figure 5.3). In general, structures with an IRMSD from $4\,\text{Å}$ to $5.5\,\text{Å}$ are likely to become one-star structures after iATTRACT refinement (Figure A.1). These results again emphasize the large-scale sampling capacity of the iATTRACT protocol. Improvements were observed across the whole benchmark, performing similarly for proteins classified as antibody, enzyme, and other, and rigid-body, medium, and hard docking cases, respectively (Table A.1).

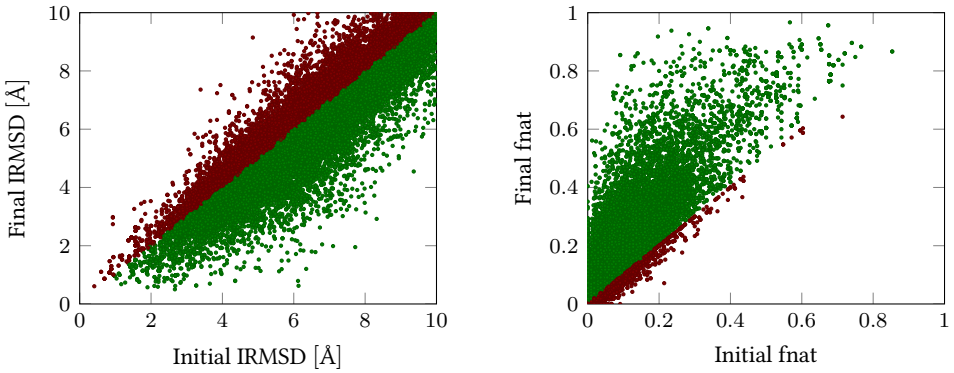
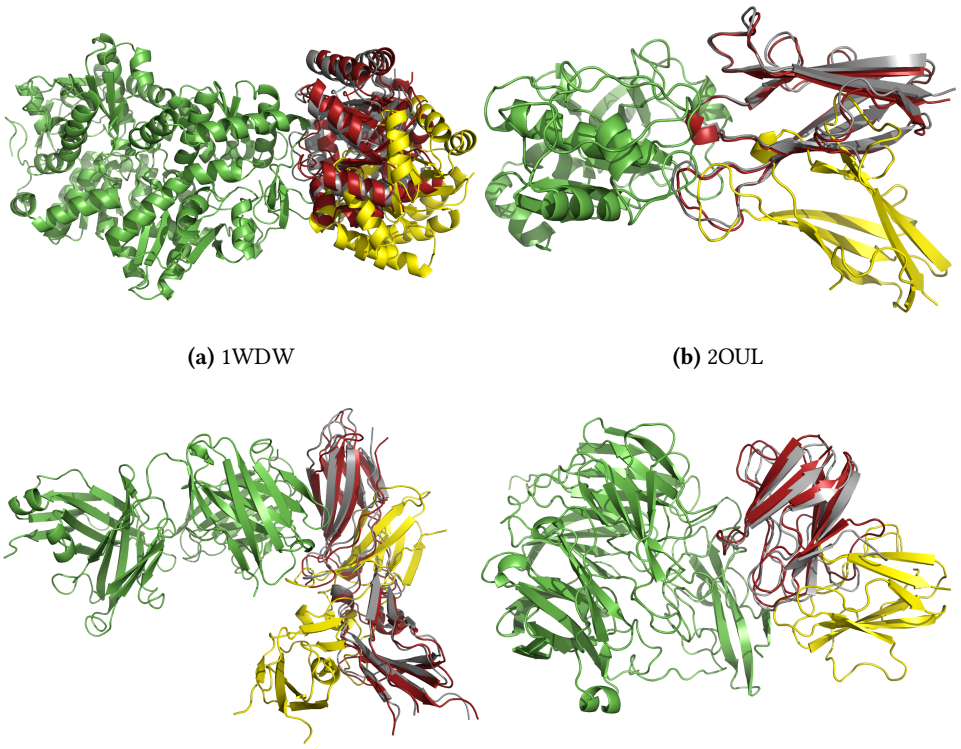


Figure 5.3. Structural change during flexible interface refinement for all complex structures. The IRMSD and fnat of the initial structures are compared with the final structures. Improvements are denoted by green markers, deterioration by red markers.

5.3.3. Sampling the bound conformation

Flexible refinement can also potentially bring individual protein partners closer to the bound form. In addition, this induced fit may also lead to an increase in the fraction of native contacts. For the average change in fraction of native contacts, there is indeed a positive contribution due to interface changes which accounts for roughly 16% of the overall increase in fnat for quality structures. For some cases, refinement resulted in a conformation closer to the bound form (Figures A.2, A.3). This was especially true for long flexible side chains that form many contacts with the partner protein and become buried at the interface. When considering refined docking solutions with an IRMSD of the protein backbone below 2 Å, we found for 23% of the refined docking solutions a closer agreement of the interface side chain and backbone conformation of individual partners with respect to the bound protein conformation compared to the unbound protein structures (Figure A.2). One should keep in mind that the improvement of the conformation of individual protein partners depends on how close the predicted complex resembles the native binding geometry and how strongly the interface restricts the possible side chain structure of interface residues.

In many cases, only small conformational changes at the interface were detected. However, these changes were decisive to remove sterical barriers for larger-scale rigid-body movements of the ligand protein. Hence, increased local flexibility can lower barriers for triggering simultaneous global large-scale motions which can result in improved surface complementarity and a larger fraction of native contacts.



(a) 1WDW **(b)** 2OUL

(c) 1AHW **(d)** 2MTA

Figure 5.4. Refinement trajectories with large-scale displacements. The initial structure of the ligand protein is drawn in yellow, the final placement in red. The reference from the crystal structure is shown in gray. The images were generated using PyMOL [404].

5.3.4. Scoring performance

Besides improvements in sampling, it is also desirable to obtain an improved scoring upon refinement. Therefore, the final iATTRACT models were ranked by their final docking energy. As illustrated in Figure A.5, a funnel-like behavior of binding energy vs. IRMSD was observed for several (but not all) complexes. To investigate the pure effect of scoring improvement without sampling improvements, the exact same models were also ranked by their energy before refinement. For both rankings, the best IRMSD within the top-ranked 5 structures was assessed. Upon refinement, this value decreased on average by $0.5\,\text{Å}$, a significant improvement ($p = 0.001$). However, no such improvement was found for the top-ranked structure. Hence, further optimization in scoring is needed to reliably distinguish structures of low IRMSD from decoys and enhance the sampling by more advanced schemes such as Monte Carlo minimization.

5.3.5. Quality of the iATTRACT models

The structures generated by iATTRACT were validated using the WHAT IF server [474, 186, 187] to assess possible errors resulting from the simple structure-based intra-protein force field. We found good quality of the models with respect to bond lengths, angles and even backbone and side chain dihedral angles (data not shown). Apparently, repulsive potentials are sufficient to ensure correct dihedral angles as noted in previous work [520]. However, we found planarity deviations for the flexible aromatic residues. The final complexes were, therefore, relaxed in a short all-atom minimization with Amber12 [61]. The quality of the Amber-reminimized final models was satisfactory with respect to all criteria, while mostly only introducing minor changes to IRMSD in the range of $0.2\,\text{Å}$ (data not shown).

5.4. Discussion

Applied to a large benchmark set of protein-protein complexes, the iATTRACT refinement method resulted in significant improvements of the quality of many docking solutions in terms of IRMSD and fnat with respect to the native complexes. A major effect on improving the fraction of native contacts up to 70% was observed. The improved packing is also reflected by the increased CAPRI quality of the solutions and the improvements with respect to the fnat of the best structure for each protein complex. In addition, a modest but significant improvement in scoring of the near-native refined solutions was observed. Still, further optimization of the force

field-based scoring is necessary to detect near-native docking solutions among alternative solutions.

When looking at the docking trajectories, we observed a coupling between the small-scale interface movements and the large-scale global displacements. The addition of local-scale degrees of freedom smoothed the energy landscape and allowed many rigidly docked complexes to escape from local minima and move further towards a near-native conformation (note that the start structures were fully converged in rigid-body space using the same force field). This effect might be enhanced by the softness of the structure-based force field and its harmonic potentials, which do not restrict torsion angles and do not include attractive non-bonded interactions and therefore allow enhanced sampling of residue conformations at the interface. Improvements on IRMSD mainly resulted from global movements of the whole partner proteins. Still, the movements of the interface atoms contribute to forming correct residue contacts between the protein partners.

Even though we obtained some promising results for refining non-quality structures to CAPRI-quality, refinement results still depend on the quality of the input structures. For complex structures beyond an initial IRMSD of 5.5 Å, the probability of improvement is low (see Figure 5.3 and Figure A.1). Also, large-scale backbone conformational changes cannot be modeled successfully. In earlier work, global normal modes were used to model backbone deformations [297]. We plan to combine the full interface mobility option with simultaneous large-scale backbone movements in future studies. Other possibilities for modeling large-scale conformational change are ensemble and multi-body docking, which are all available in ATTRACT. The combination of different flexibility mechanisms will help to tackle difficult docking cases mimicking interface flexibility, conformational selection and induced-fit processes.

It is interesting to compare the present results with the performance of other published methods. In semi-flexible refinement and subsequent refinement including explicit water with HADDOCK [95], a total average change of fnat Δfnat $= 0.11$ was achieved for the same set of complexes for one star quality structures [95]. HADDOCK runs typically use 50 to 100 CPU hours to generate 200 final models for a protein complex [98]. The results of the present computationally less demanding iATTRACT method with Δfnat $= 0.189$ for one star solutions compare very favorably, corresponding to an 80% higher increase in fnat. In another molecular dynamics-based refinement procedure, Król et al. reported increases in fnat between 0.06 and 0.11 for near-native poses compared to rigid-body results [240]. The iATTRACT refinement procedure requires only a few minutes of computer time on a single core for a typical medium sized complex, which is much less than other current all-atom molecular dynamics refinement procedures. Since refinement of a set of complexes can be performed independently, it is possible to refine hundreds of structures in par-

allel within an hour on a cluster computer. iATTRACT calculation per structure is slower than FiberDock [293], which is a highly computationally optimized program for improving the scoring of rigid-body structures, but which does not additionally sample overall protein complex conformations.

5.5. Conclusion and Outlook

Proteins often show considerable conformational flexibility upon complex formation. In particular, interface residues undergo significantly more changes than other surface residues [171]. Hence, protein-protein docking of unbound partner proteins typically yields only complexes that deviate significantly from the native complex geometry and often include only a small fraction of native contacts. A new refinement approach, iATTRACT, has been presented that combines full atomistic interface flexibility with rigid-body motions in a simultaneous energy minimization. When testing this approach on cases from benchmark 4.0 [200], we obtained significant improvements for a large variety of protein complexes and large initial deviations from bound geometries. The procedure performed especially well in cases of small or medium changes in the interface conformation upon complex formation. In many cases, large-scale improvements of IRMSD and in the number of native contacts were observed. Currently, CAPRI criteria classify structures of IRMSD $> 4\,\text{Å}$ as incorrect. However, iATTRACT refinement may allow in many cases the refinement of $4\,\text{Å}$ to $5.5\,\text{Å}$ IRMSD candidates to CAPRI quality predictions. These structures could be classified as CAPRI "half star" quality, since they can potentially be refined to quality predictions with iATTRACT. Scoring is typically optimized towards detecting docking candidates of high CAPRI quality. Instead, it might be beneficial to recalibrate towards ranking a large number of "half star" rigid-body candidates; e.g., in the top 500 for subsequent iATTRACT refinement. Recent studies indicate that 3D models for many protein-protein interactions can be generated based on sequence similarity to a known complex [512, 245]. Depending on the degree of similarity to a known template complex, such structural models may only reach acceptable quality and the iATTRACT approach could be used for further refinement of such template based model complexes. iATTRACT refinement can be selected as an option in the ATTRACT Easy protein-protein docking web-interface [99] and is also part of the pepATTRACT [391] (Chapter 6) and loopATTRACT (Chapter 7) protocols. Recently, iATTRACT was also extended towards refining protein-nucleic acids complexes and multi-body assemblies (Schindler, unpublished data). In the future, we want to systematically evaluate iATTRACT's performance on these type of interactions. Furthermore, we plan to add support for ions and cofactors during iATTRACT refinement (ATTRACT 2.0, see Chapter 4).

6. Fully Blind Peptide-Protein Docking with pepATTRACT

Peptide-protein interactions are ubiquitous in the cell and form an important part of the interactome. Computational docking methods can complement experimental characterization of these complexes, but current protocols are not applicable on the proteome scale. This chapter presents the fully blind flexible peptide-protein docking protocol pepATTRACT. pepATTRACT combines a rapid coarse-grained global peptide docking search of the entire protein surface with a two-stage atomistic flexible refinement. Global unbound-unbound docking yielded near-native models for 70% of the docking cases when testing the protocol on the largest benchmark of peptide-protein complexes available to date. This performance is similar to that of state-of-the-art local docking protocols that rely on information about the binding site. Upon restricting the search to the peptide binding region, the resulting pepATTRACT-local approach outperformed existing methods. The majority of the work presented in this chapter is published in [391].

6.1. Introduction

Peptide-mediated interactions play a dominant role in cellular processes and account for about 40% of all protein-protein interactions [343]. Peptide-protein complexes are involved in many signaling and regulatory pathways as well as the DNA replication machinery. A range of pathological disorders is related to peptide-protein interactions [319], making them interesting leads for protein drug design. However, for rational design of peptidic drugs, a thorough understanding and atomistic structural knowledge of peptide-protein complexes is necessary [461, 375]. A number of structures have been resolved experimentally and have provided important insight into the nature of peptide-mediated interactions [275, 343, 106, 426]. However, a large number of complexes is still lacking to date. Computational peptide-protein docking methods can complement experiments by providing models for the bound complex structure.

For proteome-wide applications, a peptide-protein docking method has to be fully blind, meaning that it should be based solely on the unbound (apo) structure of the protein and the peptide sequence. In other words, such an approach should predict both the peptide binding site (global search of the protein surface) and the bound peptide conformation to high precision simultaneously. A number of peptide-protein docking and binding site predictions tools have been developed to date [118, 344, 35, 90, 447, 253, 468, 384, 181, 453, 450, 370, 52, 14, 325, 425, 361, 362, 448, 13, 114, 283]. Global docking and binding site prediction methods [118, 344, 35, 90, 447, 253, 468, 384] often identify the correct binding site but do not yield high-quality models for the peptide conformation [278]. The ligand docking approach Autodock was adapted to fully blind peptide docking but is limited to short fragments [181, 453, 278]. In contrast to global prediction methods, local docking approaches sample peptide conformations at a known binding site only (and therefore are not fully blind). Local docking approaches can often yield high-quality models when tested on peptide-protein docking benchmarks [278]. Local methods include ligand docking based approaches [450] and several target-specific approaches (e.g. for MHC and PDZ domains) [370, 52, 14, 325, 425]. Several protocols are based on protein-protein docking which is also reflected by the recent addition of peptide-protein targets to the community-wide docking challenge CAPRI [257, 259, 260]. The Rosetta FlexPepDock ab-initio approach combines local docking at the binding site with peptide folding using a backbone structure library [361, 362]. Trellet et al. developed a peptide docking protocol in the data-driven docking program HADDOCK that combines the principles of conformational selection and induced fit through an ensemble of peptide conformations and flexible refinement stages [448]. They discovered that using only three distinct peptide conformations yields very good predictive performance [448], supporting earlier observations on the frequency of peptide conformations found in peptide-protein complexes [275]. The DynaDock method uses a soft-core molecular dynamics based refinement [13], whereas PepCrawler employs a fast RRT-based algorithm for local sampling of peptide conformations [114]. Recently, a molecular dynamics based approach was developed in our group that employs Hamiltonian replica exchange simulations with a variation of soft core potentials along the replicas. This method showed promising results with respect to local peptide-protein docking [283].

Here, we present pepATTRACT, a flexible peptide-protein docking approach in ATTRACT [506, 297, 393]. pepATTRACT is fast: it performs a coarse-grained ab-initio docking search within minutes, followed by atomistic refinement of only the most favorable solutions. More important, pepATTRACT is fully blind: it requires neither knowledge of the binding site nor the peptide conformation. pepATTRACT yields high-quality models of the complex, comparable to the state-of-the-art lo-

cal docking protocols Rosetta FlexPepDock ab-initio [362] and HADDOCK peptide docking [448]. We also combined pepATTRACT with ambiguous interaction restraints [322, 112] that define the peptide binding site, as used before with HADDOCK. The resulting pepATTRACT-local protocol outperformed both HADDOCK and Rosetta FlexPepDock ab-initio by a significant margin on a large variety of peptide-protein interactions.

6.2. Methods

The fully blind peptide-protein docking protocol pepATTRACT consists of the following steps (Figure 6.1). First, peptide model structures are generated from sequence [437]. Then global rigid-body docking with ATTRACT using a coarse-grained force field is performed [506, 297]. The rigid-body docking solutions are ranked by ATTRACT score and the best 1000 ranked models are selected for a subsequent atomistic refinement stage with iATTRACT [393]. The structures were then finally refined in a molecular dynamics simulation with AMBER 14 [60].

6.2.1. Peptide structures

For each peptide, we generated three conformations from its sequence using the Python library PeptideBuilder [437]. We chose backbone dihedral angles to represent α-helical ($\phi = -57°, \psi = -47°$), extended ($\phi = -139°, \psi = -135°$) and poly-proline conformation ($\phi = -78°, \psi = 149°$). This conformational selection approach is based on [448].

6.2.2. ATTRACT rigid-body docking

The protein and peptide structures were converted to the ATTRACT atom type representation [506] with the ATTRACT tool `reduce` (see Chapter 4). Starting points were generated by choosing random positions and orientations for the docking partners. For bound-bound docking, we used 100,000 starting points and tripled this number for unbound-unbound docking to account for the three possible peptide conformations. The starting structures were subjected to rigid-body optimizations in a potential energy minimization of 1000 minimization steps with the ATTRACT metric minimizer [296, 297]. Energy calculation was accelerated using a precalculated grid [99, 100] and an additional harmonic potential was applied on the protein's center of mass to draw the peptide towards it ("gravity"). A subsequent potential energy minimization of 1000 minimization steps was applied without this gravity potential. All peptide conformations were docked separately (ensemble docking). The complete

docking run takes in the order of 10 minutes to 1 h depending on the size of peptide and protein partner. Finally, the docking candidates were ranked by ATTRACT energy evaluated within a squared cutoff of $50\,\text{Å}^2$.

6.2.3. iATTRACT refinement

The protein and the peptide structures were converted into the OPLS atom type description with the ATTRACT tool `aareduce` (see Chapter 4). Missing hydrogens were built with PDB2PQR [111, 110] and protonation states were determined by PropKa[266]. For histidine protonation states, the bound structure was used as a reference to ensure that unbound and bound structures contained the same atoms. Disulfide bridges were also determined according to the bound structure based on a cutoff criterion. This was necessary for easy evaluation of the refinement results against the bound crystal structure. The peptides' termini were charged, the proteins' left uncharged. The atomistic refinement uses a physical force field based on the OPLS parameters to calculate non-bonded and electrostatic interactions (see Chapter 4). Contacts from the input structure are treated as flexible during a simultaneous potential energy minimization in rigid-body degrees of freedom and interface flexibility [393] (see Chapter 5). The best 1000 ranked models from rigid-body docking with ATTRACT were selected for iATTRACT refinement. The refinement parameters were chosen as specified in [393] with the cutoff radius for selecting the flexible interface residues set to $5\,\text{Å}$. In most cases, this meant that the entire peptide (backbone and side chains) was flexible during iATTRACT refinement. Since the global scoring performance of the OPLS-based force field was found to be worse than that of the ATTRACT force field, structures were not rescored after iATTRACT refinement (see Figure B.1).

6.2.4. AMBER refinement

The structures were converted to the AMBER atom type description using the pdb4amber tool. A Generalized-Born implicit solvent model (igb=8) was used with the newest version of the AMBER force field ff14SB [60]. The structures were first minimized with the sander program (500 steps) with a short cutoff to relax possible atom overlap and deformations resulting from the structure-based force field used in iATTRACT refinement. Then two short molecular dynamics simulations were run with the pmemd.cuda program for 1000 and 2500 steps at temperatures $T = 400\,\text{K}$ and $T = 350\,\text{K}$ respectively. During the molecular dynamics simulations, intra-molecular distances for the protein and intermolecular distances between protein and peptide backbone atoms were restrained to prevent large deformations and peptide dissocia-

tion. The intra-molecular distances were restrained with a harmonic potential to the distance found in the structure with force constant $2\,\mathrm{kcal/mol/\mathring{A}^2}$ and for deviations of larger than $3.5\,\mathring{\mathrm{A}}$ with a linear response function and force constant $2\,\mathrm{kcal/mol/\mathring{A}}$. The intermolecular distances were allowed to change by $10\,\mathring{\mathrm{A}}$ with respect to the distance found in the initial structure. Deviations from $10\,\mathring{\mathrm{A}}$ to $13.5\,\mathring{\mathrm{A}}$ were penalized by a harmonic potential with force constant $0.25\,\mathrm{kcal/mol/\mathring{A}^2}$ and further deviations by a linear potential with force constant $0.25\,\mathrm{kcal/mol/\mathring{A}}$. Then the structures were minimized for 5000 steps with a large cutoff using the pmemd.cuda program without restraints. Finally, the energy was evaluated for the complex and the individual docking partners by the sander program. The binding interaction energy score was then calculated by subtracting the energy of the free partners from the energy of the complex. The final models were ranked by their binding interaction energy score. Tests on one docking case showed that implicit solvent simulations give comparable sampling to explicit solvent and in vacuo simulations, while yielding a better scoring performance at lower computational cost. The final structures were clustered based on the fraction of common residue contacts using a cutoff of 0.6 [368] and the clusters were ranked by the average energy of their top-ranked 4 members.

6.2.5. pepATTRACT-local

To perform local docking, we repeated the pepATTRACT protocol with additional restraints (pepATTRACT-local) recreating the conditions used in previous docking procedures [448]. We used ATTRACT with ambiguous distance restraints based on active and passive residues, following their original specification in the HADDOCK method [322, 112]. The active residues on the protein were derived from the residue contacts in the bound complex structure within a cutoff of $5\,\mathring{\mathrm{A}}$. All peptide residues were treated as passive residues. The minimum effective distance was set to $3\,\mathring{\mathrm{A}}$ during the coarse-grained docking simulations and to $2\,\mathring{\mathrm{A}}$ for the atomistic refinement. For rigid-body docking, an initial rotational sampling stage was added to the protocol in which only the restraints were applied and the protein and the peptide could orient towards each other with the translational degrees of freedom fixed [112, 448]. The rotational sampling phase applied a maximum of 1000 minimization steps.

6.2.6. Test set and assessment criteria

Docking was performed on 80 peptide-protein complexes from peptiDB docking benchmark [275] for which the unbound protein structures were available including several additions of unbound structures by Trellet et al. [448]. The protein structures were downloaded from the PDB. If necessary residues were renumbered in the un-

bound structures to match the bound forms, parts in the unbound form that are not present in the bound form were removed (and vice versa), and point mutations were introduced to resolve minor differences in the protein sequences.

To classify the benchmark, we aligned the unbound protein structure and the peptide models to the bound complex and calculated the backbone interface root-mean-square deviation IRMSD$_{ub}$ for all residues within a distance cutoff of $10\,\mathrm{\AA}$ of the partner molecule. Cases were classified according to the minimal IRMSD$_{ub}$ accounting both for protein flexibility and peptide modeling quality; i.e., similarity of the peptide to one of the idealized conformations. This classification scheme is similar to the one used for the protein-protein docking benchmark [200]. We chose the following criteria to characterize the docking cases:

- easy: IRMSD$_{ub} \leqq 1.5\,\mathrm{\AA}$
- medium: $1.5\,\mathrm{\AA} <$ IRMSD$_{ub} \leqq 3\,\mathrm{\AA}$
- hard: IRMSD$_{ub} > 3\,\mathrm{\AA}$.

According to this classification the benchmark contains 31 easy, 36 medium and 13 hard cases.

The docking solutions were evaluated by interface root mean square deviation (IRMSD) [303]. Since peptide-protein interfaces are typically smaller than protein-protein interfaces, we chose the following criteria [448] to characterize the docking solutions:

- not acceptable: IRMSD $> 2\,\mathrm{\AA}$
- near-native: IRMSD $\leqq 2\,\mathrm{\AA}$
- sub-angstrom: IRMSD $\leqq 1\,\mathrm{\AA}$

The IRMSD is calculated on the backbone atoms of both protein and peptide residues that are within $10\,\mathrm{\AA}$ of the partner molecules (as defined based on the crystal structure of the complex). We further refer to as "acceptable models" any near-native or better (sub-angstrom) predictions. For evaluating the sampling and the scoring performance we calculated the percentage of successful docking cases. A docking case was deemed successful if at least one acceptable solution was found in the top N solutions.

6.2.7. Rosetta FlexPepDock refinement

Rosetta FlexPepDock refinement was run on each of the 1000 final pepATTRACT models for a subset of 14 cases from our test set. The settings were chosen as described in [361] and the final models were evaluated by the Rosetta reweighted score [362]. For each pepATTRACT model, 200 refined models were generated and the lowest energy model was selected for evaluation.

6.3. Results

In this work, we developed a fully blind peptide-protein docking protocol (pepATTRACT) and embedded it in the ATTRACT docking engine (Figure 6.1). This protocol was tested on 80 peptide-protein complexes from the peptiDB benchmark [275, 448] for which the unbound protein structures are available. Initially, peptide models were generated from the peptide sequence [437] yielding three distinct idealized conformations: an extended, an α-helical and a polyproline-II conformation. This ensemble of peptide structures was successfully used in the HADDOCK peptide-protein docking protocol [448] and is supported by experimental observations on peptide conformations found in peptide-protein complexes [275]. The ensemble of peptide structures was then first rigidly docked to the protein partner using a coarse-grained representation of the partner molecules [506]. The rigid-body docking models were ranked by their ATTRACT scores and the best 1000 models were selected for atomistic refinement using the recently developed flexible interface refinement method iATTRACT [393]. Subsequently, these 1000 models were refined in a molecular dynamics simulation with Amber 14 [60] using a Generalized Born implicit solvent model (see Methods for details). The final models were clustered by the fraction of common residue contacts [368] and ranked by the average energy of the top-ranked 4 members [448].

6.3.1. Bound-bound rigid-body docking

The coarse-grained ATTRACT force field [506] has been used successfully to predict protein-protein complex structures in the past. Although good performance was found when using ATTRACT for peptide binding site prediction [384], it has not yet been applied systematically to peptide-protein complexes. To test the performance of the ATTRACT force field with regards to sampling and scoring peptide-protein complexes, we first performed bound-bound rigid-body docking for all cases yielding a theoretical limit for the performance of unbound-unbound docking. In terms of sampling, we obtained an overall success rate of 97% with only 2 failed cases (Figure B.2).

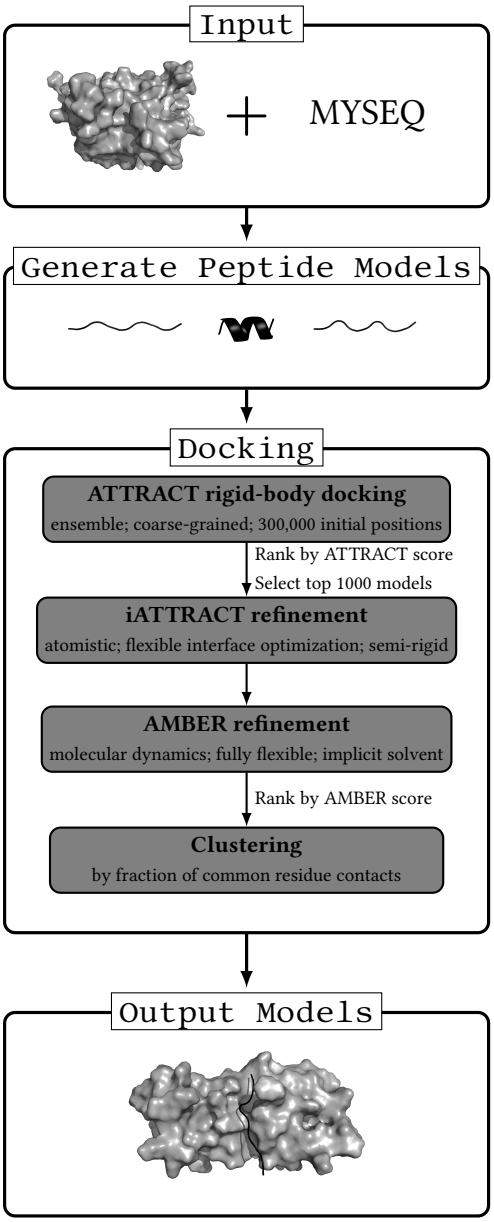


Figure 6.1. Workflow of the pepATTRACT peptide-protein docking protocol.

In both failed cases the peptide is "threaded" through a cavity in the protein and the binding site is not well accessible. 92% of the successful cases yielded models of sub-angstrom accuracy. Ranking the rigid-body docking solutions by their ATTRACT score gave a success rate of 86% in the top 1000 models and 41% for the top-ranked model. These results gave us confidence to use the coarse-grained ATTRACT force field for rigid-body sampling and scoring in the initial stage of the peptide-protein docking protocol (Figure B.2).

6.3.2. Unbound-unbound flexible docking

We then turned to the real challenge of blind unbound-unbound docking using a flexible docking approach with a coarse-grained rigid-body docking, an atomistic flexible interface refinement and a final molecular dynamics refinement stage (see Methods for details). Note, that this protocol requires neither knowledge of the binding site nor of the peptide conformation and thus represents a "worst case" scenario. Figure 6.2 shows the results for the docking success rates after the different docking stages. Overall, our protocol generated a near-native model for 70% of the 80 peptide-protein complexes (i.e., 56 complexes) when evaluating the top 1000 final docking models (see Experimental Procedures for details). 29% of these 56 successful cases yielded even sub-angstrom predictions. After clustering and ranking the clusters by the average energy of their top-ranked 4 members, the top 10 clusters contained at least one near-native cluster for 68% of the successful docking cases (48% of all cases) and the top-ranked cluster was found to be near-native in 29% of the successful cases (Figure 6.3 and Table B.1). Figure 6.4 illustrates docking models from these near-native top-ranking clusters. For the cases with sub-angstrom accuracy of the protein main chain also close agreement of predicted side chain structure with the native bound complex was observed (Figure B.1).

6.3.3. The effect of refinement

We wanted to analyze the effect of the different refinement stages considering sampling and scoring separately. When it comes to sampling, we took for each complex the full set of refined structures and computed the interface-RMSD (IRMSD) before and after refinement, without regard to any ranking. iATTRACT refinement increased the total success rate of the protocol by 10%. It succeeded both in refining structures to sub-angstrom precision as well as generating additional near-native solutions (Figure 6.2) and also helped to resolve minor clashes in transitioning from a coarse-grained to a full atomistic force field. This sampling improvement during iATTRACT refinement is also reflected by an average change in IRMSD of the structures

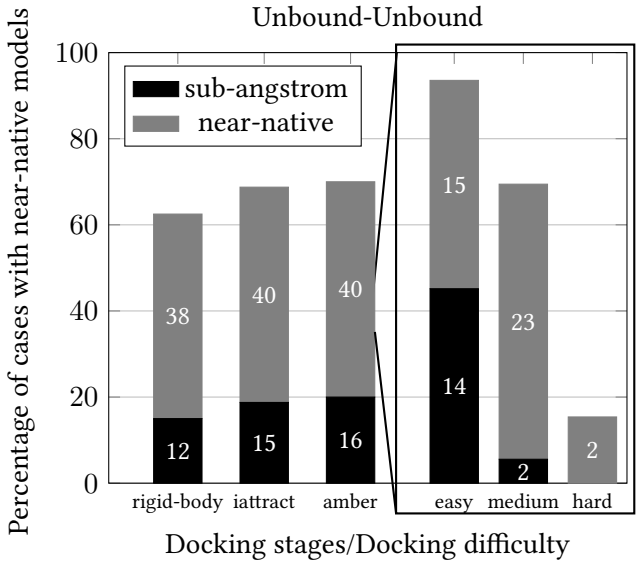


Figure 6.2. Percentage of acceptable docking cases after rigid-body docking and refinements for ab-initio unbound-unbound peptide-protein docking. A docking case was considered as a near-native/sub-angstrom hit if any of the final 1000 models was of near-native/sub-angstrom quality. The numbers on the bars report the absolute number of docking cases of near-native and sub-angstrom quality for each stage and difficulty. For a detailed list of the docking success for all cases see Table B.1. Reference data for bound-bound and unbound-bound rigid-body docking can be found in Figure B.2.

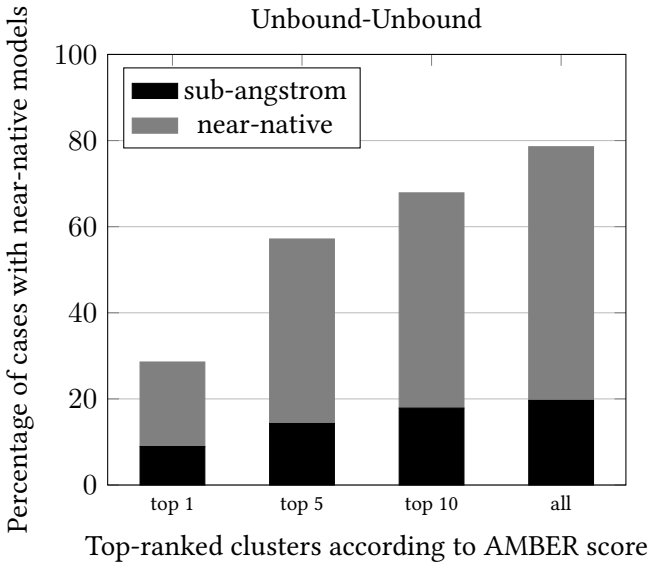
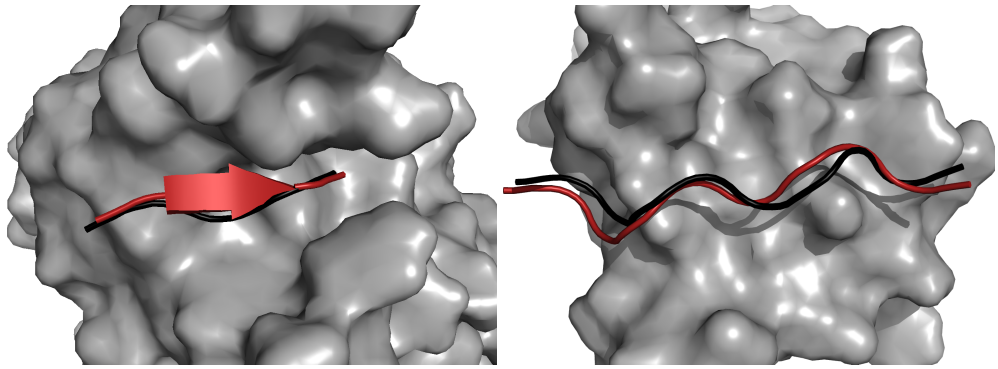
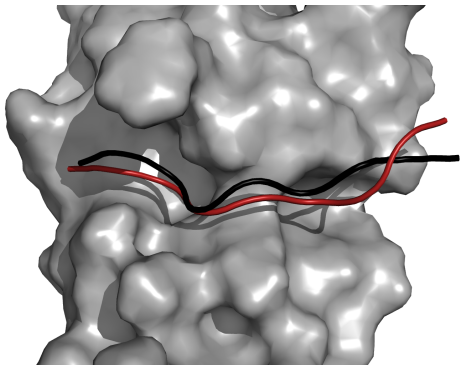


Figure 6.3. Scoring performance after clustering for the acceptable cases as a function of the number of clusters considered. A cluster is considered near-native (sub-angstrom) if any of its top-ranked 4 members is of near-native (sub-angstrom) quality. The clusters were ranked by the average energy of their top-ranked 4 members. For a comparison of the relative scoring performance of the scores before and after refinement see Figure B.1.

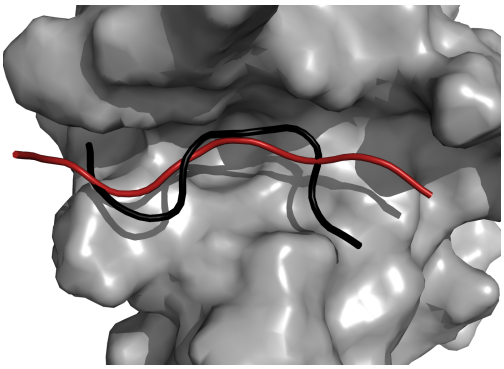


(a) 1NVR—0.7 Å—easy—3rd of 1st cluster

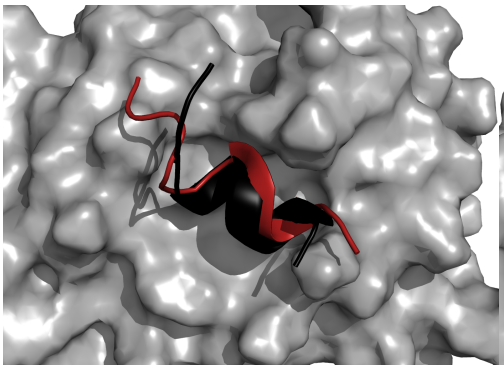
(b) 2O9V—0.6 Å—easy—4th of 1st cluster



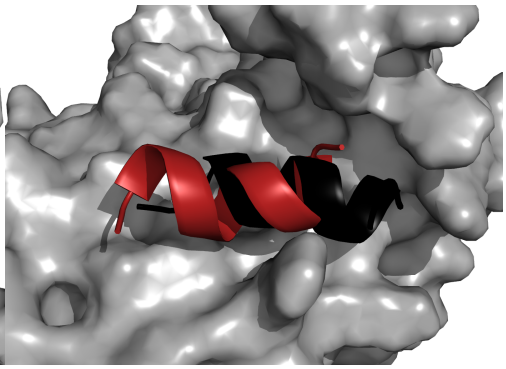
(c) 1Z9O—1.1 Å—medium—2nd of 1st cluster



(d) 1OU8—1.3 Å—medium—3rd of 1st cluster



(e) 2A3I—1.5 Å—hard—4th of 1st cluster



(f) 1NX1—1.8 Å—easy—2nd of 5th cluster

Figure 6.4. Examples of models generated by pepATTRACT unbound-unbound docking. For each complex, the PDB ID, the IRMSD, the docking difficulty and the rank of the structure after clustering are listed. The receptor is shown in gray shade, the peptide in red cartoon representation. The peptide structure from the crystallized complex is shown in black. Side chain conformations for the first two cases are shown in Figure B.3.

by $-0.10\,\text{Å}$. Note that iATTRACT refinement also allowed changes in the peptide main chain dihedral angles (Figure B.5). Compared to the results after iATTRACT refinement, AMBER refinement generated one additional successful docking case and improved the IRMSD of the structures on average by $0.44\,\text{Å}$. This clearly demonstrates that the additional flexibility and sampling of the MD refinement played a positive role, although structures that were already close to the bound form showed only little further improvement.

To capture only the effects of scoring, we ranked the final AMBER-refined docking models by different scores and then calculated for each ranking the best IRMSD of the top-ranked 10 structures for each complex. When comparing the ranking from before (ATTRACT score) and after refinement (AMBER score), we found an average improvement of $0.32\,\text{Å}$ for the AMBER-based ranking. This was connected to a 50% increase in the top 10 success rate with the AMBER-based ranking compared to the ATTRACT-based ranking (Figure B.1). Hence, the refinement yielded a significant improvement in terms of scoring.

6.3.4. Sampling the bound peptide conformation

On average, the backbone of the idealized peptide conformations deviated at least by $2.3\,\text{Å}$ from the crystallized form. We thus wanted to determine whether peptide structures moved closer to the bound form during docking. Figure 6.5 displays the IRMSD versus the change in peptide backbone-RMSD for the final near-native docking models. Interestingly, for sub-angstrom models, there was a clear tendency for the peptide structure to move closer to the bound form. A similar result was found for the RMSD calculated on all heavy atoms including side chains (Figure B.4). However, for models of only near-native quality, on average no improvement was found. This can be partly explained by the large amount of flexibility inherent to peptides and the fact that the interface does not restrict the conformation of all residues [362, 393]. Still, these results might indicate that more extensive peptide conformational sampling could be beneficial for some cases.

6.3.5. Binding site prediction

Several groups have proposed that contact analysis of docking models can be used to predict the interface of the proteins (interface post-prediction) [128, 381, 201, 258, 95, 384]. We thus also wanted to evaluate how well the binding site was predicted regardless of the peptide conformation. We analyzed the interface contacts of the top-ranked 10 final docking models and found that at least one true protein interface residue was identified in 99% of the docking cases. Furthermore, for 85% of the

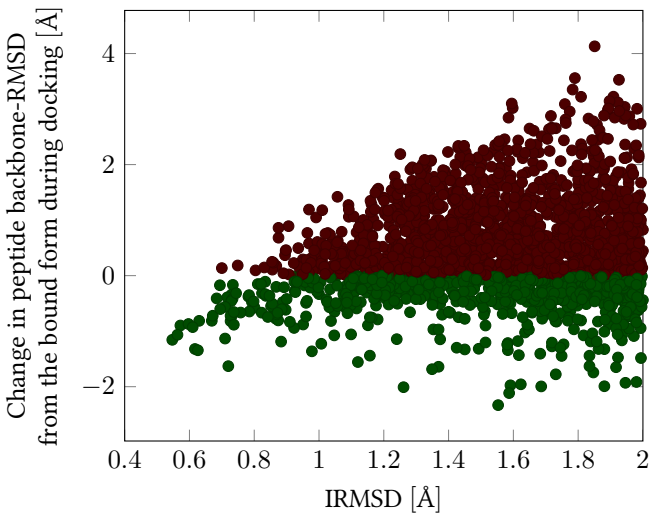


Figure 6.5. Change in peptide backbone-RMSD from the bound form for the final docking model relative to the initial idealized conformation as a function of the IRMSD of the final docking model. Only near-native models were evaluated. Green markers denote structures in which the peptide conformation moved closer to the bound form. Models of sub-angstrom quality are highlighted by a gray background. See also Figure B.4.

cases at least 50% of the correct interface residues could be recovered by these top-ranked 10 models. We compared our data to the recently published peptide binding site prediction tool PEP-SiteFinder. PEP-SiteFinder is based on the PTOOLS implementation of ATTRACT [506, 129, 383] and performs rigid-body docking with peptide structures generated by the PEP-FOLD method [295]. It was benchmarked on 41 unbound-unbound complexes from the peptiDB set. PEP-SiteFinder identified at least 50% of the correct interface residues in the top 10 poses in 71% of these cases [384], whereas our protocols was able to achieve this for 85% of these 41 complexes. In sum, the pepATTRACT protocol showed very good performance in interface post-prediction.

6.3.6. Comparison with other methods

It is interesting to compare our results to previously published methods. Here, we present a fully blind docking approach which includes a global search of the entire protein surface. Prior global methods were either limited to short peptide fragments [181] or did not yield high-quality models of the peptide conformation [278]. However, we can compare the performance of pepATTRACT to published local docking methods. In local docking, the position of the peptide is restrained towards its native binding site. The Rosetta FlexPepDock ab-initio protocol was tested on 14 unbound-unbound docking cases from our data set and achieved a docking success of 43% for the top-ranked 5 clusters (6/14) [362]. Evaluating the same set of complexes by the same criteria, we found a docking success rate of 50% (7/14) using pepATTRACT. The HADDOCK peptide docking protocol reported an overall success rate of 69% for unbound-unbound docking on 62 complexes [448]. When analyzing the data for this subset, we found an overall success rate of 73% among all final pepATTRACT models. We should however note that we did not achieve the same scoring performance: the top-ranked cluster was near-native in only 33% of the successful cases in contrast to 50% reported for the HADDOCK protocol [448]. Still, rather surprisingly, in terms of sampling pepATTRACT yielded results similar to or slightly better than the most successful local docking methods.

6.3.7. Local docking

While our blind docking results are in the range of success rates reported for the best local docking protocols FlexPepDock ab-initio and HADDOCK, we also wanted to make a direct comparison by only performing local docking. The pepATTRACT-local protocol included a set of ambiguous distance restraints [112] towards the binding site to restrict the sampling exactly matching the conditions used in the HADDOCK

protocol [448]. We applied the restraints both during the rigid-body sampling stage and the flexible interface refinement. For technical reasons, the ambiguous distance restraints were not used in the AMBER refinement and in the final scoring, which might have slightly deteriorated the results. Using pepATTRACT-local, we were able to generate a near-native solution in the top 5 clusters for 13 of the 14 cases tested in the published Rosetta FlexPepDock ab-initio protocol. The FlexPepDock ab-initio protocol itself could only achieve this result for 6 cases [362]. When limiting our data set to the 62 complexes used in [448], we obtained an overall success rate of 79% (49/62) with pepATTRACT-local and 37% of these successful cases yielded sub-angstrom models. This compares very favorably with the results from the HADDOCK peptide-protein docking protocol which achieved a 69% overall success rate (43/62) and only 23% sub-angstrom models among the successful cases [448]. Cluster-based scoring identified a near-native cluster at the top in 57% of the successful cases, which is comparable to the 50% achieved by the HADDOCK protocol [448]. The improved success rates and especially the improvements in scoring for pepATTRACT-local demonstrate the benefit of including additional information about the native binding site.

6.3.8. Combination of pepATTRACT with Rosetta FlexPepDock high-resolution refinement

In order to further increase the quality of the predictions, we tested the effect of high-resolution refinement by Rosetta FlexPepDock [361] on a subset of 14 docking cases. We ran FlexPepDock refinement as described in [361] on each of the 1000 AMBER-refined models generating 200000 FlexPepDock-refined models. We then selected the lowest-energy model as a representative and compared these 1000 models to the 1000 AMBER-refined models. The results are shown in Table 6.1 and Table B.2. In 10 of the 14 cases, the best IRMSD within the top 10 models was decreased after FlexPepDock refinement. Also in terms of overall sampling, the results were pointing towards an overall benefit with improvements in 6 cases, equal results (± 0.1 Å) in 5 cases and deterioration in IRMSD in 3 cases. However, for the top-ranked model no improvement was found. We found similar results for fnat (Table B.2), in addition here the top-ranked model also displayed a higher fraction of native contacts. In summary, it appears that FlexPepDock refinement can further improve the pepATTRACT models both in terms of sampling and scoring. These results should be validated on a larger set of cases in the future.

Table 6.1. Results for Rosetta FlexPepDock high-resolution refinement of pepATTRACT models. Refinement was performed on 1000 models for 14 cases from the docking benchmark. We ranked the pepATTRACT and FlexPepDock refined models by their AMBER and Rosetta score respectively and compared the best IRMSD within the top 1, top 10 and all sampled models. No clustering procedure was applied.

	pepATTRACT			FlexPepDock refined		
PDB	top 1 IRMSD	top 10 IRMSD	best IRMSD	top 1 IRMSD	top 10 IRMSD	best IRMSD
1AWR	1.39	1.17	0.73	0.60	0.60	0.57
1N7F	8.14	2.42	2.26	5.40	2.13	1.29
1NVR	1.03	0.68	0.68	4.97	1.29	0.62
1RXZ	3.08	2.86	1.58	4.26	3.25	1.35
1SSH	8.06	1.71	1.53	9.16	1.20	1.02
1T7R	2.66	2.66	1.64	2.76	1.46	0.99
1W9E	4.20	1.80	0.70	1.11	1.11	0.75
2A3I	12.30	1.65	1.51	14.26	2.14	1.81
2C3I	3.06	1.51	0.86	8.53	1.08	0.76
2FGR	8.73	7.89	4.60	13.54	11.42	4.80
2FMF	5.56	4.14	0.72	2.96	0.71	0.71
2O9V	0.73	0.73	0.61	0.58	0.42	0.42
2P54	10.26	6.93	1.64	14.69	4.93	1.68
2VJ0	8.13	4.06	1.16	2.80	2.59	1.50

6.4. Discussion

In this work, we developed a fully blind peptide-protein docking protocol, pepATTRACT that allows for global searches of the entire protein surface given the protein structure and the peptide sequence. It identifies the binding site and simultaneously predicts the bound peptide conformation for a large variety of complexes. This is in contrast to the previously developed binding site prediction method PEP-SiteFinder which also includes a global docking search using the PTOOLS/ATTRACT program [384]. PEP-SiteFinder only predicts the binding site but does not return structures of the peptide-protein complex. We also created the local docking protocol pepATTRACT-local which employs ambiguous interaction restraints [322, 112] to restrict the sampling towards a known binding site. pepATTRACT-local's performance clearly surpassed that of Rosetta FlexPepDock ab-initio [362] and HADDOCK [448] for a large number of peptide-protein complexes. We envision two application scenarios for pepATTRACT-local. Information about the native binding site can be obtained from experiments [81, 3] and easily included during the docking process to generate high-quality complex structures. If experimental data are unavailable, bioinformatic prediction tools could be used to identify possible binding sites [118, 344, 35, 253, 447, 468, 384]. As a special case of a bioinformatic prediction, the

contacts from the best pepATTRACT models can be extracted as an interface post-prediction (see section on binding site prediction). These predicted interface residues can then be used to restrict the sampling in a subsequent run with pepATTRACT-local and thus improve the results in terms of sampling and scoring (see section on local docking).

In the first stage of the pepATTRACT protocol, the coarse-grained ATTRACT force field was used which has been previously parameterized for protein-protein complexes [129]. The high success rates obtained already in the rigid-body sampling stage indicate that the force field is also applicable to model peptide binding. Vanhee et al. found that many of the conformations adopted by peptides in complexes are also found in monomeric proteins [462]. Recently, London et al. suggested that a large number of protein-protein interactions is dominated by the contributions of short binding motifs, so called 'hot segments' [277]. These recurrent interface design principles and thus the similarity between protein-protein and peptide-protein complexes could explain the success in applying the ATTRACT force field to the peptide-protein docking problem.

The success of the pepATTRACT protocol is based on an efficient combination of different flexibility mechanisms in the ATTRACT engine. The protocol employs a coarse-grained force field, ensemble docking, flexible interface refinement and a final molecular dynamics refinement to model protein and peptide flexibility. This versatile combination allows a high level of detail and accuracy in the final stages but at the same time is computationally efficient enough to screen 300,000 initial positions in a matter of minutes in the initial search stage. The large sampling in the rigid-body phase provides placements at the native binding site even of non-optimal peptide structures which were then relaxed to near-native models in the subsequent flexible refinement stages. Identifying many good initial global placements of the peptide and refining these is possibly more efficient than trying to sample all degrees of freedoms of the peptides right from the start due to the fact that it is easy to get stuck in local minima of the rugged docking energy landscape. Using a smoother coarse-grained energy function is certainly also helpful in this context. The coarse-grained representation of the peptide also partly compensates for inaccuracies in the initial peptide conformation.

While the overall success rate for pepATTRACT is highly encouraging, docking success still strongly depends on the quality of the peptide modeling and the range of conformational changes on the protein (Figure 6.2). For the 31 easy benchmark cases, we only had one case where we could not sample any near-native solution. Nearly half of these successful easy cases yielded sub-angstrom predictions. For docking cases of medium difficulty, we still obtained a good success rate of 69% for generating near-native solutions, however, for the hard docking cases this rate dropped

to 15% (2/13). For cases in which the best peptide model deviated by more than $5\,\text{Å}$ backbone-RMSD from the bound form, we were unable to sample any correct solution—also when using the local docking protocol. The current peptide docking protocol uses only three idealized peptide conformations and thus a very limited subset of the peptide conformational phase space. It does not take the sequence of the peptides into account; e.g., disulfide bridges and preferred backbone dihedral angles for certain residue combinations. More extensive peptide modeling could include statistical approaches [435], peptide backbone libraries [166] or even ab-initio folding in molecular dynamics simulations [184, 337]. Using more diverse peptide conformations may help to improve the sampling but also bears the risk of increasing the number of false positive solutions. It is also worth noting that there were only four cases in which the deviation of the bound peptide was greater than $5\,\text{Å}$ backbone-RMSD from the idealized conformations. This demonstrates that the idealized peptide conformations capture the main features of the bound form well. Furthermore, the correct binding site could be identified even with non-optimal peptide conformations (see Figure B.5 and section on binding site prediction).

In contrast, when examining the IRMSD between bound and unbound protein for the 24 failed docking cases, 14 cases display an IRMSD of $> 1\,\text{Å}$. To investigate the influence of the protein conformational change on docking success, we performed ab-initio rigid-body docking using the unbound structure of the protein partner and the bound form of the peptide. We found a success rate for the top-ranked 1000 models of 63% (Figure B.2) which is equal to the success rate found for unbound-unbound docking after the rigid-body stage (Figure 6.2) and significantly lower than for bound-bound docking (86%). Using the unbound protein structure prevents sampling of near-native conformations completely for 12 docking cases compared to 2 for bound-bound rigid-body docking (Figure B.2). In addition, the scoring performance deteriorated with only 74% of the successful cases ranked in the top 1000 compared to 89% in bound-bound docking (Figure B.2). Hence, the conformational change on the protein side seems to be more limiting to docking success than the accuracy of the peptide modeling. For considering receptor flexibility, the current protocol could be easily extended to include multiple conformations for the receptor in an ensemble docking approach or to approximately describe global backbone flexibility using pre-calculated normal modes [297]. Still, for the hard docking cases that include also partial refolding of the protein receptor, such a semi-rigid docking approach might not be sufficient.

6.5. Conclusion and Outlook

Peptide-protein interactions constitute a large fraction of all protein-protein interactions but due to their abundance and the inherent flexibility, many complexes have eluded experimental characterization. The high level of flexibility and the small size of the interface have proven to be obstacles for peptide-protein docking and to date many methods only perform local docking, relying on information about the peptide binding site. Previous global methods were either limited to short fragments or did not yield precise predictions for the peptide conformation [278]. To our knowledge, the pepATTRACT approach is one of the first fully blind flexible peptide-protein docking protocols for peptides of length scales typically found in peptide-protein complexes [275]. Applied to a large benchmark set of peptide-protein complexes, the pepATTRACT protocol yielded near-native models for 70% of the docking cases in a fully blind prediction manner. Its performance as a fully blind prediction method is comparable to some of the most successful local docking methods, Rosetta FlexPepDock ab-initio [362] and HADDOCK peptide docking [448]. pepATTRACT also gives very good results in interface post-prediction when compared with a state-of-the-art peptide binding site prediction tool [384]. The method could be useful for large-scale studies and the design of peptide-based inhibitors for modulating protein-protein interactions. In addition, interaction of globular proteins with disordered peptide or protein segments could also be modeled with this approach. Several peptides in the peptiDB benchmark are actually derived from disordered protein regions e.g. the cytoplasmic region of the group 1 metabotropic glutamate receptors for docking case 1DDV or the cytoplasmic tails of tumor necrosis factor receptor in docking case 1CZY. In the future, we plan to add support for modified amino acids, ions and cofactors to further enlarge the applicability of the protocol (ATTRACT 2.0, see Chapter 4).

7. Tackling large conformational changes: interface loop modeling with loopATTRACT

Accurately modeling binding-induced conformational rearrangements is a major obstacle to successful protein-protein docking. Interface loops are especially prone to such remodeling and sampling the correct conformation is difficult for loops longer than 12 residues. Here, we present a fast interface loop modeling protocol (loopATTRACT) that combines coarse-grained rigid-body docking and atomistic flexible interface refinement with a scoring function used successfully in protein structure prediction. We tested this approach on a set of challenging protein-protein docking cases. loopATTRACT improved the starting structure significantly in 7 out of 10 cases and performed as well as a more computationally demanding state-of-the-art loop modeling method. We also demonstrate the applicability of our interface loop modeling protocol in a hierarchical integrative modeling approach.

7.1. Introduction

The majority of cellular processes are carried out by interacting proteins. Protein-protein complexes are for example involved in DNA replication, protein synthesis, degradation and signal transduction. Disrupting these fine-tuned interaction networks often results in severe diseases and aberrant interactions have been identified in pathological disorders like Alzheimer's and cancer. Over the years, experimental methods like X-ray crystallography and NMR have revealed the structure of many protein-protein complexes allowing to understand their biological function in atomistic detail. These efforts have also fueled recent progress in drug design targeting protein-protein interfaces [483, 277]. However, for a large number of interactions, structural data are still unavailable and experimental structure determination for all protein-protein interaction types might not be feasible for decades [148]. Computational docking methods aim to extend the current structural coverage of the interac-

tome by predicting the structure of the complex based on the unbound structures or homology models of its constituents.

Protein-protein association is governed by the structural and physiochemical properties of the binding partners. Proteins are flexible molecules subject to thermal motions and the molecular recognition process often requires a certain degree of conformational adaptation. Binding-induced conformational changes can be small involving for example only side chain rearrangements of a limited number of interface residues. However, in many cases, conformational adaptation requires significant backbone movements. In particular, flexible interface loops often reorient upon binding. Furthermore, in the case of homology models, loop conformations are often less reliable, especially in the case of low sequence identity at the interface. Protein-protein docking methods perform satisfactorily in cases with little or no binding-induced structural changes in the protein-protein interface. However, since many algorithms employ a rigid-body approximation or consider only small-scale flexibility, it is difficult to accurately model large backbone motions like interface loop rearrangements. Consequently, many docking approaches fail to sample near-native solutions in such cases. The challenge resulting from interface loop flexibility in docking has also been documented during several rounds of the blind prediction challenge CAPRI [257, 259, 260, 262].

Several strategies have been employed to include interface loop flexibility in protein-protein docking. Multiple loop conformations can be generated prior to docking based on the unbound structures of the partners. Loop flexibility can then be included via an ensemble docking approach either docking each loop conformation separately or through a mean-field approach [30, 29]. Loop rearrangement can be sampled on-the-fly by docking flexible loops (e.g., in MD simulations) [51] or by rebuilding/perturbing the loops while exploring different overall geometries [477, 419]. However, such a direct sampling approach is usually limited to relatively short loops. Alternatively, loops can be modeled after an initial rigid-body/semi-flexible docking using the top-ranked complex geometries [478]. Such an a-posteriori strategy is particularly attractive in an integrative modeling setting; i.e., when using experimental information like cryo-EM densities to generate the initial complex model and to identify potentially flexible regions [96].

Irrespective of the specific time of loop rebuilding, for all approaches, it is necessary to identify the flexible loops and generate possible loop conformations. A range of loop modeling approaches have been developed in the protein structure prediction field [59, 290, 427, 254] and some have been made easily available via web-servers [229, 134]. Several approaches are based on robotics-inspired sampling algorithms. Canutescu and Dunbrack proposed a numerical cyclic coordinate descent algorithm to move a segment of the loop while keeping the anchors fixed [59]. Rosetta re-

builds loops based on an analytical robotics-inspired sampling approach (kinematic closure) which is applied iteratively in a Monte Carlo simulation in combination with loop backbone minimization and side chain repacking [290, 427]. FALC assembles loops from short fragments employing an analytical loop closure algorithm [254]. Das applied a step-wise enumerative approach adapted from modeling of RNA structure [92], while Olson et al. combined replica exchange coarse-grained lattice Monte Carlo with atomistic replica exchange molecular dynamics simulations to explore possible loop conformations [329]. Most of these protocols have been tested on loop lengths of up to twelve residues [329, 427] and often yield sub-angstrom precision loop models. However, these protocols have not yet been systematically applied to interface loop modeling. Interface loops can be less solvent-exposed due to contact formation with the protein partner and tend to be longer ($>$ 10 residues) than typical remodeling targets in protein structure prediction.

Here, we developed a fast interface loop modeling protocol, loopATTRACT, that combines methods from peptide-protein docking [391] and protein structure prediction [499, 498]. We tested the approach on complexes from protein-protein docking benchmark 5 [473] using the rigid-body superposition of bound and unbound protein partners and compared loopATTRACT's performance to that of the state-of-the-art loop modeling protocol in Rosetta [290, 427]. We found that despite its simplicity, loopATTRACT's performance was similar to that of the Rosetta loop modeling approach. We further tested interface loop modeling on docking models obtained from an integrative modeling approach driven by cryo-EM data [96].

7.2. Methods

The loopATTRACT protocol consists of the following steps:

1. Generating an ensemble of loop conformations.
2. ATTRACT rigid-body docking of loops to anchor points.
3. Rescoring of top-ranked 10000 models with DFIRE scoring function.
4. iATTRACT flexible interface refinement of top-ranked 1000 models.
5. Rescoring and final ranking by DFIRE scoring function.

loopATTRACT's input are the partner structures of the protein-protein complex and the start and end residue numbers specifying the loop for remodeling.

7.2.1. Loop conformations

The designated interface loop atoms were extracted from the input structure and the end-to-end distance and the length were determined from the coordinates. The BriX loop database [463, 19] was queried for loops of the same length and with a similar end-to-end distance ($\pm$ 1 Å). In case of too few (< 50) or too many matches (> 750), the end-to-end distance cutoff was adapted automatically and a new query was started. The protein structures with suitable loops were downloaded from the PDB. The backbone atoms of the available loop templates were extracted from the PDB files, the residues were mutated to the sequence of the target loop and the side chains were rebuilt with SCWRL [239]. All loop structures were fitted with an in-house tool onto the first loop conformation. This ensemble of loop structures was then used in an ensemble docking approach. For benchmarking purposes, it was assured that the loop from the bound complex was not part of the ensemble.

7.2.2. ATTRACT rigid-body docking

The protein-protein complex and the loop conformations were converted in the ATTRACT coarse-grained protein representation [506, 129] with the ATTRACT tool `reduce` (see Chapter 4). The loops were docked to the protein-protein complex using an ensemble docking approach. For each loop conformation, 1000 docking trajectories were generated starting from random positions and orientations of the loops. During docking, distance restraints were applied between the anchors on the protein and the loop. In a first minimization, the positions were fixed and the loops could orient towards the protein-protein complex. In this stage, only the distance restraints were applied as a force field ("ghost" mode). Then positions and orientation of each loop were optimized in a potential energy minimization of 1000 minimization steps using the ATTRACT metric minimizer. Energy calculations were accelerated by a pre-calculated grid [100]. During docking, interactions between the protein in which the loop resided and the loop were switched off; the ATTRACT energy was only evaluated between the loop and the protein to which the loop did not belong. Steric overlap between the loop and its native protein was prevented by an atom-density grid [96]. The generated models were ranked by their ATTRACT energy evaluated within a squared cutoff of 50 Å and the best 10000 models were selected for rescoring with DFIRE [499, 498]. The DFIRE scoring function [499, 498] was evaluated on the whole complex structure with default parameters using the DFIRE2.1 binary. The rescored structures were ranked by their DFIRE energy and the best 1000 models were selected for further refinement.

