SCUOLA DI DOTTORATO
UNIVERSITÀ DEGLI STUDI DI MILANO-BICOCCA

Department of Earth and Environmental Sciences

PhD program in Chemical, Geological and Environmental Sciences
Cycle XXXVI
Curriculum in Chemical Sciences

A DATA-DRIVEN COMPUTATIONAL STUDY OF PROTEIN-PROTEIN AND PROTEIN-GLYCAN INTERACTIONS

Anna Ranaudo

Matr. 794101

Tutor: Prof. Claudio Greco

Supervisors: Prof. Giorgio Moro, Dr. Elisabetta Moroni, Dr. Alessandro Maiocchi

Coordinator: Prof. Marco Giovanni Malusà

ACADEMIC YEAR: 2022-2023

Abstract

This PhD thesis focuses on the employment of computational approaches for the study of the interactions between biomolecules, a broad term that accounts for different molecular species ranging from proteins to small ligands. Understanding how biomolecules recognize each other, thus giving rise to complexes or assemblies, is a key point for the comprehension of biological mechanisms in living organisms and for application purposes in a variety of fields, among which drug design. These difficult and multidisciplinary issues strongly exploit in silico approaches, which, in the last decades, have become increasingly efficient and essential for supporting and guiding the experiments.

The research activity carried out during my PhD work mainly dealt with two projects. The first one revolves around protein-protein interactions and concerns a specific use-case, namely the necessity to predict how two affitins bind the human epidermal growth factor receptor 2 (HER2). This project was carried out in collaboration with Dr. Alessandro Maiocchi (Bracco S.p.A – owner of two patents that cover the use of the two affitins as molecular probes targeting HER2), and Dr. Elisabetta Moroni (SCITEC, Italian National Research Council).

The second project was conducted at the Computational Structural Biology group (Bijvoet Centre for Biomolecular Research, Universiteit Utrecht) under the supervision of Prof. Alexandre Bonvin and Dr. Marco Giulini. It aims at building a reliable protocol, based on the software HADDOCK3, which is developed at the CSB group, for the prediction of protein-glycan complexes.

The thesis is structured as follows.

Section 1 briefly covers the role of biomolecular interactions and shows how the structure of the complexes they form can be determined by means of experimental and computational approaches. It is then explained how the projects discussed in the thesis fit into the framework just presented.

Section 2 illustrates the theoretical foundations of the main methodologies used in the two projects: Molecular Mechanics, Molecular Dynamics, Molecular Docking.

Section 3 concerns the project focused on the prediction of the complexes affitins-HER2.

Section 4 covers the project aimed at developing the protocol for the prediction of protein-glycan complexes.

Section 5 aims to summarise the work done, highlighting the main results and, at the same time, the future perspectives opened by the PhD work presented here.

Contents

List of abbreviations

AI	artificial intelligence
AIRs	Ambiguous Interaction Restraints
BSA	buried surface area
CAPRI	Critical Assessment of PRediction of Interactions
CASP	Critical Assessment of Structure Prediction
CNN	convolutional neural networks
cryo-EM	cryogenic electron microscopy
DL	deep learning
ECD	extracellular domain
EPP	protein-protein interaction energy
Fab	antigen-binding fragments
f.f.	force field
FCC	fraction of common contacts
FFT	fast Fourier transform
Fnat	fraction of native contacts
GBP	glycan binding proteins
HB	hydrogen bonds
HER2	human epidermal growth factor receptor 2
IL-RMSD	interface-ligand root-mean-square deviation
I-RMSD	interface root-mean-square deviation
L-RMSD	ligand root-mean-square deviation
mAb	monoclonal antibody
MD	Molecular Dynamics
MLCE	Matrix of Local Coupling Energies
MM	Molecular Mechanics
MSA	multiple sequence alignment
NMR	Nuclear Magnetic Resonance
PCA	Principal Component Analysis
PDB	Protein Data Bank

PES	potential energy surface
pLDDT	predicted Local Distance Difference Test
PPI	protein-protein interaction
RMSD	root-mean-square deviation
RMSF	root-mean-square fluctuation
rSD	relative standard deviation
SD	standard deviation
SR	success rate

Section 1 – How do biomolecules interact? The role of computational approaches

This section aims to illustrate the importance of a detailed knowledge of the structure of complexes involving proteins, peptides, DNA, carbohydrates, and small molecules in understanding the mechanisms of interaction of biomolecular systems (**Section 1.1**).

The most common experimental methods used to solve the structures of complexes will be briefly covered (**Section 1.2**). The essential contribution given by the *in silico* approaches will then be addressed, with a focus on how the main computational methods have advanced in recent years and how they could be further improved to shed light on the intricate problem of biomolecular interactions (**Section 1.3**).

The necessity of adopting an integrative methodology, that combines diverse experimental tests with both evolution-/function- and physics-based modelling procedures, will be highlighted at the end of the section.

Finally, a paragraph will be dedicated to explaining how this PhD thesis fits into this framework (**Section 1.4**).

