On the study of biomolecular interactions at different resolutions: Does size matter?

Jorge Roel-Touris

Doctoral Thesis

On the study of biomolecular interactions at different resolutions:
Does size matter?

Jorge Roel-Touris
NMR Spectroscopy, Bijvoet Centre for Biomolecular Research
Utrecht University, The Netherlands
February 2021

Printed in The Netherlands by ProefschriftMaken

On the study of biomolecular interactions at different resolutions: Does size matter?

Over de studie van biomoleculaire interacties bij verschillende resoluties: Doet grootte er toe?

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de
Universiteit Utrecht
op gezag van de
rector magnificus, prof.dr. H.R.B.M. Kummeling,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen op

maandag 15 februari 2021 des middags te 12.45 uur

door

Jorge Luis Roel Touris

geboren op 17 juli 1990
te Barakaldo, Spanje

Promotor: Prof. dr. A. M. J. J. Bonvin

Beoordelingscommissie:

Prof. dr. F. G. Förster
Prof. dr. A. Killian
Prof. dr. A. Perrakis
Prof. dr. K. Lindorff-Larsen
Prof. dr. S. J. Marrink

Table of Contents

General Introduction

"All models are wrong, yet some are useful"

George Box, Science and Statistics, 1976

Humanity creates models in attempts to understand complex physical, economical or demographical systems. These models are approximations of reality and, therefore, do not reflect all of its complexity. Designing a model is a creative, somewhat artistic, process for which making the correct assumptions is crucial. Commonly, these models capture key aspects of systems, while, inevitably, leaving out details assumed to be inessential. The idea of building a molecular model from atoms as the basis for understanding chemical behavior was perhaps bolder than we might currently think. Back in 1860, August Wilhelm von Hofmann created the first three-dimensional topological model for methane. His design consisted of four white balls, each of them attached by a stick to a central black one with a relative angle of 90º one to another[1]. Moreover, the white balls – representing hydrogen atoms – appeared to be larger in size than the carbon atom. This primitive model was meant to illustrate that atoms adopt specific spatial arrangements in molecular environments, a concept that was still not fully consolidated at that time.

In most disciplines, there are certain rules that guide the decision making of the modeling process. It was a decade later, in 1874, when Joseph Le Bel and Jacobus Henricus van't Hoff introduced the concept of stereochemistry or chemistry in space, which was previously observed by Louis Pasteur[2]. By using optical measurements, they were able to describe the tetrahedral arrangement of the atoms bound to the carbon. It turned out that the correct angle for methane was around 109.5º, instead of 90º. These discoveries completed the model proposed by von Hofmann. Nowadays stereochemistry plays a central role in modern chemistry, and defines the rules currently used to describe the possible different three-dimensional orientations of the atoms in space.

More complex molecules, such as polypeptidic chains, adopt intricated three-dimensional arrangements, essential for their biological functions.

This process, known as folding, was already a hot topic in the 60's. At that time, Cyrus Levinthal suggested that, in nature, proteins do not sample all possible configurations in terms of spatial arrangements since the folding process mostly occurs in few seconds[3]. This idea has survived through history and it is commonly referred to as the "Levinthal paradox". Currently, it is known that there are forces driving this process and that the native state is often not the one with the lowest free energy but rather a metastable state able to survive possible perturbations. Predicting the folded state of proteins is, to date, still a hot topic, with an increasing role for machine (deep)-learning approaches, which entered the game as catalysts of development[4].

The main biomolecules of life, namely proteins, nucleic acids, carbohydrates and lipids, often function by binding one to another. Their interactions, which mediate a wide range of biological functions, have been widely studied by the so-called classical structural biology techniques, including X-ray crystallography, Nuclear Magnetic Resonance (NMR) and cryo-electron microscopy (cryo-EM). Over the recent years, computational structural biology has gained importance to understand the underlying mechanisms of biomolecular interactions, for which a variety of computational models have been developed. Atomistic models describe complex biomolecular systems by explicitly taking all of their components into account (all atoms, including hydrogen atoms). United-atom models represent a first simplification in which non-polar hydrogens are neglected. To do so, the functional chemical groups to which they belong are adapted to implicitly account for these atoms. The computational cost associated to model larger systems is often overcome by further downscaling the resolution of the biomolecules under study. These simplifications, namely coarse-graining, typically group several heavy atoms into larger pseudo-atoms or beads. As a consequence, the energy landscape of the biomolecular interaction becomes smoother, effectively allowing for an easier sampling compared to atomistic

calculations.

To date, many coarse-grained models have been proposed in the literature, among which the MARTINI model[5] is among the most popular ones. The "*MARTINI-dome"* currently features mainly lipids, proteins, carbohydrates and nucleic acids, but can also handle polymers and nanoparticles. This model maps, generally, four heavy atoms onto one coarse-grained bead and consists of generic bonded (bond, angle, dihedral and improper dihedral) and non-bonded (van der Waals and electrostatics) interaction potentials. The parametrization of the MARTINI model relies on a combination of *bottom-up* and *top-down* approaches. While the bonded terms are obtained from the underlying atomistic geometry or by comparing to atomistic simulations *(bottom-up)*, non-bonded interaction potentials are optimized to reproduce experimental thermodynamic quantities *(top-down)*.

In 1978, Shoshana Wodak and Joel Janin studied for the first time the association of two relatively small proteins – BPTI and trypsin – by computational docking[6]. Docking aims to build three-dimensional models of macromolecular complexes by first, generating thousands of possible conformations (sampling), and then dicriminating between biologically- and non-biologically relevant models (scoring). In this first docking work, a simplified or coarse-grained representation of the system was used, which allowed, given the limited computational resources at the time, to effectively screen a large number of possible interfaces. The concept of coarse-graining originates from the seminal work of Michael Levitt, Arieh Warshel and Martin Karplus in the early 70's[7], where residues were represented by only two pseudo-atoms: The Cα atom and the centroid of the side chain. Ever since, models based on the *ball-and-stick* paradigm firstly introduced by von Hofmann[1] and using (or mixing) different resolutions have become extremely popular.

Over the last 40 years, docking has been consolidated as one of the most popular computational methods for studying biomolecular association. The other methods are those typically based on either molecular dynamics (MD) simulations, Monte Carlo approaches, or, more recently, template-based modeling. Docking has been applied to a wide range of systems. Furthermore, the docking field is increasingly making use of experimental information for improving the predictions. This, broadly known as integrative modeling, is a powerful approach to determine structures of biological systems by a combination of experimental and theoretical methods. Likely, this shift from blind to data-driven predictions originates from the 2000's with HADDOCK[8] and IMP[9] as pioneer software.

The HADDOCK, High Ambiguity Driven DOCKing, software developed in Utrecht was originally designed to be used in combination with chemical shift perturbations measured from NMR experiments and mutagenesis data[8]. Over the years, several developments have extended its capabilities. It can nowadays incorporate data from a wide variety of sources (including bioinformatic predictions and cryo-EM maps) as well as perform the docking at different resolutions (atomistic and coarse-grained scales – described in this thesis). The Integrative Modeling Platform designed by Andrej Sali, however, stands out by its capabilities of mixing simultaneously different levels of resolution during the course of the simulation[9]. This is especially useful when the three-dimensional coordinates of the components are not available and/or the experimental data are very sparse.

This thesis deals with three main topics, not in order of appearance: (1) Coarse-grained and hybrid approaches for the integrative modeling of large protein-protein and protein-nucleic acid complexes (**Chapters 1**, **3** and **4)**, (2) the use of experimental information in

docking calculations with LightDock[10], another docking software (**Chapter 2**), and (3) the integrative modeling of membrane-associated protein assemblies combining LightDock and HADDOCK (**Chapter 5**). The content is presented and discussed from an integrative modeling and docking point of view.

Chapter 1 provides a review of several representative coarse-grained/hybrid approaches and parametrization strategies for the modeling of biomolecular complexes[11]. **Chapter 2** describes the implementation and use of experimental information into the LightDock docking software[12]. **Chapter 3** details the implementation of a coarse-grained docking protocol for protein-protein complexes into the information-driven software HADDOCK[13], based on the MARTINI coarse-grained force field[14], and **Chapter 4** its extension to nucleic acids[15], including specific considerations to account for Watson-Crick interactions[16]. The final chapter of this thesis, **Chapter 5**, combines various of the developments described in *Chapters 2* and *3* into a novel protocol for the integrative modeling of membrane-associated protein assemblies, which are notoriously challenging to characterize experimentally and have received little attention so far. The thesis ends with a **Conclusions and Perspectives** section.

Chapter 1

Coarse-grained (Hybrid) integrative modeling of biomolecular interactions

Jorge Roel-Touris, Alexandre M.J.J. Bonvin

Published in 2020 in Computational and Structural Biotechnology Journal, Volume 18, Pages 1182 – 1190

Abstract

The computational modeling field has vastly evolved over the past decades. The early developments of simplified protein systems represented a stepping stone towards establishing more efficient approaches to sample intricated conformational landscapes. Downscaling the level of resolution of biomolecules to coarser representations allows for studying protein structure, dynamics and interactions that are not accessible by classical atomistic approaches. The combination of different resolutions, namely hybrid modeling, has also been proved as an alternative when mixed levels of details are required. In this review, we provide an overview of coarse-grained/hybrid models focusing on their applicability in the modeling of biomolecular interactions. We give a detailed list of ready-to-use modeling software for studying biomolecular interactions allowing various levels of coarse-graining and provide examples of complexes determined by integrative coarse-grained/hybrid approaches in combination with experimental information.

1. Introduction

The chemistry that supports life is extremely sophisticated. Despite advances over the past decades, the scientific community still lacks fundamental knowledge to fully understand the biology behind the cell at atomic level. We know that basic subunit atoms (i.e. carbon, oxygen, hydrogen and nitrogen) can combine and form complex molecules such as lipids, carbohydrates, nucleic acids and proteins. At the same time, these biomolecules associate and create more intricated assemblies that adopt specific three-dimensional (3D) structures, essential for their biological functions. Their interactions mediate a wide range of biological functions such as for example signal transduction, molecular recognition or transport. Indeed, roughly 80% of the proteins might function upon association with other biomolecules[17]. It is therefore of great importance to understand how these macromolecules interact. Next to experimental methods, complementary computational approaches have been developed with the so-called integrative modeling emerging as the most promising strategy[18]. In short, integrative modeling aims at obtaining structural insights into a given system under study that cannot be revealed by a single approach alone. To do so, it combines data from multiple information sources (e.g. nuclear magnetic resonance (NMR) spectroscopy, cryo-electron microscopy (cryo-EM), mass spectrometry (MS), small angle x-ray scattering (SAXS), bioinformatics analysis. . .)[19] into computational approaches to model the assemblies. Integrative modelling has been extensively used to model increasingly larger systems in the recent past[20]. In this sense, we are probably closer than ever to construct a predictive model of an entire cell[21].

Classical atomistic computational modeling of interactions remains inefficient for many molecular assemblies. Larger systems often require longer simulations and their complex conformational landscapes cannot be

efficiently and thoroughly sampled by atomistic approaches. The simplification of large systems to coarser representations offers a valuable approach to alleviate those limitations. There is already a huge body of literature on this topic and, in the present work, we do not aspire to give the most comprehensive review covering all possible contributions, but will focus on the modeling of biomolecular interactions. i.e. complexes, involving proteins, peptides and nucleic acids (DNA and RNA). The remaining of the text is organized as follows: We first start with a brief historical overview of the development of coarse-graining. We then describe several representative designs of simplified systems and parametrization strategies and discuss how these can be implemented into the modeling of biomolecular complexes, both for the generation of possible conformations (sampling) and the discrimination between native and non-native models (scoring). Finally, we provide an overview of currently available software that support coarse-grained modeling of biomolecular complexes and highlight several representative applications.

2. Historical perspective

The structural characterization of lysozyme in 1967[22] spurred Arieh Warshel to study enzymatic reaction mechanisms. His developments in this field under the supervision of Martin Karplus, inaugurated the now well-established quantum mechanics/molecular mechanics (QM/MM) methods[23]. In parallel, Michael Levitt, a PhD student at the Medical Research Council at that time, was making significant advances for studying molecular conformations by computational approaches: Together with Shneior Lifson in 1972 at the Weizmann Institute in Israel, Levitt and Warshel started working on a simplified representation of a protein, where spheres would represent amino acids.

In fact, this project, later on in 1975, turned out in the very first computer simulation of a protein system (pancreatic trypsin inhibitor) using a coarse-grained model[7]. These simulations suggested that the protein folding process has a relatively small number of conformations, and challenged the so-called "Levinthal paradox"[3]. In this work, each residue was represented by only two beads: The Cα atom and the centroid of the side chain. Non-bonded interactions were assumed to occur only between side chains. By doing so, only torsion angles between 4 consecutive Cα atoms were considered, considerably reducing the conformational space (one degree of freedom per residue). For all these premature findings, Karplus, Levitt and Warshel were awarded with the Nobel Prize in Chemistry in 2013.

In 1975, Chothia and Janin established the structural basis of the hydrophobic effect as fundamental to the stabilization of protein association[24]. All these pioneering findings were used as a basis for the first computational analysis of a protein–protein complex: In 1978, Wodak and Janin studied the association of BPTI and trypsin using a coarse-grained representation of the system[6]. They used a combination of a simple averaged potential energy function including non-bonded (van der Waals) and residue-solvent interactions. Whilst encouraging, this early model totally neglected electrostatic interactions and was thus unable to describe hydrogen bonds and salt bridges, which, later on in 1984, were suggested to provide the specificity of the association[25]. In spite of the incompleteness of this work, they shed light on the idea that a simplified protein model could be an effective alternative to screen a relatively large number of possible interfaces, which constituted the first coarse-grained docking simulation. Ever since, coarse-grained/hybrid modeling approaches have gained importance in the computational structural biology field[26] and have become central in the study of folding, dynamics and association mechanisms of biomolecules.

3. Coarse-grained/Hybrid modeling of biomolecular interactions

In this section, we will focus on macromolecular docking approaches allowing some level of coarse-grained/hybrid representations for the modeling of interactions. These usually include two different steps: The generation of possible complex conformations, referred to as sampling, and the discrimination between biologically and non-biologically relevant models referred to as scoring. The latter might also be an integral part of the sampling process, especially when experimental or predicted information is included to bias the sampling (e.g. restraints-driven sampling). We first describe various strategies to simplify the representation of polypeptides and nucleic acids and discuss existing parametrization strategies and force fields. We then focus on how coarse-grained/hybrid approaches can be applied during the sampling and scoring steps for modeling biomolecular interactions and end with a short discussion of backmapping approaches to restore full atomistic representations.

3.1. Simplified representations and topologies

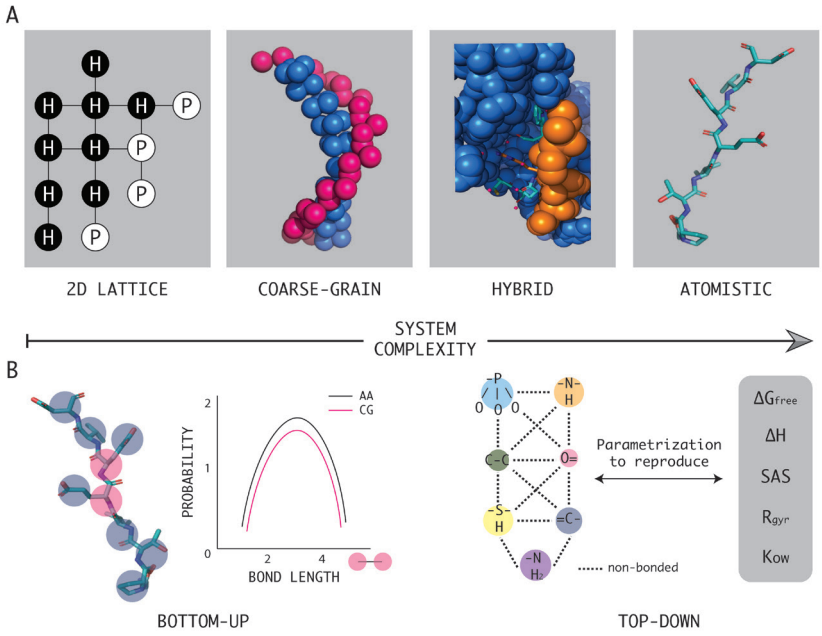
In general, a coarse-grained model aims at decreasing the complexity of a system by grouping several atoms into larger "pseudo-atoms" or "beads", thereby reducing the number of degrees of freedom. This results both in more efficient computations and a possible smoothening of the energy landscape that might facilitate the identification of relevant states of the system. In the context of proteins, the simplest models introduced are the hydrophobic/polar (HP) models (see Figure 1). These simplify the representations of a polypeptide chain[27] by considering only two type of beads (H and P), which, to some extent, approximate two types of residues:

hydrophobic (H) and polar (P)[28]. Albeit very minimalistic, HP representations have proven useful to study larger conformational changes and longer time scales. These models, and their variants, have been extensively studied in the past decade[29–32] and reviewed elsewhere[33]. Another example of a low-resolution model to represent proteins is SICHO (Side CHain Only)[34]. In the model developed by Kolinski and Skolnick[34], each amino acid is represented as a unique interaction site, located at the center of the sidechain. It is thus computationally very efficient but completely neglects backbone conformations (φ/ψ dihedrals)[35].

In order to overcome the inaccuracies of very simplistic representations, higher resolution models have been developed. PRIMO/PRIMONA, for proteins and nucleic acids, was proposed as a reduced quasi-atomistic resolution model[36]. Feig and co-workers[36] represent polypeptide backbones with three beads (Cα, N and a combined carbonyl site) and sidechains as a combination of up to five different particles. In the case of nucleotides, adenine, cytosine and uracil are represented by four coarse-grained particles, and guanine and thymine by five. The sugar-phosphate backbone of the PRIMONA model consist of eight different CG beads. In contrast, the HiRE-RNA model designed by Pasquali and Derremaux[37] only considers three of the seven backbone torsional angles (α, β and γ); each RNA nucleotide is represented by six (pyrimidine bases) or seven (purine bases) beads, allowing for a reduction of ~70% of the number of particles compared to a fully atomistic structure. Similar to PRIMO, in the SIRAH model[38]the positions of the nitrogen, carbon and oxygen from the peptide bonds are kept at pseudo-atomistic resolution, while sidechains are treated at a lower degree of detail (from one to five different beads). This model also allows for the study of protein-DNA interactions by molecular dynamics through the use of an explicit/CG solvation scheme[39,40].

Figure 1. Examples of various coarse-grained models.

A) The panels from left to right illustrate the increase in the complexity of the system (i.e. decreased coarse-graining): A 2-D lattice representation of a HP model, a coarse-grained (4:1 mapping) of a dsDNA molecule, a hybrid representation of a protein–protein interface (AA/CG) and an atomistic model of a peptide.
B) The two traditional parametrization strategies. Bottom-up: Bond-lengths are parametrized by mapping to distributions of reference atomistic simulations. Top-down: Models are designed to match specific properties (e.g. thermodynamic quantities) of the system.



Other coarse-grained models have been designed to be easily transferable and applicable to multiple systems. Among those, MARTINI is probably the most popular one. The current "MARTINIdome" includes: lipids[5], proteins[41], polymers[42,43], carbohydrate[44], water[45], glycolipids[46], nucleotides[16,47] and nanoparticles[48]. The systems are represented by four different basic particles – nonpolar (N), polar (P), apolar (C) and charged (Q) – that are further classified based on their degree of polarity and hydrogen bonding properties, giving a total of eighteen unique "building blocks". The MARTINI force field for proteins,

in its latest official release (2.2p), includes off-center charges for polar and charged residues[14]. These represent a good proxy for hydrogen bond and salt bridges formation and thus for molecular recognition. For nucleic acids, much like PRIMONA, the MARTINI model specifically accounts for Watson-Crick base pairing (eight additional beads) to stabilize the DNA double helix structure.

3.2. Parametrization of coarse-grained force fields

3.2.1. Classical parametrization strategies

In the context of molecular modeling, the set of parameters and functions used to calculate the potential energy of a system is commonly referred to as force field. Atomistic force fields provide parameters usually for every type of atom in a system (hydrogen included) but also united atom representations are often used in which non-polar hydrogens are neglected. In contrast, coarse-grained potentials are a cruder representation of the inter- and intra-molecular interactions. Regarding the latter, their parametrization follows two main routes: Hierarchical (*bottom-up*) and pragmatic (*top-down*) coarse-graining[49]. The key idea of hierarchical coarse-graining is that, the interactions at a less detailed level are the result of the collective interactions at the more detailed level[50]. As an illustration, in the 1975 abovementioned study by Levitt and Warshel[7], the interactions between coarse-grained sites were derived in a *bottom-up* way by explicitly summing up all microscopic interactions of an atomistic model. One obvious limitation of these models is that the quality of the coarse-grained model highly depends on the accuracy of the underlying atomistic one. Similarly, the seminal force-matching (FM) method proposed by Ercolesi and Adams[51] and further developed by Voth and co-workers[52,53] under the name of MS-CG (multiscale coarse-graining) uses atomistic-level inter-

actions to derive coarse-grained potentials. In short, those potentials are systematically fitted to atomistic forces by minimizing the mean-square errors between them. Much like iterative Boltzmann (IB) derived models[54], these force fields are usually more accurate as compared to more generic ones. However, they are typically less transferable and require more parametrization effort. These methods, and their extensions[55], have been recently applied to coarse-grained models for proteins such as the UNRES model[56].

Pragmatic force fields, however, are designed in such a way that they reproduce a given chosen (experimental) property[57]. The earlier lattice models (such as HP) represent a well-studied example of *top-down* coarse-graining. These models are typically cheaper to parametrize, easily transferable (to similar systems) and use rather simple analytical potentials. In a similar way and as shown in Figure 1, methodologies based on reproducing thermodynamical properties have been extensively applied in different branches of chemistry such as physical and organic chemistry. Equations of State (EoS), which are mathematical relationships between the thermodynamic variables of a given system, have been shown appropriate to accurately link the macroscopic properties of the system and the force field parameters[58]. As an example, the powerful SAFT-γ EoS, a variation of the Statistical Associating Fluid Theory (SAFT), has been used to estimate the coarse-grained potentials of the Mie force field[59]. This force field has been recently used to calculate solvation free energies of aromatic compounds, which are broadly used in the pharmaceutical industry for drug design purposes[60].

3.2.2. Machine learning-based parametrization

Machine learning, and especially deep learning, is revolutionizing in the last years many areas of science and technology. Certainly, the most significant breakthrough of the decade in the field of protein folding has

been the development of AlphaFold[61]. DeepMind, an artificial intelligence company affiliated to Google, has designed a deep learning-based method that represents a substantial advance as compared to classical modeling techniques[4,62]. These machine learning methods have been also applied in the development of force fields and are usually purely based on existing data. A general approach to design a machine (deep) learning-based force field typically includes: The generation of reference atomic configurations and forces (QM calculations), the identification of specific signatures, the selection of training and test datasets, the mapping of selected signatures to forces using specific algorithms and the assessment of the resulting predictive model[63]. Deep neural networks[64], adversarial machine learning models[65] and genetic algorithm[66] have been recently shown appropriate for the development coarse-grained force fields. Altogether, machine learning-based parametrization methodologies represent an emerging trend to automatize analytical model building from more complex data, which can deliver faster and perhaps more accurate results with minimal human intervention.

3.2.3. Combining different levels of resolution

An exhaustive, yet accurate, sampling of the conformational landscape is crucial in attempts to model biomolecular interactions and evaluate the underlying energetics. The use of simplified representations offers an effective way of sampling the landscape. However, the reduced accuracy due to the inherent simplifications still limits the systems and processes that can be studied by CG approaches. Hybrid approaches, which typically couple coarse-grained and atomistic-level representations, aim to overcome these limitations by combining different levels of resolution[67]. These combined approaches might be very helpful for quantitative studies (e.g. free energy calculations of large systems[68,69]), while still reducing the computational cost.

They are also particularly useful to include components of a system for which no or only low-resolution structural data are available. A key challenge in hybrid modeling is to integrate the different levels of resolution and to describe the AA/CG interactions. Standard mixing rules[70] have been historically very successful for this task. In short, Lennard-Jones and electrostatic interactions for mixed systems can be averaged and combined with an optimal scaling parameter depending on the size of the system[71]. Besides energetics, it still remains unclear how the interaction between two atoms might be affected by a coarse-grained surrounding as compared to its "native" environment and vice versa[72].

There are several hybrid schemes proposed in the literature, with MARTINI as a popular choice for the coarse-grained representation. One example is the PACE force field[73,74], which pairs MARTINI (water and lipids) with a united-atom protein model. In this case, the AA/CG parameters are optimized against specific thermodynamic data, which somehow limits its direct applicability to other systems. GROMOS/MARTINI coupling[72] has also been described as a potential alternative. In this work, cross-resolution interactions are calculated via virtual interactions sites on relevant atomistic groups and the standard CG beads, an approach that might lead to unbalanced electrostatics behaviors. For this reason, Wassenaar and coworkers[75] introduced an explicit electrostatic AA/CG coupling on the coarse-grained side. More recently, the CHARMM/PRIMO coupling has been proposed for single hybrid simulation purposes[76]. In the model proposed by Kar and Feig[76], the atomistic segment of the hybrid model was found to structurally deviate more than its corresponding one in a full atomistic model. This suggests that proper mixing of resolutions remains a difficult problem.

In the context of integrative modeling, the integration of experimental data at the various possible levels might have a crucial role for hybrid

representations of the system. At the sampling level, data can be used to narrow the conformational search so that binding incompetent and/or irrelevant regions are discarded *a priori*. This strategy has been shown to be best suited compared to post-simulation filtering approaches. It not only outperforms the scenario where data is solely used to discard models with a high degree of uncertainty, but also reduces significantly the computational cost[12]. Data can be also incorporated at the scoring level via a numerical penalty term or as restraining energy potential[77]. As an example, in HADDOCK[8] the distance restraints are incorporated into the scoring scheme via a soft-harmonic potential where the potential becomes linear for violations longer than 2Å[78], effectively avoiding large forces for high restraints violations. Therefore, the incorporation of data in the modeling might work as a firewall and somewhat reduce the impact of inaccuracies of hybrid schemes in terms of intra- and inter-molecular interactions.

3.2.4. Sampling and scoring schemes

Decreasing the computational cost, as well as the complexity of the system, is a major goal of coarse-grained modeling. By lowering the resolution, the energy landscape becomes smoother and it is therefore, in principle, easier to identify the global minimum. In the context of integrative modeling with HADDOCK, we recently showed that introducing the MARTINI coarse-grained force field results in a substantial increase (8–30%) in the number of near-native models generated[13]. We also find CG sampling schemes in ATTRACT[79–81] (also hybrid scoring), CABS-dock[82,83] (also scoring), FRODOCK2.0[84], InterEvDock2[85,86] (also scoring), LZerD[87,88], MAXDo[89], MCDNA[90] (also scoring), MDockPP[91] and RosettaDock[92] (also scoring in RosettaDock 4.0[93]). Some of the methods used by these software to sample the conformational landscape includes: Rigid-body energy minimization,

Fast Fourier Transformation (FFT) or Molecular Dynamics (Monte Carlo). For the purpose of scoring, coarse-grained molecular dynamics simulations have been alsoevaluated on a heterogeneous benchmark of protein–protein docking models[94]. Other modeling software such as: DOCK/PIERR[95], GALAXY[96,97], LightDock[10], MEGADOCK 4.0[98,99], PPI3D[100,101], pyDock[102,103] and V-D^{2}OCK[104] incorporate, to some extent, coarse-grained/hybrid scoring approaches for (quasi)atomistic models.

IMP[9] and PyRy3D (genesilico.pl/pyry3d) are examples of ready-to-use hybrid modeling software for predicting (sampling and scoring) biomolecular assemblies allowing to incorporate experimental data into their calculations. The Integrative Modeling Platform leans on the concept that the resolution of the representation depends on the quantity and quality of the available information. This information is also encoded in a scoring function, whose ultimate goal is to evaluate the uncertainty of the generated models. Andrej Sali and co-workers[18] understand the modeling as an endless cyclic process driven by the continuous acquisition of data. In IMP, the different subunits are represented as a combination of spherical beads of varying sizes (different levels of coarseness). The same subunits can be also be represented as 3D Gaussians (for EM map fitting) and thus combine different resolution scales simultaneously[105]. During the conformational sampling, the relative distances from all the CG beads and Gaussians are either constrained (in rigid bodies) or restrained (in flexible bodies) by the sequence connectivity.

For very high degrees of coarse-graining, only geometric considerations, e.g. exclude volume, might be used in the computations. PyRy3D allows for building low-resolution models of large macromolecular assemblies. In the software developed by Kasprzak and Bujnicki (genesilico.pl/ pyry3d), proteins and nucleic acids can be represented as rigid-bodies or as flexible shapes. A spatial restraints-driven Monte Carlo approach is used to bring the components together followed by an evaluation via a simple scoring

function. For a more detailed list of software that allow for building structural models of multi-subunit macromolecular complexes refer to Table 1.

3.2.5. Backmapping from coarse-grained to atomistic resolution

The inherent loss of accuracy of coarser representations is a limiting factor when analyzing integrative models of biomolecular complexes. Atomic details, such as specific contacts, are usually essential to understand molecular recognition and it is therefore crucial to accurately reconstruct atomistic models from their CG counterparts[106]. This process is commonly referred in the literature as reverse transformation, inverse mapping or backmapping. There is currently a number of different backmapping protocols proposed, which mostly follow two different stages: (1) The generation of an atomistic structure based on the coarse-grained coordinates, and (2) a relaxation step of the generated AA structure.

For the first step, geometrical interpolation[36,107,108], random placement[109] and fragment-based methods[92,110–112] are the most used ones. All these methods perform sufficiently well according to backbone deviations (<1.0 Å in general) but side chain reconstruction seems more problematic[113]. Side chain optimization has been extensively studied as it directly applies for protein designing purposes. The most successful methods discretize possible side chain conformations into rotamers and usually require of an exhaustive search algorithm (e.g. Monte Carlo, simulated annealing...) and an effective scoring function for selecting the proper side chain conformation. The backmapped atomistic structures can then be further improved by energy minimization[13,114] and/or more sophisticated molecular dynamics-based approaches[115]. In HADDOCK, the CG generated models are converted into atomistic resolution by using distance restraints between the atoms and their corresponding coarse-grained beads. Using those restraints, the all-atom

models of the individual components of a complex are morphed onto the coarse-grained complex by a series of energy minimizations and Cartesian molecular dynamics[13].

4. Application examples of integrative modeling of protein interactions

Ultimately, the true value of any biomolecular model is in the structural information and insights that it provides. When speaking about integrative modeling here, we refer to the branch of structural biology whose aim is to gain structural insights into biomolecular complexes by integrating a wide variety of experimental information into computational calculations. There are various challenges associated with the incorporation and use of that information for the modeling of assemblies. However, a detailed overview of those is beyond the scope of this manuscript and have been reviewed in depth elsewhere[117–119].

The relevance of integrative models is underscored by the fact that the Protein Data Bank[120,121] has now started to collect them in a new integrative model database (PDB-dev; pdb-dev.wwpdb.org)[122,123], which ultimately should be merged into the current PDB database. Since 2014, it is possible to archive structural models obtained by combining traditional structural experimental techniques such as NMR spectroscopy, electron microscopy (3DEM), small angle scattering (SAS), atomic force microscopy (AFM), chemical cross-linking, Förster resonance, energy transfer (FRET), electron paramagnetic resonance (EPR), mass spectrometry (MS), Hydrogen/Deuterium exchange (HDX) and various bioinformatic approaches, with computational methods. In this section we highlight several examples of integrative structures of protein complexes that have been determined by combining coarse-grained/hybrid computational approaches with experimental information.

Table 1.

Available software for building structural models of protein, peptide and/or DNA complexes that incorporates a coarse-grained/ hybrid approach into their protocols. Most of the listed software are available as webserver and/or standalone package.

Modeling platform	System(s)	Characteristics	Link	Reference (s)
ATTRACT	Protein, peptide and DNA	CG sampling and multiscale scoring	attract.ph.tum.de	79–81
CABS-DOCK	Peptide	CG sampling and scoring	biocomp.chem.uw.edu.pl/CABSdock	82,83
DOCK/PIERR	Protein	Multiscale scoring	clsbweb.oden.utexas.edu *	95
FRODOCK2.0	Protein	3D grid potential maps	frodock.chaconlab.org	84
GALAXY	Peptide	Multiscale scoring	galaxy.seoklab.org	96,97
HADDOCK	Protein, peptide and nucleic acids	CG sampling	wenmr.science.uu.nl/haddock2.4	13,15,116
IMP	Protein and DNA	Multiscale sampling and scoring	integrativemodeling.org	9
InterEvDock2	Protein	Sampling by FRODOCK2.0 and CG scoring	bioserv.rpbs.univ-paris-diderot.fr/services/ InterEvDock2	85,86
LightDock	Protein, peptide and DNA	Multiscale scoring	lightdock.org	10
LZerD **	Protein and peptide	3DZD representation and multiscale scoring	kiharalab.org/proteindocking	87,88
MAXDo	Protein	CG sampling	lcqb.upmc.fr/CCDMintseris	89
MCDNA	Protein and DNA	CG sampling and scoring	mmb.irbbarcelona.org/MCDNA	90
MDockPP	Protein	CG sampling	zoulab.dalton.missouri.edu	91
MEGADOCK 4.0	Protein	Multiscale scoring	bi.cs.titech.ac.jp	98,99
PPI3D	Protein	Voronoi tessellation-based scoring	bioinformatics.ibt.lt/ppi3d	100.101
pyDock	Protein	CG scoring	life.bsc.es/pid/pydockweb	102,103
PyRy3D	Protein and DNA	Multiscale sampling and scoring	genesilico.pl/pyry3d	-
RosettaDock	Protein	CG sampling	rosettacommons.org	92,93
V-D^{2}OCK	Protein	CG scoring	bioinsilico.org/cgi-bin/VD2OCK/	104

* *Submission to DOCK/PIERR webserver is no longer supported.*

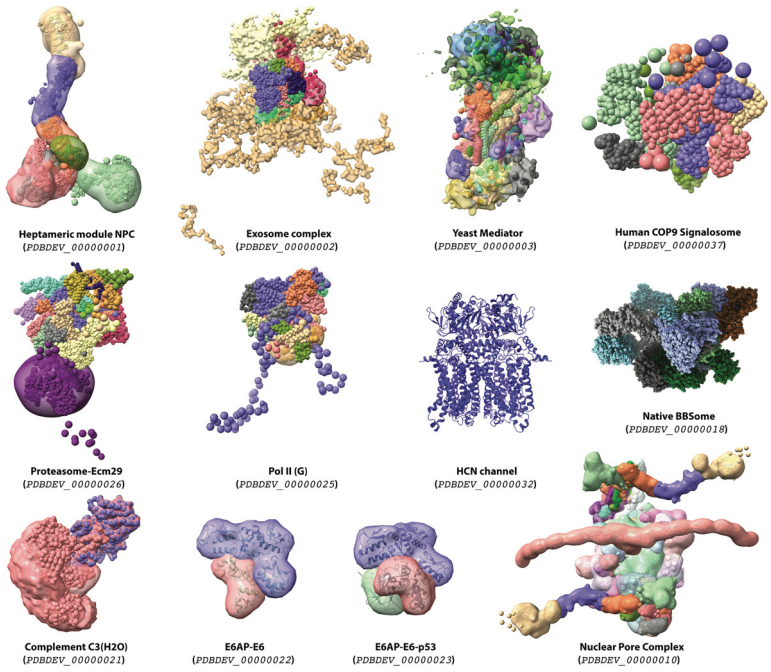
** *LZerD has an specific protocol for modeling unstructured protein–protein interactions*[88]

Among all archived structures, we find a number of them determined by coarse-grained/hybrid computational methods in combination with a wide variety of structural data (see Figure 2). Integrative structures derived from chemical cross-linking data are by far the most abundant ones, including models of the heptameric module of NPC[124], the exosome complex[125], the Complement C3 (H2O)[126], the E6AP/UBE3A-p53 enzyme-substrate complex[127], Pol II(G)[128], the Proteasome-Ecm29 complex[129] and the canonical/non-canonical human COP9 Signalosome[130]. Protein cross-links have been also combined with other types of experimental information such as three/two-dimensional Electron Microscopy (2DEM/3DEM) and/or SAS to determine structures like the yeast Mediator complex[131] or the native BBSome[132]. Other sources of information such as mutagenesis and NMR data[133] and single molecule FRET data[134] have been also used.

There are also multiple examples of integrative structures, not deposited in the PDB-dev database, which have been modelled by integrative coarse-graining methods. One of those is the ATP synthase membrane motor. Leone and Faraldo-Gómez[135] proposed a computational integrative model based on chemical cross-links, a cryo-EM map (~7Å of resolution) and evolutionary couplings. The initial homology models of either subunits were refined against the experimentally determined cryo-EM map using Rosetta, which starts its conformational exploration in coarse-grained resolution. The computationally generated models were further validated with co-evolutionary and cross-linking data and revealed important mechanistic insights into the function of the ATP synthase. Another representative example is the ISWI ATPase complex. Using upper bound distance restraints based on BS3, BS2G and UV cross-links, Harrer and coworkers[136] modelled the complex with ATTRACT, which performs a rigid-body energy minimization driven by a coarse-grained force field[110] and the distance restraints provided. The top scoring ISWI models were validated against SAXS data.

Figure 2. Examples of integrative structures determined by partial/full coarse-grained/hybrid computational approaches.

