Integrative Modelling of Biomolecular Complexes: From Small to Large

Panagiotis Koukos

ISBN/EAN 978-9-03-937247-0
Doctoral thesis
Integrative Modelling of Biomolecular Complexes: From Small to Large
Panagiotis Koukos
Computational Structural Biology group, Bijvoet Center for Biomolecular Research,
Faculty of Chemistry, Utrecht University, The Netherlands
January 2020 Copyright © 2020 Panagiotis Koukos

Printed in Greece by Kivos

Integrative Modelling of Biomolecular Complexes: From Small to Large

Integratief Modelleren van Biomoleculaire Complexen:
Van Klein tot Groot
(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 woensdag 26 februari 2020 des middags te 12.45 uur

door

Panagiotis Koukos

geboren op 2 september 1989
te Athene, Griekenland

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

Beoordelingscommissie:	Prof. dr. Syma Khalid Prof. dr. Antoinette Killian dr. Gerry Nicolaes Prof. dr. Anastassis Perrakis Prof. dr. Berend Snel

Table of Contents

General Introduction **1**

Chapter 1

Integrative modelling of biomolecular complexes **9**

Chapter 2

A membrane protein complex docking benchmark **31**

Chapter 3

Modelling of membrane protein complexes with HADDOCK **57**

Chapter 4

Performance of HADDOCK and a simple contact-based protein-ligand binding affinity predictor in the D3R Grand Challenge 2 **71**

Chapter 5

Protein-ligand pose and affinity prediction. Lessons from D3R Grand Challenge 3. **91**

Chapter 6

Shape-based modelling of protein-small molecule complexes with HADDOCK **111**

Chapter 7

Conclusions and Perspectives **125**

References **133**

Summary **153**

Samenvatting **159**

Περίληψη **165**

Acknowledgements **173**

List of Publications **181**

Curriculum Vitae **183**

General Introduction

Origins of Computational Structural Biology

The field of computational structural biology as we know it today traces its origins to the mid-to-late 1960s and early 1970s when the first computer simulations of small organic and inorganic systems became available. At the same time, another branch of the biomolecular simulation community was developing crude simulations of small proteins which, while inaccurate, paved the way for more accurate forcefields and simulation algorithms over the years that followed. The reasons behind these developments were simple enough: The first scientists – who at the time were working in a field made up of diverse disciplines and which later came to be formally known as Structural Biology – had been determining structural models of biomolecules for years, starting with the double helix model of DNA and the protein structures of myoglobin and haemoglobin. In order to create the 3D structure, parameters such as bond lengths or angles between consecutively bonded atoms were manually adjusted and combined with the primary experimental information, which, for most early biomolecular structures, was derived from X-ray crystallography diffraction data. This process was slow, painstaking and error prone, slowing down structure determination after the already arduous process of growing crystals and obtaining diffraction patterns. It was, therefore, in the best interest of the community to develop ways in which this process could be automated and sped up. At the same time, computers were becoming more powerful and common enough that chemical information such as known lengths of bonds between specific atom types, or angles, torsions between consecutive atoms and interactions between non-bonded atoms could be encoded into computer programs which could then be used in combination with the experimental data to arrive at the 3D structure.

Emergence of Biomolecular Docking

The field saw further advances in the years that followed with the development of the first docking programs in the late 1970s, early 1980s. These programs differed from the modelling software previously developed because they were intended to be used to study the interactions between biomolecules rather than single structures. Advances in docking algorithms and computer hardware and architecture increased the applicability of these programs and of biomolecular simulation in general, allowing for increasing levels of chemical correctness to be incorporated in the simulation process. By the late 1990s/early 2000s several publicly available docking programs existed. The most significant development of that era was the CAPRI initiative. CAPRI stands for **C**ritical **A**ssessment of **PR**ediction of **I**nteractions and was modelled on CASP (**C**ritical **A**ssessment of **S**tructure **P**rediction). CASP had been launched a few years earlier with the aim to assess the state-of-the-art in protein structure prediction algorithms. It ran every other year. In early iterations the emphasis was on predicting protein tertiary (and in some cases quaternary) structure from the residue sequence alone. Unlike CASP which runs every second year, CAPRI targets are announced on an irregular basis, mostly depending on their availability. The emphasis – for the most part – has been on assessing the accuracy with which human teams and docking servers can predict the structure of a complex of interacting biomolecules with a main focus on protein-protein complexes. CAPRI has been instrumental in pushing the boundaries of what was considered possible in computational modelling at the time, while also fostering communication and scientific exchange between members of the docking community.

The Advent of Integrative Modelling

Another hallmark for the field in the early 2000s was the publication of HADDOCK (**H**igh **A**mbiguity **D**riven **DOCK**ing), the first docking program which could incorporate experimental information in the simulations in order to guide those toward conformations that would satisfy those data. This was a radical departure from most docking codes published until that point, which, at that time, focused on exhaustively sampling the orientational landscape of the interacting biomolecules and relied on the use of scoring functions to discriminate between good – native-like – and bad – non-native-like – models. These two concepts of sampling (the process by which docking poses are generated) and scoring (the process by which good models are identified from a large pool of models) lie at the core of every discussion around docking and computational modelling in general. The use of experimental information enabled HADDOCK to effectively focus the sampling to the parts of the interaction landscape close to the native complex when using data that represented that native state. However, when those data are not reliable or even incorrect, using them might result in the generation of models which do not get close to the native state at all. The fact that input data have a significant impact on the outcome of the docking is one of the major challenges associated with the field of integrative modelling. For the first few years, HADDOCK could only make use of distance or residue-based information regarding the residues located at the putative interface of a given complex. Later versions of the software gained the capability to include additional information such as the relative orientation of two complexes biomolecules or shape-based information. Additionally, support was extended to biomolecules other than proteins such as nucleic acids and small molecules. These developments mirrored the rise of a new field centred around the integration of data into biomolecular simulations: Integrative Modelling. Integration of diverse data sources in a cohesive and probabilistically sound way represents another challenge unique to this field. HADDOCK is only one such Integrative Modelling software. Examples of codes which can also include data in the simulation in order to drive the sampling toward specific conformations are ROSETTA, IMP (Integrative Modelling Platform), ISD (Inferential Structure Determination) and ATTRACT. Additionally, even programs that cannot directly incorporate experimental data in the simulation to bias the sampling, can usually take advantage of those in the form of post-processing filters during the scoring of the generated poses.

Modelling of membrane protein complexes and small molecule docking

Integrative modelling approaches, docking in particular, have been applied to the study of many diverse systems. Support for the docking of transmembrane protein complexes has however been underwhelming so far as very few codes have implemented protocols tailored to the membrane environment. On one side, the very nature of the membrane bilayer restricts the translational and rotational landscape of the complex, on the other side the fact that the complexes are not surrounded – at least entirely – by water requires adaptations in energetics parameters such as empirical desolvation potentials, which have been in most cases optimised under the assumption that the complex is surrounded only by water. This apparent lack of membrane-specific optimisations is surprising, given the significance of membrane proteins in general and membrane protein interactions specifically, for the enduring survival of the cell, a significance which is also reflected in the number of drugs which target membrane proteins. One of the limiting factors has been the scarce number of membrane proteins for which 3D structures are available in the Protein Data Bank (PDB), the public repository for experimentally determined biomolecular structures. The recent explosion in the number of available membrane protein structures in the PDB, mostly as a result of the revolution the cryo-EM field has been undergoing in recent years, but also due to X-ray crystallography, has been largely responsible for a slew of related developments such as the proliferation of membrane protein-specific databases. While small molecule docking does not suffer from the same lack of attention that membrane modelling does, integrative protocols which take advantage of existing experimental information and integrate it in the simulation are still few.

Overview of thesis

This thesis mainly deals with three subjects which, in order of appearance, are: (1) data sources which can be used by integrative modelling frameworks (**Chapter 1**), (2) modelling of membrane protein complexes (**Chapters 2 and 3**) and (3) small molecule docking (**Chapters 4 through 6**). All these are examined through the prism of integrative modelling and biomolecular docking.

Chapter 1 provides an overview over some of the data sources most frequently used in integrative modelling frameworks. The experimental methods are grouped according to the nature of the information they yield in three sets: Interface-mapping methods, distance-based methods and shape-based ones. Their relative advantages and disadvantages are discussed and their future potential in the field of Integrative modelling is discussed. Recent advances in computational methods like coevolution and coarse-grained forcefields are also discussed.

Chapter 2 describes a recently published docking benchmark consisting entirely of membrane protein-protein complexes. This is the first benchmark of its kind and features ready-to-dock structures of varying difficulties (ranging from bound cases to challenging cases with significant conformational rearrangements at the binding interface). The benchmark has been used to establish the baseline performance of HADDOCK in membrane protein complexes. **Chapter 3** details a method for the docking of transmembrane protein-protein complexes. This protocol is based on the use of restraints to drive the transmembrane part of the system toward shape beads which implicitly represent the lipid bilayer. I also compare the results with runs in which only centre-of-mass restraints were used and discuss possible future avenues worth exploring.

Chapter 4 is the first chapter which deals with small molecule docking. In it, I describe the protocols we developed for dealing with protein-small molecule docking for the participation of the HADDOCK group in the 2016 iteration of the blind docking challenge organised by the D3R consortium. In **Chapter 5** I describe an improved version of the protocol described in chapter 4, which makes use of ligand and compound shape similarities to identify the best receptor template and the best generated ligand conformations prior to docking. In **Chapter 6** I describe a recently developed protocol for which shape information is extracted from suitable templates and represented as shape beads. The generated ligand conformers are driven to this shape within the protein via the use of ambiguous distance restraints. This protocol outperforms all our previous efforts.

In the final chapter, **Chapter 7**, I provide a summary of the thesis along with a critical overview of the state of the field of integrative modelling and discuss future directions which I believe to be of high importance.

Chapter 1

Integrative modelling of biomolecular complexes

P. I. Koukos, A. M. J. J. Bonvin

In press, 2019, Journal of Molecular Biology

Abstract

In recent years the use of integrative, information-driven computational approaches for modelling the structure of biomolecules has been increasing in popularity. These are now recognised as a crucial complement to experimental structural biology techniques such as X-ray crystallography, Nuclear Magnetic Resonance (NMR) spectroscopy and cryo-electron microscopy (cryo-EM). This trend can be credited to a few reasons such as the increased prominence of structures solved by cryo-EM, the improvements in proteomics approaches such as Crosslinking Mass Spectrometry (XL-MS), the drive to study systems of higher complexity in their native state and the maturation of many computational techniques combined with the wide-spread availability of information-driven integrative modelling platforms.

In this review we highlight recent works that exemplify how the use of integrative and/or information-driven approaches and platforms can produce highly accurate structural models. These examples include systems which present many challenges when studied with traditional structural biology techniques such as flexible and dynamic macromolecular assemblies and membrane associated complexes.

We also identify some key areas of interest for information-driven, integrative modelling and discuss how they relate to ongoing challenges in the fields of computational structural biology. These include the use of coarse-grained forcefields for biomolecular simulations – allowing for simulations across longer (time-) and bigger (size-dimension) scales –, the use of bioinformatics predictions to drive sampling and/or scoring in docking such as those derived from coevolution analysis, and finally the study of membrane and membrane-associated protein complexes.

Introduction

Biological macromolecules such as proteins and nucleic acids make up the majority of the machinery of life since they are responsible for performing most cellular functions. While a lot of meaningful insights about these functions can be deduced by experimental work that falls under the umbrella of functional assays, these kinds of experiments (e.g. yeast two-hybrid assays) usually fail to reveal any direct information regarding the structure of the biomolecules involved in a given process. True understanding of the mechanism of action that underlies any cellular function can however only be gained by resolving at atomic detail the molecular structures of the components and assemblies involved, allowing us a glimpse at the molecular mechanisms at play [1].

Historically, the two main techniques used for experimental structure determination of biomolecules have been X-ray crystallography and Nuclear Magnetic Resonance (NMR) spectroscopy [2]. More recently, cryo-Electron Microscopy (cryo-EM) has been added to the arsenal of structural biologists and has now overtaken NMR as the second most popular technique for obtaining molecular structures with 846 vs 395 models deposited in the PDB during 2018 for cryo-EM and NMR, respectively, although both are still lagging far behind X-ray crystallography (>10000 models deposited in 2018).

All three techniques have unique advantages and disadvantages that make them suitable for specific applications, with X-ray crystallography still being the method of choice for systems which do not contain flexible or disordered regions. On the opposite end, NMR can still capture valuable information about flexible systems as well as characterize dynamics under conditions that can be considered native-like. Solution-state NMR has, however, size limitations which only make it applicable for rather small systems when it comes to solving 3D structures. It does however allow to answer specific questions, in particular related to the dynamics of large systems like nucleosomes [3–7], proteasomes [8–12], mRNA signalling machinery [13–16] as well as systems with high clinical significance such as kinase and chaperone complexes [17–19], for which NMR has a long and well-documented history of serving as the primary data source driving the simulations [20]. While solid-state NMR [21,22] does not suffer from size limitations, it still has difficulty in yielding atomic resolution quality spatial information, especially 3D structures, despite recent methodological advancements in specific fields [23–25]. Cryo-EM is increasingly becoming one of the most popular ways of determining the structure of biomolecules and most importantly large complexes and macromolecular assemblies. However, it cannot yet routinely produce structural models of atomic resolution, the level of detail which is required to understand molecular mechanisms in depth, as can be seen from the recent statistics of the Electron Microscopy Data Resource (EMDataResource - https://www.emdataresource.org/statistics.html) and those of the European Bioinformatics Institute (EBI - https://www.ebi.ac.uk/pdbe/emdb/statistics_num_res.html/). The field is still undergoing rapid transformations reflecting its nascent state, with the absence of well-defined standard practices and ongoing instrumentation and software optimisations being highlighted as potential points of improvement that should lead to higher quality structures being made available through cryo-EM in the coming years [26]. The Electron Microscopy Data Bank (EMDB) [27] has recently sponsored two blind challenges whose stated goals were to emphasise the need for map and model validation standards and engage with the cryo-EM community towards the shared development of assessment benchmarks and best practices [28].

A careful reading of the relative strengths and weaknesses of the three techniques mentioned in the previous paragraph reveals an ideal use case for computational structural modelling which relies on the use of high-quality structural models solved with X-ray crystallography or NMR spectroscopy, for determining the finer structural details of interacting biomolecules, combined with the use of cryo-EM density maps for determining the overall topology and stoichiometry of the wider context of the complex. Indeed, we believe that the revolution cryo-EM ushered in the field of structural biology a few years ago is only going to lead to an increased demand for computational techniques that, not only can make use of the data that are being made available through cryo-EM studies, but also combine those with other types of data available through other techniques, in order to generate structural models that would normally be beyond the reach of any of those techniques taken on their own.

An additional reason necessitating the use of integrative approaches is the need to study biological systems in their proper context, not only as single-structures but as macromolecular assemblies and high-order complexes as this is seen as a stepping-stone towards realising a structural model of the cell in atomic or near atomic detail [29,30]. A complicating factor is that, in order to achieve this goal, experimental data measured under as close as possible physiological conditions needs to be captured, once again requiring robust integrative modelling frameworks and protocols to unify the various sources of experimental information in cohesive structural models.

In addition to the three aforementioned structure determination techniques a plethora of complimentary techniques is available that can provide some pieces of the puzzle for the biological systems under study. Prime examples are Cross Linking Mass Spectrometry (XL-MS) and Small Angle X-ray Scattering (SAXS). XL-MS can be used to determine distances between specific residues of biomolecular complexes that can then be used in modelling since they allow for an upper distance bound between the residues they are targeting. Variations of the technique also enable the study of dynamics of complex populations in native-like conditions or even within living cells. SAXS on the other hand, is the solution equivalent of X-ray crystallography and can provide low-resolution shape information about complexes in solution and similarly to XL-MS, can also yield information regarding dynamic populations. Both of these techniques, along with the previously mentioned ones, will be discussed in detail in the next section.

Next to experimental methods, computational and bioinformatics approaches such as coevolution analyses can be used to identify residues at protein interfaces which evolve in tandem, allowing to use those residue pairs in integrative modelling directly in the simulation or during the scoring stage. Additional developments, such as the availability of coarse-grained force fields, allow for simulations across longer (time) and bigger (size) scales, enabling multiscale studies from the quantum to various levels of coarse-grained representations. These pave the way for continuous and mesoscale studies of biomolecular systems. Some docking codes now also support modelling of membrane complexes with specially adapted implicit potentials. All these developments mean that systems of increasing complexity and high relevance can now be studied within reasonable computational costs.

Integrative and Information-driven modelling

We have so far mentioned integrative and information-driven modelling without explicitly defining what constitutes such a modelling approach and distinguishes it from de novo or first-principles modelling. The main emphasis of this mini-review is on the use of biomolecular docking for modelling the 3D structure of biomolecular protein-protein complexes with a special focus at the end on membrane protein complexes.

Molecular docking refers to a set of techniques that allows us to predict the 3D structure of a biomolecular complex via simulation when starting from the 3D structures of its unbound (free) components [31]. Unlike de-novo modelling, information-driven modelling centres on the concept of using experimentally determined (or predicted) data to guide the modelling process in the hope of sampling or selecting only the meaningful part of the conformational, interaction landscape of the complex. It thus bypasses the need to exhaustively sample the vast conformational space, which would cover a 6D space for a binary complex consisting of rigid molecules. Its complexity will, however, greatly increase when considering flexibility and/or modelling a larger number of subunits. Integrative modelling refers thus to the use of some docking protocol that combines multiple sources of information (e.g. cryo-EM density map and XL-MS derived distance restraints) in order to generate a 3D model of the assembly under study [32,33].

Docking has existed as a standalone field for close to 40 years [34,35] and is one of the main two computational methods which allows us to study the 3D structure of interacting biomolecules, the other being atomistic binding simulations based, for example, on Molecular Dynamics (MD) simulations [36,37]. Docking has seen a wide range of applications from structure-based drug design [38] to protein-protein interaction studies [39,40] and network biology [41,42]. Unlike atomistic simulations, the computational requirements of docking can be met quite easily [36], which allows us to generate models at a fraction of the time of what would be required by MD. Similar to MD and other biomolecular simulation approaches though, two factors govern its performance: Sampling and scoring. Simply put, sampling refers to the process that is used to generate the binding poses from the unbound conformations. Scoring, on the other hand, is the process which allows us to discriminate between good – or native-like – and bad – or non-native-like – models. In the context of integrative modelling, the information at hand can be used to guide the simulation towards specific conformations, thus affecting the way the sampling is performed, as a filter to select or discard models based on their agreement with the experimental data, thus affecting the way the scoring is performed, or both.

A detailed overview of the challenges of the various types of docking depending on the nature of the interacting biomolecules such as protein-protein [1,43], protein-nucleic acid [44–46], protein-small molecule [47,48] and protein-peptide [49] is beyond the scope of this mini review as are the intricate details of the algorithms used by various docking programs to achieve good sampling and scoring performance. The latter is something that has been continuously evaluated over a period spanning almost 20 years in CAPRI (Critical Assessment of Protein Interactions) [50] – the blind docking experiment [51–55]. Recent CAPRI evaluations clearly demonstrate that the best strategy to model complexes is to follow a template-based approach when homologous complexes or interfaces can be identified from the PDB database [56,57]. Among the various docking software, several are supporting the use of data directly during sampling, like HADDOCK (High Ambiguity Driven DOCKing) [58,59], the pioneer of information-driven docking, together with other widely used codes like ATTRACT [60–63], Hex [64,65], IMP [66,67], LightDock [68,69] and ROSETTA [70,71]. In general, most docking codes under active development have added support for the use of information either to drive the simulation or, more commonly, as a way to filter the generated models [1].

The next section is going to focus on the various types of experimental information that can be used by integrative modelling frameworks.

Information sources for integrative modelling

Of course, integrative modelling entirely depends on the availability of data to drive the simulation. In this section we will expand on some of the most widely used types of information that can be used in an integrative capacity starting from simple experimental setups which do not require extensive expertise or instrumentation before proceeding to more complicated ones. The most commonly used approaches are those that yield residue level information. This kind of data can be obtained from mutagenesis, crosslinking – providing upper limits to the distance between the crosslinked residues, hydrogen-deuterium exchange and NMR spectroscopy experiments. The next set of techniques can yield anything from low-resolution information to high-resolution structural models of macromolecular assemblies; Cryo-EM and Small angle X-ray Scattering belong in this category. Finally, computational techniques such as multiple sequence alignments, coevolution analysis and metagenomics sequencing can yield high-quality information and interfacial and interacting residues.

Mutagenesis

Mutagenesis experiments [72–75] rest on the hypothesis that mutation of residues that are functionally important for complex formation will prevent the biomolecules (proteins specifically in this case) from interacting with each other and thus the complex from being formed (Fig 1, panel A.1). It has been used to map the interfaces or binding sites of interacting proteins [76–78]. The benefits of mutagenesis experiments are the relative ease with which the experiment can be performed, with a large variety of detection methods possible, and the fact that it provides residue level information which constitutes high-quality data that can significantly aid the modelling process compared to assays which can only provide qualitative information with regards to whether two biomolecules are interacting or not. The main downside is that, due to the indirect nature of the experiment, it needs to be combined with functional and folding assays to ensure that lack of complex formation is a result of the mutation that was introduced and not of the incorrect folding of one of the partners. Another complicating factor are allosteric effects, which can be very challenging to detect. Recent improvements to existing high-throughput mutagenesis pipelines which minimise experimental errors should enable the rapid creation of mutant libraries, which, in turn, will allow quick screening of hundreds of mutations [79,80]. Despite these advancements and the benefits conveyed in targeted mutagenesis as demonstrated for example in the CRISPR/CAS9 system [81,82], we do not expect mutagenesis data to become a dominant part of integrative modelling protocols. It will however remain a valuable source of information, even more these days where next generation sequencing has boosted the amount of genomic information available, including the identification of disease-related mutations.

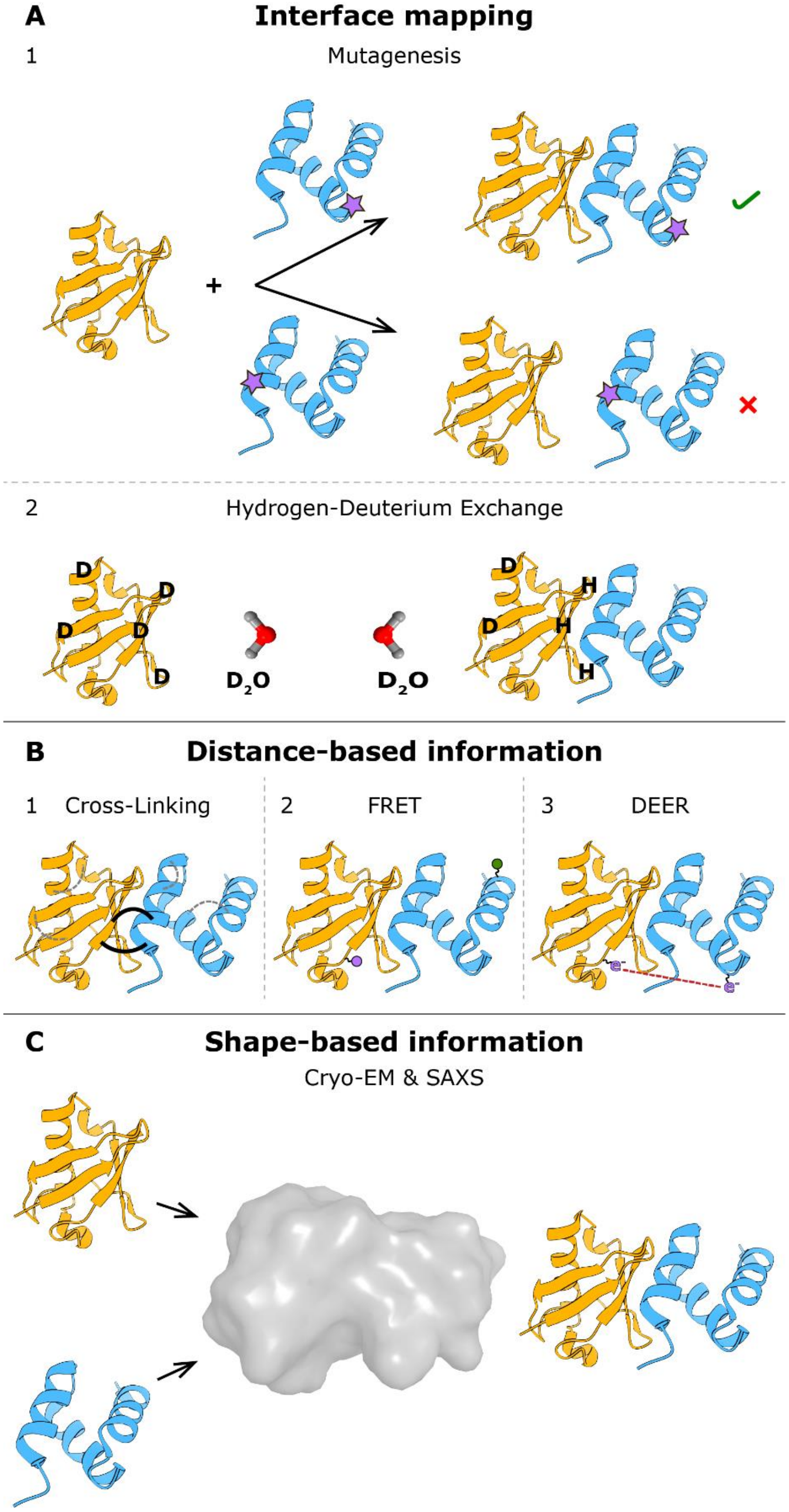


Fig 1*: Schematic representation of some of the experimental methods which can be used in integrative modelling.* ***Panel A*** *shows methods which can be used to map interfaces of interacting biomolecules.* ***Panel A.1*** *shows the experimental setup for a mutagenesis experiment combined with a binding assay. Mutations of residues which lie at the interface of the two proteins prevent complex formation. Mutated sites are shown as purple stars.* ***Panel A.2*** *shows the experimental setup for an HDX experiment during which the exchange rates for both the free forms of the proteins and their complex are compared in order to detect the regions which are occluded at the interface of the complex due to slower exchanging protons.* ***Panel B*** *shows methods which can be used to calculate residue-based distances.* ***Panel B.1*** *shows a complex which has undergone crosslinking with the intermolecular crosslinks shown as continuous black lines and the intramolecular ones as dotted grey lines.* ***Panel B.2*** *shows a complex to which fluorophore dyes have been attached at specific residues allowing us to calculate the distance between the target residues with FRET. The donor and acceptor fluorophores are shown as purple and green circles respectively.* ***Panel B.3*** *shows a complex to which spin labels have been attached enabling calculation of intermolecular distances with DEER spectroscopy. The labels are as shown as purple radicals.* ***Panel C*** *shows shape-based methods. The free structures of the complex are combined with shape-based information about the complex structure which can be derived from cryo-EM densities or SAXS shapes. The surface representation of the complex was generated with PyMOL* [83]*. All molecular graphics structures were created with ChimeraX* [84]*. The complex shown is the Ubiquituin-UBA domain from Cbl-b ubiquitin ligase (PDB entry 2oob). Ubiquitin is coloured orange and UBA light blue.*

Hydrogen-Deuterium Exchange

Hydrogen-Deuterium Exchange (HDX) is based on the principle of constant exchange of protons between biomolecules and water in which they are dissolved. In a typical HDX experiment (Fig. 1, panel A.2) the solvent (H_2O) is exchanged for D_2O, which means that the exchangeable protons of the protein (e.g. the backbone amide protons) will – at some point – exchange their proton for deuterium. The rate at which this exchange event takes place is determined by the stability of the hydrogen bond network the proton is part of, its solvent accessibility and the chemical characteristics of the residue it belongs to as well as those in its immediate vicinity [85]. Time-resolved measurements of these exchange events allow us to calculate the so-called protection factors for every exchanging amide [86], which in turn can be used to map protein-protein interfaces. These protection factors need to be determined separately for the unbound monomers and the complex. Due to the properties of deuterium, detection can be performed with either NMR or – much more commonly – Mass Spectrometry (MS). An important point regarding the mapped regions when MS is the detection method is that, due to the nature of the technique, it is not individual residues that are identified but peptide fragments whose length usually ranges between 5 to 10 residues complicating somewhat the way they can be used during computational modelling. Benefiting from methodological improvements in Liquid Chromatography (LC) and MS, along with major progress in analysis software, HDX is increasing in popularity, also aided by the relatively simple experimental setup it requires [87]. The nature of HDX experiments means these can be applied broadly for the study of any biomolecule with exchangeable protons. While HDX data are not used as often in integrative modelling of complexes we expect that situation to change in the coming years. Of particular note is the fact that the HDX community, in anticipation of this increased interest in the field, has taken steps to codify practices ranging from sample preparation, measurement and data analysis to publication and dissemination of data in standardised formats [87]. These efforts are of particular importance as they allow easier integration in modelling paradigms which combine multiple experimental sources of information [88].

Chemical cross-linking

Chemical cross linking, most often combined with Mass Spectrometry for detection purposes – Cross Linking Mass Spectrometry (XL-MS) – refers to the chemical linking of residues (most often lysine or cysteine) which are located on the surface of proteins using compounds which consist of two reactive heads and a (flexible) spacer of known maximal length [89,90]. After the crosslinking reagent has been added to the protein/cell sample and the sample has been washed, it is subjected to trypsination (or treatment with another protease) and the peptide fragments are detected via Mass Spectrometry [91] (Fig. 2, panel B.1). The benefit of XL-MS when compared to mutagenesis or HDX experiments is that the residue information is not ambiguous as it always comes in pairs (unless there are two lysines within the detected peptide fragment) and, in addition to the residues themselves, it provides information regarding their distance as the maximal spacer distance is known *a priori*. Recent improvements in crosslinking protocols and reagents along with widespread availability of proteomics facilities as well as the high-throughput nature of modern MS should also be counted among the benefits of XL-MS. All these allow for rapid and semi-automated retrieval of the distance profiles after sample preparation is complete [92]. Additional advancements have been made in spectral analysis and database search algorithms [93–95] generating high quality, quantitative XL-MS data which can be used to monitor the structure and dynamics of macromolecules and their assemblies in solution [96–98]. The wide variety of residues and chemical types which can be targeted for crosslinking ensures the wide applicability of the technique. Additionally, the combination of spacers of different lengths can provide distance information across varying scales, allowing us to capture data about both short and longer distances all of which can be used during the modelling process. Impressively, these experiments can be performed in intact cells or intact cell compartments, allowing us to extract information about the native state of the system under study [99–101]. Two major challenges of using crosslink data in docking are the fact that the crosslinks captured might reflect multiple conformational states of the complex or assembly under study, the inherent difficulty of distinguishing between intra- and inter-molecular crosslinks – or crosslinks between two residues of the same protein and residues of different proteins when dealing with symmetrical systems, and the high reactivity of the reagents which can capture non-native, encounter complexes. Despite these, we expect XL-MS to further develop in the coming years and XL-MS derived distance restraints to become increasingly prevalent in integrative modelling as integrative modelling software also develops ways of dealing with these shortcomings in a consistent way. Similar to the HDX community, multiple leading MS groups have decided to establish community guidelines regarding best practises in sample preparation, measurement, data analysis, model validation and result reporting [102–105]. Development of software which can automatically group crosslinks into the conformational states to which they correspond and identify potential false positives, would be a valuable addition. This would allow docking software which can make use of distance restraints to determine different structural models for the different states more closely reflecting the behaviour of the system under study in solution. The DisVis standalone program and webserver [106,107] can already perform the task of identifying potential false positives through enumeration of violations of distance restraints after exhaustive rotational and translational sampling and developments to allow clustering of distinct distances into conformational states are already under way.

NMR spectroscopy

Until fairly recently, NMR spectroscopy was one of two ways (the other way of course being X-ray crystallography) of obtaining high resolution structures of biomolecules. Its application to complexes was limited as solution-state NMR cannot routinely deal with complexes whose size is greater than 40kDa [108]. The most easily obtainable measurements for macromolecular complexes are Chemical Shift Perturbations (CSPs) which allows for the mapping of the interacting regions of biomolecules [109] at the residue/atomic level. In addition to CSPs additional restraints can be recorded from NMR experiments such as for example intermolecular NOEs, Residual Dipolar Couplings (RDCs) and relaxation anisotropy [110,111] which reveal details regarding the relative orientation of two interacting biomolecules. Paramagnetic probes [112] can be used to provide additional information to standard NMR experiments in the form of long-distance atomic information [113], to probe interacting surfaces of biomolecules [114,115] and study dynamics [116]. Unlike solution-NMR, solid-state NMR (ssNMR) [117,118] has no theoretical limitations on the size of the systems which can be studied, however, obtaining atomic-resolution structures can be complicated due to the spectral complexity [119,120]. More recently, ssNMR has been applied successfully for the study of transmembrane systems ranging from the Kcsa ion channel [25,121] to small peptide-based antibiotics which interfere with the lipid-II cycle [24]. Of course, the application of NMR – whether in solid- or solution state – to small and flexible systems such as peptides is not new even when considering membrane embedded peptides [122,123]. One of the recent developments in the field of NMR has been the ability to study molecules in cells allowing for qualitative comparisons between native and non-native species or analysis of conformational heterogeneity across different cell types [124–127]. A limiting factor of NMR is the often-costly procedure of preparing samples for NMR as well as the relative difficulty in analysing and interpreting the experimental measurements. In light of these observations, we expect NMR to continue to factor significantly in integrative modelling over the coming years mainly owing to the undeniable benefit and unique ability of NMR of being able to study dynamics at atomic resolution in real time across different time scales. This is particularly attractive when compared to techniques like XL-MS which can only estimate dynamics as a result of conformational heterogeneity observed in the distance profiles.

Cryo-EM

The techniques which fall under the umbrella of cryo-EM have been revolutionising structural and integrative biology for a few years now. This is for the most part due to advancements in detector technology, automation and software [128,129]. Cryo-EM derived structures can now match or even surpass structures of the same system obtained with X-ray crystallography [130]. This is also reflected in the number of near atomic-resolution structures deposited in the Electron Microscopy Data Bank (EMDB) [131] with a third of the structures deposited in 2018 having a resolution of 4Å or better (https://www.emdataresource.org) – a trend which further improved in 2019. Single particle analysis (SPA) constitutes the overwhelming majority of cryo-EM experiments undertaken in recent years [130]. The typical cryo-EM SPA experiment constitutes of loading an aqueous solution containing the biological sample on a grid mesh, blotting to remove excess solution and to form a thin layer and covering it with a thin carbon film after which the sample-loaded grid is plunge frozen. The particles are then imaged with an electron beam using sufficiently low doses to prevent radiation damage. Many 2D images are collected, aligned and then used to computationally reconstruct the molecule in 3D [132,133]. When the resolution is not sufficient to determine the molecular structure at atomic or near atomic resolution, various rigid or flexible fitting protocols can be used to fit existing structural models of the components of the assembly into the EM-derived map [134–137]. In the time period preceding the resolution revolution these inherently integrative protocols were the most common way of generating structural models with cryo-EM (Fig. 1, panel C). More recently, popular codes such as ATTRACT, IMP (which supported cryo-EM data from day 1), HADDOCK and ROSETTA have added support for cryo-EM derived density maps during the modelling process [61,67,107,138–141].

However, the ever-increasing performance gains in terms of resolution for structures solved with cryo-EM pose some interesting questions for the field of integrative modelling, specifically is there a place for integrative approaches in an era where atomic resolution models for a wide variety of systems and molecular weights can routinely be obtained with cryo-EM data alone? We believe the answer is yes, for multiple reasons. First and most importantly methodological limitations make it difficult to study small and/or flexible systems with cryo-EM. Some recently solved structures show however promising results in that direction [142,143] even potentially allowing us to study interactions between drugs and proteins [144]. Secondly, even in the cases where a high-resolution model has been obtained the resolution distribution might not be uniform with some parts of the molecule having lower resolution than others. This non-uniformity can arise as a result of structural heterogeneity, non-isotropic distribution of sampled orientations or even processing artefacts. Some groups have already suggested alternative ways of measuring resolution that take into account significant local variations from the reported mean value [145,146]. Sample structural heterogeneity is usually considered a limiting factor for cryo-EM as it makes the averaging of aligned images more difficult and results in lower-resolution models. Whereas most cryo-EM sample preparation protocols emphasise the importance of structural homogeneity in order to be able to generate high resolution models, some recently described approaches embrace the importance of structural heterogeneity as an inherent property of dynamic systems such as biological samples, allowing for identification of distinct conformational states [147–149]. It is expected that identification of these distinct states will allow to simultaneously estimate the conformational landscape and thermodynamic behaviour of the system. Such results would be very desirable when attempting to describe the intermediate states of a cellular process or when studying systems for which high structural variability is expected [150].

We conclude that despite the impressive advances made recently in the field of cryo-EM, we expect the importance of integrative approaches in the context of cryo-EM to increase. Integrative modelling might be used either as a way to validate the structural models, as a way to aid the modelling process for systems which are difficult to study with cryo-EM alone, or to model parts of the cryo-EM maps that might not reach sufficient resolution for de novo structure determination. The modelling of such systems from cryo-EM data can significantly benefit from the inclusion of additional data (e.g. from XL-MS or NMR [151,152]). The importance of integrative approaches can also be seen by recent studies which favourably compare integrative models with high-resolution structures of the same complexes made available years later by cryo-EM [33]. In light of these observations, we expect cryo-EM to play a prominent – if not dominant – role in many aspects of integrative modelling in the forthcoming years.

Small Angle X-ray Scattering

Biological Small Angle X-ray Scattering (SAXS) is the solution equivalent to X-ray crystallography. It is another field which has been undergoing a renaissance in recent years with more improvements expected in the next few years [153]. In a basic SAXS experiment a macromolecular solution is bombarded with X-ray beams and the scattering pattern is recorded by a detector placed in close proximity to the sample. The most basic information that can be extracted from the measurement is the scattering curve which is extracted from the distance profile between all sample atoms which can in turn be used to construct a low-resolution shape or envelope of the system under study [154]. The potential for SAXS data to be useful in integrative studies was realised early [155] with protocols resembling those that are used for fitting X-ray- or NMR-derived structural models into medium- to low-resolution EM density maps where the unbound structures of the components of the complex were docked against each other and the shape of the resulting complex was scored against the SAXS-calculated shape [156,157] or directly against the scattering curve [153,158–164]. More recently, protocols that can make use of the shape information to guide the docking towards conformations that agree with the SAXS shape have been described [165] (Fig. 1, panel C). The maturity of SAXS protocols, the standardisation of guidelines for publishing SAXS data, the relative ease with which samples can be prepared, the automated manner of data acquisition and analysis as well as the high-throughput nature of BIO SAXS are some of the factors which make SAXS a very attractive option for probing macromolecular interactions under solution conditions without a size limitation, but sample purity and homogeneity are important aspects in order to be able to derive reliable structural data [166]. In addition to calculating low-resolution shapes of macromolecular complexes, SAXS can be used to qualitatively and quantitatively compare samples, probe conformational differences, assembly states, folding status and in some cases even refine flexible, low resolution regions of structures determined with X-ray crystallography [154]. All these factors combine to paint a very favourable picture of SAXS in its current and future states. The ability to probe dynamics in solution without size limitations while at the same time deriving shape-based restraints which can either be used to restrain the sampling of docking simulations or filter out non-native-like solutions when scoring generated models are counted among its greatest strengths. We only expect the contribution of SAXS in the field of integrative modelling to further proliferate.

Other experimental sources of information

In addition to the experimental methods that have already been mentioned, structural biologists have access to a plethora of other methods giving various levels of experimental information about the interacting biomolecules. Covering all of these techniques as well as the ways in which the data that can be derived from them for use in integrative modelling is beyond the scope of this review. We will mention though two techniques standing out due to their high importance for the field of modelling, both of which can provide distance information between residues of the interacting biomolecules. The first is Förster Energy Resonance Transfer (FRET). It allows to detect the energy transfer between donor and acceptor fluorophores allowing for the calculation of long-range distances between those parts [167] (Fig. 1, panel B.2). It does however require covalent attachment of dyes to specific parts of the molecules. FRET data have been used successfully in integrative modelling efforts either alone or in combination with data from other sources to determine the structure of biomolecular complexes and study their dynamics [168–172]. Similarly to FRET, the Double Electron-Electron Resonance technique (DEER) is a spectroscopic approach which enables the calculation of long-distances between interacting spin labels that have been attached to specific residues (most commonly cysteines) [173,174] (Fig. 1, panel B.3). It has been applied widely to study systems of varying sizes and composition including small protein-protein complexes to large molecular machines and RNA-containing complexes [175–178].

Bioinformatics and Computational Approaches

Perhaps some of the most interesting advancements in the field of integrative modelling in recent years originate from bioinformatics and computational techniques which, on their own, cannot be classified as integrative, but whose output can be combined in integrative modelling frameworks just like experimental data. In this section we are only going to provide a succinct overview of recent developments in the field, emphasising three areas: The coming of age of coevolution, the appearance of membrane-specific modelling tools and the use of coarse graining approaches.

Coevolution

Coevolution rests on the observation that sometimes mutations at specific positions in a protein sequence correlate with mutations at other positions of the same or interacting proteins, the hypothesis being that if such residues “coevolve”, they might be in spatial proximity. When a mutation is introduced in one of the interacting pair a compensatory mutation arises in the other due to evolutionary pressure relating to functional or conformational importance of that residue pair [179]. This information can be used in the structure determination of proteins [180] but most importantly for integrative modelling purposes. The concept can be quite easily extended to protein residues which belong to different proteins forming a complex or being part of a larger molecular assembly [181]. Methods such as EVcomplex, GREMLIN [182,183] and InterEvDock [184] have been applied successfully in docking simulations [185–187]. Of particular note is the recent development of InterEvDock2, a free and fully automated webserver which allows the user to input sequences instead of structures, submit multimeric next to monomeric components and automatically derive coevolution-based restraints to use for scoring the models generated during the simulation [188]. The utility of coevolution-based data does not stop with protein folding and determination of soluble protein-protein interfaces though. More recently, it has been used to determine transmembrane protein interaction sites [189,190], identify new protein-protein interaction networks [191] and novel protein contact maps making use metagenomics data [192]. The robust state of the coevolution community in combination with the intuitive nature of the output data makes us confident that the use of coevolution-derived spatial restraints is only going to become more prevalent in the near future. One potential limiting factor for the use of coevolution-based restraints for docking is the need for extensive and diverse sequencing data in order to get deep enough alignments, although deep learning methods are becoming more robust with respect to the alignment depth [193]. This limitation can, for example, limit their applicability for the study of mammalian systems, for which sequencing data are not as exhaustive compared to bacteria and yeast.

Membrane modelling

Another field which has attracted attention recently and has seen many developments is membrane protein modelling. It is traditionally considered as one of the most difficult kinds of systems to study with experimental structural biology methods due to the nature of the lipid bilayer which requires that, either it is dissolved with detergents and reconstituted, or that native or native-like membrane mimetics are used. The former is easier and has been used with success for X-ray and NMR but raises questions about the effect the detergent has on the 3D structure. The use of native or native-like membrane mimetics is much closer to physiological conditions, which means that any structure determined this way should be closer to its counterpart in the cellular environment, but this introduces many challenges in sample preparation and measurements. A relatively recent advancement enabling studies of membrane proteins and their complexes in native-like and even native environments is the advent of

Styrene-Maleic Acid (SMA) copolymers which can be used in combination with synthetic liposomes or native membranes to solubilise patches of protein-containing lipid bilayers without the adverse effects of detergents [194]. These, so called, SMALPs (SMA-lipid particles) have already been used together with MS to determine the stoichiometry [195], acquire atomic resolution structures of membrane protein complexes using X-ray crystallography and cryo-EM [196,197] as well as study their dynamics with solid-state NMR [198,199]. Computational methods therefore remain an attractive alternative for the study of membrane bound or membrane associated proteins and their complexes [200]. The simplest – yet most effective – way in which membrane protein modelling has been made easier in the recent years comes from a higher availability of representative 3D structures in the PDB thanks, in no small part, to advances in cryo-EM [201]. This, in combination with the availability of membrane specific homology modelling tools like MEDELLER [202] and Memoir [203], which implement protocols similar to and inspired by one of the most popular homology modelling tools – MODELLER [204] – enables the creation of structural models that strongly approximate native ones [205]. These can be used to confidently model structures which have not yet been determined experimentally. These models can act as the starting point for further investigation, which usually involves some degree of integrative modelling, for example rigid/flexible fitting in low-medium resolution cryo-EM maps or embedding into membranes and studying the system by MD. This wider availability of transmembrane (TM) protein structures is also reflected in the enrichment of entries in databases that deal with membrane proteins exclusively [201], such as the manually curated *mpstruc* (membrane proteins of known 3D structure - https://blanco.biomol.uci.edu/mpstruc/), which annotates all non-redundant proteins in the PDB. The latter also serves as the starting point for the classification system the PDB uses for identifying entries as TM, for OPM (orientations of proteins in membranes - https://opm.phar.umich.edu/) [206], which computes the membrane insertion angle, tilt and width for transmembrane and membrane associated proteins, and for MemProtMD (http://memprotmd.bioch.ox.ac.uk/) which inserts proteins in lipid bilayers via self-assembly with coarse-grained MD simulations and also makes available the pre-equilibrated membrane bilayer-protein structures [207,208]. The plethora of G-protein coupled receptor (GPCR) structures which have been solved recently is of major importance not only to the structural biology community but to areas of pharmaceutical research as well owing to the importance of GPCRs in many diseases [209]. These structures, along with important details regarding the method that was used to determine them, the conditions under which the experiments were performed and various aggregated statistics and analyses are collected in GPCRdb (GPCR database -

https://gpcrdb.org/) [210]. Coarse-grained MD forcefields such as MARTINI [211–214], which was originally developed for membrane , have also been extended to include proteins and nucleic acids. These allow to simulate larger systems and/or reach longer time scales. The MARTINI force field has recently been implemented into HADDOCK for the modelling of proteins and nucleic acids complexes [215,216]. Other docking codes such as ATTRACT [61], CABS-dock [217,218], the Integrative Modeling Platform (IMP) [66,67] and PyRy3D (http://pyry3d.icm.edu.pl/) also support coarse-graining. Despite these significant advances the only docking codes which currently offer the ability to dock TM proteins with specific implicit membrane potentials are (to the best of our knowledge) ROSETTA [219,220], DOCK/PIERR [221,222] and Memdock [223]. More recently a generic, ready-to-dock benchmark of membrane protein complexes accompanied by docking decoys for the purpose of training membrane-specific scoring functions was made available [224,225]. The lack of widely available explicit support for the docking of membrane proteins has resulted in some creative integrative modelling with, for example, researchers using HADDOCK to probe the interaction between the K-RAS4B oncogenic protein when complexed with the Cmpd2 inhibitor and lipid nanodisks making use of NMR-derived restraints to drive the simulation [226]. In summary, we believe it’s high time the field of membrane complexes modelling is given the attention it deserves by the docking community as all the ingredients for successful integrative modelling are in place, with experimental methods providing good template structures for modelling as well as experimental restraints, computational tools like coevolution providing additional data to drive the docking and plentiful implicit or explicit implementations of membrane bilayers allowing for studies at different representation levels.

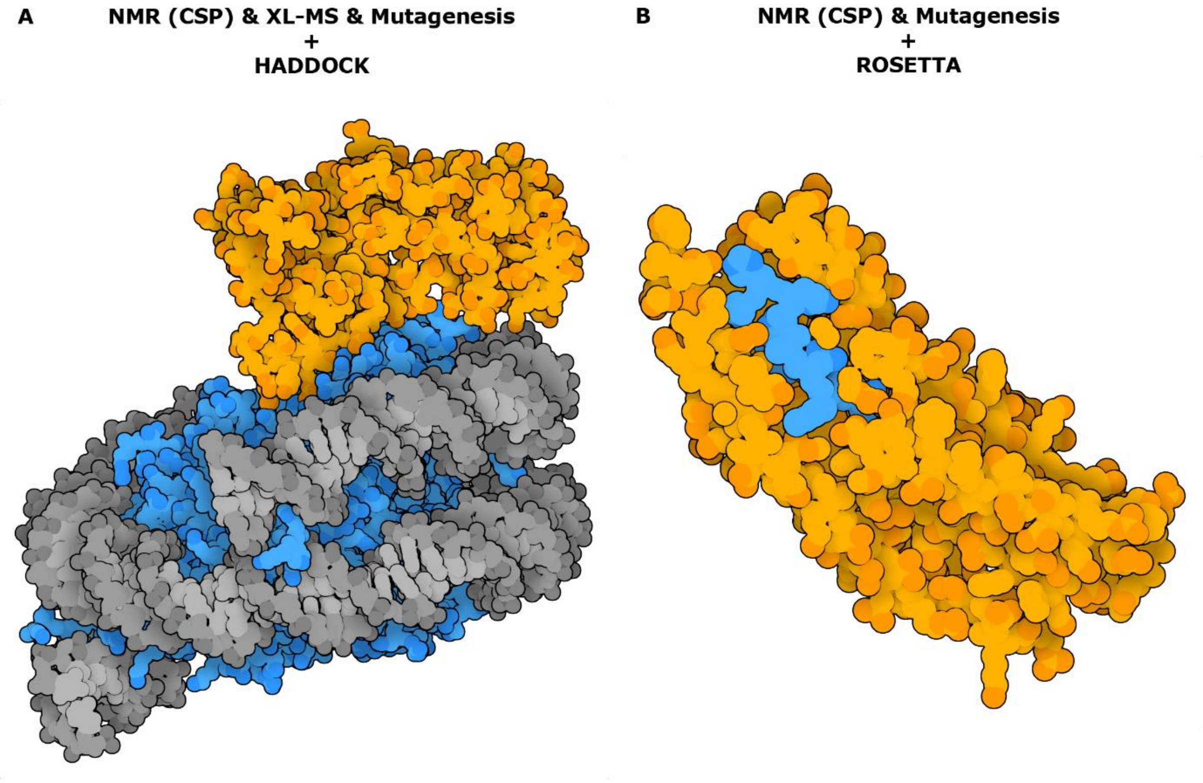


***Fig 2**: Structural models determined with integrative approaches. Panel A shows a rendering of the nucleosome complex bound to UbcH5c and RNF168-RING domain. The model was determined with HADDOCK using NMR-derived spatial restraints (CSPs) in combination with mutagenesis and XL-MS data (PDB-dev entry 29). The DNA is shown in grey, the histones in blue and the UbcH5c and RNF168-RING domain in orange. Panel B shows a rendering of the ghrelin peptide bound to its G-protein coupled receptor (GHSR) (PDB-dev entry 24). The model was determined with ROSETTA using NMR-derived spatial restraints (CSPs) and mutagenesis data. GHSR is coloured orange with the peptide in light blue with some clipping to enable visualisation of the binding pocket. Both models are illustrated with the program `illustrate` by David Goodsell. Only the top model from each submission is shown.*

Perspectives

We have highlighted some key areas of experimental and computational structural biology and identified the ones which, we believe, will factor significantly in the coming years for the field of integrative modelling in general and molecular docking specifically. Despite these advances, there are however also some areas for which we believe developments have been lacking. Chief among these is the fact that many docking codes can still not make use information during the simulation, instead only in the scoring stage, and therefore cannot be considered integrative approaches, with some exceptions existing, e.g. ATTRACT, HADDOCK, IMP, PyRy3D and RosettaDock to cite the most known ones. Another limiting factor is the fact that the number of distinct subunits which can be included in the simulation is still limited, with most codes, except a few, supporting only one receptor and one ligand [1,188,227],i.e. binary complexes.