7.2.3. iATTRACT refinement

The protein-protein complex and the loop conformations were converted in the OPLS protein representation [212, 271] with the ATTRACT tool `aareduce` (see Chapter 4). Missing atoms were built with PDB2PQR [111, 110] and protonation states were determined by PROPKA [266]. During refinement, interactions were evaluated between all residues in the system. Distance restraints towards the anchors were applied as well. Nonbonded interactions between the backbone atoms of the anchor points and the loop were set to zero to allow correct placement of the loop. iATTRACT refinement was run with parameters as described in [393]. The refined models were scored and ranked with the DFIRE scoring function using default settings [499, 498].

7.2.4. Test set

We selected a subset of 10 complexes (Table 7.1) from protein-protein docking benchmark 5 [473]. These complexes displayed significant conformational changes upon binding involving large loop rearrangements and many were classified as hard docking cases [473]. We superposed the unbound proteins on the bound complex structure (global rigid-body superposition) and selected loops that displayed large conformational changes and often steric clashes in the unbound form for remodeling. Loop regions were defined between the nearest secondary structure elements in the unbound form. The loops in the test set contained between 10 and 20 residues. The superposed structures were used as starting structures for the interface loop modeling protocol. This provides an idealized test setting where the rigid-body placement has been determined at the highest possible precision.

7.2.5. Assessment criteria

We evaluated the final models by interface root-mean-square deviation (IRMSD) and fraction of native contact (fnat) [303] (see Chapter 3). The final loopATTRACT models were compared to the bound-unbound superposition (starting structure). A loop modeling case was termed successful if at least one among the top-ranked N models had a significantly lower IRMSD (decreased by more than 0.1 Å) or significantly higher fnat (increased by more than 0.02).

7.3. Results

Here, we developed a protocol for remodeling of large interface loops (loopATTRACT) based on an initial rigid-body model of the protein-protein complex. loopATTRACT

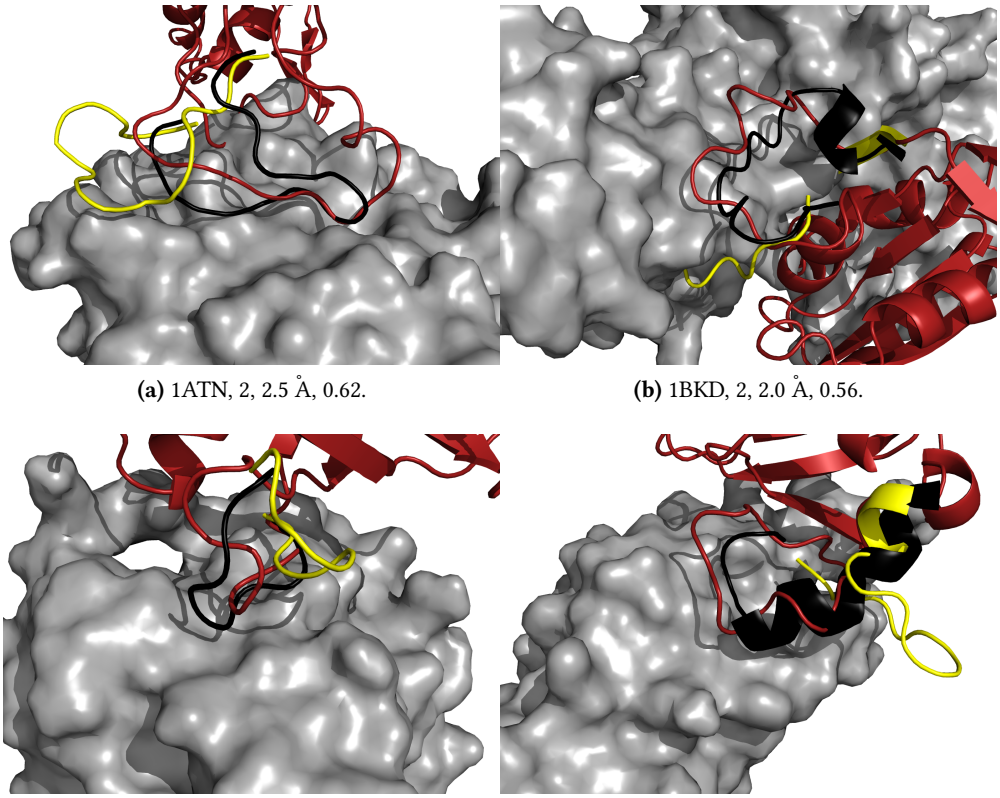
combines approaches from peptide-protein docking with a scoring function that has been previously used in protein structure prediction. First, a large ensemble of loop conformations extracted from a database of known protein fragments [463, 19] was docked rigidly against the complex using the anchor points of the loops as restraints [99]. During the initial large-scale docking, the ATTRACT coarse-grained protein representation was used [506]. The top 10,000 models were rescored with the DFIRE scoring function that has been previously developed for protein structure prediction [499, 498]. The top-ranked 1000 DFIRE models were then further optimized with the flexible interface refinement method iATTRACT [393]. Finally, the models were rescored and ranked by DFIRE.

7.3.1. Ab-initio interface loop modeling

The protocol was tested on 10 complexes from protein-protein docking benchmark 5 [473] using the unbound-bound rigid-body superpositions as starting structures (see Methods for details). These complexes display significant loop rearrangements upon binding and have been previously classified as difficult cases for protein-protein docking (medium/hard) [473]. The overall results are shown in Table 7.1 and examples of successful interface loop modeling are illustrated in Figure 7.1. In 7 out of 10 cases, we were able to generate a structure of significantly lower interface-RMSD (IRMSD) than the initial structure and rank it among the top 10 models (average Δ IRMSD $= -0.9\,\text{Å}$). Also when looking at the top-ranked model only, we found an improvement with respect to the initial model in 60 % of the cases. Furthermore, interface loop modeling increased the fraction of native contacts (fnat) of the best model in the top 10 on average by 0.11 compared to the bound-unbound superposition (initial structure). We also looked at the percentage of the 1000 final models and the top-ranked 10 models that improved by at least $0.1\,\text{Å}$ in IRMSD or by at least 0.02 in fnat. Interestingly, in the seven successful cases, the protocol indeed improved the majority of the generated structures either in terms of IRMSD (on average 82 % of the structures improved) or fnat (on average 53 %) (Table C.1). Note that improvements in IRMSD did not always result in increasing fnat (Table 7.1).

7.3.2. Effect of different protocol stages

We then analyzed the effect of the different stages on the overall loop modeling success by looking at the IRMSD and fnat distribution among subsets of 1000 models. We compared the top 1000 rigid-body models ranked by ATTRACT score, the top 1000 rigid-body models according to DFIRE score and the refined 1000 iATTRACT models. The results are shown in Figure 7.2. Re-ranking by DFIRE indeed had indeed



(a) 1ATN, 2, 2.5 Å, 0.62.

(b) 1BKD, 2, 2.0 Å, 0.56.

(c) 1LFD, 1, 1.4 Å, 0.72.

(d) 3FN1, 10, 2.2 Å, 0.39.

Figure 7.1. Interface loop modeling results. For each case, the PDB ID, the rank, IRMSD and fnat of the model are listed. The loop from the docking model is shown in red, the loop from the crystal structure of the bound complex in black and the initial loop position (unbound protein structure) in yellow.

Table 7.1. Results for interface loop modeling with loopATTRACT on 10 complexes from docking benchmark 5 [473]. For each complex, the PDB ID, the docking difficulty according to the classification in the benchmark [473], IRMSD and fnat of the initial rigid-body superposition, best IRMSD and fnat among the top-ranked and top-ranked 10 models and best sampled IRMSD and fnat are listed.

PDB	Difficulty	initial		top 1		top 10		all	
		IRMSD	fnat	IRMSD	fnat	IRMSD	fnat	IRMSD	fnat
1ATN	hard	4.54	0.35	2.97	0.49	2.02	0.62	1.75	0.62
1BKD	hard	3.15	0.53	2.47	0.54	2.03	0.56	1.78	0.60
1FQ1	hard	3.75	0.40	3.44	0.47	2.28	0.66	1.89	0.66
1LFD	medium	2.06	0.57	1.36	0.72	1.34	0.72	1.21	0.82
1PXV	hard	2.75	0.58	2.93	0.60	1.75	0.75	1.39	0.75
1R8S	hard	4.59	0.32	3.65	0.43	3.26	0.51	3.12	0.55
2NZ8	medium	2.22	0.44	3.16	0.40	2.40	0.46	1.90	0.55
2OT3	hard	2.55	0.49	2.82	0.33	2.50	0.40	2.46	0.50
3CPH	medium	2.13	0.53	3.65	0.47	2.64	0.51	2.04	0.55
3FN1	hard	3.74	0.36	2.70	0.43	2.24	0.49	1.97	0.57

a positive effect decreasing the median IRMSD and increasing median fnat compared to the ranking based on ATTRACT score. For iATTRACT refinement, we did not find an improvement in terms of IRMSD and only a slight improvement in fnat. While the maximum fnat increased from 0.75 before refinement to 0.82 after refinement, for some models fnat decreased resulting in an overall small increase of the median fnat (0.01). This is also reflected when analyzing the average change in fnat ($\overline{\Delta\text{fnat}}$) with respect to the initial structure. $\overline{\Delta\text{fnat}}$ increased by 0.005 after ATTRACT ranking, 0.015 after DFIRE ranking and 0.023 after iATTRACT refinement when evaluating on the 8 successful cases (excluding 2OT3 and 3CPH). Hence, scoring by DFIRE was overall beneficial in enriching the pool of near-native loop conformations, while iATTRACT refinement mostly optimized the packing of the loop interface and increased fnat.

7.3.3. Comparison to Rosetta loop modeling

We wanted to compare loopATTRACT's performance to a state-of-the-art loop modeling protocol. We therefore ran Rosetta next-generation KIC loop modeling on the 10 benchmark cases using the rigid-body superpositions as starting structures and evaluated the results by the same assessment criteria. Parameters for Rosetta loop modeling were taken from [427] and 2500 structures were generated for each case. The results are shown in Table 7.2 and Table C.2. For Rosetta, we found a success rate

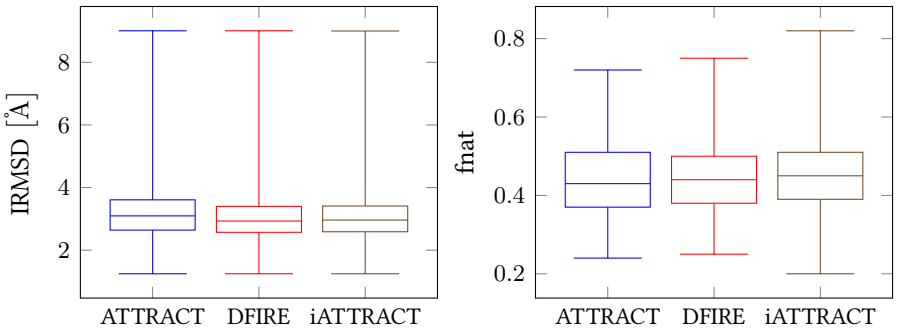


Figure 7.2. Comparison of IRMSD and fnat distribution for different stages of the loopATTRACT protocol. For each benchmark case, we evaluated the top 1000 rigid-body models ranked by their ATTRACT score, the top 1000 rigid-body models ranked by DFIRE score and the 1000 refined iATTRACT models. In total, the IRMSD and fnat distributions were evaluated with 10000 models. The whiskers mark the minimum and maximum of all the data.

of 80 % among the top-ranked 10 models. In 5 cases, the loop rebuilding even generated a CAPRI two-star quality structure among the top-ranked 10 models. The best IRMSD among the top 10 models generated by loopATTRACT was lower or equal (± 0.1 Å) to the best IRMSD among the top 10 models from Rosetta in 4 cases, in the remaining cases the IRMSD of the Rosetta models was lower. In terms of the best fnat among the top-ranked 10 models, loopATTRACT gave a better performance with higher (5 cases) or equal fnat (± 0.02, 2 cases) in 7 of the 10 cases. Similarly to loopATTRACT, improvements in IRMSD were not always accompanied by improvements in fnat (especially for 1BKD, see also Table C.2). When looking at all generated models, Rosetta achieved a significantly lower IRMSD in 4 and equal performance in 5 cases. For fnat, the overall sampling performance was very similar (3 cases where fnat was higher for loopATTRACT, equal performance in 5 cases and 2 cases where Rosetta achieved higher fnat). Both protocols failed to model the cases 2OT3 and 3CPH accurately. On average, a Rosetta run took 2682 CPU hours (which corresponds to 1072 CPU hours for generating 1000 structures). In contrast, the loopATTRACT protocol used only 20-70 CPU hours per case.

7.3.4. Refinement of ATTRACT-EM models

Recently, we showed that when using low-resolution cryo-EM maps for complex assembly (ATTRACT-EM), high precision rigid-body placements of the partners can always be obtained [96]. Furthermore, steric clashes in these models can be used to identify flexible regions undergoing conformational change upon binding [96].

Table 7.2. Results for interface loop modeling with Rosetta on 10 complexes from docking benchmark 5 [473]. For each complex, the PDB ID, the docking difficulty according to the classification in the benchmark [473], IRMSD and fnat of the initial rigid-body superposition, best IRMSD and fnat among the top-ranked and top-ranked 10 models and best sampled IRMSD and fnat are listed.

PDB	Difficulty	initial		top 1		top 10		all	
		IRMSD	fnat	IRMSD	fnat	IRMSD	fnat	IRMSD	fnat
1ATN	hard	4.54	0.35	2.32	0.47	2.17	0.47	1.50	0.79
1BKD	hard	3.15	0.53	3.05	0.46	1.77	0.56	1.60	0.60
1FQ1	hard	3.75	0.40	3.37	0.40	2.17	0.47	1.91	0.64
1LFD	medium	2.06	0.57	1.66	0.60	1.08	0.68	1.00	0.75
1PXV	hard	2.75	0.58	2.52	0.53	1.92	0.57	1.70	0.61
1R8S	hard	4.59	0.32	3.34	0.38	3.22	0.38	3.15	0.47
2NZ8	medium	2.22	0.44	1.88	0.43	1.87	0.55	1.80	0.59
2OT3	hard	2.55	0.49	2.57	0.48	2.55	0.50	2.54	0.51
3CPH	medium	2.13	0.53	2.47	0.45	2.39	0.51	2.00	0.55
3FN1	hard	3.74	0.36	2.06	0.46	1.78	0.53	1.67	0.58

Hence, ATTRACT-EM models can be further optimized by refinement procedures focusing on the identified flexible regions. Here, we tested loopATTRACT in such a hierarchical integrative modeling approach. We ran interface loop modeling on a subset of five cases using the top-ranked ATTRACT-EM model obtained with a 20 Å resolution cryo-EM density map [96]. The results are listed in Table C.3. Similarly to the refinement on the bound-unbound superpositions, the protocol was able to generate improved interface loop structures with respect to the initial structure and rank those models among the top 10 in four out of the five test cases (as in the previous benchmark, no improvement was found for 2OT3). This demonstrates that slight errors in rigid-body placement of the initial complex geometry can be tolerated by the interface loop modeling protocol. In terms of overall sampling, we found slightly worse results than for interface loop modeling on the bound-unbound superposition (Table 7.1).

7.4. Discussion

loopATTRACT is a fast protocol for generating interface loop structures on a given protein-protein complex geometry. The approach was tested on a benchmark set of 10 protein-protein complexes with long interface loops and yielded significant improvements among the top-ranked 10 models in 7 out of 10 cases. loopATTRACT also

gave promising results for loop rebuilding on complex geometries obtained from an integrative modeling approach based on low resolution cryo-EM maps (ATTRACT-EM). The success of the protocol underlines the benefit of integrating approaches from protein structure prediction and protein-protein docking.

The major bottleneck of the protocol is choosing the appropriate interface loops for remodeling; i.e., identifying loops that undergo conformational change upon complex formation. We previously showed that deep clashing atoms in ATTRACT-EM models correlate with the probability for significant motions [96] and thus can be used to select regions for flexible refinement. Loop flexibility could also be predicted by examining X-ray or NMR structural ensembles of the individual protein partners, from B-factors in crystal structures, from conformational changes in related complexes [478] and by analyzing the possible movements of the loops in molecular dynamics simulations. Loops that are part of protein-protein interfaces may be subject to evolutionary constraints and sequence alignments could possibly identify them. In addition to identifying the correct interface loop, the quality of the results may also depend on the exact choice of the loop anchors. Further tests will be necessary to quantify the influence of deviations in the anchors on the accuracy of the predicted interface loop conformation. The effect of remodeling multiple interface loops should also be explored in future work.

Despite good overall performance, loopATTRACT can be improved in several ways. First, in the setup tested here, the ensemble of loop conformations for docking was generated irrespective of the sequence. The loop models in the ensemble deviated on average by at least $2.27\,\text{Å}$ C_α-RMSD from the bound loop conformation. In test case 2OT3, the minimal RMSD was even $3.29\,\text{Å}$ causing most likely the failure of the protocol. The BriX loop library was already published in 2011 and constructing a new library might help to find more accurate loop conformations. A better selection of loop conformations could also improve the results by decreasing the search space and eliminating false positives. Still, the success of the protocol demonstrates that loop conformations are largely defined by the constraints imposed by the anchors. Hence, the phase space of loop conformations can be covered well by a few hundred models even for long loops.

Second, the scoring of near-native models in loopATTRACT still displayed considerable shortcomings. While we were able to rank improved models among the top 10, these top-ranked models were in most cases not the best sampled structures; i.e., models of lower IRMSD and higher fnat could be found at ranks higher than 10. One possible improvement might be to consider solvation effects in the scoring function. Parts of the interface loops are often exposed to the solvent and solvation should also have an impact on the preferred loop conformation. Molecular dynamics refinement in explicit solvent could help to improve the results, however since the

complexes are usually large, these approaches are still too computationally demanding for refinement of hundreds of structures. Furthermore, scoring in explicit solvent is very difficult due to large energy fluctuations in the system. Previously, Guerois and coworkers introduced an approach to score docking models by evolutionary information using the sequences from homologous complexes [11]. Since loops are often more variable in sequence than other parts of the protein, this might also be a promising approach for scoring of interface loop conformations. Furthermore, initial loop conformations could also be assessed with respect to their energy using the homologous sequences. Another possibility to improve the scoring (and the sampling) of the loopATTRACT protocol would be to incorporate experimental data. Contact and interface information from biochemical experiments (see Chapter 2) can be used during loop docking as distance or ambiguous interaction restraints [99, 96]. In addition, contact information for loops could also be derived from co-evolution analysis [330]. Models can also be scored against a cryo-EM map [94, 96] or a SAXS profile [392] to select near-native loop conformations. In general, any available ATTRACT feature can be easily included in the loopATTRACT protocol.

7.5. Conclusion and Outlook

Proteins often undergo large conformational changes upon binding. In many cases, interface loops rearrange by large backbone movements compared to the unbound structure or to the template used in homology modeling. Here, we presented an interface loop modeling protocol, loopATTRACT that can be used in a hierarchical docking approach for a-posteriori loop rebuilding. loopATTRACT collects an ensemble of loop conformations from a fragment database and uses this structural ensemble in a modeling protocol adapted from peptide-protein docking. Finally, the models are ranked using the DFIRE scoring function, which has been previously developed for protein structure prediction. The loopATTRACT protocol was tested on difficult cases from the protein-protein docking benchmark and achieved significant improvements in 7 out of 10 cases. loopATTRACT's performance was comparable to that of a state-of-the-art loop modeling protocol in Rosetta, while being far less computationally expensive. We also tested the approach for refining models derived from the ATTRACT-EM protocol and found that loopATTRACT yielded promising results in such a hierarchical integrative modeling framework. As illustrated by the most recent rounds of the blind prediction challenge CAPRI, modeling protein-protein complexes with a high degree of flexibility will become the rule rather than the exception. The loopATTRACT protocol in combination with other docking methodology in ATTRACT provides a valuable piece towards dealing with these challenging tar-

gets. In the future, interface loop modeling should be systematically tested as a part of multi-stage docking and integrative modeling protocols. The recently published benchmark for combined assessment of homology modeling and docking approaches [49] should provide an ideal testing ground.

8. SAXS Data Alone Can Generate High-Quality Models of Protein-Protein Complexes

Small angle X-ray scattering (SAXS)-driven modeling combines low-resolution data with computational modeling to predict the structure of biomolecular assemblies. A new protocol, ATTRACT-SAXS, has been developed and tested on a large protein-protein docking benchmark with simulated SAXS data. For 88% of the cases, high-quality solutions were generated using SAXS data alone without a physiochemical force field ((interface-RMSD $\leqq$ $2\,\mathring{\mathrm{A}}$ *or ligand-RMSD* $\leqq 5\,\mathring{\mathrm{A}}$*) and more than 30% native contacts). ATTRACT-SAXS gave significant improvements compared to previous approaches that filter by SAXS a-posteriori. When combining SAXS and interface properties for scoring, the protocol placed high-quality models in 79% of the successful cases among the top-ranked 100 clusters. ATTRACT-SAXS also gave excellent results when tested on experimental data if the native complex structure was compatible with the SAXS profile. Our results show that in principle, SAXS on its own can contain enough information for generating high-quality models of protein-protein complexes. The work presented in this chapter was published previously in similar form in* [392].

8.1. Introduction

Many proteins form complexes to carry out their biological function. These assemblies are involved in important cellular processes like signaling, transport and catalysis. Knowledge of the 3D atomic structure is vital for understanding their function and regulation. Integrative modeling approaches combine information from different sources such as X-ray crystallography, electron microscopy, cross-linking mass spectrometry, nuclear magnetic resonance spectroscopy (NMR) or small-angle X-ray scattering (SAXS) with computational modeling of the structure of biomolecular complexes. Integrative modeling methods have become viable alternatives for obtaining structural models and have already been applied successfully to a large vari-

ety of biological problems [353, 414, 481, 307, 355, 157]. Currently, efforts supported by the Worldwide Protein Data Bank [39] are underway to better store and share integrative models underlining their importance for the structural biology community [385].

Small-angle X-ray scattering (SAXS) is an effective way for obtaining low-resolution (10 Å to 30 Å) structural data for biomolecules in solution [340]. It has become increasingly popular in recent years [160]. In contrast to other techniques in structural biology, the molecules can be studied at near-physiological conditions and data can be collected in a few seconds on a well-equipped synchrotron line [305]. High-throughput SAXS can therefore help to examine protein-protein complexes and address current bottlenecks in studying interactions on the genome scale [198, 372].

Several integrative methods with SAXS data have been developed to date [341, 354, 398, 403, 219, 209, 491]. Most methods use a physiochemical or geometry-based force-field to generate a large number of possible models of the complex structure and applied SAXS data to filter a posteriori. The pyDockSAXS approach [354] used FTDOCK to generate rigid-body docking decoys. The FTDOCK models were then ranked by combining the pyDock energy-based scoring function and the fit to the SAXS data with CRYSOL. Likewise, Schneidman-Duhovny et al. presented the IDOCK method which samples docking models by the PatchDock method and then filters them by their capacity to describe the experimental data χ using the FoXS program [397, 399]. The filtered models are refined, clustered and finally ranked by a composite score considering both physiochemical properties and the fit to the SAXS data [403]. In addition, Karaca et al. explored the capacity of SAXS data as a scoring function and tested its performance in combination with the HADDOCK score [219]. Recently, support for SAXS data was also added to the ClusPro web-server [491]. Similarly to pyDockSAXS, this approach was based on generating models with an FFT-docking program and filtering them by their compatibility with the SAXS data. The top 1000 filtered models were finally rescored based on the PIPER-energy function alone (without the SAXS data) and clustered [491]. In summary, most of the previously published methods use SAXS data for a posteriori filtering and scoring.

The ATTRACT-SAXS approach described in the present study directly uses the SAXS data during an extensive sampling stage, repeatedly comparing generated models to the experimental profile. When testing the protocol on a large protein-protein docking benchmark, ATTRACT-SAXS yielded high-quality predictions for many complexes and showed improved performance compared to existing methods [403, 491].

8.2. Methods

The ATTRACT-SAXS protocol consists of the following stages. First, docking models are sampled exhaustively aiming to enumerate all solutions compatible with the experimental SAXS data. Then, the generated models are filtered according to a composite score considering both interface properties and fit to the SAXS data. The top-ranked 5000 models are finally clustered by pairwise ligand-RMSD and ranked by the average energy of their top-ranked 4 members.

8.2.1. Sampling stage

A flow chart of the individual sampling steps in the protocol is shown in Figure 8.1. We used a Monte Carlo based sampling scheme constrained by a so-called "atom density mask" which restricts the proteins to a certain region of space and prevents large overlaps between them (see Figure D.1 for an example). Note that the ATTRACT-SAXS protocol does not use a force field to guide the assembly of the complex, sampling is only driven by compatibility to the experimental SAXS data and soft steric repulsion. Initially, 6000 random starting positions of the protein partners were generated. These positions were optimized during 50 sampling iterations. In each iteration, the following steps were executed:

1. Monte Carlo (MC) optimization of the rigid-body placements of the proteins in the atom density mask.

2. Scoring of docking models by SAXS data.

3. Storing docking models with low E_{SAXS} in taboo pool and removing similar models from the sampling pool ("taboo clustering").

4. Swapping rigid-body placements between top-ranked 300 models ("swap-combine") and scoring by SAXS data.

5. Selection of top 6000 ranked docking models as input for next docking stage.

6. Adding previously stored docking models to the sampling pool (every 5 iterations).

On average, $\approx$ 15,000 docking models were stored per complex in this stage (Figure D.5). The entire sampling stage is typically completed in a few hours on a modern GPU.

Monte Carlo sampling For both proteins, up to 500 MC steps in rigid-body translational and orientational degrees of freedoms were performed. The step width was decreased during successive iterations. In the first iteration, an additional harmonic potential ("gravity") was imposed to pull the proteins towards the origin and into the atom density mask.

Atom density mask The atom density mask was generated from a low-resolution bead model calculated from the experimental SAXS data with the ATSAS 2.6.1 software suite [338]. The radial distribution function was obtained from datgnom [339], then 20 runs were performed with DAMMIF [140] in fast mode with default settings. The output from DAMMIF was averaged with DAMAVER [471] in automatic mode. The averaged bead model (damaver.pdb) was centered at the origin and finally converted with the ATTRACT tool `pdb2mask.py` into an atom density mask imposing an extra margin of two voxels per occupied voxel. A coarse and a fine mask were extracted using voxel dimensions of $10\,\text{Å}$ and $5\,\text{Å}$ respectively. The $10\,\text{Å}$ mask is used during the first three iterations of the protocol, in later iterations the $5\,\text{Å}$ mask is used. Note that these masks only impose very rough limits on the overall geometry of the complex. An example of atom density masks for a protein-protein complex is shown in Figure D.1.

Taboo clustering The generated docking models ("sampling pool") were compared to a set of previously stored docking models ("taboo pool") by clustering the models against this set by pairwise C_α ligand-RMSD with a cutoff of $6\,\text{Å}$. A hierarchical clustering approach was used. Similar models were removed from the sampling pool. If the E_{SAXS} score of a similar new model was lower than the score of the stored model, the old model was replaced. Non-similar models with E_{SAXS} lower than 3 were added to the taboo pool as long as the total number of stored models did not exceed 10,000. Once more than 10,000 docking models had been collected, models of higher E_{SAXS} were replaced by new models unless all the models had $E_{\text{SAXS}} < 2$. In this case, the taboo pool was expanded.

Swap-combine For increased sampling, placements of the proteins in the top-ranked 300 docking models from the sampling pool were exchanged between different models. The resulting 90,000 models were scored by E_{SAXS} and the top-ranked 6000 models were selected as input for the next docking iteration.

8.2.2. SAXS score

For each docking model, the scattering intensity $I(q)$ was calculated using the density histogram approximation of the Debye formula [102, 151]

$$\begin{aligned} I(q) &= \sum_{i,j=1}^{N} f_i(q) f_j(q) \frac{\sin(q d_{ij})}{q d_{ij}} \\ &\approx \sum_{i=1}^{n_{\text{bins}}} \widehat{F}(q, r_i) \frac{\sin(q r_i)}{q r_i} \end{aligned}$$

with $q = \frac{4\pi \sin\theta}{\lambda}$ momentum transfer and $\widehat{F}(q, r_i) = \sum_{d_{jk}=r_i} f_j(q) f_k(q)$ summed form factors of all atom pairs at distance r_i. The individual form factor f_i for each atom is given by

$$f_i(q) = \underbrace{f_i^{\text{vac}}(q)}_{\text{Protein in vacuo}} - \underbrace{C_1(q) f_i^{\text{sol}}(q)}_{\text{Displaced solvent}} + \underbrace{c_2 s_i f_i^{\text{w}}(q)}_{\text{Hydration layer}}$$

with s_i fraction of SASA of atom i [431, 141, 85]. Form factors were taken from the IMP SAXS module [378, 397, 399, 398]. The parameters C_1 and c_2 were set to 1.0 and 0.0 for all calculations [398]. We further assumed a uniform scaling

$$f_i(q) = E(q) \times f_i(0)$$

with the approximation function $E(q) = e^{-bq^2}$ [137, 399]. This then yielded the final formula for the intensity $I(q)$

$$I(q) = E(q)^2 \sum_{i=1}^{n_{\text{bins}}} F(r_i) \frac{\sin(q r_i)}{q r_i}.$$

The squared distances were binned with a bin size of $0.5\,\text{Å}^2$ [399] up to a maximum squared distance of $10.000\,\text{Å}^2$ and $40.000\,\text{Å}^2$ for smaller and larger proteins respectively.

For the experimental profile, we simulated a weighting function $w(q)$ using the formula for experimental error estimates in FoXS [397]

$$w(q) = 0.03 \times I(q) \times 5(q + 0.001) \times \left(\left|\tfrac{p}{10} - 1\right| + 1\right)$$

where p is a random number drawn from a Poisson distribution with a mean value of 10. The calculated intensity curves were scaled to the experimental curves and the

discrepancy was evaluated

$$E_{\text{SAXS}} = \sqrt{\frac{1}{N_p}\sum_{i=1}^{N_p}\left[\frac{I_{\text{exp}}(q_i) - c \times I_{\text{calc}}(q_i)}{w(q_i)}\right]^2}$$

with the scaling factor c

$$c = \frac{\sum \frac{I_{\text{exp}} I_{\text{calc}}}{w^2}}{\sum \frac{I_{\text{calc}}^2}{w^2}}$$

chosen to minimize E_{SAXS}. Note that this score is very similar to the widely used χ test. However, we decided to use a common weighting function instead of the experimental errors $\sigma_{\text{exp}}(q)$ used in χ to calculate SAXS scores under similar conditions for different complexes. We evaluated E_{SAXS} with the new ATTRACT tool `saxs-score` that calculates the pairwise distance histogram on the GPU and $I(q)$ and E_{SAXS} on the CPU. To speed up the calculations, the intra-protein distances are precalculated and only the inter-protein distances have to be evaluated for each docking model. For a medium sized complex and 50 q data points, the score was evaluated in $\approx$ 0.8 ms per structure. For comparison, the program FoXS required $\approx$ 96 ms per structure on the same complex. This fast tool allowed us to design a protocol with multiple iterations of sampling and rescoring against the experimental data. Note that the E_{SAXS} score could be easily extended to fitting against multiple data sets (e.g., for subcomplexes).

8.2.3. Filtering

All docking models in the taboo pool were filtered by a composite score:

$$E_{\text{total}} = E_{\text{interface}} + w_{\text{SAXS}} E_{\text{SAXS}}.$$

The interface energy $E_{\text{interface}}$ was calculated with an empirical step potential trained on 164 protein-protein docking cases from benchmark 4.0 [200] evaluated within cutoffs of $4\,\text{Å}$ and $6\,\text{Å}$. The set of protein-protein complexes was divided in a training and a test set of 140 and 24 complexes respectively. The step potential represents the proteins in a grouped-all-atom model and was derived by Monte Carlo Annealing with 5-fold cross-validation using a ziczac annealing scheme. As a target function, we optimized the summed linearly weighted ranks of near-native solutions which were additionally scaled by CAPRI quality [389]. The weighting factor w_{SAXS} of the composite score was optimized by systematic exploration of values in a range of 0.1 to 1000.0. We observed similar rankings for a range of parameters and finally set w_{SAXS} to 300.0. For cases in the experimental benchmark set with symmetry, we used an additional symmetry term in E_{total} [94].

8.2.4. Clustering and Final Ranking

The top-ranked 5,000 models were clustered by pairwise ligand-RMSD with a cutoff of 6.5 Å and a minimum cluster size of 1. The models were fitted onto the receptor protein structure and the RMSD between the ligand backbone atoms was evaluated. The scores of the top-ranked 4 cluster members were averaged and the clusters were ranked by the average score.

8.2.5. Test sets

The protocol was tested on two benchmark sets. The first benchmark consisted of 226 protein-protein complexes from protein-protein docking benchmark [473] using simulated SAXS data. In the second benchmark, we ran ATTRACT-SAXS with experimental data (11 cases).

Docking benchmark with simulated SAXS data We used the new version of the protein-protein docking benchmark [473] with 226 protein complexes (cases with internal symmetry and alternative binding sites were merged). The structures were downloaded from the PDB. Unbound protein structures were aligned to the bound structures with FATCAT [501]. Residues in the unbound form were renumbered, mutated and/or removed to match the bound form for easy evaluation of RMSD criteria. Missing side chain heavy atoms were built with PDB2PQR [110] when at least the backbone atoms were present, we did not add any missing residues. The proteins were converted to the ATTRACT atom type format with the ATTRACT tools `aareduce` and `reduce` (see Chapter 4). The complexes were checked for internal symmetry of the protein partners and alternative symmetry solutions were also considered for RMSD evaluation. We classified the difficulty of the docking cases as "rigid-body", "medium" and "hard" based on the IRMSD between the interface superposition of the bound and unbound structures [473]. In addition, we classified the docking cases by calculating the CAPRI star quality of the whole rigid-body and the interface superposition of unbound and bound protein structures. Cases that did not achieve CAPRI two-star quality or better in both superpositions were classified as "impossible" (16 out of 226 cases: 1ATN, 1BGX, 1DE4, 1E4K, 1F6M, 1FAK, 1GP2, 1H1V, 1JMO, 1NW9, 1Y64, 2HMI, 2I9B, 2O3B, 2VIS, 3G6D) and these cases were not considered for analysis (except for comparison to other methods).

Test cases with experimental SAXS data We tested our method on 11 cases using experimental SAXS data. The cases are listed in Table 8.1. The SAXS profiles [71, 318, 350, 525, 324, 415] were downloaded from the SASBDB (www.sasbdb.org)

[455] and the BIOISIS (www.bioisis.net) [198] databases. In most cases, corresponding crystal structures were annotated, in the others we used BLAST [210] to identify matching structures in the Protein Data Bank. The protein complex structures [358, 71, 208, 289, 74, 230, 350, 525, 324, 415, 434] were downloaded from the PDB. Missing side chains were built with PDB2PQR [110] and the structures were converted to the ATTRACT atom type description with the ATTRACT tools `aareduce` and `reduce`. For the tetrameric complexes, we performed two-body docking of the two dimers. For bovine serum albumin and ovotransferrin, we cut the protein in two parts and docked these domains, for the TG2-antibody we used the light and heavy chain as docking partners.

8.2.6. Simulating SAXS data and processing of experimental SAXS data

SAXS profiles were simulated using the bound protein-protein complex structure with the program FoXS [397] for a q range from $0.01\,\mathrm{\AA}^{-1}$ to $0.5\,\mathrm{\AA}^{-1}$. The parameters for excluded volume and hydration layer scattering were set to 1.0 and 0.0. The number of points in the profile was set to 50. All other parameters were kept at their default values. Gaussian noise with a standard deviation of 2% of the intensity was added to the curves to simulate experimental conditions [41]. For the experimental SAXS curves, we simulated the weighting function $w(q)$ and excluded data points where the experimental error was two times larger than the assigned weight $w(q)$ (we assumed that the size of the error reflects the amount of noise present in the data).

8.2.7. Assessment criteria

For benchmarking the protocol, the final docking models were evaluated using IRMSD, LRMSD and fnat criteria established in the blind protein-protein docking challenge CAPRI [303, 257, 260, 259] (see Chapter 3). The docking models were classified by the following quality criteria as:

- high/three star if (IRMSD $\leqq 1$ or LRMSD $\leqq 1$) and fnat $\geqq 0.5$
- medium/two star if (IRMSD $\leqq 2$ or LRMSD $\leqq 5$) and fnat $\geqq 0.3$
- acceptable/one star if (IRMSD $\leqq 4$ or LRMSD $\leqq 10$) and fnat $\geqq 0.1$.

A docking case was considered a one star/two star/three star success if at least one model of one star/ two star/ three star quality or better was found among the top N solutions/clusters.

8.3. Results

A new integrative modeling approach for predicting the 3D-structure of protein-protein complexes starting from unbound partner structures and small-angle X-ray scattering (SAXS) data, ATTRACT-SAXS, has been developed using the ATTRACT docking engine [506, 296, 101, 94, 393, 99]. A schematic overview of the protocol is given in Figure 8.1. It consists of an exhaustive multi-start sampling stage during which a large number of docking models compatible with the experimental data are generated (see Methods for details). The sampling is driven by maximizing the fit to the experimental scattering intensities and avoiding excessive sterical clashes, however, no specific force field is included. In addition, a new goodness-of-fit criterion, E_{SAXS}, similar to the well-known χ statistic, but with a different weighting of the individual data points has been employed (see Methods). After the sampling stage, the solutions are filtered by a composite scoring function considering both E_{SAXS} and interface properties. The interface properties are evaluated by a new empirical atomistic step potential which has been trained on known protein-protein complexes [389]. Finally, the top-ranked 5,000 models were clustered and ranked by the average energy of their top-ranked 4 cluster members. We tested the ATTRACT-SAXS approach on 226 protein-protein complexes from the recently published docking benchmark 5.0 [473] using simulated SAXS profiles (see Methods) and on 11 cases using experimental data.

8.3.1. Characteristics of SAXS data as scoring function and implications for protocol design

In order to evaluate the specificity of the scoring function E_{SAXS}, we first used the simulated SAXS profiles for the docking benchmark to evaluate the nearest-native solution; i.e., the rigid-body superposition of the unbound protein structures on the bound complex. The obtained scores are shown in Figure 8.2 (a) and demonstrate only slight sensitivity to protein conformational change (most conformational changes are well below the resolution of SAXS). For complexes with IRMSD $< 2\,\mathrm{\AA}$, 81% had an E_{SAXS} score under 1.5, 96% under 2, and all of them under 3. Therefore, in theory, one can capture the nearest-native solution by sampling the entire pool of all possible solutions below a certain E_{SAXS} threshold (e.g., $E_{\mathrm{SAXS}} < 3$). However, in practice, this is only feasible if the following two conditions are met: a) The E_{SAXS} score must have good discriminative power between nearest-native and non-native solutions, else the pool of solutions would be too large; b) Since sampling is finite, we can only approach the nearest-native structure. Therefore, the discriminative power of the scoring function must have a funnel-like behavior, extending to near-native

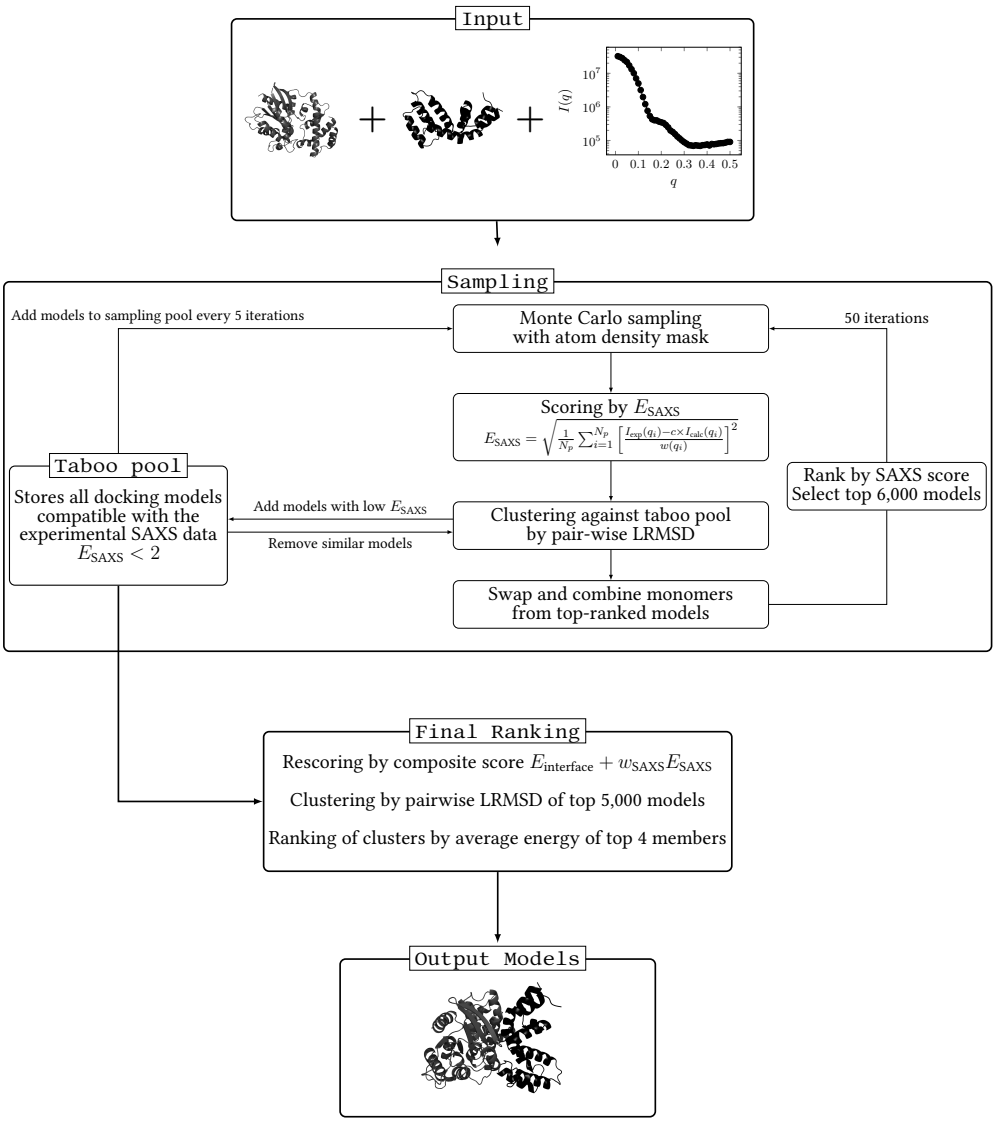
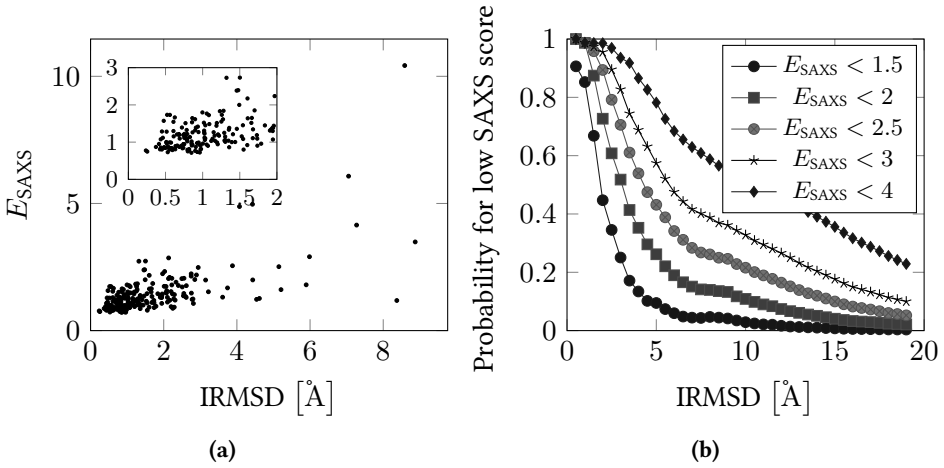


Figure 8.1. Overview of the ATTRACT-SAXS docking protocol.



(a) **(b)**

Figure 8.2. Accuracy and specificity of SAXS data as a scoring function in docking. (a) SAXS scores vs. IRMSD for the rigid-body superposition of the unbound protein structures on the bound complex for all 226 cases in the benchmark. (b) Probability for a docking model of a given IRMSD to have E_{SAXS} below a certain threshold. Results for different thresholds are shown.

structures in general. We assessed the discriminative power of E_{SAXS} by ranking a large number of decoys (near-native and non-native) obtained by standard rigid-body docking with ATTRACT [506, 99] for each complex. We calculated the probability for a docking model of given IRMSD to have an E_{SAXS} score below a certain threshold (Figure 8.2 (b)). We found that near-native models had a high probability of having a low E_{SAXS} score. With a strict threshold ($E_{\text{SAXS}} < 1.5$), this probability shows a sharp decline as a function of IRMSD. Even so, the probability does not fade to zero for large IRMSD values: in many cases, non-native solutions with a similar or even lower score existed (an example can be found in Figure D.2). Still, only a small fraction of non-native solutions yielded a good fit. Therefore, at strict thresholds, SAXS data has excellent discriminative power. However, for looser thresholds, the discriminative power quickly fades, the probability peak greatly broadens and only slight enrichment of near-native compared to non-native models was observed for a cutoff of for example $E_{\text{SAXS}} < 4$. Therefore, the appropriate E_{SAXS} threshold strongly depends on the sampling precision. At the one hand, the threshold should be as small as possible to maximize discriminative power, while at the same time, the probability peak must be broad enough to capture at least some generated near-native structures. We based the selection criteria for the sampling pool on these findings (see Methods for details).

8.3.2. Docking benchmark results

In order to interpret experimental results or to guide experimental studies, the accuracy of a docking prediction must be sufficiently high. We choose the CAPRI two-star criterion to assess the performance of ATTRACT-SAXS. We believe that structures calculated from experimental data should be held to higher standards than structures derived from ab-initio docking and therefore consider the CAPRI one-star criterion, which was used to assess the performance of previous SAXS-driven integrative modeling approaches, too lax. A two-star model has to identify at least 30% of the native intermolecular contacts, whereas a one-star model only needs to identify 10%. Correct intermolecular contacts in the model are vital to predict the effect of mutations. We limited the analysis to the 210 cases in protein-protein docking benchmark 5 [473] where a two-star model by rigid modeling is in principle possible (see Methods).

Figure 8.3 shows the docking success rate achieved among the top-ranked final clusters and among all sampled models (see also Table D.1). Overall, the protocol generated a two-star model for 88% of the 210 protein-protein complexes (i.e., 185 complexes). This sampling success rate is significantly higher than that obtained with standard ATTRACT docking [99], even with a 6-fold increased number of starting positions (77%, data not shown). After rescoring, clustering and ranking the clusters by the average energy of their top-ranked 4 members, the top-ranked 100 clusters contained at least one two-star cluster for 79% of the successful docking cases (148 complexes, Figure 8.3). Figure 8.4 illustrates docking models from top-ranked two-star clusters. For 96% of the successful cases with internal symmetry (27/28), the alternative solution was also identified at least to one-star precision, underlining the success of the exhaustive sampling stage (data not shown). We analyzed the dependence of the sampling success rate on the number of iterations during the sampling stage and found that after 40 iterations, the sampling had converged (Figure 8.5). We further compared the contributions of the different scoring terms in E_{total} (see Methods) by ranking the docking models by either the composite score E_{total} or E_{SAXS} and $E_{\text{interface}}$ separately (Figure D.4). E_{SAXS} alone already gave a good ranking, however, combination with $E_{\text{interface}}$ further improved the ranking of two-star models within the top 100 solutions compared to the individual terms. This confirms that the different terms yield complementary information for identifying near-native solutions.

We also evaluated the success rate by complex type and docking difficulty (Figure D.3). The protocol performed better for enzyme-inhibitor and antibody-antigen complexes than for the other complexes (top 10 two-star success rates of 49, 50 and 35%), although in terms of overall sampling, the results were similar. Furthermore, the differences were not as large as for other methods that employed a force field for

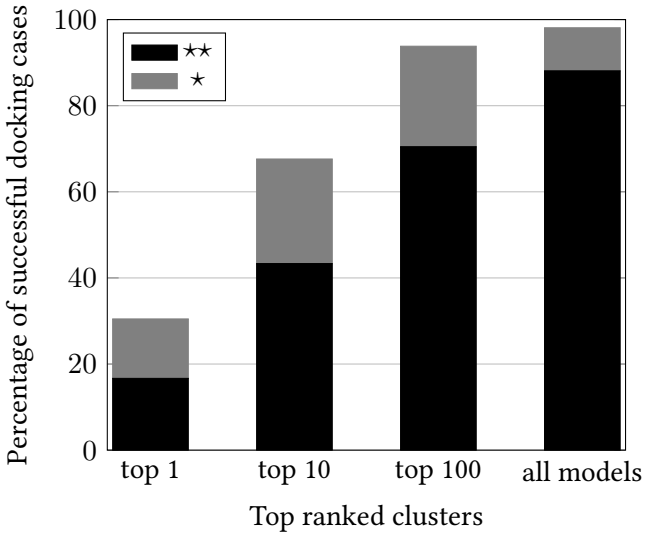
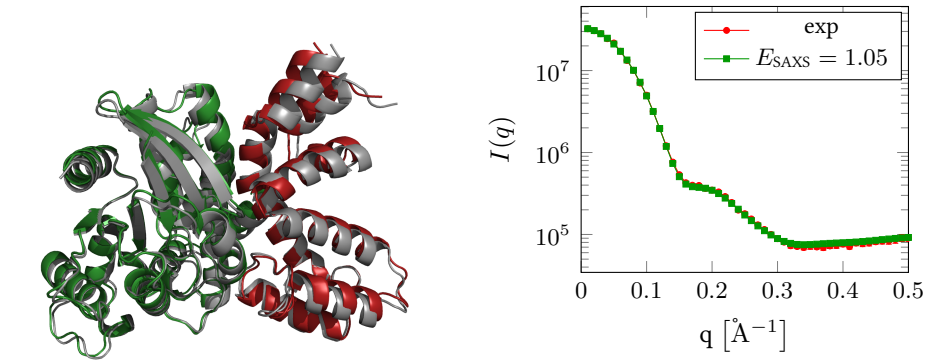
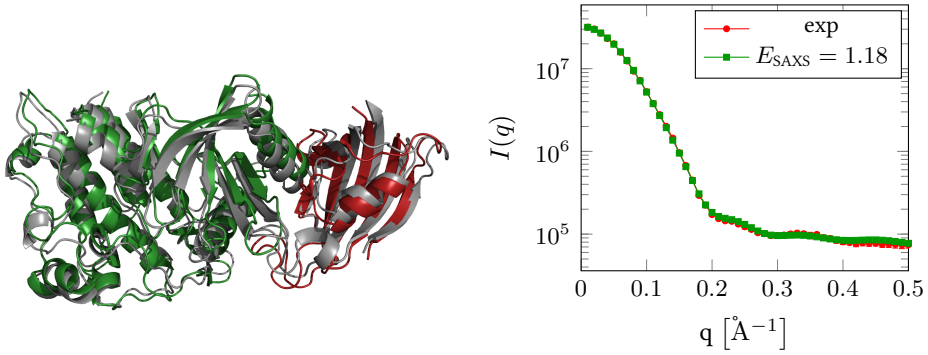


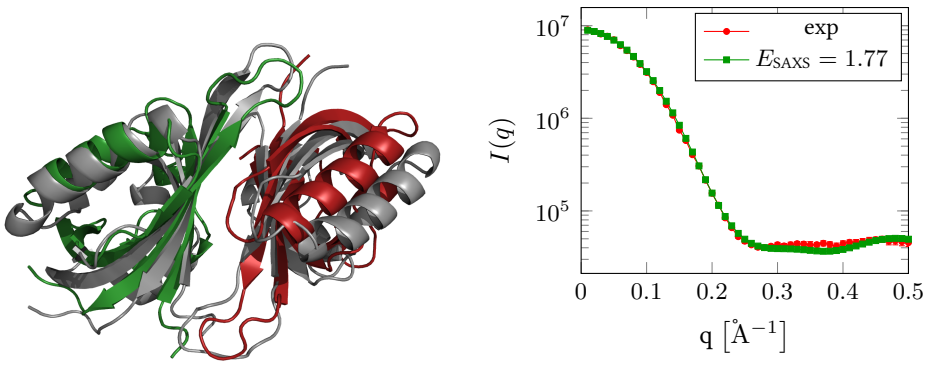
Figure 8.3. Docking results for ATTRACT-SAXS on 210 complexes from protein-protein docking benchmark 5.0 using simulated SAXS data. A cluster is considered a CAPRI one-star/two star hit if any of its top-ranked 4 members is at least of one star/two star quality. "All models" denotes the success rate considering all structures collected during the sampling stage.



(a) 2GTP, rigid-body, 1, 0.9 Å, 1.7 Å, 0.76



(b) 1B6C, medium, 1, 2.1 Å, 3.5 Å, 0.77



(c) 3F1P, hard, 4, 1.9 Å, 3.2 Å, 0.67

Figure 8.4. Examples of two-star docking models generated by ATTRACT-SAXS. The docking model is drawn in green and red, the crystal structure is shown as a reference in gray. The calculated intensity curve of the docking model is scaled to the simulated SAXS curve of the bound complex. For each case, the PDB ID of the bound complex, the docking difficulty and the cluster rank, IRMSD, LRMSD and fnat of the model are listed.

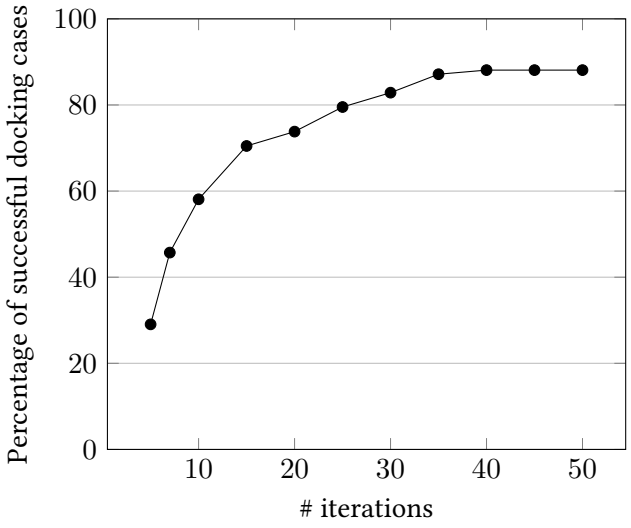


Figure 8.5. Sampling success rate for ATTRACT-SAXS vs. number of sampling iterations. The data were evaluated on 210 complexes from protein-protein docking benchmark 5.0 using simulated SAXS data. A docking case was considered successful if at least one CAPRI two-star model was generated.

sampling (e.g., [398]); force fields might be biased towards a certain complex type. Instead, this difference more likely reflects that the first two categories contain a larger fraction of rigid-body cases. When analyzing by docking difficulty, we found that the overall two-star success rate dropped from 91% for the rigid-body to 68% for the hard cases. Interestingly, for complexes of medium difficulty, we achieved an overall success rate very close to that of the rigid-body cases. Hence, in terms of sampling, the results demonstrate a robustness of the protocol to moderate conformational change. However, in terms of scoring the final models, we clearly saw a dependence on docking difficulty and further efforts to improve the ranking of near-native solutions are necessary.

Out of the 210 cases, 24 failed to sample a two-star solution at all. For half of these cases, we detected a scoring problem due to conformational changes: the superposition of the unbound protein structures on the bound complex had $E_{\mathrm{SAXS}} > 2$. For the rest, in 6 cases, either solutions very close to two-star precision or solutions of two-star quality in terms of RMSD criteria but not fnat were sampled. For these cases, additional sampling or refinement could be beneficial. In the remaining 6 cases, the interface on one of the partners was identified correctly by the top-ranked solutions, but the orientation of the second partner was wrong. In three of these cases, the

ligand protein is rather small compared to the receptor protein and of a spherical shape which made scoring with SAXS data challenging [354, 398] . Therefore, since failures are rare and mostly scoring-related, these results suggest that the sampling of our protocol is close to optimal (cf. Figure 8.5).

8.3.3. Comparison with other SAXS-driven methods

We wanted to compare ATTRACT-SAXS to previously published methods. The IDOCK approach [403] was tested on 176 complexes from protein-protein docking benchmark 4.0 [200] using simulated SAXS data. Success criteria based on IRMSD and LRMSD only were used to classify docking solutions as acceptable, medium and high [403]. For comparison, we adopted the same criteria and limited our evaluation to 174 complexes (docking case 1N8O from benchmark 4 is not available in benchmark 5; we merged docking cases 1OYV and 9OYV). For generating solutions of medium and high quality, Schneidman-Duhovny et al. obtained success rates of 13, 27 and 53% for the top-ranked 1, 10 and 100 clusters respectively. ATTRACT-SAXS yielded higher two-star success rates (16, 42 and 68%), especially for the top 10 and top 100 clusters. Also in terms of top 100 one-star success rate, ATTRACT-SAXS clearly surpassed IDOCK's performance (90% vs. 77%). Especially for docking cases of medium and hard difficulty, we found improved success rates (top 10 one-star success rate: 45% vs. 26%). Interestingly, in the initial PatchDock sampling, IDOCK achieved an overall two-star success rate of 82%, close to that of ATTRACT-SAXS (86%). However, after applying the SAXS filter, the success rate dropped to 73%. Since a high sampling precision is crucial when using SAXS data [398], structures, which were of high CAPRI quality but suboptimal for SAXS due to possibly small rotations/translations, were filtered out. This underlines the importance of using SAXS data (i.e., the scoring function) already in the sampling stage as also recommended by a recent study [454]. We further compared ATTRACT-SAXS to the recently published CLUSPRO-SAXS server [491] which was tested on 49 protein-protein complexes [70] with simulated data. When evaluating again by RMSD criteria only, CLUSPRO-SAXS placed a medium/high-quality solution among the top-ranked 10 clusters in 15 out of the 49 cases (31%) [491]. For the same subset of complexes, we obtained a top 10 two-star success rate of 41%.

8.3.4. Comparison with contact-driven modeling

It is also interesting to compare ATTRACT-SAXS to integrative modeling approaches in ATTRACT driven by other types of experimental data. In a recent study, we systematically compared three paradigms using perfect interface information, perfect

contact information and low-resolution cryo-EM maps [96]. The integrative modeling protocols were evaluated on 157 protein-protein complexes from benchmark 4.0 [200]. For contact and interface data, we extracted *all* atom-atom contacts within a cutoff of 5 Å from the bound complex; the data set contained on average 476 contacts per complex. The results for true interface and true contact docking thus represent an upper limit to what can be achieved using real data from e.g. NMR or cross-linking mass spectrometry experiments. We derived a maximal two-star success rate for true contact docking of 94% (148/157). The ATTRACT-SAXS overall two-star success rate was 91% (143/157). In terms of sampling, ATTRACT-SAXS, rather surprisingly, was close to the contact-driven protocol even when using *perfect* contact information.

8.3.5. Test cases with experimental data

We then tested the ATTRACT-SAXS protocol on 11 cases for which experimental SAXS profiles and crystal structures of the proteins were available (Table 8.1, see Methods for details). For each case, we simulated the weighting function and discarded parts of the data which were too noisy (see Methods). We also ran the protocol with simulated SAXS profiles and without SAXS data using the standard ATTRACT rigid-body docking protocol [99] for comparison. The results are summarized in Table 8.2.

In 6 of the 11 cases, the score calculated by comparing the crystal structure to the experimental SAXS profile was large ($E_{\mathrm{SAXS}} > 2.5$). These cases are essentially impossible and no CAPRI two-star solutions were generated. Still, they should not all be considered as failures: in two of these cases, the protocol correctly predicted the poor correspondence of the structures to the experimental data, since no models at all were produced in the sampling stage. For the 5 cases with good agreement between crystal structure and SAXS, ATTRACT-SAXS successfully generated two-star quality models and placed them in 4 out of 5 cases among the top-ranked 20 clusters. As a control, we ran the protocol with SAXS profiles simulated in the same way as for the docking benchmark. As expected from the good fit of the native structures ($E_{\mathrm{SAXS}} \leqq 1.5$), the protocol now succeeded in generating two-star quality predictions for all but one of the cases and ranked them even among the top 20 clusters.

In contrast to the docking benchmark, for this test set, only the bound forms of the proteins were available. Docking protocols using a force field during sampling typically perform extremely well in these situations, but struggle when structural noise caused by conformational change is introduced. This is a well-known problem in the docking field. Since ATTRACT-SAXS does not employ a force field during sampling (only a repulsive atom density voxel term), our results are not biased to-

Table 8.1. Test cases with experimental SAXS data. For each case, the name of the protein, the symmetry of the complex, the PDB ID, the ID of the SAXS database entry, the largest q value in the original and the weighted/filtered dataset used during docking are listed. We also calculated the percentage of missing residues in the crystal structure.