Structures archived in the PDB-dev datatabase[122,123] (pdb-dev.wwpdb.org). Pictures were generated with ChimeraX[137]. The experimental information used for the modeling (if included) has been omitted for visualization purposes. Models can be directly opened in ChimeraX from the command line as: *open [model_number] from pdbdev ignoreCache true*



The Nuclear Pore Complex (NPC) is probably the largest protein assembly determined by an integrative structural approach to date. It constitutes an eight-fold symmetrical cylindrical complex of 552 copies of 32 different nucleoporin proteins (Nups)[138]. With respect to the computational modeling, the NPC was represented in a multiscale fashion including multiple levels of coarseness. As an illustration, all rigid bodies derived from X-ray, NMR and integrative structures were coarse-grained into two different resolutions. They either mapped single residues or consecutive portions of up to ten different amino acids into larger beads.

The modeling was performed using the integrative modeling platform software (IMP) (integrativemodeling.org)[9]. The experimental information available included chemical cross-links, a cryo-ET density map, immuno-electron microscopy localizations, excluded volume, sequence connectivity, the shape of the pore membrane, symmetry and SAXS data, which were used to benefit the sampling, to improve the scoring, to filter out inconsistent models and/or validation purposes. By putting all these data together, they were able to fully describe, at sub-nanometer precision, the structure of the entire NPC.

5. Concluding remarks

Over the past decades, coarse-grained/hybrid modeling has been demonstrated as a powerful approach to model biomolecules and their interactions. It extends the capabilities of traditional atomistic protocols. There are multiple models to simplify the three-dimensional representation of biomolecules, each of those specifically designed to answer a specific research question. The choice between different representations directly affects the sampling and scoring capabilities of current modeling approaches. In other words, the smaller the number of pseudo-atoms or beads, the higher the increase in speed but the lower the accuracy of the resulting models. For cases where higher level of resolution is required, multiscale/hybrid modeling might help to alleviate the inherent loss of accuracy of pure coarse-grained models as demonstrated, for instance, in the modeling of the nuclear pore complex. Nevertheless, there is still an urgent need for improving interaction schemes. Coarse-grained force fields derived from classical molecular mechanics are not easily transferable and therefore, very much system-dependent. On the contrary to *bottom-up* strategies, *top-down* approaches aim to generalize structural patterns that have been

seen in thousands of known structures and/or to reproduce thermodynamic quantities. Likely, a combination of *bottom-up* and *top-down* approaches is a better option. In other words, improving *top-down* models by inferring additional interaction terms derived by *bottom-up* coarse-graining might have the most impact in future designs, increasing both their accuracy and applicability range to wider, larger and more complex assemblies. We are now approaching a time where, taking advantage of all scientific and technological advances, one might expect to build reasonable three-dimensional models of cells, which might provide insights into still unknown cellular mechanisms.

Chapter 2

LightDock goes information-driven

Jorge Roel-Touris, Alexandre M.J.J. Bonvin,
Brian Jiménez-García

Published in 2020 in Bioinformatics, Volume 36, Issue 3,
Pages 950 – 952

Abstract

The use of experimental information has been demonstrated to increase the success rate of computational macromolecular docking. Many methods use information to post-filter the simulation output while others drive the simulation based on experimental restraints, which can become problematic for more complex scenarios such as multiple binding interfaces. We present a novel method for including interface information into protein docking simulations within the LightDock framework. Prior to the simulation, irrelevant regions from the receptor are excluded for sampling (filter of initial swarms) and initial ligand poses are pre-oriented based on ligand input information. We demonstrate the applicability of this approach on the new 55 cases of the Protein–Protein Docking Benchmark 5, using different amounts of information. Even with incomplete or incorrect information, a significant improvement in performance is obtained compared to blind ab initio docking.

1. Introduction

Computational tools are essential to predict and describe three-dimensional (3D) interactions between biomolecules. In particular, integrative approaches, i.e. data- or information-driven, are broadly used in order to combine experimental data with docking simulations[9,86,103,118,139,140]. In the context of molecular docking, there are still two main challenges: (1) searching the conformational space, especially in the case of highly flexible molecules, and (2) evaluating and selecting near-native poses out of the generated conformers, which is usually referred to as scoring.

LightDock[10] is a multiscale flexible framework for the 3D determination of binary protein complexes based on the Glowworm Swarm Optimization (GSO)[141] algorithm that systematically optimizes the generated docking poses towards those energetically more favorable at every simulation step. Introducing restraints or biases in docking is a powerful mechanism to drive the simulation towards poses that satisfy those restraints[8].

Here we describe and benchmark an updated implementation of LightDock that now supports the use of information to drive or bias the docking simulation by filtering out swarms, pre-orienting ligand poses based on the available information and biasing the scoring energy upon satisfied residue contact restraints. The results on the benchmark demonstrate a high performance of LightDock when used in combination with additional information. We also explore different scenarios with less accurate or incorrect information to show the versatility and robustness of our approach.

2. Materials and Methods

2.1. Swarms selection based on receptor residue restraints

LightDock simulations are organized in swarms over the receptor surface. Given an initial number of swarms *S* (by default 400) and residue restraints *R* specified by the user, we select the ten closest swarms to each residue in *R* (Euclidean distance). The set of swarms to be simulated is therefore the union of the different swarms selected for each restraint residue, which is a subset of the initial number of swarms *S*.

2.2. Glowworms pre-orientation based on ligand residue restraints

Each glowworm in the swarm encodes a given complex pose. The poses evolve in translational (Cartesian), rotational (Quaternions) and conformational space through an Anisotropic Network Model (ANM) space. The ANM model considers (by default) the ten first non-trivial normal modes calculated on the Cα and further extended to the rest of the atoms. These are included in each glowworm optimization vector to model backbone flexibility of both receptor and ligand molecules.

For each swarm, we select from the set of input restraints, the 10 closest receptor residues with respect to the geometric center of the swarm (R_c). Then, we create random receptor-ligand restraint pairs $\{r,l\}$ where $r \in R_c$ and l is a defined restraint residue of the ligand molecule. Finally, we orient each ligand pose using the vector facing the direction given by $\{r,l\}$. Figure 1 shows the preferred orientation of yellow arrows pointing towards the receptor restraint residues.

2.3. Score bias according to percentage of satisfied residue restraints

LightDock is somehow agnostic of the scoring function as previously discussed[10]. The overall quality of the simulation will, of course, heavily depend on the capabilities of the selected scoring function to successfully describe the protein docking energetic landscape. In this new implementation, we calculate the intersection between the set of input restraints provided by the user and the set of those in contact for a given pose (3.9Å distance cutoff). The final score (eq. 1) of the complex is increased by the percentage of satisfied restraints (no penalties if none of the restraints is satisfied).

$$\varepsilon_f = \varepsilon + P_r * \varepsilon + P_l * \varepsilon \qquad \text{(eq. 1)}$$

where is the energy as calculated by the scoring function, and and are the percentage of satisfied restrained residues of the receptor and ligand respectively.

2.4. Design of artificial interfaces with the inclusion of false positive residues

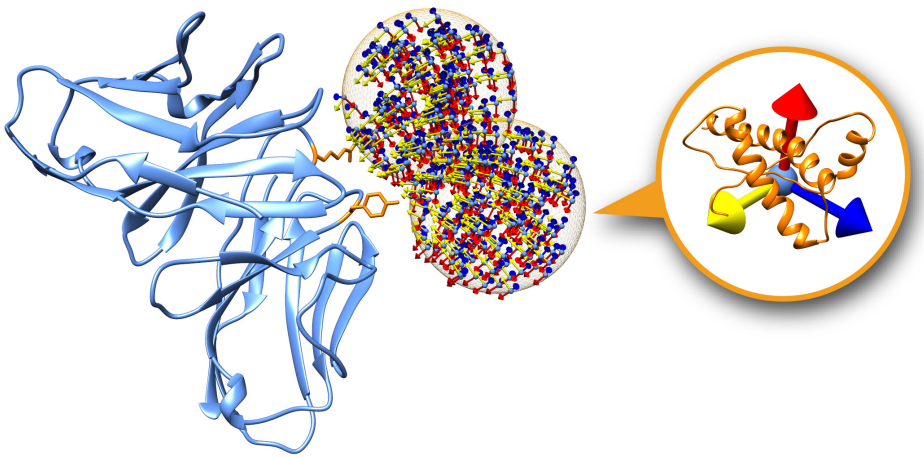
For TI_{50}, TI_{25}, $TI_{REC\text{-}50}$ and $TI_{REC\text{-}25}$ experiments, we designed artificial interfaces, of equal size as our true interface description *(TI)*, to include false positive residues in our restraint's definition. For each of the cases, we clustered the *TI* residues in either 2 (TI_{50} and $TI_{REC\text{-}50}$) or 4 clusters (TI_{25} and $TI_{REC\text{-}25}$) of equal cluster size using the *AgglomerativeClustering* algorithm included in the *scikit-learn* python package[142]. This algorithm recursively merges the pair of clusters that

minimally increases a given linkage distance, for which we used Euclidean distance in order to assure contiguity between clustered residues in space.

At this point, we expanded the clusters (receptor and ligand separate interfaces) with the inclusion of the closest surface accessible (as calculated by ProDy[143]) neighboring residues at a maximum distance of 10Å, which were obviously not part of the *TI* definition. We stopped this expansion once the size (in terms of number of residues) of the generated artificial interface (true positive and false positive residues) is equal to our original *TI* description.

Figure 1. Representation of two swarms (orange mesh) over the surface of a receptor protein (blue).

In orange, the residues considered as restraints and therefore used to filter out the initial swarms prior the simulation. The initial orientations of the ligands within the swarms are represented using an orthogonal axis (x, y, z).



3. Results

Due to the nature of the LightDock framework, information about interfacial residues can be applied at different levels depending on the availability of

information for the receptor, the ligand or both. On the receptor side, we filter out initial swarms that are not in the proximity of the defined restraints, with the collateral advantage of reducing considerably the computation time. On the ligand side, we orient initial poses based on randomly selected receptor-ligand restraint pairs. It is worth noting that these two steps (filtering and pre-orientation) are only performed at the initial setup stage of the simulation.

The latest release of LightDock (0.7.0)[144], which now supports the use of information to drive the docking in the format of residue restraints, was tested on the 55 unbound new entries of the Protein Docking Benchmark version 5[145], which represents an unbiased dataset where no software/scoring functions were trained in, and includes 16 antibody-antigen complexes. We defined various scenarios to demonstrate its versatility and robustness as follows:

1.- TI: True interface, defined as those residues at 3.9Å distance (as also defined in LIGPLOT[146] by default) from the partner molecule. This is an ideal case where a fully accurate definition of interface residues is available, but no specific contacts are defined.

2.- TI_{50}: We defined two different artificial interfaces with half of the *TI* residues and equal number of non-interfacial residues forming a contiguous patch as above described. Results are reported as averaged success rates of both runs (using each of the designed interfaces).

3.- TI_{25}: In the same way as in TI_{50}, we defined four different sets of restraints with one fourth of the original *TI* and three times more false positive residues forming a contiguous patch. Results are reported as averaged success rates of the four docking calculations (each one using a different artificially designed interface).

4.- TI_{REC}: Only the *TI* from the receptor is considered as restraints.

5.- TI_{REC-50}: As in TI_{50}, but only considering the receptor interface residues.

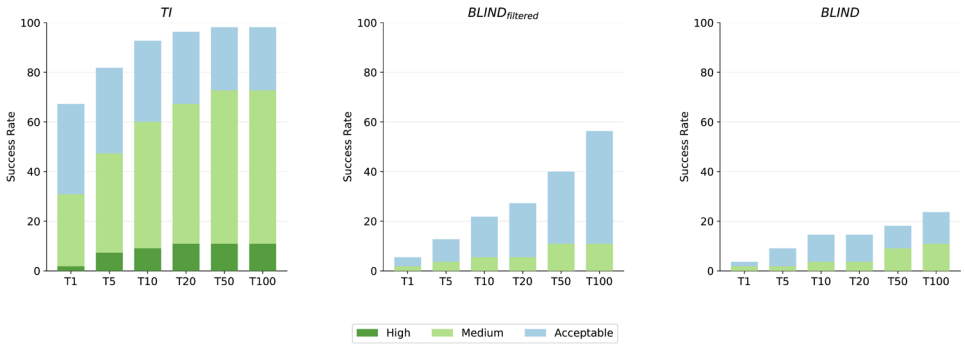
6.- TI_{REC-25}: As in TI_{25}, but only considering the receptor interface residues.

7.- *TI_{SINGLE}:* Only one receptor-ligand residue pair, making a real contact, is used as residue restraints.

8.- *TI_{ONE}:* Only one residue on the receptor, the same one as defined in *TI_{SINGLE}*, is considered as restraint, without any information on the ligand side.

Figure 2. Comparison of performance of LightDock when data is used upon docking or as a post-filtering step.

TI: All the residues from the true interface used as restraints during docking. True interface residues are calculated at 3.9Å distance.
$BLIND_{filtered}$: All the residues from the true interface are used as restraints for post-filtering BLIND predictions.
BLIND: Ab initio docking.
The results are presented according to the CAPRI quality criteria[147] and the success rate is defined as the percentage of cases with at least one acceptable or higher quality model within a given Top N (N = 1, 5, 10, 20, 50 100).



While several docking algorithms allow the use of information as *a posteriori* filter, LightDock incorporates this data *a priori*. If residue restraints are provided, irrelevant sampling regions are excluded by filtering the initial *swarms* and pre-orienting the initial poses *(glowworms)*. This method not only represents a more efficient way as compared to post-docking approaches but also leads to a higher success rate. To test this hypothesis, we have filtered the *BLIND* predictions *($BLIND_{filtered}$)* according to an accurate description of the interface (residue restraints as used in *TI*). As shown in Figure 2, post-filtering results in a clear improvement of the performance compared to ab initio docking. Nevertheless, when using this information prior the docking

(TI), the success rate considerably increases reaching a maximum of 98.2% for the Top50 (54 of 55 cases) compared to a moderate 40% in $BLIND_{filtered}$. Figure 3 shows the results for the eight scenarios described above together with *ab initio* docking, which is included as a baseline for comparison purposes. The scoring function used in these LightDock simulations is DFIRE[148]. When no prior information about the binding site is used for the docking calculations *(BLIND)*, the predictive performance of LightDock lags behind any of the other scenarios tested in this work, with a moderate 14.5 and 23.6% success rates for Top10 and Top100 respectively.

Interestingly, with the gradual use of information in the form of residue contact restraints, we find a boost in the performance up to a 92.7% for the Top10 when an accurate description of the interface *(TI)* is used. This represents an ideal case and illustrates how docking approaches can enormously benefit from integrating experimental data in their calculations. Unfortunately, structural experimental techniques rarely describe interfaces in a very accurate manner and the data produced is usually incomplete and/or incorrect, fact that heavily affect the performance of modelling approaches as previously discussed[118].

To account for inaccurate or incorrect data, we have designed artificial interfaces with false positive residues. When only 50% of the original *TI* is used *(TI50)* or 25% *(TI25)*, which represents 50 and 75% of non-interfacial residues, LightDock performance in Top10 is of 72.7 and 46.4% respectively. In the case of TI_{50}, Top100 performance compares to *TI* (94.6% versus 98.2%). This indicates that even when the information used to restrain the docking simulations in LightDock is incomplete and partially wrong, the protocol seems robust enough and still yields correct solutions for most of the cases (52 out of 55). However, the scoring becomes problematic compared to *TI* as the Top1 success rate drops from 65.5 to 33.6%.

In the scenario where only the contribution of the receptor is taken into account *(TI_{REC})*, a substantial success rate of 67.3% is obtained for the Top100. This scenario is especially interesting since it directly applies, for example, to antibody-antigen docking where no information about the epitope is known so the docking is performed exploring the whole surface of the antigen while for the antibody the HV loops are provided (see Figure 4). Moreover, when false positives are included in the *TI_{REC}* scenario (50% in *TI_{REC-50}*, 75% in *TI_{REC-25}*) the performance drops, but Top100 is still higher (46.3 and 28.2%) than *BLIND* (23.6%).

Figure 3. Performance of LightDock for the nine different scenarios.

BLIND: Ab initio docking. **TI_{REC}**: Only receptor contribution to the true interface. **TI:** All the residues from the true interface. **TI_{SINGLE}:** A single residue pair from the true interface. **TI_{REC-50}:** Half of the **TI_{REC}** and equal number of non-interfacial residues. **TI_{50}:** Half of the TI and equal number of non-interfacial residues. **TI_{ONE}:** Only one residue on the receptor, as defined in **TI_{SINGLE}**, is considered as restraint (i.e. no information on the ligand side). **TI_{REC-25}:** One fourth of the **TI_{REC}** and three times more non-interfacial residues. **TI_{25}:** One fourth of the **TI** and three times more non-interfacial residues. True interface residues are calculated at a cutoff distance of 3.9 Å. Results are presented according to the CAPRI quality criteria[147] and the success rate is defined as the percentage of cases with at least one non-incorrect model within a given Top N (N = 1, 5, 10, 20, 50, 100).

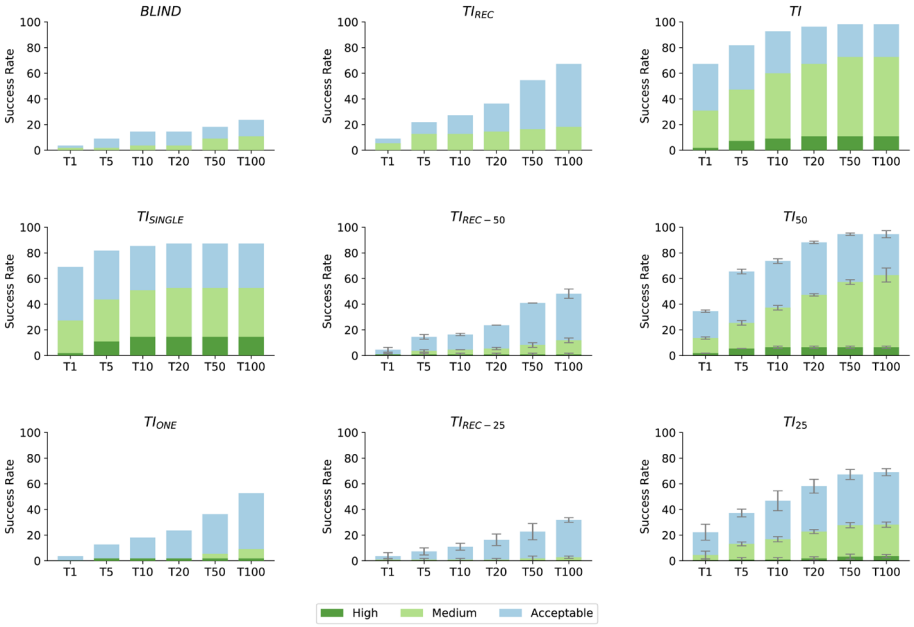
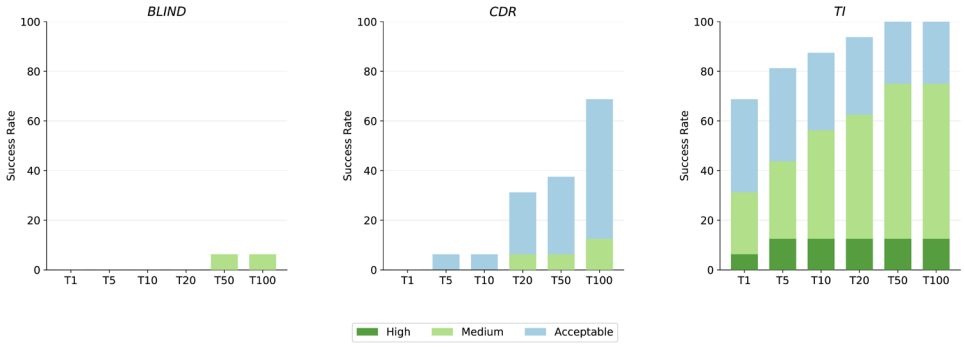


Figure 4. Performance of LightDock for three different scenarios in a subset of 16 antibody-antigen complexes of the docking benchmark 5.

BLIND: Ab initio docking.
CDR: Only antibody CDR loops residues are considered as restraints (receptor).
TI: All the residues from the true interface.
True interface residues are calculated at 3.9Å distance. We defined the CDR loops as in[149]. The results are presented according to the CAPRI quality criteria[147] and the success rate is defined as the percentage of cases with at least one acceptable or higher quality model within a given Top N (N = 1, 5, 10, 20, 50 100). Results here presented represent an important increase of success rate compared to the ones in[149], as current version (0.7.0) accounts for pre-orientation of ligand poses if information is available in contrast to version 0.5.6 where that information was not yet considered into the LightDock protocol.



Finally, we push the limits of the algorithm defining only one residue restraint on the receptor molecule (this would mimic one mutation data point for example). This effectively means that, as in TI_{SINGLE}, only the ten closest swarms to the restraint will be generated, each of them containing randomly oriented glowworm poses (200 by default). In this scenario, restricting the sampling area helps the identification of near-native models as the performance is significantly higher than *BLIND* (Figure 3). Remarkably, when we include a residue on the ligand molecule *(TI_{SINGLE})*, which is used in the pre-orienting step, LightDock predicts and scores a near-native solution in the Top1 for 69% of the cases. From the different tested scenarios, it seems reasonable

to state that our protocol enormously benefits from the additional data in form of residue restraints, even when it is incomplete and/or partially incorrect.

4. Conclusion

The new version of LightDock offers a powerful tool for modelling protein–protein complexes with high accuracy when good quality information about interfaces is available. Next to enabling the incorporation of data from mutagenesis and/or bioinformatics predictions, for example, this strategy might also be convenient in scenarios such as limiting the sampling to the solvent accessible loops of a transmembrane protein, or the CDR loops of an antibody. Moreover, when incorrect and/or incomplete data are used to restraint the simulation, LightDock is still robust enough to yield valuable predictions. While other FFT-based methods do support *a posteriori* filtering, the pre-filtering of *swarms* in LightDock does lead to a reduction of the computation time and a higher performance, which could be used to ensure a denser sampling around the binding region.

Chapter 3

Less is more: Coarse-grained integrative modeling of large biomolecular assemblies with HADDOCK

Jorge Roel-Touris, Charleen G. Don, Rodrigo V. Honorato, João P.G.L.M Rodrigues, Alexandre M.J.J. Bonvin

Published in 2019 in Journal of Chemical Theory and Computation, Volume 15, Issue 11, Pages 6358 – 6367

Abstract

Predicting the 3D structure of protein interactions remains a challenge in the field of computational structural biology. This is in part due to difficulties in sampling the complex energy landscape of multiple interacting flexible polypeptide chains. Coarse-graining approaches, which reduce the number of degrees of freedom of the system, help address this limitation by smoothing the energy landscape, allowing an easier identification of the global energy minimum. They also accelerate the calculations, allowing for modeling larger assemblies. Here, we present the implementation of the MARTINI coarse-grained force field for proteins into HADDOCK, our integrative modeling platform. Docking and refinement are performed at the coarse-grained level, and the resulting models are then converted back to atomistic resolution through a distance restraints-guided morphing procedure. Our protocol, tested on the largest complexes of the protein docking benchmark 5, shows an overall 7-fold speed increase compared to standard all-atom calculations, while maintaining a similar accuracy and yielding substantially more near-native solutions. To showcase the potential of our method, we performed simultaneous 7 body docking to model the 1:6 KaiC-KaiB complex, integrating mutagenesis and hydrogen/deuterium exchange data from mass spectrometry with symmetry restraints, and validated the resulting models against a recently published cryo-EM structure.

1. Introduction

Proteins are the workhorses of the cellular machinery. In order to function, they bind to one another, as well as to other biomolecules, to form large molecular assemblies. These interactions play a key role in all essential molecular processes within a cell. Most of these assemblies may exist as transient associations, which, together with other experimental factors, makes the characterization of their three-dimensional (3D) structure a challenge[15] for experimental metods such as nuclear magnetic resonance (NMR) spectroscopy or X-ray crystallography[151,152]. Despite recent advances in cryo-electron microscopy (cryo-EM), it is unlikely that the substantial gap between the number of estimated protein–protein interactions and those deposited in the Protein Data Bank[153] can be overcome based solely on experimental methods[154].

Computational docking has come of age as a complement to experimental methods in order to generate 3D models of protein assemblies. In particular, data- or information-driven docking and other integrative approaches are particularly appealing[118,150,155,156]. While docking performs sufficiently well for small- and medium-sized proteins, applications to large biological systems, either containing large individual molecules or a large number of interactors, are limited by the significant computational cost of thoroughly sampling complex conformational landscapes. Coarse-grained (CG) models mitigate this limitation by grouping atoms into larger pseudo atoms or beads[157–159], thus reducing the number of particles to consider in the computations. These models were used in the very first energy minimization of a protein in 1969[160] and again in the first docking simulation[6].

Since then, several CG models have been developed and applied to study different aspects of protein structural biology[26]. For protein

docking in particular, of the CG models developed over the years, three stand out for their performance and/or success in community assessment experiments: Those implemented in ATTRACT, CABS-dock, and RosettaDock. The ATTRACT model[81,140], developed by Zacharias and co-workers for flexible protein docking, represents the protein backbone by two pseudo atoms and the side chains by an additional particle (or two in the case of larger amino acids). Nonbonded interactions are described by 8–6 LJ potentials and a Coulomb type term[161], with parameters systematically optimized on both existing structures of protein–protein complexes as well as on docked models. As such, this limits the transferability of ATTRACT to other systems, such as protein-nucleic acid complexes or membrane proteins.

Another model, CABS (Cα-Cβ-Side group protein model), was originally developed for structure prediction of globular proteins[82] and later applied to protein-peptide docking[162] (CABS-dock). As in ATTRACT, protein residues are represented by a maximum of four particles: Cα, Cβ, side chain, and an extra particle representing a virtual Cα–Cα bond. Knowledge-based statistical potentials are used to describe particle interactions. The performance of CABS-dock was benchmarked on a set of protein-peptide complexes[83], with peptides of 5–15 residues in length yielding accurate predictions. Although there are no technical limitations to the application of CABS-dock to larger protein–protein systems, except the increase in computational time, this application has not been reported in the literature to date, and its performance remains thus uncertain. Moreover, given the specificity of its parameters to proteins, much like ATTRACT, the transferability of CABS to other molecular systems might be limited.

Finally, RosettaDock implements a two-step protocol with a coarse-grained global search followed by an all-atom refinement[92]. In the

coarse-grained step, the interacting proteins are represented by their backbone atoms and a single pseudo atom for the side chain. The resulting models are ranked using a combination of residue pairwise interaction terms, a contact-based term, and a term that penalizes overlapping residues. The all-atom refinement step uses the full Rosetta scoring function. As such, in the case of large assemblies, RosettaDock benefits from a smoother energy landscape during the conformational sampling, but the second all-atom refinement stage is computationally expensive.

On the other hand, some CG models were developed to be easily transferrable. MARTINI, a CG model for biomolecules, was originally applied to study lipid bilayer assembly[5] and later extended to proteins[41], carbohydrates[44], and nucleic acids[16,47]. This model maps, generally, four heavy atoms onto one coarse-grained bead. Its corresponding force field parameters have been calibrated to reproduce thermodynamic measurements. Systems are represented by 4 different basic particle types –polar (P), nonpolar (N), apolar (C), and charged (Q)– that are further divided based on their hydrogen-bonding properties and their degree of polarity, giving a total of 18 unique "building blocks". In addition to the 4 standard types of beads, the 2.2p version of MARTINI includes off-center charges for polar and charged amino acids. These extra "fake beads" improve the description of interactions between charged residues (ARG, LYS, ASP, GLU) and provide directionality/orientation in the case of polar residues, mimicking to some extent hydrogen bonds (e.g., an ASN side-chain bead has two "fake beads" associated carrying a small positive and negative charge, respectively). In addition, the MARTINI model is able to represent several types of molecules and allows for a straightforward conversion to atomistic resolution, making it ideal to use in HADDOCK for integrative modeling applications.

Here, we describe the implementation of the MARTINI CG force field for proteins[14] in our information-driven docking software HADDOCK[8]. We evaluated the performance of the coarse-grained HADDOCK protocol using the largest complexes from the protein docking benchmark 5[145], comparing it to the standard all-atom protocol. The performance increase from using a smaller set of particles to describe the molecular system allows for a substantial decrease in computational time, enabling the modeling of larger systems. As a demonstration, we modeled the heptameric KaiC-KaiB 1:6 assembly, which is part of the endogenous biological clock in cyanobacteria[163,164], by performing a simultaneous 7 body docking, guided by mass spectrometry (MS) and mutagenesis data in combination with symmetry restraints.

2. Methods

2.1. Implementation of MARTINI in HADDOCK

The integration of the MARTINI CG force field for proteins into HADDOCK focused on three key aspects: (1) converting the topology description and parametrization for each amino acid in a format suited for HADDOCK and its computational engine CNS (Crystallography and NMR System[78,165]), (2) adapting the atomic solvation parameters[166] used to calculate the desolvation energy in HADDOCK to the CG particles, and (3) developing a protocol to convert the coarse-grained system back to atomistic resolution after the semiflexible refinement stage of HADDOCK, making use of distance restraints derived from the MARTINI atoms-to-bead mapping.

As in standard MARTINI, four types of interaction sites are considered: polar (P), nonpolar (N), apolar (C), and charged (Q). The conversion of the backbone to the CG beads follows a four-to-one (4:1) mapping rule, where all four heavy atoms (N, Cα, C, O) are represented by a single bead placed at

their geometric center. The conversion of side chains varies, ranging from the same 4:1 mapping to 2:1 mapping and “small” beads in rings (HIS, PHE, TYR, TRP). We converted the topology and corresponding parameters of MARTINI 2.2p to a format compatible with CNS (see Tables SI3.1–4 in Supplementary Information). The force field, however, does not account for either the various possible histidine charge states (i.e., neutral with the proton on either the δ or ε nitrogen atom or doubly protonated and positively charged) nor for nonstandard residues (e.g., amino acids with post-translational modifications) or cofactors.

Since the amino acid backbone parameters in MARTINI are secondary structure-dependent, we use DSSP[121,167] to analyze the initial structures and encode the secondary structure in the B-factor field. Using the later information HADDOCK automatically selects the proper parameters for each backbone bead in the coarse-grained structures when building the topology of the system. This effectively restrains the existing secondary structures, which might be a limitation for docking cases with large conformational changes between the unbound and bound states. However, if no secondary structure information is encoded in the B-factor field, random coil parameters allowing for possible conformational changes will apply. Note that in contrast to standard molecular dynamics simulations of proteins using the MARTINI force field, no Go terms are used in HADDOCK since only the interface is refined, and therefore the majority of the structure is kept rigid by default. Nonbonded CG interactions are calculated using a 14 Å cutoff, as recommended, while interactions between atoms in the final stage are calculated using the OPLS force field[168] parameters with the default 8.5 Å cutoff used in HADDOCK.

2.2. Solvation parameters for the coarse-grained particles

The HADDOCK score, used to rank the predicted models, is a linear combination of energetical and empirical terms (see *Scoring* below), including a solvent-accessible surface-based desolvation energy term[166] *(E_{desolv})*. In order to score CG models using this desolvation energy, we mapped the atomistic solvation parameters onto the CG beads. For this, the solvation energy of each group of atoms belonging to a specific bead was calculated for all 20 amino acids *X* in a *GGXGG* peptide. Since the solvation energy depends on the solvent accessible surface area of an atom/bead, the total atomistic energy was divided by the solvent accessible surface area of the corresponding CG bead in a similar peptide in order to obtain the CG solvation parameters $\boldsymbol{SP^i_{cg}}$ for a specific CG particle ***i*** (eq. 1).

$$SP^i_{cg} = E^i_{desolv} / ASA^i_{cg} \qquad \text{(eq. 1)}$$

where is the atomistic solvation energy for the group of atoms belonging to a given bead ***i*** and is the accessible surface area of that bead in the *GGXGG* peptide. The all-atom and CG solvent accessible areas were calculated using CNS with an accuracy of 0.0025 using a water radius of 1.4 Å excluding all hydrogen atoms. The so-called “fake beads” are not included in the desolvation energy calculation. The resulting solvation parameters values for the MARTINI CG beads are listed in Table 1.

2.3. Pre-processing of input structures for coarse-grained docking

Setting up a CG docking run requires first converting the coordinate files, which contain information on individual atoms, into a CG representation. To this end, we adapted the *“martinize1.1.py”* (https://github.com/Tsjerk) to account for the name type extensions (i.e., *“fake beads”* present in the 2.2p

version of MARTINI) and to additionally generate distance restraints, in CNS format, between the original atoms and the corresponding CG beads, which are used in the final back-mapping stage of the protocol (see *Back-Mapping Coarse-Grained Models to Atomic Resolution by Distance Restraints* below). Since the MARTINI backbone parametrization depends on the local secondary structure, we numerically store the secondary structure assignments computed by DSSP[121,167] into the B-factor column of the resulting CG PDB files. As in the standard protocol, HADDOCK automatically builds any missing atom when creating both the topology and coordinate files from the user-provided PDB files. This procedure is done both for the starting CG and all-atom structures. The latter are used in the final back-mapping stage from CG to all-atom.

2.4. Backmapping coarse-grained models to atomic resolution by distance restraints

In order to convert the final coarse-grained models back to an all-atom representation, we make use of the ability of HADDOCK to use distance restraints to guide the modeling, using the atom-to-bead distance restraints derived during the initial setup stage. For a group of atoms belonging to a particular CG bead, we create one distance restraint with 0 length between the geometric center of the atoms and the bead to which they belong. The conversion protocol consists of the following steps:

1. *Initial fitting onto the CG model*

The all-atom structure of each molecule of the complex is fitted onto its respective CG representation in the docked CG model by rigid body energy minimization (EM) guided by the CG-to-AA distance restraints. During this

step the CG model is kept fixed, and the intermolecular interactions are scaled by a factor 0.001 to account for possible clashes between the AA molecules. No energy terms are included for the CG model, except the distance restraining potential.

Table 1. Coarse-grained solvation parameters for each amino acid, mapped from the all-atom empirical solvation parameters onto MARTINI beads.

BB: backbone beads. **SC*:** any side-chain bead. Note that "fake beads" (SCD) are not considered.

Amino Acid	Solvation Parameter	
	BB	SC*
ALA	-0.0107	-
GLY	-0.0089	-
ILE	-0.0153	0.0255
LEU	-0.0153	0.0243
VAL	-0.0158	0.0222
PRO	-0.0046	0.0230
ASN	-0.0137	-0.0192
GLN	-0.0147	-0.0135
THR	-0.0165	-0.0009
SER	-0.0154	-0.0056
MET	-0.0130	0.0202
CYS	-0.0167	0.0201
PHE	-0.0126	0.1005
TYR	-0.0134	0.0669
TRP	-0.0134	0.0872
ASP	-0.0169	-0.0360
GLU	-0.0150	-0.0301
HIS	-0.0155	0.0501
LYS	-0.0163	-0.0210
ARG	-0.0162	-0.0229

2. *Inducing conformational changes*

In order to morph the all-atom structure onto the CG model, which might have undergone conformational changes during the flexible stage of the docking protocol, we first perform two short rounds of energy minimization (500 steps), increasing the scaling factor for intermolecular interactions to 0.01 after the first minimization. Then, we perform 500 steps of Cartesian molecular dynamics (MD) at 300 K with an integration time step of 0.0005 ps and another round of EM.

3. *Clearing clashes and optimizing all-atom interactions*

We perform two rounds of energy minimization, increasing the scaling factor of the intermolecular interactions to 0.1 and 1.0, respectively, followed by another short MD (500 integration steps) and two extra minimization rounds.

In all three steps, all covalent and noncovalent energy terms are included for the AA models together with the restraint energy term for the atom-to-bead distance restraints. Once the all-atom models have been generated, the CG models are discarded, the morphing distance restraints are removed, and all other restraining energy terms representing the various data given to HADDOCK to drive the docking are reintroduced. These are used in a final round of energy minimization. Although computationally expensive for large systems, the user can then choose to follow-up with the full water refinement stage of the standard HADDOCK protocol (turned off by default).

2.5. Docking procedure

All docking calculations were performed using a local installation of the

new HADDOCK version 2.4 supporting CG docking. This protocol is also supported by the new version of our Web server[116] soon to be released. For comparison purposes, the docking was performed both with all-atom and coarse-grained representations, using the united-atom OPLS force field[168] and MARTINI 2.2p, respectively. The docking was guided by ambiguous interaction restraints (AIRs) derived from the bound complexes (true interface) by selecting all solvent accessible residues with at least one heavy atom within 3.9 Å from any heavy atom of the partner molecule. These restraints represent an ideal scenario where accurate information is available about the residues in the interface but not about their specific pairwise contacts (information that can be obtained, e.g., from NMR chemical shift perturbations, mass spectrometry hydrogen/deuterium exchange, ...)[118,156].

The sampling parameters were kept as default in HADDOCK: 1000/200/200 models were generated for the rigid body (it0), semiflexible (it1), and water refinement (itw) stages, respectively. In the CG runs, the final water refinement stage was replaced by the back-mapping from CG to all-atom as shown in Figure 1. The final models were clustered based on the fraction of common contacts (FCC)[169] using a 0.6 cutoff and a minimum number of 4 models per cluster.

2.6. Scoring

We investigated whether reparametrizing the HADDOCK-CG score led to a better scoring performance by systematically varying the weights of the scoring function. Since we did not observe significant improvements (data not shown), we kept the original HADDOCK scoring functions (HS) for the three stages of the docking protocol (rigid-body EM *(it0)*; semiflexible refinement *(it1)*; explicit solvent refinement *(itw)*):

$$HS_{it0} = 0.01*E_{vdw} + 1.0*E_{elec} + 0.01*E_{AIR} + 1.0*E_{desolv} - 0.01*BSA$$

$$HS_{it1} = 1.0*E_{vdw} + 1.0*E_{elec} + 0.1*E_{AIR} + 1.0*E_{desolv} - 0.01*BSA$$

$$HS_{itw} = 1.0*E_{vdw} + 0.2*E_{elec} + 0.1*E_{AIR} + 1.0*E_{desolv}$$

where E_{vdw} and E_{elec} are the van der Waals and electrostatic energy terms calculated using a 12-6 Lennard-Jones and Coulomb potential, respectively, with MARTINI *(it0, it1)* or OPLS *(itw)* nonbonded parameters, E_{AIR} is the ambiguous interaction restraints energy, E_{desolv} is the empirical desolvation score, and BSA is the buried surface area in $Å^2$.