1.1 – Why and how to study biomolecular interactions

The human genome, fully sequenced in 2022[1], consists of about 20000 genes encoding at least as many different proteins[2]; this number however depends on how the human proteome is defined[3].

It has been estimated that more than 80% of proteins are involved in complexes in which they bind other protein partners. The Human Reference Protein Interactome Mapping Project is trying to reach the most comprehensive picture of protein-protein interactions (PPI) in the human body[4]. To date, more than 64000 PPI have been mapped (http://www.interactome-atlas.org/).

The formation of protein–protein complexes is driven by the free energy of the process, which is mainly related to physicochemical and geometric properties of the interface. Many studies[5] have been performed with the aim of understanding what the key features of PPI are.

PPI are regulated by several mechanisms[6] and are essential for the biological functions of the organisms. The knowledge of how proteins interact with each other is essential for several reasons[7]. For example, the presence of mutations in the amino acidic sequence can lead to conformational changes that in turn could interfere with the structure of a protein-protein complex. This could give rise to diseases such as cancer pathologies[8,9]. In addition, bacteria and viruses attack host cells by interacting with the receptors on the host's cell-surface[10]. These interactions can also be mediated by other biomolecules, such as glycans. This is the case with the SARS-CoV-2 spike protein. This protein, which enters host cells by connecting to the angiotensin-converting enzyme (ACE2), is surrounded by a layer of glycans to hide from the immune system. Some specific glycans play a crucial role in the movement and structure of the part of the spike protein that binds to ACE2[11]. Removal of these sugars results in diminished binding to ACE2, highlighting potential targets on the spike protein for vaccine design. In general, knowledge of the structure of protein-protein (or protein-biomolecule) complexes is necessary for the design of modulators of the interactions[12]. Finally, characterization of PPI can also serve to understand the role of proteins whose function is unknown, if the latter is known for the protein partner[13].

There is therefore a clear need to know as many protein structures and protein-protein complexes as possible. This goal can be addressed by both experimental methods and *in silico* approaches[7], as shown graphically in *Figure 1.1*.

Experimental methods, which were obviously the only possibility until around the late 1970s[14], allow direct observation of the physical phenomenon of partners binding, thus providing unequivocal results. However, they are linked to several limitations, from sample preparation and proper set-up of instrumentation to the time and the costs required for the analysis. Issues related to waste treatment and the overall greenness of the process should be considered too.

On the other hand, *in silico* approaches represent an increasingly powerful resource in structural biology, among which artificial intelligence (AI) methods really are a breakthrough in the last few years[15].

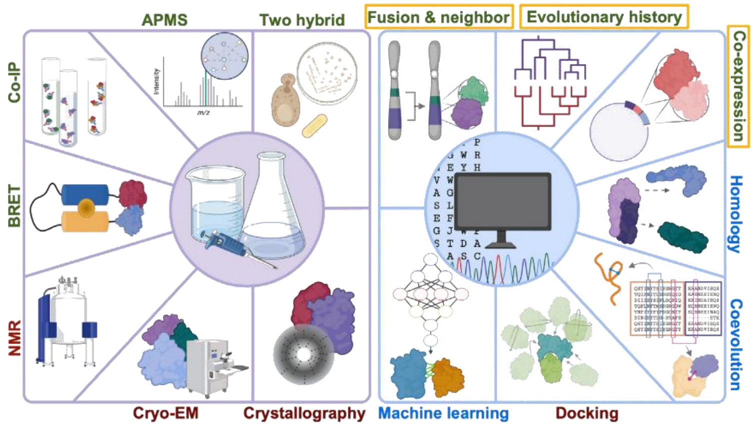
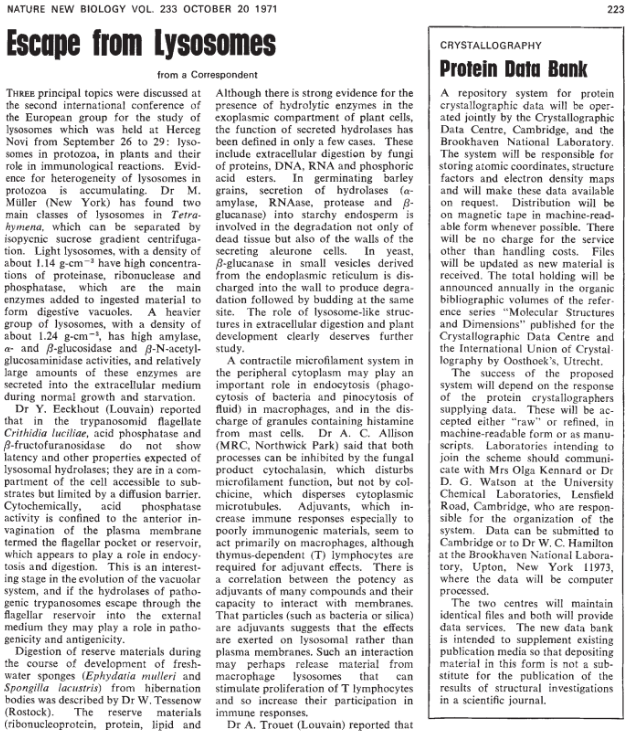


Figure 1.1 – Overview of experimental (left) and computational (right) methods for PPI detection and determination of the structure of protein complexes. The figure is reproduced from reference[7].