Another aspect of integrative modelling is that being able to combine multiple sources of information into a single docking run does not necessarily mean that the resulting models benefit from the included information. The reason for this is that information needs to be combined in a probabilistically sound way, that is in a way that reflects the uncertainty of the original measurements and properly propagates it [66,228]. Perhaps the most well-known example of software which properly accounts for this and weights the multiple data sources used in the modelling through a Bayesian framework are IMP [66,67] and the Inferential Structure Determination Software (ISD) [229]. IMP has most famously been used for the determination of an integrative model for the nuclear pore complex [230], which was validated last year when the cryo-EM structure for the entire complex was solved with a final resolution of 28Å [231]. ISD [229], originally developed for NMR [232], has recently been extended and applied to challenging systems like membrane proteins, bacterial pili and chromosomes [233–235], and also large macromolecular assemblies using shape-based (SAXS or cryo-EM) data [236].

Another alternative to one-stop integrative modelling software is the combination of multiple codes in easy-to-use and cohesive workflows which hide the technical details away from the end users and allow for seamless flow of information between different packages. Some encouraging work in this direction has already started with CROSS-ID [105], a package for the analysis and visualisation of XL-MS data which is part of XlinkX [100,104] and makes uses of DisVis for the visualisation and validation of crosslink data. Another interesting initiative is the BioExcel consortium since one of their stated goals is to promote integration among several flagship computational biology/chemistry packages such as HADDOCK and GROMACS [237,238].

Finally, next to development in integrative software, proper description and archival of integrative models is an important area which is benefiting from the advent of PDB-dev [32,239] a portal developed by wwPDB in collaboration and consultation with experts in the field of integrative modelling. Its aim is to act as a hub to collect structural models, and all their associated data, that have been determined by integrative approaches. Two examples of integrative models deposited into PDB-dev obtained with various software and data types are shown in Fig. 2.

Conclusions

In this review we have discussed aspects of integrative modelling and in particular recent developments related to the various types of information that can be used to aid the modelling process. We conclude that the future for integrative modelling software is bright as the availability and quality of data is only going to increase as will the ability of algorithms and hardware to handle that data efficiently and meaningfully. There still remain, however, long-standing challenges, such as accurate binding affinity prediction and accurately modelling large conformational changes. These have been challenges in the biomolecular simulation world since the very first days of the field [1,227]. Our ability to model the structure of biomolecules as well as biomolecular complexes has been continuously evaluated over a period spanning more than 20 years in the CASP (Critical Assessment of Structure Prediction) [240] and CAPRI (Critical Assessment of PRediction of Interactions) [241] experiments, with the first focusing on single protein structure prediction (with a multimer component) and the latter on protein complexes. In recent iterations of the challenge, CASP has featured a data-assisted category for which some information about the target system is disseminated to the participating groups thus evaluating the ability of software to incorporate information in the prediction and its outcome. In CAPRI so far, only once was a SAXS scattering profile provided. The field would clearly benefit from truly integrative blind modelling challenges as such blind challenges have been, and will remain important catalysts for further development and advances.

Chapter 2

A membrane protein complex docking benchmark

Panagiotis I. Koukos, Inge Faro, Charlotte W. van Noort, Alexandre M.J.J Bonvin

Journal of Molecular Biology, 2018, volume 430, issue 24, pages 5246-5256

Abstract

We report the first membrane protein-protein docking benchmark consisting of 37 targets of diverse functions and folds. The structures were chosen based on a set of parameters such as the availability of unbound structures, the modelling difficulty and their uniqueness. They have been cleaned and consistently numbered to facilitate their use in docking. Using this benchmark, we establish the baseline performance of HADDOCK, without any specific optimization for membrane proteins, for two scenarios: True interface-driven docking and ab-initio docking. Despite the fact that HADDOCK has been developed for soluble complexes, it shows promising docking performance for membrane systems, but there is clearly room for further optimisation. The resulting set of docking decoys, together with analysis scripts are made freely available. These can serve as a basis for the optimisation of membrane complex-specific scoring functions.

Introduction

The docking community makes extensive use of benchmarks for evaluating the performance of docking algorithms and constantly improving them. Such benchmarks are also critical to allow a fair comparison between various algorithms, next to blind docking experiments such as CAPRI [50] for protein-protein and protein-peptide docking and Drug Design Data Resource grand challenges (D3R) [242,243] for small molecule docking. Some of the most cited benchmarks are the protein-protein [244], protein-peptide [245,246], protein-DNA [247] and protein-ligand [248] ones. Several recent publications have made use of membrane protein-related benchmarks for testing and validating their software. RosettaMP [219], a recently updated addition to Rosetta's toolbox, supports a general membrane representation and can be used in combination with many of Rosetta's existing sampling and scoring protocols. The MPdock protocol is a combination of the RosettaMP and RosettaDock protocols [71] and supports docking of membrane proteins. The same publication also presents the MPsymdock protocol which can be used to assemble homomeric membrane protein complexes from their monomeric constituents using known symmetry information. The authors also tested the newly-minted protocols on membrane protein complexes, however since they intended those demonstrations as a proof-of-concept they only tested on 5 and 4 complexes for the MPdock and MPsymdock protocols respectively. Other researchers have made use of more extensive datasets for their work, such as the Memdock [223] software and the modification to the scoring schemes employed by DOCK/PIERR [221,222]. In the case of the former their training and testing sets consisted of 43 and 21 complexes (for a total of 64) obtained from the OPM database [249], however all entries are helical proteins. The same is true for DOCK/PIERR as well. Their dataset makes use of the MPSTRUC database as the primary source of data and contains 22 biological complexes as well 8 artificial complexes which have been created by separating GPCR proteins into separate parts after cutting them at one of the cytosolic/extracellular loops. This dataset mostly consists of GPCRs and small helical complexes. None of the aforementioned works make the structures they used available and therefore their datasets cannot be used as a docking benchmark. Another GPCR dataset has been recently published [250]. To the best of our knowledge, however, there is no general and non-redundant docking benchmark for membrane protein-protein complexes. This is understandable since membrane proteins are notoriously difficult to characterize experimentally [251], which limits their number in the *Protein Data Bank* (PDB) [252] and also decreases the probability of obtaining both bound and unbound conformations of the structures that make up the complex. The latter is one of the requirements of any docking benchmark to allow a realistic evaluation of docking performance. We have however reached a point where enough structures of membrane proteins have been deposited in the PDB to create a docking benchmark. This allows us here to introduce a new membrane protein-protein complex benchmark, establish the baseline performance of HADDOCK in two docking scenarios and provide a decoy dataset that will allow further optimisation of scoring functions for this specific class of complexes. This new benchmark is freely available for download. The structures have been renumbered and cleaned to facilitate immediate docking and analysis. In addition to the benchmark itself we are also providing code that can be used for the analysis of docking results as well as the decoy datasets. The content of this benchmark is more diverse than any of the datasets used in previous studies since it contains both non-helical proteins and helical proteins, and complexes that are larger than GPCRs, as well as small helices.

Materials and Methods

Data sources

The primary data source for this benchmark was the *Membrane Proteins of Known Structure* (MPSTRUC) database (http://blanco.biomol.uci.edu/mpstruc/). MPSTRUC is a manually curated database of membrane proteins. Its entries are classified into three categories:

1. Monotopic membrane proteins
2. Beta-barrel transmembrane proteins
3. Alpha-helix transmembrane proteins

We disregarded the monotopic membrane protein category since it is made up of proteins which are not embedded in the lipid bilayer but instead are only anchored to one side of it. We considered all remaining unique entries and processed them using the procedure outlined in Fig. 1.

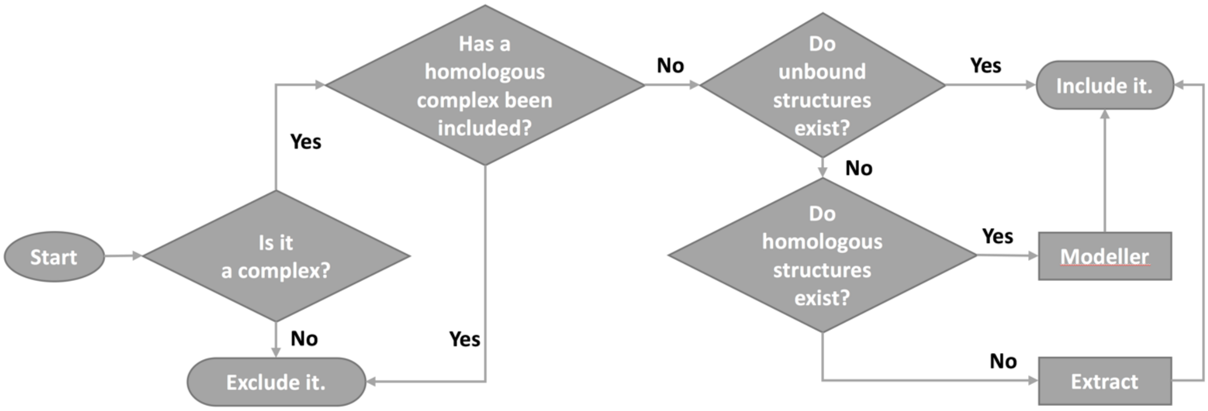

Start
Is it a complex?
Yes
No
Exclude it.
Has a homologous complex been included?
Yes
No
Do unbound structures exist?
Yes
No
Include it.
Do homologous structures exist?
Yes
Modeller
No
Extract

Fig 1: *Flowchart of the structure identification procedure.*

After identifying a complex, we searched the related structures in MPSTRUC as well as the homologous structures of that complex in the PDB to identify potential unbound structures of its components. The related MPSTRUC entries correspond to the same protein structure solved under different conditions (e.g. acidic vs basic pH), with different techniques (NMR vs X-RAY crystallography) or complexed with other biomolecules (e.g. small-molecule ligands or peptides). For the complexes where we could not identify a suitable unbound structure via MPSTRUC we turned to the PDB and made use of its precalculated sequence similarity clustering analysis results. Optimally, the structure of the complex and that of its components should have been determined independently of each other and be complete, i.e. have no missing parts or mutations close to the interface. If that is not the case, but highly homologous structures are available, those are included instead. In this case, highly homologous refers to 100% sequence identity (without gaps) of the interface region and very similar (if not identical) sequence for the remainder of the protein. In these cases, the remainder of the protein was not modelled since the overall similarity is quite high. SI table 1 lists the backbone RMSD (after optimal superimposition using backbone atoms) of all components for all entries that are not classified as "Bound" (see Table 1). The mean RMSD is 1.45 ± 0.86Å. If the homologous structures differ at the interface – due to mutations or gaps – they were modelled with modeller [204], using the *loopmodel* protocol for the cases with significant interface gaps and the *automodel* protocol for all remaining ones. We made use of homology models for 5 complexes (see Table 1) for which the sequence similarity and identity ranged between 71-96% and 56-96%, respectively. 100 models were generated and ranked according to their objective function score and the best-scoring structure that was within modelling difficulty of the complex structure was selected after visual inspection. We manually inspected the models to ensure no unnatural segments were introduced during the modelling of the gaps.

We applied additional selection criteria: we only selected heteromeric interfaces, therefore homomeric complexes that function as multi-chain proteins such as trimeric transmembrane porins (for example PDB entry 1OSM [253]), although technically transmembrane protein complexes, were not included. X-ray structures were given priority over structures determined by NMR, and higher quality structures (resolution, clashscore, R-free, Ramachandran outliers) were preferred over lower quality structures. The availability of high-quality unbound structures also influenced the inclusion of one complex over another for which no unbound structures were available or, if there were, they were of low quality (low resolution, mutations, gaps). The resulting dataset is also non-redundant in the sense that we have only included what we determined as the best complex based on the above criteria for any given protein family. In addition to the MPSTRUC classifications we also made use of a sequence identity cut-off of 30% for identifying homologous structures to ensure we only included non-redundant entries. Accordingly, no chain of any complex of the dataset has a sequence identity larger than 30% to any other. For calculating the sequence identities, we used the Needleman-Wunsch algorithm [254] with the BLOSUM62 [255] matrix, and a gap open and extend penalty of 10 and 0.5 respectively.

The entries of the benchmark have been modified to facilitate comparisons between the unbound and reference structures. The numbering and chains ids of the unbound structures have been modified to match those of the reference structures. Disordered regions were removed when near the interface or when they introduced challenging conformational rearrangements that would prevent the unbound structures from adopting a conformation close to the reference one. We have made our best efforts to include all biologically relevant ions and cofactors when they were present in both unbound and reference structures. In some cases, we have joined two or more unbound chains in a single body. Reasons for doing so are reducing the number of docking partners to two or three, since most docking codes do not support multi-body docking, the availability of unbound structures, and the topology of the complex – an example would be joining two homomeric TM subunits in a single subunit and docking that against a cytosolic partner. These cases are indicated by the presence of multiple chain ids at the end of the unbound structure id in Table 1. We consider different subunits of the complex for the four complexes (2r6g, 2zxe, 4huq, 5a63) which appear more than once (see Table 1). We only used the renamed and renumbered unbound and homology structures for docking in all cases where such structures were available. In all other cases, we used the renamed and renumbered bound structures.

Docking

The HADDOCK webserver (v2.2) (https://haddock.science.uu.nl/services/HADDOCK2.2) [59] was used for all docking runs. HADDOCK is an integrative modelling biomolecular docking platform which makes use of experimental data (mostly derived from biophysical/biochemical experiments) or bioinformatics predictions to drive the docking process. This information is typically translated into distance restraints used to drive the docking [256]. The docking consists of three stages:

i. Rigid-body energy minimisation – it0
ii. Semi-flexible refinement by simulated annealing in torsional space – it1
iii. Refinement in explicit solvent – itw

For the first stage (it0), the partners are randomly oriented and translated away from each other followed by rigid-body energy minimisation (EM). For it1, flexibility is introduced in the interface residues of the complex (defined as the set of residues whose atoms are within 5Å of any atom of any partner), first along the side-chains only and, in the final stage, including the backbone atoms as well. The last stage (itw) consists of a short molecular dynamics run in Cartesian space and explicit solvent (the docking runs were performed with the default TIP3P water model [257]).

We used two types of restraints to drive the docking: random and true interface restraints. In the case of random restraints, for each docking trial, a surface-exposed patch of residues is randomly defined on both partners of a dimeric complex and used to drive the docking by defining those patches as active residues in the HADDOCK formalism. Since this option is not supported for higher order complexes, for the three trimeric complexes in the benchmark (see Results) centre-of-mass (CM), C3 symmetry and non-crystallographic symmetry (NCS) restraints were used instead [258]. In the case of true interface restraints, we extracted the interface residues of the bound complex (at a distance cut-off of 5Å) and defined those as active in HADDOCK for the docking run.

The number of docking decoys generated was set for it0/it1/itw to 50000/400/400 and 10000/400/400, for ab-initio (random restraints) and true interface-driven docking, respectively. Additionally, since the scoring function of HADDOCK has not been optimised yet for membrane complexes, we set the number of trials in it0 to 1 and disabled the systematic sampling of 180° rotations during it0 to 1 to disable the internal scoring scheme of HADDOCK. For the cases categorised as "Buried" (see Table 1 below) we have also lowered the intermolecular energy scaling to 0.01 to allow interpenetration of chains during it0. We further kept the original scoring function of HADDOCK, defined as:

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

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

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

Where E_{vdw}, E_{elec}, E_{desolv} and E_{AIR} stand for van der Waals, electrostatic, desolvation and restraint energies. The non-bonded terms are calculated with the OPLS force field [259] with a cutoff of 8.5Å, the desolvation parameters are described in [260] and the restraint energy in [58]. BSA stands for buried surface area. It is worth noting that the desolvation potential depends on parameters that have been optimised for soluble proteins. These are thus the default scoring settings for soluble complexes.

Analysis

We report both the interface and ligand RMSD values (I/L-RMSD respectively) as used in CAPRI. For the I-RMSD, we superimpose and calculate the RMSD of the backbone atoms of the interface residues (defined at a 10Å cut-off). For the L-RMSD we superimpose on the backbone atoms of the receptor (defined as the largest of the partners) and calculate the RMSD of the backbone atoms of the ligand (defined as the smallest of the partners). For the 2 trimers (see Table 2 below), I-RMSD is calculated as described above and L-RMSD is calculated by selecting the first chain as the receptor and averaging the L-RMSD of the second and third chains. Fitting and RMSD calculations were performed using the McLachlan algorithm [261] as implemented in the program ProFit (http://www.bioinf.org.uk/software/profit/) from the SBGrid distribution [262]. All scripts used for analysis are provided together with the docking benchmark at https://github.com/haddocking/MemCplxDB.

Results and Discussion

Benchmark

Following the protocol that is outlined in the Methods section we identified 37 complexes of interest. These complexes are listed in Tables 1 and 2 (dimers and trimers respectively). An annotated version of this table, detailing the modifications that were made to the structures can be found in the SI (SI Tables 2 and 3). The tables detail the PDB id of the structure of the complex and those of the corresponding unbound entries. In the cases where all the components have been extracted from the complex, that entry is a defined as a “Bound” case. If at least one of the partners is not extracted from the complex, then that case is classified as “Unbound” and, depending on the I-RMSD of the unbound structures after optimal superposition on the reference, is classified as “Easy”, “Intermediate” or “Hard” difficulty based on I-RMSD values of less than 1 Å, between 1 and 2 Å and over 2 Å, respectively. Both trimeric entries of the benchmark (Table 2), are classified as “Unbound” because the “unbound” components originate from a different PDB entry of the same complex crystalized under different conditions (3w9h and 2qts for 2j8s and 4fz0 respectively). Those differ significantly enough in their i-RMSD (0.65 and 1.18 Å for 2j8s and 4fz0 respectively) and overall backbone RMSDs of each subunit (see Table S1) from what we define as the reference bound conformation, which justifies their inclusion in this benchmark. We have also categorized the complexes based on the nature of the interaction. Complexes whose interface is contained within the membrane are labelled “TM” for transmembrane, complexes whose interface lies between the membrane and the cytosolic/periplasmic/extracellular environment are labelled “MS” for membrane-soluble. Complexes whose interface lies in the membrane but also extends past it are labelled “Both” for both transmembrane and membrane-soluble. Complexes where one of the partners is embedded in a TM beta-barrel are labelled “Buried” and are by nature TM complexes, and complexes that involve antibodies, antibody fragments, monobodies or nanobodies are labelled “AB” and are by nature MS complexes. For details regarding the benchmark assembly refer to the “Methods” section and SI Tables 2 and 3.

Docking

To establish the baseline performance of HADDOCK on this membrane protein complex docking benchmark and generate a docking decoy dataset that can serve for further optimisation of membrane-specific scoring functions, we performed the docking using two different scenarios.

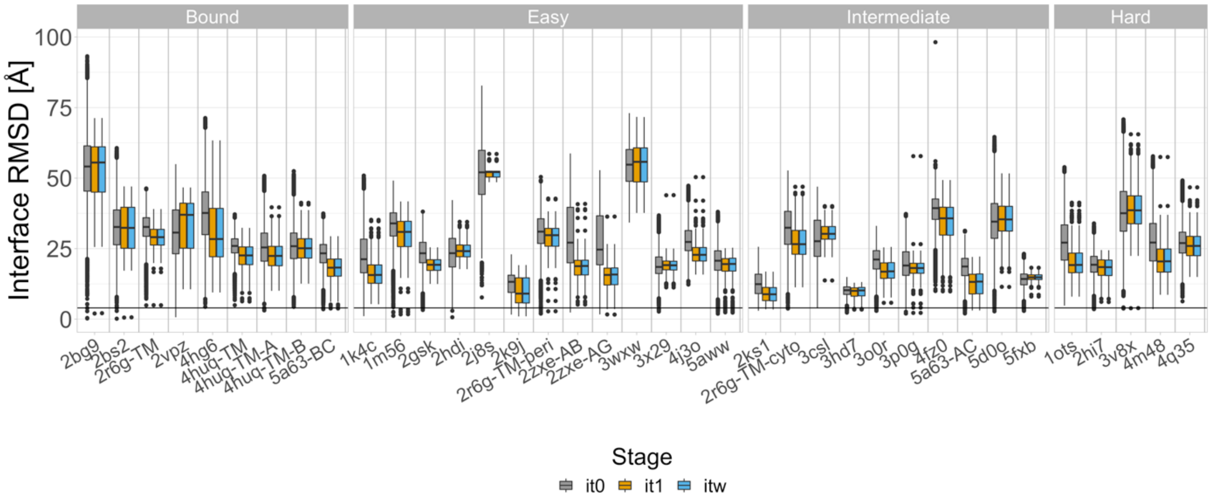


*Fig 2: I-RMSD values of the docking decoys of the membrane protein complex docking benchmark for the random-restraint driven runs. The complexes are grouped by difficulty. Each complex is represented by three boxplots, corresponding to the rigid-body (grey), semi-flexible refinement (orange) and final water refinement (blue) stages of HADDOCK. The black line represents the acceptability cut-off of 4 Å I-RMSD. The boxes of the boxplots range from the 1st to the 3rd quartile, the upper whisker extends from the hinge to the maximum value or 1.5 * Inter Quantile Range (IQR), the lower whisker extends from the hinge to the minimum value or 1.5 * IQR, outliers are shown as black points.*

Table 1: *The dimeric entries of the membrane protein complex docking benchmark. The first column is the PDB id of the complex structure, columns 2 and 3 the PDB ids of the unbound structures, category refers to the complex type, composition refers to the origin of every component of the complex, difficulty and I-RMSD reflect the difficulty of the target, secondary structure classifies the complex into one of two categories (Beta and Helical) depending on the secondary structure characteristics of its transmembrane domain, and Buried Surface Area refers to the buried surface area at the interface of every complex. The categories are Buried, MS, TM, Both and AB and they correspond to complexes whose interface lies inside a β-barrel, between cytosolic and transmembrane domains, between transmembrane domains, between transmembrane–cytosolic and transmembrane–transmembrane domains, and a complex of an antibody-like domain that stabilizes a transmembrane domain, respectively. The composition types can be BB, UB, HB, HU and UU and they stand for Bound-Bound, Unbound-Bound, Homology-Bound, Homology-Unbound, and Unbound-Unbound, respectively. BB means that both chains originate in the bound complex, UB means that one of the chains originates in the bound complex and the other in another structure, HB means one chain is a homology model based on another structure/complex and the other originates from the bound complex, HU means one chain is a homology model based on another structure/complex and the other originates from another complex or free structure, and UU means that both chains originate from another structure.*

complex	Unbound PDB id 1	Unbound PDB id 2	Category	Composition	Difficulty	i-RMSD [Å]	Buried Surface Area [Å^2]	Secondary Structure
2bg9	2bg9_ADE	2bg9_BC	Both	BB	Bound	0	5452.5	Helical
2bs2	2bs2_AB	2bs2_CD	MS	BB		0	4173.9	Helical
2r6g-TM	2r6g_F	2r6g_G	TM	BB		0	8073.3	Helical
2vpz	2vpz_AB	2vpz_CD	MS	BB		0	2064.7	Helical
4hg6	4hg6_A	4hg6_B	TM	BB		0	4704.2	Helical
4huq-TM	4huq_S	4huq_T	TM	BB		0	5202.9	Helical
4huq-TM-A	4huq_ST	4huq_A	MS	BB		0	1771.8	Helical
4huq-TM-B	4huq_ST	4huq_B	MS	BB		0	2682.6	Helical
5a63-BC	5a63_B	5a63_C	TM	BB		0	3430.2	Helical
2hdi	2hdi_A	1cii_A	Buried	UB	Easy	0.361	1925.7	Beta
4j3o	4j3o_D	3bfq_FG	Buried	UB		0.392	4681.2	Beta
1m56	2gsm_AB	1m56_CD	TM	UB		0.572	4961.5	Helical
1k4c	1k4c_A	1j95_ABCD	MS	UU		0.638	1766.9	Helical
3x29	3x29_A	2quo_A	MS	UB		0.673	2143.3	Helical
2k9j	2rmz_A	2k1a_A	TM	UU		0.678	982.0	Helical
2r6g-TM-peri	2r6g_FG	1jw4_A	MS	UB		0.716	3807.0	Helical
2gsk	2guf_A	1u07_A	MS	UU		0.86	1636.2	Beta
5aww	5aww_YG	5aww_E	TM	UB		0.868	2636.5	Helical
2zxe-AG	2zxe_A	2zxe_G	TM	UB		0.919	1528.0	Helical
2zxe-AB	2zxe_A	2zxe_B	TM	UB		0.94	1503.5	Helical

3wxw	3wxw_A	1vfa_HL	AB	HB		0.982	1672.9	Helical
3hd7	3hd7_A	3hd7_B	TM	UU	Intermediate	1.024	663.2	Helical
3csl	3csl_A	1b2v_A	MS	UB		1.065	3681.6	Beta
2ks1	2n2a_A	2m0b_A	TM	UU		1.158	662.2	Helical
5d0o	5d0o_A	2yhc_A	MS	UB		1.182	2909.4	Beta
5a63-AC	5a63_A	5a63_C	TM	UB		1.218	1953.1	Helical
3p0g	2rh1_A	4unu_A	AB	UU		1.26	1801.8	Helical
2r6g-TM-cyto	2r6g_FG	1q12_AB	MS	UB		1.363	3959.9	Helical
3o0r	3o0r_B	3o0r_C	AB	UB		1.445	2383.6	Helical
5fxb	5fxb_AB	1ttf_A	AB	HB		1.475	1716.6	Helical
4q35	4q35_A	4nhr_A	Buried	UB	Hard	2.061	5592.7	Beta
4m48	4m48_A	4dvb_HL	AB	HB		2.335	1144.1	Helical
2hi7	1ti1_A	2k73_A	MS	UU		2.588	1337.9	Helical
1ots	1kpk_AB	4nzu_HL	AB	HU		3	1327.7	Helical
3v8x	3v8x_A	4x1b_A	MS	HB		3.422	4945.0	Beta

Table 2:*The trimeric entries of the membrane protein complex docking benchmark. The first column is the PDB id of the complex structure, columns 2, 3 and 4 the PDB ids of the unbound structures, category refers to the complex type (refer to Table 1 for details), composition refers to the origin of every component of the complex (refer to Table 1 for details), difficulty and I-RMSD reflect the difficulty of the target, secondary structure classifies the complex into one of two categories (Beta and Helical) depending on the secondary structure characteristics of its transmembrane domain, and Buried Surface Area refers to the buried surface area at the interface of every complex. The unbound subunits for both entries originate from a different PDB entry of the same complex crystalized under different conditions (3w9h and 2qts for 2j8s and 4fz0 respectively). Both their individual backbone RMSDs (see Table S1) and i-RMSD values justify their classification as "UUU".*

complex	**Unbound PDB id 1**	**Unbound PDB id 2**	**Unbound PDB id 3**	**Category**	**Composition**	**Difficulty**	**i-RMSD [Å]**	**Buried Surface Area [Å^2]**	**Secondary Structure**
2j8s	3w9h_A	3w9h_B	3w9h_C	Both	UUU	Easy	0.648	10358.8	Helical
4fz0	2qts_A	2qts_B	2qts_C	Both	UUU	Intermediate	1.18	12084.0	Helical

Random Restraints

In the first scenario HADDOCK was used in its ab-initio mode with random restraints. Fig. 2 shows the distribution of I-RMSD values for all three stages of the docking runs (SI-Fig.1 shows the same plot but for L-RMSD). The RMSD values have been calculated according to CAPRI criteria as specified in the "Methods" section. The boxplots coloured grey, orange and cyan correspond to the I-RMSD values of it0, it1 and itw respectively. The horizontal black line represents the acceptability cut-off of 4 Å. Fig. 3 shows the same information but instead of displaying the raw I-RMSD values the models are classified by their quality. The figure is separated into 12 sub-graphs, each of which corresponds to one of the three docking stages (it0, it1, and itw) in one of the four difficulty groups (Bound, Easy, Intermediate, and Hard). Every sub-graph groups the performance for all complexes that have been classified into the same difficulty category for one of the three docking stages. The Y axis corresponds to the ranking of every model according to its HADDOCK score as calculated by the appropriate scoring function for every stage (see "Methods"), with models ranked near the bottom having a better score. Every model is represented by a single horizontal bar, with the colour of the bar representing the quality of the model. There is a limited number of acceptable models since we only used random restraints (ab initio docking mode) to drive the docking. Despite that, HADDOCK was able to generate at least one acceptable model in 27 of 37 (~73%) cases during it0 when considering all 50000 generated models. In 13 of those cases at least one acceptable model was also selected in the top 400 which are selected for further refinement in it1 and itw. This means that our scoring function could identify at least one acceptable model in ~48% of the cases where at least one model of acceptable quality was generated during it0. The success rate of 48% might sound less than ideal, but it becomes more impressive when one considers the number of acceptable models generated against the size of the sampling pool: In most cases only a few (<= 10) acceptable models were generated in it0 and they were correctly identified as near-native among 50000 models. SI Table 4 lists the number of acceptable structures generated during it0 for all complexes as well as the number of acceptable complexes ranked in the top 400. No more than these few acceptable models were sampled for the majority of complexes, however near-native structures were identified even when there were less than 5 of those in a pool of 50000 (1m56, 2bs2), including one case where the single near-native complex generated was selected (3v8x). The overall success rate of HADDOCK at the water stage using random restraints is ~35%, when considering all water models, as models of acceptable quality were generated in 13 of 37 cases. The difficulty or category of a complex seems to have no effect on the performance of HADDOCK with all difficulties and categories proportionately represented in the list of complexes for which no acceptable model was generated. The distribution of success rates when considering different cut-offs, for the random-restraint driven runs, can be seen in the left panel of Fig. 4. Every cell of that plot corresponds to the quality of the best model (minimum I-RMSD) when considering the top N structures (with N being 1, 5, 10, 20, 50, 100 and 400). The colour of the plot represents the quality of the minimum I-RMSD model. Even though the number of complexes for which at least one acceptable model was generated in the top400 did not change between the rigid body and refinement stages, refinement did improve the ranking of those acceptable models as well as their quality. This trend is also reflected in the mean rank of the first acceptable model which is ~110, ~22 and ~33 for it0, it1 and itw respectively. The mean I-RMSD of all acceptable models is 3.38 ± 0.5Å, 3.15 ± 0.77Å and 3.13 ± 0.76Å for it0, it1 and itw respectively.

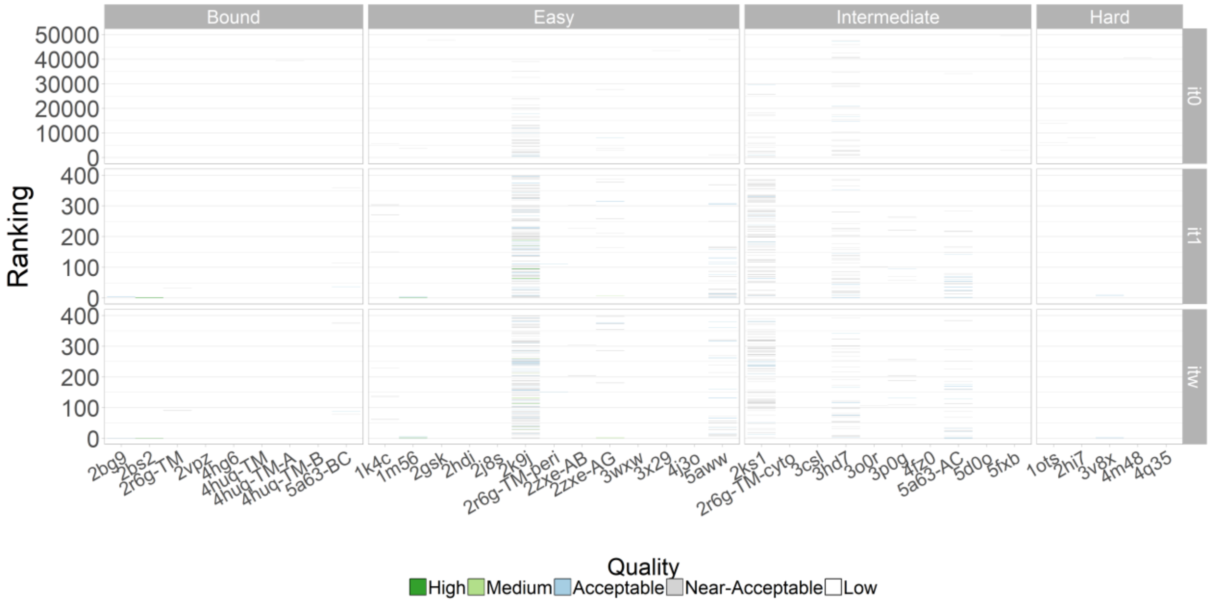


Fig 3: *Quality assessment of the generated models of the random-restraint driven runs based on I-RMSD values. The complexes are grouped by difficulty. For each, results of the rigid-body docking (it0 – top panel), semi-flexible refinement (it1 – middle panel) and water refinement (itw – bottom panel) are shown. The Y axis for all subgraphs correspond to the ranking of the models according to the default HADDOCK scoring function with models ranked near 0 having the best scores. Every model has been coloured according to its quality with high, medium, acceptable, near acceptable and low-quality models having I-RMSD values of less than 1Å (dark green), between 1 and 2Å (light green), between 2 and 4Å (light blue), between 4 and 6Å (light grey) and over 6Å (white) respectively).*

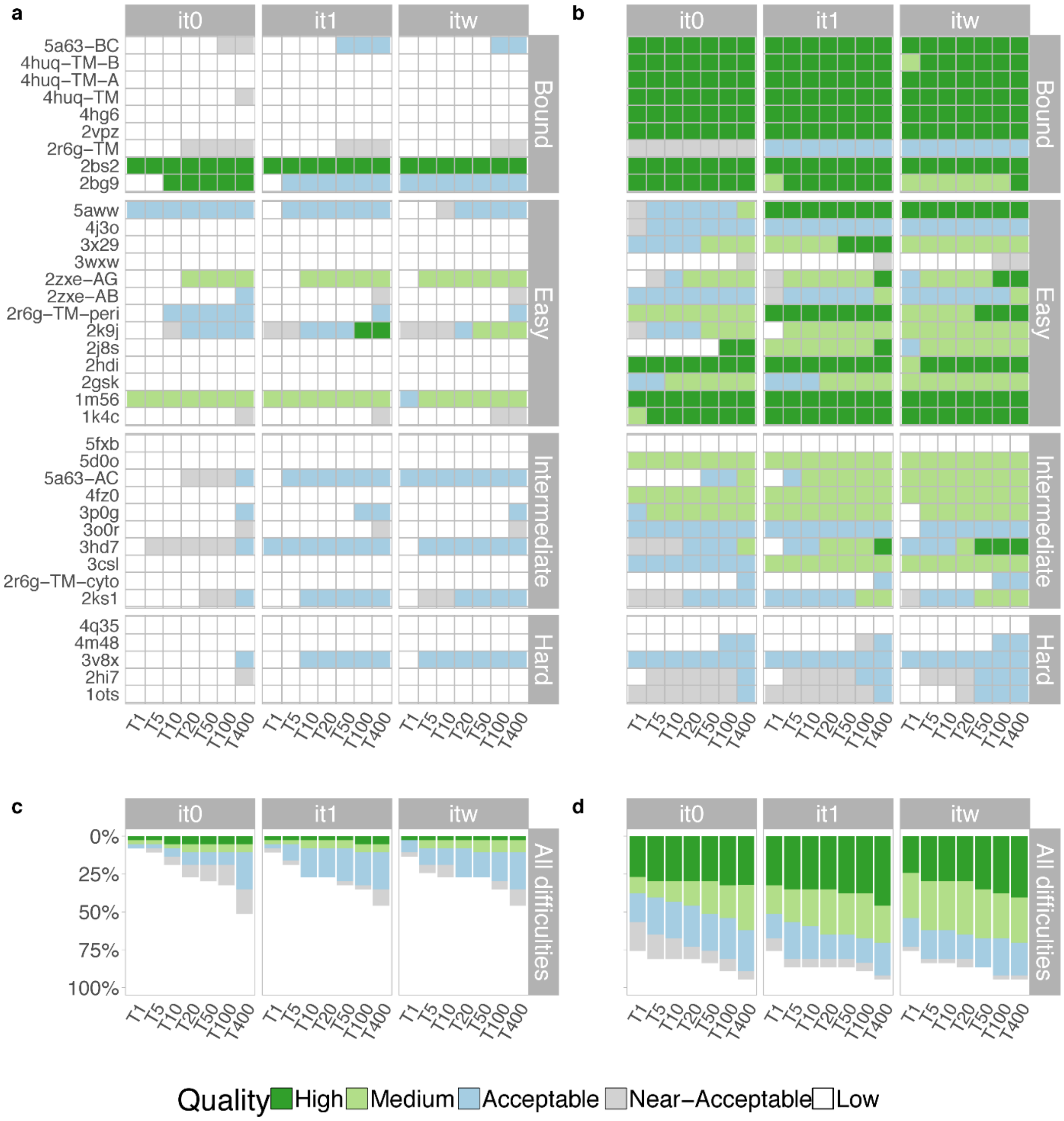


Fig 4: Evaluation of the success rate as a function of the number of models considered. Every set of horizontal cells corresponds to the performance of HADDOCK on a given complex, with the performance for random restraints being shown in panels a and c, and the performance for true interface restraints in panels b and d. For the top panels, every cell corresponds to the quality of the best model (in terms of I-RMSD) when considering the N best models (N having the values 1, 5, 10, 20, 50, 100 and 400) for all three docking stages, with the colouring of the cell representing high-, medium-, acceptable- and near acceptable-quality models (I-RMSD values of less than 1Å (dark green), between 1 and 2Å (light green), between 2 and 4Å (light blue), between 4 and 6Å (light grey) and over 6Å (white) respectively). Cells that correspond to cut-offs where only low-quality models were generated are coloured white. In the bottom panels, the success rate percentage for all complexes is shown as a function of the number of models considered. The colouring is the same as for the top panel.

True Interface Restraints

Fig. 5 shows the performance of HADDOCK when using true interface information from the native complex to drive the docking, which thus represents an ideal scenario (SI-Fig2 shows the same plot but for L-RMSD). Unlike for the random restrains runs, the difficulty of the complex is now important and is the main limiting factor for the performance of these runs. This is particularly apparent when comparing the bound and hard targets. In the bound complexes both sampling and scoring are better since a greater number of high-quality structures are generated during it0 and are scored near the top, meaning they proceed to the refinement stages. In general, the performance of HADDOCK is excellent with 36 of 37 complexes having at least one acceptable (I-RMSD <= 4Å) or near acceptable (I-RMSD <= 6Å) model in the rigid-body stage. The inclusion of near-acceptable models can be justified by the fact that when using a well-defined set of restraints, a rigid-body model in the near acceptable range might become acceptable after semi-flexible refinement, as is the case for example for complex 2r6g-TM. For that complex, no acceptable models were generated during it0, but the scoring function successfully identified the best models and, after refinement, more than half of the it1 and itw models became acceptable. Acceptable models are generated for 34 of 37 complexes during the refinement stages corresponding to an overall success rate of 92%, when considering all water models. The right panel of Fig. 4 shows the distribution of success rates for different cut-offs (see “Random Restraints” above for more details). Except for three cases (2r6g-TM-cyto, 3wxw and 5fxb), for which acceptable models were generated in it0 but not scored in the top 400 that are selected for semi-flexible refinement, our scoring function works well, ranking most near-native models higher than the non-native ones.

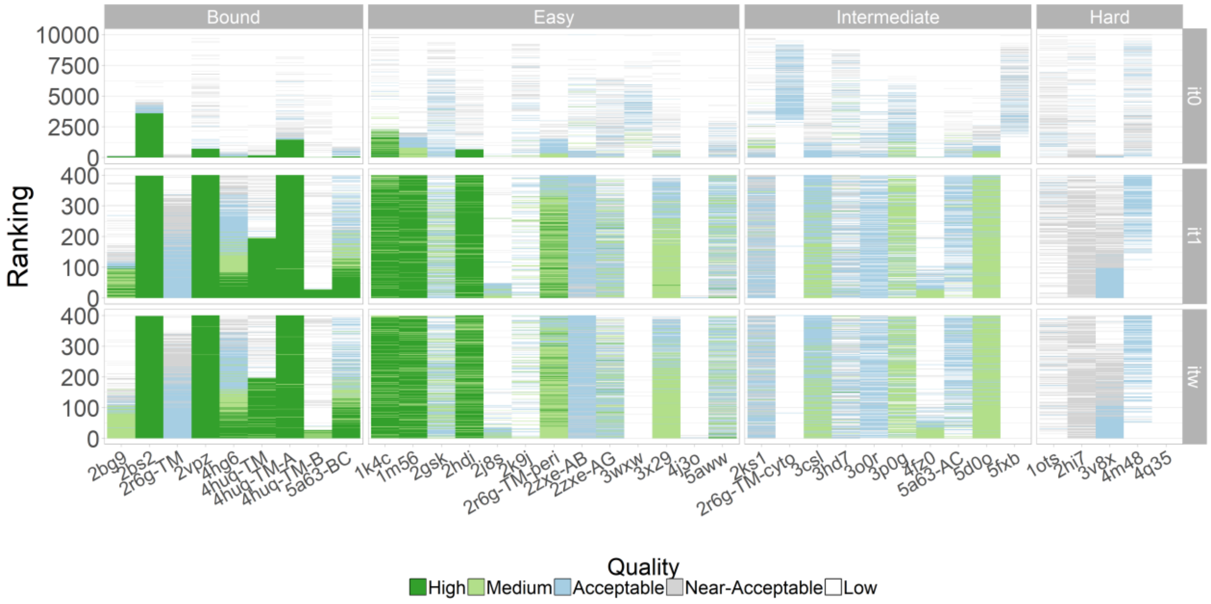


***Fig 5**: Quality assessment of the generated models of the true-interface restraint driven runs based on I-RMSD values. For details refer to the caption of Fig. 3.*

Benchmark availability

The bound and unbound structures, including the renumbered models used for docking of the membrane protein complex docking benchmark version1, along with ProFit analysis scripts can be freely downloaded at https://github.com/haddocking/MemCplxDB. The HADDOCK docking decoys are made available through the SBGrid Data Bank [263] and can be downloaded at https://data.sbgrid.com/618 [225].

Conclusion

We have assembled a membrane protein-protein docking benchmark which, to the best of our knowledge, is the first of its kind. The benchmark is freely available for download from GitHub and, in addition to the reference and unbound structures, includes renumbered, docking-ready structures, reference structures and analysis scripts for the calculation of the RMSD metrics that we are reporting in this paper. We have established the docking performance baseline of HADDOCK for two extreme scenarios. Despite the fact that HADDOCK has not been optimized for membrane proteins, it demonstrates excellent performance in the case where high-quality interface data are available, with a 92% overall success rate when considering all 400 itw models. In its ab-inito docking mode, however, the performance drops to 35% for itw models. In particular the sampling performance in the rigid body docking stage is affected, where we generate at least one acceptable model in 73% of the cases but only select at least one for further refinement in 48% of them with many near native models not being selected for the semi-flexible refinement stage as a result. This leaves room for optimization. All docking decoys for the various stages and scenarios can be freely downloaded from the SBGrid data bank. This new docking benchmark and its associated docking decoys should be a valuable resource for the community to foster the development of docking and scoring approaches for membrane protein complexes.

Acknowledgments

This work was supported by the European H2020 e-Infrastructure grants West-Life (grant no. 675858) and BioExcel (grant no. 675728) and by the Dutch Foundation for Scientific Research (NWO) (TOP-PUNT grant 718.015.001). The authors thank Dr Irina Moreira for helpful discussions. The authors declare no competing financial interest.

Supplementary Information

***Table S1**: Overall backbone RMSD values for all entries not classified as "Bound". The RMSD values have been calculated after optimal superimposition of the unbound chains on the respective reference chains using all backbone atoms for the fitting.*

complex_pdb_id	chain_identifier(s)	RMSD [Å]
1k4c	A	0.272
1k4c*	C	0.753
1m56	AB	0.560
1ots	AB	0.873
1ots	CD	2.990
2gsk*	A	2.250
2gsk	B	0.753
2hdi*	B	0.655
2hi7	A	1.446
2hi7	B	3.296
2j8s	A	0.811
2j8s	B	0.820
2j8s	C	0.722
2k9j	A	0.751
2k9j	B	0.611
2ks1	A	1.381
2ks1	B	0.810
2r6g-TM-cyto*	AB	1.149
2r6g-TM-peri	E	0.891
2zxe-AB	B	1.689
2zxe-AG	G	1.631
3csl	C	1.316
3hd7	A	1.388
3hd7	B	1.095
3o0r	C	2.326
3p0g	A	0.994
3p0g	B	0.867
3v8x	B	3.202
3wxw	HL	1.302
3x29*	B	0.879
4fz0	A	1.406
4fz0	B	1.341
4fz0	C	1.217
4j3o	G	0.667
4m48	HL	3.305
4q35	B	3.762

5a63-AC	A	2.275
5aww	E	1.749
5d0o	D	1.530
5fxb	F	2.267

* For the cases marked with an asterisk the sequence identity between the unbound and reference chains is not 100% due to differences between the reference and unbound structures. For chains 1k4c_C and 2gsk_A the sequence identity is 95.1 and 92.5% due to the presence of gaps in the alignment. For 2hdi_B, 2r6g-TM-cyto_AB and 3x29_B the sequence identity is 99.0, 99.7 and 99.1% due to mutations.

Table S2: *Dimeric entries of the membrane protein complex benchmark.*

complex	complex_pdb_id	Unbound PDB id 1	Unbound PDB id 2	Category	Composition	Difficulty	Interface RMSD [Å]	Comments
1	2bg9 [264]	2bg9_ADE	2bg9_BC	Both	BB	Bound	0	Chains A, D and E have been joined into chain A with chains D and E renumbered from 1000 and 2000 respectively. Chains B and C have been joined into chain B with chain C renumbered from 1000.
2	2bs2 [265]	2bs2_AB	2bs2_CD	MS	BB	Bound	0	Chains A and B have been joined into chain A with chain B renumbered from 1000. Chains C and D have been joined into chain B with chain D renumbered from 1000. Chains E and F have been joined into chain C with chain F renumbered from 1000.
3	2r6g-TM [266]	2r6g_F	2r6g_G	TM	BB	Bound	0	Chain F and chain G.
4	2vpz [267]	2vpz_AB	2vpz_CD	MS	BB	Bound	0	Chains A and B have been joined into chain A with chain B renumbered from 1000. Chains C and D have been joined into chain B with chain D renumbered from 1000. Chains E and F have been joined into chain C with chain F renumbered from 1000.
5	4hg6 [268]	4hg6_A	4hg6_B	TM	BB	Bound	0	-
6	4huq-TM [269]	4huq_S	4huq_T	TM	BB	Bound	0	Chain S and chain T.
7	4huq-TM-A [269]	4huq_ST	4huq_A	MS	BB	Bound	0	Chains S and T have been joined into chain S with chain T renumbered from 1000.