2.7. Protein docking benchmark

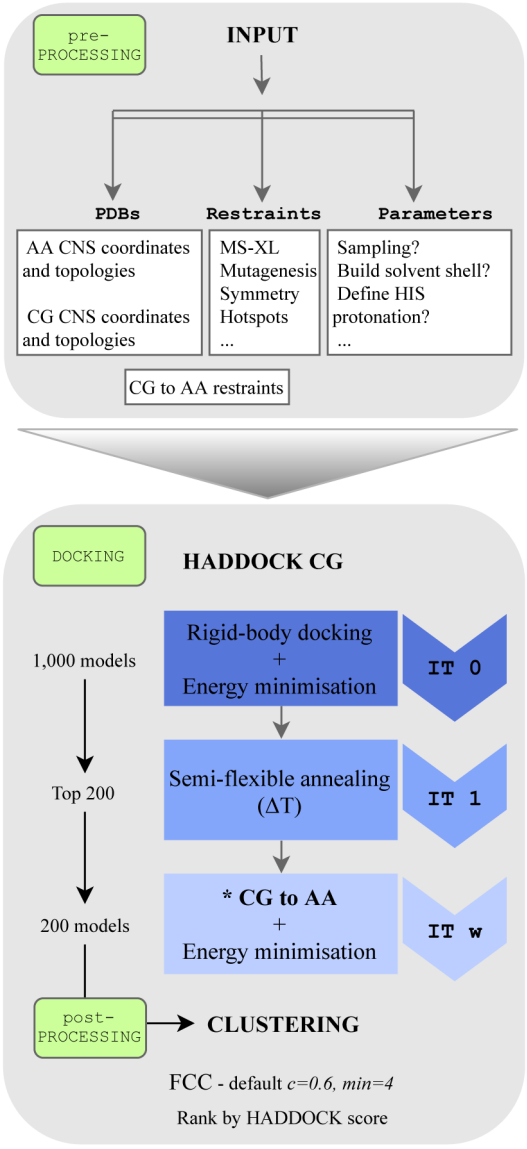
To test the performance of our HADDOCK-CG protocol, we selected a subset of complexes from the Protein–Protein Docking Benchmark version 5.0[145], consisting of all complexes with more than 5,000 heavy atoms, excluding all antibody–antigen cases. This selection yielded a benchmark set of 27 cases (see Table SI3.5 in Supplementary Information).

2.8. Metrics for the evaluation of docking success rate

The performance of the docking calculations was analyzed as follows: (1) The percentage of cases in which at least one model of a given accuracy is found within the top *N* solutions ranked by HADDOCK (*N* = 1, 5, 10, 20, 25, 50, 100, 200), and (2) the percentage of cases in which at least one acceptable or higher quality model was found in the top *T* clusters (*T* = 1, 2, 3, 4, 5).

Figure 1. HADDOCK coarse-grained flowchart.

AA = all-atom
CG = coarse-grained
FCC = fraction of common contacts.
***** Back-mapping coarse-grained models to atomic resolution by distance restraints.



2.9. Metrics for evaluation of model quality

The quality of the generated models was evaluated using standard CAPRI[147] criteria, including the fraction of native contacts (FNAT) and the positional interface (i-RMSD) and ligand (l-RMSD) root-mean-square deviations from the reference crystal structure. FNAT is calculated using all heavy atom – heavy atom intermolecular contacts using a 5 Å distance cut-off (CAPRI definition). The i-RMSD is calculated on the interface after superimposition on the interface residues, defined as those with any heavy atom within a 10 Å distance of the partner protein. The l-RMSD is calculated on the ligand (usually the smallest molecule) after superimposition on the backbone atoms of the receptor (largest molecule). For both, i-RMSD and l-RMSD, only backbone heavy atoms are considered (Cα, C, N, O). Based on these three metrics, the quality of the docking poses is classified as:

- High: FNAT ≥ 0.5 and i-RMSD ≤ 1 Å or l-RMSD ≤ 1 Å,
- Medium: FNAT ≥ 0.3 and 1 Å < i-RMSD ≤ 2 or 1 Å < l-RMSD ≤ 5 Å,
- Acceptable: FNAT ≥ 0.1 and 2 Å < i-RMSD ≤ 4 or 5 Å < l-RMSD ≤ 10 Å,
- Near-Acceptable: FNAT ≥ 0.1 and 4 Å < i-RMSD ≤ 6 Å, and
- Low quality: FNAT < 0.1 or i-RMSD > 6 Å or l-RMSD > 10 Å.

2.10. KaiC-KaiB coarse-grained integrative modeling with HADDOCK

In order to model the KaiC:KaiB 1:6 complex, we performed two different docking runs, targeting either the CI or CII domains on KaiC since the H/D exchange data from MS point to two possible interfaces (for details refer to Snijder et al.[170]). We used the crystal structure of KaiC (PDB ID: *3dvl*) consisting of 12 domains (two 6-membered rings) as a starting point for the docking. For KaiB, we used six copies of the

recent NMR structure (PDB ID: *5jyt*)[171], which shows a fold-switch at the interacting region compared to the previously determined crystal structure[172].

The regions experimentally identified by HDX-MS as protected from solvent in either the CI or CII domains of KaiC and in KaiB were specified as active residues in HADDOCK, after filtering them for solvent accessibility (relative residue solvent accessibility larger than 50% as calculated with NACCESS[173]) (see Table SI3.6 in Supplementary Information, for a detailed list of residues). For KaiB, we included three additional residues identified by mutagenesis experiments. A structural similarity analysis of KaiC revealed an asymmetrical structure with RMSD values for the interface regions between subunits in the hexamer ranging from 0.9 to 1.9 Å (see Table SI3.7 in Supplementary Information, for more details). As a result, we restrained the KaiB monomers to an approximate C6 symmetry by defining three C2 symmetry pairs (B-E/C-F/D-G) and two C3 symmetry triplets (B-D-F/C-E-G), but we did not use non-crystallographic symmetry restraints (NCS) since the interfaces are asymmetrical.

Because of the symmetry restraints, sampling of 180° rotations during the rigid-body stage was disabled. Furthermore, given the large size of the complex and the number of subunits to dock (7-body docking), the sampling was increased to *10000/400/400* models for *it0/it1/itw*, respectively. Finally, we disabled the final refinement in explicit water, only performing the back-mapping from CG to all-atom (as part of the default HADDOCK-CG pipeline). We only used the top 200 models according to the HADDOCK score for analysis and validation purposes.

3. Results and Discussion

We have integrated the MARTINI 2.2p force field for proteins into HADDOCK (see Methods; *Implementation of MARTINI in HADDOCK*), adapted the desolvation energy terms to the coarse-grained beads, and developed a distance restraints-based back-mapping method to restore the atomic resolution of the final models while accounting for possible conformational changes that took place during the CG semiflexible refinement step. In the following sections, we discuss the performance of our protocol in terms of success rate, sampling, and computational efficiency using the 27 largest complexes from the docking benchmark 5. We then showcase its potential by modeling a large heptameric complex using mass spectrometry and mutagenesis data.

3.1. Overall performance of coarse-grained HADDOCK

We compared the unbound docking performance of HADDOCK-CG with the default all-atom protocol for 27 binary complexes from the Docking Benchmark 5 (see Methods; *Protein docking benchmark*). Fourteen of those complexes were classified as easy according to the structural differences between the bound and unbound structures of the monomers, 8 as medium, and 5 as hard. The docking was performed starting from the unbound structures of each protein and driven by information from the real interface (see Methods; *Docking procedure*), mimicking an ideal scenario for HADDOCK users. The success rate was defined as the percentage of cases for which an acceptable or better model was obtained in the top *N* ranked models (for details see Methods; *Metrics for evaluation of success in docking*).

Coarse-grained docking shows a slightly better overall performance (Figure 2) in the top 1 for single structure ranking (best ranked structure) than

the standard all-atom protocol, with success rates for acceptable or higher quality models of 51.8% and 48.1%, respectively. However, this trend reverses for the performance in the top 5, with 66.6% and 77.7% success rates for coarse-grained and atomistic models, respectively. For the remaining top *N* (*N* = 10, 20, 50, 100, 200), the performance of HADDOCK-CG is comparable with that of all-atom calculations, reaching a maximum of 92.5% at *N* = 200. For the two cases with the largest conformational change (i-RMSD values of 4.69 Å/5.79 Å between unbound and bound structures for *1y64/4gam,* respectively), neither coarse-grained nor all-atom calculations generated near-acceptable solutions.

We also analyzed the success rate on a per-cluster basis, which is the standard scoring scheme of HADDOCK. Clustering models improve the success rate for both coarse-grained and all-atom simulations to 59.2% and 51.8%, respectively, for the top 1 cluster. The success rate is maximal for the top 5 clusters reaching 88.8% for acceptable or higher quality models (Figure 2B). The all-atom protocol reached the maximum success rate (88.8%) at the top 4 clusters. Compared to single structure scoring, no near-native cluster was obtained for *1ib1* due to the fact that only 3 models passed the quality thresholds and our clustering strategy requires a minimum of 4 models per cluster.

Concerning the quality of the models (see Methods; Metrics for evaluation of model quality), the all-atom runs generated higher quality solutions than CG runs (Figure 2C and 2D). For the easy cases, all-atom runs rank medium quality models in the top 10 solutions for 10 out of 14 cases and acceptable quality models for 13 out of 14 cases. For the CG runs, medium quality models are obtained in the top 10 solutions for 7 out of 14 easy cases, and acceptable quality models are obtained for all 14 cases. As for the intermediate and hard cases, the all-atom runs generate medium quality

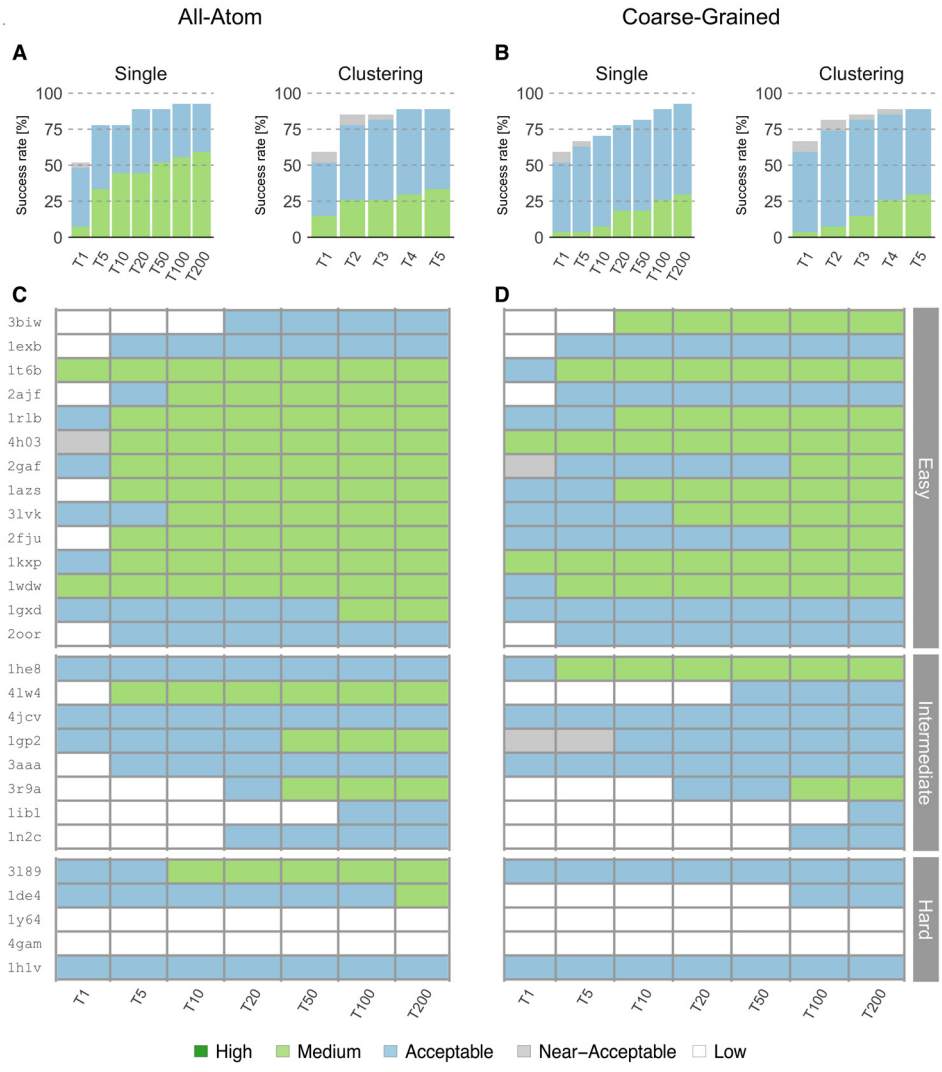
models for only 5 out of 13 cases, while CG runs generate them in 2 cases. Overall, coarse-grained HADDOCK generated medium quality solutions for 12 out of all 27 complexes including intermediate cases, slightly worse than the 16 cases for the all-atom run.

Interestingly there are 2 cases where CG docking generates better quality models than all-atom runs. For *3biw*, an easy case, coarse-grained docking generated medium quality models ranked in the top 10. The best of these models has an FNAT of 0.61 and i-RMSD of 1.9 Å, compared to an FNAT of 0.52 and i-RMSD of 3.5 Å for the all-atom run. For *1he8*, a medium difficulty case, we found a medium quality model in the top 5 with an FNAT of 0.55 and l-RMSD of 4.9 Å, while the best all-atom model has an FNAT of 0.44 and l-RMSD of 6.1 Å.

Given the back-mapping to all-atom resolution at the end of the coarse-grained protocol, we also evaluated the quality of the final models in terms of the number of atomic clashes at the interface. A clash was defined as any pair of heavy atoms belonging to different molecules within 3 Å distance, in accordance with the CAPRI assessment procedure. The number of clashes was then divided by the buried surface area of the complex, and models with more than 0.1 clashes/Å^2 were considered of poor quality. We found no model, in both CG and all-atom runs, that scored under this clash threshold. However, and interestingly, docked structures generated via coarse-graining presented, on average, half the clashes of the models from the all-atom runs, which might be explained by the multiple energy minimization rounds performed during the back-mapping protocol, compared to the default water refinement protocol.

Figure 2. Performance of the all-atom and coarse-grained protocols in HADDOCK on the 27 largest complexes of the Docking Benchmark 5.

A) Overall success rates (%) of the all-atom protocol on ranking single models (Single) or clusters (Clustering) as a function of the number of models/clusters considered.
B) Same as **(A)** but for the coarse-grained protocol.
C) and **(D)** Quality of the docking models for all 27 cases as a function of the number of models considered.
The complexes are ordered by increasing degree of difficulty (from top to bottom) for both all-atom and CG docking runs. The color coding indicates the quality of the docked models.



3.2. Reduction of the energy landscape complexity

A product of coarse-graining is a smoothening of the energy landscape, which should allow for an easier sampling compared to all-atom calculations. The coarse-grained landscape might help find energy minima, especially in cases where only few or no data are available to drive the modeling and should, therefore, contribute to a better performance of coarse-grained docking runs (i.e., an increase in the number of near-acceptable models). To test this hypothesis, we performed docking without any experimental information, using the *ab initio* mode of HADDOCK in which, for each docked model, pairs of residues on the interacting molecules are randomly selected and ambiguous interactionrestraints are defined between surface patches within 7.5 Å of those residues. In order to test whether coarse-graining improves sampling, we ran our benchmark with this type of random restraints for both all-atom and coarse-grained protocols, increasing in both cases the sampling to *10000/400/400* models for *it0/it1/itw.* We indeed observe (Table 2) a substantial increase (28.4%) in the number of models of acceptable or better quality during the rigid body stage of coarse-grained docking, compared to all-atom simulations. However, when using interface data to drive the calculations, this difference decreases to 8% more acceptable or higher quality models for the coarse-grained protocol, which is still a substantial improvement.

3.3. Computational performance

The main motivation to implement a coarse-grained force field in HADDOCK was to accelerate and enable the modeling of large biomolecular assemblies by reducing the number of particles considered during the computations. The *atom-to-bead* mapping of the MARTINI model leads to a significant reduction in the number of particles, making

the computations substantially more efficient. It was previously shown that MARTINI allows for an increase in computational efficiency by a factor 2 to 4 compared to common all-atom models[41]. In our case, integrating MARTINI into HADDOCK led to an average 7-fold speed-up in total computation time (Table 3).

Table 2.

Comparison of the total number of acceptable or higher quality models, generated over all 27 complexes at the rigid-body stage (it0), between coarse-grained and standard all-atom HADDOCK protocols in the absence of information to drive the docking (ab Initio mode) and using true interface information.

	Top 200	Top 400	Total	Ratio CG/AA
*Ab initio docking (Random patches) **				
Coarse-grained	15	16	74	**1.39**
All-atom	11	13	53	
True Interface docking				
Coarse-grained	2666	5066	9689	**1.08**
All-atom	2702	4940	8896	

** 10000 models were generated in the case of ab initio docking. For details, see Tables SI3.10–11 in Supplementary Information*

Table 3.

Comparison of average CPU times # (seconds/model) for the test benchmark
(N = 27) between the all-atom and coarse-grained HADDOCK protocols.

	it0	it1	* itw	<Ratio> AA/CG
All-atom	22.2 ± 19.8	1327.2 ± 1077	1577.4 ± 975	**6.78 ± 1.3**
Coarse-grained	2.4 ± 1.2	165.6 ± 134.4	276 ± 198.6	

 # The timings correspond to the total time reported by CNS as measured on an AMD Opteron (tm) Processor 6344.
** The coarse-grained protocol does not include refinement in explicit solvent but instead performs a back-mapping procedure to restore all-atom resolution to the final models.*

3.4. Coarse-grained integrative modeling of KaiC-KaiB

To demonstrate our coarse-grained HADDOCK protocol, we modeled the heptameric KaiC-KaiB (stoichiometry 1:6) complex by simultaneous 7 body docking using data from mutagenesis experiments and hydrogen–deuterium exchange MS[170]. The structures of KaiC and KaiB have been both characterized individually at the atomic level. KaiC forms hexamers and consists of two domains, CI and CII[174,17]. It has been shown that six KaiB monomers bind to one KaiC hexamer[163].The first published model of this complex[170] wrongly pointed to CII as binding mode, based on better agreement with collision cross section data obtained by *time-of-flight* MS. Later on, the cryo-EM structure[176] of KaiCBA revealed a CI binding mode and a different fold of KaiB corresponding to the solution NMR structure (PDB ID *5jyt*) that was solved after the initial model was published.

This NMR structure, which is also the conformation found in the cryo-EM structure, shows a fold switch compared to the crystal structure (PDB ID *4kso*) that was used in the initial modeling. The crystal structure was the only available one at the time of the first modeling. The first model was built by docking one KaiB onto two domains of KaiC (out of the 12 domains in full KaiC). We repeated here this modeling, using this time the full KaiC structure and six copies of the binding competent KaiB conformation (the NMR structure). Two 7-body docking runs targeting the CI and CII binding interfaces were performed with HADDOCK-CG. Along with the experimental data, we imposed symmetry restraints (C3 + C2, as an approximation of C6) between the 6 KaiB components. The resulting models were scored and ranked according to the HADDOCK score (see Methods; *Scoring*), including an additional energy term for the symmetry restraints. The cryo-EM map *(EMDB-3603)* was used for independent validation of the models.

Using the new, binding-competent KaiB structure we clearly identify the CI binding mode as the right answer, with a significantly lower HADDOCK score than CII: –216.7 ± 13.2 au versus +44.5 ± 19 au for the best cluster of each run (see Table SI3.8 in Supplementary Information). This model obtained based on mutagenesis and mass spectrometry data is consistent with the recent cryo-EM model of the KaiC:KaiB:KaiA complex in a fully assembled state[176] with a I-RMSD of 3.6 Å, calculated over all six interfaces, for the best model of the top scoring cluster (for more details, see Table SI3.9 in the Supplementary Information). We further validated our model by quantifying its agreement with the published cryo-EM map of the complex *(EMDB-3603)* using Chimera: The correlation score of our model is 0.82, compared to 0.84 for the original cryo-EM backbone model (PDB ID *5n8y*) as shown in Figure 3.

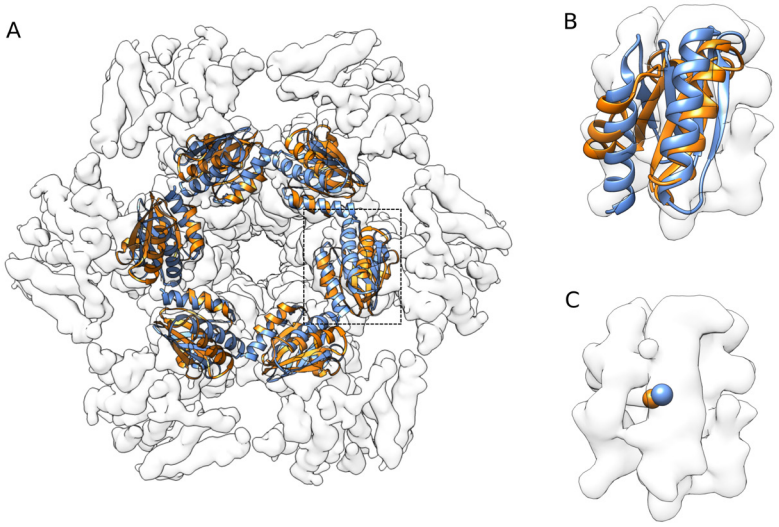
While the first all-atom model was obtained by docking a subset of the full complex, in this work we modeled here the full 1:6 KaiC-KaiB complex. By coarse-graining, we reduced the number of particles from 31726 in the original all-atom model to 9842 for the coarse-grained model, reducing the computational time by about a factor 6 times, from 4 h to 48 min, on average, per model.

4. Conclusions

In this work, we presented the integration of the MARTINI coarse-grained force field in our HADDOCK integrative modeling software. Our new docking protocol makes use of coarse-grained representations during the rigid body and semiflexible refinement stages and restores the final docked models to atomistic resolution in a final backmapping stage. By using distance restraints between beads and the atoms that belong to them, the back-mapping protocol is able to morph conformational changes that potentially took place during the coarse-grained flexible refinement.

Figure 3. Comparison of the cryo-EM model (PDB code: 5n8y, blue) and the best coarse-grained model obtained in this work (orange).

The models were fitted into the map using Chimera[177]. The correlation coefficient for our docked model is 0.82 compared to 0.84 for the cryo-EM structure. (A) Top view of the KaiB hexamer bound to KaiC CI domain. (B) Detailed view of single KaiB. (C) Comparison of centers of mass of a single KaiB monomer. Note that KaiA present in the cryo-EM model is not shown here.



The performance of coarse-grained docking is similar to that of the standard all-atom protocol in terms of success rate and quality of the generated models. In addition, it generates more near-native models when limited or no data are available and comes with the benefit of an 7-fold reduction in computing time. The power of our coarse-grained integrative modeling approach was demonstrated by modeling the structure of the heptameric KaiC:KaiB (1:6) complex, for which we obtained models in excellent agreement with the cryo-EM structure. In conclusion, the implementation of the MARTINI coarse-grained force field into HADDOCK extends its ability to model increasingly larger and more intricate biomolecular assemblies.

In the future, we plan to make use of the MARTINI models for lipids and nucleic acids and extend our protocol to allow modeling of nucleic acid complexes, as well as membrane and membrane-associated complexes, for which we recently published a new docking benchmark.

Chapter 4

MARTINI-based protein-DNA coarse-grained HADDOCKing

Rodrigo V. Honorato †, Jorge Roel-Touris †,
Alexandre M.J.J. Bonvin

Published in 2019 in Frontiers in Molecular Biosciences, Volume 6, Issue 102

† Joint first authors

Abstract

Modeling biomolecular assemblies is an important field in computational structural biology. The inherent complexity of their energy landscape and the computational cost associated with modeling large and complex assemblies are major drawbacks for integrative modeling approaches. The so-called coarse-graining approaches, which reduce the degrees of freedom of the system by grouping several atoms into larger “pseudo atoms” have been shown to alleviate some of those limitations, facilitating the identification of the global energy minima assumed to correspond to the native state of the complex, while making the calculations more efficient. Here, we describe and assess the implementation of the MARTINI force field for DNA into HADDOCK, our integrative modeling platform. We combine it with our previous implementation for protein-protein coarse-grained docking, enabling coarse-grained modeling of protein-nucleic acid complexes. The system is modeled using MARTINI topologies and interaction parameters during the rigid body docking and semi-flexible refinement stages of HADDOCK, and the resulting models are then converted back to atomistic resolution by an *atom-to-bead* distance restraints-guided protocol. We first demonstrate the performance of this protocol using 44 complexes from the protein-DNA docking benchmark, which shows an overall ~6-fold speed increase and maintains similar accuracy as compared to standard atomistic calculations. As a proof of concept, we then model the interaction between the PRC1 and the nucleosome (a former CAPRI target in round 31), using the same information available at the time the target was offered, and compare all-atom and coarse-grained models.

1. Introduction

Protein-DNA interactions play essential roles in cellular processes such as gene expression, regulation, transcription, DNA repair, or chromatin packaging in eukaryotes[178]. Computational docking, commonly referred to as prediction of the three-dimensional (3D) structure of a complex given the structures of its free constituents, has been extensively proven as an ideal complement to experimental structural methods in order to accurately model biomolecular complexes[118]. Even though computational modeling approaches have steadily progressed in the past decade[179] which starts with the free molecules and allows for conformation changes, may be used to predict the structure of a protein-protein complex. This requires at least two steps, a rigid-body search that determines the relative position and orientation of the subunits, and a refinement step. The methods developed in the past twenty years yield native-like models in most cases, but always with many false positives that must be filtered out, and they fail when the conformation changes are large. CAPRI (Critical Assessment of PRedicted Interactions, modeling large biomolecular assemblies still remains a challenge. In other words, application to either large individual or high number of interactors are limited by the significant computational cost of thoroughly sampling the complex and intricated conformational landscapes and by the increased difficulty of identifying near-native structures from the large pool of generated models[18].

Coarse-graining (CG) has been demonstrated to be a valuable alternative to standard atomistic (AA) approaches to alleviate some of those limitations and help the identification of the energy global minima by smoothing out the energy landscape[13,180]. To this end, CG approaches group several atoms (either a few atoms or entire side chains) into larger “pseudo-atoms” or

"beads," which results into a reduction in the number of degrees of freedom of the system[26]. Historically, the development of CG force fields has followed two directions: (1) Physics-based, parametrized against its atomic counterpart or (2) knowledge-based, taking advantage of the increasing growth of statistical information derived from experimentally determined structures[180]. Protein or/and protein-nucleic acid coarse-grained approaches have been implemented in several docking/ modeling software such as for example: CABS-dock[162], RosettaDock[92], IMP[9], ATTRACT[80], NPDock[181], PyRy3D (genesilico.pl/pyry3d), and more recently in HADDOCK[8,13], our integrative modeling platform.

MARTINI, a popular coarse-grained model for biomolecules, features lipids[5], proteins[41], carbohydrates[44], and nucleic acids[16,47] among others. Its DNA parametrization combines *top-down* (experimental data) and *bottom-up* (atomistic simulations) methodologies and is fully compatible with all other MARTINI models. On average, the nucleic acids' mapping follows a 1:6 ~7 rule, which means that each nucleotide is mapped onto six or seven CG beads. Bead types are selected according to partition free energies from water to chloroform or hydrated octanol. Bonded interactions have been fitted to reproduce dihedral, angle and bond distributions from atomistic simulations of short single stranded DNAs (ssDNAs)[35]. The general design and parametrization of MARTINI allow to easily combine several types of biomolecules (high transferability) as well as a straightforward conversion to atomistic resolution.

In this manuscript, we describe and benchmark the integration of the MARTINI coarse-grained force field for DNA into HADDOCK. It builds upon our recent implementation of a MARTINI coarse-grained protein-protein docking protocol[13] and is further optimized to account for Watson-Crick interactions. Prior to the docking, the input structures are converted into

their coarse-grained counterparts and hydrogen-bonding base pairs are automatically detected so that a special set of parameters and restraints are used for those during the docking. We evaluate the performance of coarse-grained protein-nucleic acid docking using 44 unbound-unbound complexes from the protein-DNA benchmark[182]. The results show a similar performance in terms of success rate and model quality while reducing the computational costs by ~6-fold compared to standard atomistic simulations. For 6 of those, we repeated the docking (both all-atom and coarse-grained) using experimental data to drive the docking as a demonstration that our coarse-grained protocol is also applicable for integrative modeling purposes. Finally, we showcase the potential of CG protein-DNA docking by revisiting the PRC1-nucleosome core particle complex[183], which was offered as a CAPRI target (Target 95 in round 31[184]) for which we failed at the time to select any near native models.

2. Methods

2.1. Integration of the MARTINI DNA Coarse-Grained Force Field Into HADDOCK

The integration of the MARTINI coarse-grained force field for nucleic acids into HADDOCK builds upon our recent HADDOCK-CG implementation for protein-protein docking[13]. We converted the MARTINI topologies and interaction parameters into a format compatible with the computational engine of HADDOCK, CNS–Crystallography and NMR System[78]. As in MARTINI, we represent the backbone of the nucleotide by three beads, one for the phosphate group, and two different beads for the sugar. Pyrimidines and purines are mapped into three and four beads, respectively. A detailed list of the topologies and parameters as

used in HADDOCK can be found in the Supplementary Information (see Tables SI4.1-2 in Supplementary Information).

The latest official release of the MARTINI force field for nucleic acids, 2.2[35], includes eight additional beads and corresponding parameters compared to previous versions. These beads specifically account for Watson-Crick base pairing and mimics, to some extent, the hydrogen bonds that are formed between complementary nucleotide base pairs. These contribute to stabilizing the DNA double helix structure. When converting atomic structures into coarse-grained models, we automatically detect base pairing by calculating the Euclidean distance between neighboring nucleic acid side-chain atoms. We also use the distance between phosphate groups to ensure that bases are paired with their counterpart on the opposite strand and not with their neighbor in the sequence. We define a base pair when two opposite bases' heavy atoms are within the well-accepted hydrogen bond length of 3.5 Å, as used for example in LIGPLOT[146], and their phosphate groups are at least 10 Å or further away from each other. If the input structures do not contain any phosphate, we use instead the center of mass of the nucleotides. By doing so, we avoid defining coupling between neighboring bases in sequence. This information is used by the HADDOCK machinery to ensure that specific interacting beads are used when necessary and the default HADDOCK DNA restraints were adapted to account for the CG beads and used to enforce correct DNA pairing (see Table SI4.3 in Supplementary Information). As recommended in MARTINI, non-bonded interactions between CG beads are calculated using a 14 Å cutoff, whilst 8.5 Å is the default value for the united-atom OPLS force field[168] used in HADDOCK. Note that 8.5 Å is a reduced cutoff compared to the recommended one for OPLS, which was chosen as a compromise between accuracy and speed.

2.2. Docking procedure

Prior to the docking, we convert the atomic PDB coordinate files containing DNA/protein into a coarse-grained representation via an updated version of our *in-house* HADDOCK script for pre-processing CG input structures. During the vacuum part of the docking protocol (*it0* and *it1*) we set the dielectric constant (epsilon) to 78.0 to screen the high DNA charge (in the all-atom representation). Epsilon is set to 1.0 for the final refinement stage in explicit solvent (water)[182]. In the CG runs, the final water refinement is replaced by the back-mapping from coarse-grained to atomistic resolution as previously described[13]. Note that in our atomistic DNA force field implementation, the charge on the backbone phosphate is reduced to 0.5 since no counter ions are included in the docking to screen its charge, while the phosphate bead in MARTINI is uncharged. The final resulting models are clustered based on the fraction of common contacts (FCC)[169] using a 0.6 cutoff (i.e., two models belonging to the same cluster share at least 60% of contacts) and a minimum of four models per cluster, which is the default clustering protocol in HADDOCK. All docking calculations were made using the latest 2.4 version of HADDOCK (still in beta version and unpublished but available upon request).

2.3. Protein-DNA docking benchmark

To systematically test the performance of our coarse-grained implementation for protein-DNA docking, we used 44 unbound-unbound cases from the protein-DNA benchmark[182]. Those are composed of 26 binary, 16 ternary, 1 quaternary (*2c5r*), and 1 pentameric (*1ddn*) complexes covering all major types of interactions[185]. We removed three cases from the original dataset (PDB codes: *1diz, 1emh,* and *4ktq*) due to the fact that the MARTINI force field does not explicitly account

for the modified nucleic bases *P2U, NRI,* and *DOC.* The benchmark is classified according to the amount of conformational changes that take place upon binding as measured by the interface positional root mean square deviation (i-RMSD) (i.e., unbound vs. bound structures) as follows:

- Easy (0 Å < i-RMSD ≤ 2 Å),
- Intermediate (2 Å < i-RMSD ≤ 5 Å), and
- Difficult (i-RMSD ≥ 5 Å).

This selection yielded 11 easy, 21 intermediate, and 12 difficult cases. For comparison purposes, we performed two different docking runs, one using the default atomistic force fields used by HADDOCK, and a second one with the parameters adapted from the MARTINI CG force field for both protein and DNA[16,41]. For the all-atom representation, OPLSX non-bonded parameters are used both for the protein[168] and DNA[186].

We used true interface information derived from the crystal structures translated into ambiguous interaction restraints (AIRs) to drive the docking calculations as previously defined[182]. The sampling parameters were kept to their default in HADDOCK: *1000/200/200* models were generated for the rigid body *(it0)*, simulated annealing *(it1)* and water refinement *(itw)* stages, respectively.

2.4. Unbound docking using experimental data

We additionally modeled six complexes from the protein-DNA benchmark for which experimental data are available. The selected cases cover the different categories from the benchmark; "easy" *(1by4, 3cro)*, "intermediate" *(1azp, 1jj4)*, and "difficult" *(1a74, 1zme)*. The available experimental information was collected from literature and include conserved residues,

mutagenesis data, ethylation interference data, methylation interference data, NMR native state amide hydrogen exchange, and Raman spectroscopy as previously described[182]. As in the previous study[182], the sampling was slightly increased to *2000/400/400* for *it0/it1/itw* docking stages, respectively.

2.5. Modeling of the PRC1 ubiquitylation module bound to the nucleosome

We modeled the interaction between the multimeric PRC1 ubiquitylation module and the nucleosome by performing both AA and CG docking runs. As starting point for the docking, we used the unbound crystal structure of the enzymatical complex (PDB code: *3rpg*) and the nucleosome particle (PDB code: *3lz0*). We followed the same docking procedure as explained above (see Methods: *Docking Procedure*) except for the sampling parameters that were increased to 100000, 400, and 400 for *it0, it1,* and *water* stages, respectively, because of the scarcity of the available information.

The docking was driven by interaction restraints obtained from the literature at the time of CAPRI Round 31: One unambiguous distance restraint between the SG atom of the catalytic cysteine 85 of PRC1 and the NZ atoms of Lys119 or Lys118 on H2A, the ubiquitination target. In addition, we included mutagenesis data on PRC1 (K62A, R64A, K97A, and R98A) shown to be crucial for the interaction with the nucleosome[187,188]. Ambiguous interaction restraints (AIRs) were defined for those (active) against all solvent accessible residues (passive) on the histones (those with either main chain or side chain relative accessibility >25% as calculated by NACCESS[173]). The list of active and passive residues used to guide the docking and the specific distance restraint can be found in Supplementary Information (see Table SI4.4).

2.6. Metrics for the evaluation of model quality

We evaluated the quality of the generated models following the standard CAPRI criteria[189]. This includes the fraction of common contacts (FNAT) and the interface (i-RMSD) and ligand (l-RMSD) positional root mean square deviations from the reference crystal structures. FNAT is calculated from all heavy atom–heavy atom intermolecular contacts using a 5 Å distance cutoff. The i-RMSD is calculated on the interface backbone atoms after superimposition on the backbone of the interface residues, defined as those with any heavy atom within 10 Å distance of the partner molecule. The l-RMSD is calculated on the ligand backbone (usually the smallest molecule) after superimposition on the backbone atoms of the receptor (largest molecule). For both i-RMSD and l-RMSD, we only considered either backbone heavy atoms for atomistic models (C-alpha, C, N, O/P, C1, C9 for protein/DNA) or backbone particles (*BB**) for coarse-grained models (in the *it0* and *it1* docking stages). The calculations were performed using ProFit and the quality of the docking poses was classified as:

- High: FNAT ≥ 0.5 and i-RMSD ≤ 1 Å or l-RMSD ≤ 1 Å,
- Medium: FNAT ≥ 0.3 and 1 Å < i-RMSD ≤ 2 or 1 Å < l-RMSD ≤ 5 Å,
- Acceptable: FNAT ≥ 0.1 and 2 Å < i-RMSD ≤ 4 or 5 Å < l-RMSD ≤ 10 Å, and
- Low quality: FNAT < 0.1 or i-RMSD > 6 Å or l-RMSD > 10 Å

2.7. Metrics for the evaluation of docking success rate

We analyzed the performance of the docking calculations as: (1) The percentage of cases in which at least one model of a given accuracy is found within the top *N* solutions ranked by HADDOCK (*N* = 1, 5, 10, 20, 25, 50, 100, 200), and (2) the percentage of cases in which at least one acceptable or higher quality model was found in the top *T* clusters (*T* = 1, 2, 3, 4, 5).