The Protein Data Bank (PDB)[16], whose birth was announced in a 1971 Nature issue[17] (*Figure 1.2*), collects since then the atomic coordinates of proteins and the complexes they form with protein partners or other biomolecules such as nucleic acids, oligosaccharides, and other small ligands.



NATURE NEW BIOLOGY VOL. 233 OCTOBER 20 1971 223

Escape from Lysosomes

from a Correspondent

THREE principal topics were discussed at the second international conference of the European group for the study of lysosomes which was held at Herceg Novi from September 26 to 29: lysosomes in protozoa, in plants and their role in immunological reactions. Evidence for heterogeneity of lysosomes in protozoa is accumulating. Dr M. Müller (New York) has found two main classes of lysosomes in *Tetrahymena*, which can be separated by isopycnic sucrose gradient centrifugation. Light lysosomes, with a density of about 1.14 g-cm^{-3} have high concentrations of proteinase, ribonuclease and phosphatase, which are the main enzymes added to ingested material to form digestive vacuoles. A heavier group of lysosomes, with a density of about 1.24 g-cm^{-3}, has high amylase, α- and β-glucosidase and β-N-acetylglucosaminidase activities, and relatively large amounts of these enzymes are secreted into the extracellular medium during normal growth and starvation.

Dr Y. Eeckhout (Louvain) reported that in the trypanosomid flagellate *Crithidia luciliae*, acid phosphatase and β-fructofuranosidase do not show latency and other properties expected of lysosomal hydrolases; they are in a compartment of the cell accessible to substrates but limited by a diffusion barrier. Cytochemically, acid phosphatase activity is confined to the anterior invagination of the plasma membrane termed the flagellar pocket or reservoir, which appears to play a role in endocytosis and digestion. This is an interesting stage in the evolution of the vacuolar system, and if the hydrolases of pathogenic trypanosomes escape through the flagellar reservoir into the external medium they may play a role in pathogenicity and antigenicity.

Digestion of reserve materials during the course of development of freshwater sponges (*Ephydatia mulleri* and *Spongilla lacustris*) from hibernation bodies was described by Dr W. Tessenow (Rostock). The reserve materials (ribonucleoprotein, protein, lipid and

Although there is strong evidence for the presence of hydrolytic enzymes in the exoplasmic compartment of plant cells, the function of secreted hydrolases has been defined in only a few cases. These include extracellular digestion by fungi of proteins, DNA, RNA and phosphoric acid esters. In germinating barley grains, secretion of hydrolases (α-amylase, RNAase, protease and β-glucanase) into starchy endosperm is involved in the degradation not only of dead tissue but also of the walls of the secreting aleurone cells. In yeast, β-glucanase in small vesicles derived from the endoplasmic reticulum is discharged into the wall to produce degradation followed by budding at the same site. The role of lysosome-like structures in extracellular digestion and plant development clearly deserves further study.

A contractile microfilament system in the peripheral cytoplasm may play an important role in endocytosis (phagocytosis of bacteria and pinocytosis of fluid) in macrophages, and in the discharge of granules containing histamine from mast cells. Dr A. C. Allison (MRC, Northwick Park) said that both processes can be inhibited by the fungal product cytochalasin, which disturbs microfilament function, but not by colchicine, which disperses cytoplasmic microtubules. Adjuvants, which increase immune responses especially to poorly immunogenic materials, seem to act primarily on macrophages, although thymus-dependent (T) lymphocytes are required for adjuvant effects. There is a correlation between the potency as adjuvants of many compounds and their capacity to interact with membranes. That particles (such as bacteria or silica) are adjuvants suggests that the effects are exerted on lysosomal rather than plasma membranes. Such an interaction may perhaps release material from macrophage lysosomes that can stimulate proliferation of T lymphocytes and so increase their participation in immune responses.

Dr A. Trouet (Louvain) reported that

CRYSTALLOGRAPHY

Protein Data Bank

A repository system for protein crystallographic data will be operated jointly by the Crystallographic Data Centre, Cambridge, and the Brookhaven National Laboratory. The system will be responsible for storing atomic coordinates, structure factors and electron density maps and will make these data available on request. Distribution will be on magnetic tape in machine-readable form whenever possible. There will be no charge for the service other than handling costs. Files will be updated as new material is received. The total holding will be announced annually in the organic bibliographic volumes of the reference series "Molecular Structures and Dimensions" published for the Crystallographic Data Centre and the International Union of Crystallography by Oosthoek's, Utrecht.