PDB	Name	Database ID	Type	$q_{\max}$ [Å^{-1}]	$q_{\max}$ [Å^{-1}]	Missing [%]
4V07	pUL26N serine protease	SASDA58	Dimer	0.443	0.131	9.2
3K3K	PYR1	1PYR1P	Dimer	0.321	0.321	8.3
1ZAH	Aldolase	SASDA68	Tetramer	0.496	0.494	0.3
4ZD3	Anti-TG2 antibody	SASDA28	Monomer	0.397	0.346	8.2
1RYX	Ovotransferrin	SASDAA2	Monomer	0.601	0.599	2.7
2R15	Myomesin-1 My12-My13	SASDAK5	Dimer	0.446	0.216	0.7
3V03	Bovine serum albumin	SASDA32	Monomer	0.601	0.600	4.4
3F7L	Superoxide dismutase	APSODP	Dimer	0.614	0.614	0.0
4W6Z	Alcohol dehydrogenase	SASDA52	Tetramer	0.601	0.599	0.3
4BLC	Catalase	SASDA92	Tetramer	0.601	0.597	5.3
1FA2	Beta-amylase	SASDA62	Tetramer	0.601	0.598	0.2

wards bound-bound docking. To illustrate this bias, we carried out standard ab-initio rigid-body docking [99] without experimental data, using the ATTRACT force field [506, 129]. For the four tetramers, the rigid-body docking of dimers failed because the dimers are intertwined strongly and correct rigid placement involves overcoming clashes during the docking trajectory. But for 6 of the 7 remaining cases, ab-initio docking generated two-star quality solutions and placed them among the top-ranked 20 models in 83% of the successful cases (we did not cluster the solutions). Unfortunately, a benchmark with experimental SAXS data and crystal structures for bound and unbound proteins is unavailable to date. Creating such a benchmark would be very important for properly testing SAXS-driven integrative modeling approaches in the future.

8.4. Discussion

Here, we developed a new SAXS-driven integrative modeling approach in ATTRACT, ATTRACT-SAXS. When tested on a large benchmark of protein-protein complexes [473] using simulated SAXS profiles, overall, ATTRACT-SAXS generated CAPRI two-star quality models for 88% of the docking cases. For 79% of the successful cases, the protocol placed near-native solutions among the top-ranked 100 clusters and for 49% even among the top-ranked 10 clusters. Using only the CAPRI one-star quality criterion, a success rate of 75% was achieved for the top-ranked 20 clusters. Hence, when considering the top-ranked 4 members of each cluster, only 80 models need to be

Table 8.2. Docking results for 11 experimental cases. For each case, the PDB ID, E_{SAXS} for the native structure in comparison with the experimental and the simulated SAXS profile and the docking results are listed. For each docking experiment, we evaluated the rank of the first two-star solution and the IRMSD, LRMSD and fnat of the best sampled model. † denotes cases in which no prediction was made; i.e., no model compatible with the SAXS data was sampled.

	Experimental SAXS data			Simulated SAXS data			Standard docking	
PDB	E_{SAXS}	Rank	Best sampled	E_{SAXS}	Rank	Best sampled	Rank	Best sampled
4V07	1.2	13	0.9/2.2/0.64	1.3	16	1.5/4.1/0.70	17	0.9/2.1/0.66
3K3K	1.8	10	1.7/4.8/0.91	1.2	1	1.5/3.4/0.72	8	0.7/2.2/0.96
1ZAH	2.1	1	1.5/4.2/0.66	1.3	1	1.2/1.7/0.62	-	4.2/9.2/0.04
4ZD3	2.4	1	1.1/3.0/0.90	1.5	1	1.7/4.4/0.61	1	0.4/1.2/0.92
1RYX	2.5	395	1.1/3.2/0.69	1.0	13	1.5/4.3/0.58	31598	1.3/4.5/0.58
2R15	2.7	-	6.3/38.5/0.09	1.4	-	2.6/11.6/0.47	-	2.5/15.3/0.44
3V03	3.4	-	2.1/3.5/0.20	1.2	1	1.4/2.6/0.73	1	0.4/1.0/0.78
3F7L	3.8	-	9.6/31.6/0.0	1.0	1	1.2/2.9/0.76	9	0.3/1.0/0.98
4W6Z	4.8	†	-/-/-	0.9	1	0.4/1.1/0.95	-	12.8/30.2/0.0
4BLC	6.0	-	10.8/26.7/0.02	1.0	1	0.7/1.5/0.67	-	23.3/49.7/0.0
1FA2	7.8	†	-/-/-/	1.5	1	0.6/2.4/0.71	-	5.9/12.8/0.08

inspected to detect a one-star quality solution with high probability which can then be validated by additional experimental data. The detected one-star model can be used to quickly filter the remaining models to identify the two-star quality cluster.

When testing ATTRACT-SAXS on a set of 11 complexes with experimental SAXS data, the protocol gave excellent results, if the crystal structures corresponded well to the experimental data. However, in 6 cases, the calculated scattering curve of the crystal structure differed significantly from the SAXS profile, much more than would be expected by simple conformational change. Consequently, this caused the protocol to fail, although it correctly identified the discrepancy in some cases. We discuss mixing/ensemble errors, systematic errors and errors resulting from the forward model as sources of this discrepancy.

In many cases, protein molecules do not adopt a single conformation but are rather represented by a structural ensemble. Protein flexibility; e.g., different loop conformations in different scattering particles, contributes to the mixing error (although this flexibility can be largely modeled by Gaussian noise [41]). Also, the dynamic equilibrium between bound proteins and free monomers cause SAXS curves being a mixture of the scattering of the individual monomers, the bound complex, and possibly intermediates. Knowing the structures of the monomers, the binding constant and the concentration of the sample, it should be possible to improve the scoring by fitting a weighted average of the intensity profiles of the monomers and the dock-

ing model. Furthermore, it might be interesting to compare an ensemble of docking models to the SAXS data instead of just a single model [41, 338, 374, 243]; e.g., in the case of fuzzy binding. Unfortunately, ensemble methods with SAXS may carry a serious risk of overfitting [38].

For integrative modeling, previously resolved structures of the individual proteins or homology models are used. In many cases, not all residues are resolved in the crystal structure, however, all residues contribute to the experimental scattering profile. Missing atoms are a source of systematic errors: for structures with more than 10% of missing residues, the calculated scattering profiles hardly fit to the experimental SAXS curve (the same is true for simulated profiles [354, 398]).

Last, we discuss errors resulting from the forward model; i.e., the calculation of the intensities from the atomic coordinates. The assumptions on which the forward model is built may not always be fulfilled and the resulting discrepancies may impair the scoring. The Debye formula assumes the absence of interactions between the individual molecules and hence the absence of aggregation. Not considering the scattering of the hydration layer; i.e., ordered water molecules on the protein surface, in cavities or at the interface, could also be a large source of error (e.g. for SASBDB entry SASDA82). Several methods for SAXS profile computation consider the solvent layer implicitly by fitting for example the solvent density or the width of the hydration shell to improve the overall fit to the experimental data [397, 431, 27, 438], however, such an approach that fits parameters for every docking model is not suitable for comparison and ranking of docking models, for which we need to calculate the scores under identical conditions [398]. The hydration shell can also be considered explicitly [336, 164, 469, 351, 274, 363, 68], but these methods are computationally expensive and therefore not yet suitable for the large-scale sampling approach used in ATTRACT-SAXS.

The success of ATTRACT-SAXS is based mainly on sampling improvements compared to standard docking. By eliminating the force field and allowing a certain degree of overlap between the protein partners, the protocol becomes less sensitive to conformational change at the interface. In addition, we introduce E_{SAXS}, an alternative statistic that does not rely on experimental errors and allows a fixed-threshold cutoff. Scoring by SAXS data requires a comparison between the experimentally obtained intensities and those computed from a model. For SAXS data, commonly the χ test is used to evaluate the similarity

$$\chi = \sqrt{\frac{1}{N_p} \sum_{i=1}^{N_p} \left[\frac{I_{\text{exp}}(q_i) - c \times I_{\text{calc}}(q_i)}{\sigma_{\text{exp}}(q_i)} \right]^2}$$

with σ_{exp} the experimental errors for the N_p data points. In order for the test to be

statistically valid, it is necessary that the experimental errors are estimated correctly. If the errors are well-determined, an accurate model should have $\chi \approx 1$ [139]. Note that χ gets smaller for large errors, however, at the same time, all discriminative power with respect to model selection is lost. The experimental errors are always unknown and have to be estimated from the data using Poisson statistics. This problematic dependence on the error estimate has been noted in the SAXS community and recently, an alternative measure based only on the experimental intensities has been proposed [139]. Here, to calculate SAXS scores under identical conditions, we modified the χ test and used a common weighting function instead of the experimental errors (see Methods). We only used the experimental errors as a measure for noise and discarded parts of the data where the experimental error was significantly larger than our weighting function. Establishing a common baseline for scoring was already recognized as an important issue in previous work [398, 491] and led to the elimination of two widely used fitting parameters in the χ test. Here, we went one step further and established a common weighting for different experimental data sets. Furthermore, instead of simply ranking generated models by their discrepancy from the SAXS profile, we established a fixed cutoff for determining which models are compatible with the SAXS profile. The advantage is that cases, in which the protein structures correspond poorly to the experimental data, can be identified by collecting no or only few structures during the sampling stage (as for 4 of the test cases with experimental profiles). The downside is that the correct model might not be sampled, if its E_{SAXS} score is slightly higher than the cutoff. However, the results on the docking benchmark and on the experimental dataset showed that ATTRACT-SAXS's benefits clearly outweigh the disadvantages.

The success of our protocol demonstrates that in principle SAXS data alone contain sufficient information to generate a rigid placement of the partners close to the native complex structure. However, in terms of scoring, SAXS is not specific enough and further information, either in the form of force fields and/or additional experimental data, has to be used. As noted before [403], combining global information from SAXS with local interface information; e.g., from cross-linking or NMR, is a very promising approach. Such local information could also improve interface modeling during flexible refinement of docked structures [96].

8.5. Conclusion and Outlook

Small angle X-ray scattering (SAXS) experiments yield low-resolution structural information for biomolecules in solution and holds the promise of fairly high-throughput characterization of protein-protein interactions. Due to their relative simplicity, SAXS

experiments have become more and more popular in recent years. Furthermore, SAXS data sets are becoming more and more available to the scientific community via specialized databases [455, 198], although at the moment the number of deposited data sets is still quite small. Integrative modeling approaches combine data from low-resolution experiments with computational optimization to predict the structure of biomolecular assemblies. In this work, we designed ATTRACT-SAXS, a new SAXS-driven integrative modeling protocol in the ATTRACT docking engine. ATTRACT-SAXS already makes use of the SAXS data during its extensive sampling stage and aims to enumerate all models compatible with the experimental data. We tested the method on a large set of protein-protein complexes with simulated profiles and on a test set with experimental SAXS data. ATTRACT-SAXS outperforms previous methods [403, 491], with an especially significant improvement for medium and hard docking cases. Comparison with other integrative modeling paradigms in ATTRACT [96] indicated a sampling performance similar to that of *perfect* atom-atom contact information. Example scripts for ATTRACT-SAXS will be available in the ATTRACT 2.0 release (see Chapter 4). In the future, the method could be expanded towards modeling other biomolecular complexes; e.g., protein-nucleic acids interactions, and towards including cofactors. After further optimization, ATTRACT-SAXS computations could also be offered to the structural biology community via a dedicated web-server. We plan to use ATTRACT-SAXS for structural modeling of chromatin remodeling enzymes based on SAXS data (as an extension to the work presented in Chapter 11).

9. ATTRACT's Performance in CAPRI Rounds 28-36

The ATTRACT coarse-grained docking approach in combination with various types of atomistic, flexible refinement has been applied to predict protein-protein and peptide-protein complexes in CAPRI rounds 28–36. For a large fraction of CAPRI targets (12 out of 18), at least one model of acceptable or better quality was generated corresponding to a success rate of 67%. In particular for several peptide-protein complexes, excellent predictions were achieved. A combination of template-based modeling and extensive molecular dynamics-based refinement often yielded medium and high-quality solutions. In one particularly challenging case, the structure of an ubiquitylation enzyme binding to the nucleosome was correctly predicted as a set of acceptable quality solutions. Based on the experience with the CAPRI targets, a new approach for flexible interface refinement (Chapter 5) and an ab-initio peptide-protein docking protocol have been developed (Chapter 6). Failures and possible improvements of the docking method with respect to scoring and protein flexibility will also be discussed. The work presented in this chapter will be published in similar form in [390].

9.1. Introduction

Protein-protein and peptide-protein interactions are abundant in the cell and are involved in virtually all important biological processes. However, so far only a small fraction of complex structures has been characterized experimentally. Especially, structure determination of transient interactions between proteins is experimentally challenging and probably in many cases impossible to obtain. Since atomic structural knowledge is vital for understanding the biological roles of these interactions, protein-protein docking methods that predict the 3D structure of complexes and efficient refinement approaches have become more and more important in structural biology. The Critical Assessment of PRediction of Interactions (CAPRI) experiment [257, 259, 260] has provided a framework for blind testing and comparative assessment of protein-protein docking and refinement methods. Apart from evaluating the

current state of the field, CAPRI's diverse, challenging targets have stimulated the development of new and more sophisticated protocols and pushed the limits of what is achievable in protein-protein docking. Our protein–protein docking approach ATTRACT [506, 393, 99] can predict protein-protein and protein-nucleic acid interactions and has already been used successfully in various rounds of CAPRI [503, 299, 261, 101]. ATTRACT's main characteristics are its coarse-grained (CG) force field, the ability to incorporate conformational flexibility already during the initial large-scale search and the possibility to dock any number of (protein) partners. The CG model was derived from statistical analysis of protein-protein and protein-nucleic acid interfaces and is intermediate between a residue/base-level and full atomistic description. It represents each amino acid by up to four pseudoatoms (two for the backbone and one or two for the side chains) [129]. A systematic docking search consists of several potential energy minimizations starting from hundreds of thousands of initial configurations. To speed up the docking calculations, the potential energy can be precalculated on a grid [100]. Global flexibility (e.g., domain–domain motion) can be included explicitly during docking by energy minimization along the directions of precalculated soft normal modes [297]. Side chain and loop conformational changes can be accommodated by a multi-copy [29] or an ensemble docking approach. The flexible interface refinement method iATTRACT [393] can be used to further optimize the rigid-body docking solutions. We have participated in CAPRI rounds 28–36 and in the following report on our predictions for targets 59–107 and related new methodological developments. This also includes our efforts to design a fully blind peptide-protein protocol and molecular dynamics-based refinement approaches.

9.2. Methods

9.2.1. ATTRACT rigid-body docking

The protein and peptide structures were converted to the ATTRACT atom type representation [506] with the ATTRACT tool `reduce` (see Chapter 4 for details). Starting points were generated by choosing random positions and orientations for the association partners with an appropriate center-of-mass distance to prevent steric overlap in the initial configuration. The starting structures were subjected to rigid-body optimizations in a potential energy minimization of 1000 minimization steps with the ATTRACT metric minimizer [296, 297]. Energy calculation was accelerated using a precalculated grid [100]. In cases where distance restraints were employed, the optimization of the six rigid-body degrees of freedom was preceded by a minimization in which the center of mass positions were fixed and the docking partners could ori-

ent towards each other. During this stage, only the restraint potentials were applied ("ghost" mode). If multiple conformations were available for the association partners, these conformations were docked separately (ensemble docking). Finally, the docking models were ranked by their ATTRACT energy evaluated within a squared cutoff of $50\,\text{Å}^2$ and highly similar models were removed with the `deredundant` tool.

9.2.2. iATTRACT flexible interface refinement

The protein and the peptide structures were converted into the OPLS atom type description with the ATTRACT tool `aareduce`. Missing hydrogens were built with PDB2PQR [111, 110] and protonation states were determined by PropKa[266]. Peptide termini were charged (unless the peptides were part of a larger protein), protein termini left uncharged. The atomistic refinement uses a physical force field based on the OPLS parameters to calculate non-bonded and electrostatic interactions between the protein partners. Contacts from the input structure are treated as flexible during a simultaneous potential energy minimization in rigid-body degrees of freedom and interface flexibility [393]. A structure-based force field is determined on-the-fly to evaluate intra-protein interactions for the flexible interface residues. Depending on the size of the target, a few hundreds to thousand models from rigid-body docking with ATTRACT were selected for iATTRACT refinement. The refinement parameters were chosen as specified in [393] and [391]. Structures were in general not rescored after iATTRACT refinement.

9.2.3. Molecular dynamics refinement

The structures were converted to the AMBER atom type description using the pdb4amber tool. A Generalized-Born implicit solvent model (igb=8) was used with the newest version of the AMBER force field ff14SB [60]. Minimization and molecular dynamics refinement in implicit solvent with AMBER 14 [60] were carried out as described in Chapter 6.
For refinement in explicit solvent with GROMACS version 4.6 (www.gromacs.org) [37, 356, 2], structures were converted into the gro format with the tool pdb2gmx. Simulations were run with explicit solvent using the TIP3P water model and the AMBER99SB-ILDN force field at a temperature of 300 K. The positions of the backbone atoms were restrained with force constant $1000\,\text{kJ}/\text{mol}/\text{Å}^2$ in x, y and z direction.

Table 9.1. Results for ATTRACT predictions in CAPRI rounds 28-36.

Target	Type	Best model	fnat	IRMSD [Å]	Classification
59	protein-protein	5	0.18	3.8	acceptable (2⋆)
60	peptide-protein	5/6	1.0/0.94	0.44	high (2⋆ ⋆ ⋆,4⋆⋆)
61	peptide-protein	1	0.76	0.5	medium (5⋆⋆)
62	peptide-protein	1	0.92	0.39	high (2⋆ ⋆ ⋆, 3⋆⋆)
63	peptide-protein	1	0.81	0.49	high (2⋆ ⋆ ⋆, 3⋆)
64	peptide-protein	2	0.87	0.42	high (3⋆ ⋆ ⋆, 2⋆⋆)
65	peptide-protein	2	0.27	3.7	incorrect
66	peptide-protein	1/1	0.5/0.75	1.3/2.1	acceptable (1⋆)
67	peptide-protein	5	0.88	0.8	medium (2⋆⋆, 8⋆)
68–94	protein-protein	-	-	-	no submission
95	protein-protein	3	0.5	3.3	acceptable (5⋆)
96	protein-protein	3	0.15	2.96	acceptable (1⋆)
97	protein-protein	8	0.05	12.2	incorrect
98–101	protein-protein	-	-	-	no submission
102	protein-protein	-	-	-	not yet available
103	protein-protein	10	0.27	12.2	incorrect
104	interfacial water	4	0.68	0.92	high (1⋆ ⋆ ⋆,9⋆⋆), water (5++)
105	interfacial water	3	0.66	1.2	medium (10⋆⋆), water (7+++)
106	protein-protein	-	-	-	not yet available
107	protein-protein	7	0.05	17.42	incorrect
108-109	peptide-protein	-	-	-	canceled

9.3. Results and Discussion

In CAPRI rounds 28–36, we submitted predictions for targets 59-67, 95-97 and 102-107 (we did not participate in the CASP-CAPRI experiment in round 30 and in CAPRI round 32; round 36 was canceled before the submission deadline). Since the templates for the targets in round 32 (targets 98–101) were all of very low sequence identity, we did not attempt to model these targets and to predict possible complex structures. Note that indeed none of the predictors achieved any successful prediction for round 32.

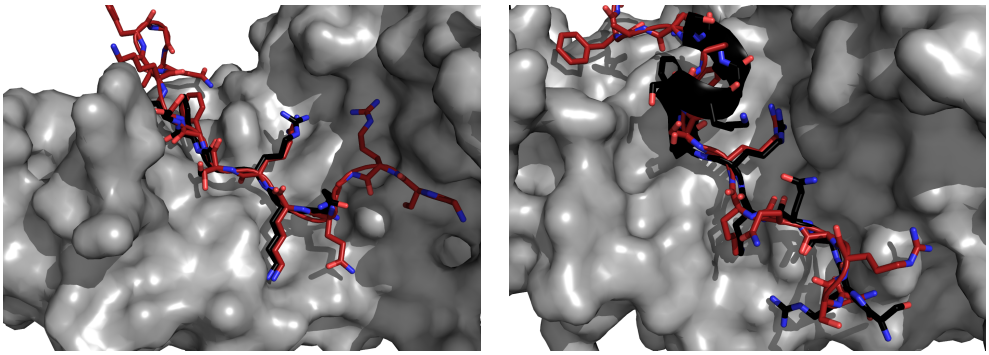
Table 9.1 shows a summary of the results. The targets vary strongly in terms of interaction type and docking difficulty. In particular, many targets of the most recent CAPRI rounds involved a high degree of flexibility and consequently, results for all predictor groups were rather poor. In the following, we discuss our predictions and the challenges we faced for some of the targets.

9.3.1. Round 28 (Targets 59–64)

Round 28 consisted of one protein-protein complex (target 59) and five peptide-protein complexes (targets 60–64). Target 59 corresponded to the Edc3 LSm domain, an activator of the mRNA decapping complex, with a motif from Rps28B. NMR ensembles of structures for the LSm domain were available both in the apo form and bound to another motif and we used all NMR models removing the highly flexible regions prior to docking [449, 142]. For Rps28B, we used an ensemble of homology models created with MODELLER [482] based on available structures [488, 15]. We performed ATTRACT ab-initio rigid-body docking and subsequent molecular dynamics refinement with AMBER [60]. For this target, we obtained two acceptable structures. The best submitted model deviated from the native structure at the interface by $3.8\,\text{Å}$ and retrieved 18% of the native contacts. To date, a more detailed analysis for this target is not possible, since the experimental structure has not yet been published. Target 60–64 were structures of importin-α binding to different peptides derived from nuclear localization signals (NLSs). NLSs contain one or two clusters of basic residues and are recognized by the import receptor importin-α [63]. Several structures of importin-alpha in complex with different peptides were available at the time of round 28 which showed two binding sites: a major and a minor site. The peptides in Round 28 were derived from nuclear localization signals and had been optimized towards binding to the minor site [63]. In the crystal structure, the major site was also occupied and so evaluation of these targets was carried out for both the major and the minor site. We generated models for the peptide based on the available crystal structures and used this in an ATTRACT rigid-body docking with ambiguous distance restraints towards the major and the minor sites. The best models were then refined by an energy minimization with AMBER [60]. For all but one target, we achieved high-quality predictions (Figure 9.1). However, we failed to predict the $\alpha-$helical turn in the peptide when binding to the minor site. Here, possibly more extensive peptide structure modeling [435] prior to complex prediction could have improved the results.

9.3.2. Round 29 (Targets 65–67)

In Round 29, three distinct peptide-protein complexes were proposed as targets. Target 65 and 66 were complexes of proteins with a SSB C-terminal peptide. For both targets, we analyzed available structures of other proteins in complex with a SSB C-terminal peptide (PDB 3UF7, 3Q8D, 3C94) [380, 280] and identified a conserved binding mode with an hydrophobic anchor at the C-terminus and a solvent exposed part of the peptide, although in general the peptide appeared relatively flexible. We

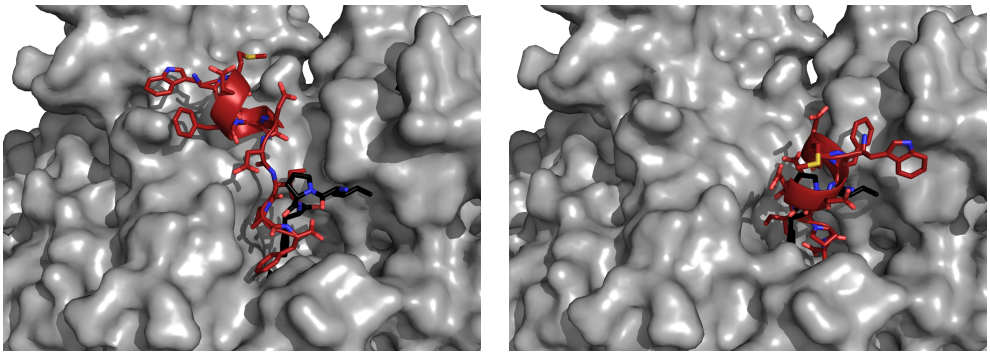


(a) Major site. **(b)** Minor site.

Figure 9.1. Peptide-protein docking model for target 60 (docked peptide indicated as red sticks, importin-α receptor shown as grey surface) superimposed on the native structure (PDB 3ZIN, bound peptide in black). Peptide binding to importin-α was modeled using a combination of homology modeling and molecular dynamics refinement. For this target, and related targets 62-64, several three-star quality models were submitted.

extracted the peptide conformations from the available complexes, recombined parts of the peptides from different crystal structures to generate additional conformations and used this peptide ensemble in a fully blind coarse-grained search of the entire protein surface combined with two stages of flexible refinement [391]. Before refinement, the rigid-body models were filtered to detect models with a buried phenylalanine at the C-terminus.

Target 65 was a complex with RNase Hi [345]. For this target we did not generate any near-native prediction, since we did not sufficiently model the protein's loop flexibility at the peptide binding site (we only used a single crystal structure for the protein during docking [204]). We therefore failed to detect the correct pocket. Target 66 corresponded to a PriA helicase in complex with a SSB C-terminal peptide [45]. We generated one model of acceptable quality (top 1) and one model (top 2) which had a very similar orientation of the peptide but an incorrect peptide conformation (α-helical turn) (Figure 9.2). In our top-ranked model, the whole peptide was in contact with the protein surface, whereas in the crystal structure the N-terminal part of the peptide is solvent-exposed. The tendency to maximize the interface between a flexible peptide and a binding region can be attributed to the use of an implicit solvent model during final refinement (see Methods). In future cases, we will consider explicit solvation during refinement of peptide-protein complexes which may help to improve the accuracy of the predictions.



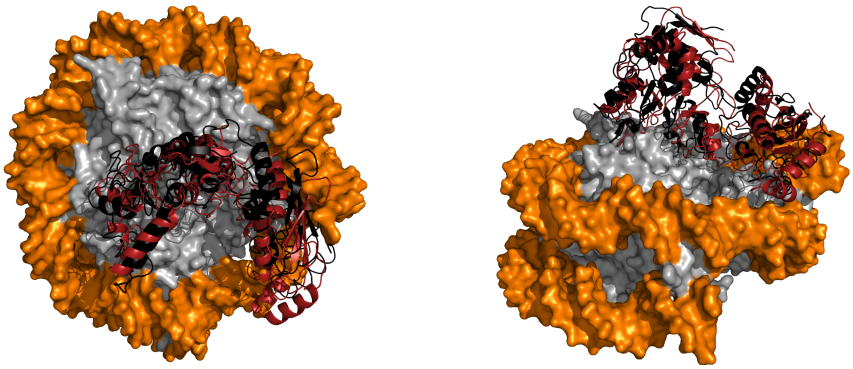
(a) top 1 **(b)** top 2

Figure 9.2. Predicted structural models (top1 and top2) for target 66 (docked peptide in red, receptor as grey surface) superimposed on the native structure (PDB 4NL8, bound peptide in black).

Target 67 corresponded to a WW domain in complex with an extended peptide containing proline (P) residues. Several structures of PPXY motifs bound to a WW domain were available in the PDB [47] and we used these to build initial models of the peptide-protein complexes. Since only a limited number of complexes were generated it was possible in this case to use restrained molecular dynamics simulations in explicit solvent with GROMACS for refinement (1 ns simulations at room temperature and normal pressure, see Methods). Finally, the models were energy-minimized with sander of the AMBER package. All our 10 submitted models were at least of acceptable quality and the best model had an IRMSD of $0.8\,\text{Å}$ and retrieved 88 % of the native contacts (medium quality).

9.3.3. Round 31 (Targets 95–97)

Round 31 comprised three very challenging protein-protein complexes as targets. For target 95, binding of an ubiquitylation enzyme PRC1 (Bmi1/Ring1b ubiquitin ligase) to the nucleosome had to be predicted [300]. PRC1 ubiquitylates the histone H2A tail at residue LYS 119 [320, 479]. Furthermore, several residues on PRC1 (ASP 56, LYS 92, LYS 93, LYS 97 and ARG 98 on Ring1b and LYS 62 and ARG 64 on Bmi1) had been identified as important for binding by mutational experiments [36]. There was also experimental evidence for PRC1 binding to DNA (although only for isolated DNA, not for the nucleosome) [36] and data pointing to an important role for the acidic patch in PRC1 function [264]. We used the unbound protein structures for PRC1 [36] and the nucleosome [464] and performed a large scale rigid-body docking search with a $10\,\text{Å}$ upper harmonic distance restraint between the C_α atoms of residue LYS



(a) Top view. **(b)** Side view.

Figure 9.3. Best model (top 3) for target 95 superimposed on the native structure (PDB 4R8P, black). PRC1 binds to the nucleosome at the histone acidic patch.

118 on the H2A tail (LYS 119 was not resolved in the structure) and residue CYS 85 in the active site of the ubiquitin ligase. Subsequently, we refined the best clusters with iATTRACT and short molecular dynamics simulations with AMBER. We found two types of solutions among the top-ranked models: one with PRCC1 contacting the acidic patch and one with PRCC1 binding to nucleosomal DNA similar to an earlier model [36]. We hence submitted 5 models for each type of solution and achieved 5 acceptable model with the best model being very close to medium quality (LRMSD 5.03 Å, fnat 0.5, Figure 9.3).

Targets 96 and 97 were complexes of designed α-repeat proteins binding to GFP (PDB 4XL5 and 4XVP) [75]. We used our standard ab-initio docking protocol in combination with additional scoring with the Rosetta force field which has been used in the past for protein design [477, 225]. We achieved acceptable quality for target 96. However, we failed to accurately predict the smaller target 97 sampling solutions where the orientation of the ligand in our models was rotated by $180°$ with respect to the native structure. This failure can be attributed to deficiencies of the scoring of the docking models selected for further refinement.

9.3.4. Round 34 (Targets 104–105)

In Round 34, the challenge was to model the placement of interfacial waters. The targets were pyocin DNase domains in complex with immunity proteins which were structurally similar to previously solved structures in colicin. We built an initial

model of the complex by homology modeling using a colicin complex [486] and refined this with backbone restrained molecular dynamics simulations in explicit solvent. We achieved a high-quality model for target 104 and medium quality models for target 105. For the prediction of the interfacial water placements we followed our molecular dynamics based protocol described previously [101, 261]. Briefly, in this protocol water molecules are placed using the AMBER leap module allowing for partial overlap of waters with solute atoms at the interface (resulting in a slight over-hydration of the interface). The hydration structure is then allowed to relax during short MD simulation including positional restraints on the protein backbone. Finally, all waters outside the interface are removed followed by energy minimization to optimize the position and orientation of each water molecule at the interface (see details in reference [261]). The prediction of interfacial waters was more accurate for target 105 (medium quality model), probably because there were less water molecules to predict for this target. Interestingly, in the overall CAPRI evaluation it was found that high accuracy water predictions were not necessarily generated only for high-quality models of the protein-protein complex (Lensink, unpublished data). This is in contrast to previous results for interfacial water predictions in CAPRI [261] and indicates that probably only certain key groups at the interface have to be in near-native position to allow correct placement of waters.

9.3.5. Round 33/35 (Targets 102–103 and 106–107)

In these two rounds, the targets were a complex of haemopexin with the haemopexin utilization protein (huxA). The complex was solved in the apo state and with haem bound. It was previously speculated that the C-terminal domain of huxA interacts with haemopexin [138]. However, at the time of submission for Round 33, structural data was only available for the N-terminal secretion domain [18]. Hence, the target was proposed again as a target in Round 35 offering an unbound structure of full-length huxA. Still, we were not able to generate any near-native predictions, since binding of huxA to haemopexin involves a conformational change in a loop on HuxA which is partly disordered in the unbound form. Unfortunately, this loop yields steric clashes with haemopexin when superimposing the unbound huxA structure on the complex. A strategy involving detailed flexibility analysis prior to docking, removal of flexible loops and a-posteriori loop rebuilding (using e.g. the protocol proposed in Chapter 7) would have certainly yielded improved results for this target.

9.4. Conclusion and Outlook

In CAPRI Rounds 28-36, we submitted predictions for 18 targets (16 distinct targets) and achieved at least acceptable predictions in 12 cases. This result can be considered as highly successful given the difficulty of many of the targets. The coarse-grained ATTRACT docking approach in combination with different refinement schemes proved to be versatile in dealing with a variety of different targets that ranged from peptide-protein interactions to docking of a large protein to the nucleosome. The CAPRI challenge has also triggered the development of several extensions of the original ATTRACT approach in the area of peptide-protein docking (pepATTRACT) [391] and of refinement (iATTRACT) [393] within the last rounds. For several CAPRI targets, we found that homology modeling in combination with restrained molecular dynamics refinement yielded medium and high-quality predictions. In some cases, significant conformational change or inaccuracies in the homology modeling have contributed to the failure of the docking search. More detailed flexibility analysis prior to docking is needed to select appropriate conformational ensembles and to identify flexible loops. It might be beneficial to eliminate such highly flexible loops prior to docking if the chances to correctly model the bound forms are small and to tackle the generation of a bound loop structure after docking (see also Chapter 7). Such approaches may also be useful for improving protein-protein complexes predicted based on sequence similarity to a known template complex. Furthermore, improvements in scoring towards distinguishing structures of medium or higher quality from just acceptable solutions are highly desirable and at the focus of our future research.

10. Peptide Recognition in the ER Associated Degration Pathway

Chaperones and co-factors like the DnaJ family recognize misfolded proteins in the endoplasmic reticulum by binding to aggregation-prone sequence stretches. These sequence motifs are typically buried when the protein is folded but remain accessible in the misfolded state. Hence, interactions with these peptidic motifs can only occur in unfolded or misfolded conformations and therefore binding to these sequences can serve to detect incorrect protein folding. While many chaperones recognize a wide range of hydrophobic sequences, several DnaJ family members display highly sequence-specific substrate binding. However, the molecular details of this specificity are not known, since atomic structural data of these interactions is lacking to date. This chapter describes an application of the pepATTRACT protocol (Chapter 6) to study peptide recognition by the DnaJ family member and chaperone cofactor ERdj5. We identified a possible peptide binding site on the Trx2b domain and proposed a set of frequently contacted residues for further mutational studies.

10.1. Introduction

In order to carry out their biological function, most proteins adopt a defined three-dimensional structure. This folding process is often controlled by molecular chaperones [176] that can recognize misfolded proteins and either refold them or mark them for degradation. Hence, chaperones exert protein quality control and prevent potentially harmful aggregation processes of misfolded proteins. The endoplasmic reticulum (ER) processes about one third of the proteins encoded in the human genome. Folded proteins are then transported along the secretory pathway and either are incorporated in the cell surface or secreted. Since detection mechanisms are limited once the proteins are secreted, asserting correct folding in the ER is vital [54] and failures in chaperone-assisted folding processes are associated with numerous human diseases [169, 183].

Several ER-associated major chaperone families have been identified in different

organisms and many of them are able to bind to a large variety of proteins that are unrelated in sequence and structure. This flexibility arises from the fact that chaperones, like the mammalian ER Hsp70 family member BiP [135], often recognize short stretches composed of hydrophobic residues that are buried if the protein folds correctly. All Hsp70 family members act as central chaperones and can further associate with multiple (co-)chaperones and folding enzymes [176] like the DnaJ cofactors. Some DnaJ cofactors can also recognize misfolded proteins and transfer them to their Hsp70 interaction partners. In addition to binding to the generic hydrophobic recognition patterns, in higher eukaryotes, different DNaJ cofactors exist that convey specificity to their Hsp70 chaperone [215, 89]. Recently, Behnke et al. investigated the sequence-specific binding preferences of different members of the ER Hsp70 chaperone system in vivo. Interestingly, they found that the DnaJ family member ERdj5 interacted specifically with aggregation-prone sequences in two secretory pathway proteins: immunoglobulin γ_1 heavy chain (mHC) and NS-1 κ light chain (Behnke et al. (2016), Mol. Cell, under review). After binding to its substrates, ERdj5 reduces disulfide bonds and can thereby assist folding or facilitate subsequent degradation of the protein [327]. However, detailed structural insights into the interaction between the peptidic recognition sequences and ERdj5 and the molecular origin of substrate specificity are unavailable to date.

Computational peptide-protein docking methods can complement experimental structure characterization by predicting the structure of the peptide-protein complex from the structure of the individual protein and the peptide sequence. The full-length ERdj5 protein has been previously characterized by X-ray crystallography [174]. The protein is organized into an N-terminal cluster and a C-terminal cluster. It contains a J-domain and in total 6 structurally similar tandem thioredoxin domains (Trx1, Trxb1, Trx2, Trxb2, Trx3, Trx4) of which two (Trxb1 and Trxb2) lack the catalytic, redox-active CXXC motif. The N-terminal cluster is composed of the J-domain and Trx1, Trxb1, Trx2 and Trxb2.

Here, we applied the peptide-protein docking protocol pepATTRACT (Chapter 6) to model the interactions between ERdj5 and various peptides. We investigated possible peptide binding sites on ERdj5 and identified ERdj5 residues that form important contacts with the peptides.

10.2. Methods

We used the pepATTRACT protocol with default parameters as described in [391] to model interactions between ERdj5 and different peptides. We used the following peptide sequences

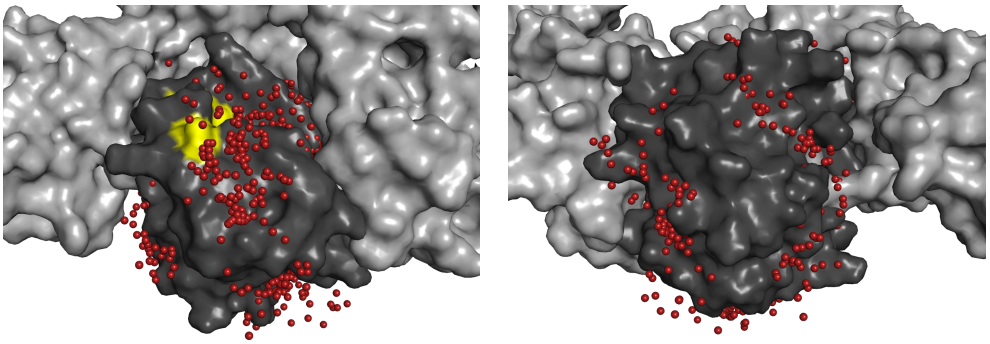
- mHC2.2: GYTFTSYWMHWV
- mHC4: KFFSYATLTVDF
- mHC6.1: CASYDYDWFAYW
- NS1-7.3: SGGASVVCFLNNF.

During docking, the peptides' termini were left uncharged, since they are usually a part of larger proteins. We extracted the structure of the Trxb2 domain from the crystal structure of the full-length ERdj5 protein (residues 351-456 of PDB 3APO) [174]. Missing residues were added to the structure with MODELLER [482] and mutated residues were changed back to cysteine using PyMOL [404]. Hydrogen bond analysis was carried out with VMD [197] with a distance cutoff of 3.5 Å and an angle cutoff of 30°.

10.3. Results and Discussion

The goal of this work was to obtain structural insights into peptide binding to the chaperone-cofactor and disulfide reductase ERdj5. We employed the pepATTRACT peptide-protein docking approach [391] to predict possible binding modes of peptides to ERdj5 [174] (see Methods). Previous truncation experiments showed that peptide recognition occurs at the N-terminal cluster. Furthermore, the mHC2.2 peptide displayed the highest binding affinity towards the Trxb2 domain (Mideksa and Feige, unpublished data). Therefore, in this study, we focused on the interactions of peptides with the Trxb2 domain.

The docking results for mHC2.2 are shown in Figure 10.1 as an example. When considering only solutions that are sterically compatible with the full-length protein, there is a clear preference for peptide binding in the vicinity of the loop in which the redox-active motif is located in other Trx domains (see also Figure E.1). Figure 10.2 displays representative models for binding in different pockets close to the redox-active motif. Interestingly, in many models, a hydrophobic residue acts as an anchor inserting itself into a small pocket on the protein surface (Figure 10.2 (a), (c), (e) and (g)). For mHC2.2, we saw that residue serine 6 often formed a large number of polar interactions with the protein. This corresponds well to a series of mutational experiments that indicated a strong contribution of this residue to the overall binding affinity in addition to the important contributions of hydrophobic residues (Behnke et al., 2016, Mol. Cell, under review). Most docking models exhibit extensive hydrogen bond formation, especially between the peptide main chain and protein side chain atoms (examples are shown in Figure 10.2 (b), (d), (f) and (h)).



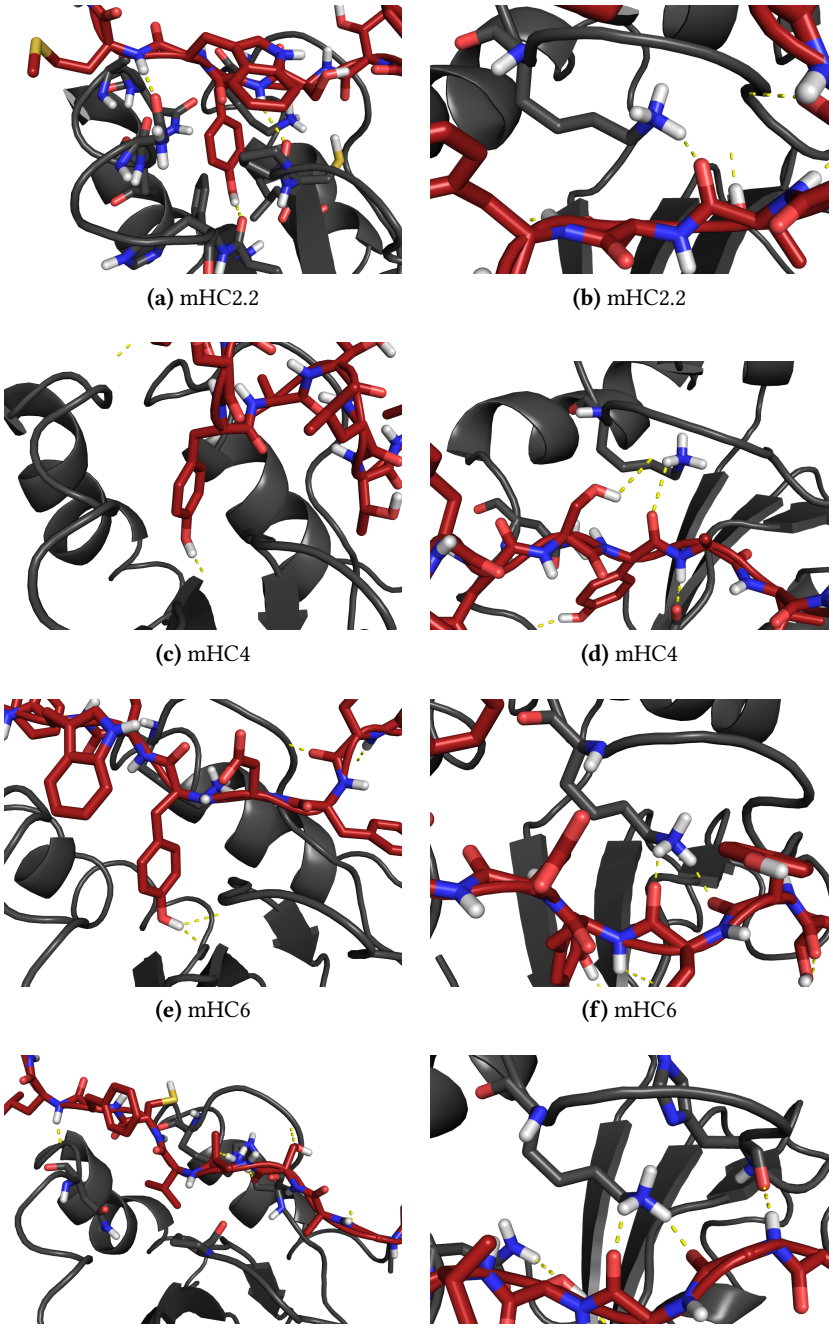
(a) Front view. **(b)** Back view.

Figure 10.1. Docking results for mHC2.2 peptide binding to ERdj5 Trxb2 domain. The position of all 1000 final models is shown by marking the position of SER6 by a red sphere. The docking models are superimposed on the full-length structure of ERdj5 [174] (gray). The Trxb2 domain is highlighted in dark gray. The positions equivalent to the redox-active motif in other Trx domains are shown in yellow.

We also found that polar peptide groups such as serine, threonine and asparagine side chains and the OH group on tyrosine residues often engaged in hydrogen bonds. These polar interactions most likely convey binding specificity. Previous experiment already confirmed the strong influence on binding when mutating serine 6 in mHC2.2 to proline (Behnke et al., 2016, Mol. Cell, under review) and we found that this side chain forms the largest number of hydrogen bonds with the protein among all peptide side chains in mHC2.2 (data not shown). Our analysis for the other peptides suggests contributions for threonine 7 and threonine 9 in mHC4 (although the overall contributions to hydrogen bonds was much lower than for serine 6 in mHC2.2), very strong contributions for aspartate 7 in mHC6 and strong contributions for serine 5 in NS1-7. Note that this result might be generally biased towards residues located in the middle of the sequence. In the future, the contribution of these residues to the overall binding affinity could be probed by mutational experiments.

In order to find protein interface residues as candidates for mutational studies, we analyzed the available models and selected residues that either formed hydrogen bonds with the peptide or interacted with hydrophobic side chains. We identified the following residues

- involved in hydrogen bonds (via side chains): 438K, 417Q, 412D, E359, 428T, 437K, 480N, 429S, 404S, 356N, 366R or 361R (potential binding pocket on the other side of the Trxb2 domain, Figure 10.1 (b))



(a) mHC2.2 **(b)** mHC2.2

(c) mHC4 **(d)** mHC4

(e) mHC6 **(f)** mHC6

(g) NS1-7 **(h)** NS1-7

Figure 10.2. Docking models illustrating recurring peptide binding motifs and important interactions. The Trxb2 domain is shown in dark gray, the peptide in dark red. Polar contacts between the peptide and the protein are drawn in yellow.

- involved in hydrophobic interactions: 373F, 416F, 370F, 420L, 367W (the last residue again corresponds to an alternative binding site).

These residues could be a starting point for further mutational studies and could help to validate the peptide binding site.

10.4. Conclusion and Outlook

Here, we described an application of the pepATTRACT peptide-protein docking protocol to studying peptide recognition and substrate specificity in the cofactor ERdj5. We found recurring binding motifs for all peptides involving a hydrophobic anchor and extensive hydrogen bond formation between peptide and protein. Based on contact analysis, we identified a set of protein residues that could be important for peptide binding. In the future, the proposed interface residues can be tested in mutational experiments to validate or falsify the binding mode. Once residues in the binding site have been confirmed experimentally, the model of the complex can be refined by docking with the pepATTRACT-local protocol [391] (see Chapter 6).

11. Integrative Modeling of ISWI Nucleosome Remodeling Enzyme

The eukaryotic genome is organized in the form of chromatin, a complex of DNA with compacting and regulatory proteins. The basic packing unit is the nucleosome where DNA is wrapped tightly around histone proteins. This packaging inevitably leads to the occlusion of DNA sequences that can no longer be accessed by regulatory proteins and the transcription machinery. To allow dynamic and regulated use of the genome, ATP-consuming nucleosome remodeling enzymes exchange histones, reposition, assemble, and disassemble nucleosomes. Structural information for nucleosome remodelers are scarce and hence, atomistic insights into the nucleosome sliding mechanism and the auto-inhibition of the enzymes are lacking to date. This chapter presents an application of different docking protocols in ATTRACT with the aim of modeling the full-length structure of the ISWI nucleosome remodeling enzyme using cross-linking/mass spectrometry (XL-MS) and small-angle X-ray scattering (SAXS) data. We obtained a model of the inactive conformation of the ISWI ATPase domain that is compatible with all available experimental data. We further identified a possible binding site for N-terminal auto-inhibitory motifs on ATPase lobe 2. Finally, the conformation of the C-terminal DNA-binding HSS domain with respect to the ATPase domain was predicted based on XL-MS data and validated by SAXS. The generated models yielded detailed insight into the regulation and auto-inhibition of ATP-dependent nucleosome remodeling enzymes. This chapter also describes a new integrative modeling approach in ATTRACT that combines XL-MS data and docking (ATTRACT-XL). The developed integrative modeling approach is broadly applicable to many transient multi-component assemblies.

11.1. Introduction

The eukaryotic genome is tightly packaged in the nucleus with the help of histone proteins. A 147 base pair DNA stretch is wrapped around the histone core octamer

forming the so-called nucleosome. However, this tight wrapping occludes access to DNA sequences by the DNA transcription machinery and proteins that regulate gene expression. This limitation is overcome by ATP-dependent nucleosome remodeling complexes that reposition, eject or modify nucleosomes. Nucleosome remodeling complexes contain a catalytic subunit that possesses ATPase activity and further regulatory subunits. The ISWI-family remodeler complexes contain the ISWI protein that can translocate DNA and slides nucleosomes [484, 524, 382]. ISWI-family remodelers can produce equally spaced nucleosome arrays [249] and have an important role in chromatin assembly after DNA replication and maintenance of higher-order chromatin structure [87]. Due to their crucial biological function, strict control of remodeling activity is vital and the ISWI protein is tightly regulated by auto-inhibitory motifs and a variety of nucleosomal epitopes [78, 77, 175, 80, 91].

The ISWI protein contains two major domains–an ATPase domain that is similar to other SF2 helicases and a DNA binding C-terminal HAND-SANT-SLIDE domain–and several regulatory regions [78] (Figure 11.1). ISWI's activity is regulated by two nucleosomal epitopes: the protein is activated by binding to a basic patch on the histone H4 tail and by extranucleosomal linker DNA [77, 175, 80, 91]. ISWI is negatively regulated by the N-terminal auto-inhibitory region AutoN region (RHRK motif) that suppresses ATP hydrolysis [78]. AutoN inhibition of ISWI's ATPase activity is lifted by binding to the histone H4 tail that contains a similar basic patch [77, 78]. Recently, it was discovered that a highly conserved, negatively charged region close to the AutoN motif ("acidic patch") also strongly inhibits ISWI ATPase activity (Figure 11.1). Mutations in either motif resulted in increased ATPase activity (Ludwigsen, unpublished data). However, combined mutations in both motifs did not increase the activity to a large extent beyond the effect of mutations in either region. Hence, these experiments suggest that these two motifs might form a functional inhibition module (Ludwigsen and Müller-Planitz, unpublished data). However, atomistic insights into the auto-inhibition mechanism of ISWI are lacking to date. Full-length nucleosome remodelers have been refractory to high-resolution structure determination by X-ray crystallography. Currently, only the structure of the isolated HSS domain [497, 168] and a few structures of the ATPase domain in other SF2 helicases have been resolved [410, 120, 436]. In particular, no structural information is available for the ISWI N-terminal region (NTR, residues 1–110), in which the AutoN and acidic patch motifs reside. Several experiments have indicated that ISWI undergoes conformational changes depending on external conditions; e.g., presence of nucleotides, and that several conformations may exist in vivo [263, 357]. Cross-linking/mass spectrometry (XL-MS) is a low-resolution technique for characterizing structures of proteins and protein complexes in solution (see also Section 2.4.5). In a XL-MS experiment, the protein is cross-linked with a chemical agent converting

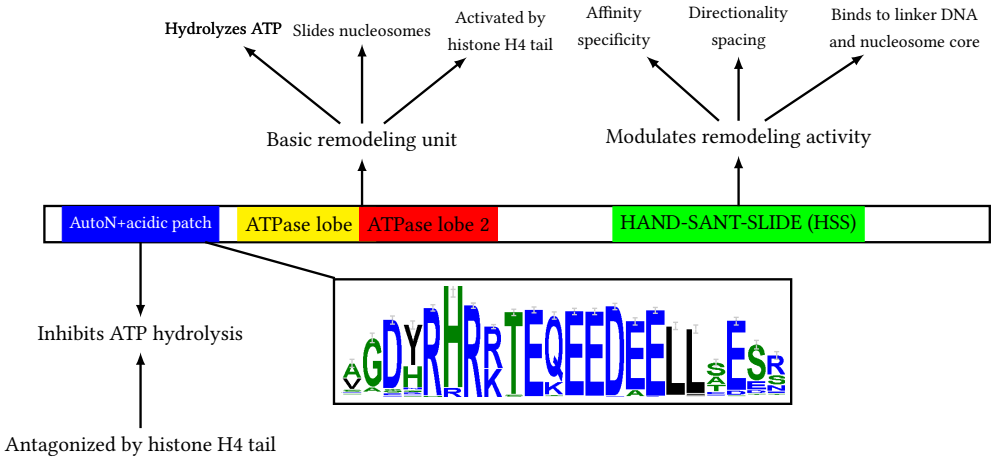


Figure 11.1. Schematic representation of the domains and regulatory regions of the ISWI nucleosome remodeling enzyme. Alignment of 380 ISWI sequences focused on the AutoN+acidic patch region was performed with ClustalOmega [417] and visualized with Weblogo [88, 396].

non-covalent interactions into covalent bonds. The cross-linked protein is then digested and the resulting peptide mix is examined by mass spectrometry to identify precisely which residues were involved in contacts. Mass spectrometers, biochemical protocols and software have advanced considerably since the first experiments at the beginning of the century [475] and numerous XL-MS experiments in conjunction with computational modeling have yielded important insights into the structures of different proteins and protein assemblies [72, 352]. This has established XL-MS as a valuable tool for the structural biology community.

Here, we set out to elucidate the structure of the full-length ISWI protein by combining data from XL-MS and SAXS with computational docking. To achieve this goal, we developed a robust integrative modeling approach, ATTRACT-XL that makes use of XL-MS data in the ATTRACT docking engine [99]. The approach was validated on XL-MS data of a known protein and then applied to study the structure of ISWI. The resulting models provide important insights into regulatory mechanisms and functional states of an ATP-dependent nucleosome remodeling enzyme.

11.2. Methods

11.2.1. Experimental data

Measurement of ISWI ATPase and remodeling activity were carried out by Johanna Ludwigsen, Nadine Harrer and Felix Müller-Planitz (BioMedical Center, Ludwigs-Maximilians-University Munich) as described in [282, 315]. Cross-linking/ mass spectrometry (XL-MS) data were collected for Drosophila melanogaster ISWI by Nadine Harrer and Felix Müller-Planitz. Three types of XL-MS experiments were performed. Cross-linking was performed in the apo state and in the presence of the nucleosome. Different sites on the ISWI ATPase domain were probed by site-specific photo-crosslinking with a genetically modified amino acid [136]. Lysine-lysine cross-linking was performed on the full-length protein with the cross-lining agent bis(sulfosuccinimidyl)glutarate (BS^2G). Cross-linking/mass spectrometry data were filtered both automatically and manually by inspection of the spectra to eliminate false positives. The size of the crosslinking agent and the size of the side chains it attaches to yield an upper limit for the distance between the C_α atoms of cross-linked residues. Considering a margin for possible conformational change and inaccuracies due to homology modeling, we obtained the following upper distance limits: $15\,\text{Å}$ for photo-XL and $25\,\text{Å}$ for BS^2G.

SAXS data were measured on full-length ISWI and ISWI ATPase (residues 26–644) and collected at beam line 12ID at Argonne National lab and at a beamline in France by Linda Bruetzel and Jan Lipfert (Physics Department, Ludwigs-Maximilians-University Munich). Measurements were essentially as described in [273, 424]. The data at different concentrations are well superposable after rescaling by intensity, indicating good data quality and the absence of aggregation. In line with this observation, the data at different concentrations give (within experimental error) the same radius of gyration obtained from Guinier analysis ($R_g = 42 \pm 0.5$Å). Note that our docking models only contain atomic coordinates for $\approx 75\%$ of the residues of full-length ISWI and $\approx 83\%$ of the residues of ISWI ATPase, which makes comparison to SAXS data challenging [398]. That is why we only used SAXS data for validating and selecting a representative model instead of using the SAXS data for sampling as described in Chapter 8.

11.2.2. Structure preparation

Sequence alignments for homology modeling were obtained from ClustalOmega web service [417, 269, 301]. We built a homology model of bovine serum albumin (PDB 3V03) [289] from the structure of human serum albumin (PDB 4G03) with the MODELLER program [482] through its UCSF Chimera [500] interface and divided it into

three parts (residues 3–206, 207–397 and 398–583). For each part, we repacked the side chains with SCWRL4 [239] to avoid bias towards the native structure. These three parts were then used as input during docking. For ISWI, only a structure of the C-terminal HAND-SANT-SLIDE domain has been crystallized so far (PDB 1OFC) [168]. We therefore modeled the ATPase domains (lobe 1 and lobe 2) by homology from the structure of Chd1 (PDB 3MWY) [177] using the MODELLER program [482, 500]. The ISWI ATPase domain shares more than 40 % sequence identity with the Chd1 ATPase domain. PDB structures of Snf2 family members were detected with blastp [58]. The hinge region between the two ATPase lobes was predicted by the HingeProt server [122] and the ATPase domain was cut into two parts at this predicted hinge (lobe 1: residues 116–351; lobe 2: residues 352–637). Flexible regions which differed significantly between different experimental structures of Snf2 ATPases were removed.

11.2.3. Ab-initio docking

Ab-initio protein-protein and peptide-protein docking was performed as described in [99] and [391]. The termini of the peptides were not charged, since they were derived from a protein region (ISWI NTR or histone H4 tail).

11.2.4. Representation of XL-MS data in ATTRACT

The cross-linked residue pairs were either restrained by harmonic distance restraints or bump restraints. Upper harmonic distance restraints between the C_α atoms are given by

$$V_{\text{rest}}^{\text{harmonic}}(r_{ij}) = \tfrac{1}{2}k\left(r_{ij} - d_{\text{max}}\right)^2 \quad \text{for } r_{ij} > d_{\text{max}}$$

with the maximum distance $d_{\text{max}} = 15\,\text{Å}$ for photo-crosslinks and $d_{\text{max}} = 25\,\text{Å}$ for BS^2G cross-links. The force constant k was set to 10. Bump restraints are formulated as

$$V_{\text{rest}}^{\text{harmonic}}(r_{ij}) = \begin{cases} \frac{1}{2}k\left[\left(r_ij - r_1\right)^2 - r_0^2\right]^2 & \text{for } r < r_1 + r_0 \\ \frac{1}{2}kr_0^4 \quad \text{for } r \leqq r_1 \text{ and } r \geqq r_2 \\ \frac{1}{2}k\left[\left(r_ij - r_1\right)^2 - r_0^2\right]^2 & \text{for } r < r_2 \text{ and } r > r_2 - r_0 \\ 0 \text{ else.} \end{cases}$$

The force constant was set to -0.1 and r_0 was set to 3 Å. We used $r_1 = 25\,\text{Å}$ for BS^2G cross-links [132]. r_2 was chosen such that $r_1 - r_2 = 7\,\text{Å}$. Due to their shorter range and higher accuracy, photo-crosslinks were only represented by harmonic potentials.

11.2.5. Integrative modeling protocol driven by XL-MS data (ATTRACT-XL)

The ATTRACT-XL protocol consists of the following steps (Figure 11.2):

1. Input data preparation and data representation. The protein structures and the restraints derived from the experimental data are converted into the ATTRACT format.

2. Repeated coarse-grained ATTRACT rigid-body docking [99] with subsets of cross-links represented as harmonic distance restraints.

3. Coarse-grained ATTRACT rigid-body docking with all cross-links represented as bump restraints.

4. iATTRACT refinement [393] with bump restraints.

5. Analysis and model validation.

The protein structures were converted into the ATTRACT representation with the ATTRACT tools `aareduce` and `reduce`. Missing atoms were built with PDB2PQR [111, 110] and protonation states determined with PROPKA [266]. Initially, for each subset of cross-links, 10,000 starting positions were generated with random orientation of the protein domains. The centers-of-mass of the domains were distributed randomly on a sphere of radius $100\,\text{Å}$ to avoid steric overlap between the domains. As a first step, each of the starting configurations was minimized in 50 step in its orientational degrees of freedom only using only the harmonic restraints derived from the cross-linking data ("ghost" mode). Then the structures were optimized in a potential energy minimization using the ATTRACT coarse-grained force field and the harmonic restraints simultaneously [99]. This process was repeated for all possible combinations of the cross-links. All the resulting models were subsequently minimized using all the available data as bump restraints (photo-crosslinks were still represented as harmonic restraints). The structures were then ranked by their ATTRACT score evaluated within a squared cutoff of $50\,\text{Å}^2$ and the bump restraint energy derived from the cross-linking data and the top-ranked 200 models were subjected to flexible refinement with the iATTRACT protocol with settings as described in [393] again using all the cross-linking data as bump restraints (and the photo-crosslinking data as harmonic restraints). In addition to the restraints derived from the XL-MS data, we also applied an upper harmonic distance restraint between the residues forming the linker between ATPase lobe 1 and lobe 2 with a maximum distance $d_{\text{max}} = 10\,\text{Å}$.

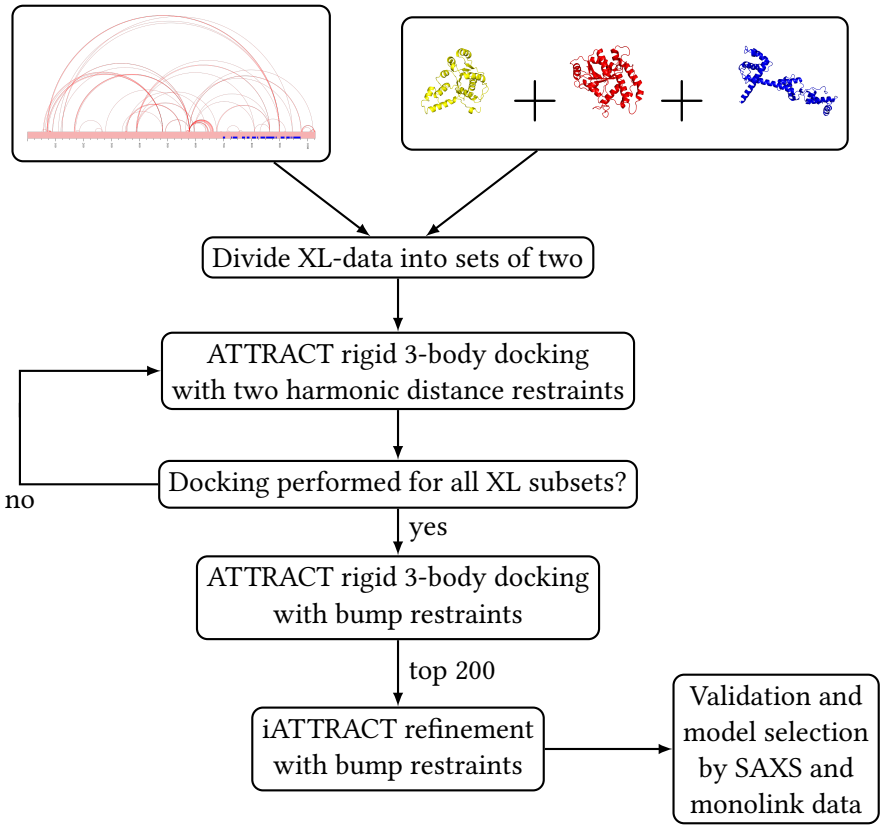


Figure 11.2. Flowchart of the XL-MS driven integrative modeling protocol in ATTRACT.