3. Results and Discussion

We have integrated the MARTINI CG force field for nucleic acids into HADDOCK version 2.4 (see Methods: *Integration of the MARTINI DNA Coarse-Grained Force Field Into HADDOCK*), combining it with our previous implementation of the protein MARTINI CG force field, enabling full coarse-grained protein-DNA docking. The AA to CG conversion scripts have been adapted to automatically account for specific Watson-Crick base pairing, which require special interacting parameters. In the following sections, we discuss the performance of our protocol for protein-DNA docking in terms of success rate and computational efficiency using 44 unbound-unbound complexes from the protein-DNA benchmark[182] with ideal interface information (see Methods; *Protein-DNA docking benchmark*). For six of them, we repeated the docking using experimental information to guide the docking.

Finally, as a proof of concept, we revisited CAPRI Target 95[184], a protein-nucleosome complex for which we failed to identify near native solutions in our original CAPRI submissions (although we did generate some). In this new modeling, our top ranked predictions are in excellent agreement with the crystal structure of the complex (not used for the docking) for both standard atomistic docking and the hereby described coarse-grained implementation.

3.1. Overall performance of coarse-grained protein-DNA docking

The docking was performed starting from the unbound structures of each molecule and driven by AIRs as defined in our previous study[182] (see Methods; *Docking procedure*). In order to evaluate the performance of our approach, we calculated the success rates of both sets of runs (AA and CG) as the percentage of cases for which an acceptable or better quality was obtained

in the top *N* ranked models (for details see Methods). Overall, coarse-grained docking generates and delivers acceptable or higher quality models for 40 out of the 44 cases after the back-mapping stage compared to 38 cases for the atomistic docking results. No near-native models are generated for four complexes; two of which are classified as difficult *(1dfm, 1o3t)*, one as intermediate *(1z9c)* and one as easy *(1tro)*. Inspection of the failed easy case reveals that it is a ternary complex (homodimer) and since no symmetry restraints were used in this case, its interface ambiguity was too high. In a previous benchmarking[190], acceptable models for this complex were obtained using a two-stage docking protocol in which a library of bent DNA conformations were given as input for the second docking run (a procedure not followed here). Among the successful CG cases, medium quality models are generated for 23 cases against 26 for the AA docking runs. Top one single structure-based ranking (best ranked structure) reaches 86.3% success rate for all-atom calculations vs. 81.8% for CG docking (Figures 1A, 1B).

The overall success rates are similar for the top 5 and becomes higher for CG docking, reaching 90.9% in the top 200 while AA docking remains at 86.3% (which corresponds to 40 vs. 38 successful cases for CG and AA docking, respectively). In contrast, the quality of the models is slightly better for AA docking as measured by the success rates (Figures 1A, 1B) and rankings of medium quality models (Figures 1C, 1D). Notably, CG docking manages to generate acceptable models for two of the difficult cases that fail at standard atomistic HADDOCK runs (*1zme* and *1qrv*). In *1zme*, we find an acceptable model at position 176 (i.e., Top 200 according to our analysis) with 0.11/7.85 Å/9.94 Å for FNAT/i-RMSD/l-RMSD while the best AA model falls out the acceptable CAPRI criteria (0.04/7.51 Å/10.3 Å).

For *1qrv*, the fourth case with the largest conformational change, the docked models generated by the standard AA HADDOCK protocol failed to satisfy the quality metric thresholds (FNAT and i-RMSD or FNAT and l-RMSD). However, several models showed a satisfactory overlap in terms of FNAT with >20% of interface contacts. With coarse-graining instead, the first acceptable model is found at rank 44 with a l-RMSD of 8.8 Å and FNAT of 0.14 (i.e., Top 50 according to our analysis).

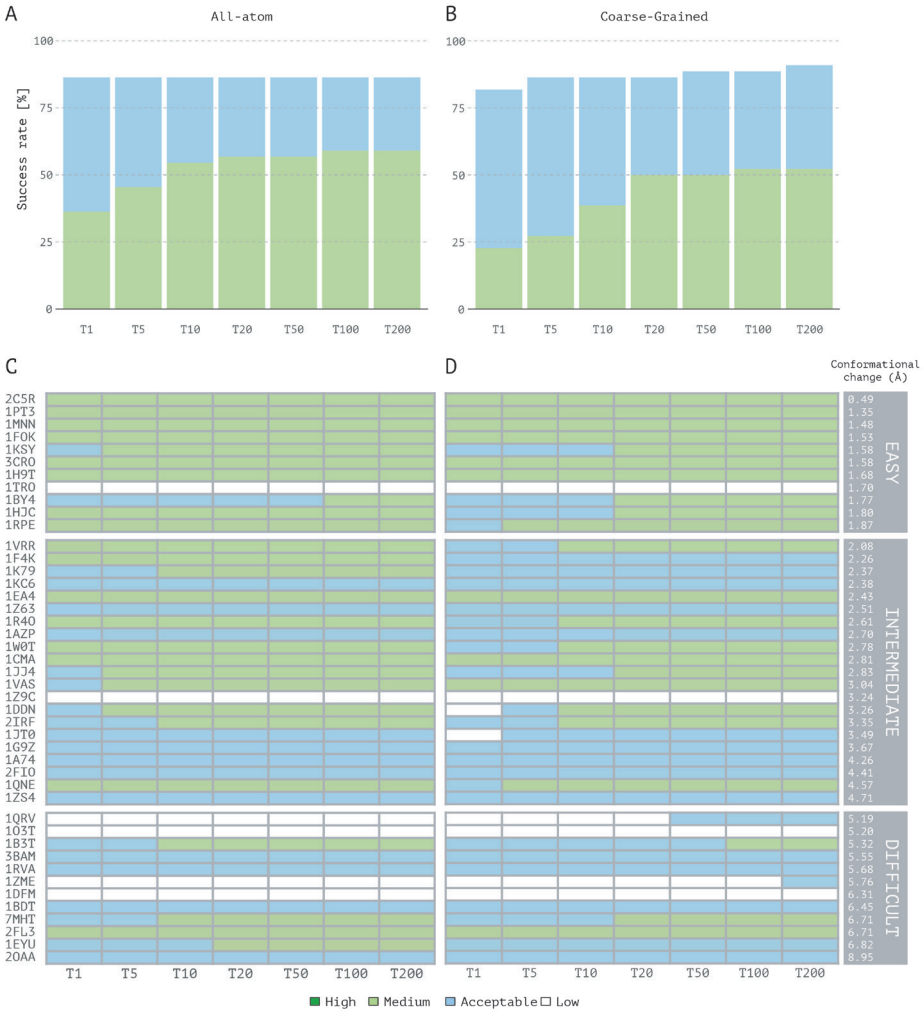
Coarse-graining approaches benefit from the reduction of the number of degrees of freedom of the systems under study and make the docking calculations computationally more efficient. The median computational time to generate one model via CG in HADDOCK is 8.6s and of 42.8s for *it0* and *it1* stages, respectively, vs. 16.5s and 115.0s for standard atomistic calculations. Overall, the use of the MARTINI force field for both proteins and nucleic acids leads to a ~6-fold speed increase during rigid-body docking and semi-flexible stage (see Table SI4.5 in Supplementary Information).

3.2. Unbound docking using experimental data

We evaluated the capabilities of our HADDOCK-CG implementation to model protein-DNA interactions when using real experimental information. We selected six representative cases[182] from the protein-DNA benchmark classified as “easy” *(1by4, 3cro)*, “intermediate” *(1azp, 1jj4)*, and “difficult” *(1a74, 1zme)* for which experimental information was available.

Figure 1. Performance of the all-atom and coarse-grained protocols in HADDOCK on the 27 largest complexes of the docking benchmark 5.

A) Overall success rates (%) of the all-atom protocol on ranking single models (Single) or clusters (Clustering) as a function of the number of models/clusters considered.
B) Same as **(A)** but for the coarse-grained protocol.
C) and **(D)** Quality of the docking models for all 27 cases as a function of the number of models considered.
The complexes are ordered by increasing degree of difficulty (from top to bottom) for both all-atom and CG docking runs. The color coding indicates the quality of the docked models.



The latter was translated into AIRs (see Methods; *Unbound docking using experimental data*) in the form of active and passive residues and two different set of docking runs were performed using either the standard all-atom or the coarse-grained protocols.

As shown in Table 1, summarizing the quality of the generated clusters, for four out of the six cases, AA docking generates better quality models. No good solution in any of the tested protocols was found for *1zme*, which undergoes a large conformational change of 4.68 Å upon binding. In terms of sampling, the standard all-atom protocol, in combination with experimental data, generates ~900 near-native models (i.e., acceptable or higher quality according to CAPRI) on average per case, while our CG approach around three times less (~300). This is somewhat surprising as the smoother energy landscape derived from the reduction of degrees of freedom might help the sampling process as previously demonstrated in our protein-protein CG implementation[13]. Despite this difference in sampling, both approaches perform rather similarly in terms of structure quality, indicating that our CG protocol is also applicable for integrative modeling of complexes in combination with real experimental data. Recent studies have indicated that the interpretation of CG models using experimental data, and in particular SAXS data, can benefit from improved forward models as demonstrated[191] for protein-DNA complexes.

3.3. Revisiting CAPRI Target 95: The PRC1 ubiquitination module bound to the nucleosome

The polycomb repressive complex 1 (PRC1) represses the expression of genes regulated by developmental processes and is responsible for the ubiquitylation of the nucleosomal histone[188]. This complex was offered as a blind target to the CAPRI experiment (Round 31, target 95), to which

we participated but failed to correctly identify near-native models out of our pool of generated complexes. Using the same information derived from the literature as used in CAPRI Round 31 (see Table SI4.4 in Supplementary Information), we repeated the docking using our MARTINI implementation in HADDOCK2.4 and validated our predictions against the crystal structure of the complex (PDB-ID: *4rp8*[183]).

Table 1. Performance of the all-atom and coarse-grained protocols in HADDOCK on six representative cases of the protein-DNA benchmark using experimental data to drive the docking.

	All-atom					Coarse-grained				
Complex	**Cluster**	**i-RMSD**	**l-RMSD**	**FNAT**	**CAPRI**	**Cluster**	**i-RMSD**	**l-RMSD**	**FNAT**	**CAPRI**
EASY										
1by4	2nd	3.66	14.37	0.18	*	1st	3.08	9.05	0.19	*
3cro	1st	1.52	2.34	0.39	**	2nd	2.77	7.35	0.22	*
INTERMEDIATE										
1azp	1st	3.14	10.16	0.11	*	1st	3.53	9.29	0.10	*
1jj4	2nd	1.98	5.71	0.25	*	1st	2.24	6.55	0.11	*
DIFFICULT										
1a74	1st	1.61	4.41	0.32	**	1st	1.83	4.54	0.24	*
1zme	1st	8.52	29.54	0.00	-	1st	8.4	30.7	0.00	-

The RMSDs (Å) and FNATs correspond to the best model of the best cluster. The ranking of the best cluster is also reported. The CAPRI column indicates the number of models per quality threshold (acceptable,*
*** medium, *** high).*

When analyzing the i-RMSD of the top-ranked model according to the HADDOCK score, the CG one is slightly closer (3.0 Å) to the reference crystal structure than the corresponding AA model (3.14 Å; Table 2A). Same behavior is observed when looking at the clustering statistics, in which the

average i-RMSD for the top four models of the best cluster for CG was 3.09 ± 0.08 Å against 3.23 ± 0.23 Å in AA. A much large difference between the two protocols is however clearly visible when comparing the number of acceptable of better models generated at the various docking stages (Table 2B) with CG docking resulting in ~1.5 times more acceptable models than AA docking.

Table 2. Sampling and quality assessment of the AA and CG PRC1 docking models

A. Number of acceptable models and time necessary to generate one model for the rigid-body and semi-flexible stages for both all-atom and coarse-grained simulations.

	# of acceptable models			Time per model [s]	
	it0[a]	it1	water	it0	it1
All-atom	360/173	169	169	138	979
Coarse-grained	536/293	290	254	27	188

[a] The first number is the total number of acceptable models within the 10000 generated and the second correspond to those in the top400 selected for further semi-flexible refinement.

B. Ranking, i-RMSD comparison and time per model of all-atom and coarse-grained simulation of CAPRI Target 95.

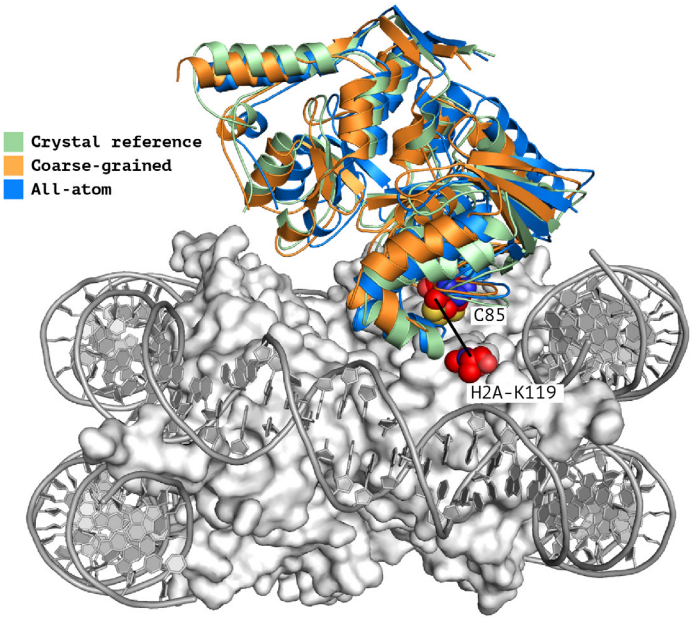
	Single structure		Cluster	
	Rank	i-RMSD [Å]	Rank	Top4 <i-RMSD> [Å]
All-atom	1	3.14	2	$3.23_{\pm 0.23}$
Coarse-grained	1	3.00	1	$3.09_{\pm 0.08}$

This improvement in the sampling is in contrast to what was observed above for the protein-DNA benchmark. As already observed for protein-protein docking[13], the impact of coarse graining is more evident when little or no information (ab initio docking) is available to drive the docking process.

Finally, a view of the top ranked models superimposed onto the reference crystal structure is shown in Figure 2. Both satisfy the distance restraint imposed to model the interaction between Cys85 of PRC1 with Lys118/119 of Histone 2A (PRC1-H2A). The proximity of those two residues was proposed[187] to be necessary to restrict the ligase complex to a single region of the nucleosome (the information we used in CAPRI), which was confirmed by the crystal structure (PDB-ID *4r8p*[183]).

Figure 2. Single structure comparison of top-ranking models predicted by HADDOCK.

Superimposition of the best models (top-ranked) predicted by HADDOCK using atomistic (blue) or coarse-grained (orange) docking onto the experimental crystal structure (PDB-ID 4r8p[183], green).
The two residues PRC1-Cys85 and H2A-Lys119 which are expected to form a covalent bond[192] (an information used to guide the docking) are shown as spheres.
The interface RMSD of the all-atom and coarse-grained top rankings models against the reference crystal structure are 3.23 and 3.0 Å, respectively.



4. Conclusion

In this work, we have presented the integration of the MARTINI coarse-grained force field for nucleic acids into our HADDOCK integrative modeling software. It builds upon our previous implementation for protein-protein docking, using a coarse-grained representation during the rigid-body and semi-flexible refinement stages, and converting back the resulting models to atomistic resolution following an atom-to-bead distance restrained-guided morphing procedure. We have shown that the performance of coarse-grained docking is similar to that of standard all-atom protocol in terms of success rate, while the quality of the generated models remains rather similar according to standard CAPRI criteria. We demonstrated that our coarse-grained protocol is perfectly suited for use with experimental or predicted data. In particular, we have revisited a challenging target of the CAPRI experiment, taking full advantage of the hereby described implementation and obtaining near-native models of PRC1 Ubiquitination module bound to the nucleosome in excellent agreement with the crystal reference. Further, by smoothening the energy landscape it also allows to generate more near native models in cases where limited information is available to guide the modeling, which should also benefit the scoring stage since it becomes easier to identify them. It also brings a significant gain in computing performance, with a ~6-fold speed increase compared to standard atomistic simulations. In conclusion, with this extension, HADDOCK has gained the capability to model significantly larger assemblies consisting of mixed protein and DNA components, in a more efficient way without compromising its overall performance.

Chapter 5

Integrative modeling of membrane-associated protein assemblies

Jorge Roel-Touris †, Brian Jiménez-García †,
Alexandre M.J.J. Bonvin

Published in 2020 in Nature Communications, Volume 12, Article Number 6210

† Joint first authors

Abstract

Historically, membrane protein systems have been considered as one of the most challenging systems to study with experimental structural biology techniques. Over the past years, increased number of experimental structures of membrane proteins have become available thanks in particular to advances in solid-state NMR spectroscopy and cryo-electron microscopy. This has opened the route to modeling the complexes that those membrane proteins form by methods such as docking. Most approaches developed to date are, however, not capable of incorporating the topological information provided by the membrane into the modeling process. Here, we present an integrative computational protocol for the modeling of membrane-associated protein assemblies, specifically complexes consisting of a membrane-embedded protein and a soluble partner. It combines efficient, artificial intelligence-based rigid-body docking by LightDock with a flexible final refinement with HADDOCK to remove potential clashes at the interface. We make use of an equilibrated coarse-grained lipid bilayer to represent the information encoded in the membrane in the form of artificial beads, which allows to target the docking towards the binding-competent regions. We demonstrate the performance of this membrane-driven protocol on eighteen membrane-associated complexes, whose interface lies between the membrane and either the cytosolic or periplasmic regions. In addition, we evaluate how different membrane definitions impact the performance of the docking protocol and provide a comparison, in terms of success rate, to another state-of-the-art docking software, ZDOCK. Finally, we discuss the quality of the generated models and propose possible future developments. Our membrane docking protocol should allow to shed light on the still rather dark fraction of the interactome consisting of membrane proteins.

1. Introduction

Membrane proteins (MPs) play crucial roles in many biological functions within the cell. Commonly, MPs are classified based on their association mode with biological membranes into two main groups: Peripheral membrane proteins that are located on either side of the membrane and are attached to it by non-covalent interactions, and integral membrane proteins (IMPs) that are inserted into the membrane and can be either exposed on only one side of the membrane (monotopic membrane proteins) or span the entire lipid bilayer. The latter, known as transmembrane proteins (TMs), are structurally categorized as α-helical bundles or β-barrels[193]. TMs mostly function as regulators of complex biochemical pathways (receptors and transducers) and/or transporters of molecules (channels and carriers). Only transmembrane proteins can function at both sides of the membrane by forming larger complexes. As such they are not simply passive membrane spanning proteins but play important roles in protein-protein interactions (PPIs), thus making them valuable targets for drug discovery (around 60% of current drug targets are MPs[194]. A well-known example are G-protein coupled receptors (GPCRs) which are involved in many diseases[195]. Those are collected in a specific database:
(GPCRdb; https://www.gpcrdb.org/)[196].

Over the past years, development of cutting-edge technologies has facilitated the study of previously inaccessible MPs, advancing the field of membrane structural biology. Obtaining high-quality crystals suitable for X-ray crystallography is still far from trivial. Solid-state NMR spectroscopy, and especially cryo-electron microscopy (cryo-EM), reaching near-atomic resolution, have become central tools to study membrane-associated protein complexes[197,198]. However, experimental conditions such as low expression profiles and/or high instability outside the native membrane still

makes their structural characterization challenging[199]. Despite their large representation in the proteome (in human, nearly a quarter of the genome encodes for MPs[200]), roughly only 1% of all deposited protein structures in the Protein Data Bank[153] (PDB) corresponds to MPs, with 1099 unique protein entries as of July 2020: blanco.biomol.uci.edu/mpstruc. Even less of those have been experimentally solved in complex with their counterpart(s). For all these reasons, membrane protein systems, which are increasingly attracting attention, have been traditionally considered as one of the most difficult type of systems to study by experimental structural biology techniques.

Computational methods offer an attractive alternative for studying membrane systems[201]. Many efforts have been made to develop efficient tools to computationally predict the 3D atomic structures of membrane-associated proteins and their complexes[202]. Some rely on secondary structure or topology prediction and make use of either knowledge-based statistics or evolutionary information to generate 3D models[203,204]. The simplest computational methods are based on homology modeling. In short, these approaches require a template structure (or multiple) with high sequence similarity to the target sequence, and usually produce very reliable "core" models (corresponding to the TM domains) and less accurate predictions for the extracellular loops. Methods such as MEDELLER[205] and Memoir[206] have greatly benefited from the increasing availability of cryo-EM derived structures in the PDB and are inspired on the well-known homology modeling tool MODELLER[207].

Another representative subset of computational methods geared towards modeling complexes are docking-based approaches. Docking commonly includes two different steps, namely sampling and scoring. Sampling is usually referred to as the process of generating (tens of) thousands

of possible conformations of a given (bio)molecular complex. This can be done through a number of well-established techniques such as Fast Fourier Transformation (FFT)-based methods included in various docking software such as GRAMM-X[208,209], ClusPro[210], pyDock[211] and ZDOCK[212]. These methods, however, do not allow for explicit flexibility of the modeled partners due to intrinsic limitations of the FFT sampling. Although this limitation can be partially solved by using ensembles of conformers, it implies higher computational cost. Energy minimization, in HADDOCK[8] and ATTRACT[110] for example, Metropolis Monte Carlo optimization, e.g. in RosettaDock[92], or artificial intelligence-based algorithms, such as implemented in Swarm-Dock[213] and LightDock[10], are also used. The sampling process is often followed by a refinement of the docked models for which molecular dynamics- or Monte-Carlo-based protocols are the most commonly used. The generated models are scored with the aim of discriminating between biologically-relevant (native) and non-relevant models. This is typically done with a scoring function, which can be based on either physico-chemical properties and/or statistical potentials[214]. Nowadays, with the increasing availability of large pools of docking models such as provided in the CAPRIDOCK[215], PPI4DOCK[216] and DOCKGROUND[217], machine(deep)-learning scoring functions are gaining interest[218]. Sampling and scoring might be coupled (scoring-driven sampling) or work as independent steps (sampling and then scoring).

In the context of membrane protein docking, software such as Rosetta[219], DOCK/PIERR[220] and Memdock[221] include built-in specific protocols to model transmembrane domains using implicit membrane potentials. Besides RosettaMP[222](for membrane protein design), none of the available membrane-specific computational methods allow for an explicit representation of the lipid bilayer and, therefore, cannot harvest the topological information encoded in it.

In this work, we present an integrative computational approach for modeling membrane-associated protein assemblies (complexes consisting of a membrane-embedded protein and a soluble partner) that combines an efficient, *swarm*-based rigid-body docking by LightDock with a flexible final refinement with HADDOCK to remove potential clashes at the interface. To introduce the topological information provided by the lipid bilayer we make use of an equilibrated coarse-grained membrane into the docking calculations. In that way we can focus the docking towards binding competent regions, excluding all irrelevant regions prior to the simulation. This membrane representation has been implemented within the LightDock framework[10]. The sampling in LightDock is based on an artificial intelligence-based *swarm* approach that relies on the metaphor that, in nature, *glowworms* (which represent ligand poses) feel attracted to each other depending on the amount of emitted light (scoring, energetic value of a docking pose). In this way, the docking poses, which constitute the *swarm* of *"glowworms"* in LightDock, are optimized towards the energetically more favorable ones through the translational, rotational and Anisotropic Network Model (ANM) spaces. The latter is, however, not available in the membrane docking mode. Sampling and scoring in LightDock are tightly interconnected since the optimization process is score-driven. In its latest official release (version 0.8.0; http://pypi.org/project/lightdock), LightDock supports the use of information such as mutagenesis and/or bioinformatic predictions to bias the sampling[12]. The LightDock-generated membrane protein models are then refined with HADDOCK via an efficient coarse-grained (CG) protocol[13]. This protocol, originally designed to backmap coarse-grained models to atomistic resolution by morphing atomistic models onto the coarse-grained ones using distance restraints, is very efficient in removing steric clashes while maintaining the original geometry of the docked models.

We demonstrate the efficiency and performance of this two-step (docking and refinement) membrane-driven protocol on the 18 membrane protein complexes from the MemCplxDB benchmark set[223] whose interface lies between the membrane and either the cytosolic or periplasmic regions. We also evaluate how different choices for defining the membrane topology affect the sampling of our protocol, and assess the quality improvement of the generated models after the HADDOCK refinement step. We compare the success rate of this integrative approach and the quality of the generated models with that of another state-of-the-art docking software, ZDOCK[212], for which we test several docking scenarios penalizing ("blocking") regions during sampling and therefore explicitly accounting for the information provided by the membrane. Finally, we discuss the quality of the side-chains at the interface of the generated models and propose future developments that could be made for improving the current results.

2. Methods

2.1. Membrane docking dataset

We selected all complexes from the MemCplxDB database[223] whose interface lies between the membrane and either cytosolic or periplasmic regions. This selection yielded a dataset of 18 cases (See Fig. 2) which were further classified into:

- α-Helical: complexes whose receptor assemblies as a α-helical bundle.
- β-Barrel: complexes whose receptor forms an antiparallel β-sheet composed tandem of repeats.
- Antibodies: complexes whose soluble ligand is an antibody or nanobody.

2.2. Pre-processing of input structures

We make use of an equilibrated coarse-grained representation of the membrane to include topological information in our modeling procedure. For this, for each benchmark case, we obtain a representative coarse-grained snapshot of the transmembrane protein inserted into a simulated lipid bilayer (MARTINI representation[41]) from the MemProtMD database (Fig.1 – *Step A*) ([224]; http://memprotmd.bioch.ox.ac.uk/). For the sake of simplicity and for saving computational resources, we remove all lipid beads except those representing the phosphate groups, which, to some extent, represent the most external layers. Then, we replace the coarse-grained TM receptors by their corresponding atomistic structure (Fig.1 – *Step B*). When needed, we remove beads overlapping or clashing (< 2.5Å distance) with any heavy atom of the transmembrane protein once inserted into the membrane (*1ots*, *2gsk*, *2hi7*, *4m48*, and *3wxw*).

2.3. Implementation of a coarse-grained membrane in LightDock

To allow for the use of a coarse-grained membrane within the LightDock framework, we added new logic for the two different stages namely: The internal preparation of the molecules (at the *setup* level) and the actual simulation (at the scoring level). In the first stage, *setup*, we have added a new flag (*-membrane*) to activate the filtering of initial *swarms* (independent centers of simulation) according to the topological information of the membrane (no *swarms* will be generated below it). The protocol will detect the number of bead membrane layers provided by the user and select the upper one. For that purpose, it is expected that the user will provide the structure in PDB format and by a cenital plane point of view (the Z-axis is perpendicular to the membrane plane, that is the default view when saved by PyMol for example[225]). In case the lower layer is the target of

interest, the system should be rotated by 180° around the X or Y axis. During the simulation, we have included a term into the scoring scheme so that docking models in which the ligand penetrates the membrane are penalized and will be forced to optimize towards more favorable poses. In our case, we have defined a very unfavorable potential value for the membrane beads in the DFIRE scoring function used by LightDock (-999.0 - the more negative the value is, the worse becomes the score), in order to penalize models whose ligand's position is incompatible with the provided membrane model.

2.4. Running LightDock in membrane mode

LightDock execution consists of two steps: *setup* and *simulation*. In the first step, *setup*, the user provides to the *lightdock3_setup.py* command line tool the receptor and ligand structures in PDB file format, together with the number of *swarms*, *glowworms* per *swarm* and other options such as removing hydrogen atoms and/or enabling ANM. In this new version of LightDock, a *-membrane* flag has been implemented in order to filter out *swarms* not compatible with the simulated coarse-grained membrane. For each of the filtered *swarms*, if residue restraints information is provided (as it is the case for the CDR loops for antibody-antigen complexes), this is used for pre-orienting the ligand poses as previously described[12].

In this work, the number of initial *swarms* used is 400 (default - many of them will be filtered by the membrane protocol) and the number of *glowworms* 200 (default). Hydrogen atoms are also removed as they are not supported by the DFIRE scoring function. Although not used in this work, the flexibility provided by the ANM implementation in LightDock is supported for the ligand (soluble) molecules (not the membrane-embedded proteins as those are considered as one entity together with the beads). This can be activated as: *-anm -anm_rec=0 -anm_lig=X*, where *X* indicates

the number of non-trivial normal modes to be considered (being 10 the recommended value). When the setup step finishes, the docking simulation is ready to be started. A second command line tool, *lightdock3.py*, performs the simulation for the number of steps provided by the user (100 in this work) using the DFIRE scoring function and running in parallel depending on the number of cores specified. Once the simulation finishes successfully, predicted poses are generated (*lgd_generate_conformations.py*) and clustered (*lgd_cluster_bsas.py*) according to the default LightDock protocol. Finally, the *lgd_rank.py* command line tool generates a ranking of the top clustered predictions according to LightDock. An exhaustive tutorial of the different steps of the protocol can be accessed online at: https://lightdock.org/tutorials/membrane.

2.5. LightDock computational time requirements

The average run time of a LightDock simulation for the benchmark set is 197min with minimum and maximum values of 22 and 427min, respectively (as measured using 48 AMD Opteron 6320 2.8GHz CPU cores). These times are for the current Python version of LightDock. A new port of the code to the Rust programming language (https://github.com/lightdock/lightdock-rust) shows a general speedup of 8 to 10 times compared to the Python version, which should make it possible to provide it as a web-based server in a near future.

2.6. Coarse-grained refinement in HADDOCK

For the local installation, models must be converted into their coarse-grained representation. This is done via an *in-home* script included in the *CGtools* directory of the HADDOCK2.4 distribution as: "*python aa2cg.py model.pdb*". As output, the script generates the MARTINI-based CG model

(*model_cg.pdb*) as well as a restraints file in the form of *model_cg_to_aa.tbl*, which includes the mapping of the generated coarse-grained beads to their corresponding atoms. The *atom-to-bead* restraints files of the different CG models must be combined into a single file (e.g. *cg-to-aa.tbl*) that will be used by HADDOCK to restore the atomistic resolution. In order to perform the refinement, a handful of parameters within the HADDOCK parameter file (*run.cns*) must be adapted as follows assuming that 100 models will be refined:

- *rotate180_it0=false* (to skip sampling 180° complementary interfaces)
- *crossdock=false* (to refine receptor – ligand from the structures provided)
- *rigidmini=false* (to skip *it0* stage)
- *randorien=false* (to skip *it0* stage)
- *rigidtrans=false* (to skip *it0* stage)
- *ntrials=1* (to skip *it0* stage)
- *structures_0=100* (for *it0* stage)
- *structures_1=100* (for *it1* stage; must always be ≤ than *structures_0*)
- *anastruc_1=100* (for analysis purposes at *it1* stage)
- *waterrefine=100* (for *itw* stage; this is the number of final output models)
- *initiosteps=0* (to skip *it1* stage)
- *cool1_steps=0* (to skip *it1* stage)
- *cool2_steps=0* (to skip *it1* stage)
- *cool3_steps=0* (to skip *it1* stage)
- *dielec_0=cdie* (to switch a constant dieletric constant when CG is used)
- *dielec_1=cdie* (to switch a constant dieletric constant when CG is used)

For setting up the refinement on the HADDOCK2.4 webserver version, a tutorial can be found at:
bonvinlab.org/software/haddock2.4/tips/advanced_refinement/
Note that on the server coarse-graining should be enabled under the *Input data* tab. The refined models are scored and ranked according to the default HADDOCK score, which is a linear weighted combination of terms as:

$$HADDOCK_{score} = 1.0*E_{vdw} + 0.2*E_{elec} + 0.1*E_{AIR} + 1.0*E_{desolv}$$

where E_{vdw} and E_{elec} are the van der Waals and electrostatic energies terms calculated using a 12-6 Lennard-Jones and Coulomb potential, respectively, with OPLS nonbonded parameters, is the ambiguous interaction restraints energy, E_{desolv} is an empirical desolvation score[226]. Note that in this protocol, since we are only refining the model and not providing any restraints, the E_{AIR} term is not contributing to the final score. An example of the HADDOCK parameter files to run the refinement (*run.param* and *run.cns*) can be found at: github.com/lightdock/membrane_docking/tree/master/refinement/example

2.7. Metrics for the evaluation of model quality and success rate

The quality of the models is assessed according to the well-accepted CAPRI criteria[147]. Docking models are classified as high (***), medium (**) or low (*) quality according to their similarities with the native structure by calculating the interface and ligand root mean square deviations (i-RMSD and l-RMSD) and the fraction of native contacts (Fnat) as:

- High: Fnat ≥ 0.5 and i-RMSD ≤ 1Å or l-RMSD ≤ 1Å,
- Medium: Fnat ≥ 0.3 and 1Å < i-RMSD ≤ 2 or 1Å < l-RMSD ≤ 5Å,
- Acceptable: Fnat ≥ 0.1 and 2Å < i-RMSD ≤ 4 or 5Å < l-RMSD ≤ 10Å and
- Incorrect: Fnat < 0.1 or i-RMSD > 6Å or l-RMSD > 10Å.

The overall success rate is defined as the percentage of benchmark cases with at least one acceptable or better model within a given Top *N* (N= 1, 5, 10, 20, 50 100).

2.8. Metrics for the determination of steric clashes

We define a steric clash as any heavy atom-heavy atom intermolecular contact shorter than 2.5Å (i.e. hydrogens excluded). Using this definition of clashes, we sought to investigate whether our coarse-grained refinement protocol in HADDOCK leads to higher quality structures (i.e. less absolute number of clashes) as compared to those generated from the docking step with LightDock. To do so, for each of the benchmarked cases we quantified and compared, on a per model basis, the number of clashes present on the top 100 docked models before and after refinement.

3. Results

3.1. Integrative modeling approach for membrane-associated protein complexes

We have developed a computational approach for modeling the interaction of membrane-associated protein complexes that accounts for the topological information encoded in the membrane. First, we insert the atomistic transmembrane protein into a pre-equilibrated coarse-grained model of the protein in a lipid bilayer provided by the MemProtMD database[224] (see Material and Methods; *Pre-processing of input structures*) and then remove all lipid beads except those representing the phosphate groups. Using this beads layer, we automatically generate a group of independent simulations known as *swarms* over the solvent-exposed receptor surface. Using of the capability of the LightDock framework,

we thus discard irrelevant sampling regions (see Fig. 1a-c where the geometrical centers of the *swarms* are depicted as blue beads). Next, each *swarm* is populated with 200 starting random orientations of the soluble ligand (200 is the defaultnumber of *glowworms*, the agents of the sampling algorithm). This procedure effectively biases the sampling toward the binding-comptent regions on the membrane protein (either cytosolic or periplasmic) and excludes those within the boundaries imposed by the membrane. While LightDock can allow for flexibility during docking through normal modes, this option is not supported for membrane-embedded proteins. In the current implementation of the protocol,the membrane-embedded proteins and their beads are considered as a single entity and as such the ANM model is not applicable. For their soluble partners, the inclusion of flexibility is completely functional and might be enabled for the docking process (See Material and Methods; *Running LightDock in membrane mode*). However, in the results hereby presented, the only limited flexibility introduced in the protocol is that of the final refinement using HADDOCK.

For the scoring during the docking simulation with LightDock, we use an adapted version of the DFIRE[148] scoring function that penalizes models penetrating the membrane, specifically those overlapping with any membrane bead (see Material and Methods; *Implementation of an explicit membrane representation into LightDock*). We select the top 100 models from the optimization of all *swarms* for a final refinement stage with HADDOCK in order to remove clashes at the interface. This is achieved using an efficient coarse-grained refinement protocol: In short, we first generate the corresponding MARTINI-based[41] coarse-grained representation for each of the docked models to be refined; then, by a combination of energy-minimizations and short molecular dynamics stages, the protocol[13] fits the atomistic structure of each of the components onto the generated CG model of the complex and optimizes the system to remove clashes. This

final refinement is performed in the absence of the membrane. The resulting models are then scored and ranked according to the HADDOCK score. Although in this work our protocol only makes use of the membrane as information source during the modeling, it is fully compatible with the use of a variety of experimental data in the form of residue restraints if this source of information is provided[12].

3.2. Overall performance on the membrane docking dataset

We have tested the performance of our membrane-driven protocol on the 18 transmembrane-soluble protein complexes of the MemCplxDB benchmark (see Material and Methods; *Membrane docking dataset* and Fig. 2) and compared it with the results of a full sampling in the absence of the topological information provided by the membrane (i.e. *Blind* docking – see next section). The docking was performed starting from the unbound structures of each constituent, except for *2bs2*, *2vpz* and *4huq* for which no unbound state structures are available. The success rate was defined as the percentage of cases for which an acceptable or better model was obtained within the top *N* ranked models (see Material and Methods; *Metrics for the evaluation of model quality and success rate*).

For the two most representative top *N (T5 and T10)*, our *Membrane* protocol shows an overall success rate of 61.1%, 11 out of 18 complexes, with 3 cases having medium quality models as shown in Fig. 3. It reaches a maximum of 88.9% for the top 100 predictions. High quality models are obtained for one α-Helical case within the top 20 *(3x29)* with the best docking pose (ranked at the 12th position) having 70% of the native contacts and 1.0Å/2.0Å i-RMSD/l-RMSD from the reference crystal structure. The highest success rate for either *T5* or *T10* is achieved for α-Helical complexes. For those, acceptable or higher quality models are

generated for 71.4% of the complexes (5 out 7 cases). This performance, however, drops for the β-Barrel category with 40% success rate for *T5/10* and 80% for *T100* (4 out of 5 cases). Not surprisingly, our protocol fails to deliver near-native models for the case with the largest conformational change (*3v8x*; i-RMSD of 3.42Å between unbound and bound structures), classified as β-Barrel. For *Antibodies* (6 cases), we used the CDR loops to pre-orient the molecules at the *setup* step[12], but these were not specifically used for the scoring. For these cases, our protocol generates acceptable and medium quality models for all complexes (100% success rate for *T50*) with a 33.3% success rate considering the top ranked model *(T1)* and 66.7% for *T5/10*.