The success of the proposed system will depend on the response of the protein crystallographers supplying data. These will be accepted either "raw" or refined, in machine-readable form or as manuscripts. Laboratories intending to join the scheme should communicate with Mrs Olga Kennard or Dr D. G. Watson at the University Chemical Laboratories, Lensfield Road, Cambridge, who are responsible for the organization of the system. Data can be submitted to Cambridge or to Dr W. C. Hamilton at the Brookhaven National Laboratory, Upton, New York 11973, where the data will be computer processed.

The two centres will maintain identical files and both will provide data services. The new data bank is intended to supplement existing publication media so that depositing material in this form is not a substitute for the publication of the results of structural investigations in a scientific journal.

Figure 1.2 – The page of the Nature issue[17] were the Protein Data Bank was first introduced.

As of the current date (October 23rd, 2023), 210836 structures are deposited in the PDB.

Most of them (179069) have been determined by X-ray crystallography, 17202 by (cryogenic) electron microscopy (cryo-EM), and 14013 by nuclear magnetic resonance spectroscopy (NMR), as reported in *Table 1.1*.

Molecular Type	X-ray	EM	NMR	Multiple methods	Neutron	Other	Total
Protein (only)	158541	11607	12285	197	73	32	182735
Protein/Oligosaccharide	9250	2042	34	8	1	0	11335
Protein/Nucleic Acid	8277	3651	284	7	0	0	12219
Nucleic Acid (only)	2727	109	1467	13	3	1	4320
Other	164	9	32	0	0	0	205
Oligosaccharide (only)	11	0	6	1	0	4	22
Total	178970	17418	14108	226	77	37	210836

Table 1.1 – PDB Data Distribution by Experimental Method and Molecular Type (https://www.rcsb.org/stats/summary, accessed on October 23rd, 2023).

As shown in *Figure 1.3*, X-ray crystallography was the only technique employed until the late 1980s, when NMR experienced significant growth that lasted until 2007-2008. Cryo-EM has been increasingly used over the past two decades, and this trend is still growing thanks to the atomic-level resolution that is now achievable[18].

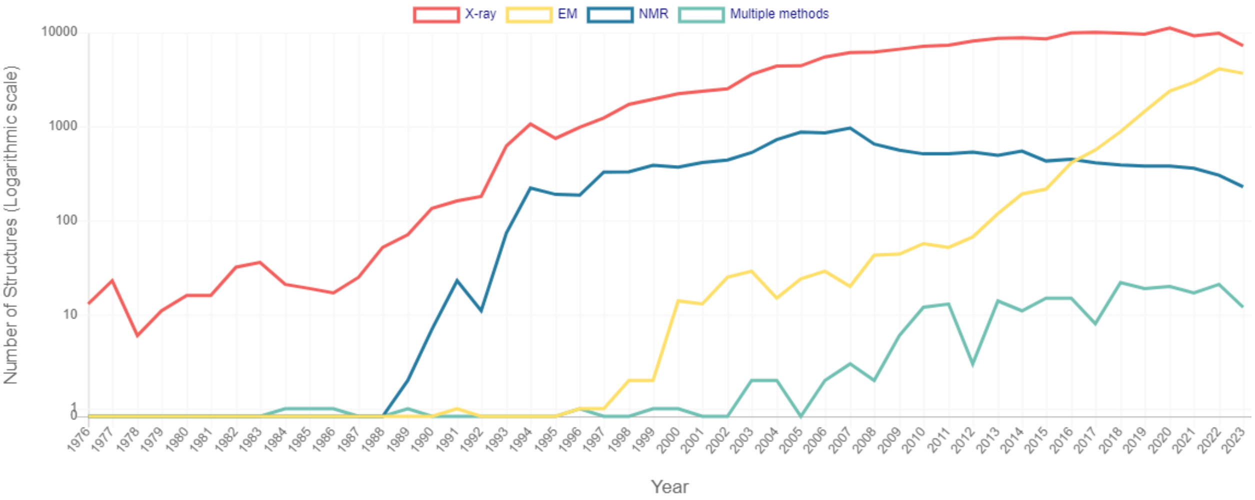


Figure 1.3 - Number of Released PDB Structures per Year (https://www.rcsb.org/stats/all-released-structures, accessed on October 23rd, 2023).

Some structures have also been determined by the use of multiple methods simultaneously (*Table 1.1*), e.g., by coupling solution NMR with solid-state NMR or with X-ray diffraction, a useful approach when the use of a single technique is not sufficient for the reliable determination of a three-dimensional structure.

In addition to the structures determined by experimental methods, since 2020 the PDB has a section called "Computed Structure Models (CSM)". Here more than 1 million structures are collected, that is, almost 5 times the experimentally determined structures, mainly retrieved from the AlphaFoldDB (https://alphafold.ebi.ac.uk/). Although some of them are predicted with low confidence (around one fifth of them shows global predicted local-distance difference (pLDDT) score[19] < 70), more than 35% have a global pLDDT > 90, corresponding to a high accuracy cut-off[20]. This highlights the undoubted central role of *in silico* approaches, and especially artificial intelligence (AI) methods, in determining biomolecular structures.