8	4huq-TM-B [269]	4huq_ST	4huq_B	MS	BB	Bound	0	Chains S and T have been joined into chain S with chain T renumbered from 1000.
9	5a63-BC [270]	5a63_B	5a63_C	TM	BB	Bound	0	-
10	2hdi [271]	2hdi_A	1cii_A [272]	Buried	UB	Easy	0.361	Residues 282-385 of 1cii_A.
11	4j3o [273]	4j3o_D	3bfq_FG [274]	Buried	UB	Easy	0.392	Chains F and G have been joined into chain F with chain G renumbered from 1000.
12	1m56 [275]	2gsm_AB [276]	1m56_CD	TM	UB	Easy	0.572	Chains A and B have been joined into chain A with chain B renumbered from 1000. Chains C and D have been joined into chain B with chain D renumbered from 1000.
13	1k4c [277]	1k4c_AB	1j95_ABCD [278]	MS	UU	Easy	0.638	Chains A, B, C and D of 1j95 have been joined into chain C with chains B, C and D renumbered from 1000, 2000, and 3000 respectively. Chains A and B of 1k4c have been joined into chain A with chain B renumbered from 1000.
14	3x29 [279]	3x29_A	2quo_A [280]	MS	UB	Easy	0.673	-
15	2k9j [281]	2rmz_A [282]	2k1a_A [283]	TM	UU	Easy	0.678	Only the first conformer has been kept for both bound and free structures.
16	2r6g-TM-peri [266]	2r6g_FG	1jw4_A [284]	MS	UB	Easy	0.716	Chains F and G have been joined into chain F with chain G renumbered from 1000.
17	2gsk [285]	2guf_A [286]	1u07_A [287]	MS	UU	Easy	0.86	Residues 153-233 of 1u07 and residues 5-594 of 2guf.
18	5aww [288]	5aww_YG	5aww_E	TM	UB	Easy	0.868	Joined chains Y and G into chain Y with chain G renumbered from 1000. Residues

								24-60 of chain E have been modelled using ideal helical backbone angles.
19	2zxe-AG [289]	2zxe_A	2zxe_G	TM	UB	Easy	0.919	The first 15 residues of chain G have been modelled using ideal backbone helical angles.
20	2zxe-AB [289]	2zxe_A	2zxe_B	TM	UB	Easy	0.94	Residues 31-61 of chain B have been modelled using ideal backbone helical angles.
21	3wxw [290]	3wxw_A	1vfa_HL [291]	AB	HB	Easy	0.982	Chains H and L have been joined into chain B with chain L renumbered from 1000.
22	3hd7 [292]	3hd7_A	3hd7_B	TM	UU	Intermediate	1.024	Chain A residues 95-116 and chain B residues 286 have been modelled using ideal helical backbone angles.
23	3csl [293]	3csl_A	1b2v_A [294]	MS	UB	Intermediate	1.065	-
24	2ks1 [295]	2n2a_A [296]	2m0b_A	TM	UU	Intermediate	1.158	-
25	5d0o [297]	5d0o_A	2yhc_A [298]	MS	UB	Intermediate	1.182	-
26	5a63-AC [270]	5a63_A	5a63_C	TM	UB	Intermediate	1.218	Residues 666-698 of chain A have been modelled using ideal helical backbone angles.
27	3p0g [299]	2rh1_A [300]	4unu_A [301]	AB	UU	Intermediate	1.26	Only the first conformer has been kept for both bound and free structures.
28	2r6g-TM-cyto [266]	2r6g_FG	1q12_AB [302]	MS	UB	Intermediate	1.363	Chains F+G & chains A+B. Chains F+G are joined into chain F and chain G is renumbered from 1000. Chains A+B are

								joined into chain A and chain B is renumbered from 1000.
29	3o0r [303]	3o0r_B	3o0r_C	AB	UB	Intermediate	1.445	Residues 5-40 of chain C have been modelled using ideal helical backbone angles.
30	5fxb [304]	5fxb_AB	1ttf_A [305]	AB	HB	Intermediate	1.475	Chains A and B of 5fxb have been joined into chain A with chain B renumbered from 1000.
31	4q35 [306]	4q35_A	4nhr_A [307]	Buried	UB	Hard	2.061	-
32	4m48 [308]	4m48_A	4dvb_HL [309]	AB	HB	Hard	2.335	Chains H and L have been joined into chain B with chain H renumbered from 1000.
33	2hi7 [310]	1ti1_A [311]	2k73_A [312]	MS	UU	Hard	2.588	-
34	1ots [313]	1kpk_AB [314]	4nzu_HL [315]	AB	HU	Hard	3	Chains A and B have been joined into chain A with chain B renumbered from 1000. Chains H and L have been joined into chain C with chain D renumbered from 1000.
35	3v8x [316]	3v8x_A	4x1b_A [317]	MS	HB	Hard	3.422	-

Table S3*: Trimeric entries of the membrane protein benchmark.*

complex	complex_pdb_id	Unbound PDB id 1	Unbound PDB id 2	Unbound PDB id 3	Composition	Category	Difficulty	Interface RMSD [A]	Comments
1	2j8s [318]	3w9h_A [319]	3w9h_B	3w9h_C	UUU	Both	Easy	0.648	-
2	4fz0 [320]	2qts_A [321]	2qts_B	2qts_C	UUU	Both	Intermediate	1.18	-

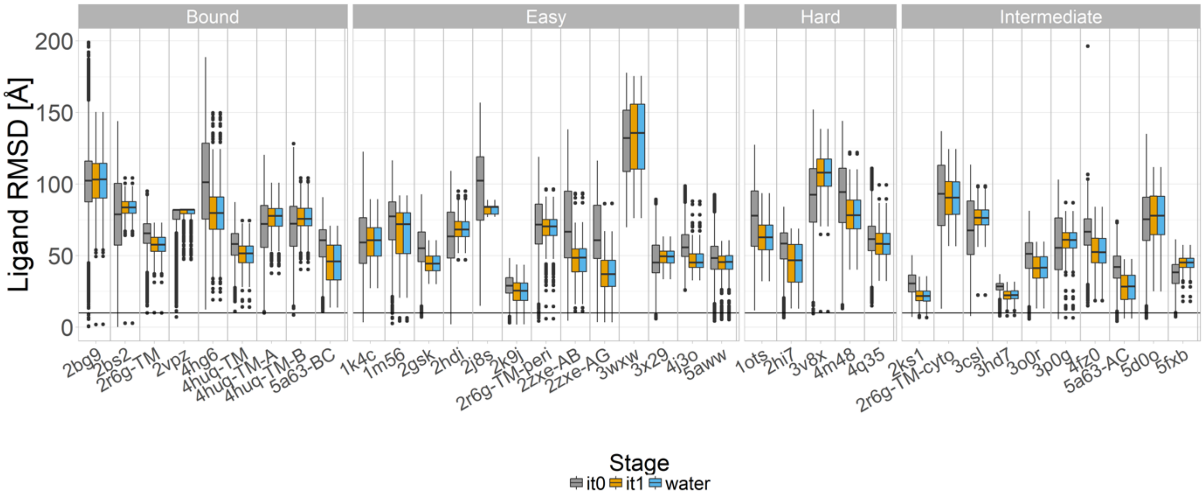


Fig S1*: Distributions of L-RMSD values for the various models of the benchmark complexes obtained using random restraints (see Methods). The complexes have been grouped by difficulty. Every complex is represented by three boxplots. The grey, orange and blue boxplots correspond to the RMSD values of all generated models for the first, second and third stage of the docking run (it0, it1, itw). The black line represents the acceptability cut-off of 10 Å. The boxes of the boxplots range from the 1st to the 3rd quartile, the upper whisker extends from the hinge to the maximum value or 1.5 * Inter Quantile Range (IQR), the lower whisker extends from the hinge to the minimum value or 1.5 * IQR, outliers are shown as black points.*

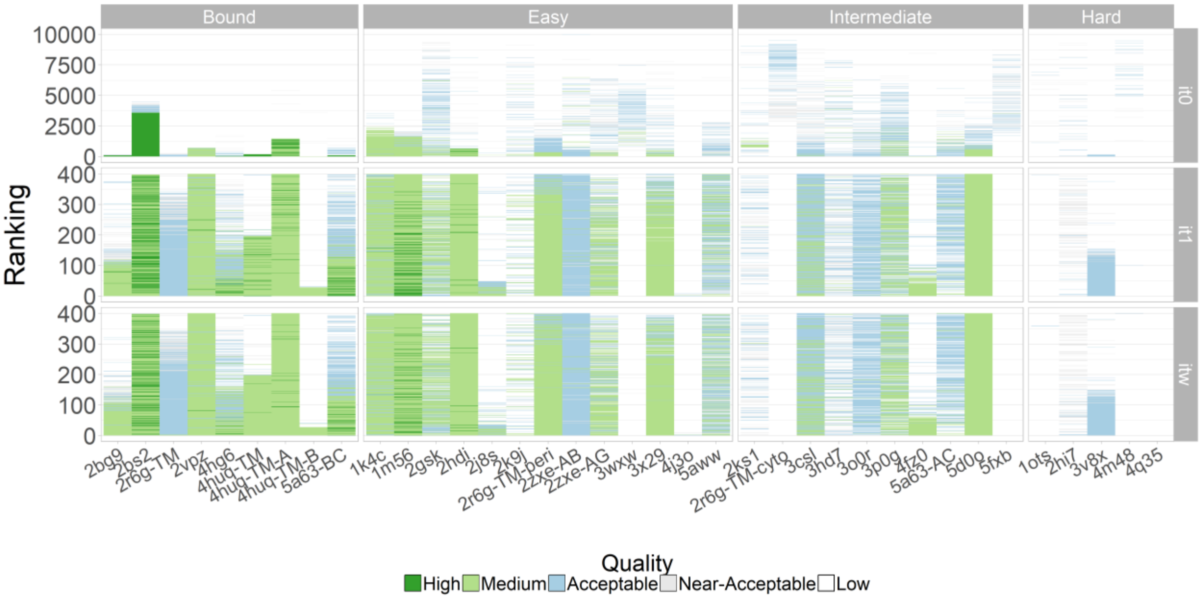


Fig S2*: Assessment of the quality (L-RMSD) and ranking (y-axis) of generated models of the true interface-driven runs. Complexes are grouped by difficulty. Results are shown for the docking stages of HADDOCK: rigid body docking (it0) – top panels; semi-flexible refinement (it1) – middle panels; final water refinement (itw) – bottom panels. The Y axis for all subgraphs correspond to the ranking of the models according to the default HADDOCK scoring function, with models ranked near 0 having the best scores. Every model has been coloured according to its quality with high, medium, acceptable, near acceptable, and low-quality models having L-RMSD values of less than 1 Å, between 1 and 5 Å, between 5 and 10 Å, between 10 and 12 Å, and more than 12 Å respectively. The high-quality models have been coloured dark green, the medium-quality ones light green, the acceptable-quality ones light blue and the near acceptable ones light grey.*

Table S4: *Number of generated (column 3) and selected models (column 4) during the rigid-body stage (it0) of the random restraints-driven runs for all entries of the benchmark.*

complex	pdb_id	acceptable models in it0	acceptable models in top400 of it0
1	1k4c	17	0
2	1m56	4	4
3	1ots	0	0
4	2bg9	5	1
5	2bs2	4	1
6	2gsk	4	0
7	2hdi	14	0
8	2hi7	6	0
9	2j8s	0	0
10	2k9j	749	49
11	2ks1	107	1
12	2r6g-TM-cyto	1	0
13	2r6g-TM	0	0
14	2r6g-TM-peri	5	1
15	2vpz	5	0
16	2zxe-AB	6	1
17	2zxe-AG	58	3
18	3csl	2	0
19	3hd7	290	3
20	3o0r	16	0
21	3p0g	6	1
22	3v8x	1	1
23	3wxw	0	0
24	3x29	6	0
25	4fz0	0	0
26	4hg6	0	0
27	4huq-TM-A	4	0
28	4huq-TM-B	0	0
29	4huq-TM	0	0
30	4j3o	0	0
31	4m48	2	0
32	4q35-cut	0	0
33	5a63-AC	24	5
34	5a63-BC	1	0
35	5aww	35	10
36	5d0o	3	0
37	5fxb	15	0

Chapter 3

Modelling of membrane protein complexes with HADDOCK

Abstract

Despite the significant role membrane proteins and their complexes play in many cellular processes like for example signal transduction and transport of nutrients, and the well-established difficulty in their experimental characterisation, most docking codes still do not support modelling of membrane protein complexes with specific adaptations tailored to the membrane environment. Here we present ongoing work, related to the development of an implicit membrane representation for use in HADDOCK. The bilayer is represented by a shape consisting of layer(s) of beads - dummy atoms - which are used as anchors to restrain the transmembrane parts of integral membrane protein complexes into the bilayer. The performance of this new protocol is compared with a simple protocol in which only one centre-of-mass restraint is defined between the core transmembrane regions of the docking partners, a minimal-data scenario, and another one in which information extracted from the interface of the native complex is used to drive the docking, a perfect-case scenario.

Introduction

Cells are the fundamental units of life. From the simplest single-cell organisms to the most complex multicellular ones, all living creatures are comprised of cells. A cell is a collection of organelles in a concentrated aqueous solution bound by a double layer of amphipathic molecules, for the most part phospholipids. It is this lipid bilayer, or plasma membrane, that allows the cell to maintain the concentration of the various substances it needs to preserve its homeostatic status, grow, divide and in the case of multicellular organisms differentiate [322]. Membranes also surround some of the organelles and compartments of the cell like the nucleus and mitochondria. In addition to the lipids, the plasma membrane is made up of proteins with multiple functions, most prominently, cellular communication via signalling cascades and transfer of substances in and out of the cell or cellular compartment [200]. The importance of membrane proteins is such that, depending on the organism, membrane proteins might make up anywhere from 20 to 30% of that organism's proteome [200,323]. They are also important for the pharmaceutical industry since, for example, G-Protein Coupled Receptors (GPCRs) alone constitute more than half of all prescription drug targets [324] and ~35% of all drug targets on the market with even more compounds targeting proteins of this family under development or in clinical testing [209].

Given the eminent difficulty of structurally charactering membrane proteins with experimental methods (see Chapters 1 and 2), computational approaches provide an attractive alternative to methods like X-ray crystallography, NMR and cryo Electron Microscopy (cryo-EM). However, despite many advances in areas like membrane protein-specific databases and recent developments in all atom and coarse-grained protein and lipid forcefields, the docking field has lagged behind with only a handful of codes having published new or adapted protocols for the docking of membrane protein complexes (see Chapter 1). In this chapter we describe ongoing work revolving around adding support for docking membrane proteins in HADDOCK [59,256].

Membrane protein complexes can be categorised in one of three groups based on the topology of the interaction:

A. Complexes which are not embedded in the membrane and only one or more of their components are (transiently) anchored to it.
B. Complexes whose interface lies at the membrane-soluble interface, and finally
C. Complexes which have at least two subunits that are entirely or partially embedded in the membrane.

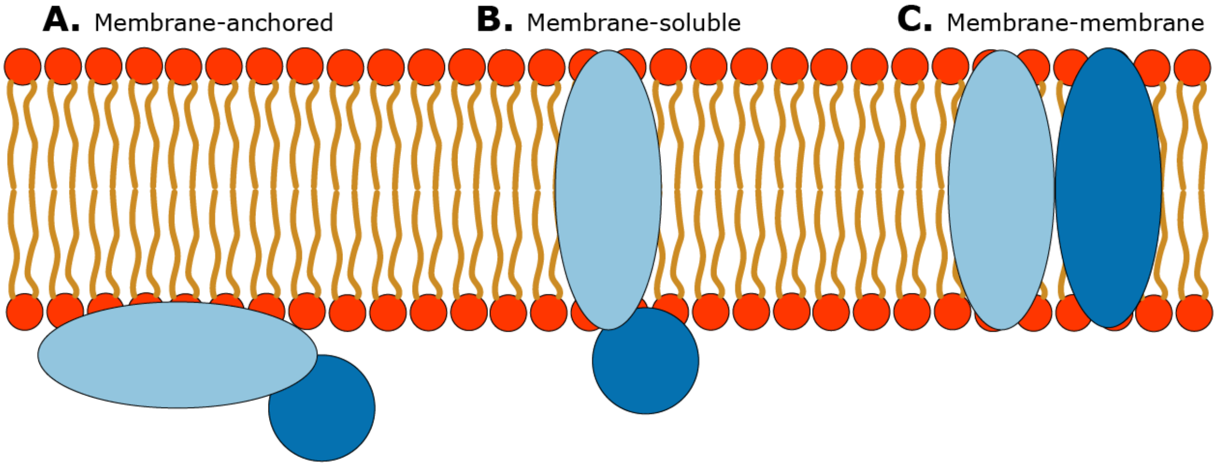


Fig 1*: Schematic representations of the three classes of membrane protein complexes. Panel A shows a protein-protein complex whose one subunit is anchored to the membrane (grey), whereas the other (purple) is entirely in solution. Panel B shows one whose one subunit is embedded in the membrane (grey) and the other (purple) lies at the membrane-soluble interface which is also where the protein-protein interface for this complex lies. Panel C shows a transmembrane protein-protein complex with both subunits entirely embedded in the membrane.*

Illustrations of these categories are shown in Fig. 1. The first two do not require in principle specific adaptations to any docking protocols as the interface of those complexes lies in the soluble region. For the second, membrane-soluble category the membrane proteins are embedded in the lipid bilayer, which constrains their rotation and translation. Also, the presence of membrane does put restraints on the area of the surface that should be sampled, an information that might benefit sampling strategies. This holds true for the complexes of the third category whose subunits are entirely or partially embedded in the membrane and whose interface lies entirely within the membrane bilayer as well (membrane-membrane). Complexes like GPCRs, ion channels and transporters all belong to this category which is going to be the focus of this chapter. The energetics for all cases might need to be revisited especially in the case of empirical desolvation potentials which have been parametrised with the assumption that the complexes are surrounded by water.

Materials and Methods

HADDOCK is an information-driven docking method which can make use of many types of data to drive the simulation. These data can be derived experimentally or computationally. These are usually translated into ambiguous distance restraints which are used to guide the simulation as well as for the scoring of the generated models (see Chapter 2 for an extended description of the method).

Membrane representation.

This concept of distance restraints is also central to the way we chose to represent the membrane in HADDOCK. As a first representation we chose an implicit model with the bilayer represented as a network of regularly spaced beads (see Fig. 2). These beads do not interact with the remaining of the system in any way. They are placed in such a way that the geometric centre of the transmembrane region of the protein and that of all the beads overlap and the primary axis of the protein is parallel to the membrane normal.

After generating a network of beads at the desired spacing, we place all subunits of the complex of interest in the "bilayer" after specifying three parameters:

1. Insertion angle,
2. Rotation angle around the membrane normal, and
3. Translation distance along the membrane normal

For those, we take the values from the OPM database [206] for the protein of interest or its closest homolog.

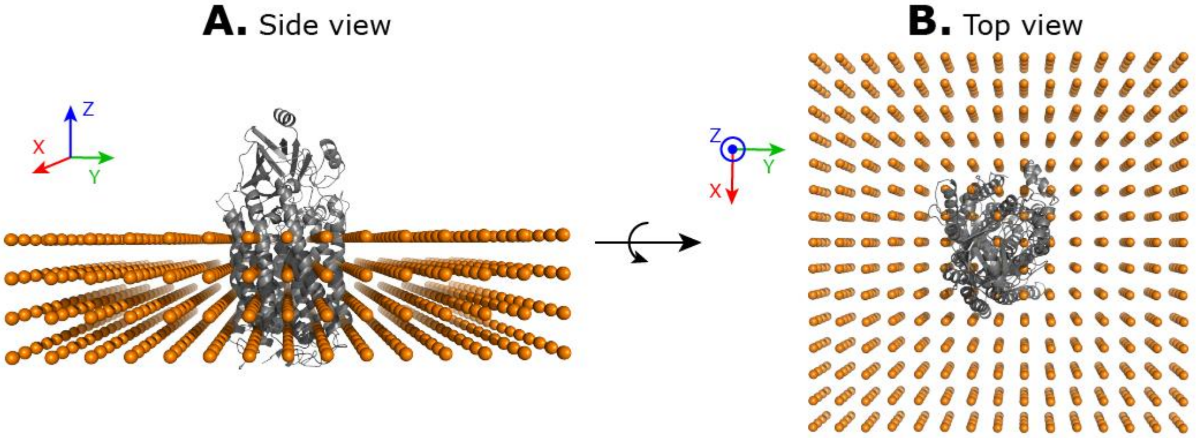


Fig 2: Subunits I and II of the cytochrome c oxidase (PDB entry 1m56) embedded in the "bilayer" which is 30Å wide and the spacing of the beads is 10Å in all three dimensions. The protein is shown as grey cartoons and the beads as orange spheres. ***Panel A*** *shows the view from the side and* ***panel B*** *from the top of the "bilayer". All molecular graphics were generated with PyMOL [83].*

Docking protocols

After insertion, ambiguous distance restraints are specified between the core transmembrane regions of every subunit and the shape beads. The way those regions are defined depends on the number of layers used to represent the membrane. We have selected to test the performance of our implementation using two representations: The first is identical to the one shown in Fig. 2 with the membrane width set to 30Å and a 10 Å spacing between the beads. For the second representation we have chosen a single layer whose Z coordinates are the same as those of the geometric centre of the transmembrane part of the protein i.e. a single plane perpendicular to the membrane normal, with the same 10 Å spacing between beads. For the first representation, ambiguous distance restraints are defined between every Cα carbon of the protein subunits within 10Å of the geometric centre along the Z axis (the core of the transmembrane region of the protein) and all beads, whereas for the second one, restraints are restricted to the Cα carbons within 5Å of the geometric centre along the Z axis. The target distance for these restraints was equal to half the spacing i.e. 5Å. In addition to the restraints driving the subunits to the beads we are also defining a centre of mass restraint between the geometric centre of the core transmembrane regions. This transmembrane (TM) centre of mass (CM) restraint is meant to bring the two subunits together in a way which only depends on knowing which are the transmembrane regions of each. We evaluate the performance of the two sets of simulations with the bead layers (bead-multiple and bead-single) and compare it with docking runs in which only the TM-CM restraint is active (TM-CM) and one in which we use the true interface information to drive the docking (True Interface – TI), which represents a best case scenario. The settings which differ compared to the default behaviour of HADDOCK are the number of models that are generated during the rigid body and refinement stages which is set to 10000 and 400 respectively, the number of rigid body energy minimisation trials which is set to 1 and the systematic sampling of 180° symmetrical solutions during the rigid body stage which is disabled. The TM-CM, Single-Layer and Multi-Layers runs were performed with the latest version of HADDOCK (v2.4) whereas the TI have been previously performed on the HADDOCK webserver (v2.2) (see Chapter 2).

Dataset

We assembled our dataset by extracting the cases classified as TM from the benchmark that was described in Chapter 2 after excluding targets with ids 2r6g-TM and 4hg6. Target 2r6g-TM was excluded because the two subunits intertwine around each other, making this an impossible target to model without the use of highly specific restraints; target 4hg6 was excluded because of the topology of the interacting surface. This resulted in a set of ten targets, details of which are shown in Table 1.

***Table 1**: The entries of the dataset used in this study. The first column is the PDB id of the complex structure, columns 2 and 3 the PDB ids of the unbound structures, composition refers to the origin of every component of the complex, difficulty and I-RMSD reflect the difficulty of the target, secondary structure classifies the complex into one of two categories (Beta and Helical) depending on the secondary structure characteristics of its transmembrane domain, and Buried Surface Area refers to the buried surface area at the interface of every complex. The composition types can be BB, UB and UU and they stand for Bound-Bound and Unbound-Unbound, respectively. BB means that both chains originate in the bound complex, UB means that one of the chains originates in the bound complex and the other in another structure and UU means that both chains originate from another structure.*

complex	Unbound PDB id 1	Unbound PDB id 2	Composition	Difficulty	i-RMSD [Å]	Buried Surface Area [Å^2]	Secondary Structure
4huq-TM	4huq_S	4huq_T	BB	Bound	0	5202.9	Helical
5a63-BC	5a63_B	5a63_C	BB		0	3430.2	Helical
1m56	2gsm_AB	1m56_CD	UB	Easy	0.572	4961.5	Helical
2k9j	2rmz_A	2k1a_A	UU		0.678	982.0	Helical
5aww	5aww_YG	5aww_E	UB		0.868	2636.5	Helical
2zxe-AG	2zxe_A	2zxe_G	UB		0.919	1528.0	Helical
2zxe-AB	2zxe_A	2zxe_B	UB		0.94	1503.5	Helical
3hd7	3hd7_A	3hd7_B	UU	Intermediate	1.024	663.2	Helical
2ks1	2n2a_A	2m0b_A	UU		1.158	662.2	Helical
5a63-AC	5a63_A	5a63_C	UB		1.218	1953.1	Helical

Results and Discussion

We only present and discuss here results related to the first stage of the docking – the rigid body stage (it0) of HADDOCK in order to focus on the impact of the bead layer representation and restraints on the sampling performance of our protocol. The rigid body stage, and the scoring happening at this stage, are the main limiting factor in terms of achieving an overall higher performance.

The TM-CM runs serve as the baseline against which we compare the two bead layer runs (Single and Multi), while the TI runs are an indication of what is achievable using high-quality interface data extracted from the native complexes. By examining Fig. 3 two observations become immediately obvious: 1) The bead layer runs do not seem to confer any obvious advantage compared to the TM-CM runs, and 2) the Single runs are performing worse than the Multi ones with the multi-layer runs outperforming the Single ones for every target and almost every number of top ranked models considered.

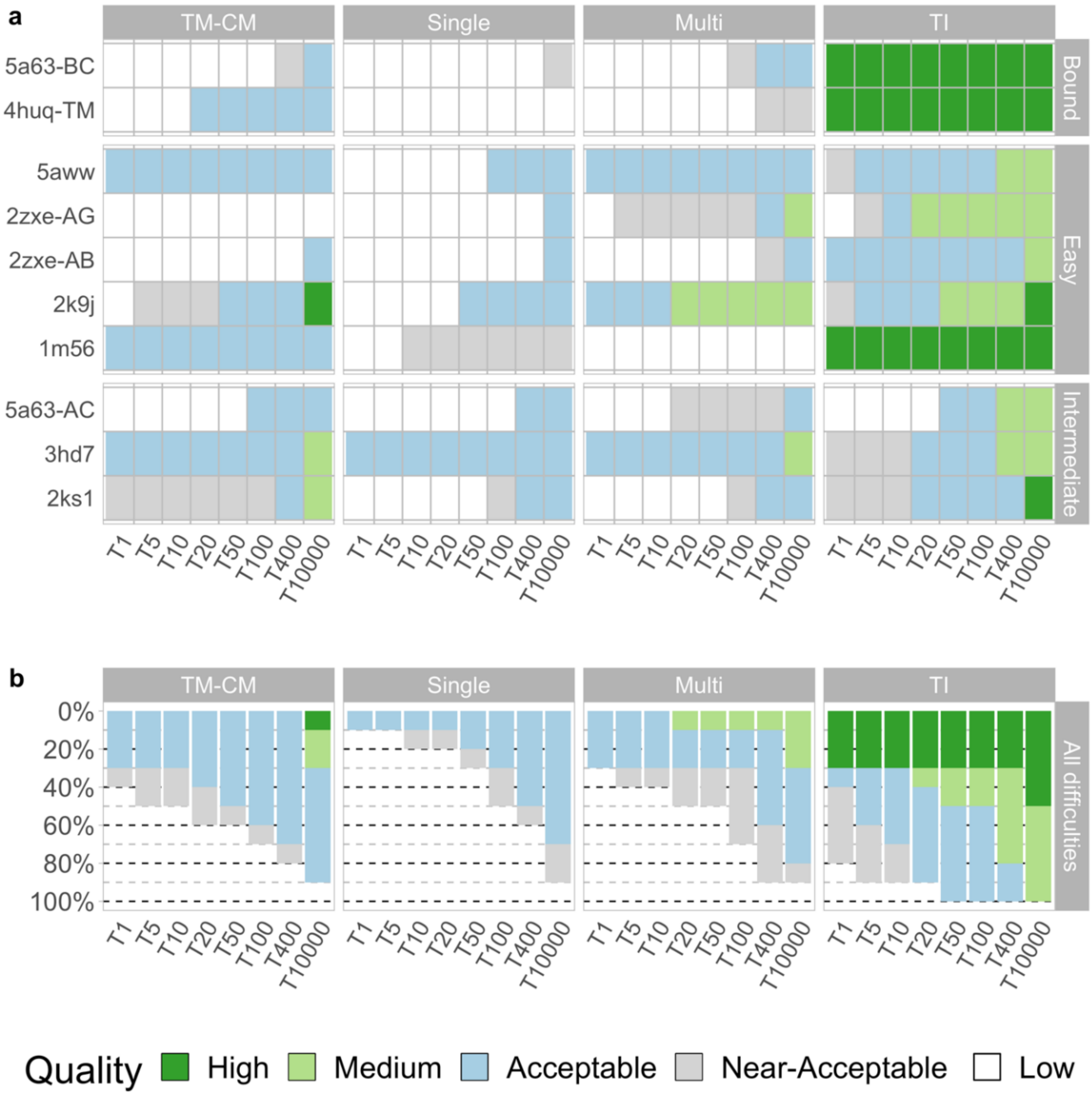


Fig 3*: Evaluation of the success rate as a function of the number of models considered. Every set of horizontal cells corresponds to the performance of HADDOCK on a given complex, with the performance for Transmembrane centre-of-mass (TM-CM, Single bead layer (Single), Multiple bead layers (Multi) and True Interface (TI) runs being shown from left to right in panel a. For panel a, every cell corresponds to the quality of the best model (in terms of I-RMSD) when considering the N best models (N having the values 1, 5, 10, 20, 50, 100, 400 and 10000 [all models generated]) for the rigid body stage (it0), with the colouring of the cell representing high-, medium-, acceptable- and near acceptable-quality models (I-RMSD values of less than 1Å (dark green), between 1 and 2Å (light green), between 2 and 4Å (light blue), between 4 and 6Å (light grey) and over 6Å (white) respectively). Cells that correspond to cut-offs where only low-quality models were generated are coloured white. In panel b, the success rate percentage for all complexes is shown as a function of the number of models considered. The colouring is the same as for the top panel.*

The success rate – defined as the number of complexes for which at least one model of acceptable- (or better) quality was generated and ranked within a given number of top ranked models – is level between the TM-CM and Multi runs when considering the top 1 to top 10 models, according to the HADDOCK Score. However, beyond that cut-off the TM-CM runs clearly perform better and we have to consider the top 400 models before the performance between the two is comparable. While the low performance of the bead layer runs is surprising, that of the TM-CM runs is probably the point which stands out the most: HADDOCK can generate acceptable- or better-quality models for 7 out of 10 complexes when considering the top 400 models. This is a significant value for the number of models to consider as in a typical docking scenario with no or very limited information regarding the interacting surface of the proteins (such as here) we would be generating 10000 models in the rigid body stage and selecting the top 400 of them for subsequent flexible refinement. Even more impressive, the success rate for the rigid body stage when considering only the top model is 30% – similar to that of the Multi runs, while the TI ones reach 40%, with high-quality models being generated for 3 of those 4 complexes.

Fig. 4 shows the distribution of the quality of the generated models over the entire ranked pool of 10000 it0 models. Those distributions highlight the fact that, in addition to the low sampling performance, scoring is also a challenge, in particular for the Single runs, which show the broadest distribution of acceptable- or better-quality models. Considering solely the sampling performance, irrespective of scoring, between the two bead layer runs the Multi protocol comes out on top with an overall success rate of 80% vs 70% for the Single protocol. For scoring we used the default rigid body scoring function of HADDOCK as described in Chapter 2, including the empirical desolvation term that uses parameters optimised for water. That particular term accounts for the displacement of surface water molecules into the bulk solvent (water). This is clearly not happening in a membrane protein – membrane protein interface as the majority of their environment is made up of the lipid bilayer. Desolvation parameters tailored to membrane protein complexes do exist [325] and can be used in the future to optimise our scoring function for the lipid environment, which might improve the scoring of these complexes.

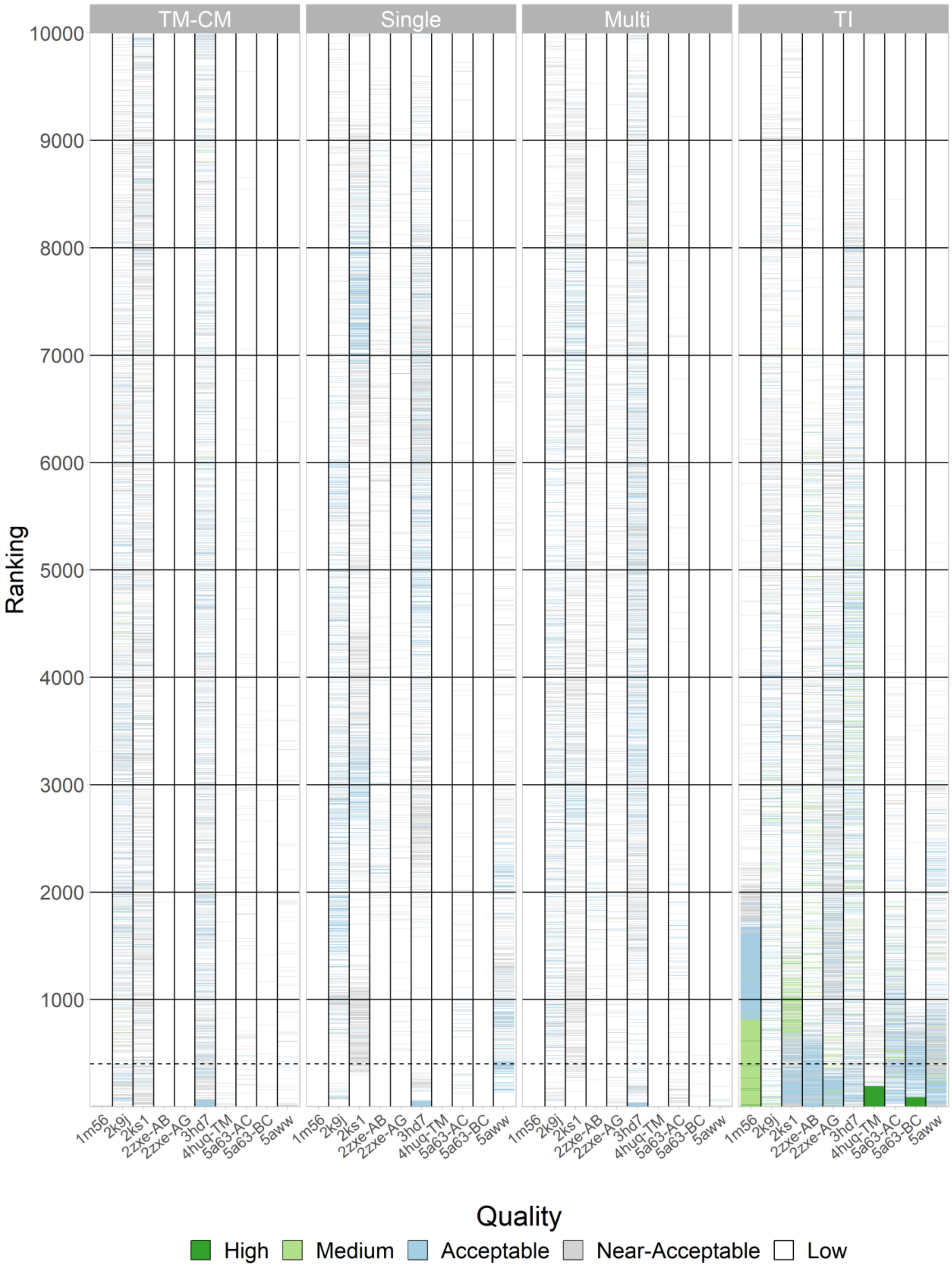


Fig 4*: Quality assessment of the generated models of all runs based on I-RMSD values. The complexes are grouped by run. The Y axis for all subgraphs corresponds to the ranking of the models according to the default HADDOCK scoring function with models ranked near 0 having the best scores. Every model has been coloured according to its quality with high, medium, acceptable, near acceptable and low-quality models having I-RMSD values of less than 1Å (dark green), between 1 and 2Å (light green), between 2 and 4Å (light blue), between 4 and 6Å (light grey) and over 6Å (white) respectively). The dotted line corresponds to the top 400 models according to HADDOCK score.*

An additional counterintuitive observation regarding the bead layer runs is the fact that the difficulty (defined in terms of conformational changes upon binding) of the various complexes of the dataset seems to have no effect on the outcome of the docking, especially when compared with the TI runs (see also Chapter 2). Rather, it seems that the size of the system and the number of restraints used during the simulation do affect the outcome to the largest degree.

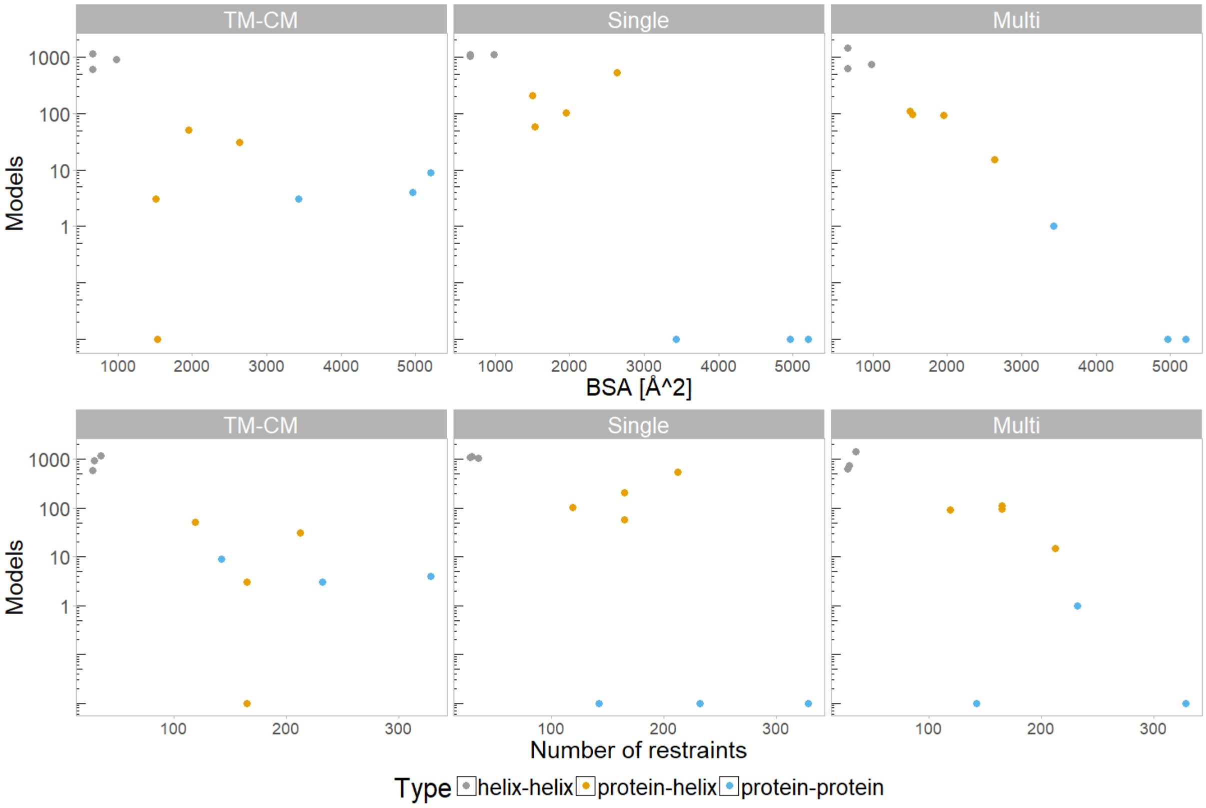


***Fig 5**: Number of acceptable- or higher quality models generated against the BSA (panel A) and number of restraints (panel B) of a given complex. Panel A: The X axis for every subplot shows the Buried Surface Area (BSA) and the Y axis the log of the number of acceptable models generated during the rigid body stage. For the complexes for which 0 acceptable models were generated, the number was set to 0.01 (log10(0.01)≈-100), to simplify the visualisation. Complexes are grouped according to complex type: Complexes where both partners are small helices (2k9j, 2ks1, 3hd7) are coloured grey, complexes where one partner is a small helix and the other a large TM receptor (2zxe-AB, 2zxe-AG, 5a63-AC, 5aww) are coloured orange and complexes where both partners are full-size proteins are coloured blue. Panel B shares the Y axis of panel A but the X axis shows the number of restraints used during the docking.*

Fig. 5 shows the relationship between the number of acceptable models generated and the size of the interface of the complexes (panel A) and the number of restraints used during the docking (panel B). The Buried Surface Area (BSA) has been used as a proxy for the size of the system (see Table 1 for the values). Our dataset is comprised of three types of complexes:

1. Protein-protein complexes; Complexes for which both partners are full size proteins – Complexes 1m56, 4huq-TM, 5a63-BC.
2. Protein-Small helix complexes; Complexes for which one partner is a full-size protein and the other a small TM helix – Complexes 2zxe-AB, 2zxe-AG, 5a63-AC, 5aww.

3. Small helix-small helix complexes; Complexes for which both partners are small TM helices – Complexes 2k9j, 2ks1, 3hd7.

Performance for the TM-CM and the layer runs is almost identical for the three small helical complexes. However, the layer runs have a significant advantage over the TM-CM runs when comparing the performance of the protein-small helix complexes for which only a handful (if any) acceptable models are generated during the TM-CM runs. The limiting factor in terms of performance for the layer protocols is the protein-protein group as only one acceptable-quality model is generated for complexes 1m56, 4huq-TM and 5a63-BC with the Multi protocol (see Figures 3 and 5), whereas the TM-CM protocol samples acceptable-quality models for all three of them when considering all models generated. The largest difference between them though is a result of the poor performance of the bead layer runs for the protein-protein complexes for which a single acceptable model is generated. This is the most significant factor preventing the bead layer runs from performing better than the TM-CM ones. The reason the bead layer protocols do not perform better on the larger complexes is not immediatelly apparent. One potential explanation would be the very large number of restraints that are defined between the beads and the Cα carbons of the TM residues, which might be preventing the models from converging. However, if this was the case, we would expect some small improvements in the Single vs the Multi layer protocol: The former only defines restraints between one layer and a smaller part of the TM region of the protein rather than four layers and the majority of the TM region. The lack of differences could indicate that the number of restraints is already too high in the Single runs. This could cause imbalances between energy terms (intermolecular van der Waals and electrostatic energies vs bead restraint energy) during the rigid-body docking stage, resulting in a poor convergence of the minimisation process.

Conclusion and Perspectives

In this Chapter, we have presented ongoing work related to the development of a protocol for the docking of TM proteins in HADDOCK. Two implementations of the suggested protocol were benchmarked on a small dataset comprised entirely of TM protein-protein complexes of different sizes. The results were compared with the performance of a protocol making use of only centre of mass restraints to drive the docking (TM-CM) and one making use of information derived from the interface of the native complex (TI), an ideal case scenario. Although the bead layer protocol was found to perform on par or favourably when compared with the TM-CM protocol for small- and medium-size complexes, it entirely failed to generate acceptable models for larger complexes. The reasons behind this are still being investigated with early indications that the number of restraints imposed on the system might be preventing the energy minimisation process from converging.

An alternative to using the aforementioned bead representations of the membrane and a simpler approach that does not require distance restraints between Cα carbons and any bead atoms, would be to use a simple harmonic potential defined along the Z-axis (parallel to the membrane normal) which would prevent TM atoms from leaving a predefined zone.

While the performance of the bead layer runs was rather disappointing, the performance of the TM-CM runs was all the more impressive, with success rates of 30 and 70% when considering the top 1 and 400 models, respectively. We are currently expanding this simple and robust protocol to define two additional distance restraints targeting pairs of Cα carbons in the top and bottom half of the proteins (along the Z-axis). This would allow to limit the orientational space available to the system when the relative orientation of the membrane proteins is approximately known (parallel or anti-parallel).

Chapter 4

Performance of HADDOCK and a simple contact-based protein-ligand binding affinity predictor in the D3R Grand Challenge 2

Zeynep Kurkcuoglu*, Panagiotis I. Koukos*, HADDOCK team, A.M.J.J. Bonvin

** These authors contributed equally to this work*

Journal of Computer-Aided Molecular Design, 2018, volume 32, issue 1, pages 175-185

Abstract

We present the performance of HADDOCK, our information-driven docking software, in the second edition of the D3R Grand Challenge. In this blind experiment, participants were requested to predict the structures and binding affinities of complexes between the Farnesoid X nuclear receptor and 102 different ligands. The models obtained in Stage1 with HADDOCK and ligand-specific protocol show an average ligand RMSD of 5.1Å from the crystal structure. Only 6/35 targets were within 2.5Å RMSD from the reference, which prompted us to investigate the limiting factors and revise our protocol for Stage2. The choice of the receptor conformation appeared to have the strongest influence on the results. Our Stage2 models were of higher quality (13 out of 35 were within 2.5Å), with an average RMSD of 4.1Å. The docking protocol was applied to all 102 ligands to generate poses for binding affinity prediction. We developed a modified version of our contact-based binding affinity predictor PRODIGY, using the number of interatomic contacts classified by their type and the intermolecular electrostatic energy. This simple structure-based binding affinity predictor shows a Kendall's Tau correlation of 0.37 in ranking the ligands (7th best out of 77 methods, 5th/25 groups). Those results were obtained from the average prediction over the top10 poses, irrespective of their similarity/correctness, underscoring the robustness of our simple predictor. This results in an enrichment factor of 2.5 compared to a random predictor for ranking ligands within the top 25%, making it a promising approach to identify lead compounds in virtual screening.

Introduction

Molecular docking is a widely-used tool in computer-aided drug design to model the three-dimensional (3D) structure of protein-ligand complexes, study their interactions and predict their binding affinities [326]. Integrated with data from the experimental techniques like X-ray crystallography and Nuclear Magnetic Resonance, docking has become a powerful tool in designing novel therapeutics [327]. Docking consists of two main steps: (i) exploration of protein-ligand binding poses (sampling) and (ii) identification of biologically relevant models (scoring). Both steps have their own challenges such as the flexibility of entities and the accuracy of the scoring functions. These have been reviewed elsewhere [327–329].

Our integrative, information-driven, flexible docking approach HADDOCK [58,59] addresses this structural modeling problem by using the available experimental and bioinformatics data to drive the docking process in combination with a simple but robust scoring function for ranking. The success of HADDOCK in modeling protein-protein, protein-nucleic acid and protein-peptide complexes has been demonstrated numerous times (for a review, see [330]). HADDOCK is also consistently among the top scorers and predictors [331] in The Critical Assessment of Predicted Interactions (CAPRI) experiment [50], where participants are expected to predict the 3D structure of an unknown biomolecular complex, given the sequence or the structure of the unbound partners.

While HADDOCK has also been used in several protein-ligand docking studies [329,332–338], no systematic benchmarking has been reported so far, making the D3R Grand Challenge 2 a perfect opportunity to assess its performance for this type of problem for which it was not originally developed. In this manuscript, we describe our strategy for predicting the binding poses of FXR ligands (Stage1), and assessing their binding affinities (Stage2), while discussing the main lessons learned from the challenge.

Materials and Methods

Data

The target of the D3R Grand Challenge 2 is the Farnesoid X nuclear receptor (FXR), which is a nuclear hormone receptor activated by bile acids [339]. FXR is highly expressed in liver, intestines and kidneys, playing an important role in the regulation of bile acid homeostasis and cholesterol, lipid and glucose metabolisms [339–341]. Due to its involvement in various diseases including inflammatory bowel disease, colorectal cancer and type 2 diabetes, FXR agonists have emerged as potential therapeutics [339–341].

In the D3R Grand Challenge 2, the FXR dataset consists of 36 crystal structures with a resolution below 2.6Å and binding data (IC50s) for 102 compounds, including the 36 for which a crystal structure is available (these were only made available in Stage2). These data have been provided by Roche and curated by D3R. The challenge consists of two stages, which are described below:

Stage1: The goal is to predict the poses of 35 ligands (one target is cancelled), and the affinities or rankings of all 102 compounds. The input files provided by organizers are the apo crystal structure of FXR and 2D ligands in SMILES and SD file formats.

Stage2: The participants are expected to predict the affinities or rankings of all 102 ligands with the 36 crystal structures of FXR-ligand complexes provided as additional input compared to Stage1.

Ligand preparation

SMILES strings of FXR-ligands were converted into 3D structures using OpenEye Omega Toolkit 2.6.4 [342]. Conformers were directly generated from SMILES by Omega torsional sampling, where the maximum number of conformers per ligand was set to 100. After this step, the conformers were clustered to select representative models to be used in the docking stage. We used for this the jclust hierarchical clustering of the MMTSB tools [343], with the maximum number of clusters set to 10 and the minimum number of structures per cluster to 4. For each ligand in Stage1, an ensemble of conformations was created by selecting a representative structure from each cluster.

Protein preparation

Docking simulations in Stage1 were run using an ensemble of 4 structures as input for the receptor. This final set of 4 receptors was selected as follows:

1. 28 homologue structures were found in the RCSB/PDB database [252] using the "Sequence" search feature with the sequence of the apo form of FXR provided by D3R and a lower limit of 80% sequence identity. All other parameters were kept as default (Search algorithm: BLAST, Expectation value: 10, Mask low complexity: yes). We also specified that structures must contain a ligand.

2. Interface residues were extracted from all homologous structures using a 5Å cutoff. All residues containing an atom located at 5Å or less from the ligand were then considered as interface. The union of all these residues was taken and matched to the target sequence. The list of residues was manually curated to remove residues on the outer surface of the receptor. We then refined the residues based on their surface accessibility (SA) in the FXR apo structure (<40% backbone or sidechain SA) using NACCESS [344].

Finally, some residues with a SA below 40% were reintroduced manually (mainly residues in loops). The identified interface residues were subsequently used for clustering the receptor (see point 3 below).

3. Any structure with one or more gaps at the interface was discarded (11) leaving 18 structures (17 homologues + 1 apo) for the calculation of a pairwise backbone-RMSD after a fitting step on the interface residues using ProFit [345]. HADDOCK's default clustering method [346] was applied on the RMSD matrix and generated 4 clusters when used with 0.5Å threshold and a minimum cluster size of 2. It is worth noting that the apo structure was not clustered with these criteria. Two other structures (1ot7_B [347] and 3p88 [348]) were not clustered as well. Cluster representatives with the best resolution and 1ot7_B were chosen as templates. 3p88 was discarded because it was too close from a representative of cluster #2.
4. Based on 4 templates (1osv [347], 1ot7_B, 3dct [349], 3olf [350]), a new set of interface residues were computed using a 4Å cutoff to define if a residue was interacting with the ligand or not. These residues were used as active residues in the docking runs (see Table S1 in Online Resource for the list).
5. For ensemble docking with HADDOCK, we mutated all residues diverging from the reference structure (apo form) to the respective residue with PyMOL [83]. Ensemble docking refers to the use of multiple starting conformations for one or more of the binding partners within the same docking run. All combinations of the various conformations are selected as starting point for the docking. How many times each conformation is sampled will thus depend on the number of conformation in the ensemble and the number of generated models at the rigid-body docking stage (see Docking below).

Revised protocol for Ligand and Protein Preparation in Stage2

In Stage2, 36 crystal structures for FXR1-36 protein-ligand complexes were provided by the organizers. We used those structures to revisit our docking protocol and identify the major limiting factor for our docking performance in Stage1. By docking with either bound ligand or receptor, we found that it is mainly the receptor conformation that limits our accuracy in generating near-native poses (see Results section). Accordingly, we identified the ligand that is most similar to FXR1-36 for targets FXR37-102 based on the Tanimoto distance calculated using fmcsR [351] and ChemmineR packages [352]. The corresponding receptor conformation was used as the protein input for all docking runs in Stage2.

As for input ligand ensemble, we followed the Stage1 protocol with an additional criterion enriching the major cluster: For the cases where less than 10 clusters were identified, remaining elements of the major cluster were additionally included in the docking ensemble, until the ensemble size reached the maximum of 10.

Access to the experimental structures of the ligands allowed us to examine the accuracy of the OMEGA generated conformers. The top panel of Fig. S1 in Online Resource provides an overview of the RMSDs of the ligand poses. The median RMSD of the generated poses for all targets was 1.9Å, the median RMSD of the poses selected for docking for stage 1 was 2.2Å and the median RMSD of the poses selected for stage 2 was 1.8Å. Overall, OMEGA generated accurate – if not quite near-native – models.

Docking

Docking was performed with the HADDOCK2.2 web server [59]. The docking protocol of HADDOCK consists of three stages: i) rigid-body docking by energy minimization from random orientations of the starting conformations – “it0” stage, ii) semi-flexible refinement of the interface by simulated annealing in torsion angle space – “it1” stage and iii) short molecular dynamics refinement in explicit solvent – “water” stage. In the semi-flexible stage (it1), protein interface residues (all those within 5Å of the ligand) and the ligand are treated as flexible. The calculations are guided by the ambiguous interaction restraints defined based on the binding pocket of the receptor (Point 2 under protein preparation above). For the D3R competition we used the buried settings of the small ligand protocol which had been benchmarked on the ASTEX dataset [353] [unpublished data]. Compared to the HADDOCK default settings, the buried binding site protocol scales the intermolecular energy terms (van der Waals and electrostatic) by a factor of 0.001 to allow penetration of the ligand into the protein binding site. This is required since the starting configurations for docking are randomly rotated and separated molecules. Accordingly, because models can contain clashes due to the scaling down of intermolecular interactions, the weight of the van der Waals energy term for scoring the initial rigid-body docking poses (it0) was set to 0.

Additionally, we fine-tuned the docking settings for Stage1 by testing on various structures of the FXR receptor bound to a plethora of ligands (namely 1osv, 1ot7, 3dct, 3hc5 [354], 3olf, 3omm [350]). Using the SMILES strings of those ligands we created ensembles of conformers as described in the ‘Ligand Preparation’ section, which we proceeded to dock against the ensemble of receptors generated during ‘Protein Preparation’ stage. The models were then compared with the bound complexes to determine the final docking settings. Based on those results, and considering the buried and rather hydrophobic nature of the binding pocket, we decided to base our selection of poses on the models obtained after the semi-flexible refinement stage (it1) of HADDOCK instead of the final, water-refined models. We increased the sampling to 10,000 and 400 poses for it0 and it1, respectively. All docking settings were left at default values except for the ones listed in Table S1 in Online Resource. The parameters and topologies for the ligands were obtained automatically by the HADDOCK server using a local version of PRODRG [355], which discards non-polar hydrogen atoms.

In both stages, two sets of restraints were provided to the server to guide the docking: 1) ambiguous interactions restraints in which the ligand and all residues in the binding pocket were defined as active to draw the ligand inside it - this was only used in it0 (50% of those restraints were randomly deleted for each docking trial); 2) unambiguous interaction restraints in which only the ligand was defined as active and the protein binding pocket as passive were used for the subsequent flexible refinement stage (it1). In this refinement phase, no energy penalty is generated if a binding pocket residue does not contact the ligand, which allows the ligand to explore the binding site. The top 5 poses from it1 stage were selected for submission.

The scoring function used for ranking the poses is the standard HADDOCK score for the flexible refinement (it1) which is defined as:

$$\mathrm{HADDOCKscore} = 1.0 * \mathrm{E_{vdW}} + 1.0 * \mathrm{E_{elec}} + 1.0 * \mathrm{E_{desol}} - 0.01 * \mathrm{BSA}$$

where BSA is the buried surface area in Å^2, $\mathrm{E_{desol}}$ an empirical desolvation energy term [260]. The intermolecular energies are calculated using the OPLS united atom force field parameters [259] for non-bonded atoms, using a 8.5Å cut-off with a shifting function for the electrostatic energy and switching function between 6.5 and 8.5Å for the van der Waals energy. For the electrostatics energy, a dielectric constant of 10 is used.

Binding Affinity Prediction

For Stage1 of the challenge, we used the HADDOCK score to rank the affinities of 102 compounds. For Stage2, we developed both a ligand-based and a structure-based binding affinity predictor, which are described below.

Ligand-based binding affinity predictor

We designed a target-specific ligand based binding affinity predictor, based on the assumption that similar ligands binding to the same protein should have similar binding affinities. From the database BindingDB [356], we retrieved 229 ligands that bind to the FXR protein with reported experimental IC50 data. We calculated the ligand similarity using Atom Pair (AP) and Maximum Common Substructure (MCS) measurements, as implemented in ChemmineR and fmcsR packages [351,352]. For this, we computed the pairwise similarity matrix among the training data (i.e., the 229 ligands). This matrix was used to train a Support Vector Regression (SVR) model using LibSVM software (version 3.21) [357]. During the training process, we transformed IC50 data into ln(IC50). We evaluated the SVR predictor on the training data using 10 repeats of 5-fold cross-validation. The AP metric outperformed the MCS metric (Table 1). We, therefore, in the subsequent analysis used AP to train our predictor. The binding affinity of the D3R ligands was then calculated using our predictor with the similarity matrix between the 102 D3R ligands and the training data (the 229 ligands from BindingDB).