Figure 1. Membrane protein integrative modeling workflow.

A) The representative coarse-grained membrane snapshot from the MemProtMD database[224] is selected.
B) The coarse-grained transmembrane receptor is replaced by its corresponding atomistic structure.
C) The binding-competent regions are sampled with LightDock using the membrane defined by beads corresponding to the phosphate positions. The resulting top 100 docked models are selected for final refinement.
D) Refinement with HADDOCK following a coarse-grained to all atom protocol and final scoring.

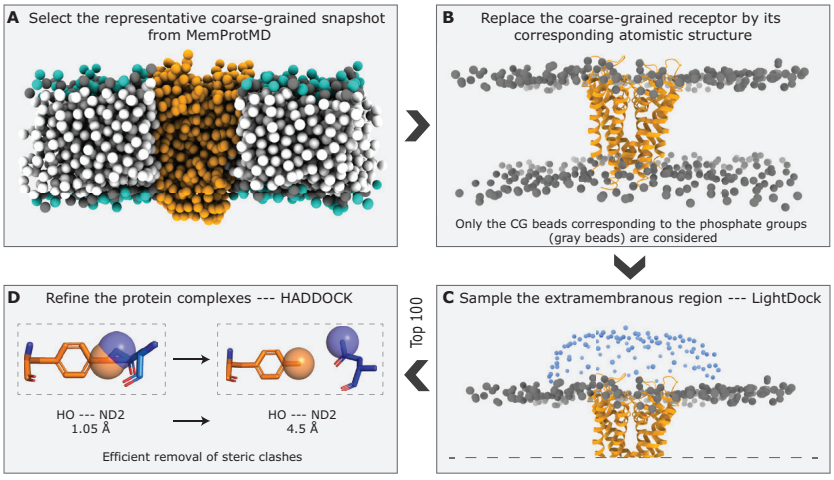
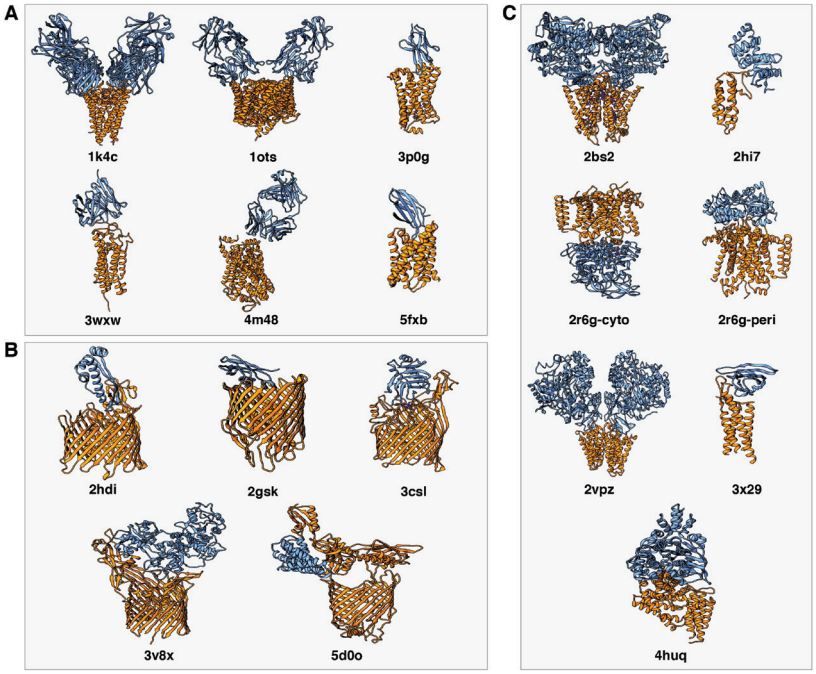


Figure 2. View of the 18 transmembrane-soluble protein complexes of the MemCplxDB database used in this work.

The cases are classified as:
A) Antibodies (the soluble partner is either an antibody or a nanobody),
B) β-Barrel (the transmembrane receptor is a β-sheet barrel) and
C) α-Helical (the transmembrane receptor consists of a bundle of α-helices).
The receptors (the membrane proteins) are depicted in orange and the ligands in blue.



3.3. The integrative modeling protocol outperforms blind predictions

For the *Blind*, membrane-free predictions, LightDock-HADDOCK reaches an overall success rate of 16.7% for the *T100* (11.1% of medium quality models), with a moderate performance for *T5* and *T10* (5.5% and 11.1%,respectively). For one bound case *(4huq)* and one with the second lowest conformational change (*3x29*; 0.67Å), the *Blind* protocol does manage to generate

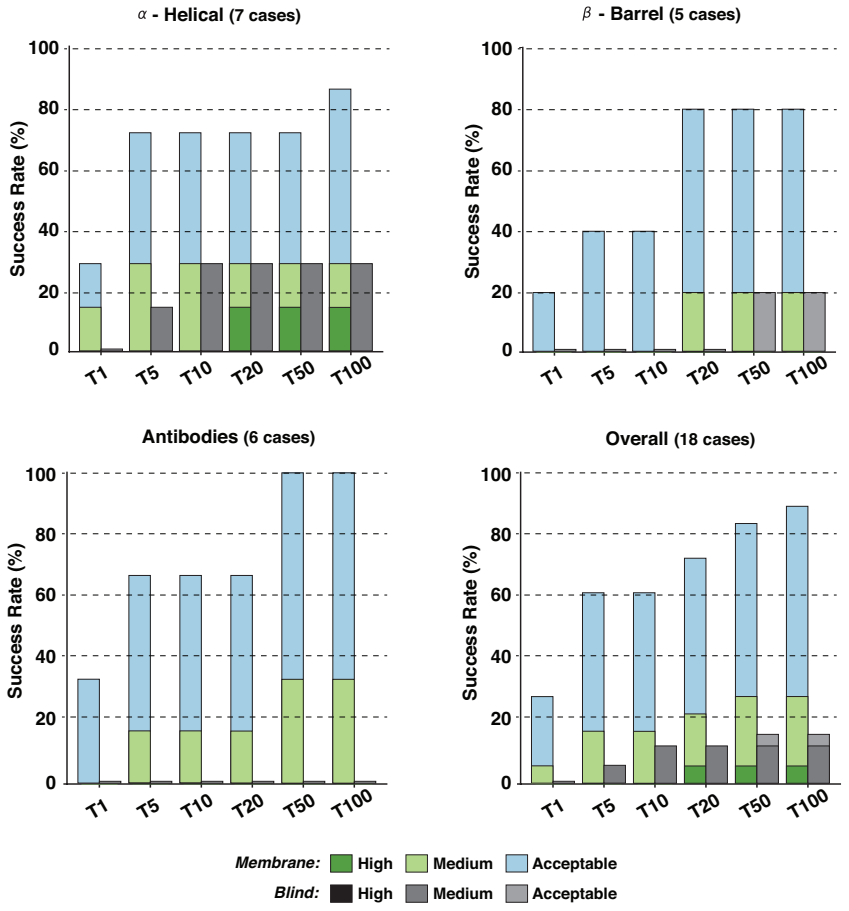
models with more than the 30% of the native contacts. For the remaining cases, acceptable models are found only for *2gsk* with the first near-native model ranked at the 39th position with a l-RMSD of 9.47Å. The top1 and top10 performance are similar to what has been reported for HADDOCK using a blind docking scenario[223]. Altogether, the results of the *Blind* predictions are considerable worse than that of our *Membrane* protocol in terms of both overall performance and CAPRI-based quality of the generated models. This clearly shows that the use of the membrane topological information to drive the modeling process has a significant impact on the docking performance.

3.4. Impact of different membrane definitions on the docking performance

The results presented so far have been obtained by either defining the membrane based on the phosphate beads positions taken from MemProtMD (Membrane) or by fully blind predictions (i.e. without any membrane). We investigate here how different definitions of the membrane might impact our docking protocol. For that purpose, we have generated two additional artificial bead representations of the membrane based on the average (Average) or minimum (Minimum) Z-axis coordinate provided by the equilibrated MemProtMD membrane model. We have compared the docking performance of those different membrane scenarios on the 18 cases from the membrane docking benchmark. As previously, we assess the performance in terms of success rate for each of the selected N tops (See Material and Methods; Metrics for the evaluation of model quality and success rate). For the sake of simplicity, we only report the success rate for acceptable or better models.

Figure 3. Performance of the membrane protein integrative modeling protocol on the 18 cases of the membrane docking benchmark.

Success rates are presented for each of the benchmark categories (α-Helical, β-Barrel and Antibodies) as well as Overall (including all three different categories). Color coding from blue to green (Membrane) and grayscale (Blind) indicates the model quality (from acceptable to high) as defined based on CAPRI criteria. Antibodies (the soluble partner is either an antibody or a nanobody).



On average, in the Membrane scenario our simulations have 99 ± 34 starting swarms (ranging from 32 to 170), while for Average and Minimum this increases to 134 ± 28 and 164 ± 34, respectively. This roughly translates into an increase of 7,000 and 13,000 in the number of *glowworms* (agents

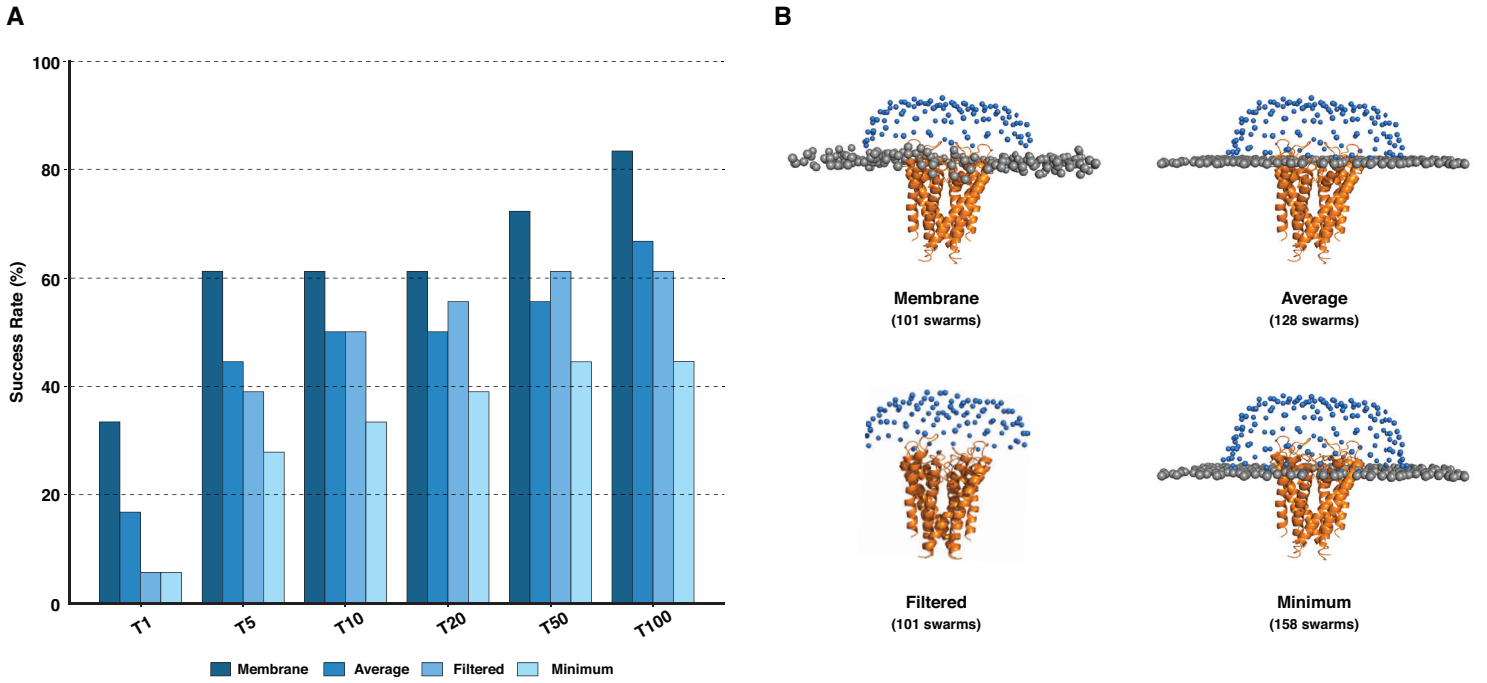
of the algorithm representing possible ligand poses) that are handled by the optimization algorithm during the sampling and scoring processes as compared to the *Membrane* scenario. As shown in Fig. 4A, for the two most representative tops (*T5* and *T10*) the success rates drop from 61.1% to 44.4% and 27.8% for *Average* and to 50% and 33.3% for *Minimum*. This pattern is observed along all selected top *N* models, which suggests that there is a negative correlation between the number of *swarms* (and *glowworms*) and the docking performance. This effect is expected, since the larger the pool of generated poses, the larger the number of possible false positives which can be selected by the scoring function. For this reason, the optimization of the poses might not always converge towards biological relevant states.

3.5. Penalizing models penetrating the membrane leads to better predictions

We have also investigated the effect of the scoring penalty on the optimization algorithm during the docking. To do so, we have designed an additional scenario (*Filtered*), in which the membrane was only used to initially filter the swarms over the receptor surface, but not considered for penalizing models penetrating the membrane (see Fig. 4B). In this case, the success rate of the top 10 is similar to that of *Average* scenario (50%). However, for higher tops such as *T1* and *T5*, the *Filtered* scenario performs considerably worse as compared to *Membrane* (5.5% and 38.8% vs 33.3% and 61.1% respectively). This clearly suggests that, while the membrane plays an important role to narrow the conformational search it has also a big impact on the scoring: First, it guides the optimization protocol towards more binding-competent regions and second, it helps identifying near-native states out of the pool of generated docked models.

Figure 4. Analysis of the impact of different membrane definitions onto the docking performance.

A) Bar plot of the performance of the different membrane setups on the 18 cases of the membrane protein docking dataset (i.e. before refinement with HADDOCK). The success rate is defined as the percentage of cases for which an acceptable or higher quality model was found within the selected top N.
B) Illustration of the different membrane setups on a representative case of the benchmark (1k4c).



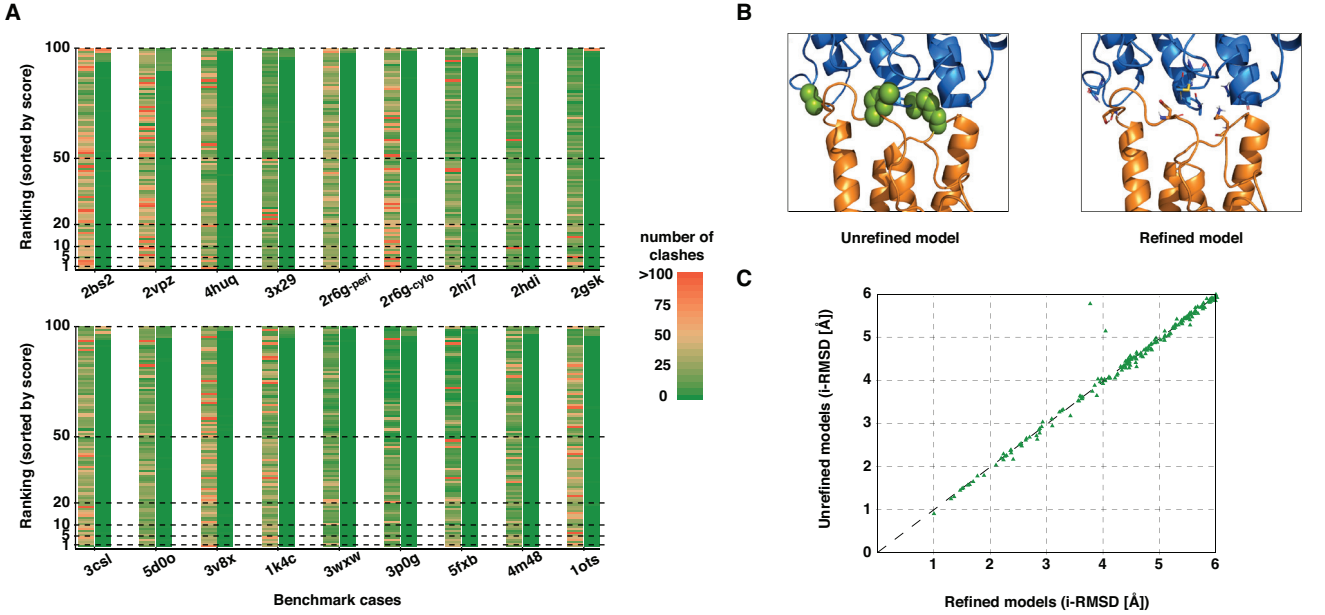
3.6 The structural quality of the docked models improves after HADDOCK refinement

We have assessed the quality of the docked models in terms of intermolecular steric clashes. To do so, we have quantified and compared the number of clashes (See Material and Methods; *Metrics for the determination of steric clashes in a protein complex*) present in our docked models before (LightDock only) and after refinement with HADDOCK. On average, the top 100 LightDock models have a significant number of clashes (28.5 ± 10.0) compared to those after refinement (0.6 ± 0.5) as shown in Fig. 5A. For some cases, *2bs2, 2gsk, 3csl* and *1ots*, few refined models ranked at positions ≥ 95 still have a moderate number of those (> 25), but these are penalized at the level of the HADDOCK score which ensures that clashing models will never be ranked at top ranking positions. Overall, this coarse-grained refinement protocol is able to refine and remove more than the 98% of the total number of clashes. As an example, a model before and after refinement is shown in Fig. 5B.

Ideally, a refined complex should not structurally deviate too much from its unrefined counterpart. If this is not the case, the refined interface might significantly differ from the predicted one and therefore loose a relevant predicted state. We have investigated whether our refined models differ from their starting conformations in terms of their interface RMSD of the backbone (i-RMSD). For this, we selected all LightDock models with an i-RSMD ≤ 6Å from the top 100 predictions for all cases (183 models in total) and compared them to their counterpart after refinement with HADDOCK. As shown in Fig. 5C, the vast majority of points are along the diagonal, which indicates that the backbone of the refined complexes has not significantly moved during the refinement. It is mainly the positions of side chains at the interface that have been optimized (See Fig. 5B and Fig. SI5.2

Figure 5. Analysis of the quality of the membrane-associated protein models before and after refinement with HADDOCK.

A) Stacked bar plot of the top 100 generated models for each of the benchmark cases (18 in total) ranked by their respective score (left – LightDock DFIRE docking score, right - HADDOCK score). For each complex the left bar corresponds to the unrefined models and the right bar to the refined models. The color coding (from green to red) indicates the number of clashes.
B) Illustration of a complex before and after refinement. Green spheres represent atomic clashes. The corresponding side chains are shown as sticks in the refined model.
C) i-RMSD comparison of all models with an i-RMSD ≤ 6Å before and after refinement (183 in total). Points above the diagonal indicate an improvement in i-RMSD value.



in Supplementary Information). Points above the diagonal, indicate models that have improved in terms of i-RMSD after refinement. The changes are however limited. Two models (from *2vpz* and *3csl*), however, show a significant improvement of 1.09Å and 1.85Å respectively. In summary, these results show that our coarse-grained refinement protocol is very efficient in removing steric clashes without compromising the quality of the backbone conformation of near-native models.

3.7. Using membrane topological information to drive the docking performs better than post-sampling filtering approaches

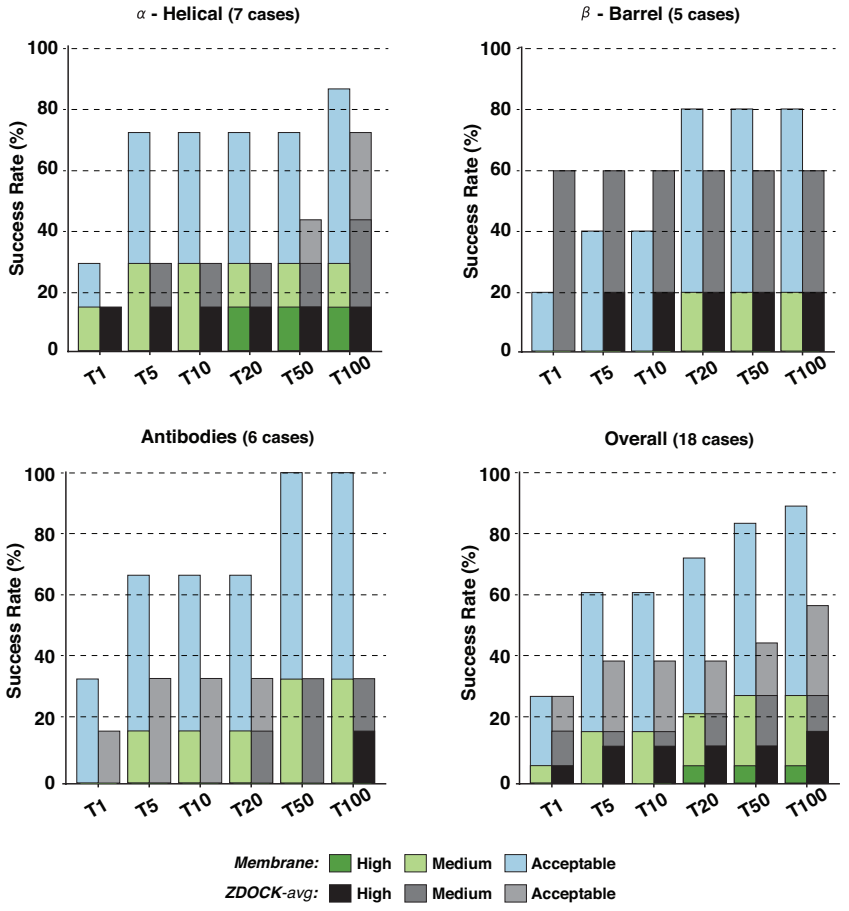
We have also analyzed how our membrane-driven protocol compares to other state-of-the-art docking software. To this end, we selected ZDOCK[212] as docking algorithm of reference for several reasons. First, it is a well-established docking program whose scoring protocol is being trained and continuously tested on a large and relatively heterogeneous benchmark of protein-protein complexes[145]. Second, it allows to mask regions not belonging to the interface. And third, its standalone version (3.0.2) is a fast and easy-to-use tool for systematic benchmarking. Despite that the current version of ZDOCK does not allow to use an explicit representation of the membrane, we have designed three different scenarios in which various levels of information are used to include information about the membrane. In order to mimic our *Membrane* scenario, we have masked all surface accessible residues below the maximum *z* coordinate provided by our membrane implementation (ZDOCK-*max*).

Similarly, to compare with our *Average* and *Minimum* scenarios, we have masked those residues below the average (ZDOCK-*avg*) or minimum (ZDOCK-*min*) *z* coordinate (Fig. SI5.1 in Supplementary Information). For the *Antibodies* subcategory, we have also masked all non CDR loops residues. Finally, we run the 18 cases of the membrane docking benchmark in fully blind (default) mode to define the baseline of ZDOCK.

The results for all those scenarios are shown in Fig. SI5.3 (Supplementary Information). The best performance is obtained for the ZDOCK-*avg* scenario. Comparing this scenario to our *Membrane* protocol shows that both protocols have an equivalent success rate of 27.7% for the top 1 model, but our protocol clearly performs best for *T5*/*T10* with 61.1% as compared to 38.8% for ZDOCK (Fig. 6). Our protocol reaches 88.8% of near-native models for the top 100 compared to the 55.5% for the best performing scenario in ZDOCK. These differences are more remarkable for α-Helical and *Antibodies* complexes with 71.4% and 66.6% for the top 5 (and top 10), respectively, compared to 28.5% and 33.3% for ZDOCK. In the case of β-Barrel, ZDOCK-*avg*, however, performs best with 60% in the top1 (3 out 5) while our protocol starts at 20% for top 1 to reach a maximum of 80% in *T20*. Based on the well-established CAPRI quality criteria (See Material and Methods; *Metrics for the evaluation of model quality and success rate*), ZDOCK builds high quality models for the 16.6% of the tested cases (3 out of 18, while only one high quality model is obtained in our case) with 2 of them ranked within the top 10 predictions. These two cases correspond to the β-Barrel complex with the smallest conformational change (*2hdi*; i-RSMD of 0.361Å between unbound and bound structures) and a α-Helical bound case *(4huq)*.

Figure 6. Comparison of the performance of the membrane protein integrative modeling protocol (*Membrane*) with the best ZDOCK scenario (ZDOCK-*avg*) on the 18 cases of the membrane docking benchmark.

Success rates are presented for each of the benchmark categories (α-Helical, β-Barrel and Antibodies) as well as Overall (including all three different categories). Color coding from blue to green (Membrane) and grayscale (ZDOCK-avg) indicates the model quality (from acceptable to high) according to CAPRI criteria.



4. Discussion

In this work, we have developed and tested a new integrative modeling protocol to build membrane-associated protein assemblies. The protocol, which specifically accounts for the topological information encoded in the membrane, combines the capability of the LightDock framework to discard non-binding regions prior to the docking, with an efficient coarse-grained refinement via HADDOCK to remove clashes. As previously demonstrated, including information during docking not only outperforms the scenario where data are only used to discard models (post-simulation approaches), but also reduces significantly the computational cost[12], in this particular case by an average factor of 75% over the 18 complexes considered. Our membrane-driven protocol shows a much better performance in generating native-like structures for the vast majority of the tested cases than when the membrane is neglected (Blind). It achieved this by both filtering the initial swarm configurations and including a membrane penalty term into the scoring which helps both the optimization algorithm and the scoring of the docked models. Altogether, our findings reinforce the well-accepted notion that the integration of (experimental) data, in this case membrane topological data, into the docking calculations improves the performance of modeling approaches.

We have also investigated how different ways of defining the membrane topological information across the z-axis affect the sampling of the conformational space. Our protocol performs best when an equilibrated and simulated bilayer is incorporated into the sampling. This limits the number of swarms (and therefore glowworms), which, in turns, allows the optimization algorithm to identify more biological relevant states compared to less restricted scenarios such as Average and Minimum. This behavior is explained by the fact that in LightDock, sampling and scoring are closely

interconnected since the optimization of the ligand poses (in rotational and translational spaces) is driven towards better scoring conformations. In other words, the reduction of potential false positives leads to an increase in the performance of the search algorithm.

We have analyzed the quality improvement of our docking predictions after refinement using a simple definition of steric clashes. We have shown that our refinement protocol leads to the removal of (almost) all clashes while keeping the backbone conformation almost unaltered, with no more than 0.25Å i-RMSD for the most altered conformations (rare cases). As a consequence of the refinement, the side chains might suffer from bad conformations introduced by the removal of clashes and move away from the native conformation in the complex. To check this, we have analyzed the impact of the coarse-grained refinement on the side chain i-RMSD and the fraction of native intermolecular contacts they form. As shown in Fig. SI5.2 (Supplementary Information), the refined models do not significantly loose native intermolecular contacts as estimated by the Fnat metric and their side-chain i-RMSDs even slightly improve.

Some knowledge of the putative binding interface is known to help the modeling of biomolecular interactions, often allowing to generate more accurate models. This information can come from a variety of experimental of bioinformatic data, such as, for example, NMR chemical shift perturbations, mutagenesis data, H/D exchange and crosslinking data obtained by mass spectroscopy or sequence conservation, among others[19]. Besides of the topological information encoded in the membrane, our protocol can also incorporate information about interfaces. As an example, we assumed that three interacting residues of the soluble partner are known and defined those in LightDock (See Table SI5.1 in Supplementary

Information). The results, shown in Fig. SI5.4, show that this does not significantly change the performance as the T100 remains constant (75% success rate) with a slight improvement in the top 10 predictions (75% as compared to 66.6%) and a slight decay in T5 and T1 (25% and 58.3% as compared to 33.3% and 66.6% for the *Membrane-rst* and *Membrane* categories respectively). It is worth noting that these variations are not that significant as they are caused by only a single case difference because of the limited size of the benchmark.

In this work, we have only focused on the modeling of membrane-associated protein assemblies. In cellular environments, however, some soluble proteins might associate with membranes in order to stabilize and/or carry out their function. These types of interactions have only been studied in a handful of systems such as signaling factors or nuclear receptors, due to the lack of more generic approaches that can be used to characterize a broader range of lipid-protein interactions[227]. Our work could be extended to build realistic models of membrane-associated protein complexes. This would require extra effort to develop a scoring function that accounts for protein-lipid interactions. Such membrane-specific scoring functions have been already shown appropriate for membrane protein structure prediction and design purposes[228] and might also represent a significant advance for membrane-associated protein docking protocols. Looking ahead, a larger benchmark set will enable broader energy function development and optimization, which should eventually cover protein-lipid interactions too. In terms of software integration, LightDock, as a sampling algorithm, could be included within the future modular version of the HADDOCK software and eventually offer an alternative to its default rigid-body sampling step. This would further extend HADDOCK modeling capabilities to account for the use of membrane-based bilayers.

Note that HADDOCK has already been used with explicit membranes (nanodisks or micelles) to study the binding and orientation of proteins onto the lipid surface[229–231]. These are however isolated cases and no systematic testing as performed here has yet been done.

In summary, we have developed an integrative modeling protocol for membrane-associated protein assemblies that accounts for the topological information provided by the membrane in the modeling process. It makes use of a membrane-derived bead bilayer during the sampling step with LightDock. Clashes resulting from the rigid-body docking are successfully removed by refinement with HADDOCK while preserving the quality of both backbone and side chains conformations at the interface. Importantly, while the present protocol only makes uses of the membrane to drive the modeling, it is fully compatible with the use of other sources of information such as mutagenesis and/or bioinformatic predictions in the form of residue restraints to further guide the docking.

Conclusions and Perspectives

"Outliers are opportunities"

Malcom Gladwell, Outliers: The story of success, 2008

This thesis addresses important, yet still open, challenges in the field of computational structural biology, namely: The integration of experimental information into the docking calculations (sampling/scoring and/or post-simulation filtering), the use of coarse-grained approaches for the modeling of large protein-protein and protein-DNA complexes and the integrative modeling of membrane-associated protein assemblies. The described research work has been inspired on the tremendous advances made by the structural biology community at large, and, for the computational parts, in particular by the work of the HADDOCK team over the years.

The advent of integrative modeling, however, cannot be understood without the progresses made in computing. Computers, as nowadays known, originate from the seminal work by Alan Turing in the 30's[232], who designed a universal machine capable of computing anything that is computable: The Turing machine. The first modern computers, a central processing unit (CPU) along with some kind of computer memory, date from the 40's and early 50's. One of the first large-scale electronic computers in the world was built at the Weizmann Institute in Israel: WEIZAC[233].This pioneer computer was later on replaced by GOLEM[233], in which the very first computer simulation of a protein system was performed in the mid 70's[7]. At that time, this task required a lot of manual work and electrical engineering/computing knowledge. Nowadays, workflows are broadly automatized but some tasks still require extensive programming skills.

Computational modeling approaches are often being developed as a complement to traditional structural biology experimental techniques. Therefore, it is reasonable to think that potential users might be either experimentalists or non-programming gurus. With the birth of the World Wide Web and all the advances in Web programming, it is essential to provide user-friendly Web-based services, which eventually will make those

computational approaches more accessible to the outside world. Albeit palpable, this does not represent a trivial task since it requires careful thinking and design as well as specialized work-force. In the context of docking, the new HADDOCK2.4 webserver represents a clear example of such online service, which is powered by a distributed cloud/grid computing infrastructure as well as local resources.

The development of novel protocols, such as the ones described in this thesis, usually requires systematic benchmarking to test their capabilities, performance and fine tune them. Over the past years, blind experiments such as CASP[234] (for protein prediction since 1994), CAPRI[235] (for macromolecular docking since 2000) or more recently, since 2015, the D3R Grand Challenge[236] (for small molecule docking and drug design) have acted as catalysts of development. These not only allow to evaluate the performance of the participating software in more realistic scenarios, but also provide new challenges to the community, hence driving innovation. In this sense, they represent an excellent framework to further develop software and tackle the most recent structural biology problems. In fact, the last assessed CAPRI experiment (rounds 38-45)[237], offered the perfect opportunity to test the coarse-grained implementation in HADDOCK. Target136, the largest target featured so far, was composed of ten protein monomers arranged in two pentameric substructures (double doughnut-like). By extracting interface restraints from a template model and applying C5 symmetry, we were able to accurately (~60% of native contacts) recreate the full decamer using five coarse-grained dimers as starting models, reducing the computational cost from 10-12 to 2 hours per generated full structure[238].

Along these lines, this thesis has focused on the modeling of biomolecular interactions by using different resolutions. The first chapter reviewed a number of different representative coarse-grained/hybrid approaches and

force field parametrization strategies. It starts with a brief historical overview on the early developments of the coarse-graining and docking fields, dating from the early 70s. In the case of hybrid systems, i.e. those mixing all-atom and coarse-grained representations simultaneously, the integration of experimental data might play an even more crucial role into the modeling process. If incorporated into the scoring function, the data might work as a firewall and alleviate the inaccuracies of hybrid schemes in terms of intra- and inter-molecular interactions. Several application examples of integrative modeling of protein interactions were described in this thesis, together with a detailed list of the available software for building structural models of macromolecular multi-subunit complexes.

The second chapter showed how experimental information can be integrated into docking in the context of the LightDock software[10]. This builds upon the well-established concept that the use of data leads to more accurate docking predictions. In general, data might be used at the various different stages (sampling, scoring or both) and prior, during and/or after the calculations. The new version of LightDock[12] incorporates such information in the form of residue restraints to, first narrow the conformational space prior the simulation, and second bias the scoring during the optimization process by measuring the amount restraints satisfaction. This protocol was demonstrated to perform better than purely post-simulations approaches, still able to yield valuable predictions with partially incorrect data.

The third chapter detailed the implementation of a coarse-grained docking protocol[13] into the information-driven software HADDOCK[8], based on the MARTINI coarse-grained force field for proteins[14]. The implementation focusses on three key aspects: (1) The topology description and parametrization for each of the amino acids, (2) the adjustment of the atomic solvation parameters to the coarse-grained particles used to calculate the desolvation

energy and (3) the design of an inverse mapping protocol to restore atomistic resolution to the generated docked complexes. The coarse-grained protocol[13], systematically tested on the largest complexes from the well-established protein-protein Docking Benchmark 5[145], yields substantially more biologically relevant solutions as compared to standard atomistic calculations, while maintaining similar accuracy and a remarkable speed increase. Finally, its capabilities are illustrated with the modeling of the heptameric KaiC:KaiB (1:6) complex, integrating mutagenesis and hydrogen-deuterium exchange data from mass spectrometry with symmetry restraints.

The fourth chapter described the integration of the MARTINI force field for DNA[16] into the HADDOCK coarse-grained protocol[13]. It leans on the implementation for proteins detailed in chapter three, and includes specific considerations to account for Watson-Crick interactions. The protocol was designed to automatically detect hydrogen-bonding base pairs and uses a special set of parameters and restraints for those during the docking[15]. Much like in the original implementation for protein-protein docking, this protocol shows a considerable speed increase and maintains similar accuracy as compared to standard atomistic calculations. As a proof of concept, a coarse-grained integrative model of the PRC1 ubiquitination module bound to the nucleosome was built, defining one ambiguous distance restraint between the catalytic pocket of PRC1 and the H2A in combination with mutagenesis data derived from the literature.

The last chapter of this thesis combined various developments described in chapters two and three[12,13] into a novel protocol for the integrative modeling of membrane-associated protein assemblies. This chapter detailed the use of an explicit membrane-based coarse-grained bilayer during docking, which allows to: (1) Narrow the conformational search towards the binding competent regions, (2) lead the optimization algorithm to putative

relevant states, and (3) help identify native poses out of the pool of generated models. The docked models undergo further refinement for clearing clashes and improving the contacts at the interface. This protocol might serve as a tool to build more realistic models of membrane-associated protein complex and opens the route to the systematic study of protein-lipid interactions in the near future.

The challenges addressed in this thesis will hopefully help shedding light onto still dark fractions of the structural biology field. In the case of large molecules, our coarse-grained implementation in HADDOCK offers a valuable tool for the modeling of large protein-protein[13] and protein-nucleic acids[15] complexes. Our back-mapping protocol to convert models to atomistic resolution appears appropriate for removing clashes at interfaces while preserving the correctly predicted geometries. In terms of sampling, using experimental information to narrow the conformational search in LightDock translates into a remarkable increase in predicting power as compared to *ab initio* or post-simulation filtering approaches[12]. Moreover, this approach was extended and demonstrated useful for modelling membrane-associated assemblies, specifically those whose interface lie between a membrane-embedded protein and their soluble counterpart.

Since the early days of docking, the modeling of protein-protein, protein-ligand and, in a lesser degree, protein-nucleic acids have attracted most of the attention within the computational structural biology field. Undoubtedly, proteins are the workhorses of the cellular machinery and their interactions one to another mediate a wide range of biological functions. Protein-ligand docking has a very special place in the general field of docking, due to its direct applications in drug discovery and therefore, in human health. In the case of nucleic acids, on the one hand, protein-DNA interactions play essential roles in cellular processes such as gene expression, regulation,

transcription, DNA repair, or chromatin packaging in eukaryotes. On the other hand, protein-RNA interactions are responsible of many post-transcriptional processes such as splicing regulation as well as have important functions in the regulation of gene expression. Nevertheless, there are still open challenges to address such as the modeling of intrinsically disordered proteins (IDPs) and their interactions as well as other types of biomolecular interactions, for instance, protein-lipids and protein-glycans interactions.