The next sections will provide a brief overview of experimental methods (**Section 1.2**) and *in silico* approaches (**Section 1.3**).

1.2 – Experimental methods

X-ray crystallography

As shown in *Table 1.1*, X-ray crystallography is by far the most employed technique. Briefly[21], the three-dimensional structure of a protein, or of a complex, is determined starting from a crystal; a highly concentrated, purified sample is thus required. The crystal is then subjected to an X-ray beam. The diffraction patterns produced are analysed, initially providing details about the symmetry of the crystal packing and the dimensions of its repeating unit, evident from the arrangement of diffraction spots. The brightness of these spots allows for the calculation of structure factors, which in turn provide a depiction of electron density. Through several enhancement techniques, this electron density map is refined to a clarity level that facilitates the construction of the molecular structure, based on the protein sequence. Finally, this derived structure undergoes further refinement to better align with the map and to assume a conformation that is thermodynamically optimal.

The limits of this technique are manifold. To start with, the crystallization of the sample already implies some difficulties in that the preparation is not always straightforward, due for example to instability issues. Moreover, it is not granted at all that the conformation assumed by protein in a crystal coincides with the conformation it would assume in physiological conditions, i.e., in its natural environment, where multiple states, equally favourable from a thermodynamic point of view, could also exist. Thus, the dynamic nature of a protein is poorly accounted for with X-ray crystallography, a limitation that does not to allow a complete view of a protein's behaviour.

Nuclear Magnetic Resonance

As depicted in *Figure 1.3*, NMR spectroscopy has been extensively used in the last ten years of the past century, before reaching a plateau.

Usually, the procedure for the determination of a protein structure with NMR involves the four following stages[22]: i) preparation of the isotope-labelled protein sample; ii) NMR data collection and analysis, with the assignment of chemical shifts of ^{1}H, ^{15}N, and ^{13}C atoms; iii) calculation of the structure and refinement using distance and/or orientation restraints, e.g., nuclear Overhauser effect (NOE)-derived restraints or residual dipolar coupling orientation restraints; iv) structural quality assessment.

One of the challenges in this well-assessed technique involves the improvement of in-cell NMR[23], which is used for studying macromolecules in living cells. It has been shown[23] that the combination of atomic-level characterization by classical solution NMR with in-cell NMR allows previously unreached insights into cellular processes and drug efficacy. Moreover, NMR still faces problems related to sensitivity and timing of data analysis[24].

Cryogenic Electron Microscopy

The 2017 Nobel Prize in Chemistry was assigned for "developing cryo-electron microscopy for the high-resolution structure determination of biomolecules in solution"[25]. This technique has been increasingly used in the last twenty years, and this trend still is growing. In 2020, the structure of apoferritin was determined with atomic resolution (1.25 Å)[18]. Cryo-EM[26] is based on 3D electron microscopy (3D-EM), where biological samples are directly visualized using transmission electron microscopy (TEM), prior to a treatment were the sample is placed on a thin support and frozen in order to minimize the damage of the radiations. TEM images correspond to 2D projections of the

3D particles in the sample, which contains multiple copies of the same particle in many different orientations. The several 2D views from various angles of the sample are merged into one 3D image thanks to the projection-slice theorem. This theorem states that the Fourier transform of a 2D projection is a central slice through the 3D Fourier transform (passing through the origin) of the underlying structure, the projection direction being orthogonal to the slice. Hence, if the angles of various 2D views are known, their respective 2D Fourier slices can be correctly placed within the 3D transform, allowing for the computation of the original 3D form using the inverse Fourier transform.

One of the main advantages, at least with respect to X-ray crystallography, is that cryo-EM does not require a crystallized sample: smaller quantities are thus necessary for performing the analysis. On the other hand[24], the low sample concentrations could cause the dissociation of weakly associated complexes, whose determination is thus more difficult. An emerging technique is the microcrystal electron diffraction (MicroED)[27], which overcomes problems related to the size and allows to study membranes and protein-drug interactions.

1.3 – *In silico* approaches

To get an idea of how *in silico* approaches are increasingly being applied in the study of proteins, a search in Scopus was conducted for articles that include the terms 'docking', 'machine learning', 'AlphaFold', and 'AlphaFold2' in the title, abstract or keywords. The trend in the occurrence of these terms, along with 'protein' and 'structure', is shown in *Figure 1.4*, for the time interval 2000-2022.