***Table 1**. Comparison of the prediction performance of Atom-Pair and Maximum Common Substructure predictors on the training dataset using 10 repeats of 5-fold cross-validation.*

	Atom-Pair	**Maximum Common Substructure**
Kendall's Tau	0.52 ± 0.01	0.50 ± 0.01
Pearson's correlation coefficient	0.70 ± 0.01	0.68 ± 0.02

Structure-based binding affinity predictor

Recently, we have introduced a residue-residue contact-based method for the prediction of the binding affinity in protein-protein complexes [358], implemented in the webserver PRODIGY (PROtein binDIng enerGY prediction) [359,360]. This simple structural-based approach has led to one of the best performing predictors so far reported on a large and heterogeneous set of data [244,361], with Pearson's Correlation of 0.73 between the predicted and the experimental values and a root mean-squared error of 1.89 kcal mol^{-1}.

For Stage2 of this D3R challenge we designed an adapted version of our contact-based prediction for protein-ligand complexes. From the 2P2I database [338], we retrieved 200 protein-ligand complexes with experimentally measured Ki (inhibition constant) and available crystal structure. Ki values were converted to free energy (ΔG) by applying the equation $\Delta G = RT\ln(Ki)$, in which R is the gas constant and T the temperature. For each entry, we ran the HADDOCK refinement protocol in order to collect the intermolecular energy terms reported in Eq. 1. This consists of the final refinement stage of HADDOCK without any initial perturbation of the starting structures. We then calculated the number of atomic contacts (ACs) within the distance threshold of 10.5Å (this cutoff was optimized to obtain the best correlation). We further classified the ACs according to the atom involved in the interaction (C=Carbon, O=Oxygen, N=Nitrogen, X=Aall other atoms). We used this combination of structural- and energy-based terms to train a multiple linear regression model with R [362] performing 4-fold cross validation. We applied Akaike's Information Criterion (AIC) stepwise selection method implemented in R to avoid overfitting and identify the significant features. The resulting binding affinity predictor ΔG_{score} model for ranking the targets based is shown in Eq. 2:

$$\Delta G_{score} = 0.343794 * E_{elec} - 0.037597 * AC_{CC} + 0.138738 * AC_{NN} + 0.160043 * AC_{OO} - 3.088861 * AC_{XX} + 187.011384$$

where AC_{CC}, AC_{NN}, AC_{OO} and AC_{XX} are the ACs between Carbon-Carbon, Nitrogen-Nitrogen, Oxygen-Oxygen and between all other atoms and polar hydrogens, respectively. E_{elec} is the electrostatic energy calculated through the HADDOCK refinement protocol.

For each of the top 10 it1 poses from the docking runs we calculated the ΔG_{score} and took the average. We finally ranked the ligands according to the predicted values of our averaged ranking-score.

Results and discussion

Binding Pose Predictions

Following the protocol described in Methods, we submitted 5 binding poses per target in Stage1. Two of the successfully predicted cases are shown in Fig. 1, where the ligand RMSD (l-RMSD, defined as the RMSD of the ligand heavy atoms after fitting on receptor backbone) is less than 2.5Å. The performance per target in the prediction phase is indicated in Fig. 2 (dark grey box plots) for our submitted five poses. We have at least one model within 2.5Å of the bound state in 6 out of 35 targets with an average l-RMSD of 5.1Å for all targets. This rather low performance encouraged us to revisit the ligand and protein preparation protocols, as described in ‘Revised protocol’ section. In particular, we investigated whether conformational changes/sampling is the limiting factor (Fig. 3). Our docking performance in Stage1 is compared to that using either the bound ligand, bound receptor or both. Our performance reaches 83% success rate for bound-bound docking. The largest improvement compared to Stage1 is obtained if the bound conformation of the receptor is used. Moreover, revisiting the ligand sampling also increased the docking success from 14% to 20% for top5 (data not shown). This prompted us to select for Stage2 the receptor conformation containing the most similar ligand to the ligand to be docked (see Material and Methods) and a resampled ensemble of ligand conformations. The resulting improvement can be easily observed in Fig. 2 (light grey box plots), where the average l-RMSD is reduced to 4.1Å and 13 out of 35 targets are within the 2.5Å cut-off. We can also clearly see that there is plenty of room for optimizing our scoring function since in most cases we did generate reasonably good predictions (shown as circles) in the pool of 400 refined models, but these did not make it in the top5.

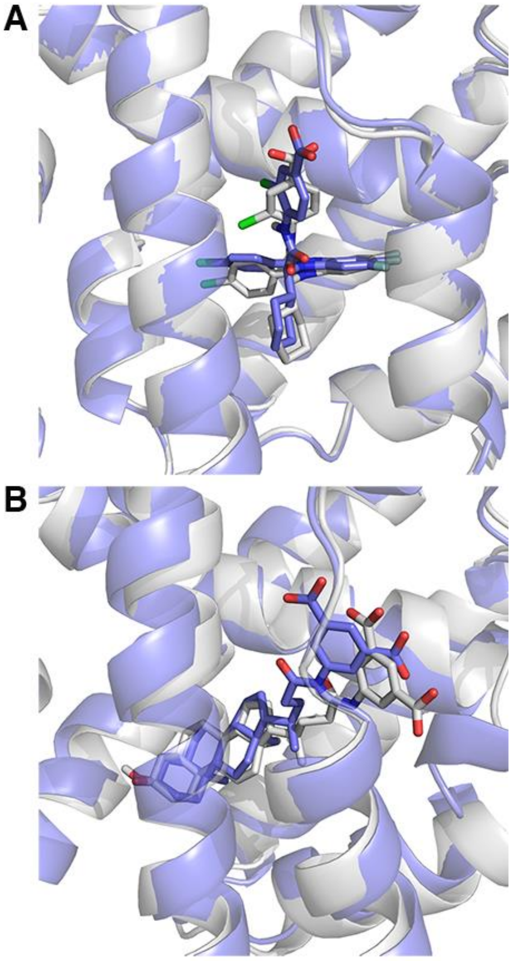


Fig 1*: Examples of successfully predicted ligand poses in Stage1 for (A) FXR-27 (B) FXR-34 with a l-RMSD of 1.27 and 1.94Å, respectively. The receptor conformations are shown as cartoon and the ligands as stick representation. The reference crystal structure is colored grey and the model as slate.*

Additionally, we investigated whether the revised protocol improves the sampling. Fig. 4 compares Stage1 and Stage2 binding poses, where the y-axis reflects the ranking of the top 100 structures at the end of it1 for each target, with higher ranked structures being close to zero. The coloring of the bars depends on the l-RMSD of the model to the bound complex, with darker shades corresponding to lower l-RMSD values. As is evident from Fig. 4, the revised protocol dramatically improves the sampling as low l-RMSD structures are identified and tend to be ranked higher.

We should also note that the ligand parameters were obtained automatically by the HADDOCK server using PRODRG – the only currently supported option on the server – with its known limitations. Especially the accuracy of the charge assignment by PRODRG can be questioned [363]. In a previous study on the prediction of the binding affinity of protein-protein interaction inhibitors [338], we have compared PRODRG and ACPYPE [364] for ligand parameter generation showing that the HADDOCK score calculated with the two parametrizations scheme are correlated (R^2=0.73). While the van der Waals and desolvation energies are essentially identical, the electrostatic energies differ substantially (R^2=0.33), which might well affect the quality of our docking poses.

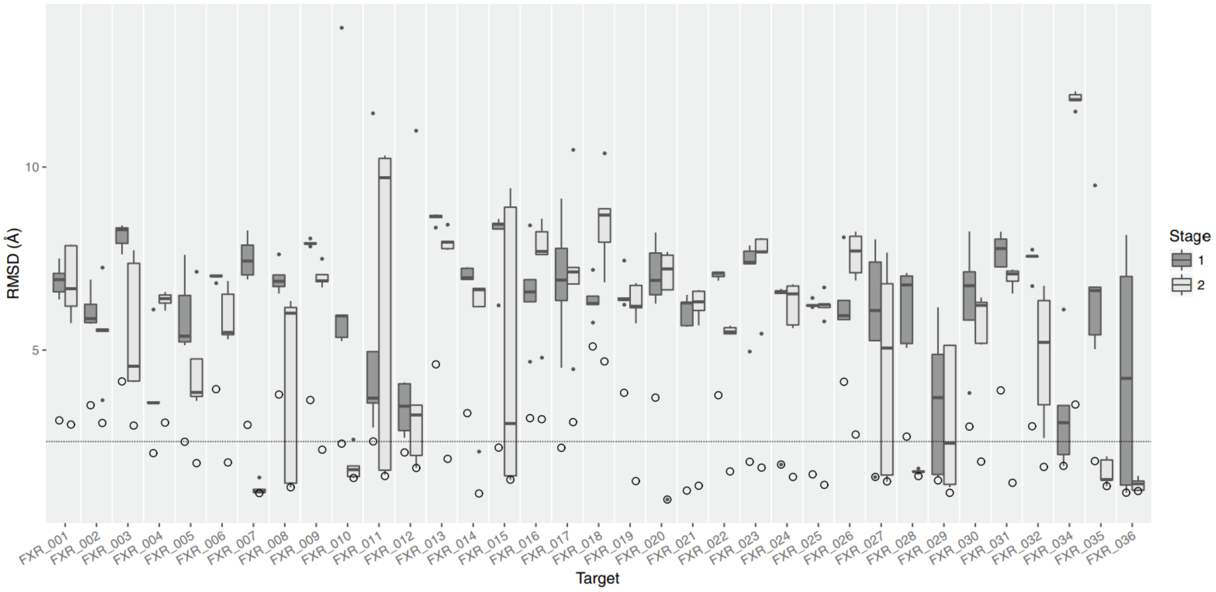


***Fig 2**: Comparison of the l-RMSDs of the top 5 scoring poses between stages 1 and 2. l-RMSD values of the top 5 poses are drawn as boxplots with the values of Stage1 colored dark gray and those of Stage2 light gray. The black line in the middle of the boxes corresponds to the median, the lower and upper hinges correspond to the 25th and 75th percentile respectively, the whiskers extend to no longer than 1.5 times the IQR from the hinge. Any point beyond that range is considered an outlier and drawn as a filled black point. The circles correspond to the overall minimum l-RMSD obtained in it1 for that target. In the cases where the circle overlaps with an outlier or a boxplot, the minimum l-RMSD structure is part of the top 5 scoring poses. The dotted line represents the l-RMSD cutoff of 2.5Å. The number of successful predictions increases from 6/35 in Stage1 to 13/35 in Stage2.*

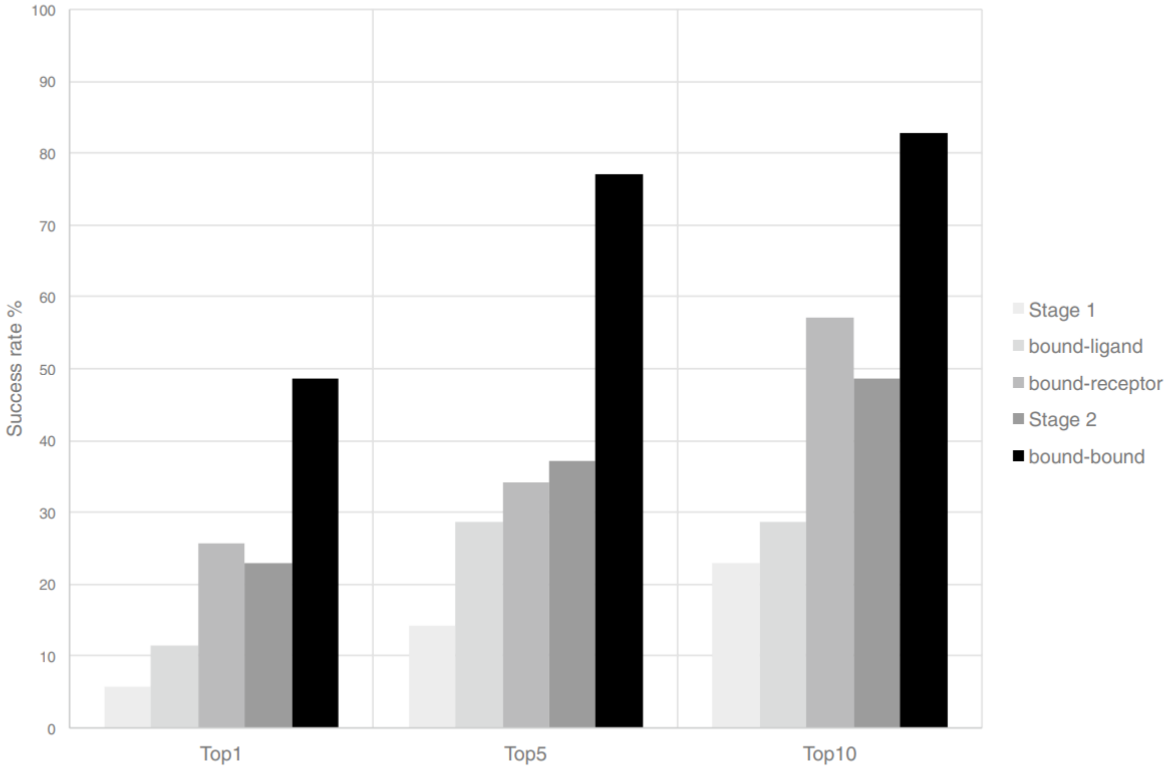


***Fig 3**: Successful prediction (l-RMSD<2.5Å) rates for top1, top5 and top10 in different docking runs for 35 targets. Bound-ligand docking refers to runs with bound ligand conformer and the ensemble of receptors used in Stage1. Bound-receptor is the one with bound receptor and the ensemble of ligands used in Stage1. Finally, bound-bound is the bound receptor-bound ligand docking runs.*

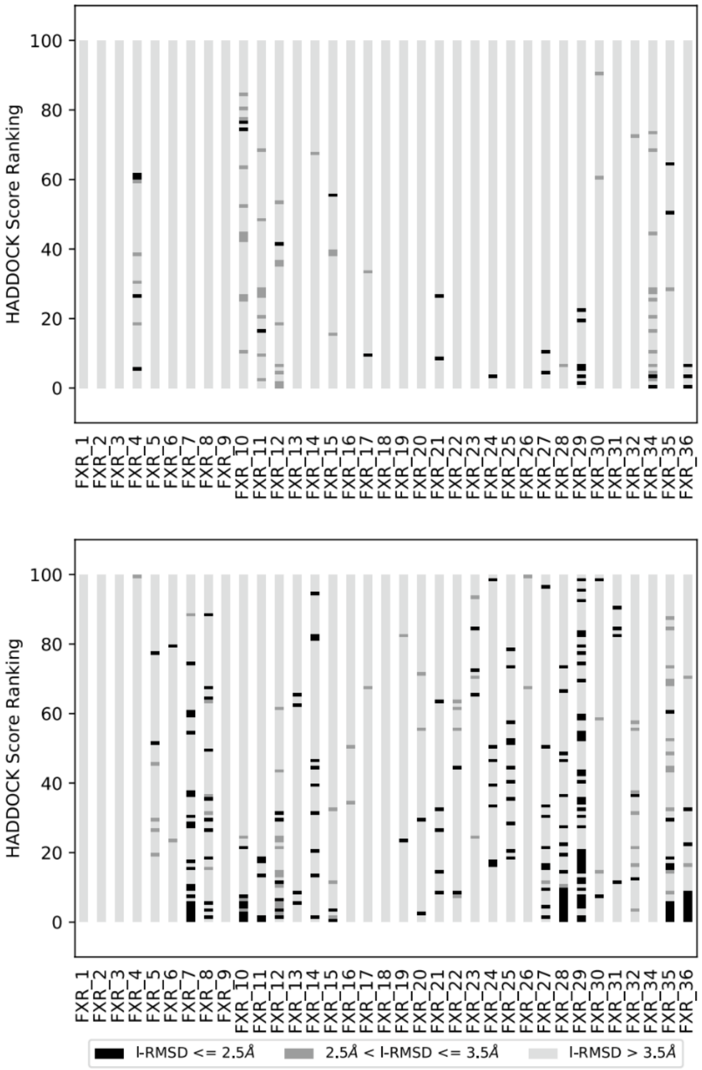


***Fig 4**: Comparison of the top100 models for the protocols used for stages 1 and 2. Each bar corresponds to structures belonging to runs for the indicated target. The coloring of the bars separates the structures in 3 classes. Structures colored black have a l-RMSD smaller than 2.5Å, structures colored dark gray have a l-RMSD between 2.5 and 3.5Å and structures with a l-RMSD of greater than 3.5Å are colored light gray. The top-ranked structures are the ones close to zero on the y-axis.*

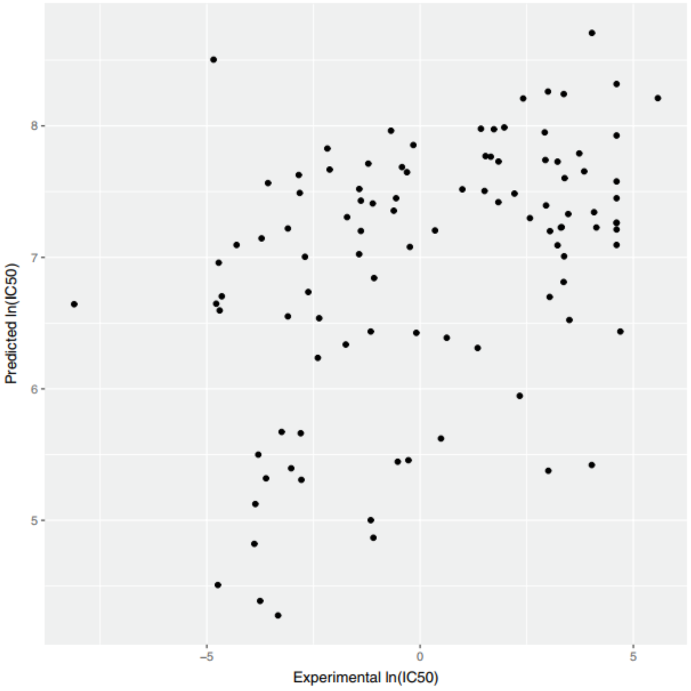


***Fig 5**: Comparison of predicted ln(IC50) with experimental ln(IC50) using our ligand-based binding affinity predictor.*

Binding Affinity

Ligand-Based Binding Affinity Prediction

A Support Vector Regression model based on ligand similarity using Atom Pair (see Material and Methods) was used for ligand-based prediction of the binding affinities. The Kendall's Tau between the ranking of the experimental and our predicted binding affinities is 0.27, which is the third best performance out of five participants. The correlation between the two sets can be visualized in Fig. 5.

Although this method does not perform as well as our structure-based predictor (see below) it has as major advantage that it does not require a structural model and is therefore extremely fast.

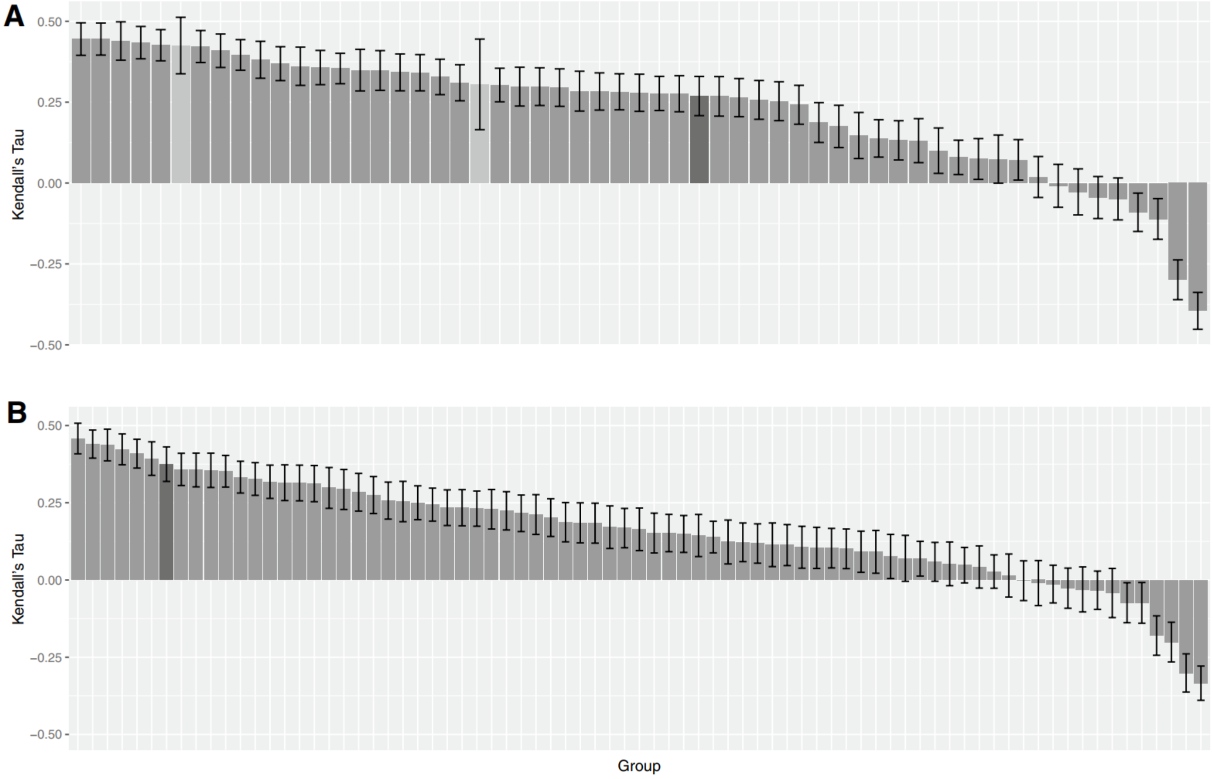


Fig 6: Ranking of binding affinity correlation per group for stages 1 and 2. The top panel reports the results of Stage1 and the bottom one of Stage2. Bars colored light gray correspond to groups which did not provide submissions for all targets. The bars colored dark gray correspond to the HADDOCK group submission.

Structure-Based Binding Affinity Prediction

The correlation scores (Kendall's Tau) of the binding affinity rankings calculated for stages 1 and 2, for all groups are summarized in Fig. 6. We clearly performed better in Stage2 with a correlation of 0.37 against 0.27 in Stage1, where we used only HADDOCK scores for ranking. In terms of Pearson's Correlation coefficient between the predicted scores and the experimental binding affinity, our prediction performance improved from 0.40 in Stage 1 to 0.51 in Stage 2 with the structure-based predictor (see Online Resource – Fig. S2). Interestingly, averaging the

ΔG_{score} over the top 10 models resulted in a correlation of 0.37 while using only the top scoring model yielded 0.28. Considering that our top 10 poses are rather heterogeneous in their conformations, our binding affinity predictor seems rather robust and not too sensitive to the exact conformation of the ligand. Further investigations are needed to further dissect those results and investigate the impact of energetics and the quality of the models on the ranking performance.

We also investigated the potential of our ranking predictor for identification of lead compounds. We defined as true positive the targets which are within the top N ranked compounds of both the predicted and experimental binding affinity rankings (N: 1,2…,102). Then, we calculated the positive predictive value (PPV), which is equal to the number of true positives divided by the number of predicted positives (top N ranked targets according to BA predictor). We plotted PPV as a function of N together with the diagonal which represents a random prediction (RP) (Fig.7). We also report the enrichment factor (PPV/RP) on the top of Fig.7. This analysis indicates that our predictor reaches a 2.5-fold improvement in correct identification of effective ligands in the top 20-25% compared to random.

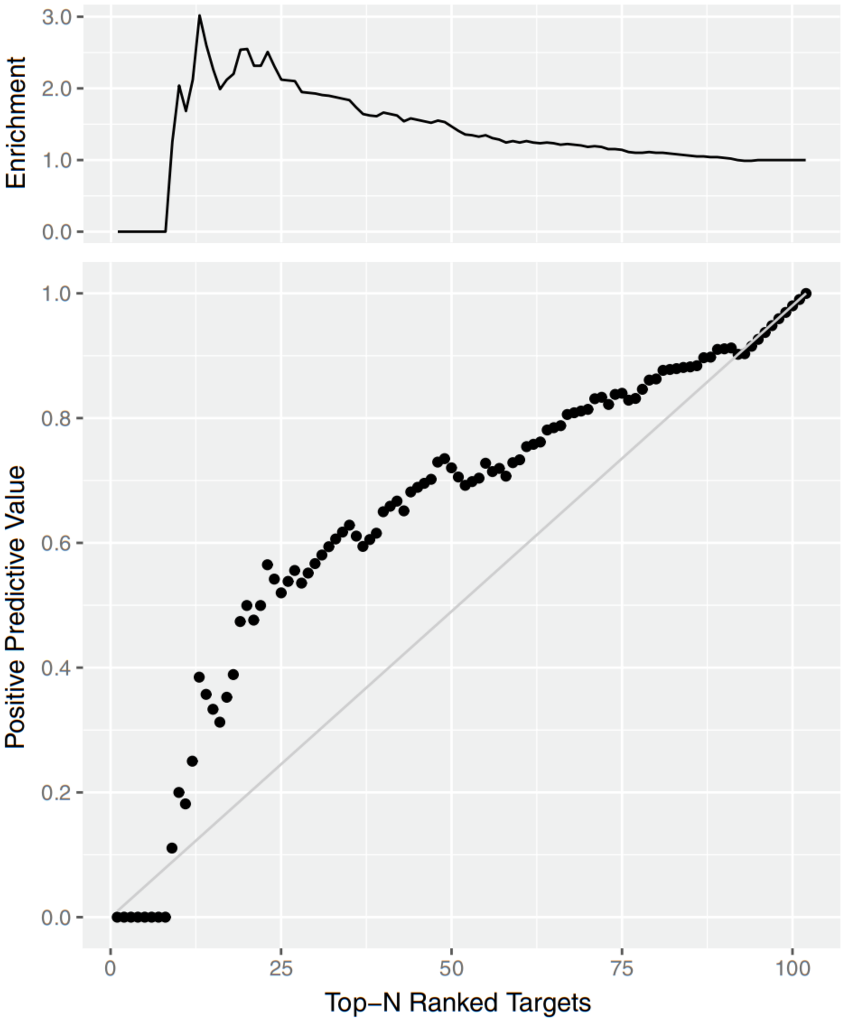


Fig 7: Positive predictive value (bottom) and enrichment factor (top) for 102 targets, using structure-based binding affinity predictor. Taking top 20-25% is associated with 2.5 enrichment factor.

Conclusions

Our participation in the D3R Grand Challenge 2 was an opportunity to evaluate and revisit our docking and ranking protocols. Our pose prediction performance in Stage1 was far from optimum, which led us to investigate the effect of ligand/protein conformer selection on the docked model quality. We identified the conformation of the receptor as main limiting factor, which led us to select receptor conformers for Stage2 based on ligand similarity, which significantly improved our pose prediction performance. This, together with a biasing of the major cluster for ligand conformers as explained in 'Revised protocol' increased our overall prediction success.

As for ranking in Stage2, we developed two different BA predictors: A ligand-based one and structure-based one. Our ligand-based predictor is computationally efficient since it does not require any 3D structural model for training. However, it does not perform as well as our structure-based predictor (Kendall's tau is 0.27 and 0.37 for ligand and structure-based, respectively). Using the structure-based predictor, which considers the number and type of interatomic contacts, for affinity ranking dramatically improved our overall performance for binding affinity prediction, with our ranking compared to the other submitted methods improving from 32nd/57 for Stage1 to 7th/77 for Stage2 (and if only considering a single submission per group per category, from 18th/27 (Stage 1) to 5th/25 (Stage 2) among all groups participating to the challenge).

As final observation, it is worth noting that our ranking was based on the average score calculated over the top 10 poses (which are heterogeneous in most cases, particularly with respect to the ligand orientation in the binding pocket – see Fig. 2). This averaging yielded better predictions than only using the top1 (Kendall's tau 0.37 and 0.28 for top10 and top1, respectively). This simple contact-based predictor seems to show promise as virtual screening tool to select a fraction of effective ligands, yielding an enrichment factor of about 2.5 for the top 25% of compounds compared to a random selection.

Acknowledgements

This work was supported by the European H2020 e-Infrastructure grants West-Life (grant no. 675858), BioExcel (grant no. 675728) and INDIGO-Datacloud (grant no 653549 and from the Dutch Foundation for Scientific Research (NWO) (TOP-PUNT grant 718.015.001). CG acknowledges financial support from the China Scholarship Council (grant no. 201406220132). LX acknowledges financial support from the Dutch Foundation for Scientific Research (Veni grant 722.014.005). AV and ISM acknowledges financial support from the European Union's H2020 Marie Skłodowska-Curie Individual Fellowships (grant no. BAP-659025 and MEMBRANEPROT-659826, respectively).

Supplementary Information

***Table S1**. HADDOCK parameters used for dockings*

Parameter	Setting
It0 sampling (rigid body docking)	10,000
It1 sampling (semi-flexible simulated annealing)	400
Delenph	False
Inter_rigid	0.001
Tadinit2_t	500
Tadfinal2_t	50
Tadinit3_t	500
Tadfinal3_t	50
Initiosteps	0
Cool1_steps	0
W_vdw_0	0
Protein interface residue list	269,273,274,277,287,288,291,292,294, 295,298,332,333,335,336,339,340,346,347, 352,356,359,361,365,369,370,373,451,454,458,469,473
amb = ExtStageConstants (firstit = 0, lastit = 0,)	

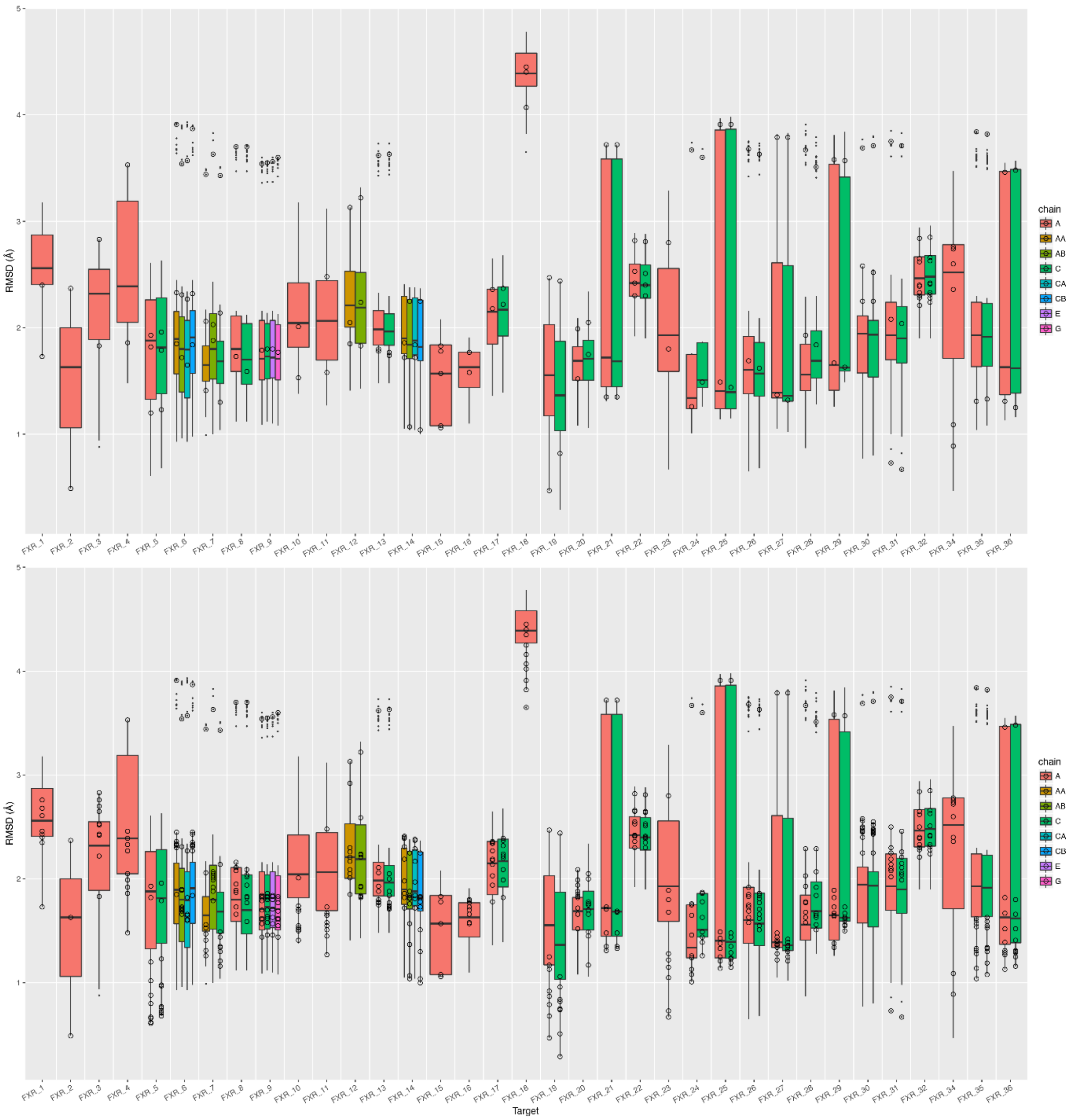
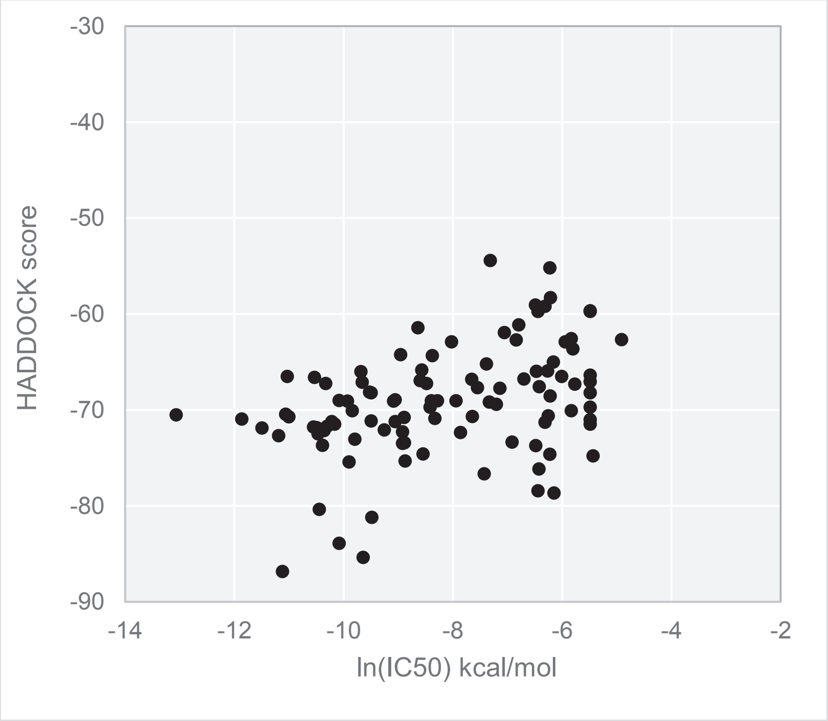


Figure S1: *RMSD values of the OMEGA generated ligand conformers against the reference structures. The boxplots are colored according to the chain ID of the reference chain used for the calculations. Circles indicate the RMSD values of the poses that were selected for docking. The top panel corresponds to the conformer selection for stage 1 and the bottom one to the conformer selection for stage 2.*



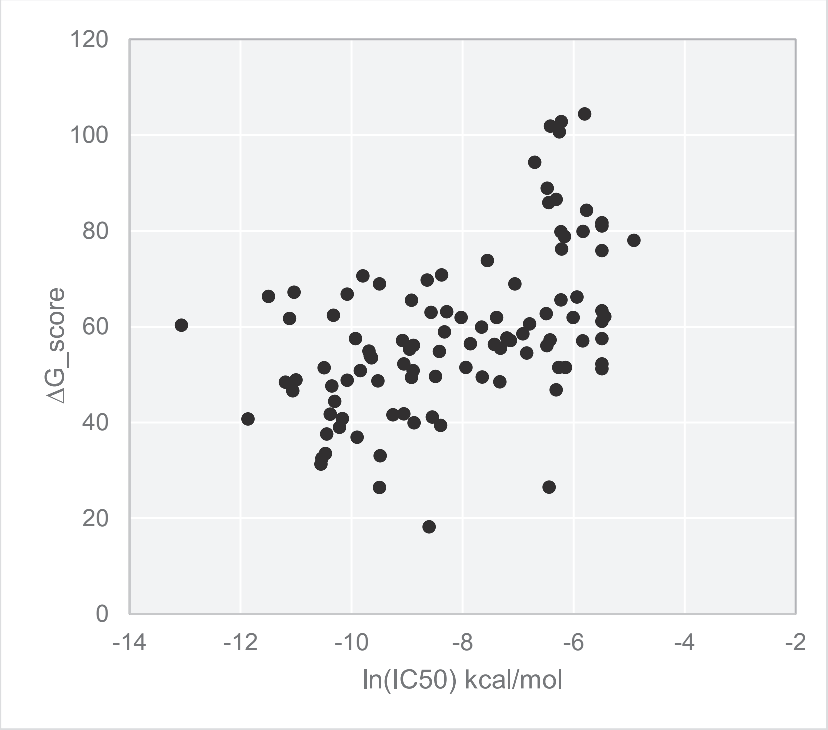


Figure S2: *Scatter plot between the HADDOCK score for Stage 1 (top panel) and ΔGscore for Stage 2 (bottom panel) versus the experimental binding affinities reported as ln(IC50). The corresponding Pearson's correlation coefficients are 0.40 and 0.51 for Stage 1 and Stage 2, respectively. The ΔG_scores have been calculated with our structure-based binding affinity predictor (see Eq. 2 in the main text), averaged over the top10 best models refined with the refinement interface of the HADDOCK2.2 web server.*

Chapter 5

Protein-ligand pose and affinity prediction. Lessons from D3R Grand Challenge 3.

Panagiotis I. Koukos, Li C. Xue, Alexandre M.J.J. Bonvin

Journal of Computer-Aided Molecular Design, 2019, volume 33, issue 1, pages 83-91

Abstract

We report the performance of HADDOCK in the 2018 iteration of the Grand Challenge organised by the D3R consortium. Building on the findings of our participation in last year's challenge, we significantly improved our pose prediction protocol which resulted in a mean RMSD for the top scoring pose of 3.04 and 2.67Å for the cross-docking and self-docking experiments respectively, which corresponds to an overall success rate of 63% and 71% when considering the top1 and top5 models respectively. This performance ranks HADDOCK as the 6th and 3rd best performing group (excluding multiple submissions from a same group) out of a total of 44 and 47 submissions respectively.

Our ligand-based binding affinity predictor is the 3rd best predictor overall, behind only the two leading structure-based implementations, and the best ligand-based one with a Kendall's Tau correlation of 0.36 for the Cathepsin challenge. It also performed well in the classification part of the Kinase challenges, with Matthews Correlation Coefficients of 0.49 (ranked 1st), 0.39 (ranked 4th) and 0.21 (ranked 4th) for the JAK2, vEGFR2 and p38a targets respectively. Through our participation in last year's competition we came to the conclusion that template selection is of critical importance for the successful outcome of the docking. This year we have made improvements in two additional areas of importance: Ligand conformer selection and initial positioning, which have been key to our excellent pose prediction performance this year.

Introduction

The D3R (**D**rug **D**esign **D**ata **R**esource) Grand Challenge of 2018 is the third iteration of the major docking competition organised by the D3R consortium [242,243] and similarly to previous years, it has two goals. The first, is the assessment of the ability of docking algorithms to accurately predict the binding poses of a protein against a diverse set of small molecules, and the second, the evaluation of the performance of binding affinity prediction algorithms.

The protein which is the focus of the pose prediction assessment is Cathepsin S – a member of the Cathepsin family. Cathepsins are proteases that are classified in three groups depending on the makeup of their catalytic site, with Cathepsin S being a member of the most populated group – cysteine proteases [365]. Its involvement in MHC class II antigen presentation is well established. Given that role, it should come as no surprise that it has been implicated in many pathological conditions such as cancer and diabetes. More recently it has been investigated for its role in pain perception [366] and cardiovascular and kidney [367] disease. It has long held an interest for the pharmaceutical industry [368] as evidenced by the plethora (more than 50 at time of writing) of human Cathepsin S structures with a bound ligand, that have been deposited in the Protein Data Bank (PDB) [252] over a time period that spans 15 years.

In addition to the Cathepsin S-centric assessment, which also includes a binding affinity prediction component, binding affinity prediction approaches are evaluated in 4 subchallenges that focus on kinases. Kinases catalyse the process of phosphorylation through which a phosphate group is covalently bound to a protein substrate. Their role in cell signalling has been well understood for decades and they are involved in many aspects of cell differentiation and growth [369]. They are a primary target for cancer-related drug development [370].

Through our participation in last year's GC [371] we came to the conclusion that template selection is of critical importance for the successful outcome of the docking. This year we have made improvements in two additional areas of importance: Ligand conformer selection and initial positioning. The impact of this is reflected in our improved performance in GC3, the results of which are presented and discussed here.

Materials and Methods

HADDOCK (**H**igh **A**mbiguity **D**riven **DOCK**ing) is our information-driven docking platform [59,256]. For an introduction to HADDOCK and small molecule docking please review the contribution we made to last year's special issue on the D3R Grand Challenge (GC) [371]. The main conclusion from our participation in last year's competition was that protein template selection is of crucial importance for the successful outcome of the docking. We used the protocol we came up with last year to select protein templates for this year's competition as well. We made improvements to the ligand conformer selection and placement protocols. Similar to last year, all new and untested parts of the protocol were benchmarked on existing protein-ligand complexes extracted from the PDB.

In a departure from previous years, this year's competition is further divided in 5 subchallenges. Subchallenge 1 is the equivalent of the GC of previous competitions and has a pose and binding affinity prediction component. Subchallenges 2 – 5 only have a binding affinity component. We participated in subchallenges 1 and 2.

Subchallenge 1

This challenge focused on Cathepsin S. For the first part of the challenge – pose prediction – we had to predict the binding pose of Cathepsin S against a set of 24 small molecules that were known to bind to it. There is a cross-docking stage, during which the structures of the target proteins are not known and a self-docking stage for which the bound protein structures – but not those of the compounds – are known. The organisers provided us initially with SMILES strings for the small molecules and the FASTA sequence of the protein, and for the self-docking stage with the coordinates of the bound receptor for each ligand. Additionally, two publicly available structures of the protein with a dimethylsulfoxide (DMSO) molecule and a sulfate ion (SO_4) placed in the binding pocket were circulated to the participants because the aforementioned molecules were detected in some of the crystal structures. For the binding affinity prediction component of the challenge we had to rank the binding affinities of 136 compounds against the protein.

Protein template selection

This part of the protocol, as well as the reasoning behind it, are described in greater detail in our previous work and so will only be covered briefly. Using the provided FASTA sequence, we identified structures of Cathepsin S that had been deposited in the PDB. We filtered the results and kept only those structures where the protein was complexed with a non-covalently bound ligand, thus identifying 36 templates. We then proceeded to compare the crystallographic ligands to the target compounds using as a similarity measure the Tanimoto distance, as implemented in the fmcsR and chemmineR packages [351,352]. In this way, we selected one protein template for each of the 24 target compounds, by identifying the template with the highest similarity ligand. The similarities of the crystallographic ligands to the prediction set compounds are shown in S.I. Fig. 1.

For the self-docking challenge, we used the provided crystallographic structures retaining crystallographic waters and DMSO (target 14) or sulphate (targets 2, 17, 20, 22, 24 and 24) molecules.

Ligand preparation

Three-dimensional (3D) conformations of the ligands were generated with OpenEye OMEGA (v20170613) [342] using the SMILES strings as input. For every molecule, we sampled up to 500 conformers. We used the TanimotoCombo metric, as implemented in OpenEye ROCS [372], to compare the generated conformers to their respective crystallographic ligand in the identified templates (see "Protein template selection"). The TanimotoCombo metric combines shape and chemical similarity and allows us to select the conformers whose shape and chemical features resemble that of the crystallographic ligands. The top 10 scoring conformers were selected for ensemble docking. Each conformer was superimposed onto the crystallographic ligand in the template using the shape toolkit of the OpenEye suite.

This protocol was benchmarked with existing Cathepsin S-ligand structures identified in the PDB. This allowed us to evaluate the impact our choices had on the quality of our poses. We used four Cathepsin S structures (PDBids: 3IEJ, 3KWN, 3MPE, 3MPF) [373–375] and their respective ligands. After selecting the protein template based on the protocol described in "Protein template selection", we selected the ligand conformers by their TanimotoCombo score and after superimposing them to the site of the crystallographic ligand, proceeded to refine them (see "Docking" below).

For the self-docking challenge, we superimposed the protein template identified during the cross-docking challenge on the prediction set crystallographic structure. That allows us to superpose the generated conformers on the crystallographic ligand which is situated in the active site of the prediction set crystallographic structure because of the first superposition.

Docking

We refined the ensemble of ligand conformations superimposed on their respective protein templates using the water refinement protocol of HADDOCK. All hydrogen atoms were kept (by default HADDOCK removes the non-polar hydrogens to save computing time). Since the ligand conformations were selected based on their similarity to the closest identified template (see above) and superimposed onto the ligand in the selected template, no exhaustive search was performed. Instead the initial poses were only subjected to a short energy minimization in which only interface residues were treated as flexible, followed by the explicit water refinement stage of HADDOCK. For this the system is solvated using an 8Å shell of TIP3P [257] water molecules. The water refinement protocol consists of a first heating phase (100 MD integration steps at 100, 200, and 300K) with weak position restraints on all atoms except those which belong to the side-chain of residues at the interface. The interface is defined as the set of residues whose atoms are within 5 Å of any atom of any binding partner. The second MD phase consists of 2500 integration steps at 300K with positional restraints on all non-Hydrogen atoms excluding the interface residues. The number of MD steps was doubled compared to HADDOCK's default value (1250) because this yielded higher quality structures during our benchmarking with the four PDB structures described in "Ligand Preparation". The last cooling phase, consists of 500 integration steps at 300, 200 and 100 K, respectively, during which positional restraints are only used for the backbone atoms of the non-interface residues. A 2fs time-step is used throughout the protocol for the integration of equation of motions. The number of water refined models was set to 200. We also modified the default HADDOCK scoring function for the refinement stage by halving the weight of the electrostatic energy term:

$$HADDOCK_{score} = 1.0 \times E_{vdw} + 0.1 \times E_{elec} + 1.0 \times E_{desolv} + 0.1 \times E_{AIR}$$

This adjustment was motivated by internal benchmarking our group has performed on small molecule-protein complexes (data not shown). This scoring function is used to rank the generated models. The various terms are the intermolecular van der Waals (E_{vdw}) and electrostatic (E_{elec}) energies calculated with the OPLS force field and an 8.5Å non-bonded cutoff [259], an empirical desolvation potential (E_{desolv}) [260] and the ambiguous interaction restraints energy (E_{AIR}) [256]. Note that in this case, since only refinement was performed without any restraints to drive the docking, E_{AIR} is effectively 0.

For the self-docking challenge, we follow the same protocol as for the cross-docking one, keeping all crystallographic waters and fixing the conformation of the protein, with the additional change of instructing HADDOCK to write PDB files containing the solvent molecules (water) present during the refinement stage.

Binding affinity

The binding affinity predictions are evaluated in two stages. The first stage takes place before the structures of the complexes (protein and ligand) are released by the organisers, which means that either only ligand information is used, or models of the complexes, and the second after, which allows participants to make use of the newly available structural information.

For the first stage, we submitted both ligand-based and structure-based rankings and for the second only a structure-based one. Both approaches are described in detail in our previous D3R paper [371]. In short, the structure-based approach consists of the PRODIGY [360] method adapted for small molecules and trained on the 2P2I dataset [338] which makes use of the following function to score protein-ligand complexes by binding affinity:

$$\Delta G_{score} = 0.343794 * E_{elec} - 0.037597 * AC_{CC} + 0.138738 * AC_{NN} + 0.160043 * AC_{OO} - 3.088861 * AC_{XX} + 187.011384$$

Where E_{elec} is the intermolecular electrostatic energy calculated by the water refinement protocol of HADDOCK (see “Docking”) and AC_{CC}, AC_{NN}, AC_{OO} and AC_{XX} are the counts of atomic contacts between Carbon-Carbon, Nitrogen-Nitrogen, Oxygen-Oxygen and all other atoms and polar Hydrogens between the protein and the ligand, within a distance cut-off of 10.5Å. We used the mean ΔG_{score} of the top 10 models of the water refinement (see “Docking”) to rank the compounds.

The ligand-based approach rests on the hypothesis that similar ligands complexed to our proteins of interest should have similar binding affinities. Using the BindingDB database [356] we identified 1839 compounds bound to Cathepsin S with IC50 values. We calculated the similarity of the prediction set to the training set using the Atom Pair (AP) measurement as a similarity measure. The similarity matrices of the BindingDB set were used to train a Support Vector Regression (SVR) model with the libSVM library for MatLab [357] that was, in turn, used to predict the binding affinities of the prediction set.

Analysis

Fitting and RMSD calculations for generating the figures were performed using the McLachlan algorithm [345] as implemented in the program ProFit (http://www.bioinf.org.uk/software/profit/) from the SBGrid distribution [262].

Subchallenge 2

Subchallenge 2 only had a binding affinity component. The participants had to predict binding affinities for three protein targets – the kinases vEGFR2, JAK2-SC2 and p38-α – and sets of 85, 89 and 72 compounds respectively. Some of the compounds were shared between the three targets. The organisers provided SMILES strings for all compounds along with FASTA sequences of the proteins.

For this challenge, we only submitted ligand-based binding affinity rankings. The method is the same as the one described in the “Binding affinity” section for Subchallenge 1. The only difference was the training data availability. Using BindingDB we identified 7049, 4582 and 4563 compounds with IC50 binding affinity measurements for the vEGFR2, JAK2-SC2 and p38a kinases respectively.

After the binding affinity rankings were released by the organisers, it quickly became apparent that for all three targets, the compounds could be classified into binding and non-binding sets since most compounds had the maximum detectable binding affinity of 10μM. This prompted the organisers to alter the way the challenge would be evaluated into a classification and regression problem, where the identification of the binding set (compounds with a Kd < 10μM) would be treated as a classification problem and the ranking of the binding compounds by binding affinity as a regression problem.

Results and Discussion

Subchallenge 1

Pose prediction

The binding pose prediction was evaluated for the cross- and self-docking experiments. Our performance in the cross-docking experiment in terms of RMSD of the five submitted poses is shown in Fig. 1.

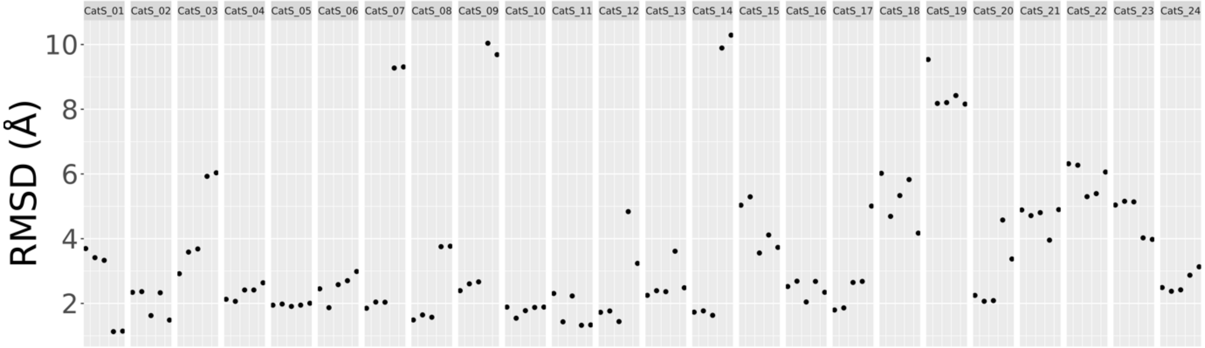


***Fig 1**: Heavy-atom RMSD values of the cross-docking models from the reference structures. Every point corresponds to one model with 5 models per target. The models are ranked by HADDOCK score with the highest scoring ones being on the left of every block.*

This analysis was carried out by superposing the interface areas of the models and their respective reference structures and calculating the Heavy-atom RMSD (excluding any halogen atoms) of the compounds. The mean RMSD values across all models and targets for this experiment are 3.04 ± 2.03 Å, whereas for the self-docking experiment, the values improved to 2.67 ± 1.63 Å. Fig. 2 highlights some of our top predictions.