In nature, a large fraction of proteomes comprises proteins that either entirely lack a well-defined three-dimensional structure (IDPs) or contain disordered regions (segments) (IDPRs – Intrinsically Disorder Protein Regions). These *"exceptionally abundant exceptions"*[239], namely IDPs/IDPRs, are involved in multiple biological functions and signaling processes. This intrinsic disorder is known to contribute to protein promiscuity and binding diversity, leading to *one-to-many* and *many-to-one* protein interactions[240]. As a result of their high degree of flexibility, when dysregulated, these disordered proteins are prone to form unwanted interactions, which can lead to severe pathological conditions including neurodegenerative diseases such as Alzheimer and Parkinson. It Is estimated that at least 15% of protein-protein interactions (PPIs) involve IDPs[88], the so-called disordered protein-protein interactions, i.e. an IDP bound to a folded protein. Due to their tendency to form weak and transient interactions as well as the highly flexible nature of IDPs, experimental techniques often have difficulties to determine their 3D structure. As such, computational modeling approaches might offer a valuable alternative.

Generally speaking, a computational model requires an (accurate) representation of the system and a set of parameters and energy functions to calculate intra- and inter-molecular interactions. The latter is commonly referred to as force field.Older versions of traditional atomistic force fields

have been shown to be biased towards α-helical or β-sheet structures[241]. However, over the past years, it is noticeable the improvement of these mathematical models to better describe important structural and dynamical properties. These include, for instance, more accurate representations of backbone conformations for both, folded proteins[242–245]. In terms of docking, current rigid-body and flexible docking methods have limitations in modeling such disordered protein-protein interactions, especially when some folding occurs upon binding. Along this line, it has been recently shown that a fragment-based docking approach might be better suited for this task as compared to protocols using the full structure of the disordered component[88]. While all these advances provide a useful toolbox to gain deeper insights into these biomolecules, and eventually design better therapeutic strategies, there are still a number of unresolved questions. For example, it still remains unclear whether fine-tuning already existing force fields represents a better strategy as a whole, versus (re)parametrizing entire force fields[249,250]. In addition, the prediction of when such conformational changes will take place, is still an open challenge in the field.

Recently, other types of biomolecular interactions have been gaining interest, such as protein-lipids interactions involved in basic structural roles as well as in highly regulated signaling events[251]. Although the importance of these interactions was recognized more than a decade ago[252], there is an emerging trend of new computational models and experimental works as protein-lipid interactions directly applies, for instance, to the design and characterization of antimicrobial peptides and therefore to antibiotic research[253]. Another type of biomolecular interactions which has historically remained in the background, are protein-glycans interactions. These are getting again increased attention because of the most devastating pandemic of the last century that the world is currently facing: The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) causing the disease known as

COVID-19. According to the World Health Organization, in October 2020 COVID-19 has been diagnosed in over 39 million people with a death rate of 2.79% approximately (>1M confirmed deaths). Since its eruption, the scientific community has joined forces to understand the biology of its infection mechanism. It is known that SARS-CoV-2 is a lipid-enveloped RNA virus that contains a set of structural proteins, among which the spike, or *S*, protein is likely the most critical one. It mediates the host cell entry by binding to the angiotensin-converting enzyme (ACE2). As many other viruses, SARS-CoV-2 utilizes a glycan shield to protect itself from the host immune response[254]. Further, those glycans might also play a role on promoting receptor binding and recent computational studies have stressed the importance of this glycan shield within the context of neutralizing antibodies[255]. All of these findings should also stimulate a flurry activity towards the developing of more specific computational methods for the functional and structural characterization of protein-glycans interactions.

Viruses, albeit biologically simple, are usually very large in size. As an illustration, influenza A viral envelope contains roughly 160 million atoms and spans around 1,150Å in diameter. A recent work showed that it is nowadays possible to study such a big systems by mesoscale all-atom simulations and provide insights into their binding mechanisms[256]. In general, the size of the system under study and the choice of the model used to represent it determine what is achievable by computer simulations. Quantum simulations typically allow to study events at the picometer to nanometer scale. Nowadays, atomistic simulations are the most common approach to study not only local motions such as side-chain arrangements, but also larger conformational re-arrangements as well as folding and unfolding events. For larger scales, coarse-grained approaches have been applied to study not only microscopic events such as global motions or

protein folding but also macroscopic events like aggregation. Moreover, multiscale models, as refer to those models combining various resolution levels simultaneously, have been proved useful to unravel critical mechanistic insights into both, Influenza[257] and the above mentioned SARS-CoV-2[258] viruses.

Thus, coming back to the initial question in the title of this thesis, "*Does size matter?* ", as stated in the general introduction of this thesis, it is crucial to design the right model to address specific research questions, where the size of systems is definitely a major player. However, with the fast advances in both theory and computing (software and hardware), the gap between size and thorough statistical exploration of the conformational and interaction landscapes, is expected to be eventually bridged in a not too far future.

Supplementary Information

Supplementary Information – Chapter 3. *Less is more: Coarse-grained integrative modeling of large biomolecular assemblies with HADDOCK.*

Table SI3.1. Backbone Particle Types.

Amino acid	Coil	Helix	Extended	Turn
ALA	F4	HPa/HP0/HP5/Hda	B0	T0
All others	F5	H0/H5/Hd/Ha	Bda	Tda

Table SI3.2. Backbone-Backbone Relations. D: Bond Length (Å). f: Bond Angle (°).Y: Bond Dihedral (°). K: Force Constant (kcal.mol-1).

Backbone SS	D_{BB}	K_{BB}	O_{BBB}	K_{BBB}	Y_{BBBB}	K_{BBB}
Coil	3.5	12.5	127	5.971	-	-
Helix	3.1	12.5	96*	167.184	-120	95.524
Extended	3.5	12.5	134	5.971	0	0.57
Turn	3.5	12.5	100	5.971	-	-

** f_{BBB} = 98 ° and K_{BBB} = 23.883 Kcal.mol-1 for PRO in Helix conformation.*

Table SI3.3. Polar and charged amino acid and corresponding parameters beads including "fake-beads" for a better description of electrostatics. BB denotes Backbone-Backbone while BS Backbone-Side-chain. Q: Charge (electron charge units). D: Bond Length (Å). f: Bond Angle (°).

Amino acid	SC bead name	Q (SCd)	D_{bs},D_{ss},D_{ss}	K_{bs},K_{ss},K_{ss}	f_{bss}	K_{bss}
ARG	RN0-AQd-SCd	+1.00	3.30,3.4,1.10	50,50,500	180	5.971
LYS	KC3-KQd-SCd	+1.00	3.30,2.80,1.10	50,50,500	180	5.971
ASN	Nda-SCd-SCd	±0.46	3.20,1.10,2.80	50,500,500	-	-
GLN	QNda-SCd-SCd	±0.46	4.00,1.10,2.80	50,500,500	-	-
SER	SN0-SCd-SCd	±0.40	2.50,1.10,2.80	75,500,500	-	-
THR	TNda-SCd-SCd	±0.31	2.50,1.10,2.80	75,500,500	-	-
GLU	Qa-SCd	-1.00	4.00,1.10	50,500	-	-
ASP	DQa-SCd	-1.00	3.20,1.10	75,500	-	-

Table SI3.4. Side-chains amino acid dependent and corresponding parameters. BB denotes Backbone-Backbone while BS Backbone - Side chain.

Amino acid	SC bead name	D_{BS},D_{SS},D_{SS},D_{SS}	K_{BS},K_{SS},K_{SS}, K_{SS}	f_{BSS},f_{BSS},f_{SSS},f_{SSS}	K_{BSS},K_{BSS},K_{SSS},K_{SSS}	Y_{BSSS}, Y_{SSSS}	K_{BSSS}, K_{SSSS}
TRP	WC4-SNd-SC5-SC5	3.00,2.70,2.70,2.70	50,500,500,500	210,90,50,50	11.942, 5.971,11.942,11.942	0, 0	11.942,11.942,47.767
TYR	YC4-SC4-SP1	3.20,2.70, -	50,500,500, -	150,150, -	11.942,11.942, -	0, -	11.942, -
PHE	FC5-SC5-SC5	3.10,2.70, -	75,500,500, -	150,150, -	11.942,11.942, -	0, -	11.942, -
HIS	HC4-SP1-SP1	3.20,2.70, -	75,500,500, -	150,150, -	11.942,11.942, -	0, -	11.942, -
CYS	C5	3.10, -	75, -	-	-	-	-
ILE	AC1	3.10, -	12.5, -	-	-	-	-
LEU	LC1	3.30, -	75, -	-	-	-	-
MET	MC5	4.00, -	25, -	-	-	-	-
PRO	PC3	3.00, -	75, -	-	-	-	-
VAL	AC2	2.65, -	12.5, -	-	-	-	-
ALA	-	-	-	-	-	-	-
GLY	-	-	-	-	-	-	-

D: Bond Length (Å). f: Bond Angle (°).Y: Bond Dihedrals and Improper (°).

Table SI3.5. Selected complexes (27) for the CG protein-protein benchmark classified in the Protein-Protein Benchmark version 5.0 as: enzyme-inhibitor, enzyme-substrate, enzyme complex or others. (PART I)

Complex	# Atoms	# Residues	MW	Resolution	Receptor	Resolution	Ligand	Resolution
1azs: Adenylyl Cyclase – GTPgammaS	5740	718	77020	2,3	1ab8	2,2	1azt	2,3
1de4: Hfe – Transferrin r	13107	1641	172501	2,8	1a6z	2,6	1cx8	3,2
1exb: T1 β – K+ channel	13256	1668	177850	2,1	1qrq	2,8	1qdv	1,6
1gp: Protein G trimer	5781	737	77145	2,3	1gia	2	1tbg	2,5
1gxd: Prommp2 – Timp2	6466	816	85521	3,1	1ck7	2,8	1br9	2,1
1h1v: Actin – Gelsolin	5414	695	71583	2,99	1ijj	2,85	1d0n	2,5
1he8: Ras – Pi3 γ kinase	7396	915	98167	3	821p	1,5	1e8z	2,4
1ib1: 14-3-3 protein – N-acetylase	5046	632	70414	2,7	1qjb	2	1kuy	2,4
1kxp: Actin – Vitamin D	6167	787	82409	2,1	1ijj	2,85	1kw2	2,15
1n2c: Nitrogenase	20058	2548	265698	3	3min	2,03	2nip	2,2
1rlb: Transthyretin – Retinol	5171	660	68709	3,1	2pab	1,8	1hbp	1,9
1t6: Anthrax – Anthrax receptor	6695	846	88168	2,5	1acc	2,1	1shu	1,5
1wdw: Tryptophan synthase	7848	1011	103279	3	1v8z	2,21	1geq	2

Table SI3.5. Selected complexes (27) for the CG protein-protein benchmark classified in the Protein-Protein Benchmark version 5.0 as: enzyme-inhibitor, enzyme-substrate, enzyme complex or others. (PART II)

1y64: Actin – Bni1	6119	767	81284	3	2fxu	1,35	1ux5	2,5
2ajf: Ace2 – SARS	6273	771	82831	2,9	1r42	2,2	2ghv	2,2
2fju: Phospolipase β2 – Rac	6993	873	92947	2,2	2zkm	1,62	1mh1	1,38
2gaf: Vp55 – Vp39	5929	723	79736	2,4	3owg	2,86	1vpt	1,8
2oor: NAD(p) α – NAD(p) β	6755	915	90178	2,32	1l7e	1,9	1e3t	NMR
3aaa: Actin – Myotrophin	5033	634	66478	2,2	3aa7	1,9	1myo	NMR
3biw: Neuroglin	5545	710	72968	3,5	3bix	2,61	2r1d	2,6
3l89: AD21 – CD46	5252	677	69446	3,5	3l88	2,5	1ckl	3,1
3lvk: IscS – TusA	6641	850	88578	2,442	3lvm	2,33	1dcj	NMR
3r9a: Alanine-Glyoxilate AT – pex5p	8199	1063	108889	2,35	1h0c	2,5	2c0m	2,5
4gam: Methane monooxygenase	18499	2262	243436	2,902	1xvb	1,8	1ckv	NMR
4h03: Ia – Actin α	6152	770	82388	1,75	1giq	1,8	1ijj	2,85
4jcv: RecR – RecO	7529	993	99383	3,34	1vdd	2,5	1w3s	2,4
4lw4: CsdA – CsdE	7095	935	93969	2,01	4lw2	1,8	1ni7	NMR

Table SI3.6. Detailed list of residues, identified by mutagenesis experiments in combination with hydrogen-deuterium exchange and mass spectroscopy (HDX-MS), used as "active" in HADDOCK-CG to drive the simulations (CI/CII).

Protein	Domain	Residues
KaiC	C I	Gly101, Leu103, Ile105, Leu106, Asp107, Ala108, Pro110, Asp111, Pro112, Glu113, Gly114, Gln115, Glu116, Val117, Val118, Gly119, Asp122, Leu123, Ser124, Ala125, Leu126, Ile130, Ala133, Ile134
	C II	Met449, Ser450, Arg451, Ala452, Ile453, Asn454, Val455, Phe456, Lys457, Met458, Arg459, Gly460, His463, Asp464, Lys465, Ala466, Ile467, Arg468, Glu469, Phe470
KaiB	Thr7, Asn17, Thr18, Pro19, Glu33, Glu35, Gly38, Lys43, Leu48, Lys49, Pro51, Gln52, Glu55, Glu56, Lys58, Leu60, Pro70, Pro71, Pro72, Val73, Arg74, Ile77, Ser81, Asn82, Glu84, Lys85, Ile88	

Table SI3.7. Paired i-RMSD values (Å) calculated after cross-superimposition of the two 6-fold rings in KaiC.

KaiC	A - 1	A - 2	A - 3	A - 4	A - 5	A - 6
A - 1	-	0.96	1.29	1.9	0.8	0.89
A - 2	0.96	-	1.23	1.89	0.91	1.01
A - 3	1.29	1.23	-	1.28	0.89	1.25
A - 4	1.9	1.89	1.28	-	1.4	1.39
A - 5	0.8	0.91	0.89	1.4	-	0.92
A - 6	0.89	1.01	1.25	1.39	0.92	-

Table SI3.8. Structural similarity assessment of the top 4 models of coarse-grained HADDOCK (from the best cluster) with respect to the cryo-EM (backbone only) model (PDB ID: 5N8Y). B/C/D/E/F/G correspond to the 6 KaiB monomers, respectively, docked onto KaiC. i-RMSD, l-RMSD and FNAT are calculated according to CAPRI criteria.

KaiB subunits	i-RMSD [Å]	l-RMSD [Å]	Fnat
OVERALL	10.1 ± 2.8	5.9 ± 1.3	0.09 ± 0.05
B	8.4 ± 2.2	3.3 ± 0.4	0.07 ± 0.07
C	8.3 ± 2.4	3.4 ± 0.4	0.05 ± 0.06
D	8.8 ± 2.1	3.5 ± 0.4	0.05 ± 0.07
E	8.3 ± 2.2	3.4 ± 0.4	0.11 ± 0.08
F	8.4 ± 2.0	3.3 ± 0.4	0.05 ± 0.05
G	6.7 ± 2.8	3.3 ± 0.4	0.12 ± 0.06

Table SI3.9. Cluster based statistics for the CI and CII docking runs based on the fraction of common contacts (0.5 cutoff). HADDOCK score single terms averaged over the top 4 members of each cluster are reported and clusters are ordered according to the averaged HADDOCK score (a.u.). E_{vdw}: Lennard-Jones potential. E_{elec}: Coulomb potential. E_{AIR}: Ambiguous interaction restraints energy. E_{desolv}: Empirical desolvation score. BSA: Buried surface area. $E_{symmetry}$: Symmetry restraints energy.

Cluster	Population	E_{vdw}	E_{elec}	E_{AIR}	E_{desolv}	BSA	$E_{symmetry}$	HADDOCK score	l-RMSD [Å]
CI Domain									
1	15	-366.3 $_{\pm 28.4}$	-809.3 $_{\pm 153}$	2899.2 $_{\pm 126.6}$	12.3 $_{\pm 12.3}$	11598.3 $_{\pm 799.5}$	91.5 $_{\pm 36.8}$	-216.7 $_{\pm 13.2}$	5.9 $_{\pm 1.3}$
3	9	-343.6 $_{\pm 18.1}$	-917.3 $_{\pm 178.2}$	2946 $_{\pm 103.3}$	29 $_{\pm 19.1}$	10805.3 $_{\pm 611}$	114.2 $_{\pm 53.7}$	-191.9 $_{\pm 30.5}$	21.1 $_{\pm 6.6}$
9	4	-327.8 $_{\pm 18.7}$	-1038.3 $_{\pm 145}$	3076.3 $_{\pm 350.6}$	27.6 $_{\pm 32.8}$	10905.4 $_{\pm 391.4}$	88.2 $_{\pm 19.6}$	-191.3 $_{\pm 43}$	18.3 $_{\pm 6.5}$
2	10	-329.3 $_{\pm 16.8}$	-1033.3 $_{\pm 122.7}$	3303.7 $_{\pm 231}$	36.3 $_{\pm 20}$	10669.7 $_{\pm 890.1}$	134.9 $_{\pm 63.5}$	-160.3 $_{\pm 16.9}$	19 $_{\pm 5.3}$
6	4	-304.9 $_{\pm 39.4}$	-897.2 $_{\pm 116.6}$	3154.9 $_{\pm 250.8}$	6.9 $_{\pm 18.2}$	10432.8 $_{\pm 996.9}$	78.2 $_{\pm 5.5}$	-154.1 $_{\pm 52.6}$	16.8 $_{\pm 2.9}$
CII Domain									
2	4	2206.6 $_{\pm 214.5}$	-265 $_{\pm 27.4}$	21375.2 $_{\pm 5169.3}$	-28.3 $_{\pm 77.8}$	14969.5 $_{\pm 5869.1}$	156.3 $_{\pm 45.48}$	+44.5 $_{\pm 19}$	-

Table SI3.10. Number of acceptable or higher quality models, for each of the protein docking benchmark complexes, generated at the rigid-body (it0) stage of coarse-grained and standard all-atom HADDOCK docking runs in the absence of information to drive the docking (ab-initio mode). 10000 models were generated in the case of ab-initio docking. (PART I)

Complex	Protocol	Top 200	Top 400	Total
1azs	Coarse-grained	0	0	4
	All-atom	0	0	1
1de4	Coarse-grained	0	0	4
	All-atom	0	0	0
1exb	Coarse-grained	0	0	0
	All-atom	0	0	0
1gp2	Coarse-grained	0	1	4
	All-atom	2	3	3
1gxd	Coarse-grained	1	1	2
	All-atom	1	1	1
1h1v	Coarse-grained	0	0	0
	All-atom	0	0	0
1he8	Coarse-grained	4	4	4
	All-atom	0	0	6
1ib1	Coarse-grained	0	0	0
	All-atom	0	0	0
1kxp	Coarse-grained	2	2	2
	All-atom	0	0	2
1n2c	Coarse-grained	0	0	0
	All-atom	0	0	0
1rlb	Coarse-grained	0	0	3
	All-atom	0	0	0
1t6b	Coarse-grained	2	2	17
	All-atom	0	0	19
1wdw	Coarse-grained	0	0	4
	All-atom	3	3	3
1y64	Coarse-grained	0	0	0
	All-atom	0	0	0

Table SI3.10. Number of acceptable or higher quality models, for each of the protein docking benchmark complexes, generated at the rigid-body (it0) stage of coarse-grained and standard all-atom HADDOCK docking runs in the absence of information to drive the docking (ab-initio mode). 10000 models were generated in the case of ab-initio docking. (PART II)

2ajf	Coarse-grained	0	0	1
	All-atom	0	0	0
2fju	Coarse-grained	0	0	4
	All-atom	0	0	1
2gaf	Coarse-grained	0	0	0
	All-atom	0	0	2
2oor	Coarse-grained	0	0	0
	All-atom	0	0	0
3aaa	Coarse-grained	0	0	1
	All-atom	0	0	0
3biw	Coarse-grained	0	0	5
	All-atom	0	0	5
3l89	Coarse-grained	1	1	1
	All-atom	0	0	2
3lvk	Coarse-grained	2	2	5
	All-atom	1	1	2
3r9a	Coarse-grained	0	0	0
	All-atom	0	0	0
4gam	Coarse-grained	0	0	0
	All-atom	0	0	0
4h03	Coarse-grained	2	2	7
	All-atom	3	4	4
4jcv	Coarse-grained	1	1	2
	All-atom	0	0	1
4lw4	Coarse-grained	0	0	4
	All-atom	1	1	1
TOTAL	**Coarse-grained**	**15**	**16**	**74**
	All-atom	**11**	**13**	**53**

Table SI3.11. Number of acceptable or higher quality models, for each of the protein docking benchmark complexes, generated at the rigid-body (it0) stage of the coarse-grained and standard all-atom HADDOCK docking runs using true interface information to drive the docking. 1000 models were generated in the case of information-driven docking. (PART I)

Complex	Protocol	Top 200	Top 400	Total
1azs	Coarse-grained	170	270	332
	All-atom	65	117	213
1de4	Coarse-grained	2	6	120
	All-atom	137	264	353
1exb	Coarse-grained	36	73	143
	All-atom	33	69	185
1gp2	Coarse-grained	111	148	155
	All-atom	152	206	215
1gxd	Coarse-grained	102	148	184
	All-atom	66	118	146
1h1v	Coarse-grained	0	0	0
	All-atom	8	26	48
1he8	Coarse-grained	157	285	491
	All-atom	53	63	212
1ib1	Coarse-grained	0	0	0
	All-atom	1	1	9
1kxp	Coarse-grained	199	395	563
	All-atom	188	337	511
1n2c	Coarse-grained	0	1	4
	All-atom	34	70	183
1rlb	Coarse-grained	160	320	618
	All-atom	190	357	734
1t6b	Coarse-grained	184	367	803
	All-atom	143	324	667
1wdw	Coarse-grained	200	397	680
	All-atom	200	399	620
1y64	Coarse-grained	0	0	0
	All-atom	0	0	0

Table SI3.11. Number of acceptable or higher quality models, for each of the protein docking benchmark complexes, generated at the rigid-body (it0) stage of the coarse-grained and standard all-atom HADDOCK docking runs using true interface information to drive the docking. 1000 models were generated in the case of information-driven docking. (PART II)

2ajf	Coarse-grained	66	156	412
	All-atom	111	257	422
2fju	Coarse-grained	126	253	694
	All-atom	165	347	703
2gaf	Coarse-grained	131	250	573
	All-atom	157	266	570
2oor	Coarse-grained	17	18	25
	All-atom	38	43	46
3aaa	Coarse-grained	52	132	289
	All-atom	128	230	369
3biw	Coarse-grained	107	245	627
	All-atom	3	21	218
3l89	Coarse-grained	109	230	488
	All-atom	132	253	430
3lvk	Coarse-grained	199	386	708
	All-atom	188	301	411
3r9a	Coarse-grained	152	327	790
	All-atom	113	260	725
4gam	Coarse-grained	0	0	0
	All-atom	0	0	0
4h03	Coarse-grained	186	376	606
	All-atom	90	215	454
4jcv	Coarse-grained	173	250	289
	All-atom	198	252	266
4lw4	Coarse-grained	27	33	95
	All-atom	109	144	186
TOTAL	**Coarse-grained**	**2666**	**5066**	**9689**
	All-atom	**2702**	**4940**	**8896**

Supplementary Information – Chapter 4. *MARTINI-based protein-DNA coarse-grained HADDOCKing.*

Table SI4.1. Nucleotide particle types.

Nucleotide	Backbone	Side-chains	H-bonding
ADE	NB1-NB2-NB3	ANS1-ANS1-ANS3-ANS2	NH1-NH2
GUA	NB1-NB2-NB3	GNS1-GNS3-GNS4-GNS2	NH3-NH4
THY	NB1-NB2-NB3	TNS1-TNS4-TNS2	NH5-NH6
CYT	NB1-NB2-NB3	CNS1-CNS4-CNS3	NH7-NH8

Table SI4.2. Nucleotide particle relations.

A) Adapted bond length parameters and corresponding force constants. D: Bond Length (Å). K: Force Constant (kcal.mol^{-1}).

Nucleotide	$D_{BB1-BB2}$ (K)	$D_{BB2-BB3}$ (K)	$D_{BB3-SC1}$ (K)	$D_{SC1-SC2}$ (K)	$D_{SC2-SC3}$ (K)	$D_{SC2-SC4}$ (K)	$D_{SC3-SC4}$ (K)	$D_{SC4-SC1}$ (K)
ADE	3.6 (47.87)	1.98 (191.479)	3.0 (71.805)	2.29 (500.0)	2.66 (500.0)	3.26 (47.87)	2.88 (500.0)	1.62 (500.0)
GUA	3.6 (47.87)	1.98 (191.479)	3.0 (71.805)	2.95 (500.0)	2.95 (500.0)	3.89 (47.87)	2.85 (500.0)	1.61 (500.0)
THY	3.6 (47.87)	1.98 (191.479)	2.7 (71.805)	2.17 (500.0)	3.22 (500.0)	-	-	2.65* (500.0)
CYT	3.6 (47.87)	1.98 (191.479)	2.7 (71.805)	2.2 (500.0)	2.85 (500.0)	-	-	2.68* (500.0)

$*D_{SC3-SC1}$

B) Adapted bond angle parameters and corresponding force constants. f: Bond Angle (°). K: Force Constant (kcal.mol-1).

Nucleotide	$f_{BB1-BB2-BB3}$ (K)	$f_{BB2-BB3-SC1}$ (K)	$f_{BB3-SC1-SC2}$ (K)	$f_{BB3-SC1-SC4}$ (K)	$f_{SC1-SC2-SC3}$ (K)	$f_{SC2-SC1-SC4}$ (K)	$f_{SC2-SC3-SC4}$ (K)	$f_{SC3-SC4-SC1}$ (K)
ADE	110 (47.87)	94 (59.84)	160 (47.87)	140 (47.87)	85 (47.87)	125 (47.87)	74 (47.87)	98 (47.87)
GUA	110 (47.87)	94 (59.84)	137 (71.8)	130 (59.84)	69 (47.87)	125 (47.87)	84 (47.87)	94 (47.87)
THY	110 (47.87)	92 (52.66)	107 (71.8)	145* (71.8)	55 (23.93)	83** (23.93)	42*** (23.93)	-
CYT	110 (47.87)	95 (50.26)	95 (71.8)	150* (71.8)	61 (47.87)	71** (47.87)	47*** (47.87)	-

$*f_{BB2-SC1-SC3}$, $**f_{SC2-SC1-SC3}$, $***f_{SC2-SC3-SC1}$

C) Adapted bond dihedral parameters and corresponding force constants. Y: Bond Dihedral (°). K: Force Constant (kcal.mol-1).

Nucleotide	$Y_{BB1-BB2-BB3-SC1}$ (K)	$Y_{BB2-BB3-SC1-SC2}$ (K)	$Y_{BB2-BB3-SC1-SC4}$ (K)
ADE	-90 (0.05)	-116 (0.0)	98 (0.04)
GUA	-90 (0.05)	-117 (0.0)	92 (0.04)
THY	-75 (0.1)	-110 (0.04)	-145* (0.16)
CYT	-78 (0.05)	-90 (0.05)	-142* (0.12)

$*Y_{BB2-BB3-SC1-SC3}$

Table SI4.3. Nucleotide particle relations for the special H-bonding beads. For the sake of simplicity, regardless the nucleotide the special beads are displayed as NH1 and NH2.
A) Adapted bond length parameters and corresponding force constants. D: Bond Length (Å). K: Force Constant (kcal.mol-1).

Nucleotide	$D_{SC2-NH1}$ (K)	$D_{SC3-NH1}$ (K)	$D_{SC4-NH1}$ (K)	$D_{SC2-NH2}$ (K)	$D_{SC4-NH2}$ (K)	$D_{NH1-NH2}$ (K)
ADE	2.29 (500.0)	2.66 (500.0)	3.26 (47.87)	2.66 (500.0)	2.88 (500.0)	2.66 (500.0)
GUA	2.95 (500.0)	2.95 (500.0)	3.89 (47.87)	2.95 (500.0)	2.85 (500.0)	2.95 (500.0)
THY	2.17 (500.0)	3.22 (500.0)	-	2.65 (500.0)	3.22 (500.0)	3.22 (500.0)
CYT	2.2 (500.0)	2.85 (500.0)	-	2.68 (500.0)	2.85 (500.0)	2.85 (500.0)

B) Adapted bond angle parameters and corresponding force constants. f: Bond Angle (°). K: Force Constant (kcal.mol^{-1}).

Nucleotide	$f_{BB3-SC1-NH1}$ (K)	$f_{SC2-SC3-NH1}$ (K)	$f_{SC2-SC4-NH1}$ (K)	$f_{SC2-NH1-NH2}$ (K)	$f_{SC3-SC4-NH1}$ (K)	$f_{SC4-NH1-NH2}$ (K)	$f_{SC1-SC2-NH2}$ (K)	$f_{SC2-SC4-NH2}$ (K)
ADE	160 (47.87)	85 (47.87)	125 (47.87)	85 (47.87)	74 (47.87)	74 (47.87)	85 (47.87)	74 (47.87)
GUA	137 (71.80)	84 (47.87)	125 (47.87)	69 (47.87)	69 (47.87)	84 (47.87)	94 (47.87)	94 (47.87)
THY	107 (71.80)	83 (23.93)	55* (23.93)	83 (23.93)	-	42** (23.93)	83 (23.93)	55*** (23.93)
CYT	95 (71.80)	71 (47.87)	47* (47.87)	71 (47.87)	-	47** (47.87)	71 (47.87)	71*** (47.87)

$*f_{SC2-SC3-NH1}$, $**f_{SC3-NH1-NH2}$, $***f_{SC2-SC3-NH2}$

C) Adapted bond dihedral parameters and corresponding force constants. Y: Bond Dihedral (°). K: Force Constant (kcal.mol^{-1}).

Nucleotide	$Y_{BB2\text{-}BB3\text{-}SC1\text{-}NH1}$ (K)	$Y_{BB2\text{-}BB3\text{-}SC1\text{-}NH2}$ (K)
ADE	-116 (0.0)	-
GUA	-117 (0.0)	-
THY	-110 (0.0)	-145 (0.16)
CYT	-90 (0.05)	-142 (0.12)

Table SI4.4. Ambiguous and unambiguous interaction restraints as defined in HADDOCK for the modeling of PRC1-nucleosome complex.

Type		Residue
Specific distance restraint		H2A: Cys85 (Atom SG) - PRC1: Lys118, Lys119 (Atom NZ). Distance range 0-2Å
Ambiguous Interaction Restraints (AIRs)	Active	62, 64, 97, 98
	Passive	38, 40, 52, 53, 56, 59, 64, 76, 77, 78, 80, 81, 86, 105, 115, 125, 129, 132, 134, 220, 221, 223, 224, 227, 231, 249, 252, 259, 260, 264, 267, 277, 284, 293, 294, 419, 422, 435, 436, 441, 468, 471, 472, 473, 474, 489, 491, 495, 498, 499, 509, 510, 513, 514, 516, 517, 518, 519, 628, 629, 631, 640, 644, 647, 648, 653, 654, 682, 696, 701, 702, 705, 706, 709, 712, 713, 716, 717, 719, 720, 722, 839, 852, 853, 856, 859, 864, 869, 876, 877, 880, 881, 886, 890, 915, 922, 925, 929, 932, 933, 934, 935, 1025, 1027, 1031, 1048, 1049, 1052, 1056, 1059, 1060, 1074, 1077, 1084, 1091, 1093, 1101, 1214, 1215, 1219, 1222, 1236, 1241, 1268, 1271, 1272, 1273, 1274, 1289, 1291, 1295, 1298, 1299, 1309, 1310, 1313, 1314, 1316, 1317, 1318, 1319, 1320, 1429, 1431, 1432, 1440, 1444, 1447, 1448, 1453, 1454, 1468, 1476, 1482, 1489, 1496, 1501, 1502, 1505, 1506, 1509, 1510, 1512, 1513, 1516, 1517, 1519, 1520, 1521

Table SI4.5. Atom/Bead count and computing time for the rigid-body and semi-flexible refinement stages (it0 + it1) for the selected cases of the protein-DNA benchmark. CPU times are averaged values (seconds/model)# (PART I)

Case	Atom count	Bead count	<AA time>	<CG time>
1azp	1175	300	32	25
1pt3	1845	513	55	27
2irf	1925	493	78	21
1qrv	2107	554	64	31
1hjc	2158	547	62	23
1vas	2275	581	61	28
1k79	2459	631	67	30
1w0t	2476	614	93	35
1zme	2558	654	92	29
1rpe	2558	643	89	45
1jj4	2603	697	78	31
1r4o	2686	652	132	33
3cro	2696	664	92	47
1qne	2703	705	247	32
1cma	2710	740	71	31
1by4	2713	699	94	40
1ea4	3003	770	99	40
2fl3	3084	862	198	41
2oaa	3169	897	90	36
1tro	3273	824	130	45
1bdt	3648	923	140	53
1b3t	3658	959	111	56

 # The timing corresponds to the total time in seconds reported by CNS as measured on an AMD Opteron (tm) Processor 6344.

Table SI4.5. Atom/Bead count and computing time for the rigid-body and semi-flexible refinement stages (it0 + it1) for the selected cases of the protein-DNA benchmark. CPU times are averaged values (seconds/model)# (PART II)

1mnn	3793	1058	157	57
1f4k	3827	995	125	56
7mht	3957	1102	369	50
1eyu	3969	1078	311	57
1ksy	4236	1140	130	55
1a74	4439	1180	138	73
2c5r	4459	1280	304	84
1zs4	4576	1163	198	51
1g9z	4593	1248	316	77
1z9c	4613	1190	168	66
1vrr	4633	1260	202	72
3bam	4733	1354	174	64
2fio	5038	1265	175	56
1dfm	5460	1540	182	67
1h9t	5577	1519	194	96
1kc6	5638	1616	263	98
1rva	5761	1672	193	104
1o3t	5930	1551	201	60
1z63	6022	1664	158	69
1fok	6854	1900	291	110
1ddn	7611	1994	710	179
1jt0	9476	2704	262	179

 # The timing corresponds to the total time in seconds reported by CNS as measured on an AMD Opteron (tm) Processor 6344.

Supplementary Information – Chapter 5. *Integrative modeling of membrane-associated protein assemblies.*

Figure SI5.1. The residues masked according to the maximum (ZDOCK-max), average (ZDOCK-avg) or minimum (ZDOCK-min) z-axis coordinate provided by the MemProtMD database are colored in dark gray. The orange regions, thus, represent those residues still allowed to contact with their counterpart.

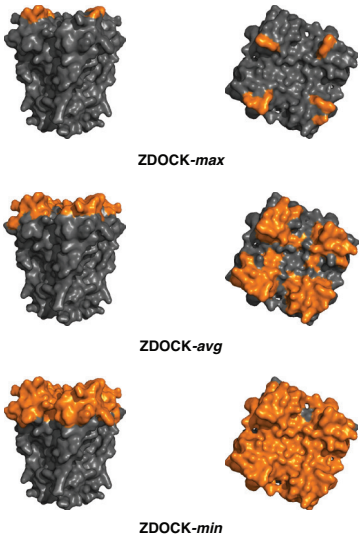


Figure SI5.2. Sidechain i-RMSD (A) and FNAT (B) comparison of all models (backbone) i-RMSD < 6Å before (y-axis) and after (x-axis) refinement (183 in total).

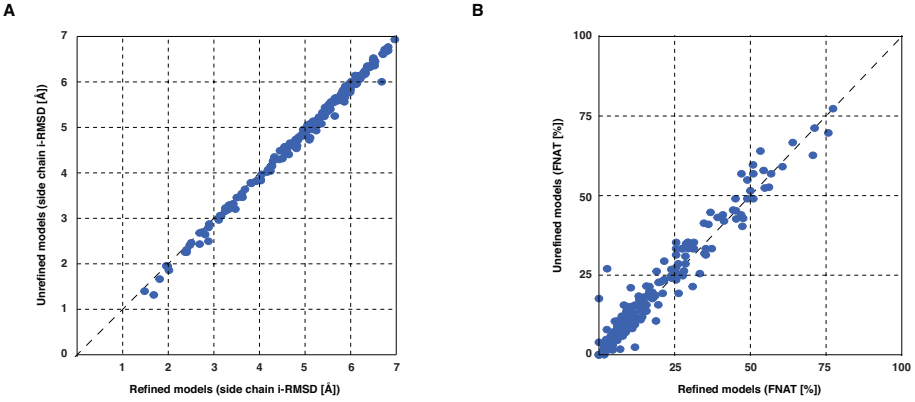


Figure SI5.3. ZDOCK success rate for the different tested (pseudo)membrane scenarios in the full dataset (18 cases).

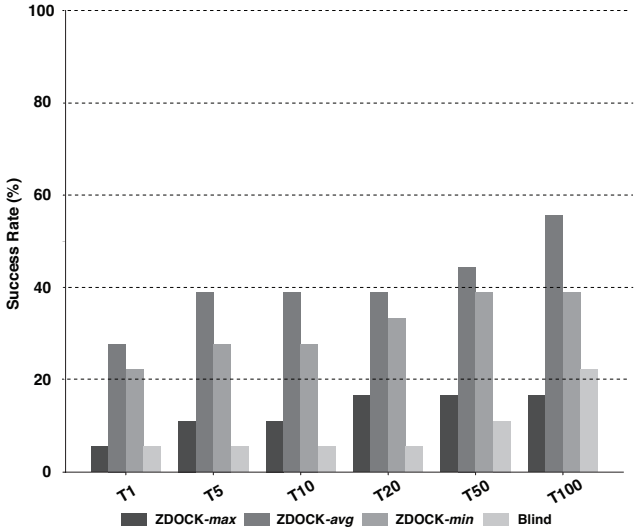


Figure SI5.4. Comparison of the membrane protein docking performance in the absence (Membrane) and presence (Membrane-rst) of experimental-like derived information (three defined interface residues (see Table S1) on the ligand (soluble) protein). The success rate is defined as the percentage of cases for which an acceptable or higher quality model was found within the selected top N models. These results include α-Helical and β-Barrel cases, 7 and 5, respectively.