11.2.6. Analysis and model validation

The final models were assessed with respect to whether they were consistent with the experimental data used during the docking and additional data. We mapped recorded mono-link sites onto the structures and also looked at the absence of inter-domain cross-links at certain sites. For ISWI, we further compared the models against the SAXS profile. The model was fitted to the experimental scattering curve using FoXS with default parameters [397, 399]. We calculated the precision of the final model similar to the procedure used in IMP [378] as the minimal average pair-wise C_α RMSD between the top-ranked models that fulfilled the experimental data. The precision among an ensemble of models that are consistent with the input data provides an estimate for the lower bound of the model error [402]. We also performed co-evolution analysis using the GREMLIN web-server with default settings [214, 331] and compared the predicted contacts to our docking models. For protocol validation with BSA, we assessed the docking models by interface root-mean-square deviation (IRMSD), ligand-RMSD (LRMSD) and fraction of native contacts (fnat) according to the criteria established in the blind docking challenge CAPRI [303, 257, 259, 260] (see Chapter 3, Table 3.2).

11.3. Results and Discussion

For predicting the full-length structure of the ISWI protein, we designed an integrative modeling approach (Figure 11.2) combining cross-linking/mass spectrometry (XL-MS) data and docking in the ATTRACT program (ATTRACT-XL) [506, 99]. The goal was to create a method that could deal with conformational heterogeneity and false positives while retaining good convergence properties that are necessary for sampling multi-body problems. ATTRACT-XL represents contacts derived from the cross-linking experiments either as upper harmonic distance restraints or bump restraints during ATTRACT rigid-body docking [99] and iATTRACT flexible refinement [393] (see Methods). Harmonic restraints are employed in the initial stages of the protocol to promote convergence of docking solutions by applying a force to the atoms that do not fulfill the restraints. In contrast, the bump potentials favor formation of contacts. However, atoms that are above the distance threshold do not experience any forces and hence, violation of XL-MS restraints is permitted yielding increased tolerance for false positives and conformational heterogeneity in the later stages of the integrative modeling protocol. Initially, several million structures were sampled. The generated models were then analyzed and validated based on additional experimental data (see Methods for details).

11.3.1. ATTRACT-XL protocol validation

To test and validate the ATTRACT-XL approach, we obtained BS^{2}G XL-MS data for bovine serum albumin (BSA) and used docking to reconstruct the structure of the full-length protein. We divided the protein into three equally sized parts and used 14 cross-links between these "domains" to model the full-length structure (see Methods). First, we validated the cross-links on the crystal structure of BSA (PDB 3V03) [289] and found a false positive rate of 21% (3 cross-linked residue pairs had a $C_\alpha - C_\alpha$ distance larger than $25\,\text{Å}$ in the crystal structure, Table F.1).

For the modeling procedure, we used the XL-MS data and two connectivity restraints for the linkers between the "domains" allowing a maximum separation of $10\,\text{Å}$ between C and N atom of residues i and $i+1$. The refined 200 models all correctly reproduced the overall topology and fulfilled all 11 true positive cross-links. The protocol also identified two of the false positive cross-links, since these cross-links were not fulfilled in any of the final solutions. We assessed our docking model against the crystal structure of BSA and found a CAPRI two-star model at rank 1 (IRMSD $= 1.6\,\text{Å}$ and fnat $= 0.69$, Figure F.1). The precision calculated over all 200 models of $5.8\,\text{Å}$ confirmed the convergence of our integrative modeling approach and gave us confidence to use ATTRACT-XL for investigating the structure of nucleosome remodeling enzymes.

11.3.2. apo ISWI ATPase domain

We wanted to model the structure of the ISWI ATPase domain and full-length ISWI to better understand ISWI's biological function and regulation mechanisms. The ATPase domain contains two RecA-like lobes connected by a flexible linker. When analyzing crystallized structures of other SF2 superfamily members [436, 177, 120], we found that the internal structure of each of the lobes is well conserved but the relative orientation of the lobes can differ drastically (Figure F.3). These different orientations could be related to different functional states of the ATPase. Previous experiments also showed that ISWI ATPase undergoes a conformational transition upon DNA binding [78]. We identified a set of high-confidence cross-links between the two ATPase lobes including previously published data by Forné et al. [136] (Table 11.1). When mapping the measured cross-links to known structures of SF2 ATPases [436, 177, 120], several lobe-lobe cross-links were violated (Figure F.3). Therefore, additional modeling was necessary to resolve the conformational state of the ISWI ATPase domain that is compatible with the available cross-linking data. We first performed a negative control and investigated the conformations predicted by the ATTRACT software without experimental information [99]. Therefore we performed a

Table 11.1. Cross-linking data between ISWI ATPase lobe 1 and lobe 2.

Residue 1	Residue 2	Upper limit for CA-CA distance [Å]
350	548	25.0
168	578	15.0
169	578	15.0
172	578	15.0
120	578	15.0
338	483	15.0

docking run in which we used only the distance between residues 351 and residues 352 as a restraint (sequence connectivity). To account for the flexibility of the linker, we restrained the N-C bond between the connecting residues to a distance of < 10 Å. The orientation of the lobes in the top scoring docking model is similar to that of the SWI2/SNF2 chromatin-remodeling domain in Rad54 (PDB 1Z3I) [436] (Figure F.2). The two conserved motifs implicated in catalysis (DEAH residues 256–259 and QAMDRAHR residues 536–543) are found in close proximity. The top-ranked model predicted by ATTRACT ab-initio docking could be similar to an active conformation of the ATPase domain. Our empirical force field favors this conformation and thus does not in general bias towards inactive conformations. We analyzed the model with respect to the distances of the residues for which cross-links had been identified (Table 11.1). Apart from the cross-link between residues 350 and 548, which just reflects the connectivity between the lobes, none of the cross-links could be fulfilled by this active ATPase state (Figure F.2).

We performed two-body docking of lobe 1 and lobe 2 using six high-confidence inter-lobe cross-links as harmonic distance restraints (Table 11.1) using the standard ATTRACT protein-protein docking approach [99]. The final 200 models were very similar to each other (precision 2.3 Å) and were compatible with the high-quality photo-crosslinks. The top-ranked model is shown as a representative of the ensemble in Figure 11.3. Note that photo-crosslinks from different sites are the results of independent experiments (e.g., BPA photo-crosslink sites at residue 483 and residue 578). Furthermore, the absence of chemical crosslinks to lobe 2 at sites 171, 174, 186, 197 and photo-crosslinks to lobe 2 at sites 190 and 199 (in the Chd1 crystal structure, these positions are close to lobe 2) and the compatibility of the model with the mono-link data (absence of cross-linking) validate the model (Figure 11.3). We also compared our docking model and homology models based on previously resolved

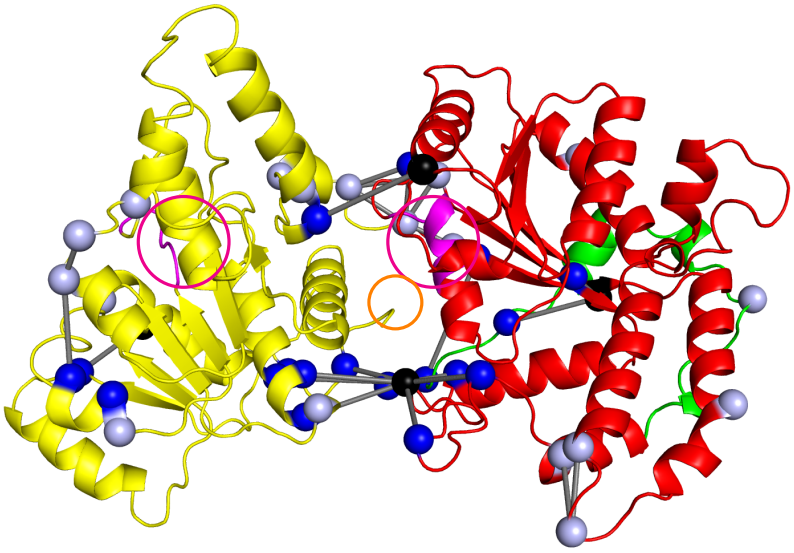


Figure 11.3. Model of ISWI ATPase domain with a novel orientation of lobe 1 (yellow) and lobe 2 (red, the C-terminal bridge domain is drawn in green). Cross-links are shown in gray, sites of photo-crosslinks are marked by black spheres. Sites where mono-links (in addition to cross-links) were detected are shown in light blue. The catalytic motifs are highlighted in magenta and the N-terminus of lobe 1 (residue 116) by an orange circle.

crystal structures of SF2 ATPases [436, 177, 120, 493] to SAXS profiles obtained for the ISWI ATPase domain (residues 26–644; Bruetzel and Lipfert, unpublished data). We used FoXS [399] and CRYSOL [431] to calculate the χ score allowing fits for excluded volume and hydration shell parameters (default settings). The results are listed in Table F.2. Indeed, our docking model fits very well to the experimental SAXS data ($\chi < 1.5$ for both FoXS and CRYSOL, Figure 11.4). Our model displayed better correspondence to the SAXS data than most other homology models (except for the model based on the crystal structure of Chd1).

The generated docking models for the ATPase domain, however, were completely different from all existing experimental structures of SF2 ATPases [436, 177, 120] (Figure F.3) and also differed from an earlier interpretation of the photo-crosslinking data (Figure F.4 (a)) [136]. Most strikingly, the catalytic DEAH and QAMDRAHR motifs on lobe 1 and lobe 2 were not oriented towards each other (highlighted in magenta in Figure 11.3). Since the conserved motifs were not in proximity, the predicted model should represent an inactive conformation. Transition to an active conforma-

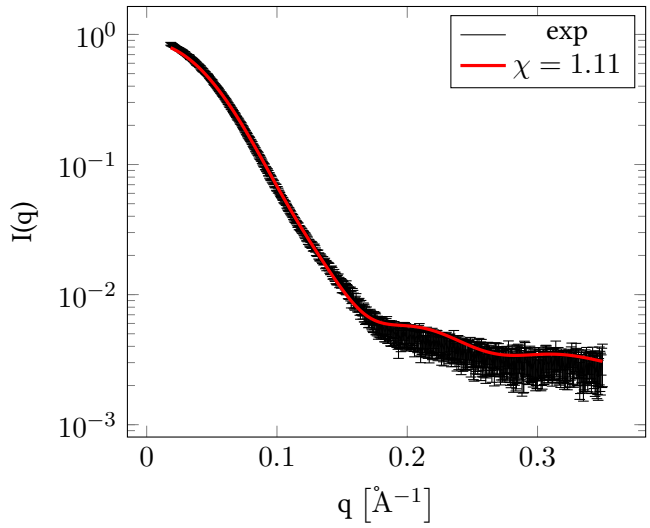


Figure 11.4. Fit of ISWI ATPase docking model to SAXS profile of ISWI ATPase (residues 26–644) using FoXS [399] with default settings. The experimental data are shown in black, the calculated scattering profile in red.

tions would require a large rotation of the lobes. We would like to emphasize that this is purely a result of the experimental data since the docking without restraints favored conformations with these motifs oriented towards each other. Interestingly, the predicted docking model placed the N-terminal part of the ATPase domain close to the lobe-lobe interface (marked orange in Figure 11.3). The cross-linking dataset contained several cross-links between residues on lobe 2 and the NTR (Harrer, unpublished data). The location of residue 116 in this docking model would permit modeling of the NTR (especially the AutoN+acidic patch motifs) close to the lobe-lobe interface as a bridge between the lobes locking the ATPase in an inactive conformation. Release of the NTR from a binding site at lobe 2 could then trigger a conformational rearrangement of the lobes.

Recently, Xia et al. published a crystal structure of the ATPase domain of Swi2/Snf2 chromatin remodeler in a resting state [493]. Swi2/Snf2 is also a member of the SF2 superfamily and its ATPase domain has approximately 38 % sequence identity with ISWI's ATPase domain. Interestingly, a previously unseen orientation of lobe 1 with respect to lobe 2 was found (Figure F.5). This conformation is distinct from our proposed docking model, however, similarly to our model, the catalytic motifs are not oriented towards each other and a transition to the active state would require a large

conformational change. The structure of Swi2/Snf2 displayed large flexibility of the loop in which residue 578 resides in ISWI, the loop was disordered in the crystal structure (this region is not very conserved). When taking large flexibility for this region into account, all cross-links to residue 578 could be fulfilled by the Swi2/Snf2 resting state conformation (C_α-C_α distances $\approx 20\,\text{Å}$). However, the cross-link between residue 483 and residue 338 is incompatible with this conformation (Figure F.5). Note that Swi2/Snf2 does not contain an AutoN or an acidic patch motif and is probably regulated in a different fashion in accordance with its different biological function [493].

In order to test experimentally whether ISWI ATPase indeed adopts an inactive conformation similar to our proposed docking model, we analyzed the interface and contacts and compared these to the results of co-evolution analysis of the ISWI ATPase domain. Co-evolution patterns were evaluated with the GREMLIN web-server using default parameters [25, 214, 330, 331] for ISWI residues 1–637 (for the full length proteins, the number of available sequence in the multiple sequence alignment was not sufficient). We identified one high-probability contact (K337-D485) that could not be mapped to the active state and corresponded well to our proposed inactive conformation. This residue pair could form a salt bridge in our model (alternatively the salt bridge could be formed between residues 337 and 484 as shown in Figure F.6). It is also located close to the photo-crosslink between residue 483 and residues 338. In the future, we will investigate ATPase activity in charge mutants involving this potential salt bridge (K337D, D484K/D485K). Increased ATP hydrolysis of the mutant protein in comparison to the wild-type would validate our model of the inactive state.

11.3.3. AutoN+acidic patch and H4 tail binding to ATPase lobe 2

Our docking results for the ISWI ATPase domain suggested that ISWI auto-inhibition could be mediated by binding of the NTR to lobe 2. In order to identify possible binding sites of the NTR, in particular the AutoN and acidic patch motifs, we used the pepATTRACT peptide-protein docking protocol [391] to model the interaction of the peptides DHRHRKTEQ (residues 89-97, contains RHRK AutoN motif) and EQEEDEELL (residues 96-104, contains EEDEE acidic patch motif) with lobe 2. We further used the PEPFOLD2, PEPFOLD3 and I-TASSER web servers [413, 435, 247, 513, 373] to predict the structure of the entire module (DHRHRKTEQEEDEELL). We found high agreement between the different methods indicating an α-helical structure that gave us increased confidence in the prediction. We then docked the resulting structural ensemble for the AutoN+acidic patch region (13 conformations) to lobe 2 [99]. The docking was performed without restraints (ab-initio docking).

Interestingly, each of the three docking runs yielded a large cluster of solutions in the vicinity of the catalytic motif on lobe 2 and residue 578 (Figure F.7). In the following, we will therefore focus on the results for docking the entire module to lobe 2. One of the top-ranked docking models is shown as an example in Figure 11.5. The AutoN and the acidic patch motif bind to a pocket in between the catalytic motif and residue 578. The selected docking model displays extensive interactions of NTR residues that had been previously identified as important for auto-inhibition (residues R91, R93, E98, D100, E102 from [78] and Ludwigsen, unpublished data) with residues of lobe 2. We analyzed the contacts between AutoN+acidic patch motif and lobe 2 in all docking models. The residues that are most frequently in contact and especially those that form contacts with important residues, are listed in Table F.3. In the future, the proposed binding site could be validated by mutating some of these possibly important lobe 2 residues. Charged residues would be especially promising candidates for such mutational studies.

Since the presence of the histone H4 tail can lift AutoN-related inhibition of ISWI ATPase activity [78], we predicted that the H4 tail also binds close to the AutoN+acidic patch binding site on lobe 2. Competitive binding to this pocket could then explain ISWI's auto-inhibition and H4 tail dependence. This hypothesis was probed in XL-MS experiments of ISWI and Snf2h (human homologue of ISWI) both in the presence of a H4 tail-derived peptide and the nucleosome. Indeed, we identified several cross-links between the histone H4 tail and residues on lobe 2 (Table F.4). We used three high-confidence crosslinks from this data set to model H4 tail (residues 1-20, TGRGKGGKGLGKGGAKRHRK) interaction with the pepATTRACT protocol using the XL-MS data as harmonic restraints [391]. A representative model is shown in Figure F.8. The docking model fulfilled the three high-confidence cross-links within a distance of 20 Å and was also compatible with several lower-confidence cross-links (Table F.4). However, a set of lower-confidence crosslinks to residues 578 and 568 was irreconcilable with those models. This indicates that the H4 tail displays high flexibility and probably binds only with a short motif involving the basic patch (residues 17–20) to lobe 2 while the rest of the tail can adopt different conformations. It is important to keep in mind that cross-linking can only capture information about (temporary) physical proximity of residues, however, does not give information on whether these residues are engaged in more permanent, biologically relevant contacts. In any case, the H4 tail-lobe 2 docking models and those that could be generated based on contacts to residues 578 and 568 are compatible with binding of the H4 basic patch to a location close to the proposed AutoN+acidic patch binding site.

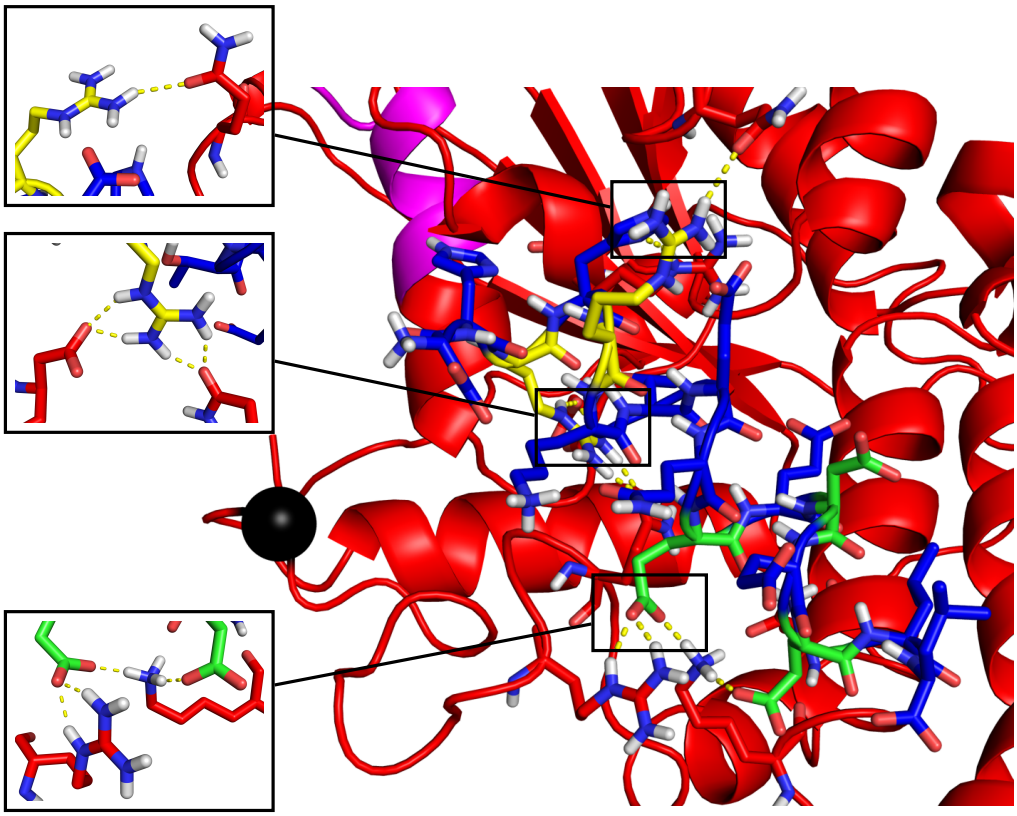


Figure 11.5. Possible AutoN+acidic patch binding mode on ISWI ATPase lobe 2 (red) identified by pepATTRACT [391]. Single point mutations R91A and R93A (yellow) decrease ISWI auto-inhibition [78]. The triple mutant E98Q D100N E102Q (green) also displayed higher ATPase activity (Ludwigsen, unpublished data). Insets show detailed interactions of these important AutoN+acidic patch residues with lobe 2.

Table 11.2. Cross-links for full length ISWI used in three-body docking of ATPase lobe 1, lobe 2 and HSS.

Residue 1	Residue 2	Upper limit for CA-CA distance
865	391	25.0
900	391	25.0
865	388	25.0
865	595	25.0
810	595	25.0
945	124	25.0
945	564	25.0
810	388	25.0
350	548	25.0

11.3.4. apo ISWI full-length structure

XL-MS experiments yielded several contacts between the ATPase lobes and the HSS domain. We used this information to predict the structure of the full-length ISWI protein with our integrative modeling approach (multi-body docking of lobe 1, lobe 2 and HSS; Figure 11.2). In addition to the photo-crosslinks (Table 11.1), we collected 9 chemical cross-links (Table 11.2). The cross-link between residue 124 and residue 945 was also found in Snf2h XL-MS experiments (Harrer, unpublished data).

During the docking run, 50 models were generated that were fully compatible with all the cross-linking restraints. These top-ranked 50 models contained only one particular arrangement of the domains showing a high degree of convergence (precision among top 50 models 4.2 Å). We scored all the top-ranked models by comparison to ISWI full-length SAXS data (Bruetzel and Lipfert, unpublished data) using FoXS [397, 399]. We then selected the docking model with the lowest χ score as a representative (rank 25, $\chi = 1.08$, Figure F.9). The representative docking model is shown in Figure 11.6. The model is compatible with the cross-linking, mono-link and SAXS data. In our model, lobe 2 contacts HSS mainly via its slide domain, the interface overlaps partially with a known DNA binding site on HSS. Figure 11.7) shows the top 6 docking model fitted manually into an envelope derived from the SAXS measurements. Recently, three additional photo-crosslinks were obtained between lobe 2 and the HSS domain (578-951, 578-952 and 578-942). These crosslinks are compatible with our current model, especially when considering increased flexibility at position 578

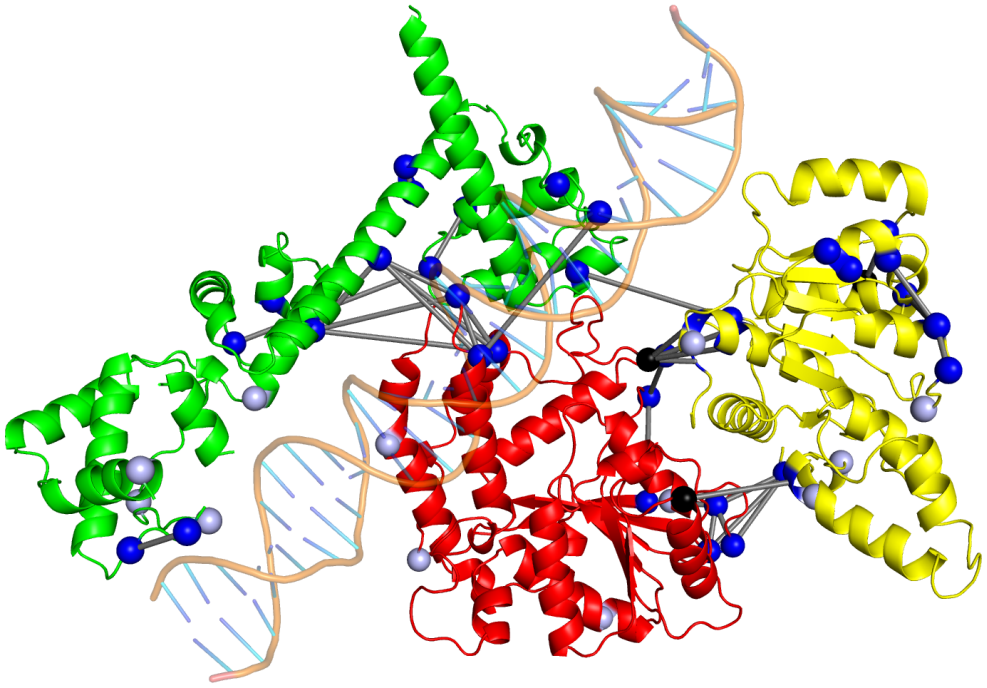


Figure 11.6. Results for 3-body docking (lobe 1 yellow, lobe 2 red, HSS green) with cross-linking restraints. Cross-links are shown in gray, sites of mono-links in light blue. A structure of HSS bound to DNA (PDB 2Y9Z) was superimposed on the docking model.

(see above).

Even though we generated a model that is compatible with the available XL-MS and SAXS data, we cannot exclude the possibility that the ATPase domain could also be bound; e.g., on the other side of the HSS domain (the dimension of HSS in one direction is comparable to the range of the cross-linking agent). It might be possible that only one or two crosslinks had a major effect on the generated conformations. To test this, we will collect a large set of lower-confidence cross-links in the future and repeat the docking with datasets of different sizes.

11.3.5. HSS binding on nucleosome

Recently, Leonard and Narlikar detected binding of the human homologue Snf2h's HSS domain to the nucleosome core in the presence of the ATP analogue ADP-BeFx [263]. Previously, it was only known that HSS could bind to linker DNA. We used

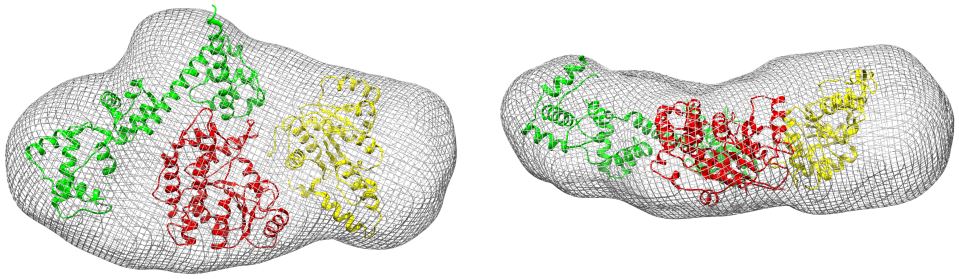
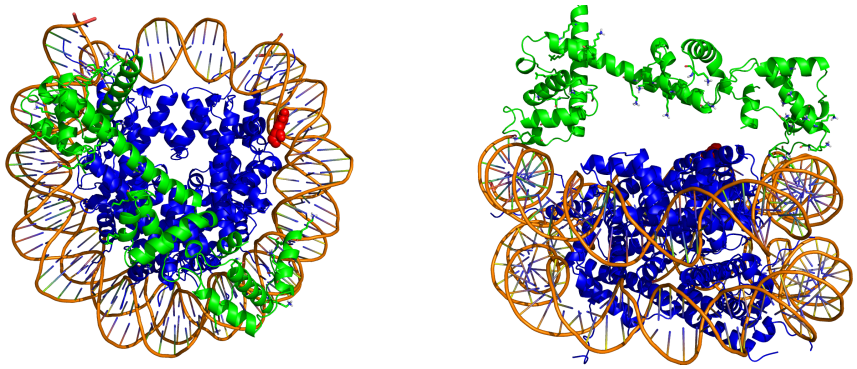


Figure 11.7. Results for 3-body docking with cross-linking restraints. The docking model was fitted manually into a SAXS envelope. The figure was created with UCSF Chimera [500].

ATTRACT ab-initio docking [99] to investigate possible binding modes. The ISWI HSS domain was docked to the Drosophila nucleosome (PDB 2PYO) [79] without experimental restraints. 500,000 initial starting positions were screened. Interestingly, among the top 50 ranked models, we found a cluster of solutions in which HSS bridges the nucleosome contacting nucleosomal DNA with its HAND and SLIDE domain and spanning histone H2A and its acidic patch (Figure 11.8). Such a position would have interesting implications for the switching of HSS between translocation and DNA length sensing phases as proposed by Leonard and Narlikar [263]. It could also explain the higher affinity of full-length ISWI to nucleosomes compared to ISWI ATPase domain without HSS [315], suggesting that HSS acts as a molecular ruler for detecting nucleosomes. The model would also provide an explanation for the observation that the histone variant H2A.Z (with an extended acidic patch) stimulates ISWI activity [155]. The model is compatible with previous FRET data [263] assuming a Foerster radius of 60 Å (the resolution of FRET experiment was too low to make more precise predictions on distances). An alternative model in which HSS binds only to the DNA is shown in Figure F.10. In this model, the binding of DNA to the SLIDE domain is similar to the crystallized structure of HSS with DNA bound (PDB 2Y9Z).

Based on the docking results, we predicted crosslinking between the acidic patch on histones H2A/H2B and HSS in cross-linking experiments of ISWI in the presence of the nucleosome and the ATP-analogue ADP-BeFx. Indeed, cross-linking of Snf2h to the nucleosome yielded several contacts between H2A and H2B (including residue H2B 105 and H2B 117 close to the acidic patch) and HAND and SLIDE domain (Harrer, unpublished data). However, these data could probably not be fulfilled by a one-state model (the distance between cross-linked residues on HAND and SLIDE domain is significantly larger than the range of the cross-linking agent). This further



(a) Top view. **(b)** Side view.

Figure 11.8. Ab-initio docking of HSS to Drosophila nucleosome core. The top ranked model is shown (HSS green). The HSS domain binds to DNA spanning across the histone core contacting nucleosomal DNA with its HAND and SLIDE domains. For reference, the H4 tail is shown in red.

underlines the dynamic nature of the system and might also point to flexibility in the HSS domain. Further analysis and experiments are necessary to distinguish between multiple states and unravel the conformational dynamics of this flexible enzyme.

11.4. Conclusion and Outlook

ATP-dependent nucleosome remodelers ensure dynamic accessibility to the genome, however, atomic structural insights into their regulation and biological function are limited to date. Here, we used different docking approaches to elucidate the structure of the full-length Drosophila melanogaster ISWI protein. We proposed a novel inactive state of the ISWI ATPase domain that is compatible with available XL-MS and SAXS data. This inactive conformation yields a structural rationale for ISWI's auto-inhibition by its N-terminal region (NTR). We further derived potential mutations for validating this ATPase structural model experimentally. Using our previously developed peptide-protein docking approach [391] (Chapter 6), we identified a binding site for the regulatory AutoN+acidic patch motif in the NTR and the histone H4 tail on lobe 2. We proposed possible mutations to validate this auto-inhibition mode experimentally. Finally, we modeled the structure of the full-length ISWI protein. In our model, the HSS domain docks against the ATPase domain contacting lobe 2 with its SLIDE domain. HSS binding to DNA would require release from the ATPase domain. Our model suggests that ISWI may need to undergo conformational changes to reach an active remodeling state. This study yields insights into the intri-

cate regulation mechanisms of nucleosome remodeling enzymes. In the future, we plan to model ISWI's conformation in the presence of different nucleotides and its nucleosomal substrate and to extend our investigations to other remodeler families. Furthermore, the ISWI protein is usually a part of larger nucleosome remodeler complexes. The structure of those complexes and ISWI's interactions with other proteins could also be the subject of future research. Our presented integrative modeling approach ATTRACT-XL can be applied to a wide range of dynamic and transient multi-component assemblies that are difficult to study by traditional structural biology methods.

12. Perspectives

Systems biology has started to move beyond genomics and proteomics and towards tackling the interactome. A large number of protein-protein interactions have been identified and interaction maps for different organisms have been explored using techniques like yeast-two hybrid assays or mass spectrometry. However, in many cases, these experiments only yield qualitative information on the nature of the interaction. In order to understand how the interaction is established, structural knowledge at the atomic level is necessary. Atomic structures allow to identify the interface between the protein partners and the residues that establish important contacts. These insights can then be used to understand the effect of mutations that occur in diseases. Identification of the interface also allows to create small molecules and peptides that can inhibit the interaction, an important strategy in current drug design efforts [16, 317]. Studying the interfaces of protein-protein complexes can further give insights into the physical forces that guide the recognition process. This knowledge can then be used to go beyond the interaction space that Nature offers us: towards re-engineering known protein-protein interactions and even designing new types of protein-protein complexes for nanotechnology. A few successful examples of such approaches have already been published to date [156, 53, 190, 26, 231]. Similar to nanomachines built from DNA [272], protein-protein interactions could be fine-tuned for desired applications and molecular machines for specific purposes could be assembled benefiting from the large variability in shape and the possibility of greater geometric control [53].

Structural genomics initiatives have revealed the folds of many proteins covering most known families. Hence, the unbound structures of individual proteins are often available or can be modeled by homology. But characterizing the structure of protein-protein complexes by techniques such as X-ray crystallography or NMR still remains a huge challenge, especially for weak binders and large multimeric assemblies. Furthermore, the sheer size of the interactome makes experimental structure determination for all complexes infeasible in the near future. Computational protein-protein docking methods can complement experiments by predicting the structure of the complex from the structures of the individual components. The protein-protein docking field has experienced tremendous progress during the last decades and genome-wide docking experiments have finally come in reach.

The performance of different methods in the docking field has been assessed and documented by the blind prediction experiment CAPRI (Critical Assessment of PRediction of Interactions, capri.ebi.ac.uk). The CAPRI challenge was established in 2001 and ATTRACT, the docking program developed in our group, has participated in CAPRI since 2003. The targets presented in CAPRI (Chapter 9) have triggered the development of new features in ATTRACT (Chapter 6 and Chapter 8). The challenges we faced in CAPRI also inspired new applications like the work presented in Chapter 11.

The progress in the docking field is also reflected by the nature of the targets in CAPRI. In the early rounds, most targets were unbound-bound complexes; i.e., a free form was only available for one of the partners. Since 2010, all the targets were unbound-unbound complexes. In many cases, the structures of the individual partners had to be modeled by homology and in some targets, the modeling was very difficult with only distant homologues available as templates. In the future, the fields of computational prediction of protein structures and protein-protein interactions will become more intertwined. This will hopefully also encourage progress for complexes that involve folding upon binding; e.g., intrinsically disordered proteins.

The last CAPRI rounds in 2015 again emphasized the importance of protein flexibility upon complexation. In several rounds, the goal was to model highly flexible peptide-protein complexes. In addition, especially in the most recent rounds, the protein-protein complexes displayed a high degree of backbone flexibility (see also Chapter 9). The aim of this thesis was to improve the ATTRACT docking methodology by explicitly considering interface flexibility. For this purpose, a new refinement method, iATTRACT, was developed that combines fully flexible interface residues with global rigid body optimization of the docking geometry (Chapter 5). This approach significantly improved initial rigid-body docking models for a large variety of protein-protein complexes. The approach was then incorporated into an ab-initio peptide-protein docking protocol (pepATTRACT) described in Chapter 6. pepATTRACT was one of the first fully blind methods for predicting peptide-protein complex structures with a performance close to two state-of-the-art local docking approaches. The previously developed iATTRACT method contributed to the success of the pepATTRACT protocol and hence allowed to extend the ATTRACT docking approach towards the important and challenging class of peptide-mediated interactions. pepATTRACT was successfully applied to ab-initio peptide-protein docking in CAPRI (Chapter 9), predicting peptide binding in the ER associated degradation pathway (Chapter 10) and modeling of auto-inhibition in the ISWI nucleosome remodeling enzyme (Chapter 11).

The peptide-protein docking protocol also formed the basis for the development of an interface loop modeling procedure (loopATTRACT, Chapter 7). We envision that

this protocol will be used to model large conformational changes of interface loops in a hierarchical integrative modeling protocol using the ATTRACT-EM [94, 96] or the ATTRACT-SAXS protocol [392] (Chapter 8) as the first modeling stage. These two protocols only use soft repulsive potentials between the protein partners that allow a certain degree of steric overlap. Previous analysis on ATTRACT-EM models demonstrated that steric clashes in the models are often linked to protein regions that undergo conformational change upon binding [96]. Hence, loop regions that have been identified as potentially flexible during the first stage of the integrative modeling protocol can be remodeled by loopATTRACT. Such a clash analysis has not yet been performed for the ATTRACT-SAXS protocol where discrimination between near-native and non-native models is generally harder than in ATTRACT-EM, however, it is likely that the same principle also applies to ATTRACT-SAXS generated models.

Apart from flexibility, a big trend in the field is using data in docking. Including external data during docking can both improve the sampling by limiting the search space and improve the scoring of near-native models compensating for inaccuracies in the approximate energy function. Data-driven docking was first started by the Bonvin group (HADDOCK) and the Sali lab (IMP) more than ten years ago. These methods were the first to systematically incorporate experimental low-resolution data into the modeling procedure. Since then many groups have adapted their software and integrative modeling/data-driven docking has become the standard in the field whenever suitable data are available. Adding support for additional types of experimental data to ATTRACT was the second strategy used in this thesis to improve the accuracy of our docking approach. Chapter 8 describes a new integrative modeling approach using small-angle X-ray scattering (SAXS) data. In contrast to previous methods that used SAXS data as an a-posteriori filter, the ATTRACT-SAXS approach directly generates docking models compatible with the experimental data. The method showed especially good performance for medium and hard docking cases; i.e., complexes that undergo conformational change upon binding, and outperformed two earlier SAXS-driven integrative modeling approaches. Another successful integrative modeling protocol in ATTRACT is described in Chapter 11. Through the combination of cross-linking/mass spectrometry, SAXS data and docking, the full-length structure of the ISWI nucleosome remodeling enzyme was predicted. We derived testable hypotheses from the docking models that have already been partly confirmed by experiments.

In the last couple of years, the concept of data-driven docking was further extended towards bioinformatics approaches making use of sequence information. de Vries et al. successfully used interface predictions from a consensus predictor as an input to the HADDOCK program instead of/complementary to experimental data

[95]. Recently, Baker and coworkers showed that information about contacts in protein-protein complexes can be extracted from residue co-variation patterns when analyzing large sequence datasets. They further demonstrated that co-evolution data can be successfully used in protein-protein complex assembly using the Rosetta program [330]. However, for extracting reliable co-evolution data, a large number of sequences are needed ($\approx 5L$ with L being the length of the target sequence) [214, 330]. Guerois and coworkers presented an approach which uses sequences from orthologous complexes to score generated docking models. Using multiple sequence alignments from only 10 to 100 different organisms, the method showed a clear improvement with respect to standard scoring functions [11]. The Guerois group was the best predictor in the 2016 CAPRI evaluation generating the largest number of medium and high quality predictions (Lensink, personal communication). It is very likely that using evolutionary information will become as much a standard in docking as using experimental data is today and that this orthogonal information will improve the accuracy of current approaches. A detailed discussion and systematic comparative evaluation on different ways of including evolutionary information in docking is, however, outside of the scope of this thesis and will be the subject of future research.

Despite the degree of success achieved in this thesis, a large obstacle to flexible docking remains to accurately predict whether and to what extent conformational changes occur upon complexation. On the level of single protein structures, a variety of simulation approaches have been developed to study flexibility. Karaca and Bonvin found that the cumulative sum of eigenvalues obtained from an elastic network calculation had some predictive power with respect to the extent of the conformational change [218]. Molecular dynamics (MD) simulations can also be used to gain insight into protein dynamics at the picosecond to microsecond timescale. However, depending on the size of the protein, these calculations may still be too computationally expensive. In addition, many conformational rearrangements like domain motions are currently beyond the time-scales accessible in MD simulations. Recent investigations of Gray and coworkers (Gray, personal communication) point towards a strong contribution of induced fit effects in the encounter complex. In other words, the conformational rearrangements and the associated protein flexibility might only become favorable when the proteins are in close proximity and cannot be sufficiently predicted from simulating single protein structures. Hence, MD simulations should in principle be performed on different encounter complexes on the multi-nanosecond time-scale. Such an approach is too computationally demanding, since trajectories for hundreds or thousands of possible docking models (encounter complexes) need to be evaluated. Recently, Oliwa and Shen proposed to apply normal mode calculations to encounter complexes and found slight improvements in approximating

the bound-unbound transition compared to conventional normal mode calculations [328]. While some of these examples are quite promising, reliable general-purpose prediction approaches for binding-induced flexibility are still lacking to date. Tackling the hard docking cases will certainly require improvements in this direction.

Another hurdle towards accurate (flexible) docking is the scoring of near-native solutions. Currently, discrimination between non-native and near-native docking models is still insufficient; i.e., in many cases non-native models that have a similar or even lower score than near-native structures can be generated. In many scoring functions, there is usually only a small energy gap between near-native and non-native docking models (if any at all). However, in vivo, there is often strong discrimination (which is also crucial to survival) and small changes via single point mutations can lead to a reduction in binding affinity by several orders of magnitude. These effects cannot be reproduced by current scoring functions. Insufficient scoring is a particularly big problem in flexible docking where an accurate energy function is crucial in order to sample the bound conformation. Incorrect scoring can in contrast drive the structure further away from the native state both in protein structure and overall complex geometry. Inaccuracies in scoring also impede further applications of docking going beyond mere structure prediction towards studying binding affinity and specificity and the effect of mutations and ultimately protein-protein interaction design.

The inaccuracies in scoring functions result from approximating and neglecting important physical interactions. In most commonly used scoring functions, solvent-solvent and solvent-protein interactions are not taken into account or approximated in a very crude fashion. Furthermore, the internal energy of the individual proteins in their respective conformational states is usually not considered. Also entropic effects are usually neglected. On a more basic level, the approximations and the simple functional forms used to represent the quantum-mechanical interactions between the electron densities might be reaching their limits. For example changes in protonation states and explicit polarizability might yield important contributions that could help to properly distinguish the native complex from other loosely bound geometries. Improvements to physics-based force fields especially with respect to including explicit polarizability are currently also an important topic for the molecular dynamics community and progress in this field will certainly help advance scoring in protein-protein docking. Note that a more detailed energy representation should be matched and adapted to an increased level of detail in sampling and flexibility. However, this might still be outside of the scope of current computing capabilities. Finally, it is possible that the shortcomings in current scoring functions indicate a partial lack of understanding in the physical principles that govern the many-body problem of protein-protein association. Using empirical data and creating knowledge-based

potentials provides a possibility to circumvent this lack of understanding. Indeed, many groups have adopted this strategy and the resulting knowledge-based scoring functions often outperform physical scoring functions. This observation supports the notion that current physics-based energy functions might still be missing some important contributions.

Last but not least, I would like to emphasize that a docking model is usually a starting point in a real-life application providing testable hypotheses that can guide further experimental studies. In other words, validation of docking results is crucial and ideally computational modeling and experimental studies should go hand in hand. Unfortunately, common standards for model validation and error estimates are still lacking to date. Model accuracy; i.e., expected deviations of docking models from the native structure, is typically inferred from benchmark tests on known structures. But for many applications, appropriate benchmarks are not yet available (see Chapter 8 and 11). Creating suitable benchmarks and reliable estimates for model quality (similar to R_{free} in X-ray crystallography) will be important steps towards establishing docking models as a standard resource in structural biology in the future [385].

A. iATTRACT Supplemental Data

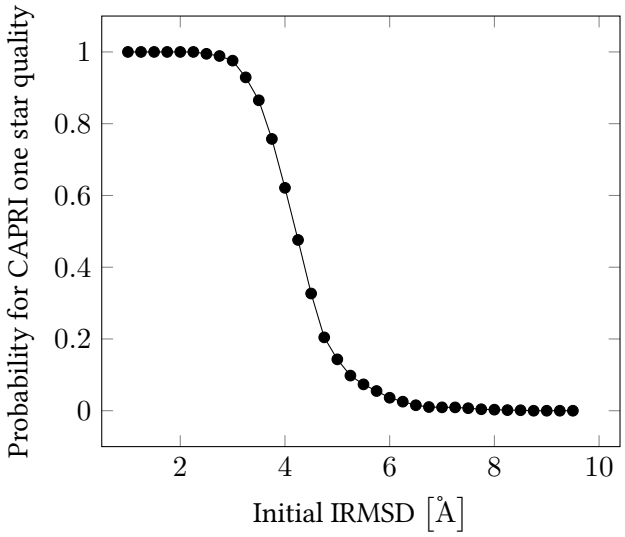


Figure A.1. Probability to yield CAPRI quality refined structures by i-ATTRACT refinement vs. initial IRMSD. The probability was calculated based on the whole set of benchmark results.

Table A.1. Average IRMSD and fnat change in interface refinement for different classes of protein complexes.

Protein type	$\overline{\Delta\text{IRMSD}}/\text{Å}$	$\overline{\Delta\text{fnat}}$
all	$-0.293(5)$	$0.071(1)$
antibody	$-0.328(11)$	$0.069(1)$
enzyme	$-0.296(10)$	$0.070(1)$
other	$-0.280(6)$	$0.071(1)$
easy	$-0.282(6)$	$0.076(1)$
medium	$-0.327(12)$	$0.072(1)$
hard	$-0.293(14)$	$0.042(1)$

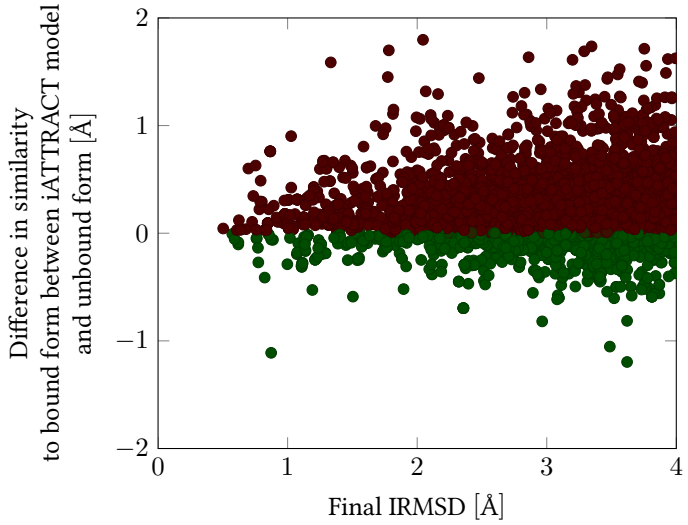
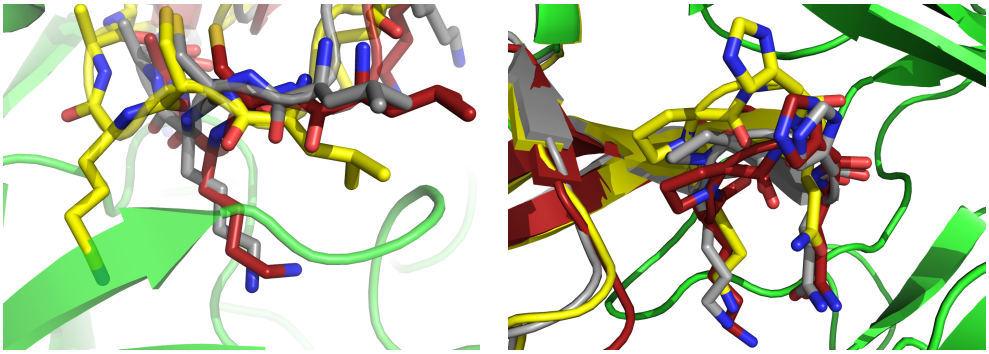


Figure A.2. Improvement of the interface conformation of protein partners after refinement with iATTRACT. The heavy-atom IRMSD of all residues within 3 Å contact distance of the interface was calculated for protein partners of each iATTRACT model and the corresponding unbound protein partners after best superposition onto the bound structures. The difference between $\text{IRMSD}_{\text{heavy}}$ of the refined models and the unbound structure is plotted on the y-axis vs. backbone IRMSD of the iATTRACT model. A negative difference indicates improved interface conformation compared to the unbound protein (opposite for positive values). iATTRACT models where protein structures have moved closer to the bound form at the interface are labeled by green marks.



(a) 2UUY **(b)** 1BJ1

Figure A.3. Examples of significant interface residue reorientations in iATTRACT which make the structure more similar to the bound form. The receptor protein of the iATTRACT model is shown in green (cartoon), the ligand protein (after iATTRACT refinement) in red. Selected interface residues are shown as sticks. The crystal structure of the bound ligand protein is shown in gray. The unbound form of the protein was superimposed with respect to the bound structure and drawn in yellow. For 2UUY, a lysine residue whose unbound conformation causes clashes has reoriented correctly. For 1BJ1, an interface histidine and a glutamine residue adopt conformations close to the bound form.

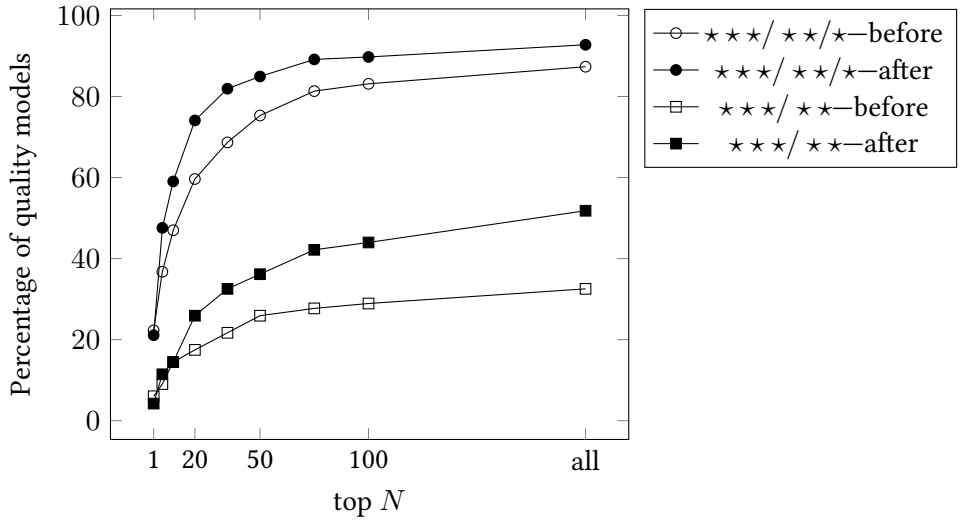


Figure A.4. Scoring of refined docking solutions for an enriched decoy set. The ranking of quality solutions by the OPLS energy function is compared before and after iATTRACT refinement.

Table A.2. Scoring before and after iATTRACT refinement. For each protein complex the best IRMSD of the top 5 ranked structures and the IRMSD of the top ranked structure before and after iATTRACT refinement is shown. The structures were ranked by their intermolecular energies based on the OPLS energy function.

	Best IRMSD in top 5		IRMSD of top 1			Best IRMSD in top 5		IRMSD of top 1	
PDB	Before	After	Before	After	PDB	Before	After	Before	After
1A2K	5.1	2.73	8.42	3.02	1NCA	0.41	3.26	0.41	6.5
1ACB	6.12	4.25	6.12	4.25	1NSN	2.54	3.14	2.54	3.14
1AHW	1.39	3.46	2.54	6.38	1NW9	5.52	5.38	7.12	7.44
1AK4	4.66	4.93	6.65	4.93	1OC0	5.85	3.34	5.85	6.46
1AKJ	2.43	2.88	3.31	3.22	1OFU	10.18	9.54	11.31	10.21
1ATN	6.05	5.64	6.89	10.67	1OPH	3.35	6.07	7.4	6.28
1AVX	1.85	1.6	2.37	1.94	1OYV	2.55	2.08	2.55	2.27
1AY7	8.42	7.73	8.42	9.87	1PPE	2.09	3.46	2.1	5.34
1AZS	6.55	2.08	6.55	6.77	1PVH	4.96	6.52	7.02	6.52
1B6C	4.91	4.85	5.06	6.81	1PXV	7.47	3.59	8.1	3.59
1BGX	6.97	6.61	7.47	6.99	1QA9	3.86	2.45	3.86	2.45
1BJ1	3.58	0.76	9.31	12.05	1QFW	0.83	1.45	0.83	4.4
1BKD	11.03	7.43	11.03	10.88	1R0R	4.53	2.25	5.58	6.25
1BUH	1.97	5.13	5.8	6.29	1R6Q	4.58	3.89	4.58	4.24
1BVK	6.33	6.06	6.89	6.06	1R8S	8.05	5.96	8.05	7.95
1BVN	3.98	3.99	6.62	10.74	1RLB	5.99	5.72	8.49	7.46
1CGI	5.84	5.27	5.84	5.88	1RV6	6.85	5.42	8.13	6.64
1CLV	5.47	3.54	5.95	4.31	1S1Q	1.4	5.37	1.4	5.8
1D6R	3.22	3.6	4.45	3.6	1SBB	7.24	6.53	7.24	6.55
1DFJ	7.73	6.94	7.73	7.34	1SYX	3.08	2.95	3.08	2.99
1DQJ	7.87	2.45	7.87	8.01	1T6B	9.77	5.79	10.57	9.62
1E4K	7.74	7.46	7.74	11.19	1TMQ	9.29	6.84	9.29	6.84
1E6E	1.88	1.46	7.32	2.17	1UDI	3.9	1.49	5.6	6.84
1E6J	4.96	5.04	4.96	5.42	1US7	5.81	4.13	6.6	6.85
1E96	5.12	6.21	6.32	6.53	1VFB	5.91	4.99	5.91	6.3
1EAW	2.15	2.81	4.56	2.81	1WDW	2.55	1.72	2.62	1.72
1EER	4.09	3.05	4.09	3.05	1WEJ	2.8	5.41	2.8	6.2
1EFN	4.47	2.8	4.47	7.92	1WQ1	6.65	4.32	7.8	4.32
1EWY	5.44	5.54	12.64	5.54	1XD3	4.76	4.13	8.68	6.21
1EZU	10.18	7.79	10.71	9.93	1XQS	2.87	2.03	2.87	2.61
1F34	12.92	12.52	12.92	14.21	1XU1	3.51	4.48	5.03	6.37
1F51	8.62	4.61	10.44	4.61	1YVB	5.55	2.52	7.55	6.9
1FC2	6.62	5.12	6.84	7.32	1Z0K	2.96	4.8	6.42	5.45
1FCC	7.18	6.49	9.07	8.88	1Z5Y	3.95	3.61	3.95	3.61
1FFW	3.68	2.15	5.74	2.15	1ZHH	8.68	7.5	10.08	9.74
1FLE	3.55	3.76	8.06	6.64	1ZHI	5.78	5.18	6.46	6.74
1FQ1	7.21	6.82	16.5	11.78	1ZLI	4.68	4.2	4.68	4.2
1FQJ	2.78	3.39	2.78	3.41	1ZM4	2.72	3.18	2.72	5.35
1FSK	0.84	2.05	1.09	7.74	2A5T	6.28	7.15	7.86	7.15
1GCQ	4.99	4.41	6.39	5.6	2A9K	4.84	4.97	5.82	4.97
1GHQ	5.42	3.8	5.81	6.35	2ABZ	6.61	2.54	6.61	7.24
1GL1	4.23	4.77	4.23	4.9	2AJF	6.87	5.01	6.87	7.61
1GLA	6.35	2.88	14.87	9.35	2AYO	4.74	5.5	4.74	6.01
1GP2	4.48	4.55	6.48	6.84	2B42	9.27	7.18	9.27	8.01
1GPW	3.18	1.82	4.38	1.83	2B4J	6.71	6.02	7.25	7.05
1GRN	2.65	2.01	2.65	5.57	2BTF	2.74	2.82	2.74	2.82
1GXD	2.76	5.15	2.76	6.98	2C0L	6.04	6.06	6.04	7.16
1H9D	4.88	5.99	7.79	5.99	2CFH	2	1.75	2	1.75
1HCF	4.97	4.96	6.41	4.99	2FD6	6.02	4.08	7.58	4.45
1HE1	2.81	2.89	2.81	2.89	2FJU	4.72	2.13	7.48	2.13
1HE8	6.11	4.97	11.12	5.15	2G77	2.58	5.23	7.97	5.23
1HIA	6.24	4.36	6.24	7.34	2H7V	6.77	6.9	7.96	6.9
1I2M	4.32	3.2	4.32	3.2	2HLE	2.5	2.63	3.42	4.34
1I4D	6.25	2.84	7.03	4.93	2HMI	10.67	3.11	13.74	12.32
1I9R	1.6	3.25	1.6	3.25	2HQS	2.5	2.41	7.3	6.07
1IB1	7.85	5.95	9.48	8.41	2HRK	3.99	1.74	7.35	3.93
1IBR	10.37	9.89	12.01	11.9	2I25	6.25	6.16	6.3	6.16
1IJK	5.6	5.2	9.84	10.22	2I9B	5.33	6.38	6.81	7.24
1IQD	5.76	2.14	6.67	6.15	2IDO	6.07	7.62	8.16	9.49
1J2J	4.02	4.16	5.67	7.96	2J0T	11.13	11.2	11.86	11.2
1JIW	5.21	4.58	5.21	6.13	2J7P	5.88	5.8	10.57	9.07
1JK9	3.08	6.3	7.75	6.87	2JEL	2.91	3.46	2.91	4.26
1JMO	6.02	4.29	16.59	4.29	2MTA	4.9	3.49	4.9	5.68
1JPS	1.56	2.02	1.56	5.57	2O3B	9.05	4.38	12.62	5.55
1JTG	1.1	5.31	1.1	5.86	2O8V	3.19	6.32	6.07	10.31
1JWH	6.01	5.89	6.84	7.5	2OOB	4.44	3.85	5.44	5.57
1JZD	5.41	7.94	10.22	7.94	2OOR	8.24	4.39	8.24	7.54
1K4C	2.15	3.71	2.15	4.7	2OT3	8.7	7.01	8.79	8.79
1K5D	6.9	4.66	7.96	11.49	2OUL	0.69	0.87	0.69	8.81
1K74	1.97	2.34	1.97	2.34	2OZA	8.76	2.85	9.71	5.7
1KAC	2.23	2.41	2.23	5.31	2PCC	4.05	3.18	4.05	3.18
1KKL	6.96	5.26	6.96	7.75	2SIC	6.19	3.88	7.35	5.82
1KLU	7.05	3.51	12.46	3.51	2SNI	5.19	1.61	5.19	5.29
1KTZ	0.89	5.62	4.17	6.11	2UUY	4.6	3.56	4.6	4.39
1KXP	8.35	1.32	8.35	1.32	2VDB	6.46	7.4	10.12	9.21
1KXQ	2.26	1.03	5.2	5.06	2VIS	3.32	3.02	6.41	3.02
1LFD	2.93	2.85	5.08	6.88	2Z0E	9.69	9.08	9.71	9.58
1M10	4.97	3.04	5.33	4.81	3BP8	3.87	3.22	6.95	7.19
1MAH	3.46	1.05	3.46	3.66	3CPH	5.46	5.28	7.67	5.28
1ML0	2.74	1.66	3.04	2.31	3D5S	3.54	1.81	3.54	2.94
1MLC	5.35	5.5	6.55	6.88	3SGQ	5.7	4.6	5.75	4.6
1MQ8	3.63	4.95	5.87	6.56	4CPA	1.65	2.13	1.65	5.51
1N8O	6.79	7.08	6.79	7.1	7CEI	2.22	1.86	2.22	1.86

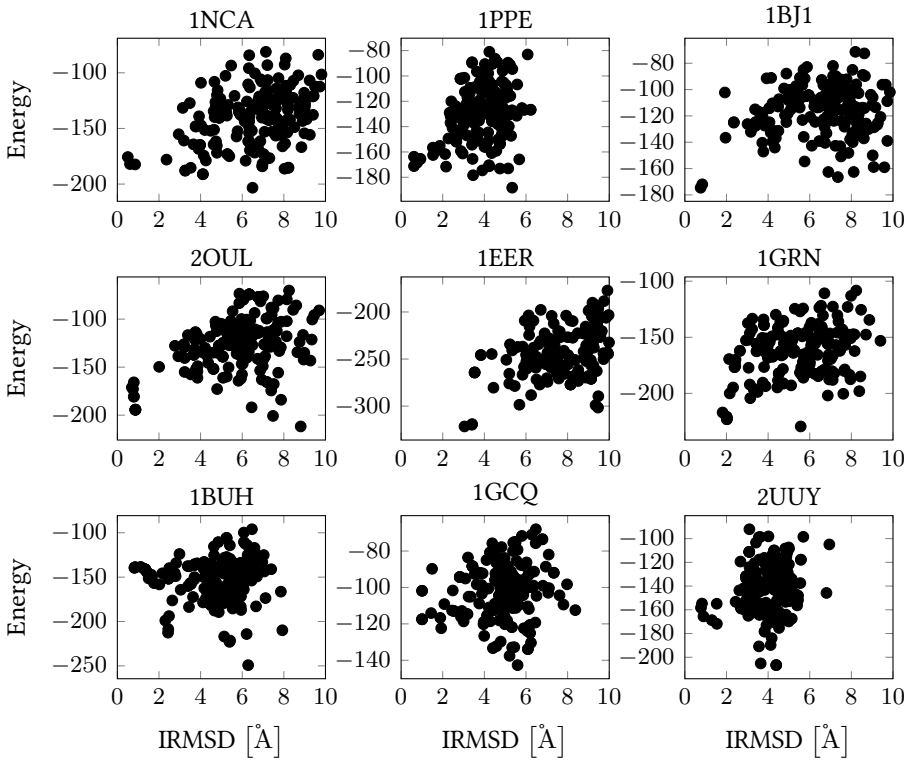


Figure A.5. Energy score vs IRMSD for some i-ATTRACT refinement examples. The score was calculated from the intermolecular energy based on the OPLS parameters.

B. pepATTRACT Supplemental Data

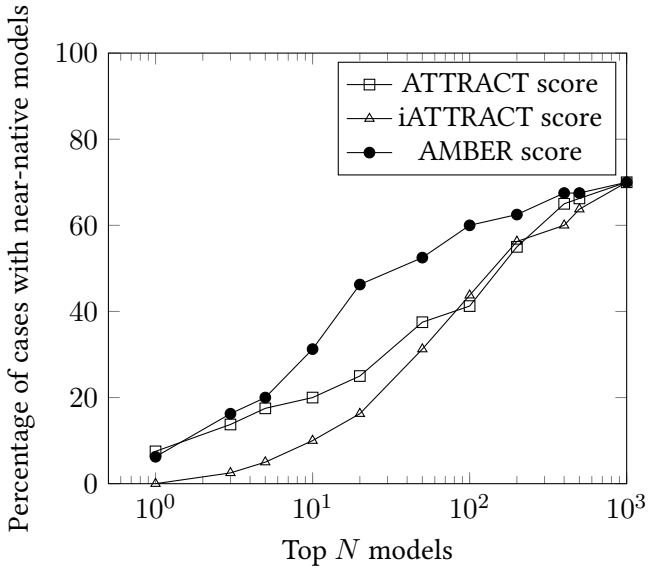


Figure B.1. Scoring performance of AMBER score for unbound-unbound docking. The final AMBER-refined docking models were ranked by their ATTRACT score (scoring before refinement), the OPLS-based scores (scoring after iATTRACT refinement) and the AMBER scores (scoring after final MD refinement). A docking case was considered successful if one of the top N solutions was of near-native or better quality.