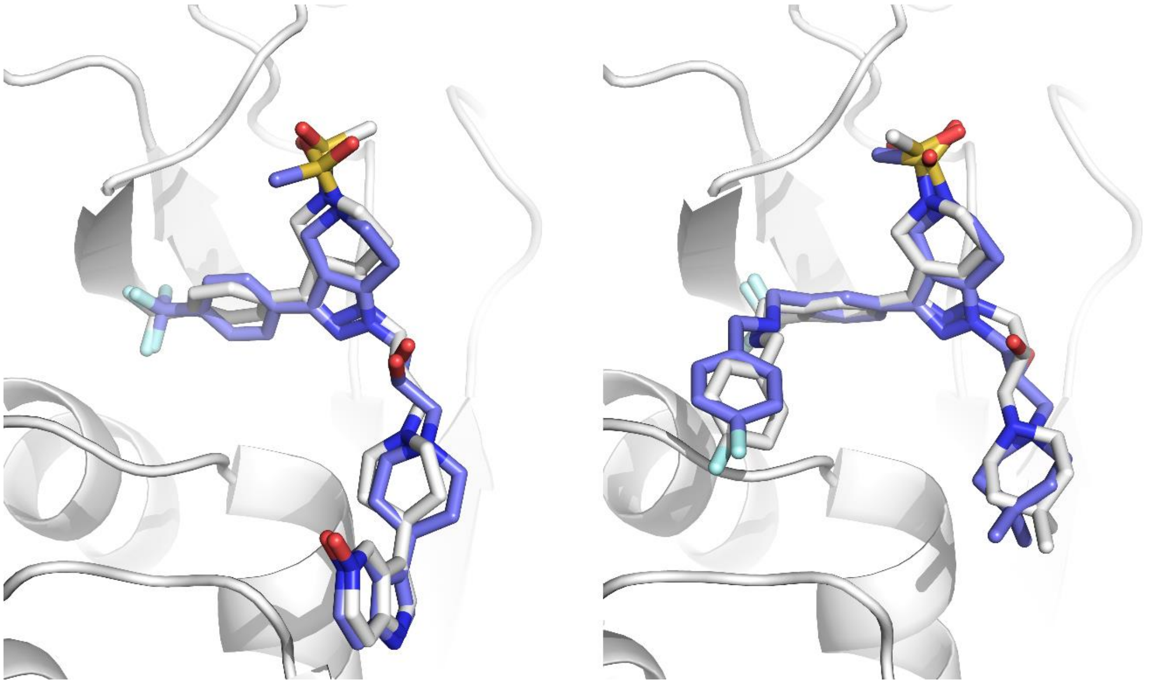


***Fig 2**: Superpositions of HADDOCK models on reference structures. Left: Model 5 from target 1 (1.1Å). Right: Model 1 from target 8 (1.5Å). The reference protein structure is shown in cartoon representation in white. The compounds are shown in stick representation in white and blue for the reference and model molecules respectively. Figure created with PyMOL* [83].

At least one of the 5 models submitted was of acceptable quality (RMSD <= 2.5Å) in 17 of the 24 targets (71% success rate top5). Our scoring function is thus able to correctly rank the near-native solutions near the top as can be seen in S.I. Fig. 2. If one considers only the top-ranked pose, the performance remains impressive with 15 out of 24 targets with an acceptable quality model (63% success rate top1). Fig 3 shows the difference between the top and bottom ranked models for target 7. Despite these excellent results, there is still room for improvement, especially in scoring: If we only consider the targets for which we generated at least one acceptable model (17 out of 24), the top-scoring pose corresponds to the best pose in 5 of the 17 targets (29%). For the remaining 12 targets, the average difference between the top scoring and best poses is 0.55 ± 0.71Å and 0.45 ± 0.61Å for the cross- and self-docking experiments respectively.

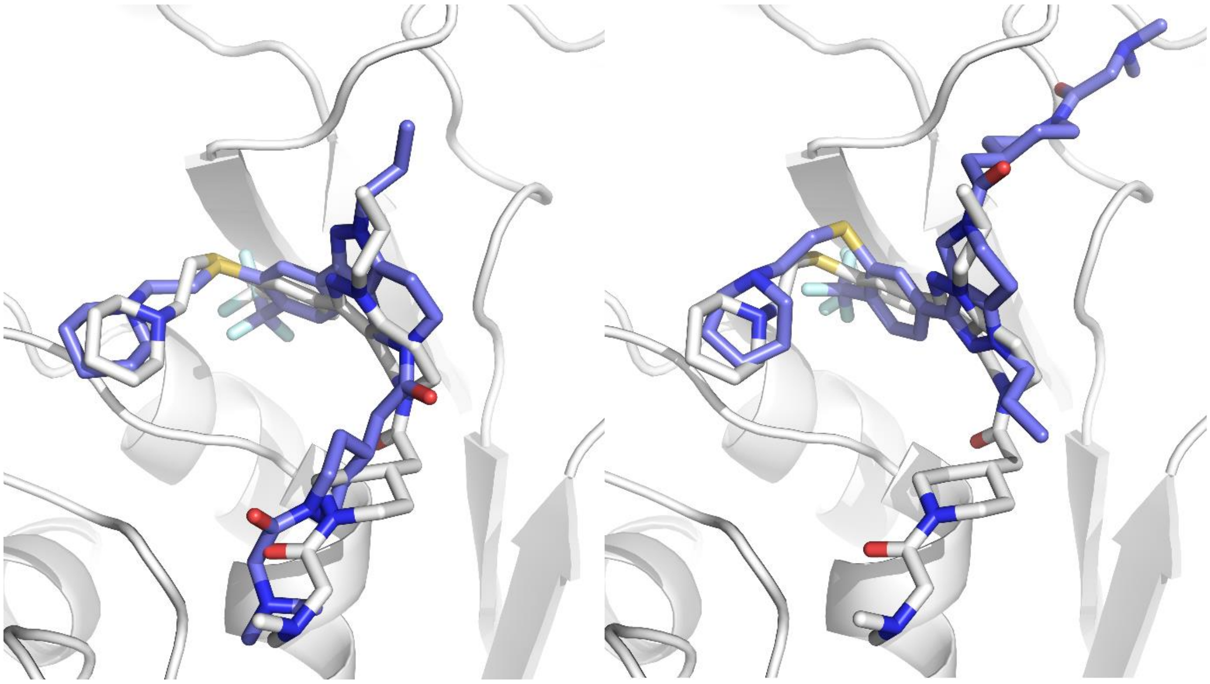


***Fig 3**: Superpositions of HADDOCK models on reference structures. Left: Model 1 from target 7 (1.85Å). Right: Model 5 from same target (9.31Å). Our scoring function can distinguish the near native model from the wrong one. The difference between the two molecules is a single torsional angle that has been rotated ~180o. Figure created with PyMOL.*

The performance of HADDOCK relative to the other participants for both experiments can be seen in Fig. 4. Note that if we would only consider one submission (the best) per group our rank would be 6th for the cross-docking experiment (top panel in Fig. 4). Our performance in the two experiments (cross- vs self-docking) is broken down by target in Fig. 5, revealing that our protocol is not very sensitive to the starting template. In most cases only rather small improvements in terms of RMSD are obtained when starting from the bound receptor including water. The single target for which we observe a significant deviation in the self-docking results compared to the cross-docking ones is the first one (see Fig. 5). The average RMSD for that target is 2.54 ± 1.29Å and 4.13 ± 3.46Å for cross- and self-docking experiments respectively. Model 5 of the self-docking experiment submission is mostly responsible for this significant change, since its RMSD is greater than 10. This is a repetition of what is shown in Fig. 3, with one of the models (model 5 in both cases) which has a torsional angle that is rotated by 180 degrees compared to the rest of the submitted models and the reference structure.

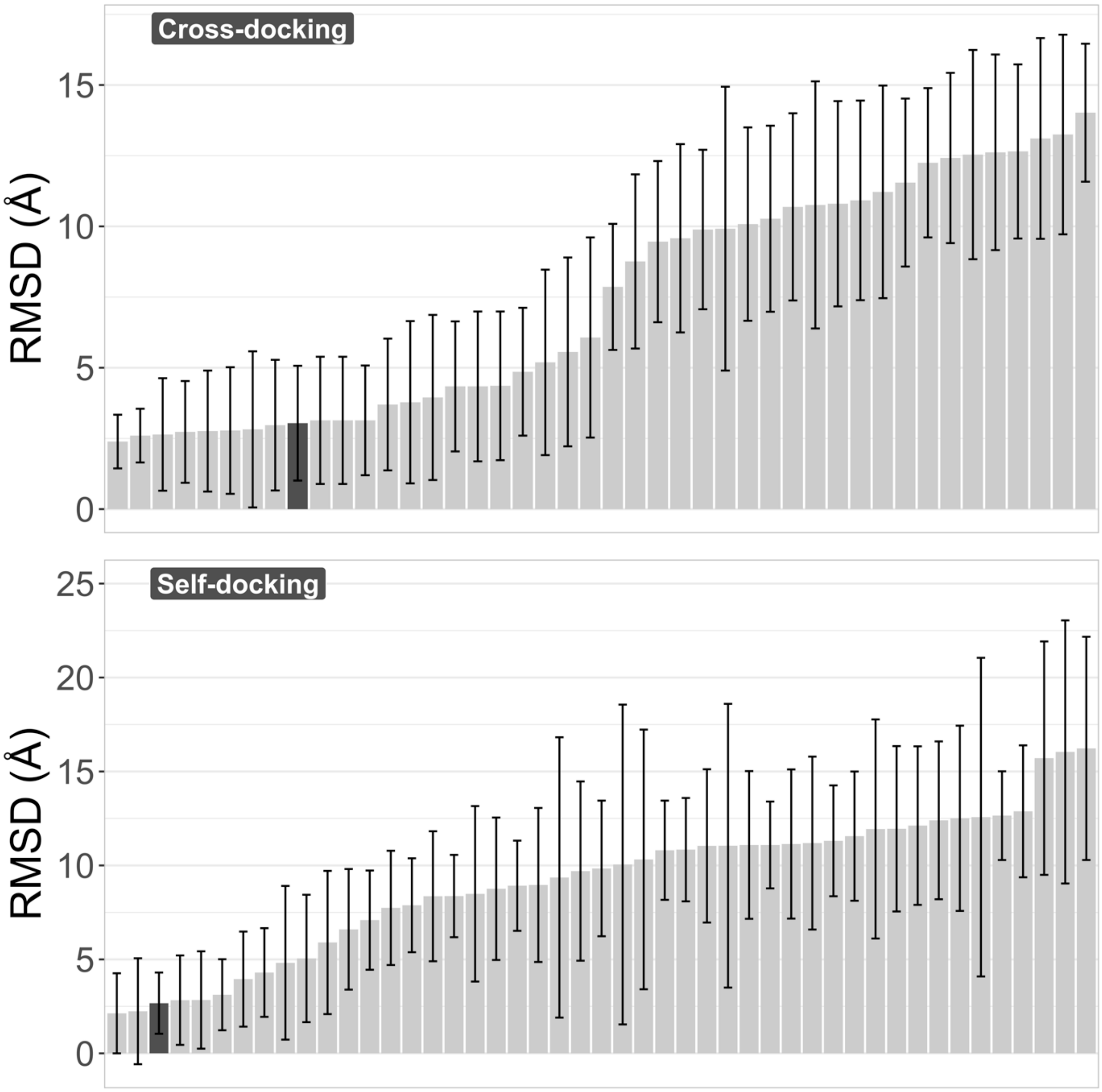


***Fig 4**: Heavy-atom RMSD values averaged over all models and all targets. Top: Cross-docking experiment. Bottom: Self-docking experiment. Every bar corresponds to a single submission. The error bars indicate the standard deviation of the mean RMSD. HADDOCK submission is represented by the dark-grey bar in both panels.*

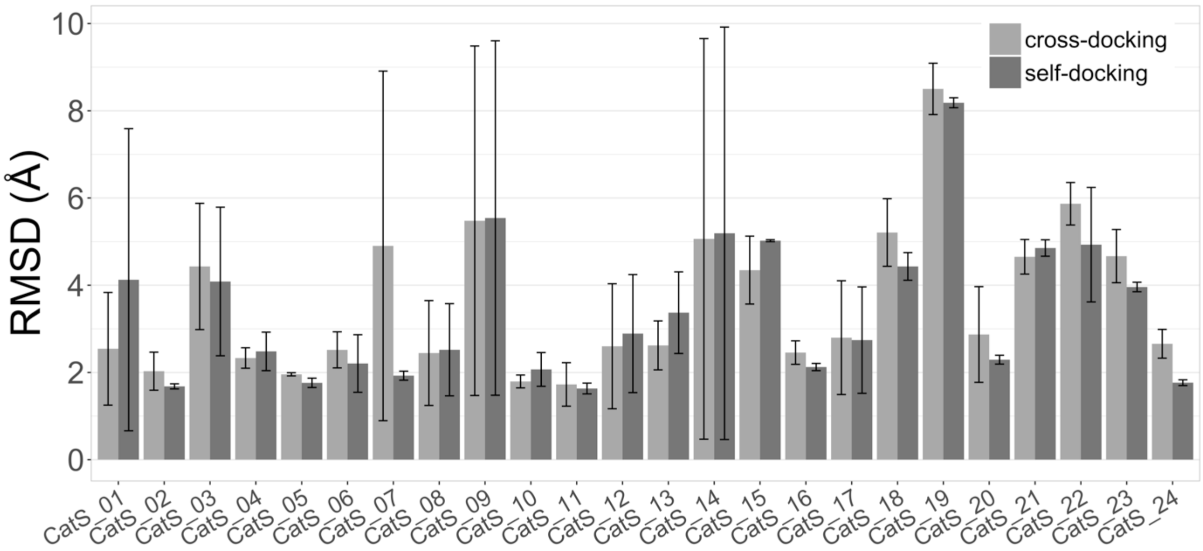


***Fig 5**: Comparison between the performance of HADDOCK in the cross-docking and self-docking stages. Every set of bars corresponds to the average heavy-atom RMSD of all 5 models for a target, with the light- and dark-grey coloured bars corresponding to the cross- and self-docking experiments respectively. The error bars indicate the standard deviation of the mean RMSD.*

Binding affinity prediction

Binding affinity predictions were performed in two stages – one before the organisers released the poses to the participants and one after. We participated in stage 1 with both ligand-based and structure-based approaches, while for stage 2 we only submitted a structure-based ranking. Fig. 6 shows our performance compared to all participants.

These results were rather surprising: The structure-based approach which was one of the top performers in last year's competition failed to produce an accurate ranking of the compounds, while our ligand-based predictor now performs as one of the best (even if the quality of the prediction is still limited). There was also no improvement for the structure-based ranking between stages 1 and 2 in contrast to GC2 where we noticed a significant improvement when using the crystallographic poses for ranking the compounds. One explanation for this could be that, compared to last year, we already had better quality poses for most of the targets for stage 1. On the other hand, our simple machine learning-based ligand-based approach is not only the most accurate ligand-based approach with a Kendall's Tau of 0.36 but the third most accurate method for both stages, behind only the top performing structure-based approaches.

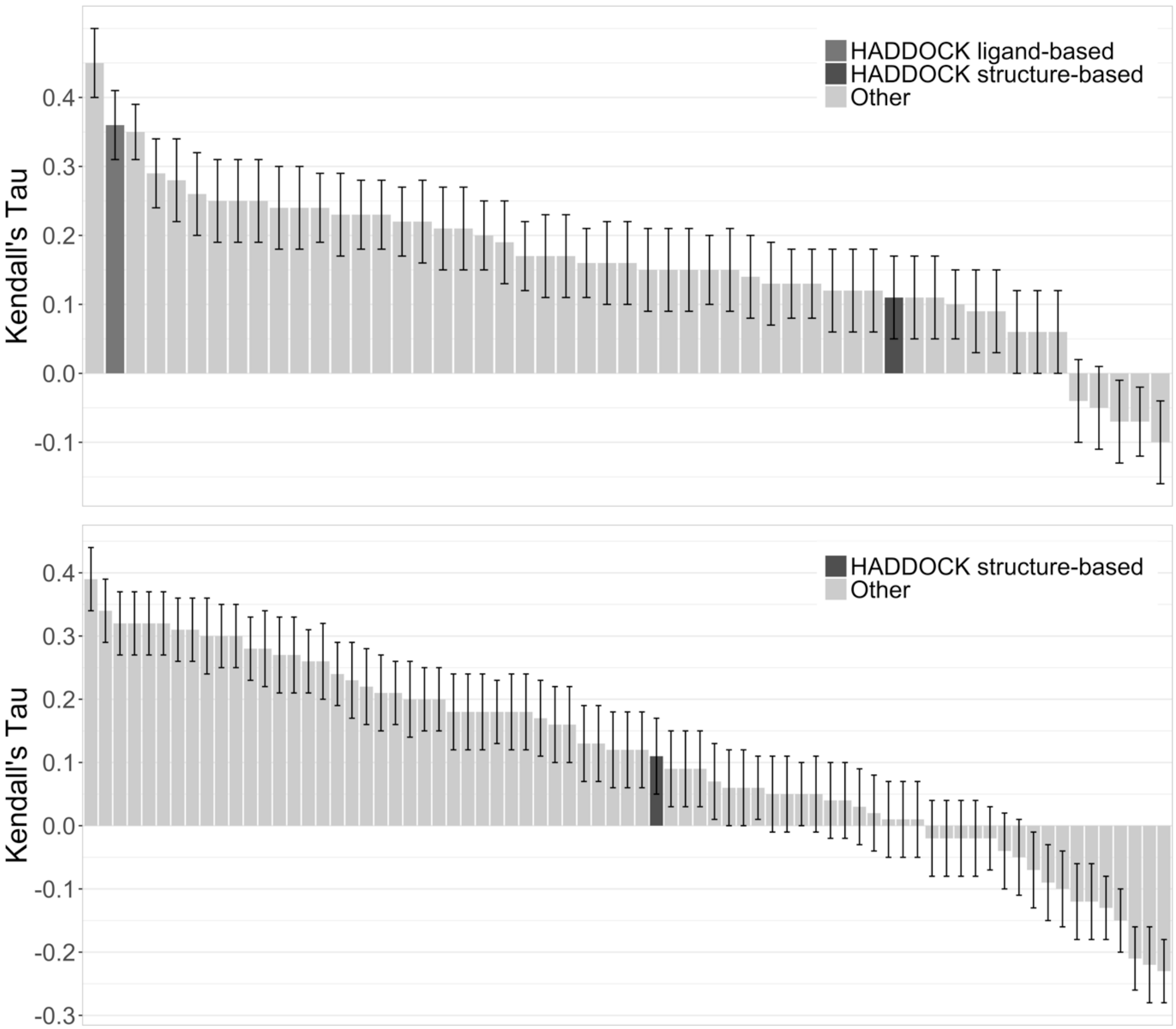


Fig 6*: Ranking of the binding affinity predictions for Cathepsin S by correlation. Top: Stage 1. Bottom: Stage 2. Every bar corresponds to one submission with our ligand-based submission having a medium and the structure-based one a dark grey colour in both panels.*

Subchallenge 2

This challenge revolved around kinase binding affinity prediction. As was mentioned in the Methods section, this is a regression-classification problem. The overall results can be seen in Fig. 7.

Despite the fact that our approach wasn't trained with classification in mind, the classification performance is better than that of the regression. Specifically, the Matthews Correlation Coefficient (MCC) values are 0.49, 0.39 and 0.21 respectively for JAK2-SC2, vEGFR2 and p38a (see S.I. fig. 3 for the classification rankings). The respective Kendall's Tau correlations are 0.15, 0.38 and 0.07. As is evident from the plot the two correlation metrics are not correlated. This means that an algorithm that accurately identifies the binders and non-binders does not necessarily rank the binders accurately. The performance differences cannot be accounted for by the difference in training set size, since we identified roughly the same number of compounds for JAK2-SC2 and p38a. Additionally, vEGFR2 had the biggest training set size but that is not translated into better performance for the classification or the regression.

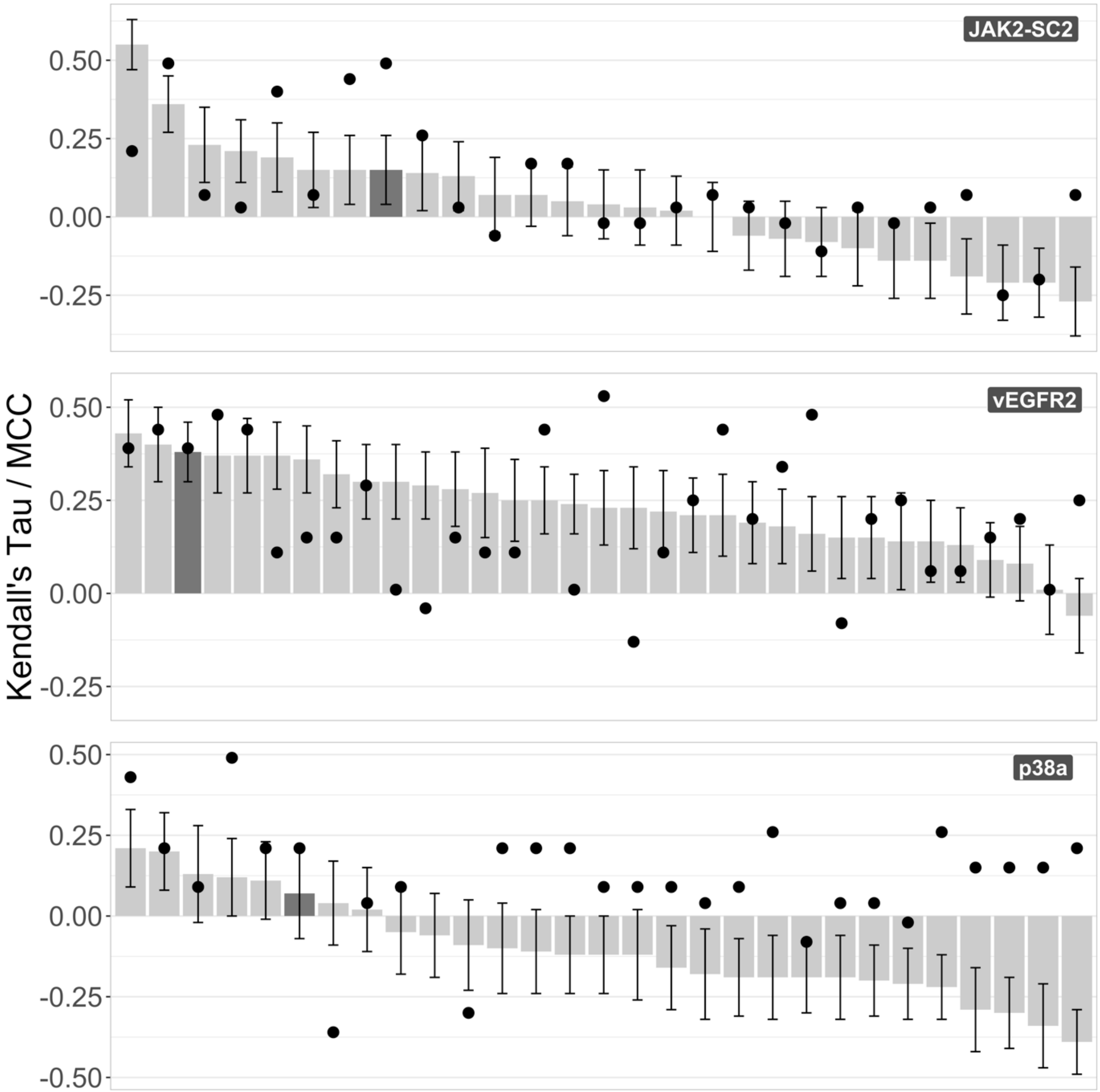


Fig 7: Binding affinity prediction correlation coefficients. Top: JAK2-SC2. Middle: vEGFR2. Bottom: p38a. The bars and the corresponding error bars represent the Kendall's Tau correlation between the binding affinity predictions and the binding set for every target. The black circles correspond to the Matthews Correlation Coefficient which was used to assess the accuracy of the classification of the compounds into binding and non-binding. The dark grey bars and their corresponding circles represent our submissions.

Conclusions

GC3 has allowed to implement the lessons that we learned by participating in GC2 and further experiment with additional optimisations. The conclusions that we can draw with regards to the pose prediction challenges are the following:

1. Selecting the protein templates accurately has the biggest effect on the outcome of the docking. By identifying templates that already have a ligand bound to them and selecting the one that is most similar to the prediction compounds, we are ensuring a protein binding interface that is highly compatible with the prediction compound. This removes the need for extensive sampling of the protein interface or ensemble docking. Moreover, this approach seems to be robust to low similarity (see S.I. Fig. 1) compounds. The majority of template ligands identified have a Tanimoto similarity of less than 0.6.
2. Selecting the ligand conformations. Identifying structures with existing compounds has the additional benefit that they can be used to select the compound structures to be used during docking. Generating 3D models of compounds from 2D information entails generating hundreds of conformers. By comparing the shape and chemical similarity of the conformers to existing compound structures we can reduce the number of conformers needed during docking and ensure the starting conformations are closer to the experimental structures.
3. Making use of the template information by positioning the conformers in the binding interface. This last observation is only relevant for molecular simulation codes that, like HADDOCK, randomise the relative orientation and position of the partners prior to docking. We can use shape similarity to position the ensemble of conformers at the binding site and bypass the first two stages of HADDOCK (rigid-body energy minimisation and flexible refinement by simulated annealing in torsion angle space) and directly refine the complexes using a longer version of our water-refinement protocol.

The applicability of our approach was demonstrated by its performance, with mean RMSD values of 3.04Å and 2.67Å for the cross-docking and self-docking experiments respectively. Our overall success rate when considering the top1 and top5 poses is 63% and 71%, respectively. These results place us as the 6th and 3rd best performers for the two challenges respectively.

The binding affinity experiments present a greater challenge to the community as whole. Despite our competitive rankings in the classification as well as the regression challenges, it appears that reliable binding affinity predictors are still not within grasp. This holds true for both ligand and structure-based approaches. However, the surprisingly good classification results (especially given that the algorithm was optimised for regression rather than classification problems) make us optimistic that this can be improved in the future.

Acknowledgments

This work was supported by the European H2020 e-Infrastructure grants West-Life (grant no. 675858), BioExcel (grant no. 675728) and from the Dutch Foundation for Scientific Research (NWO) (TOP-PUNT grant 718.015.001, Veni grant 722.014.005). The authors declare no competing financial interests. The authors thanks Adrien Melquiond, Irina Moreira, and Mikael Trellet for useful discussions.

Additional Information

The data and code used to train the ligand-based binding affinity predictor and rank the compounds are freely available on GitHub, together with our in-house scripts developed during our participation in the last two Grand Challenge competitions. These can be accessed at following URL: https://github.com/haddocking/D3R-tools.

Supplementary Information

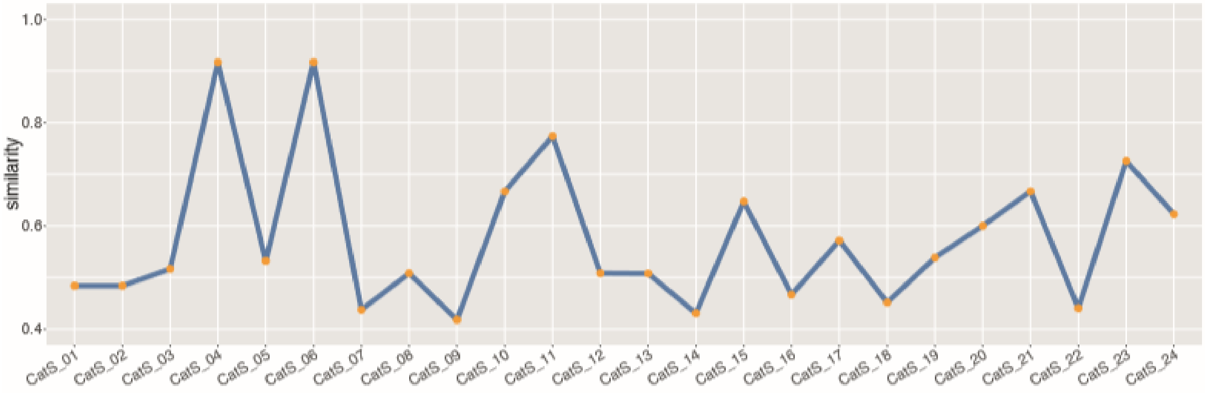


Fig S1: *Tanimoto Similarity of the prediction set compounds to the most similar crystallographic ligand of the ones identified in the PDB. A value of 1 indicates perfect similarity and a value of 0 perfect dissimilarity.*

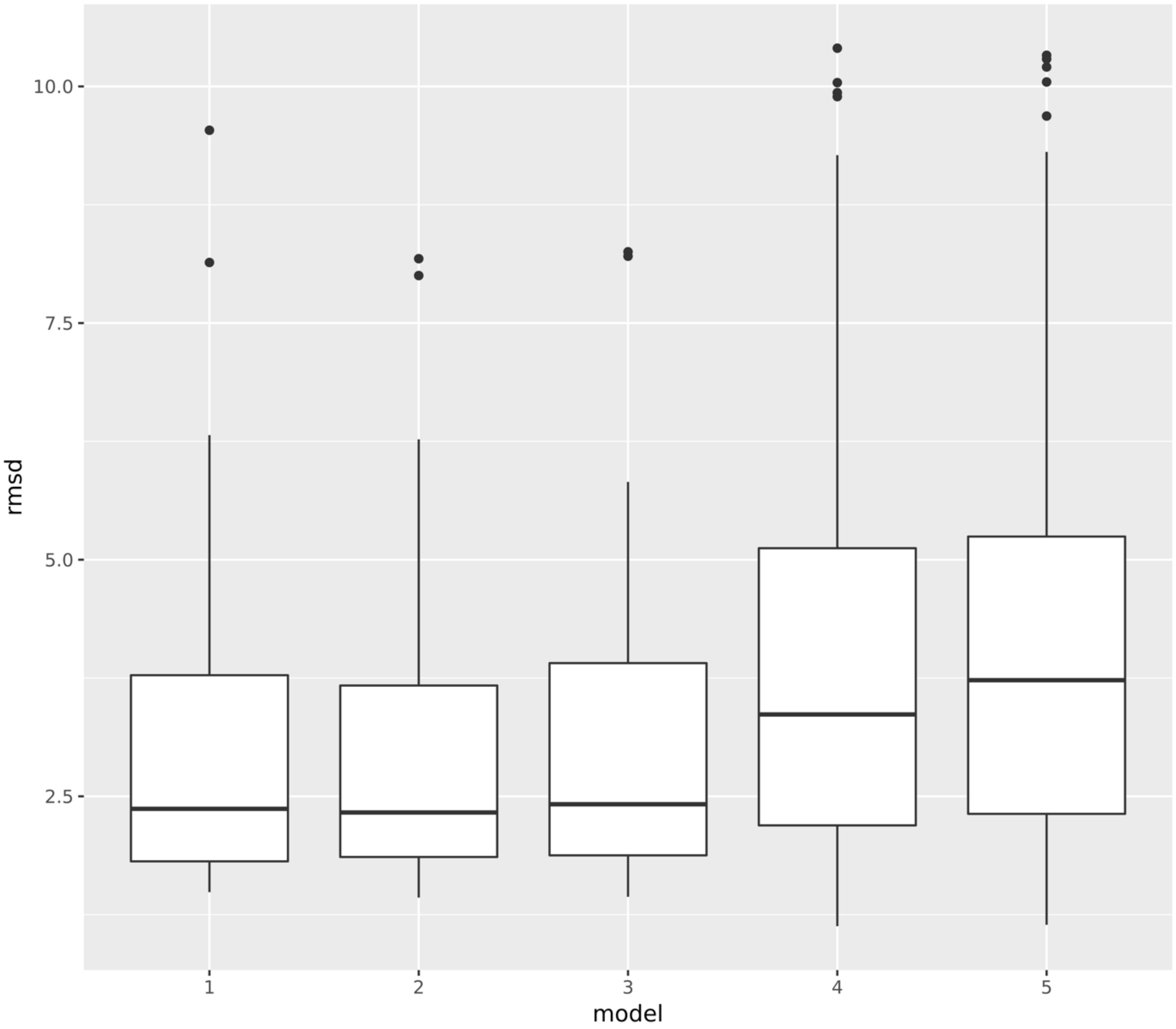


Fig S2: *Heavy-atom RMSD values grouped by ranking. The models in group 1 were the ones that were ranked at the top by HADDOCK and the ones in group 5 the ones ranked at the bottom.*

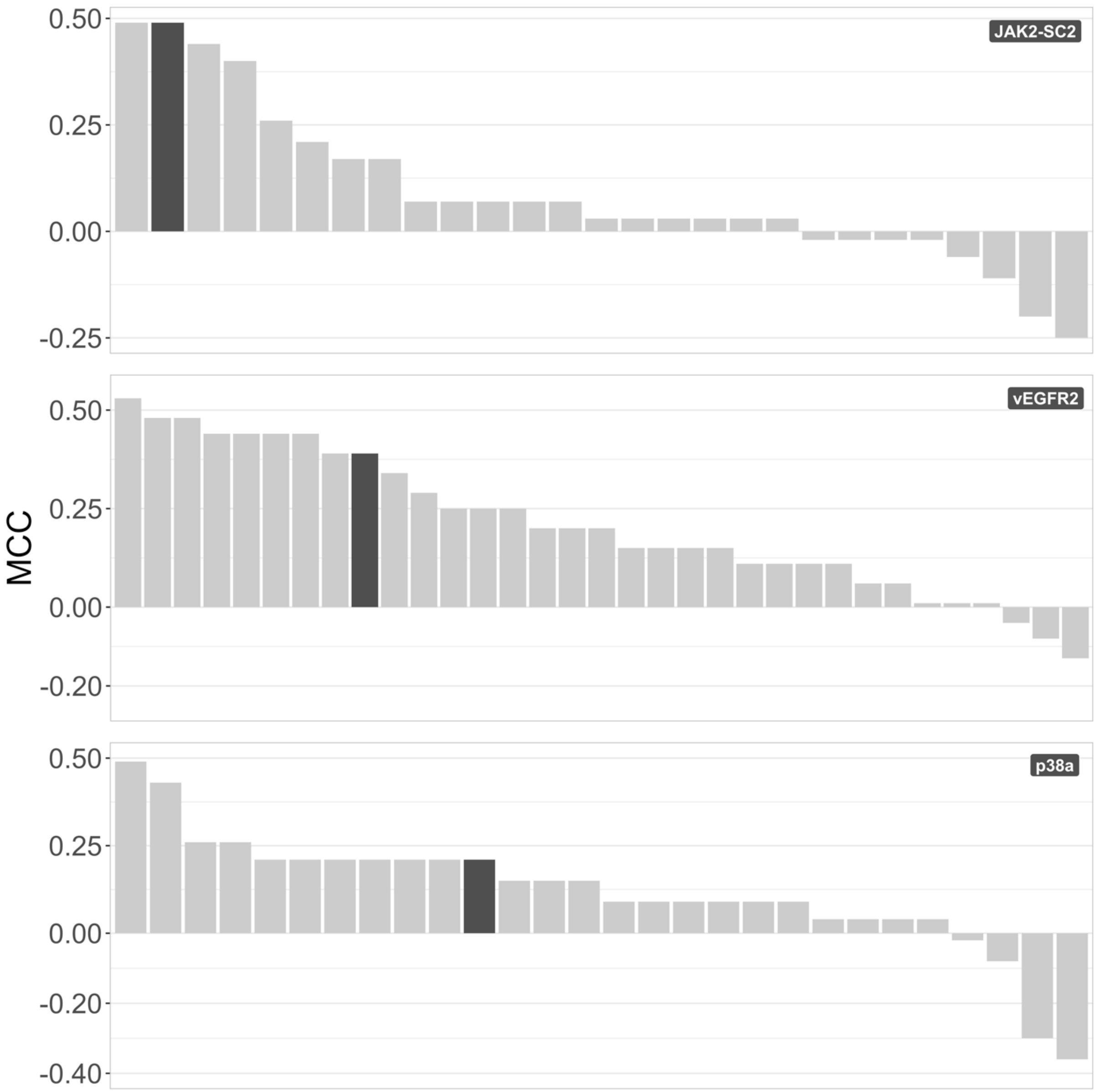


Fig S3: *Binding affinity prediction classification performance. Top: JAK2-SC2. Middle: vEGFR2. Bottom: p38a. The bars indicate the Matthews Correlation Coefficient for every submission. The dark grey bars correspond to our submissions.*

Chapter 6

Shape-based modelling of protein-small molecule complexes with HADDOCK

Abstract

Small molecule docking remains one of the most valuable techniques for the computational study of protein-small molecule complexes. It allows us to study the interactions between compounds and the protein receptors they target at atomic detail, in a timely and efficient manner. Here we present a new protocol in HADDOCK, our integrative modelling platform, which incorporates homology information for both receptor and compounds. It makes use of HADDOCK's unique ability to integrate information in the simulation to drive it toward conformations which agree with the provided data. The focal point is the use of shape restraints derived from homologous compounds bound to the target receptors. This shape is composed of fake atom beads based on the position of the heavy atoms of the template compound. Ambiguous distance restraints are subsequently defined between those beads and the heavy atoms of the ligand to be docked. We have benchmarked this protocol against another protocol developed in our group which was validated as one of the best performers in the D3R blind docking experiment. The use of shape restraints leads to an improved overall performance, being able to induce significant conformational changes on the ligand and alleviating the requirements to select a priori relevant ligand conformations for docking.

Introduction

The importance of reliable methods for the docking of small molecule compounds to receptors of pharmaceutical interest cannot be understated. The nature of modern drug development practices dictates the gradual filtering of millions (perhaps even hundreds of millions) of compounds contained in virtual libraries of large pharmaceutical corporations to, ultimately, a few dozen lead compounds that can be further optimised before their clinical potential is investigated in animal and human trials [376,377]. This set of practices – collectively known as Computer-Aided Drug Design (CADD) – encompasses a variety of methods such as virtual screening of compounds, molecular docking with recent developments making use of machine learning-based approaches [378,379] and binding affinity prediction. The two preceding chapters have detailed our recent efforts in developing approaches for the docking of protein-small molecule complexes with HADDOCK, our data-driven docking platform, spurred by our participation in the challenges organised by the D3R consortium (see Chapters 4 & 5).

Chapters 4 and 5 summarise our participation in the second and third iteration of the D3R Grand Challenge (GC), respectively. Whereas our protocol of choice for D3R GC2 did not earn us a spot among the top performing groups of that round, it did allow us to better understand the problems specific to protein-small molecule docking. The main takeaway point was that making use of the most closely related receptor for every target compound significantly improved the success of the modelling. In this case, the best receptor was identified after comparing the crystallographic compound with the target compound and selecting the receptor conformation with the most similar ligand. With that knowledge we optimised additional aspects of our approach for D3R GC3 – always prioritising the use of high-quality experimental information for every step of the process. That revised protocol resulted in HADDOCK submitting one of the most accurate predictions for the pose prediction component of the challenge. We applied the same protocol in D3R GC4 with equally good results.

Here, we present a new protocol which incorporates all the lessons that we have learned over three years of participating in the D3R blind docking experiment into a protocol tailored for HADDOCK, bypassing one of the main limitations of the previous protocols – their reliance on external software for significant parts of the ligand-based modelling process. This limitation did not allow us to use the integrative modelling capabilities of HADDOCK as the rigid-body and semi-flexible refinement stages were bypassed and only flexible refinement was performed. Also, the reliance on commercial software for shape comparison and superimposition means that the template-based protocol described in the previous chapter could never be integrated in the publicly accessible HADDOCK webserver.

Shape information was central to the protocol described in Chapter 5. Briefly, after identifying highly homologous receptors with a compound bound to them, we compare the similarity of all crystallographic compounds to all target compounds and select the receptor conformation whose compound has the highest similarity to the compound to dock. Prior to docking, we filter the generated conformers by comparing their 3D shape with that of the most similar crystallographic compound and select the 10 closest conformers in terms of shape similarity. Finally, these 10 conformers are placed into the binding pocket by superimposing their shape onto the shape of the crystallographic compound and the model was refined in HADDOCK (i.e. no docking is performed).

In the new, shape-restrained protocol presented here, the template identification and conformer generation procedures are the same as the ones described previously. After identifying a suitable receptor template for each target, its bound compound is transformed into a shape consisting of fake beads. Ambiguous distance restraints are then defined between those beads and the heavy atoms of the compound. There is no selection of conformers prior to docking, but instead all 500 conformers are docked into the receptor template. The most suited conformations are automatically selected during the docking by HADDOCK based on the shape restraints.

Materials and Methods

To validate our new approach, we compare the results obtained for GC4 using the protocol established for GC3 (see Chapter 5), with those of the shape-restrained protocol described in this Chapter. The dataset is the one provided to the GC4 participants. It consists of 20 compounds, most of which contain macrocyclic elements, to be docked against the BACE-1 receptor.

As for the previous protocols, the identification of suitable receptors is the first step. For this we use the sequence of the receptor as input and search the PDB for highly homologous structures co-crystallised with small molecules. We then match every target compound to the ones in the set of identified receptors using the Tanimoto coefficient as similarity measure. The ligand modelling begins with generating conformers for all compounds. To this end we are using OMEGA [342,380], generating up to 500 conformers per compound.

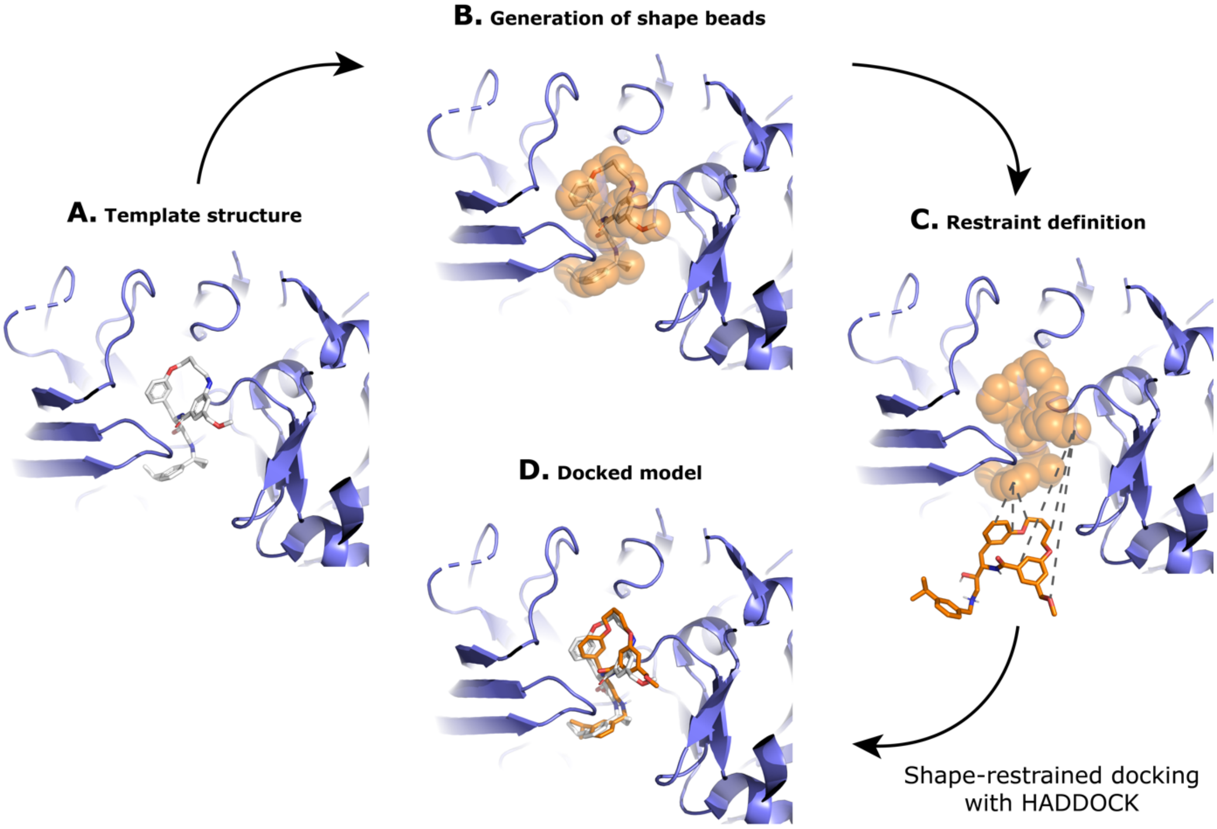

B. Generation of shape beads
A. Template structure
C. Restraint definition
D. Docked model
Shape-restrained docking with HADDOCK

Fig 1*: Illustration of the shape-restrained docking protocol. Panel A shows a suitable receptor template identified. Panel B shows the heavy atoms of the crystallographic compound transformed into shape beads. The crystallographic compound is then removed from the pocket and restraints are defined between the shape and the conformers (Panel C). Panel D shows a docked model superimposed onto the template structure. The protein receptor is shown as slate cartoon, the crystallographic compound as white sticks, the generated and docked compounds as orange sticks and the shape beads as transparent orange spheres. All molecular graphics were generated with PyMOL.*

Fig. 1 illustrates this shape-based docking protocol. After identifying one receptor template per target, we transform all heavy atoms of its compound into dummy beads (those do not interact with the remaining of the system), much in the same way as the beads described in Chapter 3. We then define ambiguous distance restraints with an upper limit of 1Å from each bead to any heavy atom of the compound to be docked. This effectively enforces that the ligand atoms must overlap with a bead. None of the restraints are discarded during the simulation (noecv = *false*). The nature of the restraints creates an additional consideration, specifically what should be the “origin” and “target” of the restraints? In this protocol we have defined the shape as the origin and the compound as the target, meaning that all shape beads should at the end be close to a ligand atom. However, in the case of a compound which would be smaller than the template shape, the optimal way of defining the restraints would be the opposite (i.e. from the compound to the shape).

For the docking we use the command-line version of HADDOCK 2.4. We sample 10000 and 400 it0 and it1 models, respectively and only perform a short energy minimisation at the end instead of the full water refinement. The positions of both the receptor and its associated shape are fixed in their original orientations. The shape is kept rigid throughout the protocol while the receptor interface becomes flexible during the refinement stage. We also scale down the intermolecular interactions during the rigid body stage to facilitate the insertion of the ligand into the binding pocket and accordingly exclude the vdW energy term during the scoring of the rigid-body models. As recommended for small ligand docking (the result of protocol optimisation during our D3R participation), we also reduce the weight of the intermolecular electrostatic energy term to 0.1 (instead of the default 0.2) in the final scoring function and use a RMSD-based clustering method with a cut-off of 1.5A.

Other than the above defined modifications, the scoring function used is the default scoring function of HADDOCK which has already been described in Chapter 2. Its functional form, specific for protein-ligand docking for the three stages is:

$$HS - it0 = 0.0 * E_{vdw} + 1.0 * E_{elec} + 1.0 * E_{desolv} + 0.01 * E_{AIR} - 0.01 * BSA$$

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

$$HS - itw = 1.0 * E_{vdw} + 0.1 * E_{elec} + 1.0 * E_{desolv} + 0.1 * E_{AIR}$$

Similarly to previous chapters, we evaluate the quality of the generated models according to their structural deviation from the reference structures. For this we use the interface-ligand RMSD (IL-RMSD), which is the RMSD calculated over all heavy atoms of the ligand after superimposing on all backbone atoms of the interface of the receptor. Models with an IL-RMSD of less than 0.5 Å, between 0.5 and 1 Å, between 1 and 2 Å and over 2 Å are classified as high-, medium-, acceptable-quality and incorrect, respectively.

Results and Discussion

We compare the results of the new shape-restrained docking protocol with those obtained applying the protocol described in Chapter 5 for the GC4 experiment. Figure 2 shows the IL-RMSD of the top 5 poses (as ranked by HADDOCK score) for the two protocols and for docking with the native (the bound crystal structure) (in the self pane) and the selected template (in the cross pane) receptors. Both protocols perform very well with overall mean IL-RMSD values of 1.57 ± 0.75 Å and 2.00 ± 1.13 Å for the new and old protocol, respectively, when docking with the template, and 1.59 ± 1.16 Å and 1.80 ± 1.12 Å when using the native receptor (i.e. no conformational changes required in the receptor) (see Table 1).

Table 1*: Median and average IL-RMSD values for both protocols when considering top5, top1 and the best poses and docking with the native or a template receptor. "new" refers to the protocol described in this Chapter and "old" to the application of the protocol described in Chapter 5 on the dataset of the GC4 blind docking experiment.*

Receptor	Protocol	Median RMSD [Å]	Mean RMSD [Å]	Cut-off
Native	New	1.22	1.59±1.16	Top5
	Old	1.31	1.80±1.21	
	New	1.17	1.35±0.58	Top1
	Old	1.26	1.58±0.80	
	New	0.91	1.02±0.38	Best
	Old	1.09	1.25±0.54	
Template	New	1.35	1.57±0.75	Top5
	Old	1.69	2.00±1.13	
	New	1.20	1.45±0.62	Top1
	Old	1.53	1.71±0.81	
	New	1.07	1.20±0.49	Best
	Old	1.20	1.46±0.75	

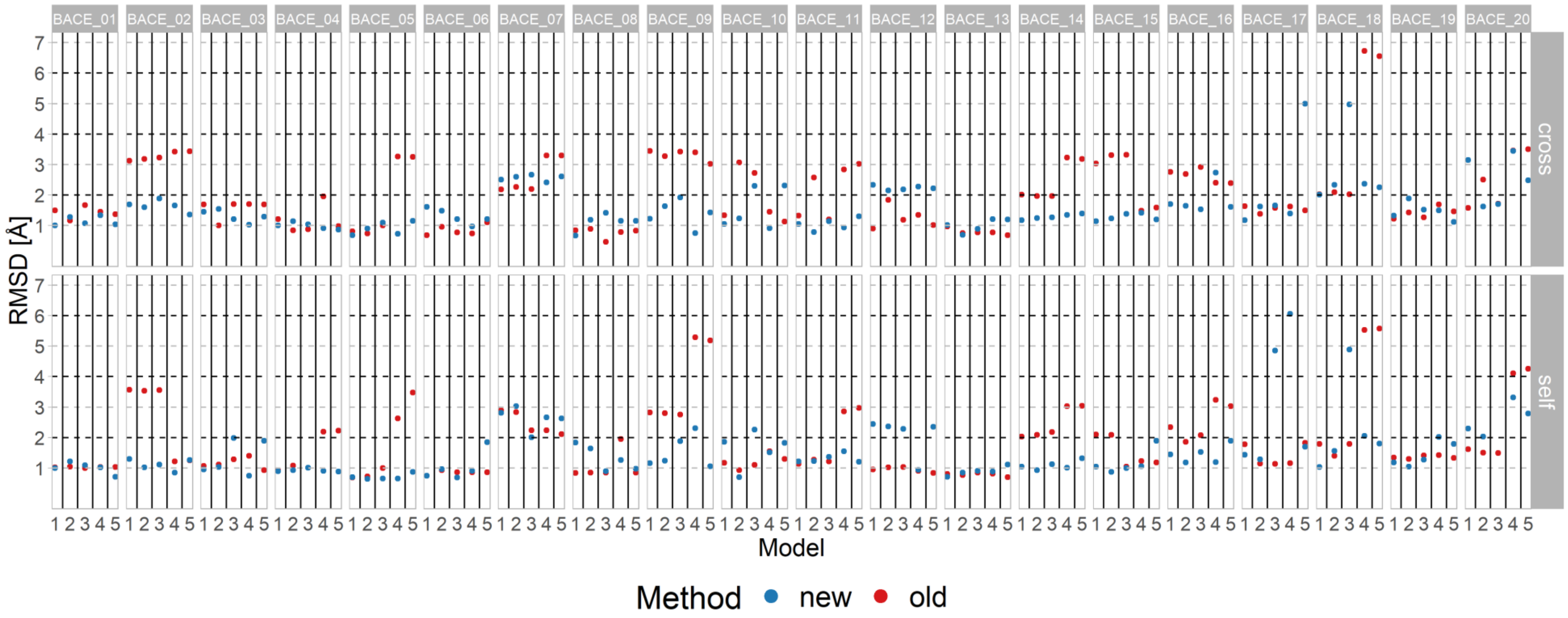


Fig 2: *Overview of the results obtained by the old and new protocols. IL-RMSD values for each of the top 5 poses are shown as red (corresponding to the old protocol) and blue (corresponding to the new shape-restrained protocol) dots. The two panels report the performance of cross- (top) and self-docking (bottom) for the various targets. Cross-docking refers to docking from the selected template while self-docking refers to docking from the bound form of the receptor.*

When considering the top 5 models generated in the cross-docking runs, the success rate is 90% (18/20) and 80% (16/20) for the new and old protocols, respectively, and 100% (20/20) and 95% (19/20) for the self-docking runs. When considering only the top model, the success rates drop to – the still impressive – 85% (17/20) and 75% (15/20), and 85% (17/20) and 80% (16/20) for the new and old protocols and for cross- and self-docking runs respectively.