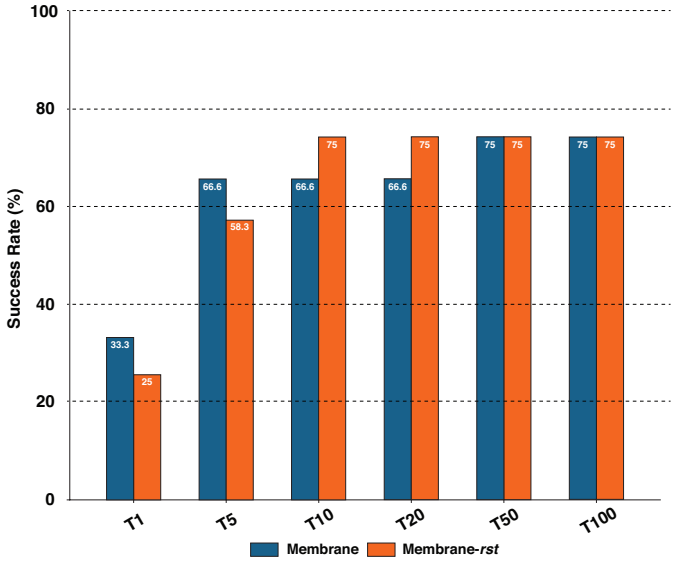


Table SI5.1. List of the residues used in LightDock as active restraints during the docking process for each of the 12 tested cases.

Case	Residues
2bs2	ARG1071, ARG1167, ARG1232
2vpz	TYR1102, GLU1189, THR1072
4huq	GLN96, VAL377, GLU390
3x29	ARG227, ASN218, TYR310
2r6g-peri	GLN49, ASP207, GLN335
2r6g-cyto	LEU52, LYS132, HIS1089
2hi7	HIS32, ARG148, VAL150
2hdi	ASP311, ARG313, GLY357
2gsk	ARG204, LYS207, ARG212
3csl	ASN41, PHE78, ASP102
5d0o	ARG60, ARG135, ASP136
3v8x	LYS365, ASP416, LYS557

References

1. Hofmann, A. W. On the Combining Power of Atoms. *Proc. R. Inst.* **4**, 401–430 (1865).
2. Ramberg, P. J. *Chemical structure, spatial arrangement: The early history of stereochemistry, 1874-1914*. *Chemical Structure, Spatial Arrangement: The Early History of Stereochemistry, 1874-1914* (2017).
3. Levinthal, C. How to fold graciously. *Mössbauer Spectrosc. Biol. Syst. Proc.* **24**, 22–24 (1969).
4. Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K. & Moult, J. Critical assessment of methods of protein structure prediction (CASP)—Round XIII. *Proteins: Structure, Function and Bioinformatics* **87,** 1011–1020 (2019).
5. Marrink, S. J., Risselada, H. J., Yefimov, S., Tieleman, D. P. & De Vries, A. H. The MARTINI force field: Coarse grained model for biomolecular simulations. *J. Phys. Chem. B* **111**, 7812–7824 (2007).
6. Wodak, S. J. & Janin, J. Computer analysis of protein-protein interaction. *J. Mol. Biol.* **124**, 323–342 (1978).
7. Levitt, M. & Warshel, A. Computer simulation of protein folding. *Nature* **253**, 694–698 (1975).
8. Dominguez, C., Boelens, R. & Bonvin, A. M. J. J. HADDOCK: A protein-protein docking approach based on biochemical or biophysical information. *J. Am. Chem. Soc.* **125**, 1731–1737 (2003).
9. Russel, D. *et al.* Putting the pieces together: Integrative modeling platform software for structure determination of macromolecular assemblies. *PLoS Biol.* **10**, (2012).
10. Jiménez-García, B. *et al.* LightDock: A new multi-scale approach to protein-protein docking. *Bioinformatics* **34**, 49–55 (2018).
11. Roel-Touris, J. & Bonvin, A. M. J. J. Coarse-grained (hybrid) integrative modeling of biomolecular interactions. *Computational and Structural Biotechnology Journal* **18**, 1182–1190 (2020).
12. Roel-Touris, J., Bonvin, A. M. J. J. & Jiménez-García, B. LightDock goes information-driven. *Bioinformatics* **36**, 950–952 (2020).

13. Roel-Touris, J., Don, C. G., Honorato, R. R., Rodrigues, J. P. G. L. M. & Bonvin, A. M. J. J. Less Is More: Coarse-Grained Integrative Modeling of Large Biomolecular Assemblies with HADDOCK. *J. Chem. Theory Comput.* **15**, 6358–6367 (2019).
14. De Jong, D. H. *et al.* Improved parameters for the martini coarse-grained protein force field. *J. Chem. Theory Comput.* **9**, 687–697 (2013).
15. Honorato, R. V., Roel-Touris, J. & Bonvin, A. M. J. J. MARTINI-Based Protein-DNA Coarse-Grained HADDOCKing. *Front. Mol. Biosci.* **6**, (2019).
16. Uusitalo, J. J., Ingólfsson, H. I., Akhshi, P., Tieleman, D. P. & Marrink, S. J. Martini Coarse-Grained Force Field: Extension to DNA. *J. Chem. Theory Comput.* **11**, 3932–3945 (2015).
17. Berggård, T., Linse, S. & James, P. Methods for the detection and analysis of protein-protein interactions. *Proteomics* **7,** 2833–2842 (2007).
18. Rout, M. P. & Sali, A. Principles for Integrative Structural Biology Studies. *Cell* **177**, 1384–1403 (2019).
19. Koukos, P. I. & Bonvin, A. M. J. J. Integrative Modelling of Biomolecular Complexes. *Journal of Molecular Biology* **432**, 2861–2881 (2020).
20. Braitbard, M., Schneidman-Duhovny, D. & Kalisman, N. Integrative Structure Modeling: Overview and Assessment. *Annu. Rev. Biochem.* **88**, 113–135 (2019).
21. Singla, J. *et al.* Opportunities and Challenges in Building a Spatio-temporal Multi-scale Model of the Human Pancreatic β Cell. *Cell* **173**, 11–19 (2018).
22. Phillips, D. C. Symposium on Three-Dimensional Structure of Macromolecules of Biological Origin. by Invitation of the Committee on Arrangements for the Autumn Meeting. Presented before the Academy on October 19, 1966. Chairman, Walter Kauzmann: THE HEN EGG-WHITE LYSOZYME. *Proc. Natl. Acad. Sci.* **57**, 483–495 (1967).
23. Warshel, A. Multiscale modeling of biological functions: From enzymes to molecular machines (nobel lecture). *Angew. Chemie - Int. Ed.* **53**, 10020–10031 (2014).

24. Chothia, C. & Janin, J. Principles of protein-protein recognition. *Nature* **256**, 705–708 (1975).
25. Fersht, A. R. *et al.* Analysis of Enzyme Structure and Activity by Protein Engineering. *Angew. Chemie Int. Ed. English* **23**, 467–473 (1984).
26. Kmiecik, S. *et al.* Coarse-Grained Protein Models and Their Applications. *Chemical Reviews* **116**, 7898–7936 (2016).
27. Lau, K. F. & Dill, K. A. A Lattice Statistical Mechanics Model of the Conformational and Sequence Spaces of Proteins. *Macromolecules* **22**, 3986–3997 (1989).
28. Dill, K. A. *et al.* Principles of protein folding — A perspective from simple exact models. *Protein Science* **4,** 561–602 (1995).
29. Šali, A., Shakhnovich, E. & Karplus, M. Kinetics of protein folding: A lattice model study of the requirements for folding to the native state. *J. Mol. Biol.* **235**, 1614–1636 (1994).
30. Dinner, A. R., Šali, A. & Karplus, M. The folding mechanism of larger model proteins: Role of native structure. *Proc. Natl. Acad. Sci. U. S. A.* **93**, 8356–8361 (1996).
31. Locker, C. R. & Hernandez, R. A minimalist model protein with multiple folding funnels. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 9074–9079 (2001).
32. Kaya, H. & Chan, H. S. Towards a consistent modeling of protein thermodynamic and kinetic cooperativity: How applicable is the transition state picture to folding and unfolding? *J. Mol. Biol.* **315**, 899–909 (2002).
33. Kolinski, A. & Skolnick, J. Reduced models of proteins and their applications. *Polymer (Guildf).* **45**, 511–524 (2004).
34. Kolinski, A. & Skolnick, J. Monte carlo simulations of protein folding. I. Lattice model and interaction scheme. *Proteins Struct. Funct. Bioinforma.* **18**, 338–352 (1994).
35. MacKerell, A. D., Feig, M. & Brooks, C. L. Improved Treatment of the Protein Backbone in Empirical Force Fields. *J. Am. Chem. Soc.* **126**, 698–699 (2004).

36. Gopal, S. M., Mukherjee, S., Cheng, Y. M. & Feig, M. PRIMO/PRIMONA: A coarse-grained model for proteins and nucleic acids that preserves near-atomistic accuracy. *Proteins Struct. Funct. Bioinforma.* **78**, 1266–1281 (2010).
37. Pasquali, S. & Derreumaux, P. HiRE-RNA: A high resolution coarse-grained energy model for RNA. *J. Phys. Chem. B* **114**, 11957–11966 (2010).
38. Darré, L. *et al.* SIRAH: A structurally unbiased coarse-grained force field for proteins with aqueous solvation and long-range electrostatics. *J. Chem. Theory Comput.* **11**, 723–739 (2015).
39. Dans, P. D., Zeida, A., MacHado, M. R. & Pantano, S. A coarse grained model for atomic-detailed DNA simulations with explicit electrostatics. *J. Chem. Theory Comput.* **6**, 1711–1725 (2010).
40. Darré, L., MacHado, M. R., Dans, P. D., Herrera, F. E. & Pantano, S. Another coarse grain model for aqueous solvation: WAT FOUR? *J. Chem. Theory Comput.* **6**, 3793–3807 (2010).
41. Monticelli, L. *et al.* The MARTINI coarse-grained force field: Extension to proteins. *J. Chem. Theory Comput.* **4**, 819–834 (2008).
42. Lee, H., De Vries, A. H., Marrink, S. J. & Pastor, R. W. A coarse-grained model for polyethylene oxide and polyethylene glycol: Conformation and hydrodynamics. *J. Phys. Chem. B* **113**, 13186–13194 (2009).
43. Gobbo, C. *et al.* MARTINI model for physisorption of organic molecules on graphite. *J. Phys. Chem. C* **117**, 15623–15631 (2013).
44. López, C. A. *et al.* Martini coarse-grained force field: Extension to carbohydrates. *J. Chem. Theory Comput.* **5**, 3195–3210 (2009).
45. Yesylevskyy, S. O., Schäfer, L. V., Sengupta, D. & Marrink, S. J. Polarizable water model for the coarse-grained MARTINI force field. *PLoS Comput. Biol.* **6**, 1–17 (2010).
46. López, C. A., Sovova, Z., Van Eerden, F. J., De Vries, A. H. & Marrink, S. J. Martini force field parameters for glycolipids. *J. Chem. Theory Comput.* **9**, 1694–1708 (2013).
47. Uusitalo, J. J., Ingólfsson, H. I., Marrink, S. J. & Faustino, I. Martini Coarse-Grained Force Field: Extension to RNA. *Biophys. J.* **113**, 246–256 (2017).

48. López, C. A. *et al.* MARTINI coarse-grained model for crystalline cellulose microfibers. *J. Phys. Chem. B* **119**, 465–473 (2015).
49. Noid, W. G. Perspective: Coarse-grained models for biomolecular systems. *Journal of Chemical Physics* **139**, 090901 (2013).
50. Saunders, M. G. & Voth, G. A. Coarse-Graining Methods for Computational Biology. *Annu. Rev. Biophys.* **42**, 73–93 (2013).
51. Ercolesi, F. & Adams, J. B. Interatomic potentials from first-principles calculations: The force-matching method. *Epl* **26**, 583–588 (1994).
52. Izvekov, S., Parrinello, M., Bumham, C. J. & Voth, G. A. Effective force fields for condensed phase systems from ab initio molecular dynamics simulation: A new method for force-matching. *J. Chem. Phys.* **120**, 10896–10913 (2004).
53. Izvekov, S. & Voth, G. A. A multiscale coarse-graining method for biomolecular systems. *J. Phys. Chem. B* **109**, 2469–2473 (2005).
54. Soper, A. K. Empirical potential Monte Carlo simulation of fluid structure. *Chem. Phys.* **202**, 295–306 (1996).
55. Lu, L., Dama, J. F. & Voth, G. A. Fitting coarse-grained distribution functions through an iterative force-matching method. *J. Chem. Phys.* **139**, 121906 (2013).
56. Liwo, A. & Czaplewski, C. Extension of the force-matching method to coarse-grained models with axially symmetric sites to produce transferable force fields: Application to the UNRES model of proteins. *J. Chem. Phys.* **152**, (2020).
57. Ingólfsson, H. I. *et al.* The power of coarse graining in biomolecular simulations. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **4**, 225–248 (2014).
58. Mejía, A., Herdes, C. & Müller, E. A. Force fields for coarse-grained molecular simulations from a corresponding states correlation. *Ind. Eng. Chem. Res.* **53**, 4131–4141 (2014).
59. Müller, E. A. & Jackson, G. Force-Field Parameters from the SAFT-γ Equation of State for Use in Coarse-Grained Molecular Simulations. *Annu. Rev. Chem. Biomol. Eng.* **5**, 405–427 (2014).

60. Matos, I. Q. & Abreu, C. R. A. Evaluation of the SAFT-γ Mie force field with solvation free energy calculations. *Fluid Phase Equilib.* **484**, 88–97 (2019).
61. Senior, A. W. *et al.* Improved protein structure prediction using potentials from deep learning. *Nature* **577**, 706–710 (2020).
62. Senior, A. W. *et al.* Protein structure prediction using multiple deep neural networks in the 13th Critical Assessment of Protein Structure Prediction (CASP13). *Proteins Struct. Funct. Bioinforma.* **87**, 1141–1148 (2019).
63. Botu, V., Batra, R., Chapman, J. & Ramprasad, R. Machine learning force fields: Construction, validation, and outlook. *J. Phys. Chem. C* **121**, 511–522 (2017).
64. Wang, J. *et al.* Machine Learning of Coarse-Grained Molecular Dynamics Force Fields. *ACS Cent. Sci.* **5**, 755–767 (2019).
65. Durumeric, A. E. P. & Voth, G. A. Adversarial-residual-coarse-graining: Applying machine learning theory to systematic molecular coarse-graining. *J. Chem. Phys.* **151**, (2019).
66. Chan, H. *et al.* Machine learning coarse grained models for water. *Nat. Commun.* **10**, (2019).
67. Ayton, G. S., Noid, W. G. & Voth, G. A. Multiscale modeling of biomolecular systems: in serial and in parallel. *Current Opinion in Structural Biology* **17**, 192–198 (2007).
68. König, G., Hudson, P. S., Boresch, S. & Woodcock, H. L. Multiscale free energy simulations: An efficient method for connecting classical MD simulations to QM or QM/MM free energies using non-Boltzmann Bennett reweighting schemes. *J. Chem. Theory Comput.* **10**, 1406–1419 (2014).
69. Lee, S., Liang, R., Voth, G. A. & Swanson, J. M. J. Computationally Efficient Multiscale Reactive Molecular Dynamics to Describe Amino Acid Deprotonation in Proteins. *J. Chem. Theory Comput.* **12**, 879–891 (2016).
70. Scott, R., Allen, M. P. & Tildesley, D. J. Computer Simulation of Liquids. *Math. Comput.* **57**, 442 (1991).

71. Michel, J., Orsi, M. & Essex, J. W. Prediction of partition coefficients by multiscale hybrid atomic-level/coarse-grain simulations. *J. Phys. Chem. B* **112**, 657–660 (2008).
72. Rzepiela, A. J., Louhivuori, M., Peter, C. & Marrink, S. J. Hybrid simulations: Combining atomistic and coarse-grained force fields using virtual sites. *Phys. Chem. Chem. Phys.* **13**, 10437–10448 (2011).
73. Wan, C. K., Han, W. & Wu, Y. D. Parameterization of PACE force field for membrane environment and simulation of helical peptides and helix-helix association. *J. Chem. Theory Comput.* **8**, 300–313 (2012).
74. Ward, M. D., Nangia, S. & May, E. R. Evaluation of the hybrid resolution PACE model for the study of folding, insertion, and pore formation of membrane associated peptides. *J. Comput. Chem.* **38**, 1462–1471 (2017).
75. Wassenaar, T. A., Ingólfsson, H. I., Prieß, M., Marrink, S. J. & Schäfer, L. V. Mixing MARTINI: Electrostatic coupling in hybrid atomistic-coarse-grained biomolecular simulations. *J. Phys. Chem. B* **117**, 3516–3530 (2013).
76. Kar, P. & Feig, M. Hybrid All-Atom/Coarse-Grained Simulations of Proteins by Direct Coupling of CHARMM and PRIMO Force Fields. *J. Chem. Theory Comput.* **13**, 5753–5765 (2017).
77. Vangone, A. *et al.* Sense and simplicity in HADDOCK scoring: Lessons from CASP-CAPRI round 1. *Proteins Struct. Funct. Bioinforma.* **85**, 417–423 (2017).
78. Brünger, A. T. *et al.* Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **54**, 905–921 (1998).
79. De Vries, S. & Zacharias, M. Flexible docking and refinement with a coarse-grained protein model using ATTRACT. *Proteins Struct. Funct. Bioinforma.* **81**, 2167–2174 (2013).
80. Setny, P., Bahadur, R. P. & Zacharias, M. Protein-DNA docking with a coarse-grained force field. *BMC Bioinformatics* **13**, (2012).
81. De Vries, S. J., Rey, J., Schindler, C. E. M., Zacharias, M. & Tuffery, P. The pepATTRACT web server for blind, large-scale peptide-protein docking. *Nucleic Acids Res.* **45**, W361–W364 (2017).

82. Kolinski, A. Protein modeling and structure prediction with a reduced representation. in *Acta Biochimica Polonica* **51**, 349–371 (2004).
83. Kurcinski, M., Jamroz, M., Blaszczyk, M., Kolinski, A. & Kmiecik, S. CABS-dock web server for the flexible docking of peptides to proteins without prior knowledge of the binding site. *Nucleic Acids Res.* **43**, W419–W424 (2015).
84. Ramírez-Aportela, E., López-Blanco, J. R. & Chacón, P. FRODOCK 2.0: Fast protein-protein docking server. *Bioinformatics* **32**, 2386–2388 (2016).
85. Andreani, J., Faure, G. & Guerois, R. InterEvScore: A novel coarse-grained interface scoring function using a multi-body statistical potential coupled to evolution. *Bioinformatics* **29**, 1742–1749 (2013).
86. Quignot, C. *et al.* InterEvDock2: An expanded server for protein docking using evolutionary and biological information from homology models and multimeric inputs. *Nucleic Acids Res.* **46**, W408–W416 (2018).
87. Esquivel-Rodriguez, J., Filos-Gonzalez, V., Li, B. & Kihara, D. Pairwise and multimeric protein–protein docking using the lzerd program suite. *Methods Mol. Biol.* **1137**, 209–234 (2014).
88. Peterson, L. X., Roy, A., Christoffer, C., Terashi, G. & Kihara, D. Modeling disordered protein interactions from biophysical principles. *PLoS Comput. Biol.* **13**, (2017).
89. Sacquin-Mora, S., Carbone, A. & Lavery, R. Identification of Protein Interaction Partners and Protein-Protein Interaction Sites. *J. Mol. Biol.* **382**, 1276–1289 (2008).
90. Walther, J. *et al.* A multi-modal coarse grained model of DNA flexibility mappable to the atomistic level. *Nucleic Acids Res.* **48**, (2020).
91. Huang, S. Y. & Zou, X. MDockPP: A hierarchical approach for protein-protein docking and its application to CAPRI rounds 15-19. *Proteins Struct. Funct. Bioinforma.* **78**, 3096–3103 (2010).
92. Gray, J. J. *et al.* Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *J. Mol. Biol.* **331**, 281–299 (2003).

93. Roy Burman, S. S. *et al.* Novel sampling strategies and a coarse-grained score function for docking homomers, flexible heteromers, and oligosaccharides using Rosetta in CAPRI rounds 37–45. *Proteins Struct. Funct. Bioinforma.* **88**, 973–985 (2019).
94. Hou, Q., Lensink, M. F., Heringa, J. & Feenstra, K. A. CLUB-MARTINI: Selecting favourable interactions amongst available candidates, a coarse-grained simulation approach to scoring docking decoys. *PLoS One* **11**, (2016).
95. Viswanath, S., Ravikant, D. V. S. & Elber, R. DOCK/PIERR: Web server for structure prediction of protein-protein complexes. *Methods Mol. Biol.* **1137**, 199–207 (2014).
96. Shin, W.-H., Lee, G. R., Heo, L., Lee, H. & Seok, C. Prediction of Protein Structure and Interaction by GALAXY Protein Modeling Programs. *Bio Des.* **2**, 01–11 (2014).
97. Lee, H., Heo, L., Lee, M. S. & Seok, C. GalaxyPepDock: A protein-peptide docking tool based on interaction similarity and energy optimization. *Nucleic Acids Res.* **43**, W431–W435 (2015).
98. Ohue, M., Matsuzaki, Y., Ishida, T. & Akiyama, Y. Improvement of the protein-protein docking prediction by introducing a simple hydrophobic interaction model: An application to interaction pathway analysis. in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **7632**, LNBI 178–187 (2012).
99. Ohue, M. *et al.* MEGADOCK 4.0: An ultra-high-performance protein-protein docking software for heterogeneous supercomputers. *Bioinformatics* **30**, 3281–3283 (2014).
100. Olechnovič, K. & Venclovas, Č. Voronota: A fast and reliable tool for computing the vertices of the Voronoi diagram of atomic balls. *J. Comput. Chem.* **35**, 672–681 (2014).
101. Dapkunas, J. *et al.* The PPI3D web server for searching, analyzing and modeling protein-protein interactions in the context of 3D structures. *Bioinformatics* **33**, 935–937 (2017).

102. Solernou, A. & Fernandez-Recio, J. PyDockCG: New coarse-grained potential for protein-protein docking. *J. Phys. Chem. B* **115**, 6032–6039 (2011).
103. Jiménez-García, B., Pons, C. & Fernández-Recio, J. pyDockWEB: A web server for rigid-body protein-protein docking using electrostatics and desolvation scoring. in *Bioinformatics* **29**, 1698–1699 (2013).
104. Segura, J., Marín-López, M. A., Jones, P. F., Oliva, B. & Fernandez-Fuentes, N. VORFFIP-driven dock: V-D2OCK, a fast, accurate protein docking strategy. *PLoS One* **10**, (2015).
105. Webb, B. *et al.* Integrative structure modeling with the Integrative Modeling Platform. *Protein Sci.* **27**, 245–258 (2018).
106. Badaczewska-Dawid, A. E., Kolinski, A. & Kmiecik, S. Computational reconstruction of atomistic protein structures from coarse-grained models. *Comput. Struct. Biotechnol. J.* **18**, 162–176 (2020).
107. Wassenaar, T. A., Pluhackova, K., Böckmann, R. A., Marrink, S. J. & Tieleman, D. P. Going backward: A flexible geometric approach to reverse transformation from coarse grained to atomistic models. *J. Chem. Theory Comput.* **10**, 676–690 (2014).
108. Machado, M. R. & Pantano, S. SIRAH tools: Mapping, backmapping and visualization of coarse-grained models. *Bioinformatics* **32**, 1568–1570 (2016).
109. Rzepiela, A. J. *et al.* Software news and update reconstruction of atomistic details from coarse-grained structures. *J. Comput. Chem.* **31**, 1333–1343 (2010).
110. Zacharias, M. Protein-protein docking with a reduced protein model accounting for side-chain flexibility. *Protein Sci.* **12**, 1271–1282 (2003).
111. Heath, A. P., Kavraki, L. E. & Clementi, C. From coarse-grain to all-atom: Toward multiscale analysis of protein landscapes. *Proteins Struct. Funct. Genet.* **68**, 646–661 (2007).
112. Shimizu, M. & Takada, S. Reconstruction of Atomistic Structures from Coarse-Grained Models for Protein-DNA Complexes. *J. Chem. Theory Comput.* **14**, 1682–1694 (2018).
113. Lombardi, L. E., Martí, M. A. & Capece, L. CG2AA: Backmapping protein coarse-grained structures. *Bioinformatics* **32**, 1235–1237 (2016).

114. Joosten, R. P., Joosten, K., Murshudov, G. N. & Perrakis, A. PDB-REDO: Constructive validation, more than just looking for errors. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **68**, 484–496 (2012).
115. Peng, J., Yuan, C., Ma, R. & Zhang, Z. Backmapping from Multiresolution Coarse-Grained Models to Atomic Structures of Large Biomolecules by Restrained Molecular Dynamics Simulations Using Bayesian Inference. *J. Chem. Theory Comput.* **15**, 3344–3353 (2019).
116. Van Zundert, G. C. P. *et al.* The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. *J. Mol. Biol.* **428**, 720–725 (2016).
117. London, N., Raveh, B. & Schueler-Furman, O. Peptide docking and structure-based characterization of peptide binding: From knowledge to know-how. *Curr. Opin. Struct. Biol.* **23**, 894–902 (2013).
118. Rodrigues, J. P. G. L. M. & Bonvin, A. M. J. J. Integrative computational modeling of protein interactions. *FEBS J.* **281**, 1988–2003 (2014).
119. Nithin, C., Ghosh, P. & Bujnicki, J. M. Bioinformatics tools and benchmarks for computational docking and 3D structure prediction of RNA-protein complexes. *Genes (Basel).* **9**, 432 (2018).
120. Joosten, R. P. *et al.* A series of PDB related databases for everyday needs. *Nucleic Acids Res.* **39**, (2011).
121. Touw, W. G. *et al.* A series of PDB-related databanks for everyday needs. *Nucleic Acids Res.* **43**, D364–D368 (2015).
122. Sali, A. *et al.* Outcome of the First wwPDB Hybrid/Integrative Methods Task Force Workshop. *Structure* **23**, 1156–1167 (2015).
123. Berman, H. M. *et al.* Federating Structural Models and Data: Outcomes from A Workshop on Archiving Integrative Structures. *Structure* **27**, 1745–1759 (2019).
124. Shi, Y. *et al.* Structural characterization by cross-linking reveals the detailed architecture of a coatomer-related heptameric module from the nuclear pore complex. *Mol. Cell. Proteomics* **13**, 2927–2943 (2014).
125. Shi, Y. *et al.* A strategy for dissecting the architectures of native macromolecular assemblies. *Nat. Methods* **12**, 1135–1138 (2015).

126. Chen, Z. A. *et al.* Structure of complement C3(H2O) revealed by quantitative cross-linking/mass spectrometry and modeling. *Mol. Cell. Proteomics* **15**, 2730–2743 (2016).
127. Sailer, C. *et al.* Structural dynamics of the E6AP/UBE3A-E6-p53 enzyme-substrate complex. *Nat. Commun.* **9**, 4441 (2018).
128. Jishage, M. *et al.* Architecture of Pol II(G) and molecular mechanism of transcription regulation by Gdown1. *Nat. Struct. Mol. Biol.* **25**, 859–867 (2018).
129. Wang, X. *et al.* The proteasome-interacting Ecm29 protein disassembles the 26S proteasome in response to oxidative stress. *J. Biol. Chem.* **292**, 16310–16320 (2017).
130. Gutierrez, C. *et al.* Structural dynamics of the human COP9 signalosome revealed by cross-linking mass spectrometry and integrative modeling. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 4088–4098 (2020).
131. Robinson, P. J. *et al.* Molecular architecture of the yeast Mediator complex. *Elife* **4**, (2015).
132. Chou, H. T. *et al.* The Molecular Architecture of Native BBSome Obtained by an Integrated Structural Approach. *Structure* **27**, 1384-1394. e4 (2019).
133. Bender, B. J. *et al.* Structural Model of Ghrelin Bound to its G Protein-Coupled Receptor. *Structure* **27**, 537-544.e4 (2019).
134. Dai, G., Aman, T. K., DiMaio, F. & Zagotta, W. N. The HCN channel voltage sensor undergoes a large downward motion during hyperpolarization. *Nat. Struct. Mol. Biol.* **26**, 686–694 (2019).
135. Leone, V. & Faraldo-Gómez, J. D. Structure and mechanism of the ATP synthase membrane motor inferred from quantitative integrative modeling. *J. Gen. Physiol.* **148**, 441–457 (2016).
136. Harrer, N. *et al.* Structural Architecture of the Nucleosome Remodeler ISWI Determined from Cross-Linking, Mass Spectrometry, SAXS, and Modeling. *Structure* **26**, 282-294.e6 (2018).
137. Goddard, T. D. *et al.* UCSF ChimeraX: Meeting modern challenges in visualization and analysis. *Protein Sci.* **27**, 14–25 (2018).
138. Kim, S. J. *et al.* Integrative structure and functional anatomy of a nuclear pore complex. *Nature* **555**, 475–482 (2018).

139. De Vries, S. J., Van Dijk, M. & Bonvin, A. M. J. J. The HADDOCK web server for data-driven biomolecular docking. *Nat. Protoc.* **5**, 883–897 (2010).
140. De Vries, S. J., Schindler, C. E. M., Chauvot De Beauchêne, I. & Zacharias, M. A web interface for easy flexible protein-protein docking with ATTRACT. *Biophys. J.* **108**, 462–465 (2015).
141. Krishnanand, K. N. & Ghose, D. Glowworm swarm optimization for simultaneous capture of multiple local optima of multimodal functions. *Swarm Intell.* **3**, 87–124 (2009).
142. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
143. Bakan, A., Meireles, L. M. & Bahar, I. ProDy: Protein dynamics inferred from theory and experiments. *Bioinformatics* **27**, 1575–1577 (2011).
144. Brian Jimenez, Vidal, M. & Roel, J. brianjimenez/lightdock: Release 0.7.0 (Version 0.7.0). *Zenodo. doi 10.5281/zenodo.3228412*.
145. Vreven, T. *et al.* Updates to the Integrated Protein-Protein Interaction Benchmarks: Docking Benchmark Version 5 and Affinity Benchmark Version 2. *J. Mol. Biol.* **427**, 3031–3041 (2015).
146. Wallace, A. C., Laskowski, R. A. & Thornton, J. M. Ligplot: A program to generate schematic diagrams of protein-ligand interactions. *Protein Eng. Des. Sel.* **8**, 127–134 (1995).
147. Lensink, M. F. & Wodak, S. J. Docking and scoring protein interactions: CAPRI 2009. *Proteins Struct. Funct. Bioinforma.* **78**, 3073–3084 (2010).
148. Zhou, H. & Zhou, Y. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Sci.* **11**, 2714–2726 (2009).
149. Ambrosetti, F., Jiménez-García, B., Roel-Touris, J. & Bonvin, A. M. J. J. Modeling Antibody-Antigen Complexes by Information-Driven Docking. *Structure* **28**, 119-129.e2 (2020).
150. Ritchie, D. Recent Progress and Future Directions in Protein-Protein Docking. *Curr. Protein Pept. Sci.* **9**, 1–15 (2008).

151. Aloy, P., Pichaud, M. & Russell, R. B. Protein complexes: Structure prediction challenges for the 21st century. *Current Opinion in Structural Biology* **15**, 15–22 (2005).
152. Wass, M. N., David, A. & Sternberg, M. J. E. Challenges for the prediction of macromolecular interactions. *Current Opinion in Structural Biology* **21**, 382–390 (2011).
153. Berman, H. M. The Protein Data Bank / Biopython. *Presentation* **28**, 235–242 (2000).
154. Mosca, R., Céol, A. & Aloy, P. Interactome3D: Adding structural details to protein networks. *Nat. Methods* **10**, 47–53 (2013).
155. Wiehe, K., Peterson, M. W., Pierce, B., Mintseris, J. & Weng, Z. Protein-protein docking: Overview and performance analysis. *Methods Mol. Biol.* **413**, 283–314 (2007).
156. Karaca, E. & Bonvin, A. M. J. J. Advances in integrative modeling of biomolecular complexes. *Methods* **59**, 372–381 (2013).
157. Rader, A. J. Coarse-grained models: Getting more with less. *Curr. Opin. Pharmacol.* **10**, 753–759 (2010).
158. Takada, S. Coarse-grained molecular simulations of large biomolecules. *Current Opinion in Structural Biology* **22**, 130–137 (2012).
159. Saunders, M. G. & Voth, G. A. Coarse-graining of multiprotein assemblies. *Current Opinion in Structural Biology* **22**, 144–150 (2012).
160. Levitt, M. & Lifson, S. Refinement of protein conformations using a macromolecular energy minimization procedure. *J. Mol. Biol.* **46**, 269–279 (1969).
161. Fiorucci, S. & Zacharias, M. Binding site prediction and improved scoring during flexible protein-protein docking with ATTRACT. *Proteins Struct. Funct. Bioinforma.* **78**, 3131–3139 (2010).
162. Blaszczyk, M. *et al.* Modeling of protein-peptide interactions using the CABS-dock web server for binding site search and flexible docking. *Methods* **93**, 72–83 (2016).
163. Ishiura, M. *et al.* Expression of a gene cluster kaiABC as a circadian feedback process in cyanobacteria. *Science.* **281**, 1519–1523 (1998).
164. Neill, J. S. O. & Reddy, A. B. Europe PMC Funders Group Circadian Clocks in Human Red Blood Cells. *Nature* **469**, 498–503 (2011).

165. Brunger, A. T. Version 1.2 of the crystallography and nmr system. *Nat. Protoc.* **2**, 2728–2733 (2007).
166. Fernández-Recio, J., Totrov, M. & Abagyan, R. Identification of protein-protein interaction sites from docking energy landscapes. *J. Mol. Biol.* **335**, 843–865 (2004).
167. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637 (1983).
168. Jorgensen, W. L. & Tirado-Rives, J. The OPLS Potential Functions for Proteins. Energy Minimizations for Crystals of Cyclic Peptides and Crambin. *J. Am. Chem. Soc.* **110**, 1657–1666 (1988).
169. Rodrigues, J. P. G. L. M. *et al.* Clustering biomolecular complexes by residue contacts similarity. *Proteins Struct. Funct. Bioinforma.* **80**, 1810–1817 (2012).
170. Snijder, J. *et al.* Insight into cyanobacterial circadian timing from structural details of the KaiB-KaiC interaction. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 1379–1384 (2014).
171. Tseng, R. *et al.* Structural basis of the day-night transition in a bacterial circadian clock. *Science.* **355**, 1174–1180 (2017).
172. Villarreal, S. A. *et al.* CryoEM and molecular dynamics of the circadian KaiB-KaiC complex indicates that KaiB monomers interact with KaiC and block ATP binding clefts. *J. Mol. Biol.* **425**, 3311–3324 (2013).
173. Lee, B. & Richards, F. M. The interpretation of protein structures: Estimation of static accessibility. *J. Mol. Biol.* **55**, (1971).
174. Hayashi, F. *et al.* ATP-induced hexameric ring structure of the cyanobacterial circadian clock protein KaiC. *Genes to Cells* **8**, 287–296 (2003).
175. Hayashi, F., Iwase, R., Uzumaki, T. & Ishiura, M. Hexamerization by the N-terminal domain and intersubunit phosphorylation by the C-terminal domain of cyanobacterial circadian clock protein KaiC. *Biochem. Biophys. Res. Commun.* **348**, 864–872 (2006).
176. Snijder, J. *et al.* Structures of the cyanobacterial circadian oscillator frozen in a fully assembled state. *Science.* **355**, 1181–1184 (2017).

177. Pettersen, E. F. *et al.* UCSF Chimera - A visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).
178. Pandey, P., Hasnain, S. & Ahmad, S. Protein-DNA interactions. in *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics* **1–3,** 142–154 (2018).
179. Janin, J. Protein-protein docking tested in blind predictions: The CAPRI experiment. *Molecular BioSystems* **6,** 2351–2362 (2010).
180. Hills, R. D., Lu, L. & Voth, G. A. Multiscale coarse-graining of the protein energy landscape. *PLoS Comput. Biol.* **6**, 1–15 (2010).
181. Tuszynska, I., Magnus, M., Jonak, K., Dawson, W. & Bujnicki, J. M. NPDock: A web server for protein-nucleic acid docking. *Nucleic Acids Res.* **43**, W425–W430 (2015).
182. van Dijk, M. & Bonvin, A. M. J. J. Pushing the limits of what is achievable in protein-DNA docking: Benchmarking HADDOCK's performance. *Nucleic Acids Res.* **38**, 5634–5647 (2010).
183. McGinty, R. K., Henrici, R. C. & Tan, S. Crystal structure of the PRC1 ubiquitylation module bound to the nucleosome. *Nature* **514**, 591–596 (2014).
184. Lensink, M. F., Velankar, S. & Wodak, S. J. Modeling protein–protein and protein–peptide complexes: CAPRI 6th edition. *Proteins Struct. Funct. Bioinforma.* **85**, 359–377 (2017).
185. Luscombe, N. M., Austin, S. E., Berman, H. M. & Thornton, J. M. An overview of the structures of protein-DNA complexes. *Genome Biol.* **1**, REVIEWS001 (2000).
186. Nozinovic, S., Fürtig, B., Jonker, H. R. A., Richter, C. & Schwalbe, H. High-resolution NMR structure of an RNA model system: The 14-mer cUUCGg tetraloop hairpin RNA. *Nucleic Acids Res.* **38**, 683–694 (2009).
187. Bentley, M. L. *et al.* Recognition of UbcH5c and the nucleosome by the Bmi1/Ring1b ubiquitin ligase complex. *EMBO J.* **30**, 3285–3297 (2011).