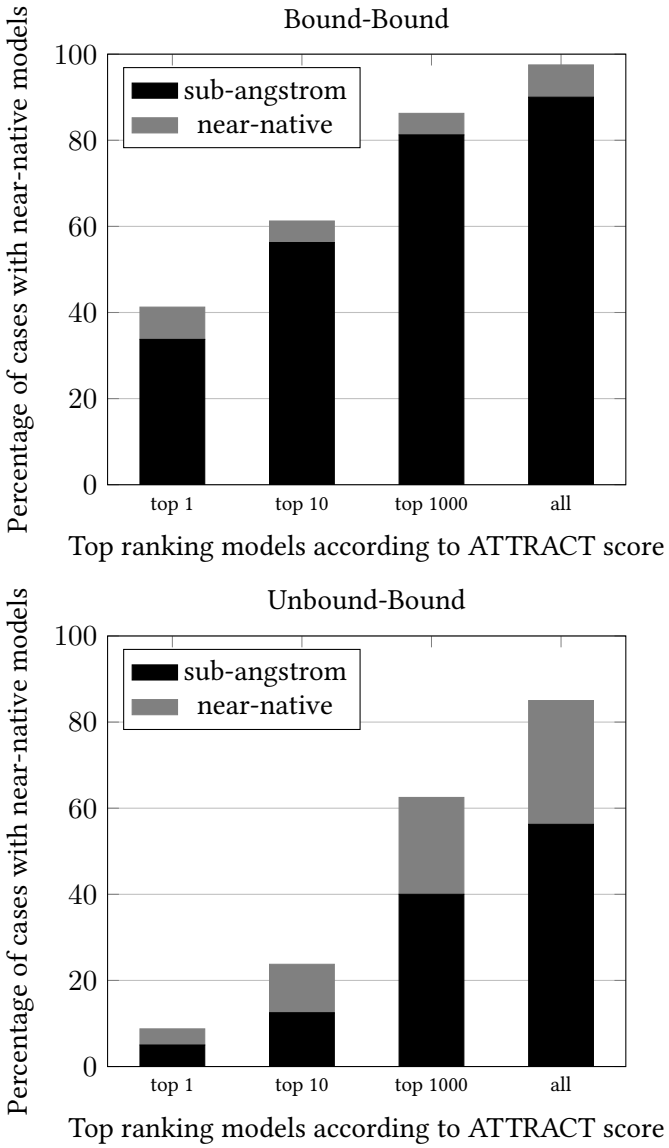
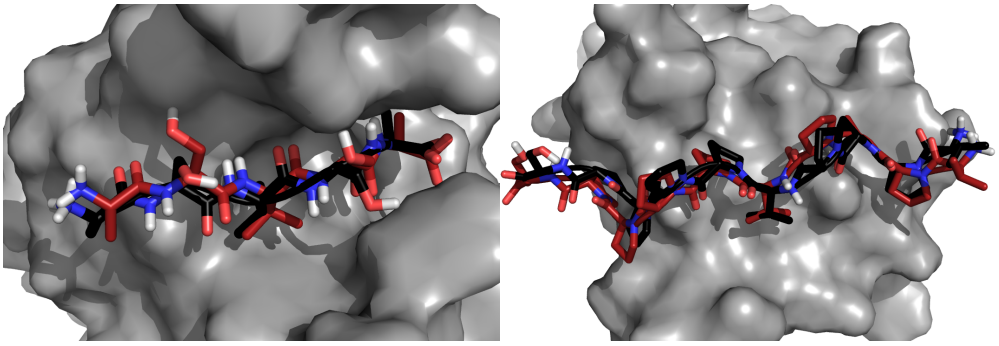


Figure B.2. Percentage of acceptable docking cases for bound-bound and unbound-bound rigid body protein-peptide docking as a function of the number of models considered. A docking case was considered as a near-native/sub-angstrom hit if any of the top N models was of near-native/sub-angstrom quality.



(a) 1NVR–0.7 Å–easy–3rd of 1st cluster **(b)** 2O9V–0.6 Å–easy–4th of 1st cluster

Figure B.3. Example of sub-angstrom models created by the protein-peptide docking protocol for unbound-unbound docking. The backbone and the side chain positions are predicted to high accuracy. The peptide from the crystal structure is shown in black.

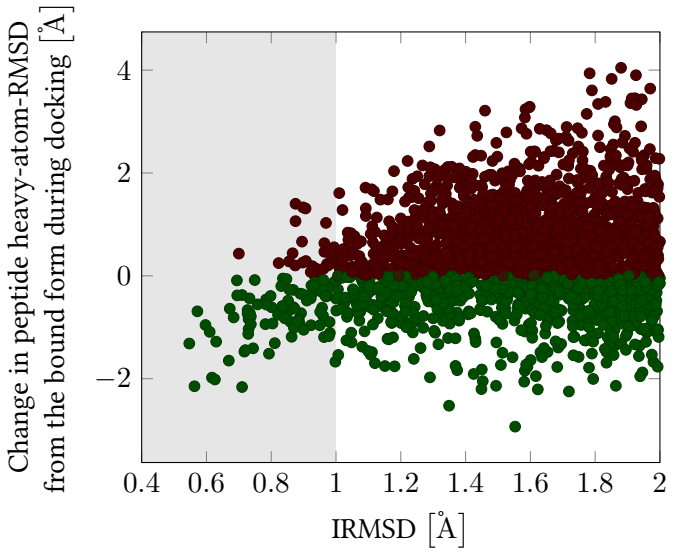
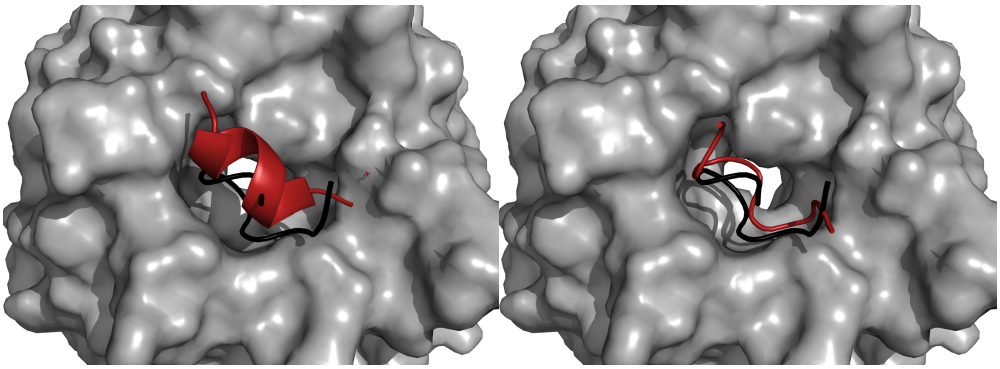
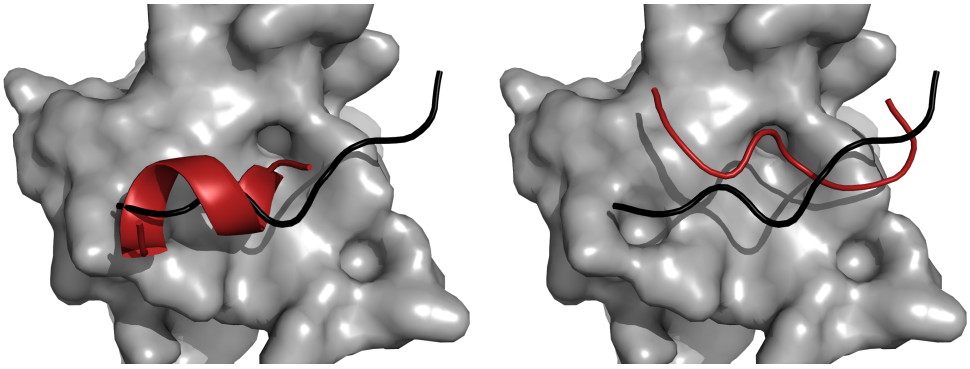


Figure B.4. Change in peptide RMSD from the bound form for the final docking model relative to the initial idealized conformation as a function of the IRMSD of the final docking model. The RMSD was calculated on all heavy atoms. Only near-native models were evaluated. Green markers denote structures in which the peptide conformation moved closer to the bound form. Models of sub-angstrom quality are highlighted by a gray background.



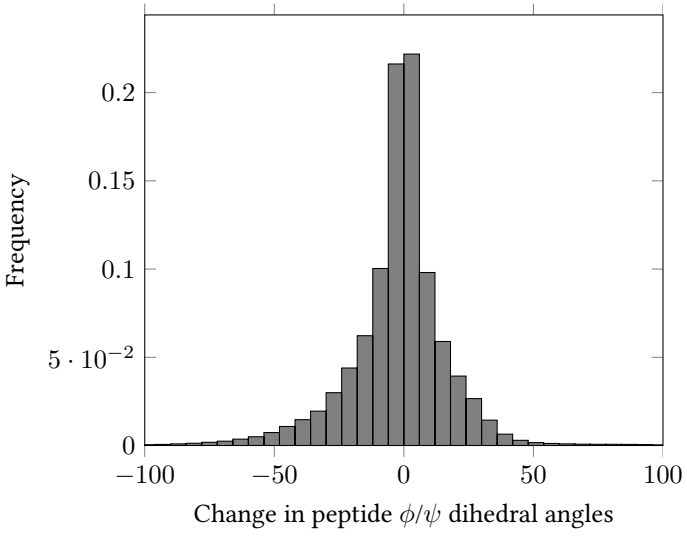
(a) 1X2R – after iATTRACT refinement

(b) 1X2R – after AMBER refinement



(c) 2J6F – after iATTRACT refinement

(d) 2J6F – after AMBER refinement



(e) Peptide main chain flexibility during iATTRACT refinement

Figure B.5. Effect of the refinement stages during docking. (A)–(D) Example of docking cases in which near-native models were generated using an incorrect initial peptide model. The peptide from the crystal structure is shown in black. (E) Change in peptide main chain dihedral angle ϕ and ψ during iATTRACT simulations. The data were evaluated for all refined structures from 80 peptide-protein complexes.

Table B.1. Performance of unbound-unbound protein-peptide docking. Best IRMSD denotes the best IRMSD sampled among all 1,000 final models. Top 10 IRMSD denotes the best IRMSD among the top 10 ranked clusters. The last column lists the rank of the first near-native cluster. A cluster is considered near-native if any of its top 4 ranking members has an IRMSD < 2 Å. The label *Single* denotes that the near-native models were not assigned to a cluster (a minimum cluster size of 4).

PDB	Best IRMSD	Top 10 IRMSD	Rank of near-native cluster
1AWR	0.72	1.06	1
1CE1	1.77	2.51	Single
1CKA	1.91	1.99	5
1CZY	0.57	0.93	7
1D4T	2.16	2.82	-
1DDV	2.02	2.55	-
1DKX	1.65	1.65	2
1EG4	2.75	3.35	-
1ELW	1.11	1.11	2
1ER8	0.82	0.82	1
1GYB	2.61	3.33	-
1HC9	4.16	4.62	-
1IHJ	1.55	1.55	2
1JBU	2.88	3.59	-
1JD5	3.83	3.83	-
1JWG	0.88	0.88	2
1KL3	2.7	2.88	-
1KLU	1.95	2.96	Single
1LVM	1.88	2.43	Single
1MFG	1.37	1.65	1
1N7F	2.26	2.26	-
1NQ7	1.58	5.73	Single
1NTV	1.74	3.33	20
1NVR	0.68	0.68	1
1NX1	1.43	1.76	5
1OAI	1.88	2.08	Single
1OU8	1.17	1.18	1
1PZ5	3.73	4.59	-
1QKZ	2.3	4	-
1RXZ	1.58	1.74	10
1SE0	1.37	1.76	8
1SFI	2.78	3.01	-
1SSH	1.16	2.2	18
1T4F	2.18	4.31	-
1T7R	1.64	3.34	Single
1TP5	0.98	0.98	1
1TW6	1.64	2.13	Single
1U00	2.78	3.16	-
1UJ0	1.51	2.38	31
1VZQ	1.14	1.27	7
1W9E	0.7	1.62	1
1X2R	1.12	1.51	1
1YMT	2.68	8.1	-
1YUC	5.09	10.39	-
1YWO	1.04	1.04	2
1Z9O	0.95	1.08	1
2A3I	1.51	1.51	1
2AK5	0.86	2.58	22
2B1Z	4.1	7.97	-
2B9H	1.58	1.58	1
2C3I	0.86	1.05	4
2CCH	1.6	1.6	3
2D0N	1.89	2.36	Single
2DS8	1.65	3.56	32
2FGR	4.6	7.13	-
2FMF	0.72	1.64	4
2FNT	1.15	1.18	2
2FOJ	0.83	0.96	1
2FVJ	0.69	0.69	7
2H9M	0.8	0.87	1
2HO2	1.18	1.3	2
2HPL	0.55	0.56	1
2IPU	3.27	4.81	-
2J6F	1.61	1.78	2
2JAM	2.7	3.88	-
2O02	1.94	2.21	Single
2O4J	1.37	1.37	1
2O9V	0.6	0.6	1
2P0W	2.52	2.96	-
2P1T	8.56	9.97	-
2P54	1.64	5.68	Single
2PUY	1.74	1.74	5
2QOS	2.37	2.84	-
2R7G	1.89	4.31	Single
2VJ0	1.16	2.21	15
2ZJD	1.61	1.65	7
3BU3	1.73	4.16	Single
3CVP	1.95	1.99	5
3D1E	1.02	1.02	4
3D9T	4.08	5.36	-

Table B.2. Results for Rosetta FlexPepDock high-resolution refinement of pepATTRACT models. Refinement was performed on 1000 models for 14 cases from the docking benchmark. We ranked the pepATTRACT and FlexPepDock refined models by their AMBER and Rosetta score respectively and compared the best fnat within the top 1, top 10 and all sampled models. No clustering procedure was applied. A contact was defined if any two heavy atoms were found within 4 Å distance.

	pepATTRACT			FlexPepDock refined		
PDB	top 1 fnat	top 10 fnat	best fnat	top 1 fnat	top 10 fnat	best fnat
1AWR	0.4	0.68	0.76	0.92	0.92	0.92
1N7F	0.04	0.15	0.58	0.04	0.23	0.58
1NVR	0.77	0.92	0.92	0.0	0.54	0.92
1RXZ	0.38	0.38	0.47	0.03	0.22	0.69
1SSH	0.0	0.60	0.65	0.20	0.70	0.80
1T7R	0.27	0.27	0.32	0.14	0.41	0.77
1W9E	0.11	0.63	0.63	0.47	0.68	0.68
2A3I	0.0	0.53	0.58	0.0	0.47	0.68
2C3I	0.19	0.19	0.67	0.0	0.59	0.78
2FGR	0.0	0.0	0.15	0.0	0.0	0.26
2FMF	0.07	0.13	0.53	0.33	0.80	0.80
2O9V	0.82	0.82	0.82	0.88	1.0	1.0
2P54	0.0	0.0	0.43	0.0	0.05	0.43
2VJ0	0.0	0.05	0.48	0.57	0.57	0.62

C. loopATTRACT Supplemental Data

Table C.1. Percentage of structures that improved more than 0.1 Å in IRMSD or by more than 0.02 in fnat during loopATTRACT loop rebuilding. The percentages were evaluated among the top-ranked 10 models and all generated 1000 models.

PDB	improved IRMSD [%]		improved fnat [%]	
	top 10	all	top 10	all
1ATN	100.0	98.7	90.0	47.1
1BKD	90.0	76.2	10.0	1.6
1FQ1	100.0	97.5	90.0	75.3
1LFD	70.0	48.9	60.0	35.9
1PXV	40.0	53.1	30.0	45.3
1R8S	100.0	100.0	100.0	95.1
2NZ8	0.0	1.5	0.0	6.6
2OT3	0.0	0.0	0.0	0.0
3CPH	0.0	0.0	0.0	0.0
3FN1	100.0	97.8	100.0	68.6

Table C.2. Percentage of structures that improved more than 0.1 Å in IRMSD or by more than 0.02 in fnat during Rosetta loopmodeling. The percentages were evaluated among the top-ranked 10 models and all generated 2500 models.

PDB	improved IRMSD [%]		improved fnat [%]	
	top 10	all	top 10	all
1ATN	100.0	95.2	100.0	96.9
1BKD	90.0	84.3	10.0	4.64
1FQ1	90.0	65.2	40.0	71.6
1LFD	100.0	95.6	20.0	16.7
1PXV	30.0	33.6	0.0	0.08
1R8S	100.0	99.5	60.0	75.5
2NZ8	80.0	77.6	50.0	49.0
2OT3	0.0	0.0	0.0	0.0
3CPH	0.0	0.3	0.0	0.0
3FN1	100.0	99.6	100.0	67.0

Table C.3. Results for loopATTRACT interface loop modeling on top-ranked ATTRACT-EM model obtained with a 20 Å resolution cryo-EM density map [96].

PDB	Difficulty	initial		top 1		top 10		all	
		IRMSD	fnat	IRMSD	fnat	IRMSD	fnat	IRMSD	fnat
1ATN	hard	4.38	0.32	2.91	0.38	2.74	0.46	2.42	0.51
1BKD	hard	3.36	0.52	2.46	0.53	2.10	0.57	1.88	0.58
1FQ1	hard	3.75	0.40	3.37	0.40	2.17	0.47	1.91	0.64
1PXV	hard	2.90	0.63	2.99	0.64	2.42	0.71	2.22	0.73
2OT3	hard	3.61	0.34	3.66	0.27	3.58	0.32	3.53	0.39

D. ATTRACT-SAXS Supplemental Data

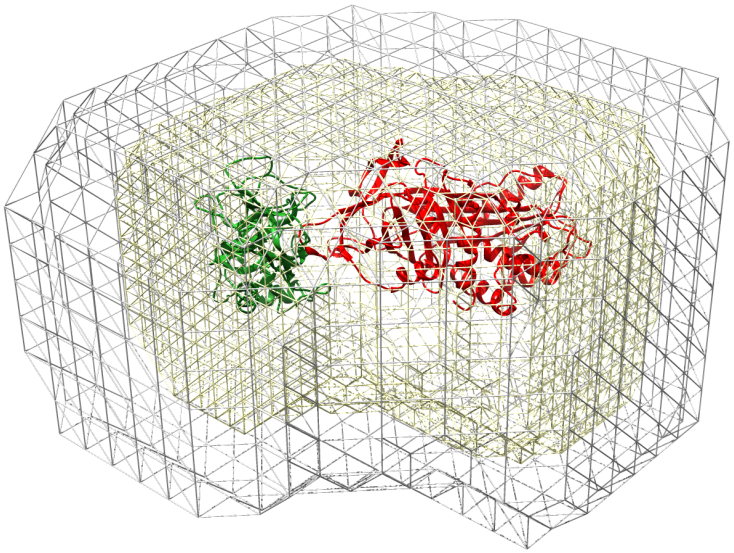


Figure D.1. Example of atom density masks with voxel sizes of 10 and 5 Å for a complex between trypsin and an inhibitor (PDB: 1OPH)[105]. The complex structure was fitted manually in the atom density mask. The figure was generated with Chimera [500].

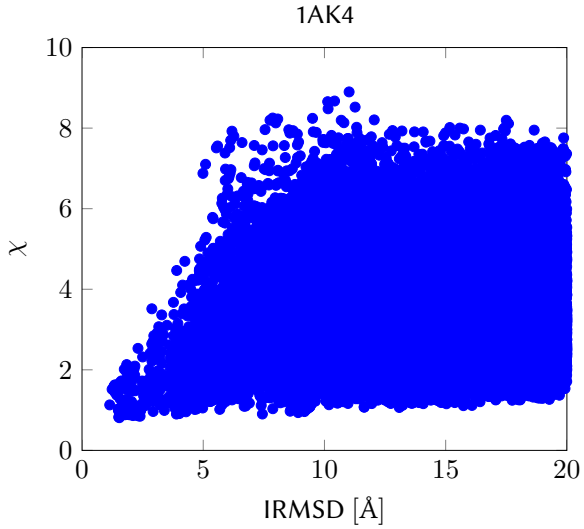


Figure D.2. Example of SAXS scores for docking decoys generated with ATTRACT standard rigid body docking for cyclophilin A bound to the N-terminal domain of HIV-1 capsid (PDB 1AK4) [145].

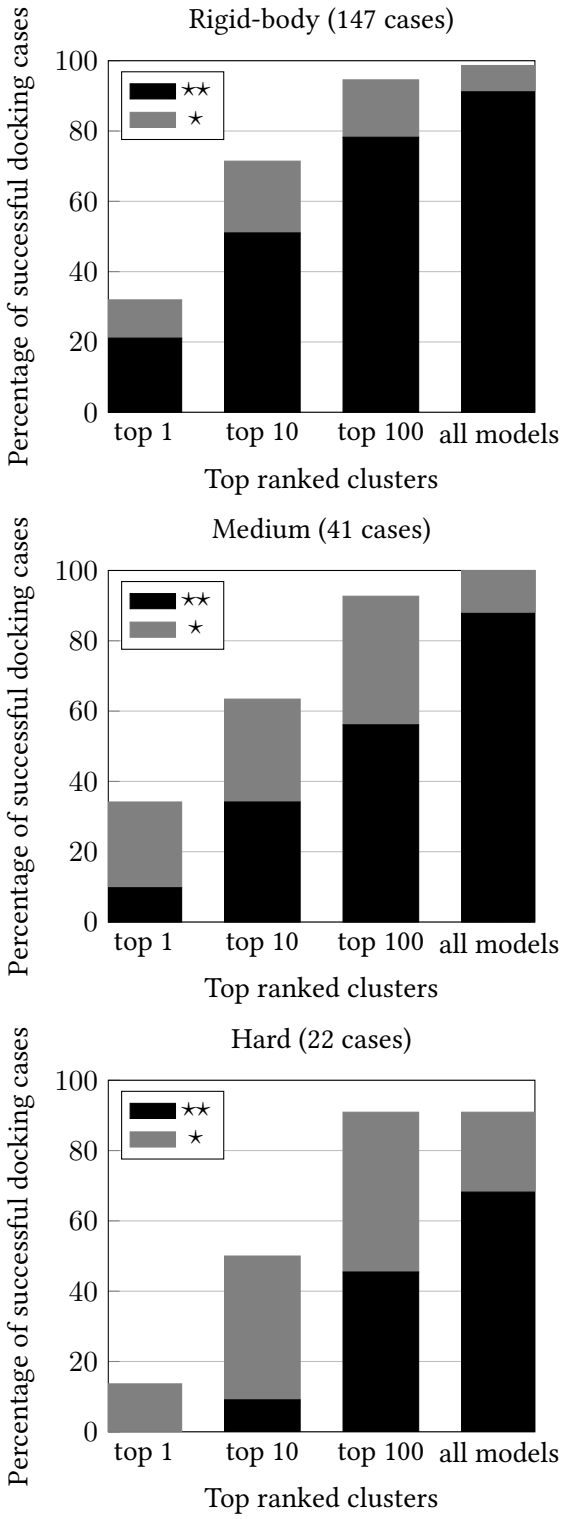


Figure D.3. Docking success rate for ATTRACT-SAXS on 210 complexes from protein-protein docking benchmark 5.0 using simulated SAXS data evaluated by complex type and docking difficulty. A cluster is considered a CAPRI one-star/two star hit if any of its top-ranked 4 members is at least of one star/two star quality. The "all models" success rate is the success rate considering all models generated during the sampling stage.

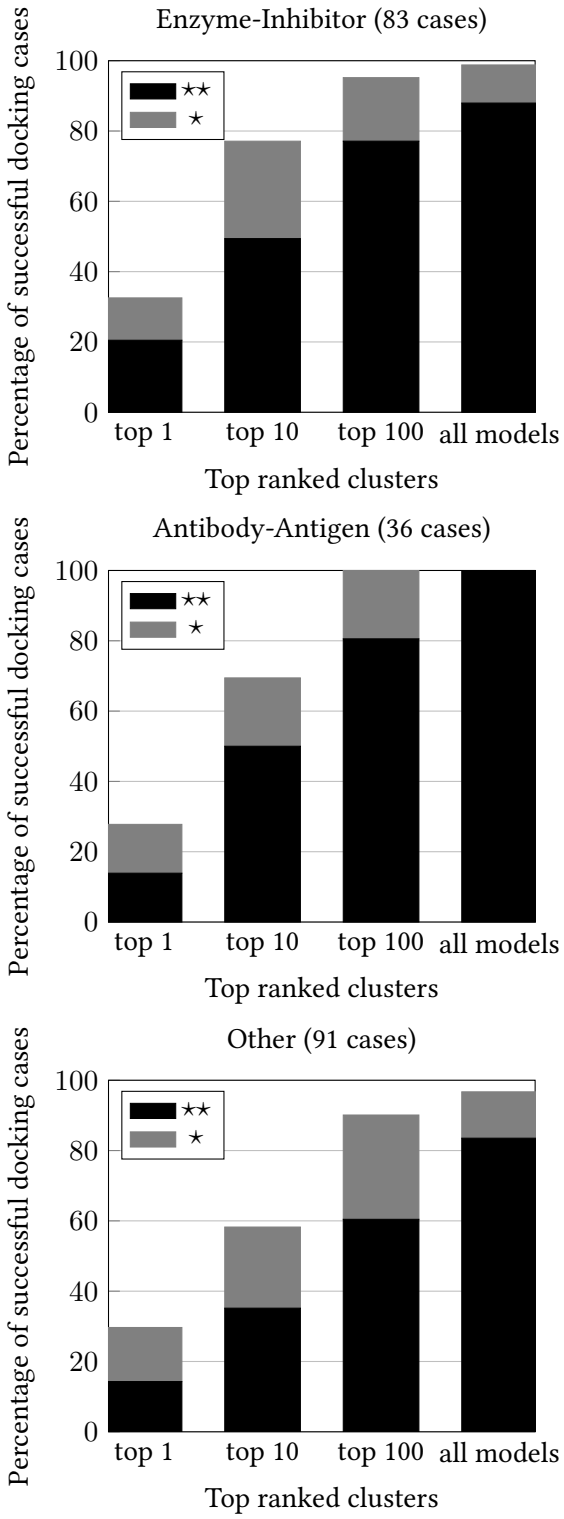


Figure D.3. Continued from previous page.

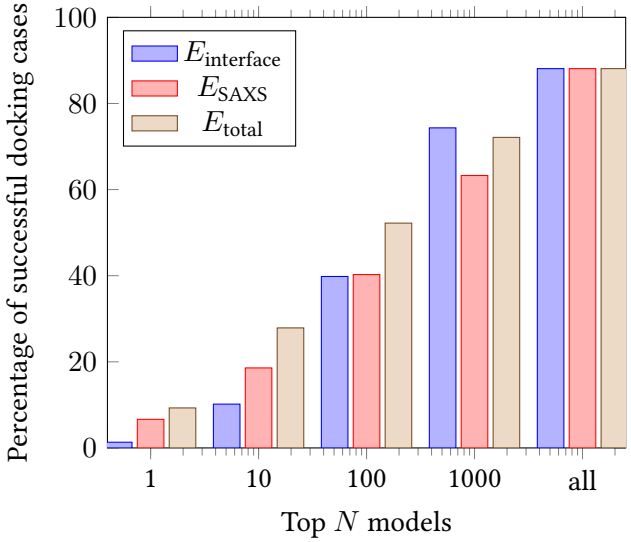


Figure D.4. Ranking of two-star models for different terms in the scoring function. The composite scoring function E_{total} is given by $E_{\text{total}} = E_{\text{interface}} + 300.0 \times E_{\text{SAXS}}$.

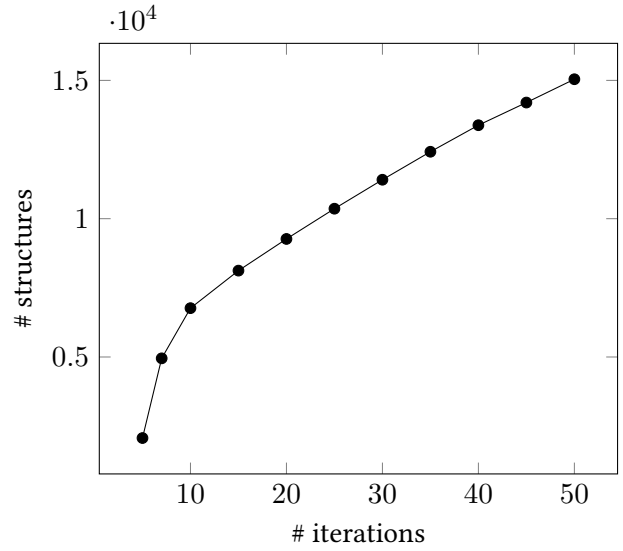


Figure D.5. Average number of structures in taboo pool vs. number of iterations.

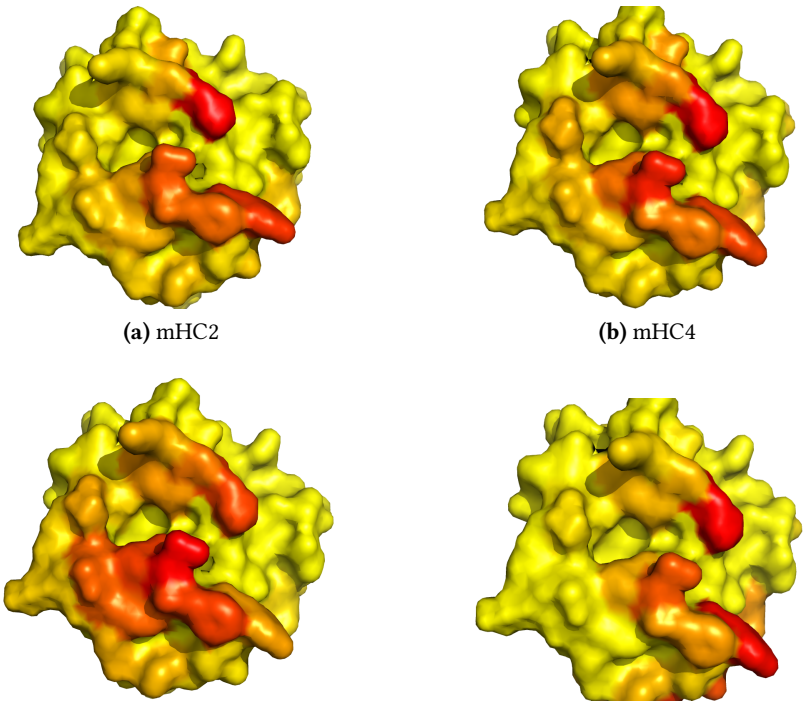
Table D.1. ATTRACT-SAXS docking results for 226 protein-protein complexes from docking benchmark 5. The best IRMSD, LRMSD, fnat and CAPRI quality for the top 1, top 10 and top 100 ranked clusters and all sampled models are listed. A cluster is considered a CAPRI one-star/two star hit if any of its top-ranked 4 members is at least of one star/two star quality. For docking case 1BGX, no models were collected during the sampling stage.

	Top 1				Top 10				Top 100				All models			
PDB	IRMSD	LRMSD	fnat	CAPRI	IRMSD	LRMSD	fnat	CAPRI	IRMSD	LRMSD	fnat	CAPRI	IRMSD	LRMSD	fnat	CAPRI
1A2K	15.3	28.0	0.00	-	6.1	19.0	0.39	-	3.1	11.3	0.63	*	3.1	11.3	0.63	*
1ACB	5.0	9.6	0.05	-	2.8	6.1	0.44	*	2.8	4.1	0.32	**	2.8	4.1	0.32	**
1AHW	4.3	12.2	0.41	-	2.0	4.5	0.73	**	2.0	4.5	0.73	**	2.0	4.5	0.73	**
1AK4	4.1	17.5	0.42	-	1.7	6.6	0.60	**	1.8	6.7	0.49	**	1.7	6.6	0.60	**
1AKJ	1.4	4.3	0.59	**	2.0	3.5	0.43	**	1.4	3.3	0.63	**	1.4	3.3	0.63	**
1ATN	9.6	32.7	0.13	-	5.0	28.4	0.22	-	5.0	28.4	0.22	-	4.2	31.5	0.21	-
1AVX	17.3	52.5	0.00	-	4.0	13.7	0.52	*	1.6	6.2	0.59	**	1.2	2.7	0.41	**
1AY7	15.1	32.8	0.00	-	10.0	16.8	0.00	-	1.6	3.2	0.62	**	1.1	3.6	0.57	**
1AZS	18.2	85.9	0.00	-	10.7	36.3	0.00	-	7.0	18.1	0.05	-	1.9	3.8	0.27	*
1B6C	2.1	3.5	0.77	**	2.1	3.5	0.77	**	2.1	3.5	0.77	**	2.1	3.5	0.77	**
1BGX	NaN	NaN	NaN	-	NaN	NaN	NaN	-	NaN	NaN	NaN	-	NaN	NaN	NaN	-
1BJ1	9.0	24.7	0.14	-	1.9	12.4	0.48	**	1.9	7.8	0.48	**	1.9	7.8	0.48	**
1BKD	23.9	55.7	0.00	-	5.3	9.8	0.35	*	4.4	6.4	0.24	*	3.0	3.5	0.55	**
1BUH	13.1	22.0	0.17	-	1.5	3.3	0.79	**	1.5	3.3	0.79	**	1.5	3.3	0.79	**
1BVK	13.8	25.1	0.00	-	13.2	25.1	0.00	-	2.3	6.1	0.38	*	1.9	5.6	0.45	**
1BVN	1.4	3.1	0.74	**	1.4	3.1	0.74	**	1.4	3.1	0.74	**	1.3	2.0	0.45	**
1CGI	7.0	13.7	0.02	-	2.4	3.4	0.53	**	2.4	3.4	0.53	**	2.3	3.7	0.53	**
1CLV	6.5	13.3	0.04	-	1.9	3.7	0.31	**	1.9	3.7	0.31	**	1.5	2.8	0.59	**
1D6R	16.0	54.3	0.00	-	3.7	9.1	0.16	*	1.5	4.6	0.64	**	1.5	4.6	0.64	**
1DE4	10.4	23.1	0.12	-	10.4	23.1	0.12	-	3.0	10.4	0.77	*	3.8	6.4	0.30	*
1DFJ	10.6	21.2	0.00	-	1.5	2.4	0.74	**	1.5	2.4	0.74	**	1.5	2.4	0.74	**
1DQJ	3.7	6.7	0.33	*	2.5	5.6	0.54	*	1.8	3.4	0.79	**	1.8	3.4	0.79	**
1E4K	14.0	32.4	0.00	-	13.7	29.8	0.00	-	11.6	35.0	0.00	-	2.2	8.0	0.37	*
1E6E	11.7	19.5	0.04	-	1.8	2.5	0.67	**	1.8	2.5	0.67	**	1.8	2.5	0.67	**
1E6J	9.1	17.2	0.14	-	2.3	5.5	0.53	*	2.1	5.8	0.37	*	1.8	5.7	0.55	**
1E96	8.1	31.5	0.03	-	5.3	13.1	0.28	-	3.9	8.6	0.40	*	1.8	4.4	0.45	**
1EAW	5.7	11.7	0.23	-	1.5	3.8	0.74	**	1.5	3.8	0.74	**	1.4	3.0	0.57	**
1EER	17.3	44.7	0.02	-	8.1	16.5	0.06	-	3.2	6.2	0.53	*	2.8	4.9	0.43	**
1EFN	9.6	16.1	0.20	-	2.9	5.5	0.54	*	1.6	5.6	0.71	**	1.3	3.9	0.49	**
1EWY	2.2	4.0	0.42	**	2.2	4.0	0.42	**	2.2	4.0	0.42	**	1.4	1.8	0.51	**
1EXB	21.0	31.8	0.00	-	5.6	6.8	0.61	*	5.8	8.0	0.54	*	5.2	5.0	0.73	**
1EZU	17.8	31.0	0.07	-	9.7	19.7	0.12	-	3.6	8.3	0.34	*	3.5	11.7	0.39	*
1F34	16.4	46.4	0.00	-	2.1	4.5	0.62	**	1.7	4.0	0.60	**	1.7	4.0	0.60	**
1F51	10.3	29.5	0.00	-	10.3	29.5	0.00	-	1.6	3.2	0.71	**	1.6	3.2	0.71	**
1F6M	18.9	55.7	0.00	-	13.2	33.1	0.16	-	7.0	8.3	0.09	-	5.7	3.1	0.14	*
1FAK	22.7	39.9	0.00	-	13.3	28.4	0.09	-	7.7	23.0	0.07	-	6.7	16.9	0.24	*
1FC2	2.2	4.3	0.63	**	2.2	4.3	0.63	**	2.2	4.3	0.63	**	1.9	3.1	0.58	**
1FCC	11.3	20.7	0.05	-	7.5	16.6	0.23	-	1.7	3.1	0.72	**	1.7	3.1	0.72	**
1FFW	12.2	44.1	0.00	-	10.0	35.5	0.00	-	8.9	28.2	0.06	-	3.8	12.4	0.17	*
1FLE	10.3	32.2	0.03	-	10.3	32.2	0.03	-	4.0	11.5	0.35	*	1.9	4.9	0.58	**
1FQ1	17.6	47.2	0.00	-	4.9	7.7	0.39	*	2.9	4.8	0.55	**	2.8	4.3	0.43	**
1FQJ	14.9	28.4	0.07	-	1.3	3.0	0.55	**	1.3	3.0	0.55	**	1.2	2.1	0.73	**
1FSK	13.7	27.4	0.08	-	1.5	3.0	0.79	**	1.5	3.0	0.79	**	1.1	3.1	0.64	**
1GCQ	10.5	44.5	0.00	-	10.2	44.7	0.00	-	4.0	8.3	0.28	*	1.4	2.8	0.56	**
1GHQ	23.6	64.3	0.00	-	12.6	58.4	0.00	-	2.5	8.9	0.56	*	1.4	3.3	0.80	**
1GL1	2.0	3.7	0.55	**	2.0	3.7	0.55	**	2.0	3.7	0.55	**	1.5	2.7	0.47	**
1GLA	2.6	5.1	0.51	*	1.4	3.7	0.51	**	1.4	3.7	0.51	**	1.3	2.9	0.44	**
1GP2	12.3	48.5	0.00	-	2.5	2.9	0.18	*	2.5	2.9	0.18	*	2.5	4.8	0.17	*
1GPW	18.5	59.7	0.00	-	2.0	5.5	0.51	*	1.3	2.4	0.73	**	1.5	2.2	0.61	**
1GRN	2.7	7.1	0.47	*	2.7	7.1	0.47	*	1.9	4.1	0.59	**	2.0	3.7	0.56	**
1GXD	6.0	20.7	0.14	-	3.7	15.9	0.32	*	3.2	19.8	0.38	*	3.8	8.1	0.24	*
1H1V	20.4	66.7	0.00	-	13.8	29.6	0.00	-	11.1	30.8	0.02	-	8.6	20.4	0.00	-
1H9D	9.7	15.8	0.00	-	3.4	6.2	0.32	*	2.0	4.4	0.59	**	1.7	3.4	0.58	**
1HCF	2.0	6.9	0.53	*	2.0	6.9	0.53	*	2.0	3.6	0.66	**	1.7	3.0	0.34	**
1HE1	1.5	2.3	0.81	**	1.5	2.3	0.81	**	1.5	2.3	0.81	**	1.5	2.3	0.81	**
1HE8	14.8	25.0	0.00	-	6.1	11.9	0.37	-	4.0	7.3	0.43	*	1.7	2.5	0.50	**
1HIA	3.1	8.0	0.51	*	2.5	6.2	0.49	*	1.8	4.7	0.60	**	1.7	2.9	0.52	**
1I2M	3.6	8.5	0.24	*	2.2	4.0	0.55	**	2.2	4.0	0.55	**	2.0	2.9	0.64	**
1I4D	10.4	41.2	0.00	-	5.0	11.6	0.13	-	1.5	2.1	0.58	**	1.5	2.1	0.58	**
1I9R	13.8	34.4	0.00	-	6.8	26.1	0.00	-	1.8	3.8	0.83	**	1.8	3.8	0.83	**
1IB1	18.8	61.6	0.00	-	18.0	61.9	0.00	-	17.2	60.7	0.00	-	2.4	3.2	0.59	**
1IBR	17.2	31.8	0.02	-	9.1	21.1	0.06	-	2.8	4.5	0.54	**	3.1	4.5	0.44	**
1IJK	17.3	33.8	0.00	-	2.0	5.9	0.52	**	2.0	5.9	0.52	**	1.1	2.8	0.52	**
1IQD	9.1	19.2	0.08	-	1.8	6.1	0.65	**	1.8	3.9	0.56	**	1.3	3.9	0.65	**
1IRA	27.3	52.9	0.00	-	23.7	55.5	0.00	-	22.5	54.1	0.00	-	16.1	27.2	0.00	-
1J2J	3.8	9.6	0.33	*	2.7	6.8	0.27	*	1.8	4.1	0.52	**	1.3	2.9	0.61	**
1JIW	10.7	15.4	0.00	-	10.4	20.5	0.14	-	8.4	24.3	0.02	-	3.3	5.2	0.60	*
1JK9	11.4	26.7	0.12	-	3.3	4.6	0.59	**	3.3	4.6	0.59	**	3.4	4.6	0.60	**
1JMO	3.7	10.0	0.19	*	3.7	10.0	0.19	*	4.8	10.0	0.15	*	5.0	8.7	0.17	*
1JPS	2.0	4.7	0.42	**	1.4	4.1	0.62	**	1.3	3.4	0.69	**	1.0	2.6	0.68	***
1JTD	12.4	40.7	0.03	-	11.9	42.9	0.00	-	1.2	3.1	0.69	**	0.6	1.3	0.77	***
1JTG	12.8	34.3	0.00	-	1.9	4.4	0.61	**	1.9	4.4	0.61	**	1.9	4.4	0.52	**
1JWH	6.8	34.4	0.13	-	5.3	20.2	0.62	-	4.0	17.2	0.51	-	4.0	16.5	0.24	*
1JZD	8.9	17.1	0.08	-	2.9	6.8	0.25	*	3.4	3.9	0.46	**	3.0	3.5	0.48	**
1K4C	3.6	15.9	0.29	*	1.8	4.3	0.74	**	1.8	4.3	0.74	**	1.9	4.7	0.31	**
1K5D	3.3	9.4	0.25	*	2.6	3.8	0.37	**	2.6	3.8	0.37	**	2.0	8.4	0.60	**
1K74	1.1	1.9	0.76	**	1.1	1.9	0.76	**	1.1	1.9	0.76	**	1.1	1.9	0.76	**
1KAC	5.7	11.6	0.20	-	2.8	5.0	0.34	**	1.6	4.1	0.64	**	1.6	4.1	0.64	**
1KKL	3.9	10.7	0.12	*	3.6	11.5	0.29	*	3.4	5.7	0.10	*	1.2	3.0	0.31	**
1KLU	15.0	29.9	0.00	-	2.5	12.5	0.33	*	2.0	5.3	0.37	**	2.0	6.3	0.37	**
1KTZ	1.1	5.4	0.86	**	1.1	5.4	0.86	**	1.0	6.0	0.83	***	1.0	6.0	0.62	***
1KXP	1.5	3.4	0.63	**	1.5	3.4	0.63	**	1.5	3.4	0.63	**	1.7	2.7	0.58	**
1KXQ	1.5	3.1	0.93	**	1.5	3.1	0.93	**	1.5	3.1	0.93	**	1.7	3.1	0.75	**
1LFD	9.4	18.1	0.23	-	4.4	10.6	0.30	-	3.1	6.0	0.42	*	2.5	4.6	0.40	**
1M10	14.8	23.7	0.10	-	10.3	21.1	0.07	-	6.7	14.9	0.07	-	3.3	4.9	0.36	**
1M27	17.6	54.6	0.00	-	3.1	9.8	0.42	*	2.5	4.1	0.29	*	2.6	5.9	0.67	*
1MAH	1.4	2.8	0.61	**	1.4	2.8	0.61	**	1.4	2.8	0.61	**	1.4	2.8	0.61	**
1ML0	12.3	21.6	0.05	-	3.5	6.9	0.27	*	1.4	2.6	0.47	**	1.6	2.8	0.48	**
1MLC	8.8	18.7	0.00	-	8.3	20.4	0.00	-	1.1	3.8	0.84	**	1.1	3.8	0.84	**
1MQ8	11.4	57.0	0.00	-	8.6	16.4	0.14	-	3.9	8.2	0.12	*	3.1	6.2	0.44	*
1N2C	43.2	100.0	0.00	-	4.3	11.3	0.37	-	3.1	8.9	0.53	*	2.6	6.8	0.62	*
1NCA	16.9	31.0	0.01	-	1.8	7.2	0.62	**	1.8	7.2	0.62	**	0.9	3.2	0.92	***
1NSN	14.5	24.0	0.02	-	1.8	3.6	0.75	**	1.8	3.6	0.75	**	1.2	3.9	0.54	**
1NW9	15.8	24.4	0.02	-	9.9	21.7	0.05	-	8.4	4.7	0.28	*	8.3	9.3	0.12	*
1OC0	5.1	8.9	0.07	-	2.1	4.9	0.37	**	2.1	4.9	0.37	**	2.0	4.3	0.22	**
1OFU	2.2	5.2	0.50	*	1.6	4.0	0.58	**	1.0	4.4	0.57	***	1.0	4.4	0.51	***
1OPH	11.6	26.6	0.23	-	3.8	5.3	0.32	*	2.3	5.1	0.74	*	2.3	5.1	0.74	*
1OYV	1.9	4.3	0.62	**	1.9	4.3	0.62	**	1.9	4.3	0.62	**	1.2	2.4	0.54	**
1PPE	2.3	6.7	0.54	*	1.5	3.9	0.61	**	1.1	2.7	0.77	**	0.7	1.4	0.73	***
1PVH	8.8	17.3	0.00	-	7.0	13.3	0.03	-	3.8	7.8	0.27	*	1.5	6.6	0.90	**
1PXV	18.1	37.4	0.00	-	6.6	18.8	0.33	-	3.2	4.1	0.64	**	3.1	4.6	0.65	**
1QA9	7.8	17.1	0.11	-	5.1	12.5	0.21	-	1.5	3.9	0.66	**	1.5	3.9	0.66	**
1QFW	20.9	48.9	0.00	-	12.3	42.2	0.08	-	1.9	4.5	0.79	**	1.9	4.5	0.79	**
1R0R	15.8	43.1	0.00	-	3.5	10.3	0.48	*	1.4	3.2	0.88	**	1.4	3.1	0.84	**
1R6Q	11.8	20.5	0.00	-	2.8	8.2	0.35	*	2.6	4.0	0.37	**	2.8	3.9	0.33	**
1R8S	6.0	10.7	0.23	-	3.7	5.8	0.45	*	3.6	6.8	0.39	*	3.6	6.8	0.39	*
1RKE	16.7	47.2	0.00	-	10.7	46.7	0.00	-	3.5	4.3	0.59	**	3.6	4.3	0.59	**
1RLB	8.5	25.0	0.18	-	4.9	16.2	0.47	-	2.9	7.5	0.47	*	1.3	5.9	0.54	**
1RV6	10.6	20.9	0.21	-	4.3	8.3	0.28	*	1.4	3.0	0.66	**	1.4	3.0	0.64	**
1S1Q	12.4	31.5	0.00	-	6.7	20.6	0.07	-	3.0	6.4	0.44	*	1.3	3.9	0.77	**

Table D.1. Continued from previous page.

	Top 1				Top 10				Top 100				All models			
PDB	IRMSD	LRMSD	fnat	CAPRI	IRMSD	LRMSD	fnat	CAPRI	IRMSD	LRMSD	fnat	CAPRI	IRMSD	LRMSD	fnat	CAPRI
1SBB	2.2	5.2	0.67	*	1.7	10.0	0.79	**	1.8	11.1	0.64	**	2.0	7.9	0.82	**
1SYX	18.1	49.0	0.00	-	13.5	47.5	0.00	-	3.0	5.4	0.35	*	2.3	3.1	0.53	**
1T6B	14.8	24.0	0.00	-	8.0	17.6	0.11	-	6.2	11.5	0.00	-	5.0	17.6	0.05	-
1TMQ	1.9	5.9	0.48	**	1.8	2.9	0.52	**	1.8	2.9	0.52	**	1.7	3.4	0.47	**
1UDI	14.9	23.1	0.07	-	1.6	4.0	0.55	**	1.6	4.0	0.55	**	1.1	1.7	0.79	**
1US7	14.6	35.7	0.06	-	2.0	3.2	0.86	**	2.0	3.2	0.86	**	1.1	2.2	0.72	**
1VFB	16.2	30.3	0.00	-	4.8	12.3	0.28	-	3.7	17.6	0.40	*	1.7	5.0	0.72	**
1WDW	15.4	43.3	0.06	-	6.8	21.0	0.16	-	1.9	5.2	0.48	**	1.3	2.6	0.55	**
1WEJ	11.6	20.5	0.00	-	1.9	6.6	0.58	**	1.7	4.7	0.51	**	1.0	3.2	0.77	***
1WQ1	13.0	23.0	0.10	-	1.5	2.4	0.47	**	1.5	2.4	0.47	**	1.6	2.9	0.44	**
1XD3	6.8	15.1	0.15	-	2.1	4.1	0.57	**	1.8	7.1	0.47	**	1.9	3.7	0.53	**
1XQS	21.5	48.6	0.00	-	3.4	11.4	0.40	*	2.3	3.4	0.45	**	2.5	3.9	0.43	**
1XU1	2.1	6.2	0.51	*	2.0	5.1	0.57	**	1.8	5.3	0.58	**	1.8	4.1	0.30	**
1Y64	21.1	33.3	0.00	-	14.7	35.1	0.00	-	11.3	33.3	0.00	-	6.0	5.2	0.17	*
1YVB	1.7	4.1	0.64	**	1.7	4.1	0.64	**	1.6	7.0	0.72	**	1.2	3.3	0.54	**
1Z0K	8.5	16.1	0.13	-	1.4	2.7	0.79	**	1.4	2.7	0.79	**	0.8	1.2	0.87	***
1Z5Y	15.7	44.8	0.00	-	1.8	3.6	0.79	**	1.8	3.6	0.79	**	1.8	3.6	0.79	**
1ZHH	10.2	23.7	0.02	-	10.2	23.7	0.02	-	3.2	7.6	0.14	*	2.5	5.8	0.26	*
1ZHI	17.5	30.1	0.00	-	7.3	32.5	0.00	-	2.3	7.1	0.41	*	2.2	4.7	0.41	**
1ZLI	14.6	24.7	0.17	-	2.9	5.2	0.41	*	2.8	4.9	0.47	**	2.8	4.9	0.47	**
1ZM4	3.5	8.4	0.46	*	3.5	8.4	0.46	*	2.8	4.1	0.68	**	2.8	4.1	0.68	**
2A1A	1.6	2.8	0.89	**	1.6	2.8	0.89	**	1.6	2.3	0.86	**	1.5	1.8	0.73	**
2A5T	1.4	1.8	0.69	**	1.4	1.8	0.69	**	1.4	1.8	0.69	**	1.5	1.7	0.69	**
2A9K	4.1	11.6	0.23	-	4.1	11.6	0.23	-	4.1	11.6	0.23	-	3.1	8.9	0.18	*
2ABZ	1.4	3.2	0.78	**	1.4	3.2	0.78	**	1.4	4.1	0.80	**	1.4	3.9	0.63	**
2AJF	1.8	5.0	0.40	**	1.6	5.7	0.71	**	1.6	5.7	0.71	**	1.6	5.7	0.71	**
2AYO	1.6	1.6	0.60	**	1.6	1.6	0.60	**	1.6	1.6	0.60	**	1.6	1.6	0.60	**
2B42	3.4	11.3	0.35	*	3.4	11.3	0.35	*	2.1	5.6	0.63	*	2.1	4.3	0.91	**
2B4J	12.5	33.5	0.00	-	9.6	35.9	0.00	-	2.1	4.9	0.81	**	1.9	6.9	0.40	**
2BTF	2.3	4.2	0.40	**	1.2	1.8	0.57	**	1.2	1.8	0.57	**	1.2	1.7	0.47	**
2C0L	11.6	27.1	0.00	-	9.3	16.8	0.00	-	3.0	6.1	0.51	*	2.6	3.3	0.44	**
2CFH	2.9	4.9	0.50	**	2.9	4.9	0.50	**	2.9	4.9	0.50	**	2.9	4.9	0.50	**
2FD6	17.4	30.6	0.00	-	2.8	12.1	0.55	*	1.8	8.9	0.74	**	1.8	8.9	0.74	**
2FJU	6.4	13.6	0.03	-	1.6	3.3	0.84	**	1.6	3.3	0.84	**	1.6	4.0	0.65	**
2G77	12.5	22.1	0.06	-	6.5	10.8	0.12	-	1.9	4.0	0.69	**	2.1	4.6	0.71	**
2GAF	22.0	63.2	0.00	-	16.5	35.9	0.00	-	2.0	4.6	0.52	**	1.6	2.6	0.56	**
2GTP	0.9	1.7	0.76	***	0.9	1.7	0.76	*	0.9	1.7	0.76	*	0.9	1.7	0.76	***
2H7V	2.3	11.7	0.61	*	2.3	11.7	0.61	*	1.9	3.5	0.52	**	2.0	4.6	0.63	**
2HLE	16.0	55.5	0.00	-	1.8	4.8	0.63	**	1.8	4.8	0.63	**	1.9	4.4	0.56	**
2HMI	24.9	69.2	0.00	-	12.0	26.9	0.00	-	10.0	42.9	0.00	-	6.2	42.1	0.00	-
2HQS	6.3	17.0	0.12	-	2.1	3.5	0.53	**	1.9	3.6	0.40	**	1.9	3.7	0.56	**
2HRK	13.3	36.7	0.00	-	4.0	10.8	0.53	-	2.7	4.8	0.72	**	2.0	3.2	0.88	**
2I25	1.5	2.9	0.52	**	1.5	2.9	0.52	**	1.5	2.9	0.52	**	1.5	2.7	0.63	**
2I9B	10.9	42.2	0.05	-	10.9	42.2	0.05	-	5.1	8.7	0.23	*	4.3	6.9	0.22	*
2IDO	3.6	8.7	0.23	*	3.6	8.7	0.23	*	2.7	4.2	0.35	**	2.7	4.2	0.35	**
2J0T	2.8	5.7	0.38	*	2.8	5.7	0.38	*	1.7	3.6	0.60	**	1.7	4.6	0.55	**
2J7P	19.8	47.5	0.00	-	5.3	11.8	0.19	-	3.3	5.1	0.60	*	3.5	5.1	0.39	*
2JEL	10.6	19.6	0.00	-	1.2	4.9	0.71	**	1.4	3.2	0.70	**	1.8	4.6	0.46	**
2MTA	1.4	4.2	0.73	**	1.4	4.2	0.73	**	1.6	3.4	0.47	**	1.4	2.8	0.49	**
2NZ8	13.8	26.9	0.09	-	5.2	7.4	0.20	*	2.9	4.4	0.37	**	2.6	4.1	0.35	**
2O3B	13.4	23.9	0.00	-	10.8	26.5	0.04	-	4.7	9.5	0.09	-	4.7	9.5	0.09	-
2O8V	16.6	64.6	0.00	-	1.9	6.6	0.55	**	1.7	4.2	0.70	**	1.8	8.4	0.90	**
2OOB	7.6	16.2	0.11	-	3.6	7.3	0.22	*	1.7	2.6	0.67	**	1.7	2.6	0.67	**
2OOR	12.6	21.4	0.16	-	12.6	21.4	0.16	-	2.6	6.8	0.59	*	2.1	5.5	0.46	*
2OT3	12.2	25.0	0.12	-	10.8	20.9	0.16	-	3.3	5.2	0.39	*	2.7	3.0	0.43	**
2OUL	1.3	2.7	0.71	**	1.3	2.7	0.71	**	1.3	2.7	0.71	**	1.2	3.6	0.65	**
2OZA	11.5	22.4	0.11	-	6.7	12.3	0.12	-	3.6	7.2	0.30	*	2.7	4.9	0.53	**
2PCC	20.2	62.4	0.00	-	1.9	3.5	0.76	**	1.9	3.5	0.76	**	1.9	3.5	0.76	**
2SIC	8.2	21.1	0.14	-	1.7	7.1	0.59	**	1.5	3.7	0.66	**	1.9	5.3	0.55	**
2SNI	14.4	44.9	0.00	-	1.0	2.8	0.79	***	1.0	2.8	0.79	***	1.0	2.8	0.79	***
2UUY	7.9	20.6	0.00	-	4.1	16.1	0.21	-	0.8	3.6	0.79	***	0.8	3.6	0.79	***
2VDB	14.9	38.0	0.00	-	12.8	31.8	0.00	-	6.0	13.0	0.13	-	2.0	3.4	0.60	**
2VIS	25.3	47.1	0.00	-	2.6	13.7	0.33	*	3.8	11.1	0.16	*	1.4	20.7	0.55	**
2VXT	8.4	19.8	0.03	-	1.9	4.2	0.63	**	1.9	4.2	0.63	**	1.9	4.2	0.63	**
2W9E	12.1	28.2	0.00	-	8.5	18.8	0.09	-	2.1	4.5	0.75	**	1.9	7.0	0.41	**
2X9A	4.6	13.3	0.33	-	2.1	2.8	0.71	**	2.1	2.8	0.71	**	2.1	2.8	0.60	**
2YVJ	2.6	4.6	0.41	**	2.6	4.6	0.41	**	2.6	4.6	0.41	**	1.4	4.9	0.51	**
2Z0E	13.8	46.4	0.01	-	13.3	47.0	0.03	-	3.8	7.6	0.20	*	2.8	4.9	0.47	**
3A4S	2.0	4.3	0.84	**	2.0	4.3	0.84	**	1.6	3.5	0.75	**	1.3	2.2	0.81	**
3AAA	14.0	25.6	0.02	-	2.4	5.3	0.64	*	2.2	3.5	0.42	**	2.2	3.5	0.42	**
3AAD	22.0	63.1	0.00	-	14.5	29.5	0.07	-	10.0	33.9	0.03	-	8.1	14.3	0.08	-
3BIW	18.7	55.7	0.00	-	2.0	4.5	0.69	**	2.0	4.5	0.69	**	1.6	5.6	0.78	**
3BP8	2.2	5.7	0.51	*	2.2	5.7	0.51	*	2.0	3.9	0.51	**	1.7	3.8	0.42	**
3BX7	2.8	7.1	0.54	*	2.5	4.8	0.36	**	2.5	4.8	0.36	**	3.1	4.7	0.40	**
3CPH	11.0	22.9	0.06	-	5.4	11.4	0.16	-	4.0	9.8	0.29	*	2.5	3.8	0.43	**
3D5S	17.8	48.8	0.00	-	12.7	39.7	0.00	-	1.8	3.5	0.78	**	1.8	3.5	0.78	**
3DAW	2.1	3.9	0.45	**	2.1	3.9	0.45	**	2.1	3.9	0.45	**	2.5	3.5	0.47	**
3EO1	2.1	8.3	0.59	*	2.0	9.4	0.36	**	2.0	14.2	0.56	**	2.0	9.4	0.36	**
3EOA	13.5	26.8	0.09	-	5.4	22.2	0.30	-	2.5	10.8	0.51	*	2.4	4.6	0.72	**
3F1P	2.7	7.2	0.37	*	1.9	3.2	0.67	**	1.9	3.2	0.67	**	1.9	3.2	0.67	**
3FN1	14.1	20.7	0.07	-	11.6	28.5	0.00	-	4.3	5.9	0.47	*	4.2	7.1	0.36	*
3G6D	8.4	19.6	0.02	-	5.1	9.5	0.08	-	1.9	5.2	0.62	**	1.9	12.9	0.62	**
3H11	16.2	23.0	0.00	-	7.5	11.1	0.26	-	6.2	8.0	0.31	*	5.6	5.1	0.47	*
3H2V	12.4	25.9	0.00	-	10.6	22.5	0.00	-	4.8	10.3	0.34	-	2.1	6.6	0.63	*
3HI6	4.1	11.1	0.34	-	2.9	10.6	0.29	*	2.5	4.6	0.45	**	3.0	4.6	0.34	**
3HMX	3.1	13.5	0.31	*	2.6	7.4	0.38	*	1.8	10.2	0.46	**	1.6	6.9	0.62	**
3K75	2.6	5.1	0.44	*	2.8	6.2	0.51	*	1.9	4.1	0.54	**	1.3	5.3	0.71	**
3L5W	2.9	11.8	0.76	*	1.4	4.9	0.90	**	1.6	4.4	0.72	**	1.3	4.7	0.76	**
3L89	2.9	5.0	0.53	*	2.9	5.0	0.53	*	2.9	5.0	0.53	*	2.8	5.4	0.41	*
3LVK	5.5	8.9	0.26	*	2.8	6.9	0.43	*	2.7	5.3	0.43	*	2.1	3.7	0.48	**
3MXW	13.4	23.0	0.00	-	2.9	5.1	0.43	*	1.5	4.3	0.62	**	1.5	4.3	0.62	**
3P57	11.8	42.1	0.00	-	3.1	17.8	0.61	*	1.9	9.9	0.74	**	1.2	3.7	0.72	**
3PC8	1.7	4.1	0.65	**	1.7	4.1	0.65	**	1.9	2.9	0.45	**	1.9	2.9	0.45	**
3R9A	18.9	34.4	0.00	-	2.7	11.8	0.58	*	2.7	11.8	0.58	*	3.2	7.6	0.32	*
3RVW	13.2	30.1	0.06	-	8.4	25.0	0.00	-	2.3	5.5	0.67	*	1.7	6.5	0.82	**
3S9D	14.1	22.7	0.00	-	3.3	13.7	0.26	*	3.1	5.5	0.32	*	2.8	3.8	0.40	**
3SGQ	7.5	15.0	0.05	-	2.0	4.0	0.45	**	2.1	4.3	0.54	**	1.6	4.5	0.39	**
3SZK	11.2	27.3	0.28	-	5.6	24.6	0.56	-	2.9	8.0	0.44	*	2.6	4.9	0.68	**
3V6Z	14.0	24.3	0.00	-	9.3	21.9	0.00	-	3.8	8.7	0.43	*	2.5	4.5	0.48	**
3VLB	17.4	28.8	0.09	-	0.9	2.5	0.79	***	0.9	2.5	0.79	***	0.9	2.5	0.79	***
4CPA	1.8	4.1	0.46	**	1.8	4.1	0.46	**	1.8	4.1	0.46	**	1.1	2.1	0.37	**
4DN4	2.5	2.9	0.47	**	2.5	2.9	0.47	**	2.5	2.9	0.47	**	2.1	4.4	0.30	**
4FQI	19.8	100.0	0.00	-	1.9	4.2	0.52	**	1.9	4.2	0.52	**	1.9	4.2	0.38	**
4FZA	14.4	27.9	0.00	-	13.4	28.5	0.00	-	3.0	5.2	0.29	*	3.0	5.2	0.29	*
4G6J	11.6	20.7	0.03	-	9.2	16.1	0.15	-	1.5	4.6	0.65	**	2.2	4.1	0.77	**
4G6M	0.9	2.4	0.83	***	0.9	2.4	0.83	**	0.9	2.4	0.83	**	1.2	2.3	0.90	**
4GAM	5.0	10.0	0.23	-	3.8	7.3	0.35	*	2.9	5.4	0.54	*	2.7	4.8	0.46	**
4GXU	13.3	24.9	0.20	-	8.2	19.4	0.18	-	3.5	10.5	0.29	*	1.5	7.1	0.50	**
4H03	17.4	34.0	0.00	-	1.9	7.5	0.40	**	1.9	7.5	0.40	**	1.5	3.9	0.52	**
4HX3	12.2	20.6	0.00	-	3.9	10.3	0.18	*	1.4	1.7	0.57	**	1.4	1.7	0.57	**
4IZ7	24.1	64.3	0.00	-	2.4	3.7	0.70	**	2.4	3.7	0.70	**	2.4	3.7	0.70	**
4JCV	3.1	5.7	0.66	*	1.3	2.9	0.71	**	1.3	2.9	0.71	**	1.8	4.0	0.67	**
4LW4	12.6	22.4	0.02	-	1.9	3.2	0.28	*	1.9	3.2	0.28	*	1.9	3.2	0.28	**
4M76	24.9	80.7	0.00	-	5.6	29.6	0.03	-	4.0	17.8	0.45	-	1.7	6.0	0.78	**
7CEI	1.5	4.5	0.73	**	1.5	4.5	0.73	**	1.0	1.7	0.81	***	1.0	1.9	0.79	***
9QFW	24.5	49.1	0.00	-	2.2	7.9	0.47	*	1.7	7.3	0.71	**	1.7	7.3	0.71	**

E. ERdj5 Supplemental Data



(a) mHC2 **(b)** mHC4

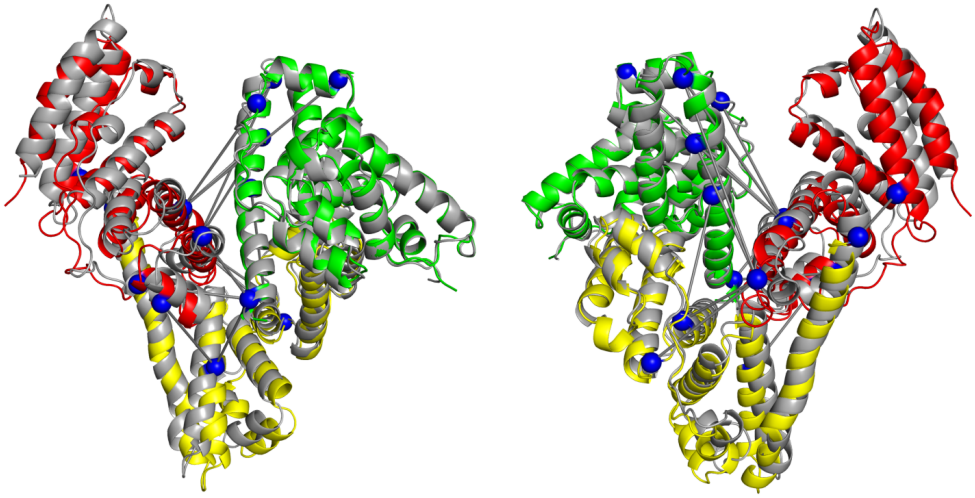
(c) mHC6 **(d)** NS1-7

Figure E.1. Contact frequency of solvent exposed residues on Trxb2 domain extracted from all 1000 final docking models. All four peptides exhibit similar binding patterns.