Fig. 3 shows the comparison between the top 5 models for the old and new protocols, with the new protocol outperforming the old one in almost every target.

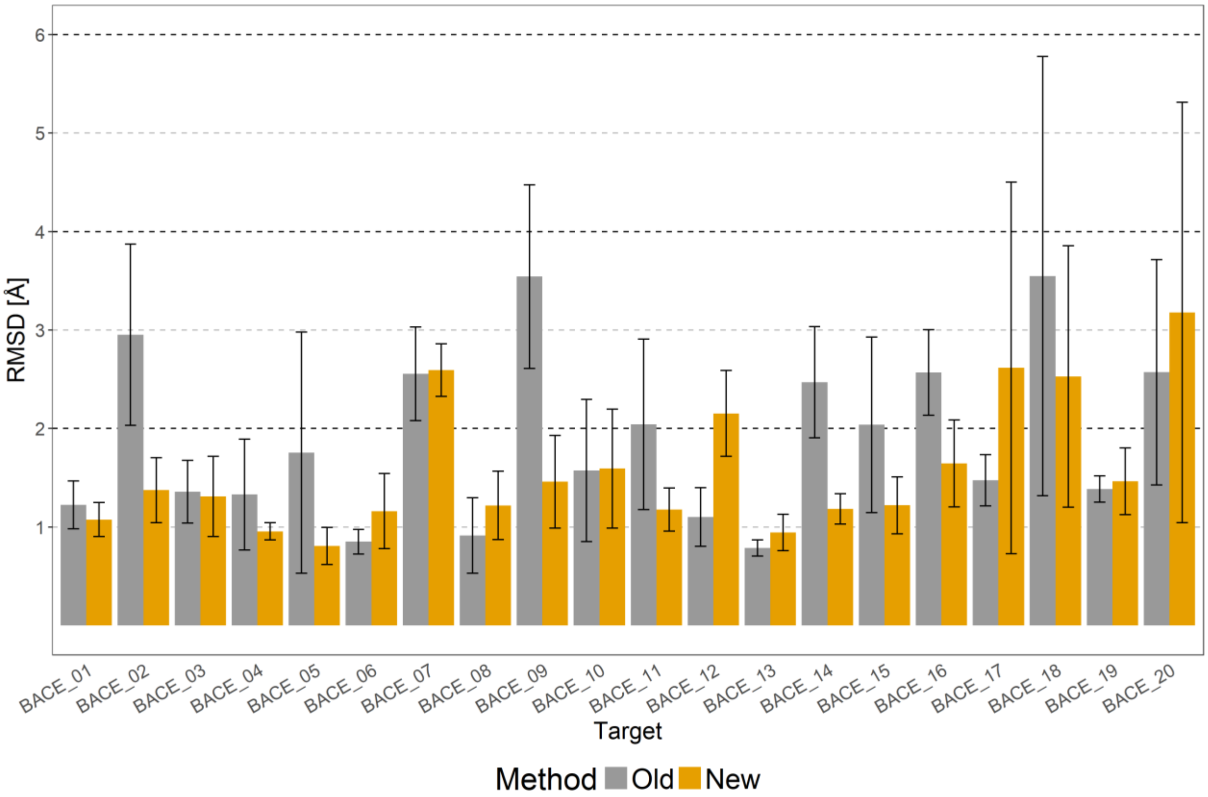


Fig 3: *Comparison of the performance of the old and new protocols by target. The IL-RMSD values have been averaged over the top 5 models and for both cross- and self-docking runs. The old protocol is shown in grey bars and the new in orange.*

Our new shape-restrained docking protocol also performs better in terms of scoring as is shown in the violin plots of Fig. 4. For example, the distribution of RMSD values for the models ranked as the third best according to HADDOK score for the old protocol clearly indicates they are of higher quality than the ones ranked as the best (they reach lower RMSD values). Excluding a handful of outliers, the new protocol performs much more consistently: Even models ranked at the fifth position are distributed in a similar fashion as the ones ranked as the top. These also sample more high-quality models as indicated by the broader distributions at lower RMSD values. These differences most likely result from the fact that – unlike the new protocol – the old one is a simple refinement. As such it cannot induce any significant conformational changes in the ligand and therefore pretty much reflects the heterogeneity of the starting conformations. The new protocol imposes restraints on the conformations that the ligands can adopt during docking due to the forces acting on them that effectively cause them to morph to the shape. This effect can clearly be seen in Fig. 5 which shows the distributions of differences in RMSDs between the rigid-body and refined models calculated over the top 200 models of all ligands. Positive differences indicate that the ligand conformation is moving toward its native bound form. The template-based protocol distributions are narrower and more symmetrical. In contrast, the shape-restrained distributions are asymmetrical and extend toward positive values. In some cases, improvements of more than 4Å RMSD are observed between the rigid-body and refined models.

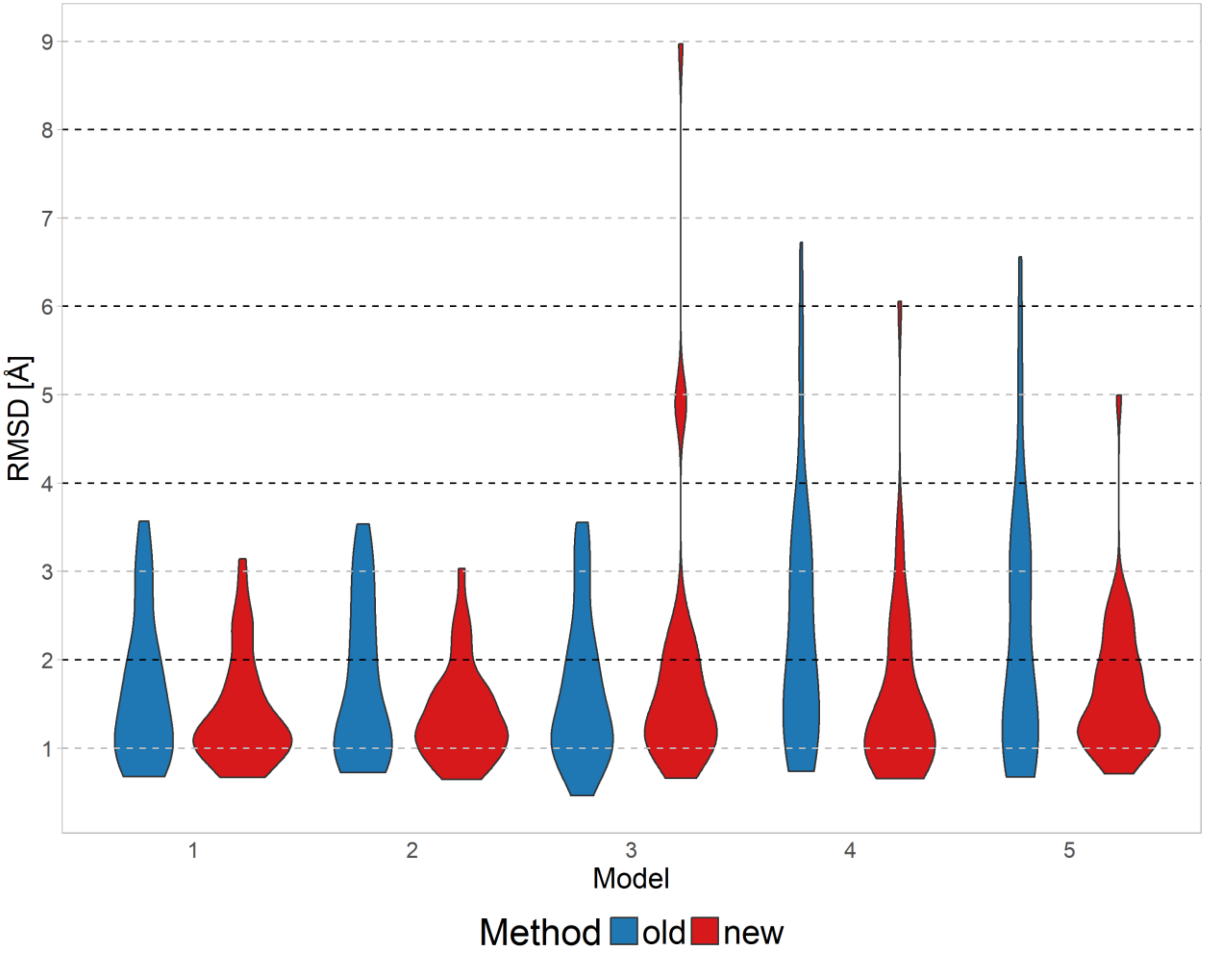


Fig 4: Violin plots of the IL-RMSD values for the new and old protocols grouped by model rank. The distributions corresponding to the old protocol are coloured blue, and those to the new red.

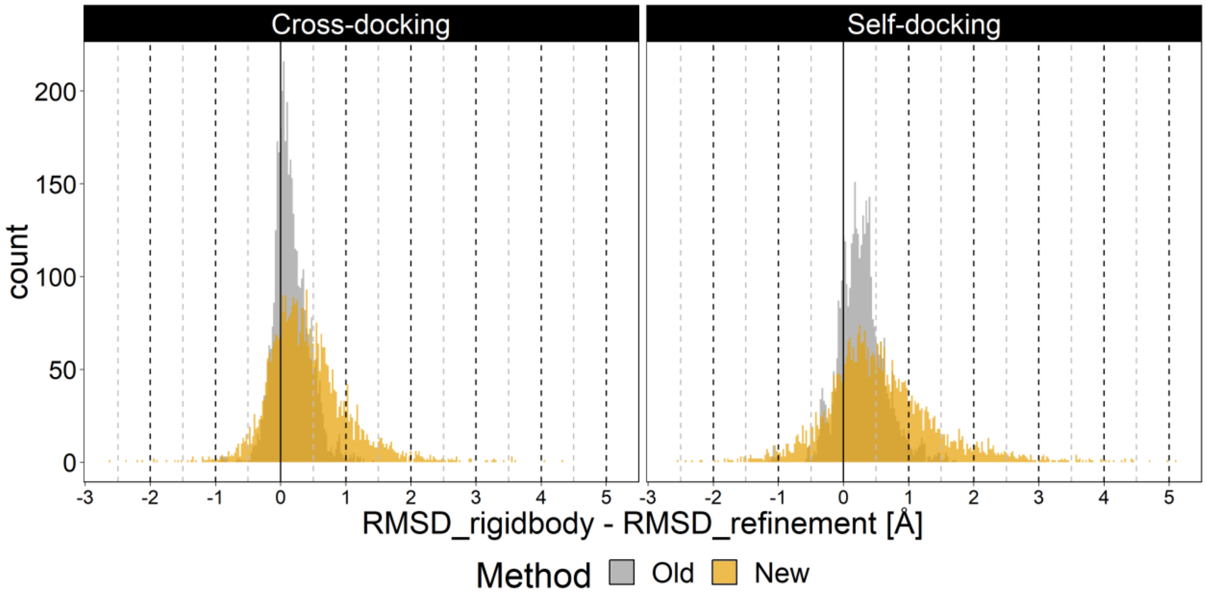


Fig 5*: Distribution of RMSD differences between the rigid body and semi-flexible refined models calculated over the top200 models of all ligands. The left and right panels show the distribution corresponding to the cross- and self-docking experiment respectively. The distribution of values corresponding to the template-based and shape-restrained protocols are coloured grey and orange, respectively. A positive RMSD difference indicate that the refinement is moving the ligand toward its native, bound conformation.*

Another remarkable point of the shape-restrained protocol is that it is much less sensitive to the quality of the starting receptor conformation: Rather minor differences are observed between the cross- and self-docking runs. As shown in Fig. 6, only slight improvements in RMSD values are observed when docking with the native receptor. This is in contrast with the old protocol, which shows clear improvement when using the native receptor for docking. Two targets (BACE 17 and 20) stand out with the performance for the self-docking runs being significantly worse than that of the cross-docking. We believe that the reason behind this are minor steric disagreements between the shape of the template and the bound receptor, forcing the binding pocket to distort during the flexible refinement stages. Self-docking in any case remains an artificial exercise only valuable to assess and compare the performance of protocols.

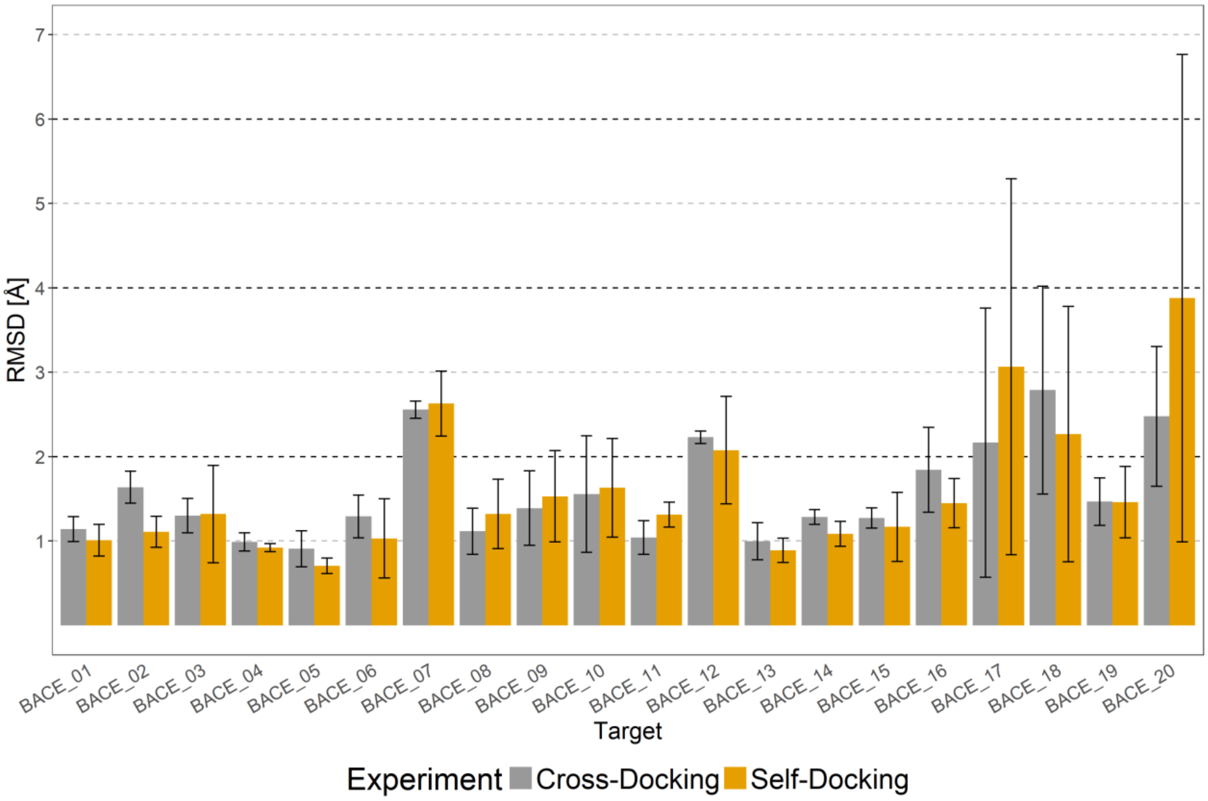


Fig 6*: Comparison of the performance between cross- (grey bars) and self-docking (orange bars) for the shape-restrained protocol. The RMSD values have been averaged over the top 5 models for each target.*

Conclusion and Perspectives

In this Chapter we have presented an original protocol for shape-restrained docking of small molecules to protein receptors. We compared its performance with the established templated-based protocol developed in our group on targets of the blind docking experiment described in the previous Chapter. This new protocol makes use of the shape of an existing compound cocrystallised with ideally the same receptor or a highly homologous one. The restraints defined between the template shape and the ligand force the ligand to adopt a conformation in the binding pocket that best matches the shape. This new protocol outperforms our old template-based refinement protocol both in terms of sampling and scoring as it produces more accurate models while at the same time ranking them better. This new protocol achieves an impressive 85% success rate when docking from a template receptor considering only the top model.

While the concept of shape is one that has already been used in the field of docking, even dating as far back as 1982 when the first docking program was published [381], we believe that the formulation put forth in this chapter, based on the use of ambiguous distance restraints to the shape, has never been used before for ligand docking. Most importantly, it allows us to combine the template information in the form of compound shape with information that might be known about the system. For example, the chemistry of enzymatic catalysis might dictate that specific atoms of a small molecule must be in close proximity to residues that are part of the catalytic triad for that enzyme to perform its function. This can be encoded as additional distance restraints acting as the same time as the shape ones. This concept becomes particularly powerful when utilised in the concept of integrative modelling frameworks like HADDOCK, which can combine information from many sources in a single simulation.

There are further avenues worth exploring to further improve the performance of this protocol. In the current implementation, the shape beads have no properties and only act as placeholders for the defined restraints. We could however define various types of beads to represent a pseudo-pharmacophore model and define custom restraints towards specific atom types of the ligand to match this pharmacophore model. This might allow to place specific atom types in an energetically favourable environment, improving thereby the docking results.

Chapter 7

Conclusions and Perspectives

Summary

The preceding chapters of this thesis have covered three main areas of research that fall under the purview of Computational Structural Biology; Specifically:

1. The types of data – experimental or computational – that can be used in Integrative Modelling approaches
2. Recent advances in the modelling of membrane protein complexes and
3. Protocols for the docking of small molecules to protein receptors.

Chapter 1 provided a detailed and comprehensive review on the types of data than can be used by Integrative Modelling software like HADDOCK, ROSETTA and IMP, with a particular emphasis on the experimental techniques which can be used to map interfaces, derive distance restraints or shape-based approaches. Another focal point of the chapter is how recent advancements have affected the field of membrane protein modelling. **Chapters 2 and 3** also relate to membrane protein modelling with the former describing a recently available docking benchmark comprised entirely of ready-to-dock membrane protein complexes as well as the baseline performance of HADDOCK for the entries of the benchmark, and the latter, ongoing work regarding development of a protocol for HADDOCK for the docking of transmembrane protein complexes.

The remaining of the thesis focused on small molecule modelling with **Chapters 4 through 6** detailing three separate protocols for the docking of small molecules and protein receptors, with every protocol and chapter reflecting methodological improvements over the previous one. In **Chapter 4**, I described the participation of the HADDOCK group in the 2016 iteration of the Grand Challenge – the blind docking experiment organised by the D3R consortium. While, despite some successes, our performance in the pose prediction component was not impressive, we could identify the main factor limiting HADDOCK's performance, namely the selection of appropriate templates for the receptor and came up with an improved way of selecting receptors. **Chapter 5** described additional improvements in our protocol related to the way the compound conformers are selected prior to docking which led to our participation in the 2017 iteration of the Grand Challenge being evaluated as one of the best in that round. **Chapter 6** detailed the development of a new protocol for protein-small molecule docking in HADDOCK, by combining the lessons and conclusions from **Chapters 4 and 5** and formalising their approaches in a method that relies on HADDOCK's main strength, its ability to incorporate information to guide the simulation. This new, shape-restrained docking protocol outperformed all our previous efforts while at the same time not relying on any external software for the docking.

A common denominator between the membrane protein work and the small ligand docking discussed in this thesis is the use of shape information. Indeed, **chapters 3 and 6** both describe applications of shape information represented as beads to drive the modelling process. In **Chapter 3** one or more layers of beads are used to implicitly represent the membrane and in **Chapter 6** ligand docking is restrained to a shape based on the structure of a homologous compound. Despite the commonalities between the two protocols, the outcome of the docking is very different between the two, with the small molecule protocol achieving high-quality results and improving upon our previous efforts in this area, whereas the membrane one achieves results which are only marginally better than defining centre-of-mass restraints between the transmembrane segments of the two partners for the docking. A main limiting factor in the case of membrane protein complexes seems to be the size of the complex, which defines the number of restraints defined between shape and molecules and negatively impacts the performance of the docking.

Challenges and future directions

Although integrative approaches have been present in the field of (computational) structural biology for a long time, it wasn't until relatively recently that they coalesced into the field of integrative modelling. Like all young disciplines, integrative modelling suffers from the growing pains associated with its nascent state. Specifically, there is a lack of interoperability and unified workflows between popular packages for molecular simulations. This is exacerbated by the lack of a common framework or format for the data that are produced by the various experimental and computational techniques which can be used in molecular modelling. As a consequence, data must be manually manipulated and transformed before they can be used in modelling workflows. This lack of appropriate data formats also extends into modelling software as most of them still need and produce – outdated and inadequate for the task – PDB-formatted files, despite the advent of the mmCIF format (https://github.com/ihmwg/IHM-dictionary) which allows for an arbitrarily high number of metadata to be associated with a single or multiple structural models in a single file. The mmCIF format has been the default format of the PDB since 2014 and some modelling software already support input and output in this format even though they might use different formats internally (HADDOCK among them). In recent years, wwPDB (http://www.wwpdb.org/), together with industrial and academic partners, has been responsible for many developments related to standardisation efforts such as the advent of PDB-dev (https://pdb-dev.wwpdb.org/)

[239], a publicly accessible repository comprised entirely of structural models obtained with integrative approaches. The primary force pushing these developments forward is the Hybrid/Integrative Methods task force assembled in 2014 to address the issues arising from the need to develop standards around integrative modelling [382]. The inaugural meeting of that task force identified five recommendations as important moving forward. (1) Archival of all the data that went into the simulation as well as the resulting models, (2) a flexible data format allowing for multi-state, multi-level and other complicated representations, (3) protocols for the estimation of uncertainty of the resulting models, (4) a system for the deposition and dissemination of the data produced and (5) publication standards for integrative models. The mmCIF format with the necessary dictionaries for integrative modelling has solved challenges (1) and (2), but it remains to be adopted by the community at large, while the advent of PDB-dev – eventually to be absorbed into the PDB itself – already serves as a repository for structural models determined with integrative methods. The development of protocols for the estimation of uncertainty in the models produced by integrative modelling is probably the point on which advances have lagged the most. Publication standards for integrative models is another active research area with more results expected soon with the publication of the outcome of the second meeting of the task force.

Another issue in most molecular modelling software is the inconsistent way in which uncertainties in the experimental data obtained by methods like, for example, SAXS or XL-MS are propagated in the simulation – if they are propagated at all. A related matter is the proper weighting of the various restraints used when modelling and using data from different sources, for example when using shape data from SAXS and residue distance information from XL-MS. Some software, like IMP and ISD already account for this, by using a Bayesian framework to weight the various terms used during docking and pave the way for the widespread development and adoption of probabilistically sound modelling protocols. This issue affects all integrative modelling software and HADDOCK is no exception. However, in HADDOCK, most experimental information is integrated in the simulation if the form of distance restraints which are then used to drive the sampling but are also part of the scoring function in the form of a restraint energy term. This applies to all interface-mapping and residue-based approaches such as mutagenesis and HDX, and NMR, XL-MS, FRET and DEER, respectively. The distance restraining function used in HADDOCK has a functional form that makes it less sensitive and more robust for large deviations. In **Chapters 3 and 6**, I presented two protocols which make use of shape information for the modelling of membrane protein and protein-small molecule complexes and while the small molecule protocol worked very well, the same cannot be said for the membrane one, for which we define a number of restraints similar to what we would define if we wanted to make use of a SAXS- or cryo-EM-derived shape represented as beads. This indicates that, for such a protocol, a reweighting of the restraint energy term against the nonbonded ones might be warranted.

Challenges related to the way the data are represented and integrated in a simulation are not the only issue though: A challenge of equal – if not greater – importance is the determination of which data should be integrated in the simulation at all. In other words, how can we determine the quality of the data that goes into a simulation to ensure the best possible outcome. This challenge is particularly important for integrative modelling software like HADDOCK which drives the sampling based on the provided data. This is highly beneficial in the case of high-quality data which represents the native state as the code will explore the part of the conformational landscape around that native state. But this can also be detrimental in cases where the data are not accurate or represent multiple states or conformations. Docking might not be the best-suited modelling approach for such cases. Some developments are already happening in this direction with the advent of software like DISVIS, which calculates the information content of XL-MS derived distance information in the context of a binary complex and aims at identifying false positive restraints. Additionally, most experimental methods now have ways of assigning confidence scores to their measurements, mainly by repeating experiments and identifying the results which are consistent between the various replicates (see Chapter 1).

Of course, the development of protocols such as those described in **Chapters 3 and 6**, is meaningless if not applied to challenging areas of structural biology like the modelling of membrane protein complexes. All primary experimental structure determination techniques – X-ray crystallography, NMR spectroscopy and cryo-EM – have contributed significantly to the recent wealth of high-quality structural models of membrane proteins. Computational approaches like molecular dynamics, membrane protein-specific databases and coarse grained forcefields have also kept up and are expanding our understanding of membrane proteins. Docking codes like ROSETTA also support the docking of transmembrane proteins with tailor-made potentials that take into account the membrane environment. The protocol that was described in **Chapter 3** is the first attempt toward providing support for the docking of transmembrane proteins in HADDOCK, however, the results indicate that significant methodological improvements are required before it can be widely applied for the study of membrane systems. However, the application of a similar, shape bead-based concept was shown to work very well when docking small molecules to protein receptors as described in

Chapter 6. The protocol still needs to be extensively benchmarked but preliminary results indicate that it can consistently yield models of high-quality. Combined with HADDOCK's innate ability to integrate diverse types of experimental data in the simulation, it opens interesting avenues for further exploration.

One development which could push the integrative modelling community forward in what the CASP and CAPRI experiments did for the protein structure prediction and protein interaction prediction communities in the early to mid-90s and early-00s, respectively. There is thus a need for a new blind challenge which would assess the state-of-the-art in integrative modelling approaches. This would galvanise the community and help create assessment criteria which would formalise the way integrative models are evaluated and disseminated, and create a stronger sense of community around the field of integrative modelling. Some steps in this direction are already taking place with the data-assisted category of CASP with one recent example being CAPRI targets T149-T151 (CASP targets T099, S099, X099) of the joint CASP-CAPRI experiment held over the summer of 2018 which featured SAXS and XL-MS data.

Finally, I would be amiss to not mention challenges that continue to affect integrative modelling and computational biology in general even if this thesis did not explicitly deal with those. Issues like the difficulty of modelling large conformational rearrangements, and predicting when those are needed for binding, and the challenge of obtaining accurate binding affinities have been present since the very first days of the field. Despite steady progress in many areas like biomolecular docking and alchemical free energy simulations, these are still open challenges in field.

References

(1) Rodrigues, J. P. G. L. M.; Bonvin, A. M. J. J. Integrative Computational Modeling of Protein Interactions. *FEBS J.* **2014**, *281* (8), 1988–2003.

(2) Rout, M. P.; Sali, A. Principles for Integrative Structural Biology Studies. *Cell* **2019**, *177* (6), 1384–1403.

(3) Kato, H. et al. Architecture of the High Mobility Group Nucleosomal Protein 2-Nucleosome Complex as Revealed by Methyl-Based NMR. *Proc. Natl. Acad. Sci.* **2011**, *108* (30), 12283–12288.

(4) Corbeski, I. et al. DNA Repair Factor APLF Acts as a H2A-H2B Histone Chaperone through Binding Its DNA Interaction Surface. *Nucleic Acids Res.* **2018**, *46* (14), 7138–7152.

(5) Horn, V.; van Ingen, H. Recognition of Nucleosomes by Chromatin Factors: Lessons from Data-Driven Docking-Based Structures of Nucleosome-Protein Complexes. In *Epigenetics [Working Title]*; IntechOpen, 2018.

(6) Xiang, S. et al. Site-Specific Studies of Nucleosome Interactions by Solid-State NMR Spectroscopy. *Angew. Chemie Int. Ed.* **2018**, *57* (17), 4571–4575.

(7) van Emmerik, C. L.; van Ingen, H. Unspinning Chromatin: Revealing the Dynamic Nucleosome Landscape by NMR. *Prog. Nucl. Magn. Reson. Spectrosc.* **2019**, *110*, 1–19.

(8) Sprangers, R.; Kay, L. E. Quantitative Dynamics and Binding Studies of the 20S Proteasome by NMR. *Nature* **2007**, *445* (7128), 618–622.

(9) Religa, T. L.; Sprangers, R.; Kay, L. E. Dynamic Regulation of Archaeal Proteasome Gate Opening As Studied by TROSY NMR. *Science (80-.).* **2010**, *328* (5974), 98–102.

(10) Ruschak, A. M.; Kay, L. E. Proteasome Allostery as a Population Shift between Interchanging Conformers. *Proc. Natl. Acad. Sci.* **2012**, *109* (50), E3454–E3462.

(11) Huang, R.; Pérez, F.; Kay, L. E. Probing the Cooperativity of Thermoplasma Acidophilum Proteasome Core Particle Gating by NMR Spectroscopy. *Proc. Natl. Acad. Sci.* **2017**, *114* (46), E9846–E9854.

(12) Kitevski-LeBlanc, J. L. et al. Investigating the Dynamics of Destabilized Nucleosomes Using Methyl-TROSY NMR. *J. Am. Chem. Soc.* **2018**, *140* (14), 4774–4777.

(13) Fromm, S. A. et al. The Structural Basis of Edc3- and Scd6-Mediated Activation of the Dcp1:Dcp2 MRNA Decapping Complex. *EMBO J.* **2012**, *31* (2), 279–290.

(14) Fromm, S. A. et al. In Vitro Reconstitution of a Cellular Phase-Transition Process That Involves the MRNA Decapping Machinery. *Angew. Chemie Int. Ed.* **2014**, *53* (28), 7354–7359.

(15) Schütz, S.; Nöldeke, E. R.; Sprangers, R. A Synergistic Network of Interactions Promotes the Formation of in Vitro Processing Bodies and Protects MRNA against Decapping. *Nucleic Acids Res.* **2017**, *45* (11), 6911–6922.

(16) Cvetkovic, M. A.; Sprangers, R. Methyl TROSY Spectroscopy to Study Large Biomolecular Complexes. In *Modern Magnetic Resonance*; Springer International Publishing: Cham, 2018; pp 453–467.

(17) Saleh, T.; Rossi, P.; Kalodimos, C. G. Atomic View of the Energy Landscape in the Allosteric Regulation of Abl Kinase. *Nat. Struct. Mol. Biol.* **2017**, *24* (11), 893–901.

(18) Saleh, T.; Kalodimos, C. G. Enzymes at Work Are Enzymes in Motion. *Science (80-.).* **2017**, *355* (6322), 247–248.

(19) Huang, C. et al. Structural Basis for the Antifolding Activity of a Molecular Chaperone. *Nature* **2016**, *537* (7619), 202–206.

(20) van Ingen, H.; Bonvin, A. M. J. J. Information-Driven Modeling of Large Macromolecular Assemblies Using NMR Data. *J. Magn. Reson.* **2014**, *241* (1), 103–114.

(21) Renault, M.; Cukkemane, A.; Baldus, M. Solid-State NMR Spectroscopy on Complex

Biomolecules. *Angew. Chemie Int. Ed.* **2010**, *49* (45), 8346–8357.

(22) Aliev, A. E.; Law, R. V. Chapter 7. Solid State NMR Spectroscopy. In *Nuclear Magnetic Resonance*; 2014; Vol. 43, pp 286–344.

(23) Marchanka, A.; Carlomagno, T. Solid-State NMR Spectroscopy of RNA. In *Methods in Enzymology*; 2019; Vol. 615, pp 333–371.

(24) Medeiros-Silva, J. et al. High-Resolution NMR Studies of Antibiotics in Cellular Membranes. *Nat. Commun.* **2018**, *9* (1), 3963.

(25) Jekhmane, S. et al. Shifts in the Selectivity Filter Dynamics Cause Modal Gating in K+ Channels. *Nat. Commun.* **2019**, *10* (1), 123.

(26) Kyrilis, F. L.; Meister, A.; Kastritis, P. L. Integrative Biology of Native Cell Extracts: A New Era for Structural Characterization of Life Processes. *Biol. Chem.* **2019**, *400* (7), 831–846.

(27) Lawson, C. L. et al. EMDataBank Unified Data Resource for 3DEM. *Nucleic Acids Res.* **2016**, *44* (D1), D396–D403.

(28) Lawson, C. L.; Chiu, W. Comparing Cryo-EM Structures. *J. Struct. Biol.* **2018**, *204* (3), 523–526.

(29) Vakser, I. A.; Deeds, E. J. Computational Approaches to Macromolecular Interactions in the Cell. *Curr. Opin. Struct. Biol.* **2019**, *55*, 59–65.

(30) Soni, N.; Madhusudhan, M. S. Computational Modeling of Protein Assemblies. *Curr. Opin. Struct. Biol.* **2017**, *44*, 179–189.

(31) Lengauer, T.; Rarey, M. Computational Methods for Biomolecular Docking. *Curr. Opin. Struct. Biol.* **1996**, *6* (3), 402–406.

(32) Vallat, B. et al. Development of a Prototype System for Archiving Integrative/Hybrid Structure Models of Biological Macromolecules. *Structure* **2018**, *26* (6), 894-904.e2.

(33) Braitbard, M.; Schneidman-Duhovny, D.; Kalisman, N. Integrative Structure Modeling: Overview and Assessment. *Annu. Rev. Biochem.* **2019**, *88* (1), 113–135.

(34) Wodak, S. J.; Janin, J. Computer Analysis of Protein-Protein Interaction. *J. Mol. Biol.* **1978**, *124* (2), 323–342.

(35) Kuntz, I. D. et al. A Geometric Approach to Macromolecule-Ligand Interactions. *J. Mol. Biol.* **1982**, *161* (2), 269–288.

(36) Tovchigrechko, A.; Wells, C. A.; Vakser, I. A. Docking of Protein Models. *Protein Sci.* **2002**, *11* (8), 1888–1896.

(37) Plattner, N. et al. Complete Protein–Protein Association Kinetics in Atomic Detail Revealed by Molecular Dynamics Simulations and Markov Modelling. *Nat. Chem.* **2017**, *9* (10), 1005–1011.

(38) Śledź, P.; Caflisch, A. Protein Structure-Based Drug Design: From Docking to Molecular Dynamics. *Curr. Opin. Struct. Biol.* **2018**, *48*, 93–102.

(39) Vakser, I. A. Protein-Protein Docking: From Interaction to Interactome. *Biophys. J.* **2014**, *107* (8), 1785–1793.

(40) Anishchenko, I. et al. Protein Models: The Grand Challenge of Protein Docking. *Proteins Struct. Funct. Bioinforma.* **2014**, *82* (2), 278–287.

(41) Mosca, R.; Céol, A.; Aloy, P. Interactome3D: Adding Structural Details to Protein Networks. *Nat. Methods* **2013**, *10* (1), 47–53.

(42) Luck, K. et al. A Reference Map of the Human Protein Interactome. *bioRxiv* **2019**, 605451.

(43) Moitessier, N. et al. Towards the Development of Universal, Fast and Highly Accurate Docking/Scoring Methods: A Long Way to Go. *Br. J. Pharmacol.* **2009**, *153* (S1), S7–S26.

(44) Nithin, C.; Ghosh, P.; Bujnicki, J. Bioinformatics Tools and Benchmarks for Computational Docking and 3D Structure Prediction of RNA-Protein Complexes. *Genes*

(Basel). **2018**, *9* (9), 432.

(45) Krüger, A. et al. Molecular Modeling Applied to Nucleic Acid-Based Molecule Development. *Biomolecules* **2018**, *8* (3), 83.

(46) Luo, J. et al. Challenges and Current Status of Computational Methods for Docking Small Molecules to Nucleic Acids. *Eur. J. Med. Chem.* **2019**, *168*, 414–425.

(47) Sousa, S. F.; Fernandes, P. A.; Ramos, M. J. Protein-Ligand Docking: Current Status and Future Challenges. *Proteins Struct. Funct. Bioinforma.* **2006**, *65* (1), 15–26.

(48) Morris, G. M.; Lim-Wilby, M. Molecular Docking. In *Methods in Molecular Biology*; 2008; pp 365–382.

(49) 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.* **2013**, *23* (6), 894–902.

(50) Janin, J. et al. CAPRI: A Critical Assessment of PRedicted Interactions. *Proteins Struct. Funct. Genet.* **2003**, *52* (1), 2–9.

(51) Méndez, R. et al. Assessment of CAPRI Predictions in Rounds 3-5 Shows Progress in Docking Procedures. *Proteins Struct. Funct. Bioinforma.* **2005**, *60* (2), 150–169.

(52) Lensink, M. F.; Méndez, R.; Wodak, S. J. Docking and Scoring Protein Complexes: CAPRI 3rd Edition. *Proteins Struct. Funct. Bioinforma.* **2007**, *69* (4), 704–718.

(53) Lensink, M. F.; Wodak, S. J. Docking and Scoring Protein Interactions: CAPRI 2009. *Proteins Struct. Funct. Bioinforma.* **2010**, *78* (15), 3073–3084.

(54) Lensink, M. F.; Wodak, S. J. Docking, Scoring, and Affinity Prediction in CAPRI. *Proteins Struct. Funct. Bioinforma.* **2013**, *81* (12), 2082–2095.

(55) Lensink, M. F.; Velankar, S.; Wodak, S. J. Modeling Protein-Protein and Protein-Peptide Complexes: CAPRI 6th Edition. *Proteins Struct. Funct. Bioinforma.* **2017**, *85* (3), 359–377.

(56) Porter, K. A. et al. What Method to Use for Protein–Protein Docking? *Curr. Opin. Struct. Biol.* **2019**, *55*, 1–7.

(57) Lensink, M. F. et al. Blind Prediction of Homo- and Hetero- Protein Complexes: The CASP13-CAPRI Experiment. *Proteins Struct. Funct. Bioinforma.* **2019**, *0* (ja), prot.25838.

(58) 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.* **2003**, *125* (7), 1731–1737.

(59) van Zundert, G. C. P. et al. The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. *J. Mol. Biol.* **2016**, *428* (4), 720–725.

(60) Zacharias, M. Protein-Protein Docking with a Reduced Protein Model Accounting for Side-Chain Flexibility. *Protein Sci.* **2003**, *12* (6), 1271–1282.

(61) de Vries, S. J.; Zacharias, M. ATTRACT-EM: A New Method for the Computational Assembly of Large Molecular Machines Using Cryo-EM Maps. *PLoS One* **2012**, *7* (12), e49733.

(62) Schindler, C. E. M. et al. SAXS Data Alone Can Generate High-Quality Models of Protein-Protein Complexes. *Structure* **2016**, *24* (8), 1387–1397.

(63) de Vries, S. J. et al. Cryo-EM Data Are Superior to Contact and Interface Information in Integrative Modeling. *Biophys. J.* **2016**, *110* (4), 785–797.

(64) Ritchie, D. W.; Venkatraman, V. Ultra-Fast FFT Protein Docking on Graphics Processors. *Bioinformatics* **2010**, *26* (19), 2398–2405.

(65) Macindoe, G. et al. HexServer: An FFT-Based Protein Docking Server Powered by Graphics Processors. *Nucleic Acids Res.* **2010**, *38* (Web Server), W445–W449.

(66) Alber, F. et al. Integrating Diverse Data for Structure Determination of Macromolecular Assemblies. *Annu. Rev. Biochem.* **2008**, *77* (1), 443–477.

(67) Russel, D. et al. Putting the Pieces Together: Integrative Modeling Platform Software for Structure Determination of Macromolecular Assemblies. *PLoS Biol.* **2012**, *10* (1), e1001244.

(68) Jiménez-García, B. et al. LightDock: A New Multi-Scale Approach to Protein–Protein Docking. *Bioinformatics* **2018**, *34* (1), 49–55.

(69) Roel-Touris, J.; Bonvin, A. M. J. J.; Jiménez-García, B. LightDock Goes Information-Driven. *Bioinformatics* **2019**.

(70) Wang, C.; Bradley, P.; Baker, D. Protein–Protein Docking with Backbone Flexibility. *J. Mol. Biol.* **2007**, *373* (2), 503–519.

(71) Chaudhury, S. et al. Benchmarking and Analysis of Protein Docking Performance in Rosetta v3.2. *PLoS One* **2011**, *6* (8), e22477.

(72) Flavell, R. A. et al. Site-Directed Mutagenesis: Effect of an Extracistronic Mutation on the in Vitro Propagation of Bacteriophage Qbeta RNA. *Proc. Natl. Acad. Sci.* **1975**, *72* (1), 367–371.

(73) Shortle, D.; Nathans, D. Local Mutagenesis: A Method for Generating Viral Mutants with Base Substitutions in Preselected Regions of the Viral Genome. *Proc. Natl. Acad. Sci.* **1978**, *75* (5), 2170–2174.

(74) Shortle, D.; DiMaio, D.; Nathans, D. Directed Mutagenesis. *Annu. Rev. Genet.* **1981**, *15* (1), 265–294.

(75) Heckman, K. L.; Pease, L. R. Gene Splicing and Mutagenesis by PCR-Driven Overlap Extension. *Nat. Protoc.* **2007**, *2* (4), 924–932.

(76) Vajdos, F. F. et al. Comprehensive Functional Maps of the Antigen-Binding Site of an Anti-ErbB2 Antibody Obtained with Shotgun Scanning Mutagenesis. *J. Mol. Biol.* **2002**, *320* (2), 415–428.

(77) Yu, E. W. et al. A Periplasmic Drug-Binding Site of the AcrB Multidrug Efflux Pump: A Crystallographic and Site-Directed Mutagenesis Study. *J. Bacteriol.* **2005**, *187* (19), 6804–6815.

(78) Ashkenazi, A. et al. Mapping the CD4 Binding Site for Human Immunodeficiency Virus by Alanine-Scanning Mutagenesis. *Proc. Natl. Acad. Sci.* **1990**, *87* (18), 7150–7154.

(79) Bill, A. et al. High Throughput Mutagenesis for Identification of Residues Regulating Human Prostacyclin (HIP) Receptor Expression and Function. *PLoS One* **2014**, *9* (6), e97973.

(80) Heydenreich, F. M. et al. High-Throughput Mutagenesis Using a Two-Fragment PCR Approach. *Sci. Rep.* **2017**, *7* (1), 6787.

(81) Luo, Y. et al. Integrative Analysis of CRISPR/Cas9 Target Sites in the Human HBB Gene. *Biomed Res. Int.* **2015**, *2015*, 1–9.

(82) Zhang, B. et al. Exploiting the CRISPR/Cas9 System for Targeted Genome Mutagenesis in Petunia. *Sci. Rep.* **2016**, *6* (1), 20315.

(83) The PyMOL Molecular Graphics System, Version 1.8 Schrödinger, LLC.

(84) Goddard, T. D. et al. UCSF ChimeraX: Meeting Modern Challenges in Visualization and Analysis. *Protein Sci.* **2018**, *27* (1), 14–25.

(85) Englander, S. W.; Kallenbach, N. R. Hydrogen Exchange and Structural Dynamics of Proteins and Nucleic Acids. *Q. Rev. Biophys.* **1983**, *16* (4), 521–655.

(86) Konermann, L.; Pan, J.; Liu, Y.-H. Hydrogen Exchange Mass Spectrometry for Studying Protein Structure and Dynamics. *Chem. Soc. Rev.* **2011**, *40* (3), 1224–1234.

(87) Masson, G. R. et al. Recommendations for Performing, Interpreting and Reporting Hydrogen Deuterium Exchange Mass Spectrometry (HDX-MS) Experiments. *Nat. Methods* **2019**, *16* (7), 595–602.

(88) Rey, M. et al. Mass Spec Studio for Integrative Structural Biology. *Structure* **2014**, *22* (10), 1538–1548.

(89) Sinz, A. et al. Chemical Cross-Linking and Native Mass Spectrometry: A Fruitful Combination for Structural Biology. *Protein Sci.* **2015**, *24* (8), 1193–1209.
(90) Leitner, A. et al. Crosslinking and Mass Spectrometry: An Integrated Technology to Understand the Structure and Function of Molecular Machines. *Trends Biochem. Sci.* **2016**, *41* (1), 20–32.
(91) Heck, A. J. R. Native Mass Spectrometry: A Bridge between Interactomics and Structural Biology. *Nat. Methods* **2008**, *5* (11), 927–933.
(92) Holding, A. N. XL-MS: Protein Cross-Linking Coupled with Mass Spectrometry. *Methods* **2015**, *89*, 54–63.
(93) Rappsilber, J. et al. A Generic Strategy To Analyze the Spatial Organization of Multi-Protein Complexes by Cross-Linking and Mass Spectrometry. *Anal. Chem.* **2000**, *72* (2), 267–275.
(94) Maiolica, A. et al. Structural Analysis of Multiprotein Complexes by Cross-Linking, Mass Spectrometry, and Database Searching. *Mol. Cell. Proteomics* **2007**, *6* (12), 2200–2211.
(95) Rappsilber, J. The Beginning of a Beautiful Friendship: Cross-Linking/Mass Spectrometry and Modelling of Proteins and Multi-Protein Complexes. *J. Struct. Biol.* **2011**, *173* (3), 530–540.
(96) Chen, Z. A.; Rappsilber, J. Protein Dynamics in Solution by Quantitative Crosslinking/Mass Spectrometry. *Trends Biochem. Sci.* **2018**, *43* (11), 908–920.
(97) Schneider, M.; Belsom, A.; Rappsilber, J. Protein Tertiary Structure by Crosslinking/Mass Spectrometry. *Trends Biochem. Sci.* **2018**, *43* (3), 157–169.
(98) Chen, Z. A.; Rappsilber, J. Quantitative Cross-Linking/Mass Spectrometry to Elucidate Structural Changes in Proteins and Their Complexes. *Nat. Protoc.* **2019**, *14* (1), 171–201.
(99) Zhang, H. et al. Identification of Protein-Protein Interactions and Topologies in Living Cells with Chemical Cross-Linking and Mass Spectrometry. *Mol. Cell. Proteomics* **2009**, *8* (3), 409–420.
(100) Liu, F. et al. Proteome-Wide Profiling of Protein Assemblies by Cross-Linking Mass Spectrometry. *Nat. Methods* **2015**, *12* (12), 1179–1184.
(101) Fasci, D. et al. Histone Interaction Landscapes Visualized by Crosslinking Mass Spectrometry in Intact Cell Nuclei. *Mol. Cell. Proteomics* **2018**, *17* (10), 2018–2033.
(102) Iacobucci, C. et al. First Community-Wide, Comparative Cross-Linking Mass Spectrometry Study. *Anal. Chem.* **2019**, *91* (11), 6953–6961.
(103) Liu, F. et al. Optimized Fragmentation Schemes and Data Analysis Strategies for Proteome-Wide Cross-Link Identification. *Nat. Commun.* **2017**, *8* (1), 15473.
(104) Klykov, O. et al. Efficient and Robust Proteome-Wide Approaches for Cross-Linking Mass Spectrometry. *Nat. Protoc.* **2018**, *13* (12), 2964–2990.
(105) de Graaf, S. C. et al. Cross-ID: Analysis and Visualization of Complex XL–MS-Driven Protein Interaction Networks. *J. Proteome Res.* **2019**, *18* (2), 642–651.
(106) van Zundert, G. C. P.; Bonvin, A. M. J. J. DisVis: Quantifying and Visualizing Accessible Interaction Space of Distance-Restrained Biomolecular Complexes: Fig. 1. *Bioinformatics* **2015**, *31* (19), 3222–3224.
(107) van Zundert, G. C. P. et al. The DisVis and PowerFit Web Servers: Explorative and Integrative Modeling of Biomolecular Complexes. *J. Mol. Biol.* **2017**, *429* (3), 399–407.
(108) Kay, L. E. NMR Studies of Protein Structure and Dynamics. *J. Magn. Reson.* **2005**, *173* (2), 193–207.
(109) Cavalli, A. et al. Protein Structure Determination from NMR Chemical Shifts. *Proc. Natl. Acad. Sci.* **2007**, *104* (23), 9615–9620.
(110) Fushman, D.; Cowburn, D. Model-Independent Analysis of 15 N Chemical Shift

Anisotropy from NMR Relaxation Data. Ubiquitin as a Test Example. *J. Am. Chem. Soc.* **1998**, *120* (28), 7109–7110.

(111) Fushman, D.; Cowburn, D. The Effect of Noncollinearity of 15N-1H Dipolar and 15N CSA Tensors and Rotational Anisotropy on 15N Relaxation, CSA/Dipolar Cross Correlation, and TROSY. *J. Biomol. NMR* **1999**, *13* (2), 139–147.

(112) Petros, A. M.; Mueller, L.; Kopple, K. D. NMR Identification of Protein Surfaces Using Paramagnetic Probes. *Biochemistry* **1990**, *29* (43), 10041–10048.

(113) Matei, E.; Gronenborn, A. M. 19 F Paramagnetic Relaxation Enhancement: A Valuable Tool for Distance Measurements in Proteins. *Angew. Chemie Int. Ed.* **2016**, *55* (1), 150–154.

(114) Öster, C. et al. Characterization of Protein–Protein Interfaces in Large Complexes by Solid-State NMR Solvent Paramagnetic Relaxation Enhancements. *J. Am. Chem. Soc.* **2017**, *139* (35), 12165–12174.

(115) Kato, K.; Yamaguchi, T. Paramagnetic NMR Probes for Characterization of the Dynamic Conformations and Interactions of Oligosaccharides. *Glycoconj. J.* **2015**, *32* (7), 505–513.

(116) Venditti, V.; Fawzi, N. L. Probing the Atomic Structure of Transient Protein Contacts by Paramagnetic Relaxation Enhancement Solution NMR. In *Methods in Molecular Biology*; 2018; Vol. 1688, pp 243–255.

(117) Griffin, R. G. Dipolar Recoupling in MAS Spectra of Biological Solids. *Nat. Struct. Biol.* **1998**, *5* (1 SUPPL. 1), 508–512.

(118) Lange, A. et al. A Concept for Rapid Protein-Structure Determination by Solid-State NMR Spectroscopy. *Angew. Chemie Int. Ed.* **2005**, *44* (14), 2089–2092.