188. Mattiroli, F., Uckelmann, M., Sahtoe, D. D., van Dijk, W. J. & Sixma, T. K. The nucleosome acidic patch plays a critical role in RNF168-dependent ubiquitination of histone H2A. *Nat. Commun.* **5**, 3291 (2014).
189. Janin, J. Assessing predictions of protein-protein interaction: The CAPRI experiment. *Protein Sci.* **14**, 278–283 (2005).
190. Van Dijk, M., Van Dijk, A. D. J., Hsu, V., Rolf, B. & Bonvin, A. M. J. J. Information-driven protein-DNA docking using HADDOCK: It is a matter of flexibility. *Nucleic Acids Res.* **34**, 3317–3325 (2006).
191. Paissoni, C., Jussupow, A. & Camilloni, C. Martini bead form factors for nucleic acids and their application in the refinement of protein–nucleic acid complexes against SAXS data. *J. Appl. Crystallogr.* **52**, 394–402 (2019).
192. Kerscher, O., Felberbaum, R. & Hochstrasser, M. Modification of Proteins by Ubiquitin and Ubiquitin-Like Proteins. *Annu. Rev. Cell Dev. Biol.* **22**, 159–180 (2006).
193. Allen, K. N., Entova, S., Ray, L. C. & Imperiali, B. Monotopic Membrane Proteins Join the Fold. *Trends in Biochemical Sciences* **44**, 7–20 (2019).
194. Overington, J. P., Al-Lazikani, B. & Hopkins, A. L. How many drug targets are there? *Nat. Rev. Drug Discov.* **5**, 993–996 (2006).
195. Hauser, A. S., Attwood, M. M., Rask-Andersen, M., Schiöth, H. B. & Gloriam, D. E. Trends in GPCR drug discovery: New agents, targets and indications. *Nat. Rev. Drug Discov.* **16**, 829–842 (2017).
196. Munk, C. *et al.* An online resource for GPCR structure determination and analysis. *Nat. Methods* **16**, 151–162 (2019).
197. Shimizu, K., Cao, W., Saad, G., Shoji, M. & Terada, T. Comparative analysis of membrane protein structure databases. *Biochimica et Biophysica Acta - Biomembranes* **1860**, 1077–1091 (2018).
198. Pinto, C. *et al.* Formation of the β-barrel assembly machinery complex in lipid bilayers as seen by solid-state NMR. *Nat. Commun.* **9**, (2018).
199. Birch, J. *et al.* The fine art of integral membrane protein crystallisation. *Methods* **147,** 150–162 (2018).
200. Sirdeshmukh, R. Indian proteomics efforts and human proteome project. *Journal of Proteomics* **127,** 147–151 (2015).

201. Koehler Leman, J., Ulmschneider, M. B. & Gray, J. J. Computational modeling of membrane proteins. *Proteins Struct. Funct. Bioinforma.* **83**, 1–24 (2015).
202. Almeida, J. G., Preto, A. J., Koukos, P. I., Bonvin, A. M. J. J. & Moreira, I. S. Membrane proteins structures: A review on computational modeling tools. *Biochimica et Biophysica Acta - Biomembranes* **1859**, 2021–2039 (2017).
203. Jones, D. T. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **292**, 195–202 (1999).
204. Viklund, H. & Elofsson, A. OCTOPUS: Improving topology prediction by two-track ANN-based preference scores and an extended topological grammar. *Bioinformatics* **24**, 1662–1668 (2008).
205. Kelm, S., Shi, J. & Deane, C. M. MEDELLER: Homology-based coordinate generation for membrane proteins. *Bioinformatics* **26**, 2833–2840 (2010).
206. Ebejer, J. P., Hill, J. R., Kelm, S., Shi, J. & Deane, C. M. Memoir: template-based structure prediction for membrane proteins. *Nucleic Acids Res.* **41**, (2013).
207. Šali, A. & Blundell, T. L. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**, 779–815 (1993).
208. Katchalski-Katzir, E. *et al.* Molecular surface recognition: Determination of geometric fit between proteins and their ligands by correlation techniques. *Proc. Natl. Acad. Sci. U. S. A.* **89**, 2195–2199 (1992).
209. Tovchigrechko, A. & Vakser, I. A. Development and testing of an automated approach to protein docking. in *Proteins: Structure, Function and Genetics* **60**, 296–301 (2005).
210. Comeau, S. R., Gatchell, D. W., Vajda, S. & Camacho, C. J. ClusPro: A fully automated algorithm for protein-protein docking. *Nucleic Acids Res.* **32**, (2004).
211. Cheng, T. M. K., Blundell, T. L. & Fernandez-Recio, J. PyDock: Electrostatics and desolvation for effective scoring of rigid-body protein-protein docking. *Proteins Struct. Funct. Genet.* **68**, 503–515 (2007).
212. Mintseris, J. *et al.* Integrating statistical pair potentials into protein complex prediction. *Proteins Struct. Funct. Genet.* **69**, 511–520 (2007).

213. Moal, I. H. & Bates, P. A. SwarmDock and the use of normal modes in protein-protein Docking. *Int. J. Mol. Sci.* **11**, 3623–3648 (2010).
214. Geng, C., Xue, L. C., Roel-Touris, J. & Bonvin, A. M. J. J. Finding the ΔΔG spot: Are predictors of binding affinity changes upon mutations in protein–protein interactions ready for it? *Wiley Interdisciplinary Reviews: Computational Molecular Science* **9**, e1410 (2019).
215. Lensink, M. F. & Wodak, S. J. Score_set: A CAPRI benchmark for scoring protein complexes. *Proteins Struct. Funct. Bioinforma.* **82**, 3163–3169 (2014).
216. Yu, J. & Guerois, R. PPI4Dock: Large scale assessment of the use of homology models in free docking over more than 1000 realistic targets. *Bioinformatics* **32**, 3760–3767 (2016).
217. Kundrotas, P. J. *et al.* Dockground: A comprehensive data resource for modeling of protein complexes. *Protein Sci.* **27**, 172–181 (2018).
218. Geng, C. *et al.* IScore: A novel graph kernel-based function for scoring protein-protein docking models. *Bioinformatics* **36**, 112–121 (2020).
219. Chaudhury, S. *et al.* Benchmarking and analysis of protein docking performance in Rosetta v3.2. *PLoS One* **6**, (2011).
220. Viswanath, S., Dominguez, L., Foster, L. S., Straub, J. E. & Elber, R. Extension of a protein docking algorithm to membranes and applications to amyloid precursor protein dimerization. *Proteins Struct. Funct. Bioinforma.* **83**, 2170–2185 (2015).
221. Hurwitz, N., Schneidman-Duhovny, Di. & Wolfson, H. J. Memdock: An α-helical membrane protein docking algorithm. *Bioinformatics* **32**, 2444–2450 (2016).
222. Alford, R. F. *et al.* An Integrated Framework Advancing Membrane Protein Modeling and Design. *PLoS Comput. Biol.* **11**, e1004398 (2015).
223. Koukos, P. I., Faro, I., van Noort, C. W. & Bonvin, A. M. J. J. A Membrane Protein Complex Docking Benchmark. *J. Mol. Biol.* **430**, 5246–5256 (2018).
224. Newport, T. D., Sansom, M. S. P. & Stansfeld, P. J. The MemProtMD database: A resource for membrane-embedded protein structures and their lipid interactions. *Nucleic Acids Res.* **47**, D390–D397 (2019).

225. DeLano, W. L. Pymol: An open-source molecular graphics tool. *CCP4 Newsl. Protein Crystallogr.* **40**, 82–92 (2002).
226. Fernández-Recio, J., Totrov, M. & Abagyan, R. Identification of protein-protein interaction sites from docking energy landscapes. *J. Mol. Biol.* **335**, 843–865 (2004).
227. Saliba, A. E., Vonkova, I. & Gavin, A. C. The systematic analysis of protein-lipid interactions comes of age. *Nature Reviews Molecular Cell Biology* **16**, 753–761 (2015).
228. Alford, R. F., Fleming, P. J., Fleming, K. G. & Gray, J. J. Protein Structure Prediction and Design in a Biologically Realistic Implicit Membrane. *Biophys. J.* **118**, 2042–2055 (2020).
229. Dancea, F., Kami, K. & Overduin, M. Lipid interaction networks of peripheral membrane proteins revealed by data-driven micelle docking. *Biophys. J.* **94**, 515–524 (2008).
230. Koppisetti, R. K. *et al.* Ambidextrous binding of cell and membrane bilayers by soluble matrix metalloproteinase-12. *Nat. Commun.* **5**, 5552 (2014).
231. Fang, Z. *et al.* Inhibition of K-RAS4B by a Unique Mechanism of Action: Stabilizing Membrane-Dependent Occlusion of the Effector-Binding Site. *Cell Chem. Biol.* **25**, 1327-1336.e4 (2018).
232. Arbib, M. Review of 'Computation: Finite and Infinite Machines' (Minsky, Marvin; 1967). *IEEE Trans. Inf. Theory* **14**, 354–355 (1968).
233. Estrin, G. The WEIZAC Years (1954-1963). *IEEE Ann. Hist. Comput.* **13**, 317–339 (1991).
234. Moult, J., Pedersen, J. T., Judson, R. & Fidelis, K. A large-scale experiment to assess protein structure prediction methods. *Proteins Struct. Funct. Genet.* **23**, ii–iv (1995).
235. Janin, J. *et al.* CAPRI: A Critical Assessment of PRedicted Interactions. *Proteins Struct. Funct. Genet.* **52**, 2–9 (2003).
236. Gathiaka, S. *et al.* D3R grand challenge 2015: Evaluation of protein–ligand pose and affinity predictions. *J. Comput. Aided. Mol. Des.* **30**, 651–668 (2016).

237. Lensink, M. F., Nadzirin, N., Velankar, S. & Wodak, S. J. Modeling protein-protein, protein-peptide, and protein-oligosaccharide complexes: CAPRI 7th edition. *Proteins Struct. Funct. Bioinforma.* **88**, 916–938 (2020).
238. Koukos, P. I. *et al.* An overview of data-driven HADDOCK strategies in CAPRI rounds 38-45. *Proteins Struct. Funct. Bioinforma.* **88**, 1029–1036 (2020).
239. Peng, Z. *et al.* Exceptionally abundant exceptions: comprehensive characterization of intrinsic disorder in all domains of life. *Cell. Mol. Life Sci.* **72**, 137–151 (2015).
240. Dunker, A. K., Cortese, M. S., Romero, P., Iakoucheva, L. M. & Uversky, V. N. Flexible nets. The roles of intrinsic disorder in protein interaction networks. *FEBS J.* **272**, 5129–5148 (2005).
241. Best, R. B., Buchete, N.-V. & Hummer, G. Are Current Molecular Dynamics Force Fields too Helical? *Biophys. J.* **95**, L07–L09 (2008).
242. Lange, O. F., Van Der Spoel, D. & De Groot, B. L. Scrutinizing molecular mechanics force fields on the submicrosecond timescale with NMR Data. *Biophys. J.* **99**, 647–655 (2010).
243. Lindorff-Larsen, K., Piana, S., Dror, R. O. & Shaw, D. E. How fast-folding proteins fold. *Science .* **334**, 517–520 (2011).
244. Lindorff-Larsen, K. *et al.* Systematic validation of protein force fields against experimental data. *PLoS One* **7**, e32131 (2012).
245. Beauchamp, K. A., Lin, Y. S., Das, R. & Pande, V. S. Are protein force fields getting better? A systematic benchmark on 524 diverse NMR measurements. *J. Chem. Theory Comput.* **8**, 1409–1414 (2012).
246. Lindorff-Larsen, K. *et al.* Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins Struct. Funct. Bioinforma.* **78**, 1950–1958 (2010).
247. Huang, J. *et al.* CHARMM36m: An improved force field for folded and intrinsically disordered proteins. *Nat. Methods* **14**, 71–73 (2016).
248. Robustelli, P., Piana, S. & Shaw, D. E. Developing a molecular dynamics force field for both folded and disordered protein states. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E4758-E4766 (2018).

249. Best, R. B. Computational and theoretical advances in studies of intrinsically disordered proteins. *Curr. Opin. Struct. Biol.* **42**, 147–154 (2017).
250. Levine, Z. A. & Shea, J. E. Simulations of disordered proteins and systems with conformational heterogeneity. *Current Opinion in Structural Biology* **43**, 95–103 (2017).
251. Corradi, V. *et al.* Emerging Diversity in Lipid–Protein Interactions. *Chem. Rev.* **119**, 5775–5848 (2019).
252. Lee, A. G. Biological membranes: the importance of molecular detail. *Trends Biochem. Sci.* **36**, 493–500 (2011).
253. Shukla, R. *et al.* Mode of action of teixobactins in cellular membranes. *Nat. Commun.* **11**, 2848 (2020).
254. Watanabe, Y. *et al.* Vulnerabilities in coronavirus glycan shields despite extensive glycosylation. *Nat. Commun.* **11**, 2688, (2020).
255. Casalino, L. *et al.* Beyond Shielding: The Roles of Glycans in the SARS-CoV-2 Spike Protein. *ACS Cent. Sci.* **6**, 1722–1734 (2020).
256. Durrant, J. D. *et al.* Mesoscale All-Atom Influenza Virus Simulations Suggest New Substrate Binding Mechanism. *ACS Cent. Sci.* **6**, 189–196 (2020).
257. Seitz, C. *et al.* Multiscale simulations examining glycan shield effects on drug binding to influenza neuraminidase. *bioRxiv* doi: 10.1101/2020.08.12.248690 (2020).
258. Yu, A. *et al.* A Multiscale Coarse-grained Model of the SARS-CoV-2 Virion. *bioRxiv* doi: 10.1101/2020.10.02.323915 (2020).

Summary

Biomolecular interactions are critical in cellular environments. In particular, proteins, which are the workhorses of the cellular machinery, mediate by their interactions a wide range of molecular processes within the cell. Structural Biology is the scientific discipline concerned with revealing the molecular functions of these macromolecules through analysis of their three-dimensional structures. Classical structural biology techniques include X-ray crystallography, Nuclear Magnetic Resonance (NMR) and cryo-Electron Microscopy (cryo-EM). As any other method, these experimental techniques have limitations that preclude their application to all biological systems. For example, large proteins (>50 kDa) are difficult to study by NMR spectroscopy and X-ray crystallography requires high quality crystals, which is not always trivial to achieve. For some specific systems, such as membrane proteins, their characterization by purely experimental techniques in their native environment is still challenging.

Computational Structural Biology is a consolidated branch of science, whose goal is to understand the role that structure and dynamics play in the definition of the function of biomolecular systems. In particular, biomolecular interactions have been a major focus of this field over the past decades. For this purpose, various computational approaches have been designed and applied to the modelling of interactions, among which molecular dynamics- Monte Carlo-, docking- and, more recently, template modeling-based methods are the most widely used ones. Roughly, docking methods aim to build three-dimensional models of macromolecular structures by first, generating thousands of possible conformations (models), and then discriminating between biologically- and non-biologically-relevant models. Docking can be performed in the absence of any experimental information *(ab initio)* or by integrating information into the calculations *(data-driven)*. In this thesis, several developments into the modeling of protein-protein and protein-nucleic acids interactions by computational integrative

modeling approaches are presented. The thesis starts with a review of various representative models for downscaling the resolution of proteins, peptides and nucleic acids for the integrative modeling of their interactions (**Chapter 1**). These simplifications have two clear advantages: (1) It is easier to identify putative binding regions and (2) the computations become much more efficient. Real applications of the use of simplified model for modelling proteins and nucleic acids complexes are described, demonstrating that coarse-graining leads to more native-like models with a remarkable speed increase (**Chapters 3** and **4**). The implementation of information into the LightDock algorithm to both drive docking and scoring is described in **Chapter 2.** In this case, the use of experimental data such as mutagenesis data, translates into an increase in performance even when the data are not accurate and/or partially incorrect. In the final chapter (**Chapter 5**), several of the developments described in previous chapters (**2** and **3**) are combined for the modeling membrane-associated assemblies which are notoriously difficult to tackle. These systems are of special importance since are directly related to many diseases and therefore, are potentials target for drug design purposes.

The thesis ends with a **Conclusions and Perspectives** section, giving a brief overview of chronological advances in both computing and Computational Structural Biology fields, together with some of the still open questions and challenges to be resolved in the near future.

Samenvatting

In de context van een levende cel, zijn biomoleculaire interacties van cruciaal belang. De werkpaarden die deze interacties mediëren zijn vooral de eiwitten. Zij zijn betrokken bij een groot scala aan cellulaire processen. De wijze waarop in de Structuur Biologie wordt geprobeerd om de functies van eiwitten in deze processen te duiden, is door de driedimensionale structuur van deze moleculen en complexen te achterhalen. De klassieke technieken binnen de structuur biologie zijn röntgenstraling eiwitkristallografie, biomoleculaire kernspinresonantie (NMR) en cryo-elektronenmicroscopie (cryo-EM). Al deze vakgebieden hebben hun specifieke limitaties, waardoor ze minder geschikt zijn om bepaalde biologische systemen te onderzoeken. Zo zijn bijvoorbeeld grote eiwitten (>50kDa) moeilijk te gebruiken voor NMR en kan het laten groeien van goede kwaliteit kristallen vaak ook zeer uitdagend zijn. In weer andere gevallen is het juist lastig om met de bovengenoemde methodes biologisch relevante condities te bereiken, zoals bijvoorbeeld de aanwezigheid van membranen.

Computationele Structuur Biologie is een tak van wetenschap die op heel andere manieren begrenst is dan de klassieke methodes, en heeft daarom grote toegevoegde waarde. Het veld heeft zich tot doel gesteld om de rol van structuur en dynamica in biomoleculaire systemen beter te begrijpen, en heeft belangrijke inzichten opgeleverd die zeer nuttig zijn gebleken in het lab. De laatste decennia heeft de focus vooral gelegen op biomoleculaire interacties tussen eiwitten en eiwitcomplexen. Hiertoe zijn verschillende computationele methodes ontwikkeld en toegepast om interacties te modelleren *(in silico)*. De meeste gebruikte zijn o.a. moleculaire dynamica-, Monte Carlo-, docking- en recenter, template-gebaseerde modelleer algoritmes. Kort door de bocht, kan over docking worden gezegd dat er geprobeerd wordt om een 3D-model te maken van macromoleculaire complexen door eerst duizenden mogelijke conformaties te generen (het maken van modellen), en hieruit de biologisch relevante modellen uit te halen

(het scoren van de modellen). Wordt dit proces uitgevoerd zonder enige experimentele informatie vooraf, dan heet dat *ab initio* modelleren. Als daarentegen bestaande informatie in het proces geïntegreerd wordt dan heet het data-driven. In dit proefschrift komen verschillende ontwikkelingen ter sprake op het gebied van integraal modelleren van eiwit-eiwit en eiwit-nucleïnezuur interacties.

Allereerst zal een overzicht gepresenteerd worden van verschillende methodes om lagere resolutie modellen te genereren van peptiden, eiwitten en nucleïnezuren om integraal hun interacties te modelleren (**Hoofdstuk 1**). Het simplificeren van het proces door middel van het gebruik van deze lagere resolutie modellen heeft twee belangrijke voordelen boven conventioneel modelleren: (1) De identificatie van de vermoedelijke binding regio's wordt vergemakkelijkt, (2) Het berekenen van de modellen wordt veel efficiënter wat tijdswinst tot gevolg heeft. Verschillende toepassingen van de lagere-resolutie modellen met eiwitten en nucleïnezuren hebben geleid tot meer biologisch relevante (native) modellen, en ook nog eens in significant minder tijd dan conventionele methodes (**Hoofdstuk 3 en 4**). De implementatie van extra informatie aan het LightDock algoritme om de docking en scoring te sturen staat beschreven in **Hoofdstuk 2**. Hier wordt aangetoond dat het gebruik van additionele experimentele informatie zoals mutagenese efficiëntie zal doen toenemen, zelfs als de extra informatie niet geheel accuraat of zelfs gedeeltelijk foutief is. In het laatste hoofdstuk (**Hoofdstuk 5**) worden de methodes omschreven die in de **Hoofdstukken 2** en **3** gebruikt zijn om enkele notoir lastige, membraangebonden complexen te modelleren. Het belang van beter begrip van zulke complexen wordt geïllustreerd door hun relevantie in vele ziektes. Meer inzicht in de structuur van zulke membraangebonden complexen zou in de toekomst kunnen helpen bij het ontwikkelen van medicijnen en behandelings strategieën.

Dit proefschrift wordt afgesloten door een hoofdstuk met **Conclusies en Perspectieven** waarin een bondig, chronologisch overzicht wordt gegeven van de ontwikkelingen in zowel de computertechnologie als de Computationele Structurele Biologie. Ook de open vragen en uitdagingen waar het veld momenteel voor staat, zullen worden besproken.

Resumen

Las relaciones entre biomoléculas juegan un papel muy importante en el funcionamiento de las células. En especial, las proteínas, y sus interacciones, arbitran una gran variedad de mecanismos moleculares. Por este motivo, se les refiere comúnmente como los motores de la maquinaria celular. La Biología Estructural es una disciplina científica enfocada en revelar las funciones moleculares que llevan a cabo estas macromoléculas a través del análisis de sus estructuras tridimensionales. Los métodos experimentales más empleados para ello son: la cristalografía de rayos X, la resonancia magnética nuclear y la microscopía electrónica criogénica. Al igual que cualquier método, estas técnicas experimentales tienen limitaciones que imposibilitan su aplicación a cualquier sistema biológico. Por ejemplo, métodos como la resonancia magnética nuclear no son adecuados para el estudio de proteínas de gran tamaño (>50 kDa), y técnicas como la cristalografía de rayos X, que requiere la obtención de cristales de alta calidad, hecho que normalmente entraña una gran dificultad. Concretamente, el estudio estructural de proteínas de membrana por métodos puramente experimentales, resulta muy complejo en condiciones naturales.

La Biología Estructural Computacional es una rama consolidada de la ciencia, cuyo objetivo principal es entender el papel que la estructura y la dinámica juegan en el funcionamiento de sistemas biomoleculares. Específicamente, durante las últimas décadas los esfuerzos se han centrado en el estudio de las relaciones entre biomoléculas. Con este fin, existen múltiples métodos computacionales entre los cuales, los basados en Dinámica Molecular, Monte Carlo, *docking*, y, más recientemente *template-based*, son ampliamente los más utilizados. Grosso modo, los métodos de *docking* tienen como finalidad la creación de modelos macromoleculares en tres dimensiones mediante, en primer lugar, la generación de miles de posibles diferentes conformaciones (modelos) para después diferenciar aquellos modelos relevantes desde el punto de vista

biológico de los irrelevantes. Los métodos de *docking* se pueden emplear en ausencia de cualquier tipo de información experimental (*ab initio*) o, por el contrario, pueden ser combinados con dicha información *(data-driven*).

La presente tesis comienza con una revisión de diferentes aproximaciones para la reducción de la complejidad de proteínas, péptidos y ácidos nucleicos, con el objetivo del estudio de sus posibles interacciones (**Capítulo 1**). Estas simplificaciones presentan dos ventajas claras: (1) Se facilita la identificación de las posibles regiones de interacción y (2), los cálculos computacionales se vuelven más eficientes. De este modo, esta tesis incluye aplicaciones reales de dichos modelos simplificados en el modelado de proteínas y ácidos nucleicos, demostrando así que estas simplificaciones conllevan la generación de un mayor número de modelos biológicamente relevantes, al igual que una remarcable ganancia en velocidad de computación (**Capítulos 3** y **4**). En el **Capítulo 2**, se describe la implementación del uso de información experimental en el algoritmo de LightDock. En este contexto, el uso de dicha información, como aquella derivada de experimentos de mutagénesis, se traduce en un incremento de la eficacia del algoritmo incluso cuando la información no es completa y/o parcialmente incorrecta. Por último, en el **Capítulo 5**, se combinan varios de los desarrollos previamente descritos (**Capítulos 2** y **3**) para dar lugar a un nuevo método aplicado al modelado de proteínas de membrana. Este tipo particular de proteínas, muy difíciles de caracterizar, son de especial importancia puesto que están directamente asociadas a múltiples enfermedades, y, por consiguiente, son de especial interés para el desarrollo de nuevos fármacos. La tesis finaliza con una sección de **Conclusiones y Perspectivas**, la cual incluye un resumen en orden cronológico de los avances tanto en computación como en el campo de la Biología Estructural Computacional, sumado a algunas de las preguntas y desafíos aún sin resolver en estos campos.

Acknowledgements

During my last year of high school, my biology teacher thought my qualifications should not get higher than 7 (out of 10) since my way of understanding biology was not *"the appropriate one"*. It is worth noting that our discrepancies usually had a not very desirable output for me and I ended up spending more time wandering around the school than attending biology lectures. I have to admit that this gave me a hard time but now, looking back with some perspective, I do believe this might have somehow contributed to where I currently stand.

My internship at the Netherlands Institute for Neuroscience was definitely a game changer for me. During my time at the Neuromodulation and Behavior group leaded at that time by Matthijs Feenstra (now happily retired) under the daily (sometimes *dad-ly*) supervision of Ralph Hamelink, I realized that wet-lab work was not meant for me. Ralph not only taught me numerous experimental techniques, but more importantly, he taught me rigorous science, good practices and time management, which I have been trying to apply during my PhD journey. I really had a great time and truly appreciate all those conversations that helped me to shape my present and recent past.

My second experience in a research group was at the *Protein Interactions and Docking* group formerly headed by Juan Fernández-Recio at the Barcelona Supercomputing Center. To be honest, the *"Supercomputing"* word was scary considering my limited programming skills at that time. However, meeting you Brian was the most enriching academic experience of that stage. Everything was so easy for you that sometimes it made me feel dumb, but your guidance and patience have been key. Indeed, you are one of the culprits for me writing these lines (more personal acknowledgements will come below). Also, especial mention to Lucía, Sergio, Chiara, Miguel, Mireia, Didier, Luis and of course Juan, for creating such a great working environment.

Once, somebody said: *"third time lucky"*, and indeed, my third experience in a research group has turned into a PhD dissertation. When I firstly emailed you Alexandre, I was not sure about enrolling into a PhD at all. However, I knew your group was the better place for it and I do not regret it. When we met in Utrecht for the first time, I had not the feeling of being interviewed and this was due to your overdeveloped abilities to relate to others, which made me feel valued from the very first moment. When it comes to work, you are always the first one doing the job and I really appreciate that you never minded to get your hands dirty every time I needed your help, as for example when those coarse-grained models were inexplicably exploding even before being docked. We had the opportunity to travel together to the beautiful island of Corsica or to the unforgettable BIOMOS meeting. Looking back with perspective, the BIOMOS meeting reminds me when you are teaching somebody to swim and you just throw them into the swimming pool with no life vest. Believe it or not, it was still an interesting experience. These trips gave me the opportunity to discover that we share hobbies like a good beer or our love for tennis, and of course, for Federer. I deeply appreciate the opportunity you have given me and your support and freedom, which have triggered my creativity and passion about science.

Big thanks to the CSB, and NMR, groups as whole for the vibrant scientific atmosphere and cordiality. In particular, to the CSB former members: Anna, Jörg, Li, Mikael… To Liang, my beloved office mate: It was an honor being your paranymph and I am very proud of what you have achieved so far. Adrien, the boss in the shadow, who is always trying to push the limits and make things better. Thanks for all the support and I wish you nothing but the best. I truly missed you at the MolMod course, and I am sure the students too. Also, to João (Azores) for your positivity and kindness and to Barend for the night life (including conferences) and for your translation support. To the current

CSB members: Panos, the *man-in-black*. I am very glad to have met you and thank you for being on my side. Charlotte, you are the next one and I wish you all the best. The late comers, Manon and Siri, a.k.a. the noisiest office mate ever and the fanciest neighbor. Best of luck for your future careers. Also, big thanks to the rest of colleagues who contributed to the content of this thesis.

To not forget the senior cluster: Rolf thank you for trying to teach me about NMR and non-NOE's. Gert, you personify the critical thinking so essential in science and unfortunately not so abundant sometimes. Hugo, your questions and feedback during the (not always appealing) NMR group meetings have been of invaluable help. Mark, thank you for all those chats trying to save the world over a couple (or three) of beers. Markus, you have a brilliant future ahead and your great work has started to pay off. Johan, the guardian angel of the NMR bunker and Marc, the head of the department, for trying to keep the mood up with all of those unexpected cakes and sweets. Finally, thanks to the best secretaries in the world. Barbara, for helping me during the always difficult beginnings. And Geeske, for your help during these last stages of my PhD.

Brian and Zuzana, my two paranymphs. I want to thank you Brian for the *crazy ideas'* moments and all those conversations about everything and nothing at the same time. You are an extraordinary scientist, but most importantly, an extraordinary person. I am sure that all the efforts you are currently making will eventually lead to a group leader position, where you will be able to shine as much as you deserve. Zuzana, when we firstly met at the unmentionable meeting, I already realized that you are one of a kind. I want to thank you for letting me be your friend and for all those coffee breaks and beers (sometimes too many). It is always so much fun to be around you. I wish you all the best and when there is a change, there is also an opportunity. I want to also take this opportunity to apologize to my

Pint of Science fellows for being a bit disconnected during this time. Here you have the reason why. You are doing an extraordinary work to bring science to the general audience and I want to thank you for letting me be part of such an amazing project. Looking really forward to #pint2021 edition!

Y ahora cambiamos de idioma. Me gustaría empezar agradeciendo a dos de las personas más importantes que tengo en mi vida. Aitas, esto es sólo un paso más en el camino que empezamos juntos hace más de 10 años. Ya sabéis que no soy de muchas palabras, pero creo que la ocasión lo merece. Quiero daros las gracias por la educación que me habéis dado, aunque muchas veces no os lo haya puesto demasiado fácil. Ama, gracias por tu infinita bondad y predisposición a ayudar. Cada vez que me encuentro con alguna dificultad, se me viene a la mente la imagen de ti y tus largas horas de estudio. Gracias por enseñarme que el trabajo y la dedicación son la única manera. Porque siempre es más fácil de resolver muchos problemas pequeños que un problema grande. Aita, más allá de tu inteligencia y talento innatos, durante todos estos años nos has demostrado que, aunque la vida no sea fácil, siempre sales adelante. Eres sin ninguna duda mi ejemplo a seguir, y gracias por tu i nfinita paciencia (incluyendo todas aquellas noches estudiando historia…). Gracias también a mi familia de Mallorca: Mónica, Pablo y Aitor, que a pesar de la distancia os tengo muy presentes, así como al resto de miembros de la familia (demasiados nombres que incluir). Por último, me gustaría recordar a las personas que ya no están con nosotros, y en especial a ti abueliña. Gracias por cuidarme siempre y por enseñarme el valor del trabajo y de la humildad.

Creo que sobran las palabras contigo Eder. Durante todos estos años hemos reído, hemos llorado, nos hemos enfadado y nos hemos perdonado. A pesar de que pueda parecer que nuestros caminos se separan más y más, siempre hemos encontrado la manera de seguir unidos. Eres sin duda un

luchador y un ejemplo de superación y valentía. Estoy muy orgulloso de poder considerarte mi amigo y estoy convencido de que lo mejor está aún por venir. Gracias a ti también Janire por aparecer en el momento oportuno. Tienes todo mi cariño y respeto como persona y también como amiga. Sin duda te llevas el premio a la jugadora revelación. No me olvido de la gente del pádel y de todos los *"clinics*", exhibiciones e inauguraciones de club por todo el país.

Canvi de terç. Abans de començar, m'agradaria demanar perdó a Pompeu Fabra per les possibles faltes d'ortografia, no es la meva intenció. A la meva família de Catalunya: Marisol (la Mili), Jordi, Nil i el Boyito *"Bunicus Guapus"*, gràcies per acollir-me com a un més des del primer moment. També, gràcies per tots els consells y per totes les provisions que han fet que la distància no semblés tanta. Clàudia, merci per les teves visites. Vull que sàpigues que ets com la meva germana petita eta Aritz, Arrasateko txirrindulari onena. Jarraitu padel praktikatzen. També, a l'Anna i el Manel per tots el bons moments a Holanda. Sense dubte heu fet que la meva, i nostra, estància aquí hagi estat molt més suportable. Us desitjo el millor i estic segur de que mantindrem el contacte.

Finalment, i no per això menys important, l'amor de la meva vida. Victòria, fa uns anys que vam decidir emprendre aquesta aventura junts i la acabarem junts també. Durant aquest temps, hem passat moments millors i pitjors tan professional com personalment. Malgrat això, hem estat capaços superar-los junts i de construir una vida conjunta. Vull donar-te les gràcies pel teu suport incondicional i per no jutjar-me mai. M'has ensenyat que les dificultats es poden superar des del positivisme i amb treball. De fet, tu ets l'exemple de constància i esforç.

Aquí i ara, puc dir que sense tu, res d'això hagués estat possible. Sense dubte, tu em fas millor persona i vull donar-te les gràcies per deixar-me formar part de la teva vida.

Que vull estar amb tu, jo vull estar amb tu.

In life, everything has an end and so this thesis. Thank you very much and see you around.

En la vida, todo tiene un final y esta tesis termina aquí. Muchas gracias a todas y todos. Nos vamos viendo.

List of Publications

12. **J. Roel-Touris**#, B. Jiménez-García# & A. M. J. J. Bonvin. Integrative Modeling of Membrane-associated Protein Assemblies. *Nature Communications.* **11**, 6210 2020)

11. **J. Roel-Touris**, & A. M. J. J. Bonvin. Coarse-Grained (Hybrid) Integrative Modeling of Biomolecular Interactions. *Computational and Structural Biotechnology Journal.* **18**, 1182-1190 (2020)

10. P.I. Koukos, **J. Roel-Touris**, F. Ambrosetti, C. Geng, J. Schaarschmidt, M.E. Trellet, A.S.J. Melquiond, L.C. Xue, R.V. Honorato, I. Moreira, Z. Kurkcuoglu, A. Vangone & A.M.J.J. Bonvin. An overview of data-driven HADDOCK strategies in CAPRI rounds 38-45. *Proteins: Struc. Funct. & Bioinformatics*. **88**, 1029-1036 (2020).

9. F. Ambrosetti, B. Jiménez-García, **J. Roel-Touris** & A. M. J. J. Bonvin. Modeling Antibody-Antigen Complexes by Information-Driven Docking. *Structure*. **28**, 119-129 (2020).

8. **J. Roel-Touris**, A.M.J.J. Bonvin & B. Jiménez-García. LightDock goes information-driven. *Bioinformatics*, **36**, 950–952 (2020).

7. M.F. Lensink, CAPRI PARTICIPANTS, P.I. Koukos, **J. Roel-Touris**, F. Ambrosetti, C. Geng. J. Schaarschmidt, M.E. Trellet, A.S.J. Melquiond, L. Xue, B. Jiménez-García, C.W. van Noort, R.V. Honorato, A.M.J.J. Bonvin & S.J. Wodak. Blind prediction of homo- and hetero-protein complexes: The CASP13-CAPRI experiment. *Proteins: Struc. Funct. & Bioinformatics*, **87**, 1200-1221 (2019)

6. R.V. Honorato#, **J. Roel-Touris**# & A.M.J.J. Bonvin. MARTINI-based protein-DNA coarse-grained HADDOCKing. *Frontiers in Molecular Biosciences*, **6**, 102 (2019).

5. **J. Roel-Touris**, C.G. Don, R.V. Honorato, J.P.G.L.M Rodrigues & A.M.J.J. Bonvin. Less is more: Coarse-grained integrative modeling of large biomolecular assemblies with HADDOCK. *J. Chem. Theory Comput.* **15**, 6358-6367 (2019).

4. Geng. L. Xue, **J. Roel-Touris** & A.M.J.J. Bonvin. Finding the ΔΔG spot: Are predictors of binding affinity changes upon mutations in protein-protein interactions ready for it? *WIREs Computational Molecular Science*. **9**, e1410 (2019).

3. A.M.J.J. Bonvin, C. Geng, M. van Dijk, E. Karaca, P. L. Kastritis, P.I. Koukos, Z. Kurkcuoglu, A.S.J. Melquiond, J.P.G.L.M. Rodrigues, J. Schaarschmidt, C. Schmitz, **J. Roel-Touris**, M.E. Trellet, S. de Vries, A. Vangone, L. Xue & G.C.P. van Zundert HADDOCK. *Encyclopedia of Biophysics*. In press (2018).

2. Z. Kurkcuoglu, P.I. Koukos, N. Citro, M.E. Trellet, J.P.G.L.M. Rodrigues, I.S. Moreira, **J. Roel-Touris**, A.S.J. Melquiond, C. Geng, J. Schaarschmidt, L.C. Xue, A. Vangone & A.M.J.J. Bonvin. Performance of HADDOCK and a simple contact-based protein-ligand binding affinity predictor in the D3R Grand Challenge 2. J. Comp. Aid. Mol. Des. 32, 175-185 (2018).

1. B. Jiménez-García, **J. Roel-Touris**, M. Romero-Durana, M. Vidal, D. Jiménez-González & J. Fernández-Recio. LightDock: a new multi-scale approach to protein–protein docking. *Bioinformatics*. **34**, 49-55 (2018).

# These authors contributed equally to the work.

Curriculum Vitae

Jorge was born on the 17th of July 1990 and raised in Santurtzi, Basque Country (Euskadi). He attended school in Santa María Ikastetxea and later on enrolled into a five-year bachelor degree in Biology at Euskal Herriko Unibertsitatea. After a year, in 2009, he moved to Barcelona for a four-year study programme in Biochemistry at Universitat Autònoma de Barcelona. In 2013, he did a one-year internship at the Netherlands Institute of Neuroscience in the pre-clinical group leaded, at that time, by Matthjis Feenstra. In 2016, he graduated from a two-year master programme in Bioinformatics for Health Science at Universitat Pompeu Fabra (Barcelona), after a one-year internship with Juan Fernández-Recio at the Barcelona Supercomputing Center. Since February 2017, Jorge has been employed by Universiteit Utrecht in the group of Alexandre Bonvin, conducting research towards a doctoral degree.