F. Nuclesome Remodeling Enzymes Supplemental Data

Table F.1. Cross-linking/mass spectrometry data for bovine serum albumin. The crosslinked residue pairs, the upper limit for the C_α-C_α distance used during modeling d_{max} and the actual distances as derived from the crystal structure (PDB 3V03) are listed. False positive crosslinks are marked in bold face.

Res1	Res2	d_{max} [Å]	PDB 3V03 [Å]
350	474	25.0	18.1
204	465	25.0	13.5
116	431	25.0	21.2
396	544	25.0	14.5
187	221	25.0	20.3
221	439	25.0	20.2
204	350	25.0	17.0
204	471	25.0	17.1
180	431	25.0	19.7
180	439	25.0	22.6
173	431	25.0	22.6
224	**439**	**25.0**	**27.0**
127	**431**	**25.0**	**35.9**
173	**439**	**25.0**	**32.8**



(a) Front view. **(b)** Back view.

Figure F.1. Integrative modeling results for bovine serum albumin. The top-ranked model is shown in cartoon representation green, yellow, red). Cross-links are drawn in gray. The crystal structure (PDB 3V03) is superimposed on the docking model (gray). The docking model is very close to the native structure (IRMSD $=$ 1.6 Å and fnat $=$ 0.69) and fulfills all the true positive cross-linking restraints within an upper limit of 25 Å. Furthermore, two of the false positive cross-links are not fulfilled in any of the 200 final models.

Table F.2. Comparison of different ISWI ATPase domain models to SAXS data of ISWI ATPase (residues 26-644; Bruetzel and Lipfert, unpublished data). Homology models were built on crystal structures of SF2 ATPases with MODELLER [482].

Name (PDB ID)	χ [FoXS]	χ^2 [CRYSOL]	Fulfills XL-MS data?
Sso (1Z6A)	2.69	6.26	no
Rad54 (1Z3I)	2.22	3.98	no
Chd1 (3MWY)	0.61	0.69	no
Swi2 (5ZHR)	3.02	7.65	partially
New dockingmodel	1.11	0.36	yes

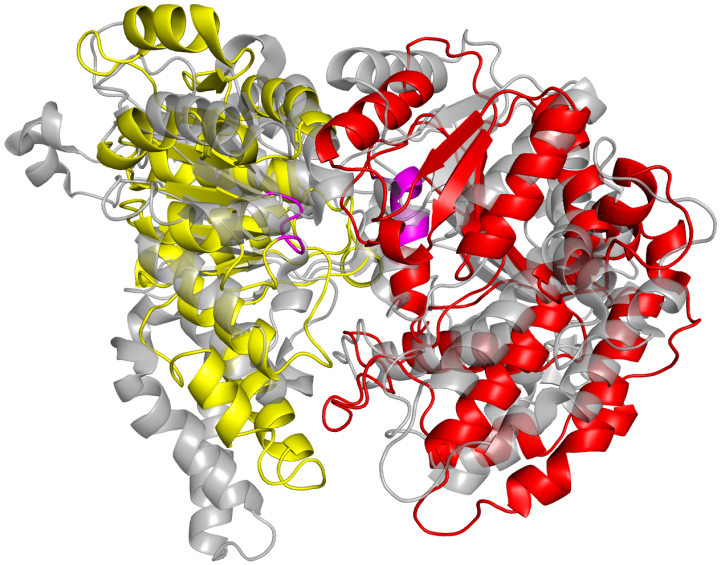


Figure F.2. Ab-initio docking results for ISWI ATPase lobe 1 (yellow) and lobe 2 (red). The distance between the connecting residues 351 and 352 was restrained. The catalytic DEAH and QAMDRAHR motifs are shown in magenta. The top-ranked model is superimposed on the crystal structure of the chromatin remodeling domain from Rad54 (PDB 1Z3I, gray).

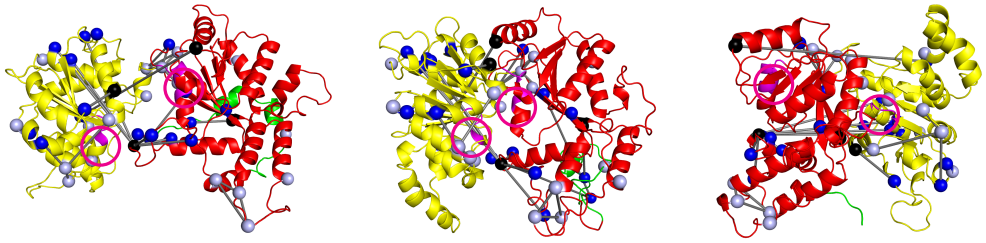
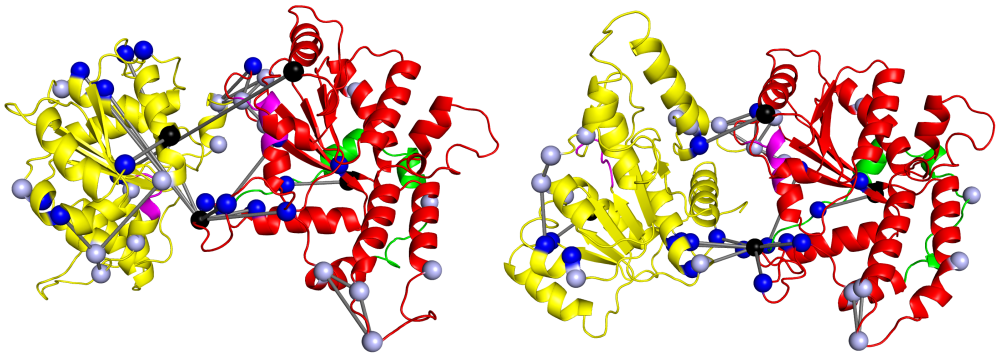


Figure F.3. Cross-linking data mapped to three available crystal structures of Snf2 ATPases. Cross-links are shown in gray, sites of photo-crosslinks are marked by black spheres. Sites where mono-links (in addition to cross-links) were detected are shown in light blue. The position of the catalytic motifs is highlighted in magenta.

Table F.3. Residues on lobe 2 which are contacted frequently by the AutoN+acidic patch in the docking models (interface post-prediction). Charged residues are highlighted in bold face.

Residue number	Residue name
508	**ARG**
591	ASN
590	SER
396	ASN
399	MET
455	GLN
457	THR
533	MET
592	GLN
403	**LYS**
458	**ARG**
395	GLN
456	MET
511	GLY
536	GLN
512	LEU
578	MET
480	GLN
532	GLN
392	MET



(a) Previous model **(b)** New docking model

Figure F.4. Comparison of old and new interpretation of ISWI ATPase cross-linking data (lobe 1 yellow, lobe 2 red, NegC green). Distances for residues with experimentally determined cross-links between the lobes are shown in gray, sites where mono-links were found in light blue. (a) Previous model [136]. The cross-links traverse the interior of the protein. (b) New model. The N-terminal part of lobe 1 (residue 116) is located at the interface between lobe 1 and lobe 2.

Table F.4. Cross-links between the histone H4 tail and the ISWI ATPase domain. High-confidence crosslinks used during docking are highlighted in bold face. The residue numbers are given according to the ISWI sequence.

H4 tail	ATPase	$d_{\max}$ [Å]
1	**482**	**20.0**
1	**495**	**20.0**
10	**468**	**20.0**
1	578	20.0
1	568	20.0
10	568	20.0
10	482	20.0
10	501	20.0
1	470	20.0
16	391	30.0
5	595	30.0

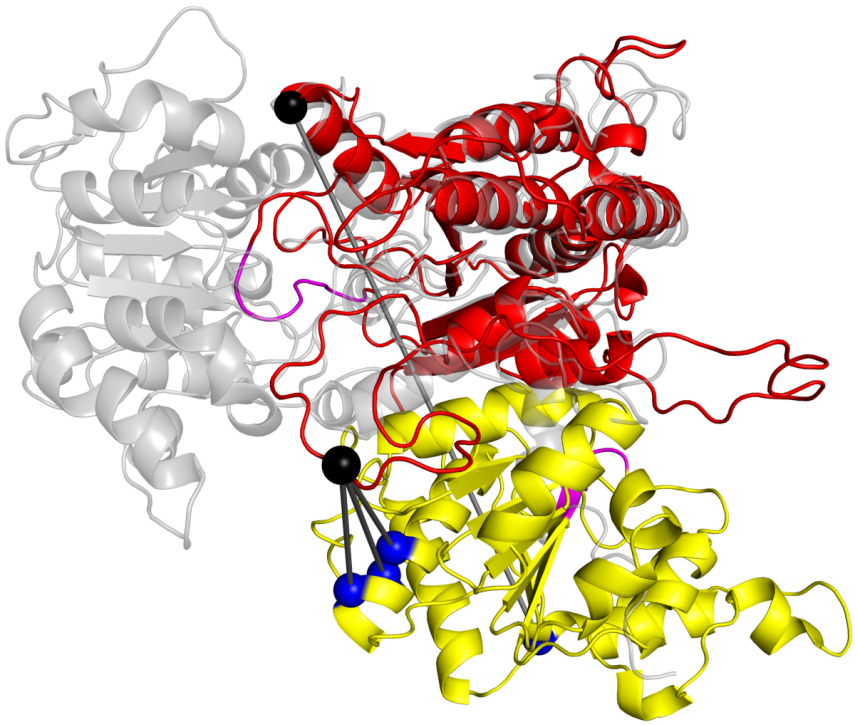


Figure F.5. Homology model of ISWI based on the crystal structure of Swi2/Snf2 chromatin remodeler (PDB 5HZR) [493]. Assuming high flexibility of the loop that contains methionine 578, all photo-crosslinks from this position could be fulfilled (distances ≈ 20 Å). Cross-links are shown in gray, sites of photo-crosslinks are marked by black spheres. As a reference, the crystal structure of Rad54 (PDB 1Z3I) is superimposed on lobe 2 and shown in gray.

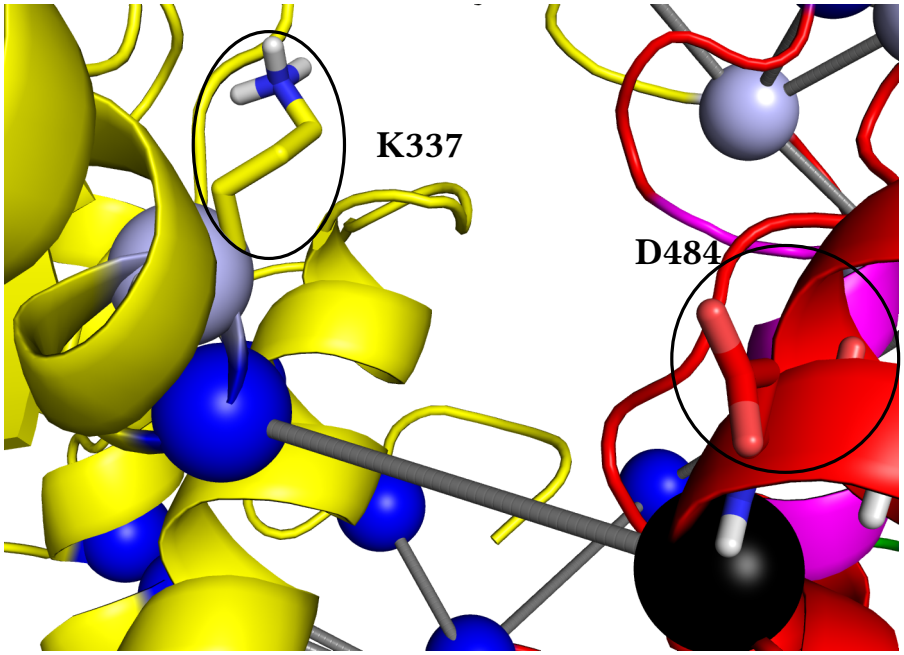
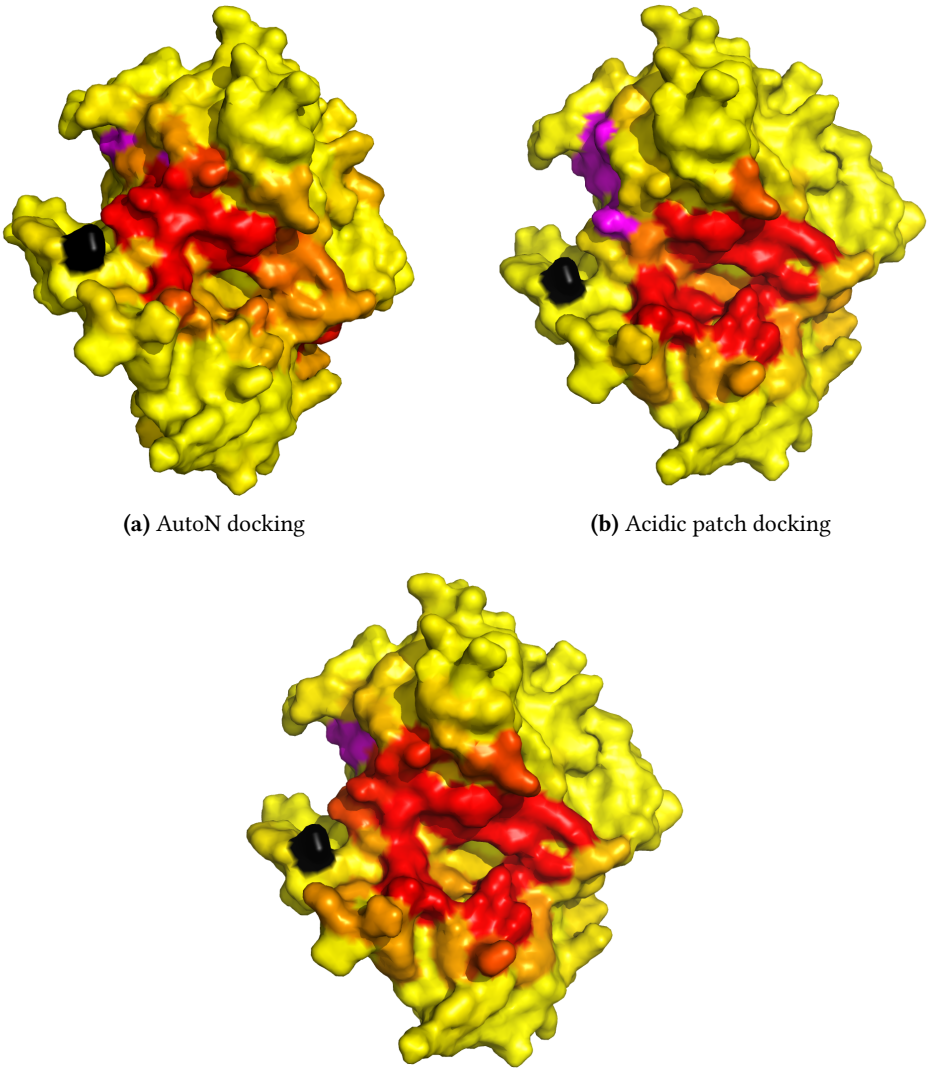


Figure F.6. Validation of our proposed ISWI ATPase model by mutating residues involved in a salt bridge at the interface. The residues involved in this ionic bond also showed strong correlations in co-evolution analysis [331]. To validate our proposed conformation, we will measure ISWI ATPase activity of charge mutants K337D and D484K/D485K.



(a) AutoN docking

(b) Acidic patch docking

(c) AutoN+acidic patch docking

Figure F.7. ISWI ATPase lobe 2 colored by contact frequency as extracted from the 1,000 final docking models. Frequently contacted residues are colored in red. Two residues were defined to be in contact if any of their heavy atoms were found within 5 Å distance. For reference, the catalytic motif on lobe 2 is shown in magenta and residue 578 is colored in black.

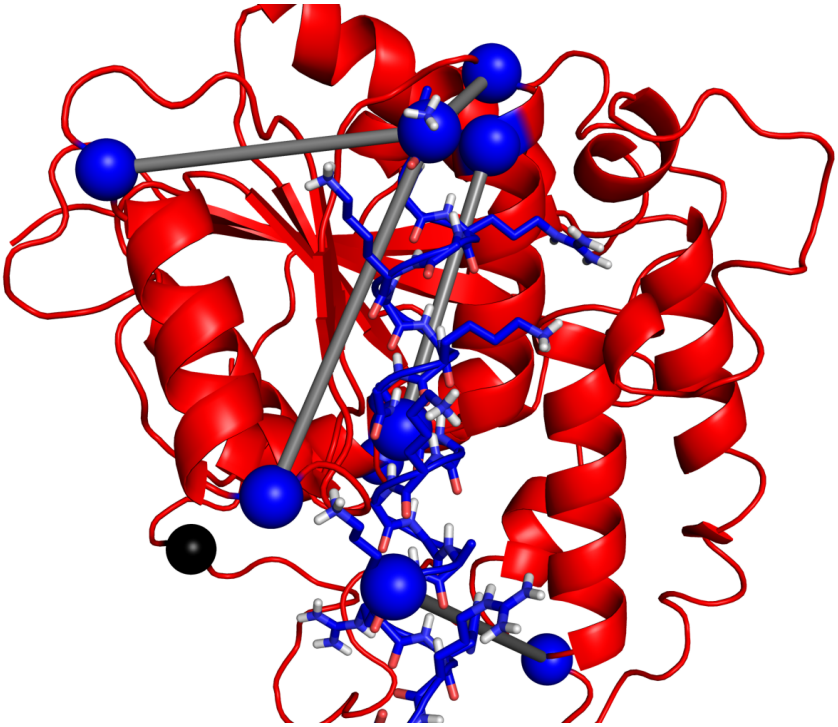


Figure F.8. Docking model of the histone H4 tail (residues 1-20) binding to ISWI ATPase lobe 2. Lobe 2 is drawn in red, the H4 tail in blue. The interaction was modeled with the pepATTRACT docking protocol [391]. Crosslinks from Table F.4 are shown in gray.

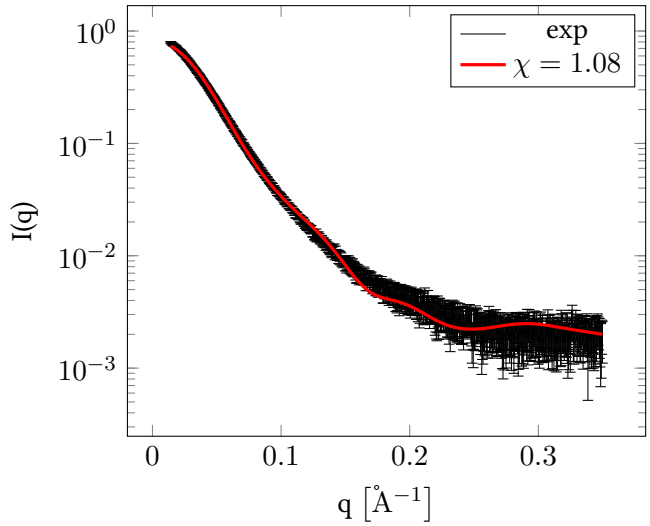


Figure F.9. Fit of full-length ISWI docking model to SAXS profile of full-length ISWI using FoXS [399] with default settings. The experimental data are shown in black, the calculated scattering profile in red.

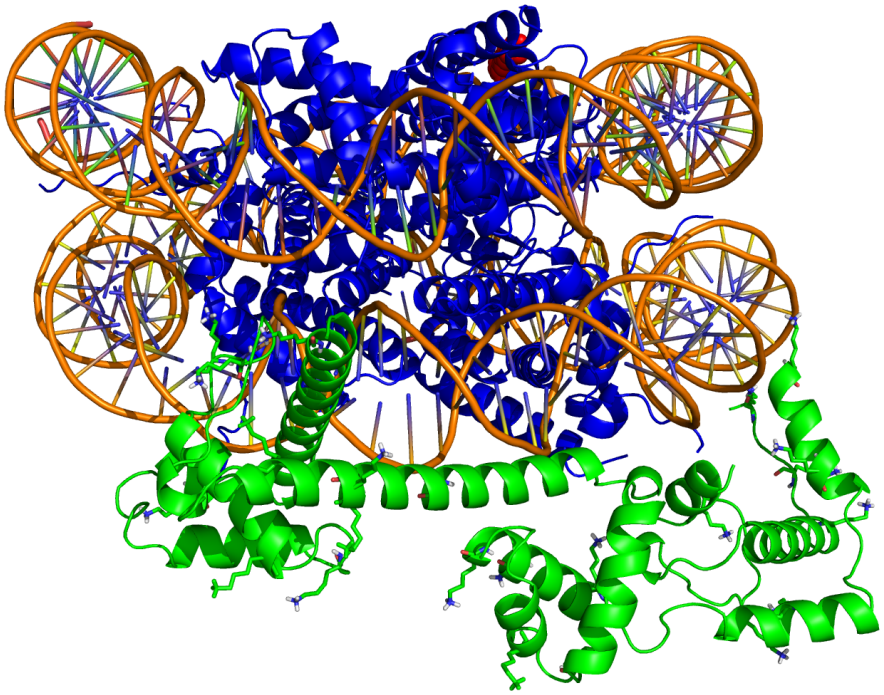


Figure F.10. HSS binding to DNA on nucleosome, alternative docking model. This model has rank 16 and shows a DNA-binding mode similar to [497].

Bibliography

[1] Ruben Abagyan, Maxim Totrov, and Dmitry Kuznetsov. "ICM–a new method for protein modeling and design: applications to docking and structure prediction from the distorted native conformation". *Journal of Computational Chemistry* 15.5 (1994), pp. 488–506. DOI: 10.1002/jcc.540150503.

[2] Mark James Abraham, Teemu Murtola, Roland Schulz, Szilárd Páll, Jeremy C Smith, Berk Hess, and Erik Lindahl. "GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers". *SoftwareX* 1 (2015), pp. 19–25. DOI: 10.1016/j.softx.2015.06.001.

[3] Chayan Acharya, Irina Kufareva, Andrey V. Ilatovskiy, and Ruben Abagyan. "PeptiSite: A structural database of peptide binding sites in 4D". *Biochemical Biophysical Research Communications* 445.4 (2014), pp. 717–723. DOI: 10.1016/j.bbrc.2013.12.132.

[4] Aashish N Adhikari, Jian Peng, Michael Wilde, Jinbo Xu, Karl F Freed, and Tobin R Sosnick. "Modeling large regions in proteins: Applications to loops, termini, and folding". *Protein Science* 21.1 (2012), pp. 107–121. DOI: 10.1002/pro.767.

[5] Mostafa H Ahmed, Francesca Spyrakis, Pietro Cozzini, Parijat K Tripathi, Andrea Mozzarelli, J Neel Scarsdale, Martin A Safo, and Glen E Kellogg. "Bound water at protein-protein interfaces: partners, roles and hydrophobic bubbles as a conserved motif". *PLOS One* 6.9 (2011), e24712. DOI: 10.1371/journal.pone.0024712.

[6] Frank Alber, Svetlana Dokudovskaya, Liesbeth M Veenhoff, Wenzhu Zhang, Julia Kipper, Damien Devos, Adisetyantari Suprapto, Orit Karni-Schmidt, Rosemary Williams, Brian T Chait, et al. "Determining the architectures of macromolecular assemblies". *Nature* 450.7170 (2007), pp. 683–694. DOI: 10.1038/nature06404.

[7] Frank Alber, Svetlana Dokudovskaya, Liesbeth M Veenhoff, Wenzhu Zhang, Julia Kipper, Damien Devos, Adisetyantari Suprapto, Orit Karni-Schmidt, Rosemary Williams, Brian T Chait, et al. "The molecular architecture of the nuclear pore complex". *Nature* 450.7170 (2007), pp. 695–701. DOI: 10.1038/nature06405.

[8] Patrick Aloy and Robert B Russell. "Ten thousand interactions for the molecular biologist". *Nature biotechnology* 22.10 (2004), pp. 1317–1321. DOI: 10.1038/nbt1018.

[9] Nozomi Ando, Yan Kung, Mehmet Can, Güneş Bender, Stephen W Ragsdale, and Catherine L Drennan. "Transient B12-dependent methyltransferase complexes revealed by small-angle X-ray scattering". *Journal of the American Chemical Society* 134.43 (2012), pp. 17945–17954. DOI: 10.1021/ja3055782.

[10] Ingemar André, Charlie EM Strauss, David B Kaplan, Philip Bradley, and David Baker. "Emergence of symmetry in homooligomeric biological assemblies". *Proceedings of the National Academy of Sciences* 105.42 (2008), pp. 16148–16152. DOI: 10.1073/pnas.0807576105.

[11] Jessica Andreani, Guilhem Faure, and Raphael Guerois. "InterEvScore: a novel coarse-grained interface scoring function using a multi-body statistical potential coupled to evolution". *Bioinformatics* 29 (14 2013), pp. 1742–1749. DOI: 10.1093/bioinformatics/btt260.

[12] Nelly Andrusier, Efrat Mashiach, Ruth Nussinov, and Haim J. Wolfson. "Principles of flexible protein–protein docking". *Proteins: Structure, Function, and Bioinformatics* 73.2 (2008), pp. 271–289. DOI: 10.1002/prot.22170.

[13] Iris Antes. "DynaDock: A new molecular dynamics-based algorithm for protein–peptide docking including receptor flexibility". *Proteins: Structure, Function, and Bioinformatics* 78.5 (2010), pp. 1084–1104. DOI: 10.1002/prot.22629.

[14] Iris Antes, Shirley W. I. Siu, and Thomas Lengauer. "DynaPred: A structure and sequence based method for the prediction of MHC class I binding peptide sequences and conformations". *Bioinformatics* 22.14 (2006), e16–e24. DOI: 10.1093/bioinformatics/btl216.

[15] James M Aramini, Yuanpeng J Huang, John R Cort, Sharon Goldsmith-Fischman, Rong Xiao, Liang-Yu Shih, Chi K Ho, Jinfeng Liu, Burkhard Rost, Barry Honig, et al. "Solution NMR structure of the 30S ribosomal protein S28E from Pyrococcus horikoshii". *Protein Science* 12.12 (2003), pp. 2823–2830. DOI: 10.1110/ps.03359003.

[16] Michelle Arkin. "Protein–protein interactions and cancer: small molecules going in for the kill". *Current Opinion in Chemical Biology* 9.3 (2005), pp. 317–324.

[17] Varsha D Badal, Petras J Kundrotas, and Ilya A Vakser. "Text mining for protein docking". *PLOS Computational Biology* 11.12 (2015), e1004630. DOI: 10.1371/journal.pcbi.1004630.

[18] Stéphanie Baelen, Frédérique Dewitte, Bernard Clantin, and Vincent Villeret. "Structure of the secretion domain of HxuA from Haemophilus influenzae". *Acta Crystallographica Section F: Structural Biology and Crystallization Communications* 69.12 (2013), pp. 1322–1327. DOI: 10.1107/S174430911302962X.

[19] Lies Baeten, Joke Reumers, Vicente Tur, François Stricher, Tom Lenaerts, Luis Serrano, Frederic Rousseau, and Joost Schymkowitz. "Reconstruction of protein backbones from the BriX collection of canonical protein fragments". *PLOS Computational Biology* 4.5 (2008), e1000083. DOI: 10.1371/journal.pcbi.1000083.

[20] Ranjit Prasad Bahadur, Pinak Chakrabarti, Francis Rodier, and Joël Janin. "Dissecting subunit interfaces in homodimeric proteins". *Proteins: Structure, Function, and Bioinformatics* 53.3 (2003), pp. 708–719. DOI: 10.1002/prot.10461.

[21] Ivet Bahar, Chakra Chennubhotla, and Dror Tobi. "Intrinsic dynamics of enzymes in the unbound state and relation to allosteric regulation". *Current opinion in structural biology* 17.6 (2007), pp. 633–640. DOI: 10.1016/j.sbi.2007.09.011.

[22] Xiao-Chen Bai, Greg McMullan, and Sjors HW Scheres. "How cryo-EM is revolutionizing structural biology". *Trends in Biochemical Sciences* 40.1 (2015), pp. 49–57. DOI: 10.1016/j.tibs.2014.10.005.

[23] Chandrajit L Bajaj, Rezaul Chowdhury, and Vinay Siddahanavalli. "*F*2 Dock: Fast Fourier Protein-Protein Docking". *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)* 8.1 (2011), pp. 45–58. DOI: 10.1109/TCBB.2009.57.

[24] Ahmet Bakan and Ivet Bahar. "The intrinsic dynamics of enzymes plays a dominant role in determining the structural changes induced upon inhibitor binding". *Proceedings of the National Academy of Sciences* 106.34 (2009), pp. 14349–14354. DOI: 10.1073/pnas.0904214106.

[25] Sivaraman Balakrishnan, Hetunandan Kamisetty, Jaime G Carbonell, Su-In Lee, and Christopher James Langmead. "Learning generative models for protein fold families". *Proteins: Structure, Function, and Bioinformatics* 79.4 (2011), pp. 1061–1078. DOI: 10.1002/prot.22934.

[26] J. B. Bale, S. Gonen, Y. Liu, W. Sheffler, D. Ellis, C. Thomas, D. Cascio, T. O. Yeates, T. Gonen, N. P. King, and D. Baker. "Accurate design of megadalton-scale two-component icosahedral protein complexes". *Science* 353.6297 (July 2016), pp. 389–394. DOI: 10.1126/science.aaf8818.

[27] Jaydeep Bardhan, Sanghyun Park, and Lee Makowski. "SoftWAXS: a computational tool for modeling wide-angle X-ray solution scattering from biomolecules". *J. Appl. Crystallogr.* 42.5 (2009), pp. 932–943. DOI: 10.1107/S0021889809032919.

[28] Amita Barik, Ranjit Prasad Bahadur, et al. "A protein–RNA docking benchmark (i): Nonredundant cases". *Proteins: Structure, Function, and Bioinformatics* 80.7 (2012), pp. 1866–1871. DOI: 10.1002/prot.24083.

[29] Karine Bastard, Chantal Prévost, and Martin Zacharias. "Accounting for loop flexibility during protein–protein docking". *Proteins: Structure, Function, and Bioinformatics* 62.4 (2006), pp. 956–969. DOI: 10.1002/prot.20770.

[30] Karine Bastard, Aurelien Thureau, Richard Lavery, and Chantal Prevost. "Docking macromolecules with flexible segments". *Journal of computational chemistry* 24.15 (2003), pp. 1910–1920. DOI: 10.1002/jcc.10329.

[31] Davide Baù, Amartya Sanyal, Bryan R Lajoie, Emidio Capriotti, Meg Byron, Jeanne B Lawrence, Job Dekker, and Marc A Marti-Renom. "The three-dimensional folding of the α-globin gene domain reveals formation of chromatin globules". *Nature structural & molecular biology* 18.1 (2011), pp. 107–114. DOI: 10.1038/nsmb.1936.

[32] Jutta Beneken, Jian Cheng Tu, Bo Xiao, Mutsuo Nuriya, Joseph P Yuan, Paul F Worley, and Daniel J Leahy. "Structure of the Homer EVH1 domain-peptide complex reveals a new twist in polyproline recognition". *Neuron* 26.1 (2000), pp. 143–154. DOI: 10.1016/S0896-6273(00)81145-9.

[33] Pascal Benkert, Silvio CE Tosatto, and Dietmar Schomburg. "QMEAN: A comprehensive scoring function for model quality assessment". *Proteins: Structure, Function, and Bioinformatics* 71.1 (2008), pp. 261–277. DOI: 10.1002/prot.21715.

[34] Melanie J Bennett, Michael P Schlunegger, and David Eisenberg. "3D domain swapping: a mechanism for oligomer assembly". *Protein Science* 4.12 (1995), pp. 2455–2468. DOI: 10.1002/pro.5560041202.

[35] Avraham Ben-Shimon and Miriam Eisenstein. "Computational Mapping of Anchoring Spots on Protein Surfaces". *Journal of Molecular Biology* 402.1 (2010), pp. 259–277. DOI: 10.1016/j.jmb.2010.07.021.

[36] Matthew L Bentley, Jacob E Corn, Ken C Dong, Qui Phung, Tommy K Cheung, and Andrea G Cochran. "Recognition of UbcH5c and the nucleosome by the Bmi1/Ring1b ubiquitin ligase complex". *The EMBO journal* 30.16 (2011), pp. 3285–3297. DOI: 10.1038/emboj.2011.243.

[37] Herman JC Berendsen, David van der Spoel, and Rudi van Drunen. "GROMACS: a message-passing parallel molecular dynamics implementation". *Computer Physics Communications* 91.1 (1995), pp. 43–56. DOI: 10.1016/0010-4655(95)00042-E.

[38] Konstantin Berlin, Carlos A Castaneda, Dina Schneidman-Duhovny, Andrej Sali, Alfredo Nava-Tudela, and David Fushman. "Recovering a representative conformational ensemble from underdetermined macromolecular structural data". *Journal of the American Chemical Society* 135.44 (2013), pp. 16595–16609. DOI: 10.1021/ja4083717.

[39] Helen Berman, Kim Henrick, and Haruki Nakamura. "Announcing the worldwide protein data bank". *Nature Structural & Molecular Biology* 10.12 (2003), pp. 980–980. DOI: 10.1038/nsb1203-980.

[40] Helen Berman, Kim Henrick, Haruki Nakamura, and John L Markley. "The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data". *Nucleic Acids Research* 35.suppl 1 (2007), pp. D301–D303. DOI: 10.1093/nar/gkl971.

[41] Pau Bernadó, Efstratios Mylonas, Maxim V Petoukhov, Martin Blackledge, and Dmitri I Svergun. "Structural characterization of flexible proteins using small-angle X-ray scattering". *Journal of the American Chemical Society* 129.17 (2007), pp. 5656–5664. DOI: 10.1021/ja069124n.

[42] Carrie Bernecky, Franz Herzog, Wolfgang Baumeister, Jürgen M Plitzko, and Patrick Cramer. "Structure of transcribing mammalian RNA polymerase II". *Nature* 529 (2016), pp. 551–554. DOI: 10.1038/nature16482.

[43] Matthew J Betts and Michael JE Sternberg. "An analysis of conformational changes on protein–protein association: implications for predictive docking". *Protein Engineering* 12.4 (1999), pp. 271–283. DOI: 10.1093/protein/12.4.271.

[44] Dipankar Bhandari, Tobias Raisch, Oliver Weichenrieder, Stefanie Jonas, and Elisa Izaurralde. "Structural basis for the Nanos-mediated recruitment of the CCR4–NOT complex and translational repression". *Genes & development* 28.8 (2014), pp. 888–901. DOI: 10.1101/gad.237289.113.

[45] Basudeb Bhattacharyya, Nicholas P George, Tiffany M Thurmes, Ruobo Zhou, Niketa Jani, Sarah R Wessel, Steven J Sandler, Taekjip Ha, and James L Keck. "Structural mechanisms of PriA-mediated DNA replication restart". *Proceedings of the National Academy of Sciences* 111.4 (2014), pp. 1373–1378. DOI: 10.1073/pnas.1318001111.

[46] Maciej Blaszczyk, Mateusz Kurcinski, Maksim Kouza, Lukasz Wieteska, Aleksander Debinski, Andrzej Kolinski, and Sebastian Kmiecik. "Modeling of protein–peptide interactions using the CABS-dock web server for binding site search and flexible docking". *Methods* 93 (2016), pp. 72–83. DOI: 10.1016/j.ymeth.2015.07.004.

[47] Romel Bobby, Karima Medini, Philipp Neudecker, Tet Verne Lee, Margaret A Brimble, Fiona J McDonald, J Shaun Lott, and Andrew J Dingley. "Structure and dynamics of human Nedd4-1 WW3 in complex with the αENaC PY motif". *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics* 1834.8 (2013), pp. 1632–1641. DOI: 10.1016/j.bbapap.2013.04.031.

[48] Andrew A Bogan and Kurt S Thorn. "Anatomy of hot spots in protein interfaces". *Journal of molecular biology* 280.1 (1998), pp. 1–9. DOI: 10.1006/jmbi.1998.1843.

[49] Tanggis Bohnuud, Lingqi Luo, Shoshana J Wodak, Alexandre MJJ Bonvin, Zhiping Weng, Sandor Vajda, Ora Schueler-Furman, and Dima Kozakov. "A benchmark testing ground for integrating homology modeling and protein docking". *Proteins: Structure, Function, and Bioinformatics* (2016). DOI: 10.1002/prot.25063.

[50] Massimiliano Bonomi, Riccardo Pellarin, Seung Joong Kim, Daniel Russel, Bryan A Sundin, Michael Riffle, Daniel Jaschob, Richard Ramsden, Trisha N Davis, Eric GD Muller, et al. "Determining Protein Complex Structures Based on a Bayesian Model of in Vivo Förster Resonance Energy Transfer (FRET) Data". *Molecular & Cellular Proteomics* 13.11 (2014), pp. 2812–2823. DOI: 10.1074/mcp.M114.040824.

[51] Alexandre MJJ Bonvin. "Flexible protein–protein docking". *Current Opinion in Structural Biology* 16.2 (2006), pp. 194–200. DOI: 10.1016/j.sbi.2006.02.002.

[52] Andrew J. Bordner and Ruben Abagyan. "Ab initio prediction of peptide-MHC binding geometry for diverse class I MHC allotypes". *Proteins: Structure, Function, and Bioinformatics* 63.3 (2006), pp. 512–526. DOI: 10.1002/prot.20831.

[53] S. E. Boyken, Z. Chen, B. Groves, R. A. Langan, G. Oberdorfer, A. Ford, J. M. Gilmore, C. Xu, F. DiMaio, J. H. Pereira, B. Sankaran, G. Seelig, P. H. Zwart, and D. Baker. "De novo design of protein homo-oligomers with modular hydrogen-bond network-mediated specificity". *Science* 352.6286 (May 2016), pp. 680–687. DOI: 10.1126/science.aad8865.

[54] Ineke Braakman and Daniel N Hebert. "Protein folding in the endoplasmic reticulum". *Cold Spring Harbor perspectives in biology* 5.5 (2013), a013201. DOI: 10.1101/cshperspect.a013201.

[55] G Patrick Brady and Kim A Sharp. "Entropy in protein folding and in protein-protein interactions". *Current opinion in structural biology* 7.2 (1997), pp. 215–221. DOI: 10.1016/S0959-440X(97)80028-0.

[56] Ryan Brenke, David R Hall, Gwo-Yu Chuang, Stephen R Comeau, Tanggis Bohnuud, Dmitri Beglov, Ora Schueler-Furman, Sandor Vajda, and Dima Kozakov. "Application of asymmetric statistical potentials to antibody–protein docking". *Bioinformatics* 28.20 (2012), pp. 2608–2614. DOI: 10.1093/bioinformatics/bts493.

[57] Alan Brown, Alexey Amunts, Xiao-chen Bai, Yoichiro Sugimoto, Patricia C Edwards, Garib Murshudov, Sjors HW Scheres, and V Ramakrishnan. "Structure of the large ribosomal subunit from human mitochondria". *Science* 346.6210 (2014), pp. 718–722. DOI: 10.1126/science.1258026.

[58] Christiam Camacho, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L Madden. "BLAST+: architecture and applications". *BMC Bioinformatics* 10.1 (2009), p. 1. DOI: 10.1186/1471-2105-10-421.

[59] Adrian A Canutescu and Roland L Dunbrack. "Cyclic coordinate descent: A robotics algorithm for protein loop closure". *Protein Science* 12.5 (2003), pp. 963–972. DOI: 10.1110/ps.0242703.

[60] DA Case, TA Darden, TE Cheatham III, CL Simmerling, J Wang, RE Duke, R Luo, RC Walker, W Zhang, KM Merz, F Paesani, DR Roe, A Roitberg, C Sagui, R Salomon-Ferrer, G Seabra, CL Simmerling, W Smith, J Swails, RC Walker, J Wang, RM Wolf, X Wu, and PA Kollman. "AMBER 14". *University of California, San Francisco* (2014).

[61] DA Case, TA Darden, TE Cheatham III, CL Simmerling, J Wang, RE Duke, R Luo, RC Walker, W Zhang, KM Merz, et al. "AMBER 12". *University of California, San Francisco* (2012).

[62] Pinak Chakrabarti and Joel Janin. "Dissecting protein–protein recognition sites". *Proteins: Structure, Function, and Bioinformatics* 47.3 (2002), pp. 334–343. DOI: 10.1002/prot.10085.

[63] Chiung-Wen Chang, Rafael Miguez Couñago, Simon J Williams, Mikael Bodén, and Bostjan Kobe. "Distinctive conformation of minor site-specific nuclear localization signals Bound to importin-α". *Traffic* 14.11 (2013), pp. 1144–1154. DOI: 10.1111/tra.12098.

[64] Henry N Chapman, Anton Barty, Michael J Bogan, Sébastien Boutet, Matthias Frank, Stefan P Hau-Riege, Stefano Marchesini, Bruce W Woods, Saša Bajt, W Henry Benner, et al. "Femtosecond diffractive imaging with a soft-X-ray free-electron laser". *Nature Physics* 2.12 (2006), pp. 839–843. DOI: 10.1038/nphys461.

[65] Sidhartha Chaudhury and Jeffrey J. Gray. "Conformer Selection and Induced Fit in Flexible Backbone Protein–Protein Docking Using Computational and NMR Ensembles". *Journal of Molecular Biology* 381.4 (2008), pp. 1068–1087. DOI: 10.1016/j.jmb.2008.05.042.

[66] Sidhartha Chaudhury, Berrondo Monica, Weitzner Brian D., Muthu Pravin, Bergman Hannah, and Gray Jeffrey J. "Benchmarking and Analysis of Protein Docking Performance in Rosetta v3.2". *PLOS One* 6.8 (Aug. 2011), e22477. DOI: 10.1371/journal.pone.0022477.

[67] Isaure Chauvot de Beauchêne, Sjoerd J de Vries, and Martin Zacharias. "Binding Site Identification and Flexible Docking of Single Stranded RNA to Proteins Using a Fragment-Based Approach". *PLOS Computational Biology* 12.1 (2016), e1004697. DOI: 10.1371/journal.pcbi.1004697.

[68] Po-chia Chen and Jochen S Hub. "Interpretation of Solution X-Ray Scattering by Explicit-Solvent Molecular Dynamics". *Biophysical Journal* 108.10 (2015), pp. 2573–2584. DOI: 10.1016/j.bpj.2015.03.062.

[69] Rong Chen, Li Li, and Zhiping Weng. "ZDOCK: an initial-stage protein-docking algorithm". *Proteins: Structure, Function, and Bioinformatics* 52.1 (2003), pp. 80–87. DOI: 10.1002/prot.10389.

[70] Rong Chen, Julian Mintseris, Joël Janin, and Zhiping Weng. "A protein–protein docking benchmark". *Proteins: Structure, Function, and Bioinformatics* 52.1 (2003), pp. 88–91. DOI: 10.1002/prot.10390.

[71] Xi Chen, Kathrin Hnida, Melissa Ann Graewert, Jan Terje Andersen, Rasmus Iversen, Anne Tuukkanen, Dmitri Svergun, and Ludvig M Sollid. "Structural basis for antigen recognition by transglutaminase 2-specific autoantibodies in celiac disease". *Journal of Biological Chemistry* 290.35 (2015), pp. 21365–21375. DOI: 10.1074/jbc.M115.669895.

[72] Zhuo Angel Chen, Anass Jawhari, Lutz Fischer, Claudia Buchen, Salman Tahir, Tomislav Kamenski, Morten Rasmussen, Laurent Lariviere, Jimi-Carlo Bukowski-Wills, Michael Nilges, et al. "Architecture of the RNA polymerase II–TFIIF complex revealed by cross-linking and mass spectrometry". *The EMBO Journal* 29.4 (2010), pp. 717–726. DOI: 10.1038/emboj.2009.401.

[73] Tammy Man-Kuang Cheng, Tom L Blundell, and Juan Fernandez-Recio. "pyDock: Electrostatics and desolvation for effective scoring of rigid-body protein–protein docking". *Proteins: Structure, Function, and Bioinformatics* 68.2 (2007), pp. 503–515. DOI: 10.1002/prot.21419.

[74] Cheom Gil Cheong, Soo Hyun Eom, Changsoo Chang, Dong Hae Shin, Hyun Kyu Song, Kyeongsik Min, Jin Ho Moon, Kyeong Kyu Kim, Kwang Yeon Hwang, and Se Won Suh. "Crystallization, molecular replacement solution, and refinement of tetrameric β-amylase from sweet potato". *Proteins: Structure, Function, and Bioinformatics* 21.2 (1995), pp. 105–117. DOI: 10.1002/prot.340210204.

[75] Anne Chevrel, Agathe Urvoas, Ines Li De La Sierra-gallay, Magali Aumont-Nicaise, Sandrine Moutel, Michel Desmadril, Franck Perez, Alexis Gautreau, Herman van Tilbeurgh, Philippe Minard, et al. "Specific GFP-binding artificial proteins (αRep): a new tool for in vitro to live cell applications". *Bioscience Reports* 35.4 (2015), e00223. DOI: 10.1042/BSR20150080.

[76] Tim Clackson and James A Wells. "A hot spot of binding energy in a hormone-receptor interface". *Science* 267.5196 (1995), pp. 383–386.

[77] C. R. Clapier, G. Längst, D. F. Corona, P. B. Becker, and K. P. Nightingale. "Critical role for the histone H4 N terminus in nucleosome remodeling by ISWI." eng. *Molecular and Cellular Biology* 21.3 (Feb. 2001), pp. 875–883. DOI: 10.1128/MCB.21.3.875-883.2001.

[78] Cedric R. Clapier and Bradley R. Cairns. "Regulation of ISWI involves inhibitory modules antagonized by nucleosomal epitopes." eng. *Nature* 492.7428 (Dec. 2012), pp. 280–284. DOI: 10.1038/nature11625.

[79] Cedric R. Clapier, Srinivas Chakravarthy, Carlo Petosa, Carlos Fernández-Tornero, Karolin Luger, and Christoph W. Müller. "Structure of the Drosophila nucleosome core particle highlights evolutionary constraints on the H2A-H2B histone dimer." eng. *Proteins* 71.1 (Apr. 2008), pp. 1–7. DOI: 10.1002/prot.21720.

[80] Cedric R. Clapier, Karl P. Nightingale, and Peter B. Becker. "A critical epitope for substrate recognition by the nucleosome remodeling ATPase ISWI." eng. *Nucleic Acids Research* 30.3 (Feb. 2002), pp. 649–655. DOI: 10.1093/nar/30.3.649.

[81] DavidJ. Clarke, Euan Murray, Ted Hupp, C.Logan Mackay, and PatR.R. Langridge-Smith. "Mapping a Noncovalent Protein–Peptide Interface by Top-Down FTICR Mass Spectrometry Using Electron Capture Dissociation". English. *Journal of the American Sociecty of Mass Spectrometry* 22.8 (2011), pp. 1432–1440. DOI: 10.1007/s13361-011-0155-3.

[82] G Marius Clore and Charles D Schwieters. "Docking of protein-protein complexes on the basis of highly ambiguous intermolecular distance restraints derived from 1HN/15N chemical shift mapping and backbone 15N-1H residual dipolar couplings using conjoined rigid body/torsion angle dynamics". *Journal of the American Chemical Society* 125.10 (2003), pp. 2902–2912. DOI: 10.1021/ja028893d.

[83] Stephen R Comeau, David W Gatchell, Sandor Vajda, and Carlos J Camacho. "ClusPro: an automated docking and discrimination method for the prediction of protein complexes". *Bioinformatics* 20.1 (2004), pp. 45–50. DOI: 10.1093/bioinformatics/btg371.

[84] Michael L Connolly. "Analytical molecular surface calculation". *Journal of Applied Crystallography* 16.5 (1983), pp. 548–558.

[85] Michael L Connolly. "Solvent-accessible surfaces of proteins and nucleic acids". *Science* 221.4612 (1983), pp. 709–713.

[86] Loredana Lo Conte, Cyrus Chothia, and Joël Janin. "The atomic structure of protein-protein recognition sites". *Journal of Molecular Biology* 285.5 (1999), pp. 2177–2198. DOI: 10.1006/jmbi.1998.2439.

[87] Davide FV Corona and John W Tamkun. "Multiple roles for ISWI in transcription, chromosome organization and DNA replication". *Biochimica et Biophysica Acta (BBA)–Gene Structure and Expression* 1677.1 (2004), pp. 113–119. DOI: 10.1016/j.bbaexp.2003.09.018.

[88] Gavin E Crooks, Gary Hon, John-Marc Chandonia, and Steven E Brenner. "WebLogo: a sequence logo generator". *Genome Research* 14.6 (2004), pp. 1188–1190. DOI: 10.1101/gr.849004.

[89] Douglas M Cyr and Carlos H Ramos. "Specification of Hsp70 function by type I and type II HSP40". *The Networking of Chaperones by Co-chaperones*. Springer, 2015, pp. 91–102. DOI: 10.1007/978-3-319-11731-7_4.

[90] Onur Dagliyan, Elizabeth A Proctor, Kevin M D'Auria, Feng Ding, and Nikolay V Dokholyan. "Structural and dynamic determinants of protein–peptide recognition". *Structure* 19.12 (2011), pp. 1837–1845. DOI: 10.1016/j.str.2011.09.014.

[91] Weiwei Dang, Mohamedi N Kagalwala, and Blaine Bartholomew. "Regulation of ISW2 by concerted action of histone H4 tail and extranucleosomal DNA". *Molecular and cellular biology* 26.20 (2006), pp. 7388–7396. DOI: 10.1128/MCB.01159-06.

[92] Rhiju Das. "Atomic-accuracy prediction of protein loop structures through an RNA-inspired ansatz". *PLOS One* 8.10 (2013), e74830. DOI: 10.1371/journal.pone.0074830.

[93] Rhiju Das, Ingemar André, Yang Shen, Yibing Wu, Alexander Lemak, Sonal Bansal, Cheryl H Arrowsmith, Thomas Szyperski, and David Baker. "Simultaneous prediction of protein folding and docking at high resolution". *Proceedings of the National Academy of Sciences* 106.45 (2009), pp. 18978–18983. DOI: 10.1073/pnas.0904407106.

[94] SJ de Vries and M Zacharias. "ATTRACT-EM: a new method for the computational assembly of large molecular machines using cryo-EM maps." *PLOS One* 7.12 (2012), e49733–e49733. DOI: 10.1371/journal.pone.0049733.

[95] Sjoerd J de Vries and Alexandre MJJ Bonvin. "CPORT: a consensus interface predictor and its performance in prediction-driven docking with HADDOCK". *PLOS One* 6.3 (2011), e17695. DOI: 10.1371/journal.pone.0017695.

[96] Sjoerd J de Vries, Isaure Chauvot de Beauchêne, Christina EM Schindler, and Martin Zacharias. "Cryo-EM data is superior to contact or interface information in integrative modeling". *Biophysical Journal* 110.4 (2016), pp. 462–465. DOI: 10.1016/j.bpj.2015.12.038.

[97] Sjoerd J. de Vries, Aalt D. J. van Dijk, Mickaël Krzeminski, Mark van Dijk, Aurelien Thureau, Victor Hsu, Tsjerk Wassenaar, and Alexandre M. J. J. Bonvin. "HADDOCK versus HADDOCK: New features and performance of HADDOCK2.0 on the CAPRI targets". *Proteins: Structure, Function, and Bioinformatics* 69.4 (2007), pp. 726–733. DOI: 10.1002/prot.21723.

[98] Sjoerd J de Vries, Marc van Dijk, and Alexandre MJJ Bonvin. "The HADDOCK web server for data-driven biomolecular docking". *Nature Protocols* 5.5 (2010), pp. 883–897. DOI: 10.1038/nprot.2010.32.

[99] Sjoerd J de Vries, Christina EM Schindler, Isaure Chauvot de Beauchêne, and Martin Zacharias. "A web interface for easy flexible protein–protein docking with ATTRACT". *Biophysical Journal* 108.3 (2015), pp. 462–465. DOI: 10.1016/j.bpj.2014.12.015.

[100] Sjoerd J de Vries and Martin Zacharias. "Grid-accelerated docking in ATTRACT". *In preparation.* (2016).

[101] Sjoerd de Vries and Martin Zacharias. "Flexible docking and refinement with a coarse-grained protein model using ATTRACT". *Proteins: Structure, Function, and Bioinformatics* 81.12 (2013), pp. 2167–2174. DOI: 10.1002/prot.24400.

[102] Peter Debye. "Zerstreuung von Röntgenstrahlen". *Annalen der Physik* 351.6 (1915), pp. 809–823.

[103] Yves Dehouck, Dimitri Gilis, and Marianne Rooman. "A new generation of statistical potentials for proteins". *Biophysical Journal* 90.11 (2006), pp. 4010–4017. DOI: 10.1529/biophysj.105.079434.

[104] Warren L DeLano. "Unraveling hot spots in binding interfaces: progress and challenges". *Current Opinion in Structural Biology* 12.1 (2002), pp. 14–20. DOI: 10.1016/S0959-440X(02)00283-X.

[105] Alexey Dementiev, Miljan Simonovic, Karl Volz, and Peter GW Gettins. "Canonical inhibitor-like interactions explain reactivity of α1-proteinase inhibitor Pittsburgh and antithrombin with proteinases". *Journal of Biological Chemistry* 278.39 (2003), pp. 37881–37887. DOI: 10.1074/jbc.M305195200.

[106] Francesca Diella, Niall Haslam, Claudia Chica, Aidan Budd, Sushama Michael, Nigel P Brown, Gilles Travé, and Toby J Gibson. "Understanding eukaryotic linear motifs and their role in cell signaling and regulation". *Front Biosci* 13 (2008), pp. 6580–6603. DOI: 10.2741/3175.

[107] Frank DiMaio, Michael D Tyka, Matthew L Baker, Wah Chiu, and David Baker. "Refinement of protein structures into low-resolution density maps using Rosetta". *Journal of Molecular Biology* 392.1 (2009), pp. 181–190. DOI: 10.1016/j.jmb.2009.07.008.

[108] Holger Dinkel, Kim Van Roey, Sushama Michael, Norman E Davey, Robert J Weatheritt, Diana Born, Tobias Speck, Daniel Krüger, Gleb Grebnev, Marta Kubań, et al. "The eukaryotic linear motif resource ELM: 10 years and counting". *Nucleic Acids Research* 42 (D1 2013), pp. D259–D266. DOI: 10.1093/nar/gkt1047.

[109] Sara E Dobbins, Victor I Lesk, and Michael JE Sternberg. "Insights into protein flexibility: the relationship between normal modes and conformational change upon protein–protein docking". *Proceedings of the National Academy of Sciences* 105.30 (2008), pp. 10390–10395. DOI: 10.1073/pnas.0802496105.

[110] Todd J. Dolinsky, Paul Czodrowski, Hui Li, Jens E. Nielsen, Jan H. Jensen, Gerhard Klebe, and Nathan A. Baker. "PDB2PQR: expanding and upgrading automated preparation of biomolecular structures for molecular simulations". *Nucleic Acids Research* 35.suppl 2 (2007), W522–W525. DOI: 10.1093/nar/gkm276.

[111] Todd J. Dolinsky, Jens E. Nielsen, J. Andrew McCammon, and Nathan A. Baker. "PDB2PQR: an automated pipeline for the setup of Poisson–Boltzmann electrostatics calculations". *Nucleic Acids Research* 32.suppl 2 (2004), W665–W667. DOI: 10.1093/nar/gkh381.

[112] Cyril Dominguez, Rolf Boelens, and Alexandre M. J. J. Bonvin. "HADDOCK: A protein-protein docking approach based on biochemical or biophysical information". *Journal of the American Chemical Society* 125.7 (2003), pp. 1731–1737. DOI: 10.1021/ja026939x.

[113] Guang Qiang Dong, Hao Fan, Dina Schneidman-Duhovny, Ben Webb, and Andrej Sali. "Optimized atomic statistical potentials: assessment of protein interfaces and loops". *Bioinformatics* 29 (24 2013), pp. 3158–3166. DOI: 10.1093/bioinformatics/btt560.

[114] Elad Donsky and Haim J Wolfson. "PepCrawler: a fast RRT-based algorithm for high-resolution refinement and binding affinity estimation of peptide inhibitors". *Bioinformatics* 27.20 (2011), pp. 2836–2842. DOI: 10.1093/bioinformatics/btr498.

[115] Julia Drebes, Madeleine Künz, Björn Windshügel, Alexey G Kikhney, Ingrid B Müller, Raphael J Eberle, Dominik Oberthür, Huaixing Cang, Dmitri I Svergun, Markus Perbandt, et al. "Structure of ThiM from Vitamin B1 biosynthetic pathway of Staphylococcus aureus–Insights into a novel pro-drug approach addressing MRSA infections". *Scientific Reports* 6.22871 (2016). DOI: 10.1038/srep22871.

[116] Jose M Duarte, Adam Srebniak, Martin A Schärer, and Guido Capitani. "Protein interface classification by evolutionary analysis". *BMC Bioinformatics* 13.1 (2012), p. 334. DOI: 10.1186/1471-2105-13-334.

[117] Mathieu Dubé, Mariana G Bego, Catherine Paquay, and Éric A Cohen. "Modulation of HIV-1-host interaction: role of the Vpu accessory protein". *Retrovirology* 7 (2010), p. 114. DOI: 10.1186/1742-4690-7-114.

[118] Joe Dundas, Zheng Ouyang, Jeffery Tseng, Andrew Binkowski, Yaron Turpaz, and Jie Liang. "CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues". *Nucleic Acids Research* 34.suppl 2 (2006), W116–W118. DOI: 10.1093/nar/gkl282.

[119] Matthew Dunne, Stefan Leicht, Boris Krichel, Haydyn DT Mertens, Andrew Thompson, Jeroen Krijgsveld, Dmitri I Svergun, Natalia Gómez-Torres, Sonia Garde, Charlotte Uetrecht, et al. "Crystal structure of the CTP1L endolysin reveals how its activity is regulated by a secondary translation product". *Journal of Biological Chemistry* 291.10 (2015), pp. 4882–4893. DOI: 10.1074/jbc.M115.671172.