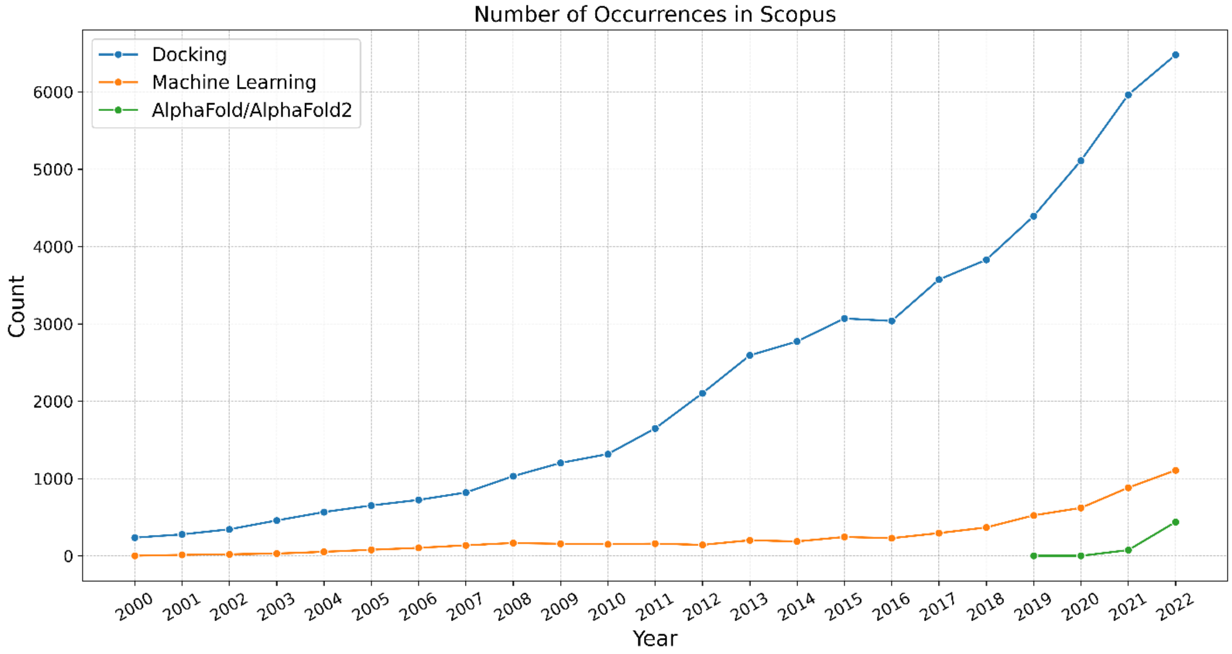


Figure 1.4 – Number of occurrences in Scopus (https://www.scopus.com/) of the terms Docking, Machine Learning, and AlphaFold or AlphaFold2, in conjunction with the term 'protein' and 'structure'. The Scopus database was accessed on October 25th, 2023.

In silico approaches can be coarsely divided into function-/ evolution-based methods and physics-based methods; however, this classification is not rigid, as the two can be combined to achieve greater accuracy[7]. Function-/ evolution-based methods are presented in the next paragraph, while physics-based methods, which are the approach adopted in this thesis, will be briefly introduced here and described in detail in **Section 2**.

Function-/ evolution-based methods

Function-/ evolution-based methods rely on the following points[7]. First, interacting proteins tend to be encoded by genes that are located nearby in the genome, and they co-occur in similar species, showing similar evolutionary rates. Interacting proteins are also co-expressed in the same tissues, at the same time. Moreover, the same protein-protein interactions occur in different species, thus allowing homology modelling to be exploited for their detection. Finally, coevolution, i.e., the process whereby two distinct residues within a protein or between two proteins mutually influence their evolutionary paths, is typically observed for residues that are in direct contact.

The prediction of a protein structure can thus exploit the knowledge that coevolving residues are usually close in space[28]. The use of appropriate global statistical methods, able to analyse coevolutionary signals from deep multiple sequence alignments (MSA), has led, since the 11th round of the Critical Assessment of Structure Prediction (CASP), to significant improvement in the *de novo* prediction of complex protein structures[29].

Further progress was achieved a few years later, when convolutional neural networks (CNNs) were introduced to translate MSA covariation into the likelihood of interactions between residues[30].

This advancement notably increased the accuracy of structure prediction, inspiring deep learning (DL)-based methods such as the first version of AlphaFold[31], which was the top performer in CASP13[32], or RoseTTAFold[33]. With the growth of DL techniques, the version 2 of AlphaFold[20], sometimes found as AlphaFold2, achieved near-atomic precision in structure prediction during the CASP14-CAPRI experiment[34]. The CASP-CAPRI experiment will be covered in detail in **Section 2.3**.

DL-based methods for protein structure prediction, such as AlphaFold, are probably the biggest revolution in structural biology since a long time. They are expected to be a key topic in research for the years to come[15], also considering that there is still room for improvement.

For example, a current limitation observed in AlphaFold is the dependence of the accuracy of the model on the depth of MSA, i.e., on the number of sequences. Although MSA information is essential for the coarse definition of the structure, it is not the only factor that influences the refinement of the models[20].

In general, predictions may fail when the amount of coevolution information is insufficient or not available at all, as is the case, for instance, with proteins engineered to bind a specific target. The prediction of the binding of small ligands (glycans, small molecules and cofactors) to proteins has not yet been solved. This problem is currently being addressed by AlphaFill[35], which adds the ligands to the protein models predicted by AlphaFold based on sequence and structure similarity to experimentally known structures.