(119) Samoson, A.; Lippmaa, E.; Pines, A. High Resolution Solid-State N.M.R. *Mol. Phys.* **1988**, *65* (4), 1013–1018.

(120) Kaptein, R.; Wagner, G. Integrative Methods in Structural Biology. *J. Biomol. NMR* **2019**, *73* (6–7), 261–263.

(121) Visscher, K. M. et al. Supramolecular Organization and Functional Implications of K + Channel Clusters in Membranes. *Angew. Chemie Int. Ed.* **2017**, *56* (43), 13222–13227.

(122) Cross, T. A.; Opella, S. J. Solid-State NMR Structural Studies of Peptides and Proteins in Membranes. *Curr. Opin. Struct. Biol.* **1994**, *4* (4), 574–581.

(123) McDermott, A. Structure and Dynamics of Membrane Proteins by Magic Angle Spinning Solid-State NMR. *Annu. Rev. Biophys.* **2009**, *38* (1), 385–403.

(124) Freedberg, D. I.; Selenko, P. Live Cell NMR. *Annu. Rev. Biophys.* **2014**, *43* (1), 171–192.

(125) Selenko, P.; Wagner, G. Looking into Live Cells with In-Cell NMR Spectroscopy. *J. Struct. Biol.* **2007**, *158* (2), 244–253.

(126) Theillet, F.-X. et al. Structural Disorder of Monomeric α-Synuclein Persists in Mammalian Cells. *Nature* **2016**, *530* (7588), 45–50.

(127) Narasimhan, S. et al. DNP-Supported Solid-State NMR Spectroscopy of Proteins Inside Mammalian Cells. *Angew. Chemie Int. Ed.* **2019**, *58* (37), 12969–12973.

(128) Kuhlbrandt, W. The Resolution Revolution. *Science (80-.).* **2014**, *343* (6178), 1443–1444.

(129) McMullan, G.; Faruqi, A. R.; Henderson, R. Direct Electron Detectors. In *The Resolution Revolution: Recent Advances In cryoEM*; 2016; pp 1–17.

(130) Kim, D. N.; Sanbonmatsu, K. Y. Tools for the Cryo-EM Gold Rush: Going from the Cryo-EM Map to the Atomistic Model. *Biosci. Rep.* **2017**, *37* (6), BSR20170072.

(131) Lawson, C. L. et al. EMDataBank.Org: Unified Data Resource for CryoEM. *Nucleic Acids Res.* **2011**, *39* (Database), D456–D464.

(132) Cheng, Y. et al. A Primer to Single-Particle Cryo-Electron Microscopy. *Cell* **2015**, *161*

(3), 438–449.
(133) Cheng, Y. Single-Particle Cryo-EM at Crystallographic Resolution. *Cell* **2015**, *161* (3), 450–457.
(134) Orzechowski, M.; Tama, F. Flexible Fitting of High-Resolution X-Ray Structures into Cryoelectron Microscopy Maps Using Biased Molecular Dynamics Simulations. *Biophys. J.* **2008**, *95* (12), 5692–5705.
(135) Trabuco, L. G. et al. Flexible Fitting of Atomic Structures into Electron Microscopy Maps Using Molecular Dynamics. *Structure* **2008**, *16* (5), 673–683.
(136) Villa, E.; Lasker, K. Finding the Right Fit: Chiseling Structures out of Cryo-Electron Microscopy Maps. *Curr. Opin. Struct. Biol.* **2014**, *25*, 118–125.
(137) Brown, A. et al. Tools for Macromolecular Model Building and Refinement into Electron Cryo-Microscopy Reconstructions. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2015**, *71* (1), 136–153.
(138) C.P.van Zundert, G.; M.J.J. Bonvin, A. Fast and Sensitive Rigid-Body Fitting into Cryo-EM Density Maps with PowerFit. *AIMS Biophys.* **2015**, *2* (2), 73–87.
(139) van Zundert, G. C. P.; Melquiond, A. S. J.; Bonvin, A. M. J. J. Integrative Modeling of Biomolecular Complexes: HADDOCKing with Cryo-Electron Microscopy Data. *Structure* **2015**, *23* (5), 949–960.
(140) van Zundert, G. C. P.; Bonvin, A. M. J. J. Defining the Limits and Reliability of Rigid-Body Fitting in Cryo-EM Maps Using Multi-Scale Image Pyramids. *J. Struct. Biol.* **2016**, *195* (2), 252–258.
(141) Wang, R. Y. R. et al. Automated Structure Refinement of Macromolecular Assemblies from Cryo-EM Maps Using Rosetta. *Elife* **2016**, *5* (September2016).
(142) Fan, X. et al. Single Particle Cryo-EM Reconstruction of 52 KDa Streptavidin at 3.2 Angstrom Resolution. *Nat. Commun.* **2019**, *10* (1), 2386.
(143) Liu, Y.; Huynh, D. T.; Yeates, T. O. A 3.8 Å Resolution Cryo-EM Structure of a Small Protein Bound to an Imaging Scaffold. *Nat. Commun.* **2019**, *10* (1), 1864.
(144) Herzik, M. A.; Wu, M.; Lander, G. C. High-Resolution Structure Determination of Sub-100 KDa Complexes Using Conventional Cryo-EM. *Nat. Commun.* **2019**, *10* (1), 1032.
(145) Cardone, G.; Heymann, J. B.; Steven, A. C. One Number Does Not Fit All: Mapping Local Variations in Resolution in Cryo-EM Reconstructions. *J. Struct. Biol.* **2013**, *184* (2), 226–236.
(146) Kucukelbir, A.; Sigworth, F. J.; Tagare, H. D. Quantifying the Local Resolution of Cryo-EM Density Maps. *Nat. Methods* **2014**, *11* (1), 63–65.
(147) Bonomi, M.; Vendruscolo, M. Determination of Protein Structural Ensembles Using Cryo-Electron Microscopy. *Curr. Opin. Struct. Biol.* **2019**, *56*, 37–45.
(148) Moscovich, A. et al. Cryo-EM Reconstruction of Continuous Heterogeneity by Laplacian Spectral Volumes. **2019**.
(149) Bonomi, M.; Pellarin, R.; Vendruscolo, M. Simultaneous Determination of Protein Structure and Dynamics Using Cryo-Electron Microscopy. *Biophys. J.* **2018**, *114* (7), 1604–1613.
(150) Frank, J. Single-Particle Imaging of Macromolecules by Cryo-Electron Microscopy. *Annu. Rev. Biophys. Biomol. Struct.* **2002**, *31* (1), 303–319.
(151) Kaplan, M. et al. Probing a Cell-Embedded Megadalton Protein Complex by DNP-Supported Solid-State NMR. *Nat. Methods* **2015**, *12* (7), 649–652.
(152) Bardiaux, B. et al. Dynamics of a Type 2 Secretion System Pseudopilus Unraveled by Complementary Approaches. *J. Biomol. NMR* **2019**, *73* (6–7), 293–303.
(153) Mertens, H. D. T.; Svergun, D. I. Structural Characterization of Proteins and Complexes Using Small-Angle X-Ray Solution Scattering. *J. Struct. Biol.* **2010**, *172* (1), 128–141.
(154) Brosey, C. A.; Tainer, J. A. Evolving SAXS Versatility: Solution X-Ray Scattering for

Macromolecular Architecture, Functional Landscapes, and Integrative Structural Biology. *Curr. Opin. Struct. Biol.* **2019**, *58*, 197–213.

(155) Putnam, C. D. et al. X-Ray Solution Scattering (SAXS) Combined with Crystallography and Computation: Defining Accurate Macromolecular Structures, Conformations and Assemblies in Solution. *Q. Rev. Biophys.* **2007**, *40* (3), 191–285.

(156) Schneidman-Duhovny, D.; Kim, S.; Sali, A. Integrative Structural Modeling with Small Angle X-Ray Scattering Profiles. *BMC Struct. Biol.* **2012**, *12* (1), 17.

(157) Schneidman-Duhovny, D.; Hammel, M. Modeling Structure and Dynamics of Protein Complexes with SAXS Profiles. In *Methods in Molecular Biology*; 2018; Vol. 1764, pp 449–473.

(158) Svergun, D. I.; Koch, M. H. J. Small-Angle Scattering Studies of Biological Macromolecules in Solution. *Reports Prog. Phys.* **2003**, *66* (10), 1735–1782.

(159) Petoukhov, M. V.; Svergun, D. I. Global Rigid Body Modeling of Macromolecular Complexes against Small-Angle Scattering Data. *Biophys. J.* **2005**, *89* (2), 1237–1250.

(160) Tria, G. et al. Advanced Ensemble Modelling of Flexible Macromolecules Using X-Ray Solution Scattering. *IUCrJ* **2015**, *2* (2), 207–217.

(161) Franke, D.; Jeffries, C. M.; Svergun, D. I. Correlation Map, a Goodness-of-Fit Test for One-Dimensional X-Ray Scattering Spectra. *Nat. Methods* **2015**, *12* (5), 419–422.

(162) Karaca, E.; Bonvin, A. M. J. J. On the Usefulness of Ion-Mobility Mass Spectrometry and SAXS Data in Scoring Docking Decoys. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2013**, *69* (5), 683–694.

(163) Pons, C. et al. Structural Characterization of Protein–Protein Complexes by Integrating Computational Docking with Small-Angle Scattering Data. *J. Mol. Biol.* **2010**, *403* (2), 217–230.

(164) Jiménez-García, B. et al. PyDockSAXS: Protein–Protein Complex Structure by SAXS and Computational Docking. *Nucleic Acids Res.* **2015**, *43* (W1), W356–W361.

(165) Karaca, E. et al. M3: An Integrative Framework for Structure Determination of Molecular Machines. *Nat. Methods* **2017**, *14* (9), 897–902.

(166) Hura, G. L. et al. Robust, High-Throughput Solution Structural Analyses by Small Angle X-Ray Scattering (SAXS). *Nat. Methods* **2009**, *6* (8), 606–612.

(167) Dimura, M. et al. Quantitative FRET Studies and Integrative Modeling Unravel the Structure and Dynamics of Biomolecular Systems. *Curr. Opin. Struct. Biol.* **2016**, *40*, 163–185.

(168) Bonomi, M. et al. Determining Protein Complex Structures Based on a Bayesian Model of in Vivo Förster Resonance Energy Transfer (FRET) Data. *Mol. Cell. Proteomics* **2014**, *13* (11), 2812–2823.

(169) Kilic, S. et al. Single-Molecule FRET Reveals Multiscale Chromatin Dynamics Modulated by HP1α. *Nat. Commun.* **2018**, *9* (1), 235.

(170) Lehmann, K. et al. Multiple Interaction Modes of the Nucleosomal Histone H3 N-Terminal Tail Revealed by High Precision Single-Molecule FRET. *Biophys. J.* **2019**, *116* (3), 468a-469a.

(171) Hellenkamp, B. et al. Precision and Accuracy of Single-Molecule FRET Measurements—a Multi-Laboratory Benchmark Study. *Nat. Methods* **2018**, *15* (9), 669–676.

(172) Möckel, C. et al. Integrated NMR, Fluorescence, and Molecular Dynamics Benchmark Study of Protein Mechanics and Hydrodynamics. *J. Phys. Chem. B* **2019**, *123* (7), 1453–1480.

(173) Pannier, M. et al. Dead-Time Free Measurement of Dipole–Dipole Interactions between Electron Spins. *J. Magn. Reson.* **2000**, *142* (2), 331–340.

(174) Jeschke, G. DEER Distance Measurements on Proteins. *Annu. Rev. Phys. Chem.* **2012**,

63 (1), 419–446.

(175) Fehr, N. et al. Modeling of the N-Terminal Section and the Lumenal Loop of Trimeric Light Harvesting Complex II (LHCII) by Using EPR. *J. Biol. Chem.* **2015**, *290* (43), 26007–26020.

(176) Stadtmueller, B. M. et al. DEER Spectroscopy Measurements Reveal Multiple Conformations of HIV-1 SOSIP Envelopes That Show Similarities with Envelopes on Native Virions. *Immunity* **2018**, *49* (2), 235-246.e4.

(177) Glaenzer, J.; Peter, M. F.; Hagelueken, G. Studying Structure and Function of Membrane Proteins with PELDOR/DEER Spectroscopy – The Crystallographers' Perspective. *Methods* **2018**, *147*, 163–175.

(178) Masliah, G. et al. Structural Basis of Si RNA Recognition by TRBP Double-stranded RNA Binding Domains. *EMBO J.* **2018**, *37* (6), e97089.

(179) Szurmant, H.; Weigt, M. Inter-Residue, Inter-Protein and Inter-Family Coevolution: Bridging the Scales. *Curr. Opin. Struct. Biol.* **2018**, *50*, 26–32.

(180) Sutto, L. et al. From Residue Coevolution to Protein Conformational Ensembles and Functional Dynamics. *Proc. Natl. Acad. Sci.* **2015**, *112* (44), 13567–13572.

(181) Uguzzoni, G. et al. Large-Scale Identification of Coevolution Signals across Homo-Oligomeric Protein Interfaces by Direct Coupling Analysis. *Proc. Natl. Acad. Sci.* **2017**, *114* (13), E2662–E2671.

(182) Balakrishnan, S. et al. Learning Generative Models for Protein Fold Families. *Proteins Struct. Funct. Bioinforma.* **2011**, *79* (4), 1061–1078.

(183) Kamisetty, H.; Ovchinnikov, S.; Baker, D. Assessing the Utility of Coevolution-Based Residue-Residue Contact Predictions in a Sequence- and Structure-Rich Era. *Proc. Natl. Acad. Sci.* **2013**, *110* (39), 15674–15679.

(184) Yu, J. et al. InterEvDock: A Docking Server to Predict the Structure of Protein–Protein Interactions Using Evolutionary Information. *Nucleic Acids Res.* **2016**, *44* (W1), W542–W549.

(185) Hopf, T. A. et al. Sequence Co-Evolution Gives 3D Contacts and Structures of Protein Complexes. *Elife* **2014**, *3*.

(186) Yu, J. et al. Lessons from (Co-)Evolution in the Docking of Proteins and Peptides for CAPRI Rounds 28-35. *Proteins Struct. Funct. Bioinforma.* **2017**, *85* (3), 378–390.

(187) Ovchinnikov, S.; Kamisetty, H.; Baker, D. Robust and Accurate Prediction of Residue–Residue Interactions across Protein Interfaces Using Evolutionary Information. *Elife* **2014**, *3* (3).

(188) 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.* **2018**, *46* (W1), W408–W416.

(189) Zeng, B.; Hönigschmid, P.; Frishman, D. Residue Co-Evolution Helps Predict Interaction Sites in α-Helical Membrane Proteins. *J. Struct. Biol.* **2019**, *206* (2), 156–169.

(190) Stansfeld, P. J. Computational Studies of Membrane Proteins: From Sequence to Structure to Simulation. *Curr. Opin. Struct. Biol.* **2017**, *45*, 133–141.

(191) Cong, Q. et al. Protein Interaction Networks Revealed by Proteome Coevolution. *Science (80-.).* **2019**, *365* (6449), 185–189.

(192) Wu, Q. et al. Protein Contact Prediction Using Metagenome Sequence Data and Residual Neural Networks. *Bioinformatics* **2019**.

(193) Shrestha, R. et al. Assessing the Accuracy of Contact Predictions in CASP13. *Proteins Struct. Funct. Bioinforma.* **2019**, *0* (ja), prot.25819.

(194) Dörr, J. M. et al. The Styrene–Maleic Acid Copolymer: A Versatile Tool in Membrane Research. *Eur. Biophys. J.* **2016**, *45* (1), 3–21.

(195) Hellwig, N. et al. Native Mass Spectrometry Goes More Native: Investigation of Membrane Protein Complexes Directly from SMALPs. *Chem. Commun.* **2018**, *54* (97), 13702–13705.

(196) Simon, K. S.; Pollock, N. L.; Lee, S. C. Membrane Protein Nanoparticles: The Shape of Things to Come. *Biochem. Soc. Trans.* **2018**, *46* (6), 1495–1504.

(197) Parmar, M. et al. Using a SMALP Platform to Determine a Sub-Nm Single Particle Cryo-EM Membrane Protein Structure. *Biochim. Biophys. Acta - Biomembr.* **2018**, *1860* (2), 378–383.

(198) Dörr, J. M. et al. Detergent-Free Isolation, Characterization, and Functional Reconstitution of a Tetrameric K + Channel: The Power of Native Nanodiscs. *Proc. Natl. Acad. Sci.* **2014**, *111* (52), 18607–18612.

(199) Radoicic, J.; Park, S. H.; Opella, S. J. Macrodiscs Comprising SMALPs for Oriented Sample Solid-State NMR Spectroscopy of Membrane Proteins. *Biophys. J.* **2018**, *115* (1), 22–25.

(200) Almeida, J. G. et al. Membrane Proteins Structures: A Review on Computational Modeling Tools. *Biochim. Biophys. Acta - Biomembr.* **2017**, *1859* (10), 2021–2039.

(201) Shimizu, K. et al. Comparative Analysis of Membrane Protein Structure Databases. *Biochim. Biophys. Acta - Biomembr.* **2018**, *1860* (5), 1077–1091.

(202) Kelm, S.; Shi, J.; Deane, C. M. MEDELLER: Homology-Based Coordinate Generation for Membrane Proteins. *Bioinformatics* **2010**, *26* (22), 2833–2840.

(203) Ebejer, J.-P. et al. Memoir: Template-Based Structure Prediction for Membrane Proteins. *Nucleic Acids Res.* **2013**, *41* (W1), W379–W383.

(204) Šali, A.; Blundell, T. L. Comparative Protein Modelling by Satisfaction of Spatial Restraints. *J. Mol. Biol.* **1993**, *234* (3), 779–815.

(205) Nikolaev, D. M. et al. A Comparative Study of Modern Homology Modeling Algorithms for Rhodopsin Structure Prediction. *ACS Omega* **2018**, *3* (7), 7555–7566.

(206) Lomize, M. A. et al. OPM: Orientations of Proteins in Membranes Database. *Bioinformatics* **2006**, *22* (5), 623–625.

(207) Stansfeld, P. J. et al. MemProtMD: Automated Insertion of Membrane Protein Structures into Explicit Lipid Membranes. *Structure* **2015**, *23* (7), 1350–1361.

(208) 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.* **2019**, *47* (D1), D390–D397.

(209) Hauser, A. S. et al. Trends in GPCR Drug Discovery: New Agents, Targets and Indications. *Nat. Rev. Drug Discov.* **2017**, *16* (12), 829–842.

(210) Munk, C. et al. An Online Resource for GPCR Structure Determination and Analysis. *Nat. Methods* **2019**, *16* (2), 151–162.

(211) Marrink, S. J. et al. The MARTINI Force Field: Coarse Grained Model for Biomolecular Simulations. *J. Phys. Chem. B* **2007**, *111* (27), 7812–7824.

(212) Monticelli, L. et al. The MARTINI Coarse-Grained Force Field: Extension to Proteins. *J. Chem. Theory Comput.* **2008**, *4* (5), 819–834.

(213) de Jong, D. H. et al. Improved Parameters for the Martini Coarse-Grained Protein Force Field. *J. Chem. Theory Comput.* **2013**, *9* (1), 687–697.

(214) Arnarez, C. et al. Dry Martini, a Coarse-Grained Force Field for Lipid Membrane Simulations with Implicit Solvent. *J. Chem. Theory Comput.* **2015**, *11* (1), 260–275.

(215) Roel-Touris, J. et al. Less Is More: Coarse-Grained Integrative Modeling of Large Biomolecular Assemblies with HADDOCK. *J. Chem. Theory Comput.* **2019**, *In press*, acs.jctc.9b00310.

(216) Honorato, R. V.; Roel-Touris, J.; Bonvin, A. M. J. J. MARTINI-Based Protein-DNA Coarse-Grained HADDOCKing. *Front. Mol. Biosci.* **2019**, *In press*.

(217) Kurcinski, M. et al. CABS-Dock Web Server for the Flexible Docking of Peptides to Proteins without Prior Knowledge of the Binding Site. *Nucleic Acids Res.* **2015**, *43* (W1), W419–W424.

(218) Kurcinski, M. et al. CABS-Dock Standalone: A Toolbox for Flexible Protein–Peptide Docking. *Bioinformatics* **2019**, *35* (20), 4170–4172.

(219) Alford, R. F. et al. An Integrated Framework Advancing Membrane Protein Modeling and Design. *PLOS Comput. Biol.* **2015**, *11* (9), e1004398.

(220) Koehler Leman, J.; Mueller, B. K.; Gray, J. J. Expanding the Toolkit for Membrane Protein Modeling in Rosetta. *Bioinformatics* **2016**, *33* (5), btw716.

(221) Viswanath, S.; Ravikant, D. V. S.; Elber, R. DOCK/PIERR: Web Server for Structure Prediction of Protein–Protein Complexes. In *Methods in Molecular Biology*; 2014; pp 199–207.

(222) Viswanath, S. et al. Extension of a Protein Docking Algorithm to Membranes and Applications to Amyloid Precursor Protein Dimerization. *Proteins Struct. Funct. Bioinforma.* **2015**, *83* (12), 2170–2185.

(223) Hurwitz, N.; Schneidman-Duhovny, Di.; Wolfson, H. J. Memdock: An α-Helical Membrane Protein Docking Algorithm. *Bioinformatics* **2016**, *32* (16), 2444–2450.

(224) Koukos, P. I. et al. A Membrane Protein Complex Docking Benchmark. *J. Mol. Biol.* **2018**, *430* (24), 5246–5256.

(225) Koukos, P.; Bonvin, A. HADDOCK Membrane Protein-Protein Complex Models. SBGrid Data Bank 2018.

(226) 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.* **2018**, *25* (11), 1327-1336.e4.

(227) Pagadala, N. S.; Syed, K.; Tuszynski, J. Software for Molecular Docking: A Review. *Biophys. Rev.* **2017**, *9* (2), 91–102.

(228) Bonomi, M. et al. Bayesian Weighing of Electron Cryo-Microscopy Data for Integrative Structural Modeling. *Structure* **2019**, *27* (1), 175-188.e6.

(229) Rieping, W. Inferential Structure Determination. *Science (80-.).* **2005**, *309* (5732), 303–306.

(230) Tamura, K.; Hara-Nishimura, I. The Molecular Architecture of the Plant Nuclear Pore Complex. *J. Exp. Bot.* **2013**, *64* (4), 823–832.

(231) Kim, S. J. et al. Integrative Structure and Functional Anatomy of a Nuclear Pore Complex. *Nature* **2018**, *555* (7697), 475–482.

(232) Rieping, W.; Nilges, M.; Habeck, M. ISD: A Software Package for Bayesian NMR Structure Calculation. *Bioinformatics* **2008**, *24* (8), 1104–1105.

(233) Shahid, S. A. et al. Membrane-Protein Structure Determination by Solid-State NMR Spectroscopy of Microcrystals. *Nat. Methods* **2012**, *9* (12), 1212–1217.

(234) Habenstein, B. et al. Hybrid Structure of the Type 1 Pilus of Uropathogenic Escherichia Coli. *Angew. Chemie Int. Ed.* **2015**, *54* (40), 11691–11695.

(235) Carstens, S.; Nilges, M.; Habeck, M. Inferential Structure Determination of Chromosomes from Single-Cell Hi-C Data. *PLOS Comput. Biol.* **2016**, *12* (12), e1005292.

(236) Chen, Y.-L.; Habeck, M. Data-Driven Coarse Graining of Large Biomolecular Structures. *PLoS One* **2017**, *12* (8), e0183057.

(237) Van Der Spoel, D. et al. GROMACS: Fast, Flexible, and Free. *J. Comput. Chem.* **2005**, *26* (16), 1701–1718.

(238) Abraham, M. J. et al. GROMACS: High Performance Molecular Simulations through Multi-Level Parallelism from Laptops to Supercomputers. *SoftwareX* **2015**, *1–2*, 19–25.

(239) Burley, S. K. et al. PDB-Dev: A Prototype System for Depositing Integrative/Hybrid

Structural Models. *Structure* **2017**, *25* (9), 1317–1318.

(240) Moult, J. et al. Critical Assessment of Methods of Protein Structure Prediction (CASP)-Round XII. *Proteins Struct. Funct. Bioinforma.* **2018**, *86*, 7–15.

(241) Janin, J. et al. CAPRI: A Critical Assessment of PRedicted Interactions. *Proteins Struct. Funct. Genet.* **2003**, *52* (1), 2–9.

(242) Gathiaka, S. et al. D3R Grand Challenge 2015: Evaluation of Protein–Ligand Pose and Affinity Predictions. *J. Comput. Aided. Mol. Des.* **2016**, *30* (9), 651–668.

(243) Gaieb, Z. et al. D3R Grand Challenge 2: Blind Prediction of Protein–Ligand Poses, Affinity Rankings, and Relative Binding Free Energies. *J. Comput. Aided. Mol. Des.* **2018**, *32* (1), 1–20.

(244) Vreven, T. et al. Updates to the Integrated Protein–Protein Interaction Benchmarks: Docking Benchmark Version 5 and Affinity Benchmark Version 2. *J. Mol. Biol.* **2015**, *427* (19), 3031–3041.

(245) Trellet, M.; Melquiond, A. S. J.; Bonvin, A. M. J. J. A Unified Conformational Selection and Induced Fit Approach to Protein-Peptide Docking. *PLoS One* **2013**, *8* (3), e58769.

(246) London, N.; Movshovitz-Attias, D.; Schueler-Furman, O. The Structural Basis of Peptide-Protein Binding Strategies. *Structure* **2010**, *18* (2), 188–199.

(247) van Dijk, M.; Bonvin, A. M. J. J. A Protein-DNA Docking Benchmark. *Nucleic Acids Res.* **2008**, *36* (14), e88–e88.

(248) Hartshorn, M. J. et al. Diverse, High-Quality Test Set for the Validation of Protein−Ligand Docking Performance. *J. Med. Chem.* **2007**, *50* (4), 726–741.

(249) Lomize, M. A. et al. OPM: Orientations of Proteins in Membranes Database. *Bioinformatics* **2006**, *22* (5), 623–625.

(250) Weiss, D. R. et al. GPCR-Bench: A Benchmarking Set and Practitioners' Guide for G Protein-Coupled Receptor Docking. *J. Chem. Inf. Model.* **2016**, *56* (4), 642–651.

(251) Almeida, J. G. et al. Membrane Proteins Structures: A Review on Computational Modeling Tools. *Biochim. Biophys. Acta - Biomembr.* **2017**, *1859* (10), 2021–2039.

(252) Berman, H. M. et al. The Protein Data Bank, 1999–. In *International Tables for Crystallography*; International Union of Crystallography: Chester, England, 2006; Vol. 28, pp 675–684.

(253) Dutzler, R. et al. Crystal Structure and Functional Characterization of OmpK36, the Osmoporin of Klebsiella Pneumoniae. *Structure* **1999**, *7* (4), 425–434.

(254) Needleman, S. B.; Wunsch, C. D. A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins. *J. Mol. Biol.* **1970**, *48* (3), 443–453.

(255) Henikoff, S.; Henikoff, J. G. Amino Acid Substitution Matrices from Protein Blocks. *Proc. Natl. Acad. Sci.* **1992**, *89* (22), 10915–10919.

(256) 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.* **2003**, *125* (7), 1731–1737.

(257) Jorgensen, W. L. et al. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983**, *79* (2), 926–935.

(258) Karaca, E. et al. Building Macromolecular Assemblies by Information-Driven Docking. *Mol. Cell. Proteomics* **2010**, *9* (8), 1784–1794.

(259) Jorgensen, W. L.; Tirado-Rives, J. The OPLS [Optimized Potentials for Liquid Simulations] Potential Functions for Proteins, Energy Minimizations for Crystals of Cyclic Peptides and Crambin. *J. Am. Chem. Soc.* **1988**, *110* (6), 1657–1666.

(260) Fernández-Recio, J.; Totrov, M.; Abagyan, R. Identification of Protein-Protein Interaction Sites from Docking Energy Landscapes. *J. Mol. Biol.* **2004**, *335* (3), 843–865.

(261) McLachlan, A. D. Rapid Comparison of Protein Structures. *Acta Crystallogr. Sect. A* **1982**, *38* (6), 871–873.

(262) Morin, A. et al. Collaboration Gets the Most out of Software. *Elife* **2013**, *2* (2).

(263) Meyer, P. A. et al. Data Publication with the Structural Biology Data Grid Supports Live Analysis. *Nat. Commun.* **2016**, *7* (1), 10882.

(264) Unwin, N. Refined Structure of the Nicotinic Acetylcholine Receptor at 4Å Resolution. *J. Mol. Biol.* **2005**, *346* (4), 967–989.

(265) Madej, M. G. et al. Evidence for Transmembrane Proton Transfer in a Dihaem-Containing Membrane Protein Complex. *EMBO J.* **2006**, *25* (20), 4963–4970.

(266) Oldham, M. L. et al. Crystal Structure of a Catalytic Intermediate of the Maltose Transporter. *Nature* **2007**, *450* (7169), 515–521.

(267) Jormakka, M. et al. Molecular Mechanism of Energy Conservation in Polysulfide Respiration. *Nat. Struct. Mol. Biol.* **2008**, *15* (7), 730–737.

(268) Morgan, J. L. W.; Strumillo, J.; Zimmer, J. Crystallographic Snapshot of Cellulose Synthesis and Membrane Translocation. *Nature* **2013**, *493* (7431), 181–186.

(269) Xu, K. et al. Crystal Structure of a Folate Energy-Coupling Factor Transporter from Lactobacillus Brevis. *Nature* **2013**, *497* (7448), 268–271.

(270) Bai, X. et al. An Atomic Structure of Human γ-Secretase. *Nature* **2015**, *525* (7568), 212–217.

(271) Buchanan, S. K. et al. Structure of Colicin I Receptor Bound to the R-Domain of Colicin Ia: Implications for Protein Import. *EMBO J.* **2007**, *26* (10), 2594–2604.

(272) Wiener, M. et al. Crystal Structure of Colicin Ia. *Nature* **1997**, *385* (6615), 461–464.

(273) Geibel, S. et al. Structural and Energetic Basis of Folded-Protein Transport by the FimD Usher. *Nature* **2013**, *496* (7444), 243–246.

(274) Puorger, C. et al. Infinite Kinetic Stability against Dissociation of Supramolecular Protein Complexes through Donor Strand Complementation. *Structure* **2008**, *16* (4), 631–642.

(275) Svensson-Ek, M. et al. The X-Ray Crystal Structures of Wild-Type and EQ(I-286) Mutant Cytochrome c Oxidases from Rhodobacter Sphaeroides. *J. Mol. Biol.* **2002**, *321* (2), 329–339.

(276) Qin, L. et al. Identification of Conserved Lipid/Detergent-Binding Sites in a High-Resolution Structure of the Membrane Protein Cytochrome c Oxidase. *Proc. Natl. Acad. Sci.* **2006**, *103* (44), 16117–16122.

(277) Zhou, Y. et al. Chemistry of Ion Coordination and Hydration Revealed by a K+ Channel–Fab Complex at 2.0 Å Resolution. *Nature* **2001**, *414* (6859), 43–48.

(278) Zhou, M. et al. Potassium Channel Receptor Site for the Inactivation Gate and Quaternary Amine Inhibitors. *Nature* **2001**, *411* (6838), 657–661.

(279) Saitoh, Y. et al. Structural Insight into Tight Junction Disassembly by Clostridium Perfringens Enterotoxin. *Science (80-.).* **2015**, *347* (6223), 775–778.

(280) Van Itallie, C. M. et al. Structure of the Claudin-Binding Domain of Clostridium Perfringens Enterotoxin. *J. Biol. Chem.* **2008**, *283* (1), 268–274.

(281) Lau, T.-L. et al. The Structure of the Integrin AIIbβ3 Transmembrane Complex Explains Integrin Transmembrane Signalling. *EMBO J.* **2009**, *28* (9), 1351–1361.

(282) Lau, T.-L. et al. Structure of the Integrin B3 Transmembrane Segment in Phospholipid Bicelles and Detergent Micelles †. *Biochemistry* **2008**, *47* (13), 4008–4016.

(283) Lau, T.-L.; Dua, V.; Ulmer, T. S. Structure of the Integrin AIIb Transmembrane Segment. *J. Biol. Chem.* **2008**, *283* (23), 16162–16168.

(284) Duan, X.; Quiocho, F. A. Structural Evidence for a Dominant Role of Nonpolar Interactions in the Binding of a Transport/Chemosensory Receptor to Its Highly Polar Ligands † , ‡. *Biochemistry* **2002**, *41* (3), 706–712.

(285) Shultis, D. D. et al. Outer Membrane Active Transport: Structure of the BtuB:TonB Complex. *Science (80-.).* **2006**, *312* (5778), 1396–1399.

(286) Cherezov, V. et al. In Meso Structure of the Cobalamin Transporter, BtuB, at 1.95 Å Resolution. *J. Mol. Biol.* **2006**, *364* (4), 716–734.

(287) Ködding, J. et al. Crystal Structure of a 92-Residue C-Terminal Fragment of TonB from Escherichia Coli Reveals Significant Conformational Changes Compared to Structures of Smaller TonB Fragments. *J. Biol. Chem.* **2005**, *280* (4), 3022–3028.

(288) Tanaka, Y. et al. Crystal Structures of SecYEG in Lipidic Cubic Phase Elucidate a Precise Resting and a Peptide-Bound State. *Cell Rep.* **2015**, *13* (8), 1561–1568.

(289) Shinoda, T. et al. Crystal Structure of the Sodium–Potassium Pump at 2.4 Å Resolution. *Nature* **2009**, *459* (7245), 446–450.

(290) Tanabe, H. et al. Crystal Structures of the Human Adiponectin Receptors. *Nature* **2015**, *520* (7547), 312–316.

(291) Bhat, T. N. et al. Bound Water Molecules and Conformational Stabilization Help Mediate an Antigen-Antibody Association. *Proc. Natl. Acad. Sci.* **1994**, *91* (3), 1089–1093.

(292) Stein, A. et al. Helical Extension of the Neuronal SNARE Complex into the Membrane. *Nature* **2009**, *460* (7254), 525–528.

(293) Krieg, S. et al. Heme Uptake across the Outer Membrane as Revealed by Crystal Structures of the Receptor-Hemophore Complex. *Proc. Natl. Acad. Sci.* **2009**, *106* (4), 1045–1050.

(294) Arnoux, P. et al. The Crystal Structure of HasA, a Hemophore Secreted by Serratia Marcescens. *Nat. Struct. Biol.* **1999**, *6* (6), 516–520.

(295) Mineev, K. S. et al. Spatial Structure of the Transmembrane Domain Heterodimer of ErbB1 and ErbB2 Receptor Tyrosine Kinases. *J. Mol. Biol.* **2010**, *400* (2), 231–243.

(296) Bragin, P. E. et al. HER2 Transmembrane Domain Dimerization Coupled with Self-Association of Membrane-Embedded Cytoplasmic Juxtamembrane Regions. *J. Mol. Biol.* **2016**, *428* (1), 52–61.

(297) Gu, Y. et al. Structural Basis of Outer Membrane Protein Insertion by the BAM Complex. *Nature* **2016**, *531* (7592), 64–69.

(298) Albrecht, R.; Zeth, K. Structural Basis of Outer Membrane Protein Biogenesis in Bacteria. *J. Biol. Chem.* **2011**, *286* (31), 27792–27803.

(299) Rasmussen, S. G. F. et al. Structure of a Nanobody-Stabilized Active State of the B2 Adrenoceptor. *Nature* **2011**, *469* (7329), 175–180.

(300) Cherezov, V. et al. High-Resolution Crystal Structure of an Engineered Human 2-Adrenergic G Protein-Coupled Receptor. *Science (80-.).* **2007**, *318* (5854), 1258–1265.

(301) Brumshtein, B. et al. Formation of Amyloid Fibers by Monomeric Light Chain Variable Domains. *J. Biol. Chem.* **2014**, *289* (40), 27513–27525.

(302) Chen, J. et al. A Tweezers-like Motion of the ATP-Binding Cassette Dimer in an ABC Transport Cycle. *Mol. Cell* **2003**, *12* (3), 651–661.

(303) Hino, T. et al. Structural Basis of Biological N2O Generation by Bacterial Nitric Oxide Reductase. *Science (80-.).* **2010**, *330* (6011), 1666–1670.

(304) Stockbridge, R. B. et al. Crystal Structures of a Double-Barrelled Fluoride Ion Channel. *Nature* **2015**, *525* (7570), 548–551.

(305) Main, A. L. et al. The Three-Dimensional Structure of the Tenth Type III Module of Fibronectin: An Insight into RGD-Mediated Interactions. *Cell* **1992**, *71* (4), 671–678.

(306) Dong, H. et al. Structural Basis for Outer Membrane Lipopolysaccharide Insertion. *Nature* **2014**, *511* (7507), 52–56.

(307) Malojčić, G. et al. LptE Binds to and Alters the Physical State of LPS to Catalyze Its Assembly at the Cell Surface. *Proc. Natl. Acad. Sci.* **2014**, *111* (26), 9467–9472.

(308) Penmatsa, A.; Wang, K. H.; Gouaux, E. X-Ray Structure of Dopamine Transporter Elucidates Antidepressant Mechanism. *Nature* **2013**, *503* (7474), 85–90.

(309) Jiang, L. et al. Rezymogenation of Active Urokinase Induced by an Inhibitory Antibody. *Biochem. J.* **2013**, *449* (1), 161–166.

(310) Inaba, K. et al. Crystal Structure of the DsbB-DsbA Complex Reveals a Mechanism of Disulfide Bond Generation. *Cell* **2006**, *127* (4), 789–801.

(311) Ondo-Mbele, E. et al. Intriguing Conformation Changes Associated with the Trans/Cis Isomerization of a Prolyl Residue in the Active Site of the DsbA C33A Mutant. *J. Mol. Biol.* **2005**, *347* (3), 555–563.

(312) Zhou, Y. et al. NMR Solution Structure of the Integral Membrane Enzyme DsbB: Functional Insights into DsbB-Catalyzed Disulfide Bond Formation. *Mol. Cell* **2008**, *31* (6), 896–908.

(313) Dutzler, R. Gating the Selectivity Filter in ClC Chloride Channels. *Science (80-.).* **2003**, *300* (5616), 108–112.

(314) Dutzler, R. et al. X-Ray Structure of a ClC Chloride Channel at 3.0 Å Reveals the Molecular Basis of Anion Selectivity. *Nature* **2002**, *415* (6869), 287–294.

(315) Grover, R. K. et al. A Structurally Distinct Human Mycoplasma Protein That Generically Blocks Antigen-Antibody Union. *Science (80-.).* **2014**, *343* (6171), 656–661.

(316) Noinaj, N. et al. Structural Basis for Iron Piracy by Pathogenic Neisseria. *Nature* **2012**, *483* (7387), 53–58.

(317) Wang, M. et al. “Anion Clamp” Allows Flexible Protein to Impose Coordination Geometry on Metal Ions. *Chem. Commun.* **2015**, *51* (37), 7867–7870.

(318) Sennhauser, G. et al. Drug Export Pathway of Multidrug Exporter AcrB Revealed by DARPin Inhibitors. *PLoS Biol.* **2006**, *5* (1), e7.

(319) Nakashima, R. et al. Structural Basis for the Inhibition of Bacterial Multidrug Exporters. *Nature* **2013**, *500* (7460), 102–106.

(320) Baconguis, I.; Gouaux, E. Structural Plasticity and Dynamic Selectivity of Acid-Sensing Ion Channel–Spider Toxin Complexes. *Nature* **2012**, *489* (7416), 400–405.

(321) Jasti, J. et al. Structure of Acid-Sensing Ion Channel 1 at 1.9 Å Resolution and Low PH. *Nature* **2007**, *449* (7160), 316–323.

(322) Alberts, B. et al. *Molecular Biology of the Cell*; Wilson, J., Hunt, T., Eds.; Garland Science, 2017.

(323) Doerr, A. Membrane Protein Structures. *Nat. Methods* **2008**, *6*, 35.

(324) Heifetz, A. et al. GPCR Structure, Function, Drug Discovery and Crystallography: Report from Academia-Industry International Conference (UK Royal Society) Chicheley Hall, 1–2 September 2014. *Naunyn. Schmiedebergs. Arch. Pharmacol.* **2015**, *388* (8), 883–903.

(325) Bordner, A. J.; Zorman, B.; Abagyan, R. Efficient Molecular Mechanics Simulations of the Folding, Orientation, and Assembly of Peptides in Lipid Bilayers Using an Implicit Atomic Solvation Model. *J. Comput. Aided. Mol. Des.* **2011**, *25* (10), 895–911.

(326) Andrusier, N. et al. Principles of Flexible Protein-Protein Docking. *Proteins Struct. Funct. Bioinforma.* **2008**, *73* (2), 271–289.

(327) Ferreira, L. et al. Molecular Docking and Structure-Based Drug Design Strategies. *Molecules* **2015**, *20* (7), 13384–13421.

(328) Karaca, E.; Bonvin, A. M. J. J. Advances in Integrative Modeling of Biomolecular Complexes. *Methods* **2013**, *59* (3), 372–381.

(329) Kastritis, P. L.; Bonvin, A. M. J. J. Predicting and Dissecting High-Order Molecular Complexity by Information-Driven Biomolecular Docking. In *Antimicrobial drug discovery: emerging strategies*; Tegos, A., Mylonakis, E., Eds.; CABI: Wallingford,

2012; pp 232–246.
(330) Moreira, I. S.; Fernandes, P. A.; Ramos, M. J. Protein-Protein Docking Dealing with the Unknown. *J. Comput. Chem.* **2010**, *31* (2), 317–342.
(331) Vangone, A. et al. Sense and Simplicity in HADDOCK Scoring: Lessons from CASP-CAPRI Round 1. *Proteins Struct. Funct. Bioinforma.* **2017**, *85* (3), 417–423.
(332) Rutten, L. et al. Crystal Structure and Catalytic Mechanism of the LPS 3-O-Deacylase PagL from Pseudomonas Aeruginosa. *Proc. Natl. Acad. Sci.* **2006**, *103* (18), 7071–7076.
(333) Tomaselli, S. et al. NMR-Based Modeling and Binding Studies of a Ternary Complex between Chicken Liver Bile Acid Binding Protein and Bile Acids. *Proteins Struct. Funct. Bioinforma.* **2007**, *69* (1), 177–191.
(334) Wu, A. M. et al. Activity–Structure Correlations in Divergent Lectin Evolution: Fine Specificity of Chicken Galectin CG-14 and Computational Analysis of Flexible Ligand Docking for CG-14 and the Closely Related CG-16. *Glycobiology* **2007**, *17* (2), 165–184.
(335) Arnusch, C. J. et al. The Vancomycin−Nisin(1−12) Hybrid Restores Activity against Vancomycin Resistant Enterococci †. *Biochemistry* **2008**, *47* (48), 12661–12663.
(336) Rutten, L. et al. Active-Site Architecture and Catalytic Mechanism of the Lipid A Deacylase LpxR of Salmonella Typhimurium. *Proc. Natl. Acad. Sci.* **2009**, *106* (6), 1960–1964.
(337) Schneider, T. et al. Plectasin, a Fungal Defensin, Targets the Bacterial Cell Wall Precursor Lipid II. *Science (80-.).* **2010**, *328* (5982), 1168–1172.
(338) Kastritis, P. L.; Rodrigues, J. P. G. L. M.; Bonvin, A. M. J. J. HADDOCK 2P2I : A Biophysical Model for Predicting the Binding Affinity of Protein–Protein Interaction Inhibitors. *J. Chem. Inf. Model.* **2014**, *54* (3), 826–836.
(339) Zheng, W. et al. A Novel Class of Natural FXR Modulators with a Unique Mode of Selective Co-Regulator Assembly. *ChemBioChem* **2017**, *18* (8), 721–725.
(340) Ding, L. et al. Bile Acid Nuclear Receptor FXR and Digestive System Diseases. *Acta Pharm. Sin. B* **2015**, *5* (2), 135–144.
(341) Ali, A. H.; Carey, E. J.; Lindor, K. D. Recent Advances in the Development of Farnesoid X Receptor Agonists. *Ann. Transl. Med.* **2015**, *3* (1), 5.
(342) Omega Toolkit 2.6.4 OpenEye Scientific Software, Santa Fe, NM, USA. 2017.
(343) Feig, M.; Karanicolas, J.; Brooks, C. L. MMTSB Tool Set: Enhanced Sampling and Multiscale Modeling Methods for Applications in Structural Biology. *J. Mol. Graph. Model.* **2004**, *22* (5), 377–395.
(344) Hubbard, S. J.; Thornton, J. M. 'NACCESS' Computer Program. Department of Biochemistry and Molecular Biology, University College London. *Http://Www.Bioinf.Manchester.Ac.Uk/Naccess*. 1993.
(345) McLachlan, A. D. Rapid Comparison of Protein Structures. *Acta Crystallogr. Sect. A* **1982**, *38* (6), 871–873.
(346) Daura, X. et al. Peptide Folding: When Simulation Meets Experiment. *Angew. Chemie Int. Ed.* **1999**, *38* (1–2), 236–240.
(347) Mi, L.-Z. et al. Structural Basis for Bile Acid Binding and Activation of the Nuclear Receptor FXR. *Mol. Cell* **2003**, *11* (4), 1093–1100.
(348) Bass, J. Y. et al. Conformationally Constrained Farnesoid X Receptor (FXR) Agonists: Heteroaryl Replacements of the Naphthalene. *Bioorg. Med. Chem. Lett.* **2011**, *21* (4), 1206–1213.
(349) Akwabi-Ameyaw, A. et al. Conformationally Constrained Farnesoid X Receptor (FXR) Agonists: Naphthoic Acid-Based Analogs of GW 4064. *Bioorg. Med. Chem. Lett.* **2008**, *18* (15), 4339–4343.
(350) Richter, H. G. F. et al. Optimization of a Novel Class of Benzimidazole-Based Farnesoid

X Receptor (FXR) Agonists to Improve Physicochemical and ADME Properties. *Bioorg. Med. Chem. Lett.* **2011**, *21* (4), 1134–1140.

(351) Wang, Y. et al. FmcsR: Mismatch Tolerant Maximum Common Substructure Searching in R. *Bioinformatics* **2013**, *29* (21), 2792–2794.

(352) Cao, Y. et al. ChemmineR: A Compound Mining Framework for R. *Bioinformatics* **2008**, *24* (15), 1733–1734.

(353) Nissink, J. W. M. et al. A New Test Set for Validating Predictions of Protein-Ligand Interaction. *Proteins Struct. Funct. Bioinforma.* **2002**, *49* (4), 457–471.

(354) Akwabi-Ameyaw, A. et al. FXR Agonist Activity of Conformationally Constrained Analogs of GW 4064. *Bioorg. Med. Chem. Lett.* **2009**, *19* (16), 4733–4739.

(355) Schüttelkopf, A. W.; van Aalten, D. M. F. PRODRG : A Tool for High-Throughput Crystallography of Protein–Ligand Complexes. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2004**, *60* (8), 1355–1363.

(356) Gilson, M. K. et al. BindingDB in 2015: A Public Database for Medicinal Chemistry, Computational Chemistry and Systems Pharmacology. *Nucleic Acids Res.* **2016**, *44* (D1), D1045–D1053.

(357) Chang, C.-C.; Lin, C.-J. LIBSVM. *ACM Trans. Intell. Syst. Technol.* **2011**, *2* (3), 1–27.

(358) Vangone, A.; Bonvin, A. M. J. J. Contacts-Based Prediction of Binding Affinity in Protein–Protein Complexes. *Elife* **2015**, *4* (JULY2015), e07454.

(359) Xue, L. C. et al. PRODIGY: A Web Server for Predicting the Binding Affinity of Protein–Protein Complexes. *Bioinformatics* **2016**, *32* (23), btw514.

(360) Vangone, A.; Bonvin, A. PRODIGY: A Contact-Based Predictor of Binding Affinity in Protein-Protein Complexes. *BIO-PROTOCOL* **2017**, *7* (3), e2124.

(361) Kastritis, P. L. et al. A Structure-Based Benchmark for Protein-Protein Binding Affinity. *Protein Sci.* **2011**, *20* (3), 482–491.

(362) R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria 2016.

(363) Lemkul, J. A.; Allen, W. J.; Bevan, D. R. Practical Considerations for Building GROMOS-Compatible Small-Molecule Topologies. *J. Chem. Inf. Model.* **2010**, *50* (12), 2221–2235.

(364) Sousa da Silva, A. W.; Vranken, W. F. ACPYPE - AnteChamber PYthon Parser InterfacE. *BMC Res. Notes* **2012**, *5* (1), 367.

(365) Wilkinson, R. D. A. et al. Cathepsin S: Therapeutic, Diagnostic, and Prognostic Potential. *Biol. Chem.* **2015**, *396* (8), 867–882.

(366) Ye, L. et al. Cathepsin S in the Spinal Microglia Contributes to Remifentanil-Induced Hyperalgesia in Rats. *Neuroscience* **2017**, *344*, 265–275.

(367) Sena, B. F.; Figueiredo, J. L.; Aikawa, E. Cathepsin S As an Inhibitor of Cardiovascular Inflammation and Calcification in Chronic Kidney Disease. *Front. Cardiovasc. Med.* **2018**, *4*, 88.

(368) J. M. Wiener, J.; Sun, S.; L. Thurmond, R. Recent Advances in the Design of Cathepsin S Inhibitors. *Curr. Top. Med. Chem.* **2010**, *10* (7), 717–732.

(369) Arkhipov, A. et al. Architecture and Membrane Interactions of the EGF Receptor. *Cell* **2013**, *152* (3), 557–569.

(370) Roskoski, R. The ErbB/HER Family of Protein-Tyrosine Kinases and Cancer. *Pharmacol. Res.* **2014**, *79*, 34–74.

(371) Kurkcuoglu, Z. et al. Performance of HADDOCK and a Simple Contact-Based Protein–Ligand Binding Affinity Predictor in the D3R Grand Challenge 2. *J. Comput. Aided. Mol. Des.* **2018**, *32* (1), 175–185.

(372) Hawkins, P. C. D.; Skillman, A. G.; Nicholls, A. Comparison of Shape-Matching and Docking as Virtual Screening Tools. *J. Med. Chem.* **2007**, *50* (1), 74–82.

(373) Ameriks, M. K. et al. Diazinones as P2 Replacements for Pyrazole-Based Cathepsin S Inhibitors. *Bioorg. Med. Chem. Lett.* **2010**, *20* (14), 4060–4064.
(374) Ameriks, M. K. et al. Pyrazole-Based Cathepsin S Inhibitors with Arylalkynes as P1 Binding Elements. *Bioorg. Med. Chem. Lett.* **2009**, *19* (21), 6131–6134.
(375) Wiener, D. K. et al. Thioether Acetamides as P3 Binding Elements for Tetrahydropyrido-Pyrazole Cathepsin S Inhibitors. *Bioorg. Med. Chem. Lett.* **2010**, *20* (7), 2379–2382.
(376) Schneider, G.; Fechner, U. Computer-Based de Novo Design of Drug-like Molecules. *Nat. Rev. Drug Discov.* **2005**, *4* (8), 649–663.
(377) Mandal, S.; Moudgil, M.; Mandal, S. K. Rational Drug Design. *Eur. J. Pharmacol.* **2009**, *625* (1–3), 90–100.
(378) Schneider, G. Automating Drug Discovery. *Nat. Rev. Drug Discov.* **2018**, *17* (2), 97–113.
(379) Vamathevan, J. et al. Applications of Machine Learning in Drug Discovery and Development. *Nat. Rev. Drug Discov.* **2019**, *18* (6), 463–477.
(380) Hawkins, P. C. D. et al. Conformer Generation with OMEGA: Algorithm and Validation Using High Quality Structures from the Protein Databank and Cambridge Structural Database. *J. Chem. Inf. Model.* **2010**, *50* (4), 572–584.
(381) Kuntz, I. D. et al. A Geometric Approach to Macromolecule-Ligand Interactions. *J. Mol. Biol.* **1982**, *161* (2), 269–288.
(382) Sali, A. et al. Outcome of the First WwPDB Hybrid/Integrative Methods Task Force Workshop. *Structure* **2015**, *23* (7), 1156–1167.