[120] Harald Dürr, Christian Körner, Marisa Müller, Volker Hickmann, and Karl-Peter Hopfner. "X-ray structures of the Sulfolobus solfataricus SWI2/SNF2 ATPase core and its complex with DNA". *Cell* 121.3 (2005), pp. 363–373. DOI: 10.1016/j.cell.2005.03.026.

[121] Uwe Ehmann. "Accelerating protein-protein docking calculations using graphics processing units". MA thesis. Technical University of Munich, 2015.

[122] Ugur Emekli, Dina Schneidman-Duhovny, Haim J Wolfson, Ruth Nussinov, and Turkan Haliloglu. "HingeProt: automated prediction of hinges in protein structures". *Proteins: Structure, Function, and Bioinformatics* 70.4 (2008), pp. 1219–1227. DOI: 10.1002/prot.21613.

[123] Juan Esquivel-Rodrıguez, Yifeng David Yang, and Daisuke Kihara. "Multi-LZerD: Multiple protein docking for asymmetric complexes". *Proteins: Structure, Function, and Bioinformatics* 80.7 (2012), pp. 1818–1833. DOI: 10.1002/prot.24079.

[124] Volker A Eyrich, Marc A Martı-Renom, Dariusz Przybylski, Mallur S Madhusudhan, Andras Fiser, Florencio Pazos, Alfonso Valencia, Andrej Sali, and Burkhard Rost. "EVA: continuous automatic evaluation of protein structure prediction servers". *Bioinformatics* 17.12 (2001), pp. 1242–1243. DOI: 10.1093/bioinformatics/17.12.1242.

[125] Hao Fan, Dina Schneidman-Duhovny, John J Irwin, Guangqiang Dong, Brian K Shoichet, and Andrej Sali. "Statistical potential for modeling and ranking of protein–ligand interactions". *Journal of Chemical Information and Modeling* 51.12 (2011), pp. 3078–3092. DOI: 10.1021/ci200377u.

[126] Javier Fernandez-Martinez, Jeremy Phillips, Matthew D Sekedat, Ruben Diaz-Avalos, Javier Velazquez-Muriel, Josef D Franke, Rosemary Williams, David L Stokes, Brian T Chait, Andrej Sali, et al. "Structure–function mapping of a heptameric module in the nuclear pore complex". *The Journal of Cell Biology* 196.4 (2012), pp. 419–434. DOI: 10.1083/jcb.201109008.

[127] Juan Fernández-Recio, Maxim Totrov, and Ruben Abagyan. "ICM-DISCO docking by global energy optimization with fully flexible side-chains". *Proteins: Structure, Function, and Bioinformatics* 52.1 (2003), pp. 113–117. DOI: 10.1002/prot.10383.

[128] Juan Fernández-Recio, Maxim Totrov, and Ruben Abagyan. "Identification of Protein–Protein Interaction Sites from Docking Energy Landscapes". *Journal of Molecular Biology* 335.3 (2004), pp. 843–865. DOI: 10.1016/j.jmb.2003.10.069.

[129] Sébastien Fiorucci and Martin Zacharias. "Binding site prediction and improved scoring during flexible protein–protein docking with ATTRACT". *Proteins: Structure, Function, and Bioinformatics* 78.15 (2010), pp. 3131–3139. DOI: 10.1002/prot.22808.

[130] Daniel Fischer, Shuo Liang Lin, Haim L Wolfson, and Ruth Nussinov. "A geometry-based suite of molecular docking processes". *Journal of Molecular Biology* 248.2 (1995), pp. 459–477. DOI: 10.1016/S0022-2836(95)80063-8.

[131] Emil Fischer. "Einfluss der Configuration auf die Wirkung der Enzyme". *Berichte der deutschen chemischen Gesellschaft* 27.3 (1894), pp. 2985–2993.

[132] Lutz Fischer, Zhuo Angel Chen, and Juri Rappsilber. "Quantitative cross-linking/mass spectrometry using isotope-labelled cross-linkers". *Journal of proteomics* 88 (2013), pp. 120–128. DOI: 10.1016/j.jprot.2013.03.005.

[133] Niels Fischer, Andrey L Konevega, Wolfgang Wintermeyer, Marina V Rodnina, and Holger Stark. "Ribosome dynamics and tRNA movement by time-resolved electron cryomicroscopy". *Nature* 466.7304 (2010), pp. 329–333. DOI: 10.1038/nature09206.

[134] András Fiser and Andrej Sali. "ModLoop: automated modeling of loops in protein structures". *Bioinformatics* 19.18 (2003), pp. 2500–2501. DOI: 10.1093/bioinformatics/btg362.

[135] Gregory C Flynn, Jan Pohl, Mark T Flocco, James E Rothman, et al. "Peptide-binding specificity of the molecular chaperone BiP". *Nature* 353.6346 (1991), pp. 726–730. DOI: 10.1038/353726a0.

[136] Ignasi Forné, Johanna Ludwigsen, Axel Imhof, Peter B Becker, and Felix Mueller-Planitz. "Probing the conformation of the ISWI ATPase domain with genetically encoded photoreactive crosslinkers and mass spectrometry". *Molecular & Cellular Proteomics* 11.4 (2012), pp. M111–012088. DOI: 10.1074/mcp.M111.012088.

[137] Friedrich Förster, Benjamin Webb, Kristin A Krukenberg, Hiro Tsuruta, David A Agard, and Andrej Sali. "Integration of small-angle X-ray scattering data into structural modeling of proteins and their assemblies". *Journal of Molecular Biology* 382.4 (2008), pp. 1089–1106. DOI: 10.1016/j.jmb.2008.07.074.

[138] Clémence Fournier, Ann Smith, and Philippe Delepelaire. "Haem release from haemopexin by HxuA allows Haemophilus influenzae to escape host nutritional immunity". *Molecular Microbiology* 80.1 (2011), pp. 133–148. DOI: 10.1111/j.1365-2958.2011.07562.x.

[139] Daniel Franke, Cy M Jeffries, and Dmitri I Svergun. "Correlation Map, a goodness-of-fit test for one-dimensional X-ray scattering spectra". *Nature Methods* 12 (2015), pp. 419–422. DOI: 10.1038/nmeth.3358.

[140] Daniel Franke and Dmitri I Svergun. "DAMMIF, a program for rapid ab-initio shape determination in small-angle scattering". *Journal of Applied Crystallography* 42.2 (2009), pp. 342–346. DOI: 10.1107/S0021889809000338.

[141] RDB Fraser, TP MacRae, and E Suzuki. "An improved method for calculating the contribution of solvent to the X-ray diffraction pattern of biological molecules". *Journal of Applied Crystallography* 11.6 (1978), pp. 693–694. DOI: 10.1107/S0021889878014296.

[142] Simon A Fromm, Vincent Truffault, Julia Kamenz, Joerg E Braun, Niklas A Hoffmann, Elisa Izaurralde, and Remco Sprangers. "The structural basis of Edc3-and Scd6-mediated activation of the Dcp1: Dcp2 mRNA decapping complex". *The EMBO Journal* 31.2 (2012), pp. 279–290. DOI: 10.1038/emboj.2011.408.

[143] Naoshi Fukuhara and Takeshi Kawabata. "HOMCOS: a server to predict interacting protein pairs and interacting sites by homology modeling of complex structures". *Nucleic Acids Research* 36.suppl 2 (2008), W185–W189. DOI: 10.1093/nar/gkn218.

[144] Henry A Gabb, Richard M Jackson, and Michael JE Sternberg. "Modelling protein docking using shape complementarity, electrostatics and biochemical information". *Journal of Molecular Biology* 272.1 (1997), pp. 106–120. DOI: 10.1006/jmbi.1997.1203.

[145] T. R. Gamble, F. F. Vajdos, S. Yoo, D. K. Worthylake, M. Houseweart, W. I. Sundquist, and C. P. Hill. "Crystal structure of human cyclophilin A bound to the amino-terminal domain of HIV-1 capsid." eng. *Cell* 87.7 (Dec. 1996), pp. 1285–1294.

[146] Ying Gao, Dominique Douguet, Andrey Tovchigrechko, and Ilya A Vakser. "DOCKGROUND system of databases for protein recognition studies: Unbound structures for docking". *Proteins: Structure, Function, and Bioinformatics* 69.4 (2007), pp. 845–851. DOI: 10.1002/prot.21714.

[147] Eleanor J Gardiner, Peter Willett, and Peter J Artymiuk. "GAPDOCK: A genetic algorithm approach to protein docking in CAPRI round 1". *Proteins: Structure, Function, and Bioinformatics* 52.1 (2003), pp. 10–14. DOI: 10.1002/prot.10386.

[148] Leonardo Garma, Srayanta Mukherjee, Pralay Mitra, and Yang Zhang. "How many protein-protein interactions types exist in nature?" *PLOS One* 7.6 (2012), e38913. DOI: 10.1371/journal.pone.0038913.

[149] José Ignacio Garzon, José Ramón Lopéz-Blanco, Carles Pons, Julio Kovacs, Ruben Abagyan, Juan Fernandez-Recio, and Pablo Chacon. "FRODOCK: a new approach for fast rotational protein–protein docking". *Bioinformatics* 25.19 (2009), pp. 2544–2551. DOI: 10.1093/bioinformatics/btp447.

[150] Hamed Tabatabaei Ghomi, Jared J Thompson, and Markus A Lill. "Are distance-dependent statistical potentials considering three interacting bodies superior to two-body statistical potentials for protein structure prediction?" *Journal of Bioinformatics and computational Biology* 12.05 (2014), p. 1450022. DOI: 10.1142/S021972001450022X.

[151] Otto Glatter. "Computation of distance distribution-functions and scattering functions of models for small-angle scattering experiments". *Acta Physica Austriaca* 52.3-4 (1980), pp. 243–256.

[152] Christoph Göbl, Tobias Madl, Bernd Simon, and Michael Sattler. "NMR approaches for structural analysis of multidomain proteins and complexes in solution". *Progress in Nuclear Magnetic Resonance Spectroscopy* 80 (2014), pp. 26–63. DOI: 10.1016/j.pnmrs.2014.05.003.

[153] Holger Gohlke and David A Case. "Converging free energy estimates: MM-PB (GB) SA studies on the protein–protein complex Ras–Raf". *Journal of Computational Chemistry* 25.2 (2004), pp. 238–250. DOI: 10.1002/jcc.10379.

[154] Holger Gohlke and Gerhard Klebe. "Statistical potentials and scoring functions applied to protein–ligand binding". *Current Opinion in Structural Biology* 11.2 (2001), pp. 231–235. DOI: 10.1016/S0959-440X(00)00195-0.

[155] Joseph A Goldman, Joseph D Garlick, and Robert E Kingston. "Chromatin remodeling by imitation switch (ISWI) class ATP-dependent remodelers is stimulated by histone variant H2A.Z". *Journal of Biological Chemistry* 285.7 (2010), pp. 4645–4651. DOI: 10.1074/jbc.M109.072348.

[156] S. Gonen, F. DiMaio, T. Gonen, and D. Baker. "Design of ordered two-dimensional arrays mediated by noncovalent protein-protein interfaces". *Science* 348.6241 (June 2015), pp. 1365–1368. DOI: 10.1126/science.aaa9897.

[157] Zhou Gong, Charles D Schwieters, and Chun Tang. "Conjoined Use of EM and NMR in RNA Structure Refinement". *PLOS One* 10.3 (2015), e0120445. DOI: 10.1371/journal.pone.0120445.

[158] David S Goodsell and Arthur J Olson. "Structural symmetry and protein function". *Annual Review of Biophysics and Biomolecular Structure* 29.1 (2000), pp. 105–153. DOI: 10.1146/annurev.biophys.29.1.105.

[159] Christy R Grace, David Ban, Jaeki Min, Anand Mayasundari, Lie Min, Kristin E Finch, Lyra Griffiths, Nagakumar Bharatham, Donald Bashford, R Kiplin Guy, et al. "Monitoring ligand-induced protein ordering in drug discovery". *Journal of Molecular Biology* 428.6 (2016), pp. 1290–1303. DOI: 10.1016/j.jmb.2016.01.016.

[160] Melissa A Graewert and Dmitri I Svergun. "Impact and progress in small and wide angle X-ray scattering (SAXS and WAXS)". *Current Opinion in Structural Biology* 23.5 (2013), pp. 748–754. DOI: 10.1016/j.sbi.2013.06.007.

[161] Jeffrey J Gray, Stewart Moughon, Chu Wang, Ora Schueler-Furman, Brian Kuhlman, Carol A Rohl, and David Baker. "Protein–protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations". *Journal of Molecular Biology* 331.1 (2003), pp. 281–299. DOI: 10.1016/S0022-2836(03)00670-3.

[162] Andrei Grigoriev. "On the number of protein–protein interactions in the yeast proteome". *Nucleic Acids Research* 31.14 (2003), pp. 4157–4161. DOI: 10.1093/nar/gkg466.

[163] Sam Z Grinter and Xiaoqin Zou. "A Bayesian statistical approach of improving knowledge-based scoring functions for protein–ligand interactions". *Journal of Computational Chemistry* 35.12 (2014), pp. 932–943. DOI: 10.1002/jcc.23579.

[164] Alexander Grishaev, Liang Guo, Thomas Irving, and Ad Bax. "Improved fitting of solution X-ray scattering data to macromolecular structures and structural ensembles by explicit water modeling". *Journal of the American Chemical Society* 132.44 (2010), pp. 15484–15486. DOI: 10.1021/ja106173n.

[165] Alexander Grishaev, Vitali Tugarinov, Lewis E Kay, Jill Trewhella, and Ad Bax. "Refined solution structure of the 82-kDa enzyme malate synthase G from joint NMR and synchrotron SAXS restraints". *Journal of Biomolecular NMR* 40.2 (2008), pp. 95–106. DOI: 10.1007/s10858-007-9211-5.

[166] Dominik Gront, Daniel W Kulp, Robert M Vernon, Charlie EM Strauss, and David Baker. "Generalized fragment picking in Rosetta: design, protocols and applications". *PLOS One* 6.8 (2011), e23294. DOI: 10.1371/journal.pone.0023294.

[167] Raik Grünberg, Michael Nilges, and Johan Leckner. "Flexibility and conformational entropy in protein-protein binding". *Structure* 14.4 (2006), pp. 683–693.

[168] Tim Grüne, Jan Brzeski, Anton Eberharter, Cedric R Clapier, Davide FV Corona, Peter B Becker, and Christoph W Müller. "Crystal structure and functional analysis of a nucleosome recognition module of the remodeling factor ISWI". *Molecular cell* 12.2 (2003), pp. 449–460.

[169] Christopher J Guerriero and Jeffrey L Brodsky. "The delicate balance between secreted protein folding and endoplasmic reticulum-associated degradation in human physiology". *Physiological Reviews* 92.2 (2012), pp. 537–576. DOI: 10.1152/physrev.00027.2011.

[170] Mainak Guharoy and Pinak Chakrabarti. "Conservation and relative importance of residues across protein-protein interfaces". *Proceedings of the National Academy of Sciences* 102.43 (2005), pp. 15447–15452. DOI: 10.1073/pnas.0505425102.

[171] Mainak Guharoy, Joël Janin, and Charles H. Robert. "Side-chain rotamer transitions at protein-protein interfaces". *Proteins: Structure, Function, and Bioinformatics* 78.15 (2010), pp. 3219–3225. DOI: 10.1002/prot.22821.

[172] Peter Güntert. "Automated NMR structure calculation with CYANA". *Protein NMR Techniques* (2004), pp. 353–378.

[173] Stefan Günther, Patrick May, Andreas Hoppe, Cornelius Frömmel, and Robert Preissner. "Docking without docking: ISEARCH–prediction of interactions using known interfaces". *Proteins: Structure, Function, and Bioinformatics* 69.4 (2007), pp. 839–844.

[174] Masatoshi Hagiwara, Ken-ichi Maegawa, Mamoru Suzuki, Ryo Ushioda, Kazutaka Araki, Yushi Matsumoto, Jun Hoseki, Kazuhiro Nagata, and Kenji Inaba. "Structural basis of an ERAD pathway mediated by the ER-resident protein disulfide reductase ERdj5". *Molecular Cell* 41.4 (2011), pp. 432–444. DOI: 10.1016/j.molcel.2011.01.021.

[175] Ali Hamiche, Ju-Gyeong Kang, Cynthia Dennis, Hua Xiao, and Carl Wu. "Histone tails modulate nucleosome mobility and regulate ATP-dependent nucleosome sliding by NURF". *Proceedings of the National Academy of Sciences* 98.25 (2001), pp. 14316–14321. DOI: 10.1073/pnas.251421398.

[176] F Ulrich Hartl. "Chaperone-assisted protein folding: the path to discovery from a personal perspective". *Nature Medicine* 17.10 (2011), pp. 1206–1210. DOI: 10.1038/nm.2467.

[177] Glenn Hauk, Jeffrey N McKnight, Ilana M Nodelman, and Gregory D Bowman. "The chromodomains of the Chd1 chromatin remodeler regulate DNA access to the ATPase motor". *Molecular Cell* 39.5 (2010), pp. 711–723. DOI: 10.1016/j.molcel.2010.08.012.

[178] Janosch Hennig, Sjoerd J de Vries, Klaus DM Hennig, Leah Randles, Kylie J Walters, Maria Sunnerhagen, and Alexandre MJJ Bonvin. "MTMDAT-HADDOCK: high-throughput, protein complex structure modeling based on limited proteolysis and mass spectrometry". *BMC Structural Biology* 12.29 (2012). DOI: 10.1186/1472-6807-12-29.

[179] Torsten Herrmann, Peter Güntert, and Kurt Wüthrich. "Protein NMR structure determination with automated NOE assignment using the new software CANDID and the torsion angle dynamics algorithm DYANA". *Journal of Molecular Biology* 319.1 (2002), pp. 209–227. DOI: 10.1016/S0022-2836(02)00241-3.

[180] Franz Herzog, Abdullah Kahraman, Daniel Boehringer, Raymond Mak, Andreas Bracher, Thomas Walzthoeni, Alexander Leitner, Martin Beck, Franz-Ulrich Hartl, Nenad Ban, et al. "Structural probing of a protein phosphatase 2A network by chemical cross-linking and mass spectrometry". *Science* 337.6100 (2012), pp. 1348–1352. DOI: 10.1126/science.1221483.

[181] Csaba Hetényi and David van der Spoel. "Efficient docking of peptides to proteins without prior knowledge of the binding site". *Protein Sci* 11.7 (2002), pp. 1729–1737. DOI: 10.1110/ps.020230.

[182] Konrad Hinsen. "Analysis of domain motions by approximate normal mode calculations". *Proteins: Structure, Function, and Genetics* 33.3 (1998), pp. 417–429.

[183] Mark S Hipp, Sae-Hun Park, and F Ulrich Hartl. "Proteostasis impairment in protein-misfolding and-aggregation diseases". *Trends in Cell Biology* 24.9 (2014), pp. 506–514. DOI: 10.1016/j.tcb.2014.05.003.

[184] Bosco K Ho and Ken A Dill. "Folding very short peptides using molecular dynamics". *PLOS Computational Biology* 2.4 (2006), e27. DOI: 10.1371/journal.pcbi.0020027.

[185] Daniel Holtby, Shuai Cheng Li, and Ming Li. "LoopWeaver: loop modeling by the weighted scaling of verified proteins". *Journal of Computational Biology* 20.3 (2013), pp. 212–223. DOI: 10.1089/cmb.2012.0078.

[186] R. W. W. Hooft, C. Sander, and G. Vriend. "Verification of Protein Structures: Side-Chain Planarity". *Journal of Applied Crystallography* 29.6 (1996), pp. 714–716. DOI: 10.1107/S0021889896008631.

[187] Rob W.W. Hooft, Chris Sander, and Gerrit Vriend. "Objectively judging the quality of a protein structure from a Ramachandran plot". *Computer Applications in the Biosciences: CABIOS* 13.4 (1997), pp. 425–430. DOI: 10.1093/bioinformatics/13.4.425.

[188] Viktor Hornak, Robert Abel, Asim Okur, Bentley Strockbine, Adrian Roitberg, and Carlos Simmerling. "Comparison of multiple Amber force fields and development of improved protein backbone parameters". *Proteins: Structure, Function, and Bioinformatics* 65.3 (2006), pp. 712–725. DOI: 10.1002/prot.21123.

[189] Raghavendra Hosur, Jinbo Xu, Jadwiga Bienkowska, and Bonnie Berger. "iWRAP: an interface threading approach with application to prediction of cancer-related protein–protein interactions". *Journal of Molecular Biology* 405.5 (2011), pp. 1295–1310. DOI: 10.1016/j.jmb.2010.11.025.

[190] Yang Hsia, Jacob B. Bale, Shane Gonen, Dan Shi, William Sheffler, Kimberly K. Fong, Una Nattermann, Chunfu Xu, Po-Ssu Huang, Rashmi Ravichandran, Sue Yi, Trisha N. Davis, Tamir Gonen, Neil P. King, and David Baker. "Design of a hyperstable 60-subunit protein icosahedron". *Nature* 535.7610 (June 2016), pp. 136–139. DOI: 10.1038/nature18010.

[191] Shang-Te D Hsu, Christine Peter, Wilfred F van Gunsteren, and Alexandre MJJ Bonvin. "Entropy calculation of HIV-1 Env gp120, its receptor CD4, and their complex: an analysis of configurational entropy changes upon complexation". *Biophysical Journal* 88.1 (2005), pp. 15–24. DOI: 10.1529/biophysj.104.044933.

[192] Sheng-You Huang. "Search strategies and evaluation in protein–protein docking: principles, advances and challenges". *Drug Discovery Today* 19.8 (2014), pp. 1081–1096. DOI: 10.1016/j.drudis.2014.02.005.

[193] Sheng-You Huang and Xiaoqin Zou. "A nonredundant structure dataset for benchmarking protein-RNA computational docking". *Journal of Computational Chemistry* 34.4 (2013), pp. 311–318. DOI: 10.1002/jcc.23149.

[194] Sheng-You Huang and Xiaoqin Zou. "An iterative knowledge-based scoring function for protein–protein recognition". *Proteins: Structure, Function, and Bioinformatics* 72.2 (2008), pp. 557–579. DOI: 10.1002/prot.21949.

[195] Sheng-You Huang and Xiaoqin Zou. "An iterative knowledge-based scoring function to predict protein–ligand interactions: I. Derivation of interaction potentials". *Journal of Computational Chemistry* 27.15 (2006), pp. 1866–1875. DOI: 10.1002/jcc.20504.

[196] Sheng-You Huang and Xiaoqin Zou. "Ensemble docking of multiple protein structures: considering protein structural variations in molecular docking". *Proteins: Structure, Function, and Bioinformatics* 66.2 (2007), pp. 399–421. DOI: 10.1002/prot.21214.

[197] William Humphrey, Andrew Dalke, and Klaus Schulten. "VMD: visual molecular dynamics". *Journal of Molecular Graphics* 14.1 (1996), pp. 33–38. DOI: 10.1016/0263-7855(96)00018-5.

[198] Greg L Hura, Angeli L Menon, Michal Hammel, Robert P Rambo, Farris L Poole Ii, Susan E Tsutakawa, Francis E Jenney Jr, Scott Classen, Kenneth A Frankel, Robert C Hopkins, et al. "Robust, high-throughput solution structural analyses by small angle X-ray scattering (SAXS)". *Nature Methods* 6.8 (2009), pp. 606–612. DOI: 10.1038/nmeth.1353.

[199] Howook Hwang, Brian Pierce, Julian Mintseris, Joël Janin, and Zhiping Weng. "Protein–protein docking benchmark version 3.0". *Proteins: Structure, Function, and Bioinformatics* 73.3 (2008), pp. 705–709. DOI: 10.1002/prot.22106.

[200] Howook Hwang, Thom Vreven, Joël Janin, and Zhiping Weng. "Protein–protein docking benchmark version 4.0". *Proteins: Structure, Function, and Bioinformatics* 78.15 (2010), pp. 3111–3114. DOI: 10.1002/prot.22830.

[201] Howook Hwang, Thom Vreven, Brian G Pierce, Jui-Hung Hung, and Zhiping Weng. "Performance of ZDOCK and ZRANK in CAPRI rounds 13–19". *Proteins: Structure, Function, and Bioinformatics* 78.15 (2010), pp. 3104–3110. DOI: 10.1002/prot.22764.

[202] Lilia M Iakoucheva, Predrag Radivojac, Celeste J Brown, Timothy R O'Connor, Jason G Sikes, Zoran Obradovic, and A Keith Dunker. "The importance of intrinsic disorder for protein phosphorylation". *Nucleic Acids Research* 32.3 (2004), pp. 1037–1049. DOI: 10.1093/nar/gkh253.

[203] Yuval Inbar, Hadar Benyamini, Ruth Nussinov, and Haim J Wolfson. "Prediction of multimolecular assemblies by multiple docking". *Journal of Molecular Biology* 349.2 (2005), pp. 435–447. DOI: 10.1016/j.jmb.2005.03.039.

[204] Kohki Ishikawa, Haruki Nakamura, Kosuke Morikawa, and Shigenori Kanaya. "Stabilization of Escherichia coli ribonuclease HI by cavity-filling mutations within a hydrophobic core". *Biochemistry* 32.24 (1993), pp. 6171–6178. DOI: 10.1021/bi00075a009.

[205] Takashi Ito, Tomoko Chiba, Ritsuko Ozawa, Mikio Yoshida, Masahira Hattori, and Yoshiyuki Sakaki. "A comprehensive two-hybrid analysis to explore the yeast protein interactome". *Proceedings of the National Academy of Science* 98.8 (2001), pp. 4569–4574. DOI: 10.1073/pnas.061034498.

[206] Joël Janin. "The targets of CAPRI rounds 6–12". *Proteins: Structure, Function, and Bioinformatics* 69.4 (2007), pp. 699–703. DOI: 10.1002/prot.21689.

[207] Joël Janin. "Wet and dry interfaces: the role of solvent in protein–protein and protein–DNA recognition". *Structure* 7.12 (1999), R277–R279. DOI: 10.1016/S0969-2126(00)88333-1.

[208] Miguel St-Jean, Julien Lafrance-Vanasse, Brigitte Liotard, and Jurgen Sygusch. "High Resolution Reaction Intermediates of Rabbit Muscle Fructose-1,6-bisphosphate Aldolase–substrate cleavage and induced fit". *Journal of Biological Chemistry* 280.29 (2005), pp. 27262–27270. DOI: 10.1074/jbc.M502413200.

[209] Brian Jiménez-Garcıa, Carles Pons, Dmitri I. Svergun, Pau Bernadó, and Juan Fernández-Recio. "pyDockSAXS: protein–protein complex structure by SAXS and computational docking". *Nucleic Acids Research* 43 (2015), W356–W361. DOI: 10.1093/nar/gkv368.

[210] Mark Johnson, Irena Zaretskaya, Yan Raytselis, Yuri Merezhuk, Scott McGinnis, and Thomas L Madden. "NCBI BLAST: a better web interface". *Nucleic Acids Research* 36.suppl 2 (2008), W5–W9. DOI: 10.1093/nar/gkn201.

[211] Pall F Jonsson and Paul A Bates. "Global topological features of cancer proteins in the human interactome". *Bioinformatics* 22.18 (2006), pp. 2291–2297. DOI: 10.1093/bioinformatics/btl390.

[212] William L. Jorgensen and Julian. Tirado-Rives. "The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin". *Journal of the American Chemical Society* 110.6 (1988), pp. 1657–1666. DOI: 10.1021/ja00214a001.

[213] Abdullah Kahraman, Franz Herzog, Alexander Leitner, George Rosenberger, Ruedi Aebersold, and Lars Malmström. "Cross-link guided molecular modeling with Rosetta". *PLOS One* 8.9 (2013), e73411. DOI: 10.1371/journal.pone.0073411.

[214] Hetunandan Kamisetty, Sergey Ovchinnikov, and David Baker. "Assessing the utility of coevolution-based residue–residue contact predictions in a sequence-and structure-rich era". *Proceedings of the National Academy of Sciences* 110.39 (2013), pp. 15674–15679. DOI: 10.1073/pnas.1314045110.

[215] Harm H Kampinga and Elizabeth A Craig. "The HSP70 chaperone machinery: J proteins as drivers of functional specificity". *Nature Reviews Molecular Cell Biology* 11.8 (2010), pp. 579–592. DOI: 10.1038/nrm2941.

[216] Srinivasaraghavan Kannan and Martin Zacharias. "Enhanced sampling of peptide and protein conformations using replica exchange simulations with a peptide backbone biasing-potential". *Proteins: Structure, Function, and Bioinformatics* 66.3 (2007), pp. 697–706. DOI: 10.1002/prot.21258.

[217] Gozde Kar, Attila Gursoy, and Ozlem Keskin. "Human cancer protein-protein interaction network: a structural perspective". *PLOS Computational Biology* 5.12 (2009), e1000601. DOI: 10.1371/journal.pcbi.1000601.

[218] Ezgi Karaca and Alexandre MJJ Bonvin. "A multidomain flexible docking approach to deal with large conformational changes in the modeling of biomolecular complexes". *Structure* 19.4 (2011), pp. 555–565. DOI: 10.1016/j.str.2011.01.014.

[219] Ezgi Karaca and Alexandre MJJ Bonvin. "On the usefulness of ion-mobility mass spectrometry and SAXS data in scoring docking decoys". *Acta Crystallographica, Section D: Biological Crystallography* 69.5 (2013), pp. 683–694. DOI: 10.1107/S0907444913007063.

[220] Ezgi Karaca, Adrien SJ Melquiond, Sjoerd J de Vries, Panagiotis L Kastritis, and Alexandre MJJ Bonvin. "Building macromolecular assemblies by information-driven docking introducing the HADDOCK multibody docking server". *Molecular & Cellular Proteomics* 9.8 (2010), pp. 1784–1794. DOI: 10.1074/mcp.M000051-MCP201.

[221] Panagiotis L Kastritis and Alexandre MJJ Bonvin. "Are scoring functions in protein-protein docking ready to predict interactomes? Clues from a novel binding affinity benchmark". *Journal of Proteome Research* 9.5 (2010), pp. 2216–2225. DOI: 10.1021/pr9009854.

[222] Panagiotis L Kastritis, Aalt DJ van Dijk, and Alexandre MJJ Bonvin. "Explicit treatment of water molecules in data-driven protein–protein docking: the solvated HADDOCKing approach". *Computational Drug Discovery and Design*. Vol. 819. Methods in Molecular Biology. Springer, 2012. Chap. 22, pp. 355–374. DOI: 10.1007/978-1-61779-465-0_22.

[223] Panagiotis L Kastritis, Koen M Visscher, Aalt DJ van Dijk, and Alexandre MJJ Bonvin. "Solvated protein–protein docking using Kyte-Doolittle-based water preferences". *Proteins: Structure, Function, and Bioinformatics* 81.3 (2013), pp. 510–518. DOI: 10.1002/prot.24210.

[224] Ephraim Katchalski-Katzir, Isaac Shariv, Miriam Eisenstein, Asher A Friesem, Claude Aflalo, and Ilya A Vakser. "Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques". *Proceedings of the National Academy of Sciences* 89.6 (1992), pp. 2195–2199.

[225] Kristian W Kaufmann, Gordon H Lemmon, Samuel L DeLuca, Jonathan H Sheehan, and Jens Meiler. "Practically useful: what the Rosetta protein modeling suite can do for you". *Biochemistry* 49.14 (2010), pp. 2987–2998. DOI: 10.1021/bi902153g.

[226] Ozlem Keskin. "Binding induced conformational changes of proteins correlate with their intrinsic fluctuations: a case study of antibodies". *BMC structural biology* 7.31 (2007). DOI: 10.1186/1472-6807-7-31.

[227] Kyong-Rim Kieffer-Kwon, Zhonghui Tang, Ewy Mathe, Jason Qian, Myong-Hee Sung, Guoliang Li, Wolfgang Resch, Songjoon Baek, Nathanael Pruett, Lars Grøntved, et al. "Interactome maps of mouse gene regulatory domains reveal basic principles of transcriptional regulation". *Cell* 155.7 (2013), pp. 1507–1520. DOI: 10.1016/j.cell.2013.11.039.

[228] Tatsiana Kirys, Anatoly M Ruvinsky, Deepak Singla, Alexander V Tuzikov, Petras J Kundrotas, and Ilya A Vakser. "Simulated unbound structures for benchmarking of protein docking in the Dockground resource". *BMC bioinformatics* 16.243 (2015). DOI: 10.1186/s12859-015-0672-3.

[229] Junsu Ko, Dongseon Lee, Hahnbeom Park, Evangelos A Coutsias, Julian Lee, and Chaok Seok. "The FALC-Loop web server for protein loop modeling". *Nucleic Acids Research* 39.suppl 2 (2011), W210–W214. DOI: 10.1093/nar/gkr352.

[230] T-P Ko, John Day, Alexander J Malkin, and Alexander McPherson. "Structure of orthorhombic crystals of beef liver catalase". *Acta Crystallographica, Section D: Biological Crystallography* 55.8 (1999), pp. 1383–1394. DOI: 10.1107/S0907444999007052.

[231] Merika Treants Koday, Jorgen Nelson, Aaron Chevalier, Michael Koday, Hannah Kalinoski, Lance Stewart, Lauren Carter, Travis Nieusma, Peter S. Lee, Andrew B. Ward, Ian A. Wilson, Ashley Dagley, Donald F. Smee, David Baker, and Deborah Heydenburg Fuller. "A Computationally Designed Hemagglutinin Stem-Binding Protein Provides In Vivo Protection from Influenza Independent of a Host Immune Response". *PLOS Pathogens* 12.2 (Feb. 2016). Ed. by George F. Gao, e1005409. DOI: 10.1371/journal.ppat.1005409.

[232] Oliver Korb, Thomas Stutzle, and Thomas E Exner. "Empirical scoring functions for advanced protein-ligand docking with PLANTS". *Journal of Chemical Information and Modeling* 49.1 (2009), pp. 84–96. DOI: 10.1021/ci800298z.

[233] Tanja Kortemme and David Baker. "Computational design of protein-protein interactions". *Current Opinion in Chemical Biology* 8.1 (2004), pp. 91–97. DOI: 10.1016/j.cbpa.2003.12.008.

[234] Tanja Kortemme, Alexandre V Morozov, and David Baker. "An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein–protein complexes". *Journal of Molecular Biology* 326.4 (2003), pp. 1239–1259. DOI: 10.1016/S0022-2836(03)00021-4.

[235] Max Kotlyar, Chiara Pastrello, Flavia Pivetta, Alessandra Lo Sardo, Christian Cumbaa, Han Li, Taline Naranian, Yun Niu, Zhiyong Ding, Fatemeh Vafaee, et al. "In silico prediction of physical protein interactions and characterization of interactome orphans". *Nature Methods* 12.1 (2015), pp. 79–84. DOI: 10.1038/nmeth.3178.

[236] Noga Kowalsman and Miriam Eisenstein. "Inherent limitations in protein–protein docking procedures". *Bioinformatics* 23.4 (2007), pp. 421–426. DOI: 10.1093/bioinformatics/btl524.

[237] Dima Kozakov, Ryan Brenke, Stephen R Comeau, and Sandor Vajda. "PIPER: an FFT-based protein docking program with pairwise potentials". *Proteins: Structure, Function, and Bioinformatics* 65.2 (2006), pp. 392–406. DOI: 10.1002/prot.21117.

[238] Evgeny Krissinel and Kim Henrick. "Inference of macromolecular assemblies from crystalline state". *Journal of Molecular Biology* 372.3 (2007), pp. 774–797. DOI: 10.1016/j.jmb.2007.05.022.

[239] Georgii G. Krivov, Maxim V. Shapovalov, and Roland L. Dunbrack. "Improved prediction of protein side-chain conformations with SCWRL4". *Proteins: Structure, Function, and Bioinformatics* 77.4 (2009), pp. 778–795. DOI: 10.1002/prot.22488.

[240] Marcin Król, Alexander L. Tournier, and Paul A. Bates. "Flexible relaxation of rigid-body docking solutions". *Proteins: Structure, Function, and Bioinformatics* 68.1 (2007), pp. 159–169. DOI: 10.1002/prot.21391.

[241] Dennis M Krüger, José Ignacio Garzón, Pablo Chacón, and Holger Gohlke. "DrugScore PPI knowledge-based potentials used as scoring and objective function in protein-protein docking". *PLOS One* 9.2 (2014), e89466. DOI: 10.1371/journal.pone.0089466.

[242] Florian Krull. "Improving scoring functions for protein docking by machine learning and learning data". PhD thesis. Freie Universität Berlin, 2016.

[243] Mickaël Krzeminski, Joseph A Marsh, Chris Neale, Wing-Yiu Choy, and Julie D Forman-Kay. "Characterization of disordered proteins with ENSEMBLE". *Bioinformatics* 29.3 (2013), pp. 398–399. DOI: 10.1093/bioinformatics/bts701.

[244] Werner Kühlbrandt. "The resolution revolution". *Science* 343.6178 (2014), pp. 1443–1444. DOI: 10.1126/science.1251652.

[245] Petras J. Kundrotas, Zhengwei Zhu, Joël Janin, and Ilya A. Vakser. "Templates are available to model nearly all complexes of structurally characterized proteins". *Proceedings of the National Academy of Sciences* (2012), pp. 9438–9441. DOI: 10.1073/pnas.1200678109.

[246] Mateusz Kurcinski and Andrzej Kolinski. "Hierarchical modeling of protein interactions". *Journal of Molecular Modeling* 13.6-7 (2007), pp. 691–698. DOI: 10.1007/s00894-007-0177-8.

[247] Alexis Lamiable, Pierre Thévenet, Julien Rey, Marek Vavrusa, Philippe Derreumaux, and Pierre Tufféry. "PEP-FOLD3: faster de novo structure prediction for linear peptides in solution and in complex". *Nucleic Acids Research* 44.W1 (2016), W449–W454. DOI: 10.1093/nar/gkw329.

[248] Oliver F Lange, Paolo Rossi, Nikolaos G Sgourakis, Yifan Song, Hsiau-Wei Lee, James M Aramini, Asli Ertekin, Rong Xiao, Thomas B Acton, Gaetano T Montelione, et al. "Determination of solution structures of proteins up to 40 kDa using CS-Rosetta with sparse NMR data from deuterated samples". *Proceedings of the National Academy of Sciences* 109.27 (2012), pp. 10873–10878. DOI: 10.1073/pnas.1203013109.

[249] Gernot Längst and Peter B Becker. "Nucleosome mobilization and positioning by ISWI-containing chromatin-remodeling factors". *Journal of Cell Science* 114.14 (2001), pp. 2561–2568.

[250] Keren Lasker, Friedrich Förster, Stefan Bohn, Thomas Walzthoeni, Elizabeth Villa, Pia Unverdorben, Florian Beck, Ruedi Aebersold, Andrej Sali, and Wolfgang Baumeister. "Molecular architecture of the 26S proteasome holocomplex determined by an integrative approach". *Proceedings of the National Academy of Sciences* 109.5 (2012), pp. 1380–1387. DOI: 10.1073/pnas.1120559109.

[251] Keren Lasker, Jeremy L Phillips, Daniel Russel, Javier Velazquez-Muriel, Dina Schneidman-Duhovny, Elina Tjioe, Ben Webb, Avner Schlessinger, and Andrej Sali. "Integrative structure modeling of macromolecular assemblies from proteomics data". *Molecular & Cellular Proteomics* 9.8 (2010), pp. 1689–1702. DOI: 10.1074/mcp.R110.000067.

[252] Keren Lasker, Maya Topf, Andrej Sali, and Haim J Wolfson. "Inferential optimization for simultaneous fitting of multiple components into a CryoEM map of their assembly". *Journal of Molecular Biology* 388.1 (2009), pp. 180–194. DOI: 10.1016/j.jmb.2009.02.031.

[253] Assaf Lavi, Chi Ho Ngan, Dana Movshovitz-Attias, Tanggis Bohnuud, Christine Yueh, Dmitri Beglov, Ora Schueler-Furman, and Dima Kozakov. "Detection of peptide-binding sites on protein surfaces: The first step toward the modeling and targeting of peptide-mediated interactions". *Proteins: Structure, Function, and Bioinformatics* 81.12 (2013), pp. 2096–2105. DOI: 10.1002/prot.24422.

[254] Julian Lee, Dongseon Lee, Hahnbeom Park, Evangelos A Coutsias, and Chaok Seok. "Protein loop modeling by using fragment assembly and analytical loop closure". *Proteins: Structure, Function, and Bioinformatics* 78.16 (2010), pp. 3428–3436. DOI: 10.1002/prot.22849.

[255] Alexander Leitner, Marco Faini, Florian Stengel, and Ruedi Aebersold. "Crosslinking and Mass Spectrometry: An Integrated Technology to Understand the Structure and Function of Molecular Machines". *Trends in Biochemical Sciences* 41.1 (2016), pp. 20–32. DOI: 10.1016/j.tibs.2015.10.008.

[256] Alexander Leitner, Thomas Walzthoeni, Abdullah Kahraman, Franz Herzog, Oliver Rinner, Martin Beck, and Ruedi Aebersold. "Probing native protein structures by chemical cross-linking, mass spectrometry, and bioinformatics". *Molecular & Cellular Proteomics* 9.8 (2010), pp. 1634–1649. DOI: 10.1074/mcp.R000001-MCP201.

[257] Marc F. Lensink, Raúl Méndez, and Shoshana J. Wodak. "Docking and scoring protein complexes: CAPRI 3rd Edition". *Proteins: Structure, Function, and Bioinformatics* 69.4 (2007), pp. 704–718. DOI: 10.1002/prot.21804.

[258] Marc F Lensink and Shoshana J Wodak. "Blind predictions of protein interfaces by docking calculations in CAPRI". *Proteins: Structure, Function, and Bioinformatics* 78.15 (2010), pp. 3085–3095. DOI: 10.1002/prot.22850.

[259] Marc F. Lensink and Shoshana J. Wodak. "Docking and scoring protein interactions: CAPRI 2009". *Proteins: Structure, Function, and Bioinformatics* 78.15 (2010), pp. 3073–3084. DOI: 10.1002/prot.22818.

[260] Marc F. Lensink and Shoshana J. Wodak. "Docking, scoring, and affinity prediction in CAPRI". *Proteins: Structure, Function, and Bioinformatics* 81.12 (2013), pp. 2082–2095. DOI: 10.1002/prot.24428.

[261] Marc F Lensink, Iain H Moal, Paul A Bates, Panagiotis L Kastritis, Adrien SJ Melquiond, Ezgi Karaca, Christophe Schmitz, Marc Dijk, Alexandre MJJ Bonvin, Miriam Eisenstein, et al. "Blind prediction of interfacial water positions in CAPRI". *Proteins: Structure, Function, and Bioinformatics* 82.4 (2014), pp. 620–632. DOI: 10.1002/prot.24439.

[262] Marc F Lensink, Sameer Velankar, Andriy Kryshtafovych, Shen-You Huang, Dina Schneidman-Duhovny, Andrej Sali, Joan Segura, Narcis Fernandez-Fuentes, Shruthi Viswanath, Ron Elber, et al. "Prediction of homo-and hetero-protein complexes by protein docking and template-based modeling: a CASP-CAPRI experiment". *Proteins: Structure, Function, and Bioinformatics* (2016), In press. DOI: 10.1002/prot.25007.

[263] John D Leonard and Geeta J Narlikar. "A nucleotide-driven switch regulates flanking DNA length sensing by a dimeric chromatin remodeler". *Molecular Cell* 57.5 (2015), pp. 850–859. DOI: 10.1016/j.molcel.2015.01.008.

[264] Justin W Leung, Poonam Agarwal, Marella D Canny, Fade Gong, Aaron D Robison, Ilya J Finkelstein, Daniel Durocher, and Kyle M Miller. "Nucleosome acidic patch promotes RNF168-and RING1B/BMI1-dependent H2AX and H2A ubiquitination and DNA damage signaling". *PLOS Genetics* 10.3 (2014), e1004178. DOI: 10.1371/journal.pgen.1004178.

[265] Yaakov Levy and José N Onuchic. "Water mediation in protein folding and molecular recognition". *Annual Review of Biophysics and Biomolecular Structure* 35 (2006), pp. 389–415. DOI: 10.1146/annurev.biophys.35.040405.10213.

[266] Hui Li, Andrew D. Robertson, and Jan H. Jensen. "Very fast empirical prediction and rationalization of protein pKa values". *Proteins: Structure, Function, and Bioinformatics* 61.4 (2005), pp. 704–721. DOI: 10.1002/prot.20660.

[267] Lin Li, Dachuan Guo, Yangyu Huang, Shiyong Liu, and Yi Xiao. "ASPDock: protein-protein docking algorithm using atomic solvation parameters model". *BMC Bioinformatics* 12.36 (2011). DOI: 10.1186/1471-2105-12-36.

[268] Siming Li, Christopher M Armstrong, Nicolas Bertin, Hui Ge, Stuart Milstein, Mike Boxem, Pierre-Olivier Vidalain, Jing-Dong J Han, Alban Chesneau, Tong Hao, et al. "A map of the interactome network of the metazoan C. elegans". *Science* 303.5657 (2004), pp. 540–543. DOI: 10.1126/science.1091403.

[269] Weizhong Li, Andrew Cowley, Mahmut Uludag, Tamer Gur, Hamish McWilliam, Silvano Squizzato, Young Mi Park, Nicola Buso, and Rodrigo Lopez. "The EMBL-EBI bioinformatics web and programmatic tools framework". *Nucleic Acids Research* 43.W1 (2015), W580–W584. DOI: 10.1093/nar/gkv279.

[270] Yunqi Li, Jian Zhang, David Tai, C Russell Middaugh, Yang Zhang, and Jianwen Fang. "Prots: A fragment based protein thermo-stability potential". *Proteins: Structure, Function, and Bioinformatics* 80.1 (2012), pp. 81–92. DOI: 10.1002/prot.23163.

[271] J.P. Linge and M. Nilges. "Influence of non-bonded parameters on the quality of NMR structures: A new force field for NMR structure calculation". English. *Journal of Biomolecular NMR* 13.1 (1999), pp. 51–59. DOI: 10.1023/A:1008365802830.

[272] Veikko Linko and Hendrik Dietz. "The enabled state of DNA nanotechnology". *Current Opinion in Biotechnology* 24.4 (Aug. 2013), pp. 555–561. DOI: 10.1016/j.copbio.2013.02.001.

[273] Jan Lipfert, Ian S Millett, Sönke Seifert, and Sebastian Doniach. "Sample holder for small-angle X-ray scattering static and flow cell measurements". *Review of Scientific Instruments* 77.4 (2006), p. 046108. DOI: 10.1063/1.2194484.

[274] Haiguang Liu, Richard J Morris, Alexander Hexemer, Scott Grandison, and Peter H Zwart. "Computation of small-angle scattering profiles with three-dimensional Zernike polynomials". *Acta Crystallographica, Section A: Foundations of Crystallography* 68.2 (2012), pp. 278–285. DOI: 10.1107/S010876731104788X.

[275] Nir London, Dana Movshovitz-Attias, and Ora Schueler-Furman. "The structural basis of peptide-protein binding strategies". *Structure* 18.2 (2010), pp. 188–199. DOI: 10.1016/j.str.2009.11.012.

[276] Nir London, Barak Raveh, Dana Movshovitz-Attias, and Ora Schueler-Furman. "Can self-inhibitory peptides be derived from the interfaces of globular protein–protein interactions?" *Proteins: Structure, Function, and Bioinformatics* 78.15 (2010), pp. 3140–3149. DOI: 10.1002/prot.22785.

[277] Nir London, Barak Raveh, and Ora Schueler-Furman. "Druggable protein-protein interactions – from hot spots to hot segments". *Current Opinion in Chemical Biology* 17.6 (2013), pp. 952–959. DOI: 10.1016/j.cbpa.2013.10.011.

[278] Nir London, Barak Raveh, and Ora Schueler-Furman. "Peptide docking and structure-based characterization of peptide binding: from knowledge to know-how". *Current Opinion in Structural Biology* 23.6 (2013), pp. 894–902. DOI: 10.1016/j.sbi.2013.07.006.

[279] Antoine Loquet, Nikolaos G Sgourakis, Rashmi Gupta, Karin Giller, Dietmar Riedel, Christian Goosmann, Christian Griesinger, Michael Kolbe, David Baker, Stefan Becker, et al. "Atomic model of the type III secretion system needle". *Nature* 486.7402 (2012), pp. 276–279. DOI: 10.1038/nature11079.

[280] Duo Lu and James L Keck. "Structural basis of Escherichia coli single-stranded DNA-binding protein stimulation of exonuclease I". *Proceedings of the National Academy of Sciences* 105.27 (2008), pp. 9169–9174. DOI: 10.1146/annurev.bi.63.070194.002523.

[281] Mingyang Lu, Athanasios D Dousis, and Jianpeng Ma. "OPUS-PSP: an orientation-dependent statistical all-atom potential derived from side-chain packing". *Journal of Molecular Biology* 376.1 (2008), pp. 288–301. DOI: 10.1016/j.jmb.2007.11.033.

[282] Johanna Ludwigsen, Henrike Klinker, and Felix Mueller-Planitz. "No need for a power stroke in ISWI-mediated nucleosome sliding". *EMBO reports* 14.12 (2013), pp. 1092–1097. DOI: 10.1038/embor.2013.160.

[283] Manuel P. Luitz and Martin Zacharias. "Protein-ligand docking using Hamiltonian Replica Exchange simulations with soft core potentials". *Journal of Chemical Information and Modeling* 54.6 (2014), pp. 1669–1675. DOI: 10.1021/ci500296f.

[284] Manuel Luitz, Rainer Bomblies, Katja Ostermeir, and Martin Zacharias. "Exploring biomolecular dynamics and interactions using advanced sampling methods". *Journal of Physics: Condensed Matter* 27.32 (2015), p. 323101.

[285] Sergey Lyskov and Jeffrey J. Gray. "The RosettaDock server for local protein–protein docking". *Nucleic Acids Research* 36.suppl 2 (2008), W233–W238. DOI: 10.1093/nar/gkn216.

[286] Artem Y Lyubimov, Thomas D Murray, Antoine Koehl, Ismail Emre Araci, Monarin Uervirojnangkoorn, Oliver B Zeldin, Aina E Cohen, S Michael Soltis, Elizabeth L Baxter, Aaron S Brewster, et al. "Capture and X-ray diffraction studies of protein microcrystals in a microfluidic trap array". *Acta Crystallographica Section D: Biological Crystallography* 71.4 (2015), pp. 928–940. DOI: 10.1107/S1399004715002308.

[287] Gary Macindoe, Lazaros Mavridis, Vishwesh Venkatraman, Marie-Dominique Devignes, and David W Ritchie. "HexServer: an FFT-based protein docking server powered by graphics processors". *Nucleic Acids Research* 38.suppl 2 (2010), W445–W449. DOI: 10.1093/nar/gkq311.

[288] Andi Mainz, Tomasz L Religa, Remco Sprangers, Rasmus Linser, Lewis E Kay, and Bernd Reif. "NMR spectroscopy of soluble protein complexes at one mega-dalton and beyond". *Angewandte Chemie International Edition* 52.33 (2013), pp. 8746–8751. DOI: 10.1002/anie.201301215.

[289] Karolina A Majorek, Przemyslaw J Porebski, Arjun Dayal, Matthew D Zimmerman, Kamila Jablonska, Alan J Stewart, Maksymilian Chruszcz, and Wladek Minor. "Structural and immunologic characterization of bovine, horse, and rabbit serum albumins". *Molecular Immunology* 52.3 (2012), pp. 174–182. DOI: 10.1016/j.molimm.2012.05.011.

[290] Daniel J Mandell, Evangelos A Coutsias, and Tanja Kortemme. "Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling". *Nature Methods* 6.8 (2009), pp. 551–552. DOI: 10.1038/nmeth0809-551.

[291] Binchen Mao, Roberto Tejero, David Baker, and Gaetano T Montelione. "Protein NMR structures refined with Rosetta have higher accuracy relative to corresponding X-ray crystal structures". *Journal of the American Chemical Society* 136.5 (2014), pp. 1893–1906. DOI: 10.1021/ja409845w.

[292] Joseph A Marsh and Sarah A Teichmann. "Structure, dynamics, assembly, and evolution of protein complexes". *Annual Review of Biochemistry* 84 (2015), pp. 551–575. DOI: 10.1146/annurev-biochem-060614-034142.

[293] Efrat Mashiach, Ruth Nussinov, and Haim J. Wolfson. "FiberDock: a web server for flexible induced-fit backbone refinement in molecular docking". *Nucleic Acids Research* 38.suppl 2 (2010), W457–W461. DOI: 10.1093/nar/gkq373.

[294] Efrat Mashiach, Ruth Nussinov, and Haim J Wolfson. "FiberDock: Flexible induced-fit backbone refinement in molecular docking". *Proteins: Structure, Function, and Bioinformatics* 78.6 (2010), pp. 1503–1519. DOI: 10.1002/prot.22668.

[295] Julien Maupetit, Philippe Derreumaux, and Pierre Tufféry. "A fast method for large-scale De Novo peptide and miniprotein structure prediction". *Journal of Computational Chemistry* 31.4 (2010), pp. 726–738. DOI: 10.1002/jcc.21365.

[296] Andreas May and Martin Zacharias. "Accounting for global protein deformability during protein–protein and protein–ligand docking". *Biochimica and Biophysica Acta, Proteins Proteomics* 1754.1–2 (2005), pp. 225–231. DOI: 10.1016/j.bbapap.2005.07.045.

[297] Andreas May and Martin Zacharias. "Energy minimization in low-frequency normal modes to efficiently allow for global flexibility during systematic protein–protein docking". *Proteins: Structure, Function, and Bioinformatics* 70.3 (2008), pp. 794–809. DOI: 10.1002/prot.21579.

[298] Andreas May and Martin Zacharias. "Protein-ligand docking accounting for receptor side chain and global flexibility in normal modes: evaluation on kinase inhibitor cross docking". *Journal of Medicinal Chemistry* 51.12 (2008). PMID: 18517186, pp. 3499–3506. DOI: 10.1021/jm800071v.

[299] Andreas May and Martin Zacharias. "Protein–protein docking in CAPRI using ATTRACT to account for global and local flexibility". *Proteins: Structure, Function, and Bioinformatics* 69.4 (2007), pp. 774–780. DOI: 10.1002/prot.21735.

[300] Robert K McGinty, Ryan C Henrici, and Song Tan. "Crystal structure of the PRC1 ubiquitylation module bound to the nucleosome". *Nature* 514.7524 (2014), pp. 591–596. DOI: 10.1038/nature13890.

[301] Hamish McWilliam, Weizhong Li, Mahmut Uludag, Silvano Squizzato, Young Mi Park, Nicola Buso, Andrew Peter Cowley, and Rodrigo Lopez. "Analysis tool web services from the EMBL-EBI". *Nucleic Acids Research* 41.W1 (2013), W597–W600. DOI: 10.1093/nar/gkt376.

[302] Adrien S.J. Melquiond, Ezgi Karaca, Panagiotis L. Kastritis, and Alexandre M.J.J. Bonvin. "Next challenges in protein–protein docking: from proteome to interactome and beyond". *Wiley Interdisciplinary Reviews: Computational Molecular Science* 2.4 (2012), pp. 642–651. DOI: 10.1002/wcms.91.

[303] Raúl Méndez, Raphaël Leplae, Marc F. Lensink, and Shoshana J. Wodak. "Assessment of CAPRI predictions in rounds 3–5 shows progress in docking procedures". *Proteins: Structure, Function, and Bioinformatics* 60.2 (2005), pp. 150–169. DOI: 10.1002/prot.20551.

[304] Xuan-Yu Meng, Hong-Xing Zhang, Mihaly Mezei, and Meng Cui. "Protein-Protein Docking Using a Brownian Dynamics Simulations Approach". *Biophysical Journal* 98.3 (2010), 455a. DOI: 10.1016/j.bpj.2009.12.2474.

[305] Haydyn DT Mertens and Dmitri I Svergun. "Structural characterization of proteins and complexes using small-angle X-ray solution scattering". *Journal of Structural Biology* 172.1 (2010), pp. 128–141. DOI: 10.1016/j.jsb.2010.06.012.

[306] Julian Mintseris, Kevin Wiehe, Brian Pierce, Robert Anderson, Rong Chen, Joël Janin, and Zhiping Weng. "Protein–protein docking benchmark 2.0: an update". *Proteins: Structure, Function, and Bioinformatics* 60.2 (2005), pp. 214–216. DOI: 10.1002/prot.20560.

[307] Yasuyuki Miyazaki, Rossitza N Irobalieva, Blanton S Tolbert, Adjoa Smalls-Mantey, Kilali Iyalla, Kelsey Loeliger, Victoria D'Souza, Htet Khant, Michael F Schmid, Eric L Garcia, et al. "Structure of a conserved retroviral RNA packaging element by NMR spectroscopy and cryo-electron tomography". *Journal of Molecular Biology* 404.5 (2010), pp. 751–772. DOI: 10.1016/j.jmb.2010.09.009.

[308] Sanzo Miyazawa and Robert L Jernigan. "Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation". *Macromolecules* 18.3 (1985), pp. 534–552. DOI: 10.1021/ma00145a039.

[309] Sanzo Miyazawa and Robert L Jernigan. "Residue–residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading". *Journal of Molecular Biology* 256.3 (1996), pp. 623–644. DOI: 10.1006/jmbi.1996.0114.

[310] Iain H Moal and Paul A Bates. "SwarmDock and the use of normal modes in protein-protein docking". *International Journal of Molecular Sciences* 11.10 (2010), pp. 3623–3648. DOI: 10.3390/ijms11103623.

[311] Iain H Moal and Juan Fernandez-Recio. "Intermolecular contact potentials for protein–protein interactions extracted from binding free energy changes upon mutation". *Journal of Chemical Theory and Computation* 9.8 (2013), pp. 3715–3727. DOI: 10.1021/ct400295z.

[312] Jacque Monod, Jeffries Wyman, and Jean-Pierre Changeux. "On the nature of allosteric transitions: a plausible model". *Journal of Molecular Biology* 12.1 (1965), pp. 88–118. DOI: 10.1016/S0022-2836(65)80285-6.

[313] Wijnand Mooij and Marcel L Verdonk. "General and targeted statistical potentials for protein–ligand interactions". *Proteins: Structure, Function, and Bioinformatics* 61.2 (2005), pp. 272–287. DOI: 10.1002/prot.20588.

[314] Roberto Mosca, Arnaud Céol, and Patrick Aloy. "Interactome3D: adding structural details to protein networks". *Nature Methods* 10.1 (2013), pp. 47–53. DOI: 10.1038/nmeth.2289.

[315] Felix Mueller-Planitz, Henrike Klinker, Johanna Ludwigsen, and Peter B Becker. "The ATPase domain of ISWI is an autonomous nucleosome remodeling machine". *Nature Structural & Molecular Biology* 20.1 (2013), pp. 82–89. DOI: 10.1038/nsmb.2457.

[316] Srayanta Mukherjee and Yang Zhang. "Protein-protein complex structure predictions by multimeric threading and template recombination". *Structure* 19.7 (2011), pp. 955–966. DOI: 10.1016/j.str.2011.04.006.

[317] Asher Mullard. "Protein–protein interaction inhibitors get into the groove". *Nature Reviews Drug Discovery* 11.3 (2012), pp. 173–175. DOI: 10.1038/nrd3680.

[318] Efstratios Mylonas and Dmitri I Svergun. "Accuracy of molecular mass determination of proteins in solution by small-angle X-ray scattering". *Journal of Applied Crystallography* 40.s1 (2007), pp. 245–249. DOI: 10.1107/S002188980700252X.

[319] Fred Naider and Jacob Anglister. "Peptides in the treatment of AIDS". *Current Opinion in Structural Biology* 19.4 (2009), pp. 473–482. DOI: 10.1016/j.sbi.2009.07.003.

[320] Mariana de Napoles, Jacqueline E Mermoud, Rika Wakao, Y Amy Tang, Mitusuhiro Endoh, Ruth Appanah, Tatyana B Nesterova, Jose Silva, Arie P Otte, Miguel Vidal, et al. "Polycomb group proteins Ring1A/B link ubiquitylation of histone H2A to heritable gene silencing and X inactivation". *Developmental Cell* 7.5 (2004), pp. 663–676. DOI: 10.1016/j.devcel.2004.10.005.

[321] Richard Neutze, Gisela Brändén, and Gebhard FX Schertler. "Membrane protein structural biology using X-ray free electron lasers". *Current Opinion in Structural Biology* 33 (2015), pp. 115–125. DOI: 10.1016/j.sbi.2015.08.006.

[322] Michael Nilges. "A calculation strategy for the structure determination of symmetric demers by 1H NMR". *Proteins: Structure, Function, and Bioinformatics* 17.3 (1993), pp. 297–309. DOI: 10.1002/prot.340170307.

[323] Jerome Nilmeier, Lan Hua, Evangelos A Coutsias, and Matthew P Jacobson. "Assessing protein loop flexibility by hierarchical Monte Carlo sampling". *Journal of Chemical Theory and Computation* 7.5 (2011), pp. 1564–1574. DOI: 10.1021/ct1006696.

[324] Noriyuki Nishimura, Kenichi Hitomi, Andrew S Arvai, Robert P Rambo, Chiharu Hitomi, Sean R Cutler, Julian I Schroeder, and Elizabeth D Getzoff. "Structural mechanism of abscisic acid binding and signaling by dimeric PYR1". *Science* 326.5958 (2009), pp. 1373–1379. DOI: 10.1126/science.1181829.

[325] Masha Y Niv and Harel Weinstein. "A flexible docking procedure for the exploration of peptide binding selectivity to known structures and homology models of PDZ domains". *Journal of the American Chemical Society* 127.40 (2005), pp. 14072–14079. DOI: 10.1021/ja054195s.

[326] Masahito Ohue, Takehiro Shimoda, Shuji Suzuki, Yuri Matsuzaki, Takashi Ishida, and Yutaka Akiyama. "MEGADOCK 4.0: an ultra–high-performance protein–protein docking software for heterogeneous supercomputers". *Bioinformatics* 30.22 (2014), pp. 3281–3283. DOI: 10.1093/bioinformatics/btu532.