In addition to predicting particularly tricky structures, one of the greatest challenges that methods such as AlphaFold need to address is probably how to account for the intrinsic dynamic nature of biomolecules, which is essential in their biological activity. The problem lies in the data used to train the neural networks on which these methods are based. These data, which consist mainly of the atomic coordinates of experimentally determined structures, provide only a "static" view of the structures. However, it is known that proteins and biomolecules in general should be better represented as an ensemble of conformations.

Therefore, even though we are in the era of AI methods, physics-based methods are still necessary to address the above-mentioned issues.

Physics-based methods

Physics-based methods study the interactions between biomolecules by modelling the physical forces that are responsible for the biding between the two partners. These methods rely on a classical description of the system and include molecular mechanics (MM), molecular dynamics (MD), and molecular docking; they are described in detail in **Section 2**.

Their peculiarity, compared to methods based only on evolutionary information, is that they can take into account the dynamic nature of biomolecules and their interactions, which in many cases is not negligible.

To conclude, function-/ evolution-based methods and physics-based methods can be coupled to achieve greater accuracy[36]. This has been done, for instance, with iScore[37], which combines HADDOCK energy terms[38], accounting for the empirical / physical part, with a score obtained using a graph representation of protein–protein interfaces and a measure of evolutionary conservation. Instead, in Feig's work[39,40], MD simulations are used, among other things, to refine AlphaFold models.

1.4 – Thesis framework

As remarked in a very recent Nature Viewpoint article[24], each of the most common experimental techniques provides an incomplete picture of the biomolecule one tries to visualise. On the other hand, *in silico* approaches, while showing increasingly impressive performance, cannot give unambiguous answers to the intricate puzzle of biomolecular interactions. Therefore, they cannot substitute experiments.

Instead, different methods, both experimental and *in silico*, should be combined, thus applying an integrative approach[41] that would provide a more comprehensive view of the object of the study.

This PhD thesis fits in the picture just drawn, in the sense that physics-based computational studies of different biomolecules and their interactions are carried out, while at the same time accounting for some experimentally derived information.

This certainly applies to the project conducted in collaboration with Bracco S.p.A. With the aim of understanding how two affitins bind the human epidermal growth factor receptor 2 (HER2), information on competition for specific HER2 sites is exploited to guide docking calculations. Once docking models are obtained, the availability of three-dimensional structures of HER2 with other protein partners is used to perform targeted experimental tests, which are necessary for an unambiguous determination of the binding interface.

The project carried out at the Computational Structural Biology group (Universiteit Utrecht) aims to build a reliable protocol for the prediction of protein-glycan complexes by making use of the in-house developed HADDOCK3 docking programme. Here, docking calculations are not fully blind either: the study is conducted on a dataset of known complexes, thus information about the protein interface is used as a restraint to drive the docking calculations. In a realistic scenario, where the three-dimensional structures would not be available, the interface residues could be identified experimentally by, for example, the analysis of NMR chemical shift perturbation.

References

(1) Nurk, S.; Koren, S.; Rhie, A.; Rautiainen, M.; Bzikadze, A. V.; Mikheenko, A.; Vollger, M. R.; Altemose, N.; Uralsky, L.; Gershman, A.; Aganezov, S.; Hoyt, S. J.; Diekhans, M.; Logsdon, G. A.; Alonge, M.; Antonarakis, S. E.; Borchers, M.; Bouffard, G. G.; Brooks, S. Y.; Caldas, G. V.; Chen, N.-C.; Cheng, H.; Chin, C.-S.; Chow, W.; de Lima, L. G.; Dishuck, P. C.; Durbin, R.; Dvorkina, T.; Fiddes, I. T.; Formenti, G.; Fulton, R. S.; Fungtammasan, A.; Garrison, E.; Grady, P. G. S.; Graves-Lindsay, T. A.; Hall, I. M.; Hansen, N. F.; Hartley, G. A.; Haukness, M.; Howe, K.; Hunkapiller, M. W.; Jain, C.; Jain, M.; Jarvis, E. D.; Kerpedjiev, P.; Kirsche, M.; Kolmogorov, M.; Korlach, J.; Kremitzki, M.; Li, H.; Maduro, V. V.; Marschall, T.; McCartney, A. M.; McDaniel, J.; Miller, D. E.; Mullikin, J. C.; Myers, E. W.; Olson, N. D.; Paten, B.; Peluso, P.; Pevzner, P. A.; Porubsky, D.; Potapova, T.; Rogaev, E. I.; Rosenfeld, J. A.; Salzberg, S. L.; Schneider, V. A.; Sedlazeck, F. J.; Shafin, K.; Shew, C. J.; Shumate, A.; Sims, Y.; Smit, A. F. A.; Soto, D. C.; Sović, I.; Storer, J. M.; Streets, A.; Sullivan, B. A.; Thibaud-Nissen, F.; Torrance, J.; Wagner, J.; Walenz, B. P.; Wenger, A.; Wood, J. M. D.; Xiao, C.; Yan, S. M.; Young, A. C.; Zarate, S.; Surti, U.; McCoy, R. C.; Dennis, M. Y.; Alexandrov, I. A.; Gerton, J. L.; O'Neill, R. J.; Timp, W.; Zook, J. M.; Schatz, M. C.; Eichler, E. E.; Miga, K. H.; Phillippy, A. M. The Complete Sequence of a Human Genome. *Science (80-.).* **2022**, *376* (6588), 44–53. https://doi.org/10.1126/science.abj6987.