Summary

All cells, whether prokaryotic or eukaryotic, are finely tuned biochemical machines. In broad terms, genetic information is encoded in the nucleic acid sequence and is translated in functionally active biomolecules (proteins or other nucleic acids). These biomolecules then perform the multitude of functions the cell needs in order to maintain its homeostatic status. Biomolecules do not exist or perform their functions in isolation: They always act on – or together with – other molecules whether that is an enzyme catalysing a reaction involving a substrate, an activator protein acting on its target or a large collection of biomolecules coming together to create a large macromolecular machine such as the ribosome. Understanding cellular mechanisms in depth, therefore, requires understanding the makeup and function of these biomolecular complexes. For most types of complexes, truly understanding their function relies upon being able to obtain high-quality structures or models of the complex.

Traditional structure determination techniques such as X-ray crystallography, Nuclear Magnetic Resonance (NMR) spectroscopy and cryo-Electron Microscopy (cryo-EM) have been used to determine the structure of thousands of biomolecules and biomolecular complexes. As of October 2019, the Protein Data Bank (PDB), the public repository of solved structures, counts more than 156000 entries. However, if one were to determine the unique entries in the database and then further focus on protein-protein complexes rather than free structures, the resulting number would only be a fraction of that (around 6000-7000 non-redundant biologically relevant complexes). With the protein-protein interactions in the cell estimated to be in the hundreds of thousands it quickly becomes clear that there is a significant gap between the number of biomolecular complexes with solved structures and the total number of complexes identified from interactome high-throughput studies. Next to the experimental methods mentioned above, another way of obtaining structural models for these complexes, the necessary step to understand the molecular mechanisms at play, is computational modelling.

The field of computational modelling that deals with biomolecular complexes, which is the subject of this thesis, is integrative modelling, and in particular, biomolecular docking. These, like all subfields of computational simulation, share some of the same challenges, specifically sampling – or how to generate poses which resemble those of native complexes – and scoring – or how to identify good (or near-native) from wrong models in a large pool of models. Another challenge is about the way in which data are integrated into the simulations, or the need to weight those data in a way that allows for multiple data sources to be efficiently used in the same simulation while also reflecting the experimental uncertainties. This thesis focuses on two additional areas of interest: Docking of transmembrane TM protein complexes and protein-small molecule docking. Both areas are of great interest, both for academic and pharmaceutical research, as TM receptors constitute most drug targets and the majority of drugs on the market today are small compounds. The first half of this thesis pertains to membrane protein modelling, whereas the second half focuses on the modelling of protein-small molecule complexes. In the introductory Chapter, I provide an overview of the state of the integrative modelling field, of the ways in which data from diverse experimental sources can be used by modelling frameworks, and of recent advances in specific areas of interest.

The thesis begins with the **General Introduction** which gently introduces some of the core concepts that are later expanded upon in the following chapters. It also briefly introduces the subject of each chapter. The main part of the thesis consists of **Chapters 1 through 6**.

In **Chapter 1**, I provide an overview of the state of the field of integrative modelling with a particular emphasis on the types of data that can be used by integrative modelling frameworks such as HADDOCK (**H**igh **A**mbiguity **D**riven **DOCK**ing), ROSETTA or IMP (**I**ntegrative **M**odelling **P**latform). The experimental methods that are discussed fall into one of three broad categories depending on the type of data that can be obtained from them: Interface-mapping techniques, techniques providing some kind of distance information between residues and shape-based techniques. Mutagenesis, HDX and NMR (when deriving chemical shift perturbations from titration experiments) are the interface-mapping techniques that are discussed, crosslinking, FRET and DEER the distance-based ones and cryo-EM and SAXS the shape-based techniques. For all these, I evaluate their relevance for the field of integrative modelling and provide examples of their application in modelling interesting and challenging targets. An additional focal point of this chapter is the evaluation of some computational methods in which significant progress has been made recently, namely the use of evolutionary information in the form of coevolution data in docking, advances related to the modelling of membrane proteins and applications of coarse-grained forcefields.

In **Chapter 2**, I describe a recently published benchmark of membrane protein complexes. It is the first, and to the best of my knowledge, the only one of its kind, thus addressing a key missing element for further development of membrane protein docking algorithms. This non-redundant dataset consists entirely of transmembrane α-helical and β-barrel complexes, covering varying difficulty ranges from bound complexes (cases in which both bound components were extracted from the reference complex) to difficult, unbound cases with significant conformational rearrangements at the interface. Using this dataset, we define the baseline performance of HADDOCK for this type of complexes. In addition to the dataset itself we also make available a decoy set consisting of HADDOCK models produced during the benchmarking process.

Chapter 3 is the last chapter which focuses on membrane protein modelling. In that Chapter I describe a protocol for HADDOCK, still under development, that implicitly represents the membrane bilayer by a shape consisting of layers of beads. Restraints are defined between these beads and Cα carbons of the subunits of the complex to drive them to the "membrane". I compare the performance of these shape-restrained runs with one where a single centre-of-mass restraint is defined between the transmembrane segments of the two subunits. The performance of the shape runs is lower than expected compared to the simple transmembrane centre-of-mass restraint ones. Further work will be required to optimise this new approach.

Chapters 4 through 6 constitute the second half of the thesis and revolve around protein-small molecule docking. In **Chapter 4**, I discuss our participation in a blind docking experiment – the 2016 iteration of the Grand Challenge experiment organised by the D3R consortium in which we had to model 36 protein-ligand complexes for the pharmaceutically relevant Farnesoid X receptor. Our small molecule docking protocol consists of the following steps: *i)* Identification of relevant protein receptor templates in the PDB and creation of an ensemble after clustering and selecting representative structures, *ii)* creation of a ligand ensemble after conformer generation and clustering of the resulting conformers and *iii)* docking using residues identified from the receptor templates. Despite excellent results for some cases, our overall performance was not so good compared to the other participants. We could identify the main limiting factor affecting the performance – namely the selection of the receptor template related to conformational changes taking place upon binding. Replacing the ensemble part of the protocol with the selection of a single template based on the similarity of its bound ligand with the target compound to dock does indeed leads to better results.

In the following Chapter – **Chapter 5** – I describe our participation in the following years' Grand Challenge. In addition to selecting a single receptor template based on compound similarity we also revisit the conformer selection procedure: Ligand conformations for docking are selected based on their 3D shape similarity with the bound ligand present in the selected template. We also use the shape similarity to superimpose the selected conformers in the binding pocket of the receptor, bypass the initial stage of the docking protocol and directly proceed to refine the models using the water refinement stage of HADDOCK. The incorporation of shape information in our protocol has a significant impact on our success rate: Our submission (24 protein-ligand complexes predicted) was evaluated as one of the best within standard deviation of the top performing participant.

Chapter 6 represents the logical conclusion of the small molecule docking part of the thesis. In that Chapter I describe a new HADDOCK protocol that makes use of the shape information which was identified as highly relevant in the preceding chapter. The innovative aspect of this protocol is the way the ligand shape is represented with the heavy atoms of the template compound being transformed into shape beads as used for the representation of the membrane described in **Chapter 3**. Similar to that protocol, restraints are then defined between the shape beads and the atoms of the generated ligand conformers. With the use of those shape restraints we don't need to pre-select conformations, but instead use all 500 generated ones, increase the sampling and let HADDOCK select the near-native ones. This protocol outperforms the one described in the preceding chapter. The shape restraints allow to induce rather large conformational changes in the ligand toward its bound form. In addition, since we are now performing a full docking run, we can integrate any additional information in the simulation – something which was not possible in the previous protocol as it was a simple refinement.

In the last Chapter – **Chapter 7** – I summarise the main findings of this thesis and offer critical perspectives for the challenges the field is facing as well as some potential avenues worth exploring in the future.

Samenvatting

Alle cellen, zowel prokaryoot als eukaryoot, zijn voorzichtig afgestelde biochemische machines. Kort samengevat ligt genetische informatie vast in DNA-sequenties die vertaald worden naar functionele biomoleculen (proteïnen of andere nucleotiden). Deze biomoleculen vervullen een groot aantal verschillende functies die nodig zijn om de homeostase van de cel te handhaven. Biomoleculen bestaan, noch vervullen ze hun taken, in isolatie: ze werken altijd aan – of samen met – andere moleculen, of het nou een enzym is dat een reactie met een substraat katalyseert, een activator die invloed uitoefent op zijn *target*, of een grote groep biomoleculen die samen een grote macromoleculaire machine zoals een ribosoom vormen. Diepgaand inzicht in cellulaire mechanismen vergt daarom kennis van de samenstelling en functie van deze biomoleculaire complexen. Om de functie van de meeste complexen volledig te kunnen begrijpen is het nodig structuren of modellen van hoge kwaliteit van het complex te construeren.

Traditionele technieken voor het bepalen van structuren, zoals röntgenkristallografie, kernspinresonantie (NMR-spectroscopie), en cryo-elektronenmicroscopie (cryo-EM), zijn gebruikt om de structuur van duizenden biomoleculen en biomoleculaire complexen te bepalen. In oktober 2019 telde de Protein Data Bank (PDB), de openbare database van opgeloste structuren, meer dan 156.000 structuren. Als men echter kijkt welke vermeldingen uniek zijn en zich verder concentreert op eiwit-eiwit complexen in plaats van losse structuren, wordt het aantal vele malen kleiner (ongeveer 6.000-7.000 unieke, biologisch-relevante complexen). Aangezien het geschatte aantal eiwit-eiwit interacties in de cel in de honderdduizenden ligt, is het duidelijk dat er een grote kloof zit tussen het aantal biomoleculaire complexen waarvoor een structuur beschikbaar is en het totale aantal complexen dat is geïdentificeerd door middel van high-throughput studies. Naast de bovengenoemde experimentele methodes om structuurmodellen van deze complexen te creëren is een andere benadering nodig om de relevante moleculaire mechanismen te begrijpen, door middel van computationeel modelleren. Het vakgebied binnen computationeel modelleren dat zich bezighoudt met biomoleculaire complexen, het onderwerp van dit proefschrift, is integratief modelleren en, specifieker, biomoleculair docking. Deze gebieden, zoals alle vakgebieden die zich bezighouden met computersimulaties, hebben bepaalde uitdagingen gemeen: *sampling* – of hoe poses te genereren die lijken op die van de natieve staat (*native state*) – en *scoring* – of hoe de goede (*near-native*) van de foute modellen te onderscheiden in een grote verzameling modellen. Een andere uitdaging ligt in de integratie van data in de simulaties, of hoe deze data zo mee te wegen dat data uit meerdere bronnen efficiënt gebruikt kunnen worden in dezelfde simulatie en tegelijkertijd rekening te houden met experimentele onzekerheden. Dit proefschrift concentreert zich op twee verdere onderwerpen: het docken van transmembraan (TM) proteïnecomplexen en eiwit-kleine molecuul docking. Beide onderwerpen zijn van groot belang, zowel voor academisch- als voor farmaceutisch onderzoek, omdat TM-receptoren het merendeel van de *targets* van medicijnen uitmaken en de meerderheid van de medicijnen die momenteel op de markt zijn kleine moleculen zijn. De eerste helft van het proefschrift heeft betrekking op het modelleren van membraanproteïnen, terwijl de tweede helft focust op het modelleren van eiwit-kleine molecuul complexen. In het introductiehoofdstuk geef ik een overzicht van de huidige staat van het gebied van modelleren, van de manieren waarop data uit diverse experimentele bronnen kunnen worden gebruikt door modellering-*frameworks*, en van recente ontwikkelingen op specifieke vlakken.

Het proefschrift begint met een **Algemene Introductie** (General Introduction) die een aantal centrale concepten introduceert die in latere hoofdstukken verder worden toegelicht. Ook wordt kort het onderwerp van elk hoofdstuk beschreven. Het centrale gedeelte van het proefschrift bestaat uit hoofdstukken 1 t/m 6.

In **Hoofdstuk 1** geef ik een overzicht van de huidige staat van het integratief modelleren, met speciale aandacht voor de soorten data die gebruikt kunnen worden door integratieve modellering-*frameworks* zoals HADDOCK (**H**igh **A**mbiguity **D**riven **DOCK**ing), ROSETTA, of IMP (**I**ntegrative **M**odeling **P**latform). De experimentele methodes die hier besproken worden kunnen ingedeeld worden in drie categorieën op basis van het type data die ze verschaffen: *interface-mapping* technieken, technieken die informatie geven over de afstand tussen bepaalde residuen, en vorm-gerelateerde technieken. Mutagenese, HDX en NMR (in het geval van verandering van chemische verschuiving in titratie-experimenten) zijn de *interface-mapping* technieken die hier worden besproken; crosslinking, FRET en DEER geven informatie over afstanden; cryo-EM en SAXS over vorm. Van al deze technieken beschrijf ik het belang voor integratief modelleren en geef voorbeelden van hun toepassing in het modelleren van interessante en gecompliceerde *targets*. Een verdere focus in dit hoofdstuk is de evaluatie van een aantal computationele methodes waarin recent vooruitgang is geboekt, namelijk het gebruik van evolutionaire informatie in de vorm van co-evolutie data in docking, ontwikkelingen op het gebied van het modelleren van membraanproteïnen, en toepassingen van *coarse-grained* representaties.

In **Hoofdstuk 2** beschrijf ik een recent gepubliceerde benchmark van membraanproteïnecomplexen. Het is de eerste en, naar mijn weten, de enige in zijn soort en dus het antwoord op een tot dusver ontbrekend element voor het verder ontwikkelen van algoritmes voor het docken van membraanproteïnes. Deze niet-redundante dataset bestaat geheel uit α-helix en β-barrel transmembraancomplexen van verschillende moeilijkheidsgraden, variërend van gebonden complexen (gevallen waarin beide componenten uit de gebonden vorm van het complex komen) tot moeilijke, ongebonden gevallen met significante conformatieveranderingen in het interactieoppervlak. Met gebruik van deze dataset leggen we vast hoe HADDOCK presteert met dit type complexen. Naast de dataset zelf is ook een set *decoys* beschikbaar, bestaande uit HADDOCK-modellen die voortkomen uit het benchmarkingproces.

Hoofdstuk 3 is het laatste hoofdstuk met betrekking tot het modelleren van membraanproteïnecomplexen. In dat hoofdstuk beschrijf ik een protocol voor HADDOCK, nog in ontwikkeling, waarin de lipide dubbellaag wordt vertegenwoordigd door lagen van bollen (*dummy atoms*). Om de subeenheden van het complex naar het “membraan” te brengen worden maximumafstanden (*restraints*) gebruikt tussen deze bollen en Cα atomen van de eiwitsubeenheden van het complex. Ik vergelijk de resultaten van deze runs, met vorm-*restraints*, met runs waarbij een enkel massamiddelpunt-*restraint* tussen het transmembraangedeelte van de twee subeenheden is gebruikt. Verder werk zal nodig zijn om deze nieuwe aanpak te optimaliseren.

Hoofdstukken 4 t/m 6 vormen de tweede helft van het proefschrift en gaan over eiwit-kleine molecuul docking. In **Hoofdstuk 4** bespreek ik onze deelname aan een blind docking experiment – de 2016 editie van de Grand Challenge georganiseerd door het D3R consortium – waarvoor we 36 eiwit-ligand complexen moesten modelleren van de farmaceutisch-relevante Farnesoid X receptor. Ons kleine-molecuul dockingprotocol bestaat uit de volgende stappen: *i)* Identificatie van relevante eiwitreceptor templates in de PDB en het creëren van een ensemble van representatieve structuren na het clusteren van de templates, *ii)* creatie van een ligand ensemble na verschillende conformeren te hebben gegenereerd en geclusterd en *iii)* docking met gebruik van residuen geïdentificeerd in de receptor templates. Ondanks zeer goede resultaten in sommige gevallen was onze prestatie in het algemeen niet zo goed als die van andere deelnemers. We waren in staat de belangrijkste beperkende factor te identificeren, de selectie van receptor templates had namelijk te maken met conformatieveranderingen na het binden van de ligand. Door het ensemble-gedeelte van het protocol te vervangen met de selectie van een enkel template gebaseerd op de overeenkomst van de gebonden ligand met de te docken *target*-verbinding verbeterden de resultaten inderdaad.

In het volgende hoofdstuk, **Hoofdstuk 5**, beschrijf ik onze deelname aan de Grand Challenge van het daaropvolgende jaar. Behalve dit keer één enkel receptor template te selecteren gebaseerd op ligand-overeenkomst herzien we ook de procedure voor selectie van de ligand conformeer: conformaties worden geselecteerd op basis van de overeenkomst van hun 3D-vorm met die van de gebonden ligand in het template. Ook gebruiken we de vormovereenkomsten om de geselecteerde conformeer in de *binding pocket* van de receptor te superimposeren, slaan de eerste fase van het dockingprotocol over en gaan direct over tot het verfijnen van de modellen door middel van het waterverfijning gedeelte van HADDOCK. Het gebruik van vorminformatie in ons protocol heeft een significante impact op het succespercentage: onze inzending (24 voorspelde eiwit-ligand complexen) kwam als een van de besten uit de evaluatie, binnen standaarddeviatie van de best-presterende deelnemer.

Hoofdstuk 6 is de logische conclusie van het kleine-molecuul docking gedeelte van dit proefschrift. In dat hoofdstuk beschrijf ik een nieuw HADDOCK-protocol dat de vorminformatie gebruikt die in het voorgaande hoofdstuk als zeer relevant werd geïdentificeerd. Het innovatieve aspect van dit protocol is de manier waarop de vorm van de ligand wordt gerepresenteerd: door de zware atomen van het template te veranderen in bollen, zoals eerder beschreven voor de vertegenwoordiging van het membraan in **Hoofdstuk 3**. Net zoals in dat protocol worden hier *restraints* gedefinieerd tussen de bollen en de atomen van de ligand conformeren. Door dit soort vorm-*restraints* te gebruiken hoeven we niet van tevoren conformaties te selecteren, maar kunnen we alle 500 gebruiken en met verhoogde *sampling* HADDOCK de *near-native* conformaties laten selecteren. Dit protocol presteert beter dan dat beschreven in het voorgaande hoofdstuk. De vorm-*restraints* maken relatief grote conformatieveranderingen in de ligand mogelijk, om dichter bij de gebonden vorm te kunnen komen. Daarnaast, omdat we nu een volledige docking run uitvoeren, kunnen we aanvullende informatie in de simulatie integreren – iets dat niet mogelijk was met het vorige protocol dat slechts uit een verfijning bestond.

In het laatste hoofdstuk, **Hoofdstuk 7**, vat ik de belangrijkste bevindingen van het proefschrift samen en geef ik kritisch commentaar op de uitdagingen waar het vakgebied voor staat, alsmede enkele mogelijke richtingen die de moeite waard zijn in de toekomst te exploreren.

Περίληψη

Όλα τα κύτταρα, είτε είναι προκαρυωτικά είτε ευκαρυωτικά, αποτελούν καλά συντονισμένες βιοχημικές μηχανές. Με ευρείς όρους, η γενετική πληροφορία είναι κωδικοποιημένη στην νουκλεϊκή αλληλουχία και μεταφράζεται σε λειτουργικά ενεργά βιομόρια (πρωτεΐνες η άλλα νουκλεϊκά οξέα). Αυτά τα βιομόρια με την σειρά τους πραγματοποιούν την πληθώρα των λειτουργειών το κύτταρο χρειάζεται προκειμένου να διατηρήσει την ομοιοστατική του ισορροπία. Τα βιομόρια δεν υπάρχουν ούτε πραγματοποιούν τις λειτουργίες τους σε απομόνωση: Πάντα επιδρούν σε – η μαζί – με άλλα μόρια ανεξάρτητα από το αν είναι ένζυμα που καταλύουν κάποια χημική αντίδραση η οποία περιλαμβάνει το υπόστρωμα τους, κάποια πρωτεΐνη η οποία ενεργοποιεί άλλες πρωτεΐνες η μία μεγάλη ομάδα βιομορίων τα οποία συσχετίζονται προκειμένου να δημιουργήσουν μια μεγάλη μακρομοριακή μηχανή όπως το ριβόσωμα. Η κατανόηση των μοριακών μηχανισμών απαιτεί κατανόηση της σύστασης και λειτουργίας αυτών των βιομοριακών συμπλόκων. Για τους περισσότερους τύπους συμπλόκων, η πλήρης κατανόηση της λειτουργίας τους απαιτεί υψηλής ποιότητας δομές η μοντέλα του συμπλόκου.

Παραδοσιακές τεχνικές δομικής βιολογίας όπως η κρυσταλλογραφία ακτίνων Χ (X-ray crystallography), ο πυρηνικός μαγνητικός συντονισμός (NMR) και η κρυο-ηλεκτρονική μικροσκοπία (cryo-EM) έχουν χρησιμοποιηθεί για τον προσδιορισμό των δομών χιλιάδων βιομορίων και βιομοριακών συμπλόκων. Μέχρι τον Οκτώβριο του 2019, η πρωτεϊνική βάση δεδομένων (PDB – Protein Data Bank), η δημόσια βάση δεδομένων μοριακών δομών, αριθμούσε περισσότερες από 156.000 καταχωρήσεις. Ωστόσο, ο αριθμός των μοναδικών καταχωρήσεων και ο αριθμός των δομών που αντιπροσωπεύουν μοριακά σύμπλοκα αποτελούν ένα κλάσμα του συνολικού αριθμού (περίπου 6.000-7.000 μοναδικά, βιολογικά ενεργά σύμπλοκα). Με τις εκτιμήσεις για τον αριθμό των αλληλεπιδράσεων ανάμεσα σε πρωτεΐνες στο κυτταρικό περιβάλλον να αγγίζουν τις εκατοντάδες χιλιάδες, εύκολα προκύπτει η μεγάλη ανακολουθία ανάμεσα στον αριθμό των βιομοριακών συμπλόκων με διαθέσιμες μοριακές δομές και τον συνολικό αριθμό συμπλόκων που έχουν διαπιστωθεί σε υψηλής διακίνησης (high-throughput) μελέτες του «αλληλεπιδρώματος» (interactome). Πέρα από τις πειραματικές τεχνικές που αναφέρθηκαν παραπάνω, ένας εναλλακτικός τρόπος προσδιορισμού μοριακών δομών, βήμα απαραίτητο για την κατανόηση των μοριακών μηχανισμών, είναι ο υπολογιστικές προσεγγίσεις (computational modelling).

Το πεδίο των υπολογιστικών προσεγγίσεων το οποίο ασχολείται με βιομοριακά συμπλέγματα, το οποίο είναι το αντικείμενο αυτής της διατριβής, είναι η ολοκληρωτική μοντελοποίηση (integrative modelling) και συγκεκριμένα η μοριακή αγκυροβόληση (biomolecular docking). Όπως και όλα τα πεδία της υπολογιστικής προσομοίωσης, μοιράζονται μερικές από τις ίδιες δυσκολίες, και συγκεκριμένα την δημιουργία μοριακών δομών οι οποίες να προσεγγίζουν τις φυσικές (sampling) καθώς και την αναγνώριση καλών (ή δομές οι οποίες να προσεγγίζουν τις φυσικές) και κακών μοντέλων (scoring). Μία άλλη δυσκολία είναι ο τρόπος με τον οποίο δεδομένα ενσωματώνονται στις προσομοιώσεις, ή η ανάγκη εξισορρόπησης αυτών των δεδομένων με τέτοιο τρόπο ώστε πολλαπλές πηγές δεδομένων να μπορούν να χρησιμοποιηθούν αποτελεσματικά στην ίδια προσομοίωση ενώ οι αρχικές πειραματικές αβεβαιότητες τηρούνται. Η παρούσα διατριβή εστιάζει σε δύο επιπλέον περιοχές ενδιαφέροντος: Την αγκυροβόληση διαμεμβρανικών πρωτεϊνικών συμπλεγμάτων και αγκυροβόληση πρωτεϊνών και μικρών μορίων. Και οι δύο παρουσιάζουν μεγάλο ενδιαφέρον, τόσο για ακαδημαϊκή όσο και φαρμακευτική έρευνα, καθώς οι διαμεμβρανικοί υποδοχείς αποτελούν τους περισσότερους φαρμακευτικούς στόχους και η πλειοψηφία των φαρμάκων τα οποία είναι σήμερα διαθέσιμα στην αγορά είναι μικρά μόρια. Στο πρώτο μισό της διατριβής η έμφαση είναι στην μοντελοποίηση διαμεμβρανικών πρωτεϊνών, ενώ στο δεύτερο στην μοντελοποίηση συμπλεγμάτων πρωτεϊνών και μικρών μορίων. Στο εισαγωγικό κεφάλαιο, παρέχω μία επισκόπηση της τρέχουσας κατάστασης του πεδίου της ολοκληρωτικής μοντελοποίησης, των τρόπων με τους οποίους δεδομένα από ποικίλες πειραματικές πηγές μπορούν να ενσωματωθούν σε υπολογιστικές μελέτες και πρόσφατων εξελίξεων σε συγκεκριμένες περιοχές ενδιαφέροντος.

Η διατριβή ξεκινά με την **Γενική Εισαγωγή** η οποία εισαγάγει με ήπιο τρόπο μερικές από τις θεμελιώδεις ιδέες οι οποίες εξερευνώνται περαιτέρω στα υπόλοιπα κεφάλαια. Επίσης, εισαγάγει το θεματικό αντικείμενο του κάθε κεφαλαίου. Τα **Κεφάλαια 1 έως και 6** απαρτίζουν το κύριο κομμάτι της διατριβής.

Στο **Κεφάλαιο 1**, παρέχω μια επισκόπηση της τρέχουσας κατάστασης του πεδίου της ολοκληρωτικής μοντελοποίησης με ιδιαίτερη έμφαση στους τύπους δεδομένων που μπορούν να χρησιμοποιηθούν από υπολογιστικές μεθόδους όπως τα HADDOCK (**H**igh **A**mbiguity **D**riven **DOCK**ing), ROSETTA ή IMP (**I**ntegrative **M**odelling **P**latform). Οι πειραματικές μέθοδοι οι οποίες αναλύονται ανήκουν σε μία από τρεις ευρείες κατηγορίες: Τεχνικές οι οποίες μας επιτρέπουν να προσδιορίσουμε την επιφάνεια αλληλεπίδρασης ανάμεσα σε βιομόρια, τεχνικές οι οποίες παρέχουν κάποιου είδους πληροφορία σχετικά με την απόσταση συγκεκριμένων αμινοξικών καταλοίπων και τεχνικές οι οποίες προσδιορίζουν το σχήμα των βιομορίων. Η μεταλλαξιγένεση (mutagenesis), η ανταλλαγή υδρογόνου-δευτέριου (Hydrogen-deuterium exchange – HDX) και ο πυρηνικός μαγνητικός συντονισμός (όταν υπολογίζονται χημικές μετατοπίσεις [chemical shift perturbations] μετά από ογκομετρική ανάλυση [titration]) είναι οι τεχνικές οι οποίες μας επιτρέπουν να εντοπίσουμε την αλληλεπιδρούσα επιφάνεια βιομορίων, η χημική διασύνδεση (chemical crosslinking), η μεταβίβαση ενεργειακού συντονισμού Förster (FRET) και ο διπλός συντονισμός ηλεκτρονίων-ηλεκτρονίων (DEER) μας επιτρέπουν να υπολογίσουμε αποστάσεις ανάμεσα σε αμινοξικά κατάλοιπα και τέλος η κρυο-ηλεκτρονική μικροσκοπία και η σκέδαση ακτίνων Χ μικρών γωνιών (SAXS) είναι οι τεχνικές μέσω των οποίων μπορούμε να αποκτήσουμε πληροφορίες για το σχήμα βιομορίων. Αναλύω την σχετικότητα όλων των προαναφερθέντων τεχνικών για το πεδίο της ολοκληρωτικής μοντελοποίησης και παραθέτω παραδείγματα εφαρμογών τους σε περιπτώσεις μοντελοποίησης ενδιαφερόντων ή προκλητικών συστημάτων. Ένα επιπλέον εστιακό σημείο αυτού του κεφαλαίου είναι η επισκόπηση κάποιων υπολογιστικών μεθόδων οι οποίες έχουν σημειώσει σημαντική πρόοδο πρόσφατα, και συγκεκριμένα γύρω από την χρήση εξελικτικών δεδομένων με τη μορφή πληροφοριών συνεξέλιξης (coevolution), εξελίξεις οι οποίες σχετίζονται με την μοντελοποίηση μεμβρανικών πρωτεϊνών και εφαρμογές αδρών (coarse-grained) πεδίων ισχύος (force fields).

Στο **Κεφάλαιο 2**, περιγράφω ένα πρόσφατα δημοσιευμένο σετ δεδομένων το οποίο απαρτίζεται εξ ολοκλήρου από σύμπλοκα μεμβρανικών πρωτεϊνών. Είναι το πρώτο, και από όσο γνωρίζω, μοναδικό του είδους του και με αυτόν τον τρόπο παρέχει ένα στοιχείο κομβικής σημασίας για περαιτέρω εξέλιξη αλγορίθμων μοριακής αγκυροβόλησης μεμβρανικών πρωτεϊνών. Αυτό το σετ το οποίο απαρτίζεται από μοναδικά (μη επαναλαμβανόμενα) σύμπλοκα αποτελείται εξ ολοκλήρου από διαμεμβρανικά σύμπλοκα α-ελίκων και β-βαρελιών, και καλύπτει μία ευρεία γκάμα δυσκολίας, από σύμπλοκα στην προσδεδεμένη τους κατάσταση (περιπτώσεις στις οποίες και τα δύο αλληλεπιδρώντα μόρια έχουν απομονωθεί από το σύμπλοκο αναφοράς) μέχρι σύμπλοκα υψηλής δυσκολίας, μη προσδεδεμένες περιπτώσεις με σημαντικές αλλαγές στην αλληλεπιδρώσα επιφάνεια. Με βάση αυτό το σετ, προσδιορίσαμε την απόδοση του HADDOCK για σύμπλοκα τέτοιου τύπου. Πέρα από το σετ, προσφέρουμε επίσης όλα τα μοντέλα τα οποία δημιουργήθηκαν με το HADDOCK κατά την διάρκεια της διαδικασίας αυτής.

Το **Κεφάλαιο 3** είναι το τελευταίο το οποίο αφορά μοντελοποίηση μεμβρανικών πρωτεϊνών. Σε αυτό το κεφάλαιο περιγράφω ένα πρωτόκολλο για το HADDOCK, το οποίο είναι ακόμα υπό βελτίωση, το οποίο αναπαριστά την κυτταρική μεμβράνη με ένα σχήμα το οποίο αποτελείται από στρώματα από σφαίρες. Ορίζουμε περιορισμούς ανάμεσα στις σφαίρες αυτές και τους Cα άνθρακες των υπομονάδων του συμπλόκου ώστε να τις οδηγήσουμε στην «μεμβράνη». Συγκρίνω την απόδοση αυτού του πρωτοκόλλου με ένα στο οποίο χρησιμοποιούμε μόνο έναν περιορισμό ανάμεσα στο κέντρο βάρους των διαμεμβρανικών τμημάτων των δύο υπομονάδων. Η απόδοση του πρωτοκόλλου είναι χαμηλότερη των προσδοκιών συγκριτικά με την απλή μέθοδο η οποία χρησιμοποιεί μόνο τον ένα απλό περιορισμό ανάμεσα στις δύο υπομονάδες και θα χρειαστούν επιπλέον δοκιμές προκειμένου να οριστικοποιηθεί.

Τα **Κεφάλαια 4 ως και 6** αποτελούν το δεύτερο μισό της διατριβής και περιστρέφονται γύρω από την μοριακή αγκυροβόληση πρωτεϊνών και μικρών μορίων. Στο **Κεφάλαιο 4**, συζητώ την συμμετοχή μας σε ένα τυφλό πείραμα – το Grand Challenge που οργανώθηκε το 2016 από την ομάδα D3R (**D**rug **D**esign **D**ata **R**esource) και στο οποίο ο στόχος ήταν η πρόβλεψη της δομής 36 συμπλόκων πρωτεϊνών-μικρών μορίων για τον φαρμακευτικού ενδιαφέροντος υποδοχέα Farnesoid X. Το πρωτόκολλο μας απαρτιζόταν από τα ακόλουθα στάδια: *i)* Εντοπισμός σχετικών πρωτεϊνικών δομών στην PDB και δημιουργία ενός σετ αντιπροσωπευτικών δομών μετά από ανάλυση, *ii)* δημιουργία ενός σετ για τα μικρά μόρια ακολουθώντας παρόμοια ανάλυση και *iii)* μοριακή αγκυροβόληση χρησιμοποιώντας κατάλοιπα τα οποία αναγνωρίστηκαν στο στάδιο *(i)*. Παρά τα εξαιρετικά αποτελέσματα για μερικές περιπτώσεις, η συνολική μας απόδοση δεν ήταν ιδιαίτερα καλή σε σχέση με τους υπόλοιπους συμμετέχοντες. Ο κύριος παράγοντας ο οποίος επηρέασε αρνητικά την απόδοση μας ήταν το πρωτόκολλο επιλογής πρωτεϊνικών υποδοχέων πριν την μοριακή αγκυροβόληση. Αντικαθιστώντας αυτό το κομμάτι του πρωτοκόλλου με την επιλογή ενός υποδοχέα με βάση την ομοιότητα ανάμεσα στο προσδεδεμένο του μικρό μόριο και τα μικρά μόρια ενδιαφέροντος η διαδικασία της μοριακής αγκυροβόλησης οδήγησε σε καλύτερα αποτελέσματα.

Στο ακόλουθο κεφάλαιο – το **Κεφάλαιο 5** – περιγράφω την συμμετοχή μας στο τυφλό πείραμα το οποίο διοργανώθηκε για την χρονιά 2017. Πέρα από την επιλογή ενός μόνο υποδοχέα με βάση την ομοιότητα των μικρών μορίων, βελτιώσαμε την διαδικασία επιλογής δομών μικρών μορίων πριν την μοριακή αγκυροβόληση: Οι δομές μικρών μορίων επιλέχθηκαν με βάση την ομοιότητα του σχήματος του με το σχήμα του προσδεδεμένου μικρού μορίου σε κάθε υποδοχέα. Επίσης χρησιμοποιήσαμε την σχηματική ομοιότητα για να τοποθετήσουμε τις επιλεγμένες δομές στην περιοχή πρόσδεσης του υποδοχέα, υπερπηδώντας το πρώτο στάδιο της μοριακής αγκυροβόλησης και χρησιμοποιώντας το τελευταίο στάδιο του HADDOCK το οποίο απλά βελτιώνει τις δομές. Η ενσωμάτωση των σχηματικών δεδομένων στο πρωτόκολλο μας έχει ξεκάθαρη επιρροή στην απόδοσή του: Η πρόβλεψη μας (24 σύμπλοκα πρωτεϊνών-μικρών μορίων) αξιολογήθηκε ως μία από τις καλύτερες και εντός τυπικής απόκλισης από την κορυφαία.

Το **Κεφάλαιο 6** αντιπροσωπεύει την λογική κατάληξη του κομματιού της διατριβής το οποίο ασχολείται με την μοριακή αγκυροβόληση μικρών μορίων. Σε αυτό το κεφάλαιο περιγράφω ένα καινούριο πρωτόκολλο για το HADDOCK το οποίο χρησιμοποιεί σχηματικά δεδομένα, τα οποία αναγνωρίστηκαν ως πολύ σημαντικά στο προηγούμενο κεφάλαιο. Το ριζοσπαστικό στοιχείο αυτού του πρωτοκόλλου είναι ο τρόπος με τον οποίο το σχήμα των μικρών μορίων αναπαρίσταται, με τα βαρέα άτομα (όλα πλην του υδρογόνου) του προσδεδεμένου μικρού μορίου να μεταμορφώνονται σε σφαίρες όπως αυτές που χρησιμοποιούνται για την αναπαράσταση της κυτταρικής μεμβράνης στο **Κεφάλαιο 3**. Όπως και σε εκείνο το πρωτόκολλο, δημιουργούνται περιορισμοί ανάμεσα στα άτομα των μικρών μορίων και τις σφαίρες του σχήματος. Με την χρήση αυτών των περιορισμών δεν χρειάζεται να επιλέξουμε δομές μικρών μορίων πριν την μοριακή αγκυροβόληση, αλλά χρησιμοποιούμε όλες τις δομές που δημιουργήθηκαν, αυξάνουμε τον αριθμό των μοντέλων που το HADDOCK θα δημιουργήσει και το αφήνουμε να επιλέξει αυτές που μοιάζουν περισσότερο με τις φυσικές. Αυτό το πρωτόκολλο λειτουργεί καλύτερα από αυτό το οποίο περιεγράφηκε στο προηγούμενο κεφάλαιο. Οι σχηματικοί περιορισμοί μας επιτρέπουν να επιβάλλουμε σημαντικές αλλαγές στην δομή των μικρών μορίων προς την κατεύθυνση των φυσικών δομών. Καθώς τώρα πραγματοποιούμε ολόκληρη την διαδικασία της μοριακής αγκυροβόλησης μπορούμε να ενσωματώσουμε επιπλέον πληροφορίες στην προσομοίωση – κάτι το οποίο δεν ήταν δυνατόν με το πρωτόκολλο που περιγράφεται στο προηγούμενο κεφάλαιο καθώς ήταν μία απλή βελτίωση των δομών.

Στο τελευταίο κεφάλαιο – **το Κεφάλαιο 7** – πραγματοποιώ μία ανασκόπηση των κύριων συμπερασμάτων της παρούσας διατριβής και προσφέρω μερικές κριτικές απόψεις για τις δυσκολίες τις οποίες αντιμετωπίζει το πεδίο και προτείνω ερευνητικές προοπτικές οι οποίες θεωρώ ότι χρήζουν προσοχής στο μέλλον.

Acknowledgements

I am not sure where one begins when there are so many people to acknowledge and things to be thankful for, but I might as well start at the beginning. None of this would have been remotely possible without the love and support of my family and in particular my parents. Their unwavering support and belief in my scientific undertakings have been the backbone of all my endeavours.

Science, however, does not happen *in vacuo*, but is a group effort instead, and one couldn't ask for a better group than the one I found myself to be a part of in March 2016, when I joined the Computational Structural Biology group. The group – at the time – was very different to its current line-up as I am the one that, as of December 2019, has been around the longest. I immediately felt very welcome (although most people that were around back then would probably agree it didn't necessarily look that way) and the person that bears most of the responsibility for that is **Anna**. You patiently introduced me to all kinds of modelling I had no experience with before coming to Utrecht and also to the sheer joy that is CAPRI, all done in your unique Neapolitan way. Both you and **Li**, created a very warm environment (not only figuratively) in which I very quickly felt at ease. Some of my happiest work-related memories from that time come from working side with side with **Zeynep** trying (and initially failing) to figure out small-molecule docking. Something unusual about the group at the time was its composition, with seven post-docs and just two PhD candidates – me and **Liang** (aka Dr Geng) who has since become a scientific software developer, a role that suits him well. **Irina** joined the group around the same time I did and has since been in the running for busiest person of the year with running her group in Portugal, visiting our group as a post-doc, applying for grants and tenure-track positions, studying for her second PhD, raising two children, etc… Despite our differences of opinion (which include my future work prospects) I have nothing but respect for you and consider myself lucky I got to know you both professionally and personally. Rounding out the group in terms of gender balance were, of course, the occupants of office 1.16. **Jörg**, you show nothing but professionalism and characteristic German efficiency in everything you do: coding, science, house repairs, beer drinking and, most importantly, tolerating and even thriving in an office dominated by the inimitable French Duo. **Mikael** and **Adrien** also (re)joined the group within a couple of weeks of me and strongly imbued it with French flair and double entendres. **Mikael**, your time in Utrecht (your second time that is) was marked by profound changes in your personal life as you became a father (twice!). Despite all the associated challenges that go hand in hand with such a big change both you and Sophie handle it beautifully and throughout all three years you remained filled with positivity and energy. **Adrien**, if I had to choose one person who wasn't my boss and I've learned from the most then that person would be you. Hands down. Your uncanny ability to balance family life, an impressive array of sports-related activities, teaching and research while at the same time maintaining genuine interest, curiosity and inquisitiveness about the world around you is something I can only hope to one day match.

Of course, over time the group changed, and new people took the place of those who had since moved on to bigger and better things. The Mediterranean element of the group was reinvigorated with the arrival of **Jorge** and **Francesco** who joined the group as PhD candidates in a desperate attempt to contain the French contingent. Perhaps surprisingly, your drinks of choice would feature prominently in various aspects and on multiple occasions over the coming years. In the case of Francesco that was Caffè Borbone; In the case of Jorge that drink was Pint of Science, and that wasn't the end of the social activities you organised in your time in Utrecht. Had it not been for you, the group outing that we coorganised with Miranda would have been a disaster. Building on the Mediterranean connections, **Brian** joined the group as the resident software engineer and metalhead. The number of fires that you are putting out on a daily basis so that our services keep running smoothly while at the same time maintaining and expanding LightDock, is a testament to your skills and tenacity. However, there is no creative process (and regardless of commonly held beliefs software development is such a process) that cannot benefit from a dash of madness and that – in addition to a zen attitude, genuine enthusiasm and skills – is exactly what **Rodrigo** is infusing the next generation of HADDOCK with. Rodrigo, you're a long way from home but you liked us well enough the first time, that you decided to take the plunge and move to cold and rainy Utrecht. It's a bold move and I hope everything goes well here for both you and Gabi (and the cat). In addition to Brian and Rodrigo, **Zuzana** is the third occupant of – present day – office 1.16 and one of the most recent additions to our group. The CSB group made an excellent first impression on her when we first met during BIOMOS just over two years ago when she thought both me and Jorge were depressed and possibly non-verbal. Despite that, she chose to join the group and brings some much-needed expertise in Free Energy simulations and interpretive dance as a means of communication. It's not as interesting when you're not around Zuzana; Never change. Rounding out the recent arrivals, is **Siri** who chose to come back to the Netherlands all the way from sunny California – a choice which I guess you might be questioning during the Dutch winter – and is developing new ways of making use of crosslink data in simulations. You have integrated into the group very quickly and even introduced boardgame nights, which is already something that plenty of us are looking forward to and regularly participating in. Our group has, at times, also hosted plenty of MSc students and I've been lucky enough to supervise a few of them, namely

Zhengyue, **Charlotte** and presently **Sam** and in the immediate future **Etjen**. In addition to the material help on the various projects that you've all worked on, the thing I've grown to appreciate the most about supervising and teaching is that it's a process in which both parties have things to teach each other. **Charlotte**, in particular, I guess you enjoyed your time in our group during your research project so much – despite me never introducing you to the members of the group – that you decided to come back for your PhD which is now almost a year in the making and already showing promising results. Not only that, but we are currently sharing an office – which to the untrained eye might appear dark and silent but is, in reality, energy-efficient and conducive to contemplation – and you were also kind enough to translate my summary in Dutch. Thanks for all the help.

Finally, **Alexandre**. I've been going over the acknowledgements sections of old group members theses' and one thing becomes immediately clear. Your PhD students are all very happy to have had you as their supervisor and I am no exception. As I am writing this, I am flying over the mountains of the Alps that I have so often heard you use as examples when lecturing about energy minimisation and it occurs to me, that although you've never formally taught me, the things I've learned just by being around you over the last four years are profound. You conduct science in a way that is filled with integrity, curiosity and finesse. I can't recall an occasion on which you lost your patience or even your cool including unfair reviews, rejections and even PhD students choosing to ignore the project for which they were hired for two years *and even* after the PhD student who was previously supposed to work on said project chose to also ignore it. By far your most impressive quality – next to your surreal multitasking skills – is your talent or skill (or both?) when it comes to picking people for your group that can stand on their own but also be part of a team rather than individuals. My only regret about joining your group is one I imagine many of your PhDs and former PhDs share, namely that, in terms of future bosses, it's probably downhill from this point on.

Our group, however, is but one of a few occupying the building and groups come with group leaders: **Marc**, **Hugo** and **Markus**. We haven't interacted much over my time in Utrecht but at the very least I am thankful to you for always keeping us in touch with the experimental side of Structural Biology but more importantly for your people. **Gert**, you keep everyone in line with the biological questions and are always happy to join us for a beer during Friday borrel (unless it's ping pong night of course). **Geeske**, doing the impossible task of keeping all of our paperwork in order and doing so gracefully, having taken over from **Barbara** close to two years ago; Everything would grind to a halt without you. **Johan**, the unsung hero of all my NMR colleagues making sure every machine in the building is running smoothly.

As I already mentioned though, more important than the exchange of information and expertise facilitated by the fact we share our building with our NMR colleagues (or rather, them sharing their building with us) are the very people themselves; People that I've grown to befriend over the past four years.

Starting with the solution group, **Ulric** the great, explainer in chief, pointer extraordinaire and purveyor of encyclopaedic knowledge on obscure and archaic topics such as the reproductive system of hyenas or medieval folklore involving unicorns. We share an interest in the noble pastime of getting lost in Wikipedia articles and for that alone I consider you a kindred spirit. **Heyi** the indomitable, always shouting at people for touching you and effortlessly swag-ing your way through life, **Klara**, possibly the most patient person I know and **Vincenzo**, your name and Dutch sandwiches you have during lunch always at odds. **Reinier**, a man with a plan (weekly, monthly, possibly yearly?) and always in control, and the most recent additions to the solid-state group **David** and **Agnes**, you both only just started but I'm sure you'll do great. **Rhythm**, same goes for you. You have some big shoes to fill but I'm sure you'll rise to the challenge. Also, I really am sorry you thought I was scary when you first met me. It's an occupational hazard when one engages in competitive frowning. Of course, all the people that have since moved on to new things, **Klaartje**, **Cecilia**, **Jon**, **Deni**, **Velten**.

Moving downstairs, to the part of the building that some people might have (affectionately) referred to as a dungeon, one will find possibly the happiest office in the entire building. **Alessandra**, your name, although beautiful, is nowhere near as representative of your true self as your nickname so I think you should stick to that from now on. You are beaming happiness in all directions around you and your excitedness towards small everyday things is infectious to the point where I can even forgive your subpar taste in pizza. **Miranda**, I'll do you the favour of refraining from using your nickname to refer to you here. I will say though that it is representative of many things which I like about you but most importantly your ability to prioritise the things which are important to you and compromise on less significant aspects of your life. However, the things I like the most about you, by far, are your steadfastness and impressive capacity for joy and laughter even when things were not going great. You'll never know how much of a relief it was when things were not great with me either. Never lose that. **Leona**, not only is your second name infinitely cooler than your first one, lions are traditionally used to represent courage and it takes a lot of it to go through with the changes that you have identified are required for your wellbeing. Same goes for you **Barend**.

Last, but certainly not least, of the Weingarth group, **João** and also last but equally not least of the Baldus group, **Sid**. It's fair to say that without you two my time in the Netherlands would have been significantly worse. I think you both know how I feel so I'll keep it short. **João**, you are one of the most gentle and kind people I know, despite your best efforts to conceal those aspects and by far the most hard-working one, especially considering your survival does not depend on it. As a matter of fact, you are so kind you could see flowers around my head. Although that could have been the Ginjinha too, not sure. You deserve every great thing that's coming your way – professionally and personally – and make no mistake, good things are coming your way. Just give it a bit of time. **Sid**, you are a testament to my constant amazement that people want to get to know me and are also patient enough for me to relax enough to actually let it happen. Even more importantly, when I moved to the Netherlands, I never expected I would meet someone who was interested in many of the same nerdy things I am. I have witnessed you undergoing some fairly impressive transformations, both physically and mentally, and I think you are all the better for them and look forward to the changes yet to come.

Panos

List of Publications

[1] **P.I. Koukos**, N.M. Glykos, Grcarma: A fully automated task-oriented interface for the analysis of molecular dynamics trajectories, J. Comput. Chem. 34 (2013) 2310–2312.

[2] **P.I. Koukos**, N.M. Glykos, On the application of good-turing statistics to quantify convergence of biomolecular simulations, J. Chem. Inf. Model. 54 (2014) 209–217.

[3] **P.I. Koukos**, N.M. Glykos, Folding molecular dynamics simulations accurately predict the effect of mutations on the stability and structure of a vammin-derived peptide, J. Phys. Chem. B. 118 (2014) 10076–10084.

[4] I.S. Moreira*, **P.I. Koukos***, R. Melo, J.G. Almeida, A.J. Preto, J. Schaarschmidt, M. Trellet, Z.H. Gümüş, J. Costa, A.M.J.J. Bonvin, SpotOn: High Accuracy Identification of Protein-Protein Interface Hot-Spots, Sci. Rep. (2017). doi:10.1038/s41598-017-08321-2.

[5] J.G. Almeida, A.J. Preto, **P.I. Koukos**, A.M.J.J. Bonvin, I.S. Moreira, Membrane proteins structures: A review on computational modeling tools, Biochim. Biophys. Acta - Biomembr. (2017). doi:10.1016/j.bbamem.2017.07.008.

[6] 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. Comput. Aided. Mol. Des. (2018). doi:10.1007/s10822-017-0049-y.

[7] **P.I. Koukos**, I. Faro, C.W. van Noort, A.M.J.J. Bonvin, A Membrane Protein Complex Docking Benchmark, J. Mol. Biol. (2018). doi:10.1016/j.jmb.2018.11.005.

[8] A.J. Preto, P. Matos-Filipe, **P.I. Koukos**, P. Renault, S.F. Sousa, I.S. Moreira, Structural Characterization of Membrane Protein Dimers, in: A.E. Kister (Ed.), Protein Supersecondary Struct. Methods Protoc., Springer New York, New York, NY, 2019: pp. 403–436. doi:10.1007/978-1-4939-9161-7_21.

[9] **P.I. Koukos**, L.C. Xue, A.M.J.J. Bonvin, Protein–ligand pose and affinity prediction: Lessons from D3R Grand Challenge 3, J. Comput. Aided. Mol. Des. (2019). doi:10.1007/s10822-018-0148-4.

[10] A. Vangone*, J. Schaarschmidt*, **P.I. Koukos***, C. Geng, N. Citro, M.E. Trellet, L.C. Xue, A.M.J.J. Bonvin, Large-scale prediction of binding affinity in protein-small ligand complexes: The PRODIGY-LIG web server, Bioinformatics. (2019). doi:10.1093/bioinformatics/bty816.

[11] **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 Struct. Funct. Bioinforma. (2019)., *Submitted*.

[12] A. Basciu, **P.I. Koukos**, G. Malloci, A.M.J.J. Bonvin, A.V. Vargiu, Coupling enhanced sampling of the apo-receptor with template-based ligand conformers selection: Performance in pose prediction in the D3R Grand Challenge 4, J. Comput. Aided. Mol. Des. (2019)., *Submitted*.

[13] **P.I. Koukos**, A.M.J.J. Bonvin, Integrative modelling of biomolecular complexes, J. Mol. Biol. (2019)., *Submitted*.

*: These authors contributed equally

Curriculum Vitae

Panagiotis Koukos was born on September 2nd 1989 in Athens, Greece. After graduating high school, he enrolled in the Molecular Biology and Genetics department of Democritus University of Thrace from which he graduated in October 2013. After a short research assistant placement, he enrolled in the Bioinformatics and Systems Biology MSc programme at Imperial College London in September 2014 from which he graduated one year later. Following a few months working as a developer he joined the Computational Structural Biology group of Alexandre Bonvin as a PhD candidate in March 2016 and for the four years that followed, he split his time between research, teaching and supervising students, and avoiding his PhD project. He submitted his thesis in October 2019 and is scheduled to defend it on the 26th of February 2020.