[327] Ojore Benedict Valentine Oka, Marie Anne Pringle, Isabel Myriam Schopp, Ineke Braakman, and Neil John Bulleid. "ERdj5 is the ER reductase that catalyzes the removal of non-native disulfides and correct folding of the LDL receptor". *Molecular Cell* 50.6 (2013), pp. 793–804. DOI: 10.1016/j.molcel.2013.05.014.

[328] Tomasz Oliwa and Yang Shen. "cNMA: a framework of encounter complex-based normal mode analysis to model conformational changes in protein interactions". *Bioinformatics* 31.12 (2015), pp. i151–i160. DOI: 10.1093/bioinformatics/btv252.

[329] Mark A Olson, Michael Feig, and Charles L Brooks. "Prediction of protein loop conformations using multiscale modeling methods with physical energy scoring functions". *Journal of Computational Chemistry* 29.5 (2008), pp. 820–831. DOI: 10.1002/jcc.20827.

[330] Sergey Ovchinnikov, Hetunandan Kamisetty, and David Baker. "Robust and accurate prediction of residue–residue interactions across protein interfaces using evolutionary information". *Elife* 3 (2014), e02030. DOI: 10.7554/eLife.02030.

[331] Sergey Ovchinnikov, Lisa Kinch, Hahnbeom Park, Yuxing Liao, Jimin Pei, David E Kim, Hetunandan Kamisetty, Nick V Grishin, and David Baker. "Large-scale determination of previously unsolved protein structures using evolutionary information". *Elife* 4 (2015), e09248. DOI: 10.7554/eLife.09248.

[332] P Nuno Palma, Ludwig Krippahl, John E Wampler, and José JG Moura. "BiGGER: a new (soft) docking algorithm for predicting protein interactions". *Proteins: Structure, Function, and Bioinformatics* 39.4 (2000), pp. 372–384. DOI: 10.1002/(SICI)1097-0134(20000601)39:4⟨372::AID-PROT100⟩3.0.CO;2-Q.

[333] Arun Prasad Pandurangan, Daven Vasishtan, Frank Alber, and Maya Topf. "γ-TEMPy: simultaneous fitting of components in 3D-EM maps of their assembly using a genetic algorithm". *Structure* 23.12 (2015), pp. 2365–2376. DOI: 10.1016/j.str.2015.10.013.

[334] Hardik I Parikh and Glen E Kellogg. "Intuitive, but not simple: Including explicit water molecules in protein–protein docking simulations improves model quality". *Proteins: Structure, Function, and Bioinformatics* 82.6 (2014), pp. 916–932. DOI: 10.1002/prot.24466.

[335] Hahnbeom Park, Gyu Rie Lee, Lim Heo, and Chaok Seok. "Protein loop modeling using a new hybrid energy function and its application to modeling in inaccurate structural environments". *PLOS One* 9.11 (2014), e113811. DOI: 10.1371/journal.pone.0113811.

[336] Sanghyun Park, Jaydeep P Bardhan, Benoıt Roux, and Lee Makowski. "Simulated X-ray scattering of protein solutions using explicit-solvent models". *Journal of Chemical Physics* 130.13 (2009), p. 134114. DOI: 10.1063/1.3099611.

[337] Ilias Patmanidis and Nicholas M. Glykos. "As good as it gets? Folding molecular dynamics simulations of the LytA choline-binding peptide result to an exceptionally accurate model of the peptide structure". *Journal of Molecular Graphics and Modelling* 41 (2013), pp. 68–71. DOI: 10.1016/j.jmgm.2013.02.004.

[338] Maxim V Petoukhov, Daniel Franke, Alexander V Shkumatov, Giancarlo Tria, Alexey G Kikhney, Michal Gajda, Christian Gorba, Haydyn DT Mertens, Petr V Konarev, and Dmitri I Svergun. "New developments in the ATSAS program package for small-angle scattering data analysis". *Journal of Applied Crystallography* 45.2 (2012), pp. 342–350. DOI: 10.1107/S0021889812007662.

[339] Maxim V Petoukhov, Peter V Konarev, Alexey G Kikhney, and Dmitri I Svergun. "ATSAS 2.1-towards automated and web-supported small-angle scattering data analysis". *Journal of Applied Crystallography* 40.s1 (2007), s223–s228. DOI: 10.1107/S0021889807002853.

[340] Maxim V Petoukhov and Dmitri I Svergun. "Applications of small-angle X-ray scattering to biomacromolecular solutions". *International Journal of Biochemistry & Cell Biology* 45.2 (2013), pp. 429–437. DOI: 10.1016/j.biocel.2012.10.017.

[341] Maxim V Petoukhov and Dmitri I Svergun. "Global rigid body modeling of macromolecular complexes against small-angle scattering data". *Biophysical Journal* 89.2 (2005), pp. 1237–1250. DOI: 10.1529/biophysj.105.064154.

[342] Paula Petrone and Vijay S Pande. "Can conformational change be described by only a few normal modes?" *Biophysical Journal* 90.5 (2006), pp. 1583–1593. DOI: 10.1529/biophysj.105.070045.

[343] Evangelia Petsalaki and Robert B Russell. "Peptide-mediated interactions in biological systems: new discoveries and applications". *Current Opinion in Biotechnology* 19.4 (2008), pp. 344–350. DOI: 10.1016/j.copbio.2008.06.004.

[344] Evangelia Petsalaki, Alexander Stark, Eduardo Garcıa-Urdiales, and Robert B Russell. "Accurate prediction of peptide binding sites on protein surfaces". *PLOS Computational Biology* 5.3 (2009), e1000335. DOI: 10.1371/journal.pcbi.1000335.

[345] Christine Petzold, Aimee H Marceau, Katherine H Miller, Susan Marqusee, and James L Keck. "Interaction with single-stranded DNA-binding protein stimulates Escherichia coli Ribonuclease HI enzymatic activity". *Journal of Biological Chemistry* 290.23 (2015), pp. 14626–14636. DOI: 10.1074/jbc.M115.655134.

[346] Brian G Pierce, Yuichiro Hourai, and Zhiping Weng. "Accelerating protein docking in ZDOCK using an advanced 3D convolution library". *PLOS One* 6.9 (2011), e24657. DOI: 10.1371/journal.pone.0024657.

[347] Brian G. Pierce, Kevin Wiehe, Howook Hwang, Bong-Hyun Kim, Thom Vreven, and Zhiping Weng. "ZDOCK server: interactive docking prediction of protein-protein complexes and symmetric multimers". *Bioinformatics* 30.12 (2014), pp. 1771–1773. DOI: 10.1093/bioinformatics/btu097.

[348] Brian Pierce, Weiwei Tong, and Zhiping Weng. "M-ZDOCK: a grid-based approach for C_n symmetric multimer docking". *Bioinformatics* 21.8 (2005), pp. 1472–1478. DOI: 10.1093/bioinformatics/bti229.

[349] Brian Pierce and Zhiping Weng. "ZRANK: reranking protein docking predictions with an optimized energy function". *Proteins: Structure, Function, and Bioinformatics* 67.4 (2007), pp. 1078–1086. DOI: 10.1002/prot.21373.

[350] Nikos Pinotsis, Stephan Lange, Jean-Claude Perriard, Dmitri I Svergun, and Matthias Wilmanns. "Molecular basis of the C-terminal tail-to-tail assembly of the sarcomeric filament protein myomesin". *The EMBO Journal* 27.1 (2008), pp. 253–264. DOI: 10.1038/sj.emboj.7601944.

[351] Frédéric Poitevin, Henri Orland, Sebastian Doniach, Patrice Koehl, and Marc Delarue. "AquaSAXS: a web server for computation and fitting of SAXS profiles with non-uniformally hydrated atomic models". *Nucleic Acids Research* 39.suppl 2 (2011), W184–W189. DOI: 10.1093/nar/gkr430.

[352] Argyris Politis, Carla Schmidt, Elina Tjioe, Alan M Sandercock, Keren Lasker, Yuliya Gordiyenko, Daniel Russel, Andrej Sali, and Carol V Robinson. "Topological models of heteromeric protein assemblies from mass spectrometry: application to the yeast eIF3: eIF5 complex". *Chemistry & Biology* 22.1 (2015), pp. 117–128. DOI: 10.1016/j.chembiol.2014.11.010.

[353] Argyris Politis, Florian Stengel, Zoe Hall, Helena Hernández, Alexander Leitner, Thomas Walzthoeni, Carol V Robinson, and Ruedi Aebersold. "A mass spectrometry-based hybrid method for structural modeling of protein complexes". *Nature Methods* 11.4 (2014), pp. 403–406. DOI: 10.1038/nmeth.2841.

[354] Carles Pons, Marco D'Abramo, Dmitri I. Svergun, Modesto Orozco, Pau Bernadó, and Juan Fernández-Recio. "Structural Characterization of Protein–Protein Complexes by Integrating Computational Docking with Small-angle Scattering Data". *Journal of Molecular Biology* 403.2 (2010), pp. 217–230. DOI: 10.1016/j.jmb.2010.08.029.

[355] Filippo Prischi, Petr V Konarev, Clara Iannuzzi, Chiara Pastore, Salvatore Adinolfi, Stephen R Martin, Dmitri I Svergun, and Annalisa Pastore. "Structural bases for the interaction of frataxin with the central components of iron-sulphur cluster assembly". *Nature Communications* 1.95 (2010). DOI: 10.1038/ncomms1097.

[356] Sander Pronk, Szilárd Páll, Roland Schulz, Per Larsson, Pär Bjelkmar, Rossen Apostolov, Michael R Shirts, Jeremy C Smith, Peter M Kasson, David van der Spoel, et al. "GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit". *Bioinformatics* 29.7 (2013), pp. 845–854. DOI: 10.1093/bioinformatics/btt055.

[357] Lisa R Racki, Nariman Naber, Ed Pate, John D Leonard, Roger Cooke, and Geeta J Narlikar. "The histone H4 tail regulates the conformation of the ATP-binding pocket in the SNF2h chromatin remodeling enzyme". *Journal of Molecular Biology* 426.10 (2014), pp. 2034–2044. DOI: 10.1016/j.jmb.2014.02.021.

[358] Savarimuthu Baskar Raj, Subramanian Ramaswamy, and Bryce V Plapp. "Yeast alcohol dehydrogenase structure and catalysis". *Biochemistry* 53.36 (2014), pp. 5791–5803. DOI: 10.1021/bi5006442.

[359] Erney Ramırez-Aportela, José Ramón López-Blanco, and Pablo Chacón. "FRODOCK 2.0: fast protein–protein docking server". *Bioinformatics* 32.15 (2016), pp. 2386–2388. DOI: 10.1093/bioinformatics/btw141.

[360] Juri Rappsilber. "The beginning of a beautiful friendship: cross-linking/mass spectrometry and modelling of proteins and multi-protein complexes". *Journal of Structural Biology* 173.3 (2011), pp. 530–540. DOI: 10.1016/j.jsb.2010.10.014.

[361] Barak Raveh, Nir London, and Ora Schueler-Furman. "Sub-angstrom modeling of complexes between flexible peptides and globular proteins". *Proteins: Structure, Function, and Bioinformatics* 78.9 (2010), pp. 2029–2040. DOI: 10.1002/prot.22716.

[362] Barak Raveh, Nir London, Lior Zimmerman, and Ora Schueler-Furman. "Rosetta FlexPepDock ab-initio: simultaneous folding, docking and refinement of peptides onto their receptors". *PLOS One* 6.4 (2011), e18934. DOI: 10.1371/journal.pone.0018934.

[363] Krishnakumar M Ravikumar, Wei Huang, and Sichun Yang. "Fast-SAXS-pro: a unified approach to computing SAXS profiles of DNA, RNA, protein, and their complexes". *Journal of Chemical Physics* 138.2 (2013), p. 024112. DOI: 10.1063/1.4774148.

[364] David W Ritchie, Dima Kozakov, and Sandor Vajda. "Accelerating and focusing protein–protein docking correlations using multi-dimensional rotational FFT generating functions". *Bioinformatics* 24.17 (2008), pp. 1865–1873. DOI: 10.1093/bioinformatics/btn334.

[365] David W Ritchie and Vishwesh Venkatraman. "Ultra-fast FFT protein docking on graphics processors". *Bioinformatics* 26.19 (2010), pp. 2398–2405. DOI: 10.1093/bioinformatics/btq444.

[366] Philip J Robinson, Michael J Trnka, Riccardo Pellarin, Charles H Greenberg, David A Bushnell, Ralph Davis, Alma L Burlingame, Andrej Sali, and Roger D Kornberg. "Molecular architecture of the yeast Mediator complex". *eLife* 4 (2015), e08719. DOI: 10.7554/eLife.08719.

[367] Natacha Rochel, Fabrice Ciesielski, Julien Godet, Edelmiro Moman, Manfred Roessle, Carole Peluso-Iltis, Martine Moulin, Michael Haertlein, Phil Callow, Yves Mély, et al. "Common architecture of nuclear receptor heterodimers on DNA direct repeat elements with different spacings". *Nature Structural & Molecular Biology* 18.5 (2011), pp. 564–570. DOI: 10.1038/nsmb.2054.

[368] João P. G. L. M. Rodrigues, Mikaël Trellet, Christophe Schmitz, Panagiotis Kastritis, Ezgi Karaca, Adrien S. J. Melquiond, and Alexandre M. J. J. Bonvin. "Clustering biomolecular complexes by residue contacts similarity". *Proteins: Structure, Function, and Bioinformatics* 80.7 (2012), pp. 1810–1817. DOI: 10.1002/prot.24078.

[369] Thomas Rolland, Murat Taşan, Benoit Charloteaux, Samuel J Pevzner, Quan Zhong, Nidhi Sahni, Song Yi, Irma Lemmens, Celia Fontanillo, Roberto Mosca, et al. "A proteome-scale map of the human interactome network". *Cell* 159.5 (2014), pp. 1212–1226. DOI: 10.1016/j.cell.2014.10.050.

[370] Rakefet Rosenfeld, Qiang Zheng, Sandor Vajda, and Charles DeLisi. "Flexible docking of peptides to class I major-histocompatibility-complex receptors". *Genetic Analysis: Biomolecular Engineering* 12.1 (1995), pp. 1–21. DOI: 10.1016/1050-3862(95)00107-7.

[371] Paolo Rossi, Lei Shi, Gaohua Liu, Christopher M Barbieri, Hsiau-Wei Lee, Thomas D Grant, Joseph R Luft, Rong Xiao, Thomas B Acton, Edward H Snell, et al. "A hybrid NMR/SAXS-based approach for discriminating oligomeric protein interfaces using Rosetta". *Proteins: Structure, Function, and Bioinformatics* 83.2 (2015), pp. 309–317. DOI: 10.1002/prot.24719.

[372] AR Round, D Franke, S Moritz, R Huchler, M Fritsche, D Malthan, R Klaering, DI Svergun, and M Roessle. "Automated sample-changing robot for solution scattering experiments at the EMBL Hamburg SAXS station X33". *Journal of Applied Crystallography* 41.5 (2008), pp. 913–917. DOI: 10.1107/S0021889808021018.

[373] Ambrish Roy, Alper Kucukural, and Yang Zhang. "I-TASSER: a unified platform for automated protein structure and function prediction". *Nature Protocols* 5.4 (2010), pp. 725–738. DOI: 10.1038/nprot.2010.5.

[374] Bartosz Różycki, Young C Kim, and Gerhard Hummer. "SAXS ensemble refinement of ESCRT-III CHMP3 conformational transitions". *Structure* 19.1 (2011), pp. 109–116. DOI: 10.1016/j.str.2010.10.006.

[375] Mor Rubinstein and Masha Y Niv. "Peptidic modulators of protein-protein interactions: Progress and challenges in computational design". *Biopolymers* 91.7 (2009), pp. 505–513. DOI: 10.1002/bip.21164.

[376] Manuel Rueda, Giovanni Bottegoni, and Ruben Abagyan. "Consistent improvement of cross-docking results using binding site ensembles generated with elastic network normal modes". *Journal of Chemical Information and Modeling* 49.3 (2009), pp. 716–725. DOI: 10.1021/ci8003732.

[377] Helmut Ruska, Bodo v Borries, and Ernst Ruska. "Die Bedeutung der übermikroskopie für die Virusforschung". *Archiv für die gesamte Virusforschung* 1.1 (1939), pp. 155–169.

[378] Daniel Russel, Keren Lasker, Ben Webb, Javier Velázquez-Muriel, Elina Tjioe, Dina Schneidman-Duhovny, Bret Peterson, and Andrej Sali. "Putting the pieces together: integrative modeling platform software for structure determination of macromolecular assemblies". *PLOS Biology* 10.1 (2012), e1001244. DOI: 10.1371/journal.pbio.1001244.

[379] Daniel P Ryan and Jacqueline M Matthews. "Protein–protein interactions in human disease". *Current Opinion in Structural Biology* 15.4 (2005), pp. 441–446. DOI: 10.1016/j.sbi.2005.06.001.

[380] Mikhail Ryzhikov, Olga Koroleva, Dmitri Postnov, Andrew Tran, and Sergey Korolev. "Mechanism of RecO recruitment to DNA by single-stranded DNA binding protein". *Nucleic Acids Research* 39.14 (2011), pp. 6305–6314. DOI: 10.1093/nar/gkr199.

[381] Sophie Sacquin-Mora, Alessandra Carbone, and Richard Lavery. "Identification of protein interaction partners and protein-protein interaction sites". *Journal of Molecular Biology* 382.5 (2008), pp. 1276–1289. DOI: 10.1016/j.jmb.2008.08.002.

[382] Anjanabha Saha, Jacqueline Wittmeyer, and Bradley R Cairns. "Chromatin remodeling through directional DNA translocation from an internal nucleosomal site". *Nature Structural & Molecular Biology* 12.9 (2005), pp. 747–755. DOI: 10.1038/nsmb973.

[383] Adrien Saladin, Sébastien Fiorucci, Pierre Poulain, Chantal Prévost, and Martin Zacharias. "PTools: an opensource molecular docking library". *BMC Structural Biology* 9.1 (2009), p. 27. DOI: 10.1186/1472-6807-9-27.

[384] Adrien Saladin, Julien Rey, Pierre Thévenet, Martin Zacharias, Gautier Moroy, and Pierre Tufféry. "PEP-SiteFinder: a tool for the blind identification of peptide binding sites on protein surfaces". *Nucleic Acids Research* 42.W1 (2014), W221–W226. DOI: 10.1093/nar/gku404.

[385] Andrej Sali, Helen M Berman, Torsten Schwede, Jill Trewhella, Gerard Kleywegt, Stephen K Burley, John Markley, Haruki Nakamura, Paul Adams, Alexandre MJJ Bonvin, et al. "Outcome of the First wwPDB Hybrid/Integrative Methods Task Force Workshop". *Structure* 23.7 (2015), pp. 1156–1167. DOI: 10.1016/j.str.2015.05.013.

[386] Parthasarathy Sampathkumar, Seung Joong Kim, Paula Upla, William J Rice, Jeremy Phillips, Benjamin L Timney, Ursula Pieper, Jeffrey B Bonanno, Javier Fernandez-Martinez, Zhanna Hakhverdyan, et al. "Structure, dynamics, evolution, and function of a major scaffold component in the nuclear pore complex". *Structure* 21.4 (2013), pp. 560–571. DOI: 10.1016/j.str.2013.02.005.

[387] Sergey Samsonov, Joan Teyra, and M Teresa Pisabarro. "A molecular dynamics approach to study the importance of solvent in protein interactions". *Proteins: Structure, Function, and Bioinformatics* 73.2 (2008), pp. 515–525. DOI: 10.1002/prot.22076.

[388] Gilberto Sánchez-González, Jae-Kwan Kim, Deok-Soo Kim, and Ramón Garduño-Juárez. "A beta-complex statistical four body contact potential combined with a hydrogen bond statistical potential recognizes the correct native structure from protein decoy sets". *Proteins: Structure, Function, and Bioinformatics* 81.8 (2013), pp. 1420–1433. DOI: 10.1002/prot.24293.

[389] Alexander Sasse, Sjoerd J de Vries, Christina EM Schindler, Isaure Chauvot de Beauchêne, and Martin Zacharias. "Rapid design of knowledge-based scoring potentials for enrichment of near-native geometries in protein docking with the GRADSCOPT tool kit". *PLOS One* 12.1 (2017), e0170625. DOI: 10.1371/journal.pone.0170625.

[390] Christina EM Schindler, Isaure Chauvot de Beauchêne, Sjoerd J de Vries, and Martin Zacharias. "Protein-protein and peptide-protein docking and refinement using ATTRACT in CAPRI". *Proteins: Structure, Function, and Bioinformatics* (2016), In press. DOI: 10.1002/prot.25196.

[391] Christina EM Schindler, Sjoerd J de Vries, and Martin Zacharias. "Fully blind peptide-protein docking with pepATTRACT". *Structure* 23.8 (2015), pp. 1507–1515. DOI: 10.1016/j.str.2015.05.021.

[392] Christina EM Schindler, Sjoerd J de Vries, Alexander Sasse, and Martin Zacharias. "SAXS data alone can generate high-quality models of protein-protein complexes". *Structure* 24.8 (2016), pp. 1387–1397. DOI: 10.1016/j.str.2016.06.007.

[393] Christina EM Schindler, Sjoerd J de Vries, and Martin Zacharias. "iATTRACT: Simultaneous global and local interface optimization for protein–protein docking refinement". *Proteins: Structure, Function, and Bioinformatics* 83.2 (2015), pp. 248–258. DOI: 10.1002/prot.24728.

[394] Christina EM Schindler and Martin Zacharias. "Application of the ATTRACT coarse-grained docking and atomistic refinement for predicting peptide-protein interactions". Ed. by Ora Schueler-Furman and Nir London. Methods in Molecular Biology. Springer Press, 2016. Chap. 7, In review.

[395] Emmanuelle Schmitt, Michel Panvert, Christine Lazennec-Schurdevin, Pierre-Damien Coureux, Javier Perez, Andrew Thompson, and Yves Mechulam. "Structure of the ternary initiation complex aIF2–GDPNP–methionylated initiator tRNA". *Nature Structural & Molecular biology* 19.4 (2012), pp. 450–454. DOI: 10.1038/nsmb.2259.

[396] Thomas D Schneider and R Michael Stephens. "Sequence logos: a new way to display consensus sequences". *Nucleic Acids Research* 18.20 (1990), pp. 6097–6100. DOI: 10.1093/nar/18.20.6097.

[397] Dina Schneidman-Duhovny, Michal Hammel, and Andrej Sali. "FoXS: a web server for rapid computation and fitting of SAXS profiles". *Nucleic Acids Research* 38.suppl 2 (2010), W540–W544. DOI: 10.1093/nar/gkq461.

[398] Dina Schneidman-Duhovny, Michal Hammel, and Andrej Sali. "Macromolecular docking restrained by a small angle X-ray scattering profile". *Journal of Structural Biology* 173.3 (2011), pp. 461–471. DOI: 10.1016/j.jsb.2010.09.023.

[399] Dina Schneidman-Duhovny, Michal Hammel, John A Tainer, and Andrej Sali. "Accurate SAXS profile computation and its assessment by contrast variation experiments". *Biophysical Journal* 105.4 (2013), pp. 962–974. DOI: 10.1016/j.bpj.2013.07.020.

[400] Dina Schneidman-Duhovny, Yuval Inbar, Ruth Nussinov, and Haim J Wolfson. "Geometry-based flexible and symmetric protein docking". *Proteins: Structure, Function, and Bioinformatics* 60.2 (2005), pp. 224–231. DOI: 10.1002/prot.20562.

[401] Dina Schneidman-Duhovny, Yuval Inbar, Ruth Nussinov, and Haim J. Wolfson. "PatchDock and SymmDock: servers for rigid and symmetric docking". *Nucleic Acids Research* 33.suppl 2 (2005), W363–W367. DOI: 10.1093/nar/gki481.

[402] Dina Schneidman-Duhovny, Riccardo Pellarin, and Andrej Sali. "Uncertainty in integrative structural modeling". *Current Opinion in Structural Biology* 28 (2014), pp. 96–104. DOI: 10.1016/j.sbi.2014.08.001.

[403] Dina Schneidman-Duhovny, Andrea Rossi, Agustin Avila-Sakar, Seung Joong Kim, Javier Velázquez-Muriel, Pavel Strop, Hong Liang, Kristin A Krukenberg, Maofu Liao, Ho Min Kim, et al. "A method for integrative structure determination of protein-protein complexes". *Bioinformatics* 28.24 (2012), pp. 3282–3289. DOI: 10.1093/bioinformatics/bts628.

[404] Schrödinger, LLC. *The PyMOL Molecular Graphics System, Version 1.7r0.* 2014.

[405] Charles D Schwieters, John J Kuszewski, Nico Tjandra, and G Marius Clore. "The Xplor-NIH NMR molecular structure determination package". *Journal of Magnetic Resonance* 160.1 (2003), pp. 65–73. DOI: 10.1016/S1090-7807(02)00014-9.

[406] Piotr Setny, Ranjit Bahadur, and Martin Zacharias. "Protein-DNA docking with a coarse-grained force field". *BMC Bioinformatics* 13.228 (2012). DOI: 10.1186/1471-2105-13-228.

[407] Piotr Setny and Martin Zacharias. "A coarse-grained force field for protein-RNA docking". *Nucleic Acids Research* 39.21 (2011), pp. 9118–9129. DOI: 10.1093/nar/gkr636.

[408] Nikolaos G Sgourakis, Kannan Natarajan, Jinfa Ying, Beat Vogeli, Lisa F Boyd, David H Margulies, and Ad Bax. "The structure of mouse cytomegalovirus m04 protein obtained from sparse NMR data reveals a conserved fold of the m02-m06 viral immune modulator family". *Structure* 22.9 (2014), pp. 1263–1273. DOI: 10.1016/j.str.2014.05.018.

[409] Nikolaos G Sgourakis, Wai-Ming Yau, and Wei Qiang. "Modeling an in-register, parallel "iowa" Aβ fibril structure using solid-state NMR data from labeled samples with Rosetta". *Structure* 23.1 (2015), pp. 216–227. DOI: 10.1016/j.str.2014.10.022.

[410] Amit Sharma, Katherine R Jenkins, Annie Héroux, and Gregory D Bowman. "Crystal structure of the chromodomain helicase DNA-binding protein 1 (Chd1) DNA-binding domain in complex with DNA". *Journal of Biological Chemistry* 286.49 (2011), pp. 42099–42104. DOI: 10.1074/jbc.C111.294462.

[411] Maxim Shatsky, Ruth Nussinov, and Haim J Wolfson. "MultiProt–a multiple protein structural alignment algorithm". *Algorithms in Bioinformatics*. Vol. 2452. Lecture Notes in Computer Science. Springer, 2002, pp. 235–250. DOI: 10.1007/3-540-45784-4{_}18.

[412] Min-yi Shen and Andrej Sali. "Statistical potential for assessment and prediction of protein structures". *Protein Science* 15.11 (2006), pp. 2507–2524. DOI: 10.1110/ps.062416606.

[413] Yimin Shen, Julien Maupetit, Philippe Derreumaux, and Pierre Tuffe ry. "Improved PEP-FOLD approach for peptide and miniprotein structure prediction". *Journal of Chemical Theory and Computation* 10.10 (2014), pp. 4745–4758. DOI: 10.1021/ct500592m.

[414] Yi Shi, Javier Fernandez-Martinez, Elina Tjioe, Riccardo Pellarin, Seung Joong Kim, Rosemary Williams, Dina Schneidman-Duhovny, Andrej Sali, Michael P Rout, and Brian T Chait. "Structural characterization by cross-linking reveals the detailed architecture of a coatomer-related heptameric module from the nuclear pore complex". *Molecular & Cellular Proteomics* 13.11 (2014), pp. 2927–2943. DOI: 10.1074/mcp.M114.041673.

[415] David S Shin, Michael DiDonato, David P Barondeau, Greg L Hura, Chiharu Hitomi, J Andrew Berglund, Elizabeth D Getzoff, S Craig Cary, and John A Tainer. "Superoxide dismutase from the eukaryotic thermophile Alvinella pompejana: structures, stability, mechanism, and insights into amyotrophic lateral sclerosis". *Journal of Molecular Biology* 385.5 (2009), pp. 1534–1555. DOI: 10.1016/j.jmb.2008.11.031.

[416] Brian K Shoichet and Irwin D Kuntz. "Matching chemistry and shape in molecular docking". *Protein Engineering* 6.7 (1993), pp. 723–732. DOI: 10.1093/protein/6.7.723.

[417] Fabian Sievers, Andreas Wilm, David Dineen, Toby J Gibson, Kevin Karplus, Weizhong Li, Rodrigo Lopez, Hamish McWilliam, Michael Remmert, Johannes Söding, et al. "Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega". *Molecular Systems Biology* 7.1 (2011). DOI: 10.1038/msb.2011.75.

[418] Rohita Sinha, Petras J Kundrotas, and Ilya A Vakser. "Docking by structural similarity at protein-protein interfaces". *Proteins: Structure, Function, and Bioinformatics* 78.15 (2010), pp. 3235–3241. DOI: 10.1002/prot.22812.

[419] Aroop Sircar and Jeffrey J Gray. "SnugDock: paratope structural optimization during antibody-antigen docking compensates for errors in antibody homology models". *PLOS Computational Biology* 6.1 (2010), e1000644. DOI: 10.1371/journal.pcbi.1000644.

[420] Devika Sirohi, Zhenguo Chen, Lei Sun, Thomas Klose, Theodore C Pierson, Michael G Rossmann, and Richard J Kuhn. "The 3.8 Å resolution cryo-EM structure of Zika virus". *Science* 352.6284 (2016), pp. 467–470. DOI: 10.1126/science.aaf5316.

[421] Soren Skou, Richard E Gillilan, and Nozomi Ando. "Synchrotron-based small-angle X-ray scattering (SAXS) of proteins in solution". *Nature Protocols* 9.7 (2014), p. 1727. DOI: 10.1038/nprot.2014.116.

[422] Albert Solernou and Juan Fernandez-Recio. "pyDockCG: new coarse-grained potential for protein–protein docking". *The Journal of Physical Chemistry B* 115.19 (2011), pp. 6032–6039. DOI: 10.1021/jp112292b.

[423] Velin Z Spassov, Paul K Flook, and Lisa Yan. "LOOPER: a molecular mechanics-based algorithm for protein loop prediction". *Protein Engineering Design and Selection* 21.2 (2008), pp. 91–100. DOI: 10.1093/protein/gzm083.

[424] Benjamin J Spink, Sivaraj Sivaramakrishnan, Jan Lipfert, Sebastian Doniach, and James A Spudich. "Long single α-helical tail domains bridge the gap between structure and function of myosin VI". *Nature Structural & Molecular Biology* 15.6 (2008), pp. 591–597. DOI: 10.1038/nsmb.1429.

[425] Iskra Staneva and Stefan Wallin. "All-Atom Monte Carlo Approach to Protein–Peptide Binding". *Journal of Molecular Biology* 393.5 (2009), pp. 1118–1128. DOI: 10.1016/j.jmb.2009.08.063.

[426] Amelie Stein and Patrick Aloy. "Novel peptide-mediated interactions derived from high-resolution 3-dimensional structures". *PLOS Computational Biology* 6.5 (2010), e1000789. DOI: 10.1371/journal.pcbi.1000789.

[427] Amelie Stein and Tanja Kortemme. "Improvements to robotics-inspired conformational sampling in Rosetta". *PLOS One* 8.5 (2013), e63090. DOI: 10.1371/journal.pone.0063090.

[428] Ulrich Stelzl, Uwe Worm, Maciej Lalowski, Christian Haenig, Felix H Brembeck, Heike Goehler, Martin Stroedicke, Martina Zenkner, Anke Schoenherr, Susanne Koeppen, et al. "A human protein-protein interaction network: a resource for annotating the proteome". *Cell* 122.6 (2005), pp. 957–968. DOI: 10.1016/j.cell.2005.08.029.

[429] Michael P. H. Stumpf, Thomas Thorne, Eric de Silva, Ronald Stewart, Hyeong Jun An, Michael Lappe, and Carsten Wiuf. "Estimating the size of the human interactome". *Proceedings of the National Academy of Sciences* 105.19 (2008), pp. 6959–6964. DOI: 10.1073/pnas.0708078105.

[430] Karsten Suhre and Yves-Henri Sanejouand. "ElNémo: a normal mode web server for protein movement analysis and the generation of templates for molecular replacement". *Nucleic Acids Research* 32.suppl 2 (2004), W610–W614. DOI: 10.1093/nar/gkh368.

[431] D Svergun, C Barberato, and MHJ Koch. "CRYSOL–a program to evaluate X-ray solution scattering of biological macromolecules from atomic coordinates". *Journal of Applied Crystallography* 28.6 (1995), pp. 768–773. DOI: 10.1107/S0021889895007047.

[432] Andras Szilagyi and Yang Zhang. "Template-based structure modeling of protein–protein interactions". *Current Opinion in Structural Biology* 24 (2014), pp. 10–23. DOI: 10.1016/j.sbi.2013.11.005.

[433] Mohamed Tagari, Richard Newman, Monica Chagoyen, Jose-Maria Carazo, and Kim Henrick. "New electron microscopy database and deposition system". *Trends in Biochemical Sciences* 27.11 (2002), p. 589. DOI: 10.1016/S0968-0004(02)02176-X.

[434] Piyali Guha Thakurta, Debi Choudhury, Rakhi Dasgupta, and JK Dattagupta. "Tertiary structural changes associated with iron binding and release in hen serum transferrin: a crystallographic and spectroscopic study". *Biochemical and Biophysical Research Communications* 316.4 (2004), pp. 1124–1131. DOI: 10.1016/j.bbrc.2004.02.165.

[435] Pierre Thévenet, Yimin Shen, Julien Maupetit, Frédéric Guyon, Philippe Derreumaux, and Pierre Tufféry. "PEP-FOLD: an updated de novo structure prediction server for both linear and disulfide bonded cyclic peptides". *Nucleic Acids Research* 40.W1 (2012), W288–W293. DOI: 10.1093/nar/gks419.

[436] Nicolas H Thomä, Bryan K Czyzewski, Andrei A Alexeev, Alexander V Mazin, Stephen C Kowalczykowski, and Nikola P Pavletich. "Structure of the SWI2/SNF2 chromatin-remodeling domain of eukaryotic Rad54". *Nature Structural & Molecular Biology* 12.4 (2005), pp. 350–356. DOI: 10.1038/nsmb919.

[437] Matthew Z Tien, Dariya K Sydykova, Austin G Meyer, and Claus O Wilke. "PeptideBuilder: a simple Python library to generate model peptides". *PeerJ* 1 (2013), e80. DOI: 10.7717/peerj.80.

[438] Elina Tjioe and William T Heller. "ORNL_SAS: software for calculation of small-angle scattering intensities of proteins and protein complexes". *Journal of Applied Crystallography* 40.4 (2007), pp. 782–785. DOI: 10.1107/S002188980702420X.

[439] Dror Tobi. "Designing coarse grained-and atom based-potentials for protein-protein docking". *BMC Structural Biology* 10.1 (2010), p. 1. DOI: 10.1186/1472-6807-10-40.

[440] Dror Tobi and Ivet Bahar. "Structural changes involved in protein binding correlate with intrinsic motions of proteins in the unbound state". *Proceedings of the National Academy of Sciences of the United States of America* 102.52 (2005), pp. 18908–18913. DOI: 10.1073/pnas.0507603102.

[441] Peter Tompa, Norman E Davey, Toby J Gibson, and M Madan Babu. "A million peptide motifs for the molecular biologist". *Molecular Cell* 55.2 (2014), pp. 161–169. DOI: 10.1016/j.molcel.2014.05.032.

[442] Peter Tompa, Eva Schad, Agnes Tantos, and Lajos Kalmar. "Intrinsically disordered proteins: emerging interaction specialists". *Current Opinion in Structural Biology* 35 (2015), pp. 49–59. DOI: 10.1016/j.sbi.2015.08.009.

[443] Maya Topf, Keren Lasker, Ben Webb, Haim Wolfson, Wah Chiu, and Andrej Sali. "Protein structure fitting and refinement guided by cryo-EM density". *Structure* 16.2 (2008), pp. 295–307. DOI: 10.1016/j.str.2007.11.016.

[444] Mieczyslaw Torchala, Iain H. Moal, Raphael A. G. Chaleil, Juan Fernandez-Recio, and Paul A. Bates. "SwarmDock: a server for flexible protein–protein docking". *Bioinformatics* 29.6 (2013), pp. 807–809. DOI: 10.1093/bioinformatics/btt038.

[445] Maxim Totrov and Ruben Abagyan. "Flexible ligand docking to multiple receptor conformations: a practical alternative". *Current Opinion in Structural Biology* 18.2 (2008), pp. 178–184. DOI: 10.1016/j.sbi.2008.01.004.

[446] Andrey Tovchigrechko and Ilya A. Vakser. "GRAMM-X public web server for protein–protein docking". *Nucleic Acids Research* 34.suppl 2 (2006), W310–W314. DOI: 10.1093/nar/gkl206.

[447] Leonardo G Trabuco, Stefano Lise, Evangelia Petsalaki, and Robert B Russell. "PepSite: prediction of peptide-binding sites from protein surfaces". *Nucleic Acids Research* 40.W1 (2012), W423–W427. DOI: 10.1093/nar/gks398.

[448] Mikael Trellet, Adrien SJ Melquiond, and Alexandre MJJ Bonvin. "A unified conformational selection and induced fit approach to protein–peptide docking". *PLOS One* 8.3 (2013), e58769. DOI: 10.1371/journal.pone.0058769.

[449] Felix Tritschler, Ana Eulalio, Vincent Truffault, Marcus D Hartmann, Sigrun Helms, Steffen Schmidt, Murray Coles, Elisa Izaurralde, and Oliver Weichenrieder. "A divergent Sm fold in EDC3 proteins mediates DCP1 binding and P-body targeting". *Molecular and Cellular Biology* 27.24 (2007), pp. 8600–8611. DOI: 10.1128/MCB.01506-07.

[450] Ivan Tubert-Brohman, Woody Sherman, Matt Repasky, and Thijs Beuming. "Improved docking of polypeptides with Glide". *Journal of Chemical Information and Modeling* 53.7 (2013), pp. 1689–1699. DOI: 10.1021/ci400128m.

[451] Nurcan Tuncbag, Attila Gursoy, Ruth Nussinov, and Ozlem Keskin. "Predicting protein-protein interactions on a proteome scale by matching evolutionary and structural similarities at interfaces using PRISM". *Nature Protocols* 6.9 (2011), pp. 1341–1354. DOI: 10.1038/nprot.2011.367.

[452] Eldon L Ulrich, Hideo Akutsu, Jurgen F Doreleijers, Yoko Harano, Yannis E Ioannidis, Jundong Lin, Miron Livny, Steve Mading, Dimitri Maziuk, Zachary Miller, et al. "BioMagResBank". *Nucleic Acids Research* 36.suppl 1 (2008), pp. D402–D408. DOI: 10.1093/nar/gkm957.

[453] E Besray Unal, Attila Gursoy, and Burak Erman. "VitAL: Viterbi algorithm for de novo peptide design". *PLOS One* 5.6 (2010), e10926. DOI: 10.1371/journal.pone.0010926.

[454] Sandor Vajda, David R Hall, and Dima Kozakov. "Sampling and scoring: A marriage made in heaven". *Proteins: Structure, Function, and Bioinformatics* 81.11 (2013), pp. 1874–1884. DOI: 10.1002/prot.24343.

[455] Erica Valentini, Alexey G Kikhney, Gianpietro Previtali, Cy M Jeffries, and Dmitri I Svergun. "SASBDB, a repository for biological small-angle scattering data". *Nucleic Acids Research* 43 (2015), pp. D357–D363. DOI: 10.1093/nar/gku1047.

[456] Aalt DJ van Dijk and Alexandre MJJ Bonvin. "Solvated docking: introducing water into the modelling of biomolecular complexes". *Bioinformatics* 22.19 (2006), pp. 2340–2347. DOI: 10.1093/bioinformatics/btl395.

[457] Marc van Dijk and Alexandre MJJ Bonvin. "A protein–DNA docking benchmark". *Nucleic Acids Research* 36.14 (2008), e88. DOI: 10.1093/nar/gkn386.

[458] Kim van Roey, Bora Uyar, Robert J Weatheritt, Holger Dinkel, Markus Seiler, Aidan Budd, Toby J Gibson, and Norman E Davey. "Short linear motifs: ubiquitous and functionally diverse protein interaction modules directing cell regulation". *Chemical Reviews* 114.13 (2014), pp. 6733–6778. DOI: 10.1021/cr400585q.

[459] Gydo CP van Zundert, Adrien SJ Melquiond, and Alexandre MJJ Bonvin. "Integrative modeling of biomolecular complexes: HADDOCKing with cryo-electron microscopy data". *Structure* 23.5 (2015), pp. 949–960. DOI: 10.1016/j.str.2015.03.014.

[460] Anna Vangone and Alexandre MJJ Bonvin. "Contacts-based prediction of binding affinity in protein–protein complexes". *eLife* 4 (2015), e07454. DOI: 10.7554/eLife.07454.001.

[461] Peter Vanhee, Almer M van der Sloot, Erik Verschueren, Luis Serrano, Frederic Rousseau, and Joost Schymkowitz. "Computational design of peptide ligands". *Trends in Biotechnology* 29.5 (2011), pp. 231–239. DOI: 10.1016/j.tibtech.2011.01.004.

[462] Peter Vanhee, Francois Stricher, Lies Baeten, Erik Verschueren, Tom Lenaerts, Luis Serrano, Frederic Rousseau, and Joost Schymkowitz. "Protein-peptide interactions adopt the same structural motifs as monomeric protein folds". *Structure* 17.8 (2009), pp. 1128–1136. DOI: 10.1016/j.str.2009.06.013.

[463] Peter Vanhee, Erik Verschueren, Lies Baeten, Francois Stricher, Luis Serrano, Frederic Rousseau, and Joost Schymkowitz. "BriX: a database of protein building blocks for structural analysis, modeling and design". *Nucleic Acids Research* 39.suppl 1 (2011), pp. D435–D442. DOI: 10.1093/nar/gkq972.

[464] Dileep Vasudevan, Eugene YD Chua, and Curt A Davey. "Crystal structures of nucleosome core particles containing the '601' strong positioning sequence". *Journal of Molecular Biology* 403.1 (2010), pp. 1–10. DOI: 10.1016/j.jmb.2010.08.039.

[465] Javier Velázquez-Muriel, Keren Lasker, Daniel Russel, Jeremy Phillips, Benjamin M Webb, Dina Schneidman-Duhovny, and Andrej Sali. "Assembly of macromolecular complexes by satisfaction of spatial restraints from electron microscopy images". *Proceedings of the National Academy of Sciences* 109.46 (2012), pp. 18821–18826. DOI: 10.1073/pnas.1216549109.

[466] Vishwesh Venkatraman and David W. Ritchie. "Flexible protein docking refinement using pose-dependent normal mode analysis". *Proteins: Structure, Function, and Bioinformatics* 80.9 (2012), pp. 2262–2274. DOI: 10.1002/prot.24115.

[467] Vishwesh Venkatraman, Yifeng D Yang, Lee Sael, and Daisuke Kihara. "Protein-protein docking using region-based 3D Zernike descriptors". *BMC bioinformatics* 10.407 (2009). DOI: 10.1186/1471-2105-10-407.

[468] Erik Verschueren, Peter Vanhee, Frederic Rousseau, Joost Schymkowitz, and Luis Serrano. "Protein–peptide complex prediction through fragment interaction patterns". *Structure* 21.5 (2013), pp. 789–797. DOI: 10.1016/j.str.2013.02.023.

[469] Jouko Juhani Virtanen, Lee Makowski, Tobin R Sosnick, and Karl F Freed. "Modeling the hydration layer around proteins: applications to small-and wide-angle x-ray scattering". *Biophysical Journal* 101.8 (2011), pp. 2061–2069. DOI: 10.1016/j.bpj.2011.09.021.

[470] Shruthi Viswanath, DVS Ravikant, and Ron Elber. "Improving ranking of models for protein complexes with side chain modeling and atomic potentials". *Proteins: Structure, Function, and Bioinformatics* 81.4 (2013), pp. 592–606. DOI: 10.1002/prot.24214.

[471] Vladimir V Volkov and Dmitri I Svergun. "Uniqueness of ab initio shape determination in small-angle scattering". *Journal of Applied Crystallography* 36.3 (2003), pp. 860–864. DOI: 10.1107/S0021889803000268.

[472] Thom Vreven, Howook Hwang, Brian G Pierce, and Zhiping Weng. "Evaluating template-based and template-free protein–protein complex structure prediction". *Briefings in Bioinformatics* 15.2 (2014), pp. 169–176. DOI: 10.1093/bib/bbt047.

[473] Thom Vreven, Iain H Moal, Anna Vangone, Brian G Pierce, Panagiotis L Kastritis, Mieczyslaw Torchala, Raphael Chaleil, Brian Jiménez-Garcıa, Paul A Bates, Juan Fernandez-Recio, et al. "Updates to the Integrated Protein–Protein Interaction Benchmarks: Docking Benchmark Version 5 and Affinity Benchmark Version 2". *Journal of Molecular Biology* 427.19 (2015), pp. 3031–3041. DOI: 10.1016/j.jmb.2015.07.016.

[474] Gert Vriend. "WHAT IF: a molecular modeling and drug design program". *Journal of Molecular Graphics* 8.1 (1990), pp. 52–56. DOI: 10.1016/0263-7855(90)80070-V.

[475] Thomas Walzthoeni, Alexander Leitner, Florian Stengel, and Ruedi Aebersold. "Mass spectrometry supported determination of protein complex structure". *Current Opinion in Structural Biology* 23.2 (2013), pp. 252–260. DOI: 10.1016/j.sbi.2013.02.008.

[476] Ruixue Wan, Chuangye Yan, Rui Bai, Lin Wang, Min Huang, Catherine CL Wong, and Yigong Shi. "The 3.8 Å structure of the U4/U6. U5 tri-snRNP: Insights into spliceosome assembly and catalysis". *Science* (2016), aad6466. DOI: 10.1126/science.aad6466.

[477] Chu Wang, Philip Bradley, and David Baker. "Protein–protein docking with backbone flexibility". *Journal of Molecular Biology* 373.2 (2007), pp. 503–519. DOI: 10.1016/j.jmb.2007.07.050.

[478] Chu Wang, Ora Schueler-Furman, Ingemar Andre, Nir London, Sarel J Fleishman, Philip Bradley, Bin Qian, and David Baker. "RosettaDock in CAPRI rounds 6–12". *Proteins: Structure, Function, and Bioinformatics* 69.4 (2007), pp. 758–763. DOI: 10.1002/prot.21684.

[479] Hengbin Wang, Liangjun Wang, Hediye Erdjument-Bromage, Miguel Vidal, Paul Tempst, Richard S Jones, and Yi Zhang. "Role of histone H2A ubiquitination in Polycomb silencing". *Nature* 431.7010 (2004), pp. 873–878. DOI: 10.1038/nature02985.

[480] Ray Yu-Ruei Wang, Mikhail Kudryashev, Xueming Li, Edward H Egelman, Marek Basler, Yifan Cheng, David Baker, and Frank DiMaio. "De novo protein structure determination from near-atomic-resolution cryo-EM maps". *Nature Methods* 12.4 (2015), pp. 335–338. DOI: 10.1038/nmeth.3287.

[481] Andrew B Ward, Andrej Sali, and Ian A Wilson. "Integrative structural biology". *Science* 339.6122 (2013), p. 913. DOI: 10.1126/science.1228565.

[482] Benjamin Webb and Andrej Sali. "Comparative protein structure modeling using Modeller". *Current Protocols in Bioinformatics* (2014), pp. 5–6. DOI: 10.1002/0471250953.bi0506s47.

[483] James A Wells and Christopher L McClendon. "Reaching for high-hanging fruit in drug discovery at protein–protein interfaces". *Nature* 450.7172 (2007), pp. 1001–1009. DOI: 10.1038/nature06526.

[484] Iestyn Whitehouse, Chris Stockdale, Andrew Flaus, Mark D Szczelkun, and Tom Owen-Hughes. "Evidence for DNA translocation by the ISWI chromatin-remodeling enzyme". *Molecular and Cellular Biology* 23.6 (2003), pp. 1935–1945. DOI: 10.1128/MCB.23.6.1935-1945.2003.

[485] Shoshana J Wodak and Joël Janin. "Computer analysis of protein-protein interaction". *Journal of Molecular Biology* 124.2 (1978), pp. 323–342. DOI: 10.1016/0022-2836(78)90302-9.

[486] Justyna Aleksandra Wojdyla, Sarel J Fleishman, David Baker, and Colin Kleanthous. "Structure of the ultra-high-affinity colicin E2 DNase–Im2 complex". *Journal of Molecular Biology* 417.1 (2012), pp. 79–94. DOI: 10.1016/j.jmb.2012.01.019.

[487] Peter E Wright and H Jane Dyson. "Intrinsically disordered proteins in cellular signalling and regulation". *Nature Reviews Molecular Cell Biology* 16.1 (2015), pp. 18–29. DOI: 10.1038/nrm3920.

[488] Bin Wu, Adelinda Yee, Antonio Pineda-Lucena, Anthony Semesi, Theresa A Ramelot, John R Cort, Jin-Won Jung, Aled Edwards, Weontae Lee, Michael Kennedy, et al. "Solution structure of ribosomal protein S28E from Methanobacterium thermoautotrophicum". *Protein science* 12.12 (2003), pp. 2831–2837. DOI: 10.1110/ps.03358203.

[489] Sitao Wu, Jeffrey Skolnick, and Yang Zhang. "Ab initio modeling of small proteins by iterative TASSER simulations". *BMC Biology* 5.1 (2007), p. 17. DOI: 10.1186/1741-7007-5-17.

[490] Yinghao Wu, Mingyang Lu, Mingzhi Chen, Jialin Li, and Jianpeng Ma. "OPUS-Ca: A knowledge-based potential function requiring only Cα positions". *Protein Science* 16.7 (2007), pp. 1449–1463. DOI: 10.1110/ps.072796107.

[491] Bing Xia, Artem Mamonov, Seppe Leysen, Karen N. Allen, Sergei V. Strelkov, Ioannis Ch. Paschalidis, Sandor Vajda, and Dima Kozakov. "Accounting for observed small angle X-ray scattering profile in the protein–protein docking server CLUSPRO". *Journal of Computational Chemistry* 36.20 (2015), pp. 1568–1572. DOI: 10.1002/jcc.23952.

[492] Bing Xia, Sandor Vajda, and Dima Kozakov. "Accounting for pairwise distance restraints in FFT-based protein–protein docking". *Bioinformatics* (2016), In press. DOI: 10.1093/bioinformatics/btw306.

[493] Xian Xia, Xiaoyu Liu, Tong Li, Xianyang Fang, and Zhucheng Chen. "Structure of chromatin remodeler Swi2/Snf2 in the resting state". *Nature Structural & Molecular Biology* (2016), In press. DOI: 10.1038/nsmb.3259.

[494] Qifang Xu, Adrian A Canutescu, Guoli Wang, Maxim Shapovalov, Zoran Obradovic, and Roland L Dunbrack. "Statistical analysis of interface similarity in crystals of homologous proteins". *Journal of Molecular Biology* 381.2 (2008), pp. 487–507. DOI: 10.1016/j.jmb.2008.06.002.

[495] Qifang Xu and Roland L Dunbrack. "The protein common interface database (ProtCID)–a comprehensive database of interactions of homologous proteins in multiple crystal forms". *Nucleic Acids Research* 39.suppl 1 (2011), pp. D761–D770. DOI: 10.1093/nar/gkq1059.

[496] Li C Xue, João PGLM Rodrigues, Drena Dobbs, Vasant Honavar, and Alexandre MJJ Bonvin. "Template-based protein–protein docking exploiting pairwise interfacial residue restraints". *Briefings in Bioinformatics* (2016), In press. DOI: 10.1093/bib/bbw027.

[497] Kazuhiro Yamada, Timothy D. Frouws, Brigitte Angst, Daniel J. Fitzgerald, Carl DeLuca, Kyoko Schimmele, David F. Sargent, and Timothy J. Richmond. "Structure and mechanism of the chromatin remodelling factor ISW1a." eng. *Nature* 472.7344 (Apr. 2011), pp. 448–453. DOI: 10.1038/nature09947.

[498] Yuedong Yang and Yaoqi Zhou. "Ab initio folding of terminal segments with secondary structures reveals the fine difference between two closely related all-atom statistical energy functions". *Protein Science* 17.7 (2008), pp. 1212–1219. DOI: 10.1110/ps.033480.107.

[499] Yuedong Yang and Yaoqi Zhou. "Specific interactions for ab initio folding of protein terminal regions with secondary structures". *Proteins: Structure, Function, and Bioinformatics* 72.2 (2008), pp. 793–803. DOI: 10.1002/prot.21968.

[500] Zheng Yang, Keren Lasker, Dina Schneidman-Duhovny, Ben Webb, Conrad C Huang, Eric F Pettersen, Thomas D Goddard, Elaine C Meng, Andrej Sali, and Thomas E Ferrin. "UCSF Chimera, MODELLER, and IMP: an integrated modeling system". *Journal of Structural Biology* 179.3 (2012), pp. 269–278. DOI: 10.1016/j.jsb.2011.09.006.

[501] Yuzhen Ye and Adam Godzik. "FATCAT: a web server for flexible structure comparison and structure similarity searching". *Nucleic Acids Research* 32.suppl 2 (2004), W582–W585. DOI: 10.1093/nar/gkh430.

[502] Martin Zacharias. "Accounting for conformational changes during protein–protein docking". *Current Opinion in Structural Biology* 20.2 (2010). Theory and simulation / Macromolecular assemblages, pp. 180–186. DOI: 10.1016/j.sbi.2010.02.001.

[503] Martin Zacharias. "ATTRACT: protein–protein docking in CAPRI using a reduced protein model". *Proteins: Structure, Function, and Bioinformatics* 60.2 (2005), pp. 252–256. DOI: 10.1002/prot.20566.

[504] Martin Zacharias. "Combining coarse-grained nonbonded and atomistic bonded interactions for protein modeling". *Proteins: Structure, Function, and Bioinformatics* 81.1 (2013), pp. 81–92. DOI: 10.1002/prot.24164.

[505] Martin Zacharias, ed. *Protein-protein complexes: Analysis, modeling and drug design.* London, United Kingdom: Imperial College Press, 2010.

[506] Martin Zacharias. "Protein–protein docking with a reduced protein model accounting for side-chain flexibility". *Protein Science* 12.6 (2003), pp. 1271–1282. DOI: 10.1110/ps.0239303.

[507] Chao Zhang, George Vasmatzis, James L Cornette, and Charles DeLisi. "Determination of atomic desolvation energies from the structures of crystallized proteins". *Journal of Molecular Biology* 267.3 (1997), pp. 707–726. DOI: 10.1006/jmbi.1996.0859.

[508] Chi Zhang, Song Liu, Qianqian Zhu, and Yaoqi Zhou. "A knowledge-based energy function for protein-ligand, protein-protein, and protein-DNA complexes". *Journal of Medicinal Chemistry* 48.7 (2005), pp. 2325–2335. DOI: 10.1021/jm049314d.

[509] Jian Zhang and Yang Zhang. "A novel side-chain orientation dependent potential derived from random-walk reference state for protein fold selection and structure prediction". *PLOS One* 5.10 (2010), e15386. DOI: 10.1371/journal.pone.0015386.

[510] Junjie Zhang, Boxue Ma, Frank DiMaio, Nicholai R Douglas, Lukasz A Joachimiak, David Baker, Judith Frydman, Michael Levitt, and Wah Chiu. "Cryo-EM structure of a group II chaperonin in the prehydrolysis ATP-bound state leading to lid closure". *Structure* 19.5 (2011), pp. 633–639. DOI: 10.1016/j.str.2011.03.005.

[511] Qiangfeng Cliff Zhang, Donald Petrey, José Ignacio Garzón, Lei Deng, and Barry Honig. "PrePPI: a structure-informed database of protein–protein interactions". *Nucleic Acids Research* 41.D1 (2013), pp. D828–D833. DOI: 10.1093/nar/gks1231.

[512] Qiangfeng Cliff Zhang, Donald Petrey, Lei Deng, Li Qiang, Yu Shi, Chan Aye Thu, Brygida Bisikirska, Celine Lefebvre, Domenico Accili, Tony Hunter, et al. "Structure-based prediction of protein-protein interactions on a genome-wide scale". *Nature* 490.7421 (2012), pp. 556–560. DOI: 10.1038/nature11503.

[513] Yang Zhang. "I-TASSER server for protein 3D structure prediction". *BMC Bioinformatics* 9.40 (2008). DOI: 10.1186/1471-2105-9-40.

[514] Yang Zhang, Andrzej Kolinski, and Jeffrey Skolnick. "TOUCHSTONE II: a new approach to ab initio protein structure prediction". *Biophysical Journal* 85.2 (2003), pp. 1145–1164. DOI: 10.1016/S0006-3495(03)74551-2.

[515] Yang Zhang and Jeffrey Skolnick. "Automated structure prediction of weakly homologous proteins on a genomic scale". *Proceedings of the National Academy of Sciences of the United States of America* 101.20 (2004), pp. 7594–7599. DOI: 10.1073/pnas.0305695101.

[516] Zhe Zhang and Oliver F Lange. "Replica exchange improves sampling in low-resolution docking stage of RosettaDock". *PLOS One* 8.8 (2013), e72096. DOI: 10.1371/journal.pone.0072096.

[517] Zhe Zhang, Christina EM Schindler, Oliver F Lange, and Martin Zacharias. "Application of enhanced sampling Monte Carlo methods for high-resolution protein-protein docking in Rosetta". *PLOS One* 10.6 (2015), e0125941. DOI: 10.1371/journal.pone.0125941.

[518] Chunxiang Zheng, Chad R Weisbrod, Juan D Chavez, Jimmy K Eng, Vagisha Sharma, Xia Wu, and James E Bruce. "XLink-DB: database and software tools for storing and visualizing protein interaction topology data". *Journal of Proteome Research* 12.4 (2013), pp. 1989–1995. DOI: 10.1021/pr301162j.

[519] Zheng Zheng and Kenneth M Merz Jr. "Development of the knowledge-based and empirical combined scoring algorithm (kecsa) to score protein–ligand interactions". *Journal of Chemical Information and Modeling* 53.5 (2013), pp. 1073–1083. DOI: 10.1021/ci300619x.

[520] Alice Qinhua Zhou, Corey O'Hern, and Lynne Regan. "The power of hard-sphere models for proteins: Understanding side-chain conformations and predicting thermodynamic stability". *Bulletin of the American Physical Society* 58 (2013), Z46.00011.

[521] Hongyi Zhou and Jeffrey Skolnick. "GOAP: a generalized orientation-dependent, all-atom statistical potential for protein structure prediction". *Biophysical Journal* 101.8 (2011), pp. 2043–2052. DOI: 10.1016/j.bpj.2011.09.012.

[522] Hongyi Zhou and Yaoqi Zhou. "Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction". *Protein Science* 11.11 (2002), pp. 2714–2726. DOI: 10.1110/ps.0217002.

[523] Michael T Zimmermann, Sumudu P Leelananda, Andrzej Kloczkowski, and Robert L Jernigan. "Combining statistical potentials with dynamics-based entropies improves selection from protein decoys and docking poses". *The Journal of Physical Chemistry B* 116.23 (2012), pp. 6725–6731. DOI: 10.1021/jp2120143.

[524] Martin Zofall, Jim Persinger, Stefan R Kassabov, and Blaine Bartholomew. "Chromatin remodeling by ISW2 and SWI/SNF requires DNA translocation inside the nucleosome". *Nature Structural & Molecular Biology* 13.4 (2006), pp. 339–346. DOI: 10.1038/nsmb1071.

[525] Martin Zühlsdorf, Sebastiaan Werten, Barbara G Klupp, Gottfried J Palm, Thomas C Mettenleiter, and Winfried Hinrichs. "Dimerization-Induced Allosteric Changes of the Oxyanion-Hole Loop Activate the Pseudorabies Virus Assemblin pUL26N, a Herpesvirus Serine Protease". *PLOS Pathogens* 11.7 (2015), e1005045. DOI: 10.1371/journal.ppat.1005045.

Danksagungen

Ich möchte mich bei allen bedanken, ohne deren Unterstützung diese Dissertation nicht möglich gewesen wäre:

- meinem Doktorvater Prof. Martin Zacharias, der mich nicht nur fachlich exzellent unterstützt hat und mir vielfältigste Möglichkeiten gegeben hat spannenden wissenschaftlichen Fragestellungen in verschiedensten Richtungen nachzugehen, sondern auch für viele andere Belange als Mentor immer ein offenes Ohr und eine offene Tür für mich hatte. Besonders möchte ich mich bedanken für die Unterstützung im Bezug auf Karriereplanung, die Gespräche, vor allem beim Biophysical Society Meeting in Los Angeles, waren sehr bereichernd für mich.
- Dr. Sjoerd de Vries als zweites wissenschaftliches Standbein dieser Dissertation. Vom Einarbeiten in ATTRACT über die diversen Projekte konnte ich mich immer an ihn wenden. Danke für die vielen spannenden Diskussionen und intensiven gemeinschaftlichen Programmiersessions!
- meinen Kollaborationspartnern Dr. Felix Müller-Planitz, Nadine Harrer, Johanna Ludwigsen, Dr. Jan Lipfert, Linda Bruetzel, Yonatan Mideksa und Dr. Matthias Feige für spannende Projekte und biologische Fragestellungen, auf die ich meine entwickelten Methoden anwenden konnte und tolle Meetings mit regem Gedankenaustausch.
- Sonja Ortner für Unterstützung in administrativen Fragen.
- allen Mitglieder des Lehrstuhls T38 für die Gemeinschaft und den Austausch, die meine Zeit hier zu etwas Besonderem gemacht hat.
- meiner Familie, meinem Freund Felix und meinen Freunden, die mich immer tatkräftig und moralisch unterstützt haben und geholfen haben auch mal hakelige Phasen während der Doktorarbeit zu bewältigen.
- dem Center for Integrated Protein Science Munich für die Finanzierung dieses Projekts.