(2) Ponomarenko, E. A.; Poverennaya, E. V.; Ilgisonis, E. V.; Pyatnitskiy, M. A.; Kopylov, A. T.; Zgoda, V. G.; Lisitsa, A. V.; Archakov, A. I. The Size of the Human Proteome: The Width and Depth. *Int. J. Anal. Chem.* **2016**, *2016*, 1–6. https://doi.org/10.1155/2016/7436849.

(3) Aebersold, R.; Agar, J. N.; Amster, I. J.; Baker, M. S.; Bertozzi, C. R.; Boja, E. S.; Costello, C. E.; Cravatt, B. F.; Fenselau, C.; Garcia, B. A.; Ge, Y.; Gunawardena, J.; Hendrickson, R. C.; Hergenrother, P. J.; Huber, C. G.; Ivanov, A. R.; Jensen, O. N.; Jewett, M. C.; Kelleher, N. L.; Kiessling, L. L.; Krogan, N. J.; Larsen, M. R.; Loo, J. A.; Ogorzalek Loo, R. R.; Lundberg, E.; MacCoss, M. J.; Mallick, P.; Mootha, V. K.; Mrksich, M.; Muir, T. W.; Patrie, S. M.; Pesavento, J. J.; Pitteri, S. J.; Rodriguez, H.; Saghatelian, A.; Sandoval, W.; Schlüter, H.; Sechi, S.; Slavoff, S. A.; Smith, L. M.; Snyder, M. P.; Thomas, P. M.; Uhlén, M.; Van Eyk, J. E.; Vidal, M.; Walt, D. R.; White, F. M.; Williams, E. R.;

Wohlschlager, T.; Wysocki, V. H.; Yates, N. A.; Young, N. L.; Zhang, B. How Many Human Proteoforms Are There? *Nat. Chem. Biol.* **2018**, *14* (3), 206–214. https://doi.org/10.1038/nchembio.2576.

(4) Luck, K.; Kim, D.-K.; Lambourne, L.; Spirohn, K.; Begg, B. E.; Bian, W.; Brignall, R.; Cafarelli, T.; Campos-Laborie, F. J.; Charloteaux, B.; Choi, D.; Coté, A. G.; Daley, M.; Deimling, S.; Desbuleux, A.; Dricot, A.; Gebbia, M.; Hardy, M. F.; Kishore, N.; Knapp, J. J.; Kovács, I. A.; Lemmens, I.; Mee, M. W.; Mellor, J. C.; Pollis, C.; Pons, C.; Richardson, A. D.; Schlabach, S.; Teeking, B.; Yadav, A.; Babor, M.; Balcha, D.; Basha, O.; Bowman-Colin, C.; Chin, S.-F.; Choi, S. G.; Colabella, C.; Coppin, G.; D'Amata, C.; De Ridder, D.; De Rouck, S.; Duran-Frigola, M.; Ennajdaoui, H.; Goebels, F.; Goehring, L.; Gopal, A.; Haddad, G.; Hatchi, E.; Helmy, M.; Jacob, Y.; Kassa, Y.; Landini, S.; Li, R.; van Lieshout, N.; MacWilliams, A.; Markey, D.; Paulson, J. N.; Rangarajan, S.; Rasla, J.; Rayhan, A.; Rolland, T.; San-Miguel, A.; Shen, Y.; Sheykhkarimli, D.; Sheynkman, G. M.; Simonovsky, E.; Taşan, M.; Tejeda, A.; Tropepe, V.; Twizere, J.-C.; Wang, Y.; Weatheritt, R. J.; Weile, J.; Xia, Y.; Yang, X.; Yeger-Lotem, E.; Zhong, Q.; Aloy, P.; Bader, G. D.; De Las Rivas, J.; Gaudet, S.; Hao, T.; Rak, J.; Tavernier, J.; Hill, D. E.; Vidal, M.; Roth, F. P.; Calderwood, M. A. A Reference Map of the Human Binary Protein Interactome. *Nature* **2020**, *580* (7803), 402–408. https://doi.org/10.1038/s41586-020-2188-x.

(5) Chakrabarti, P.; Janin, J. Dissecting Protein–Protein Recognition Sites. *Proteins Struct. Funct. Bioinforma.* **2002**, *47* (3), 334–343. https://doi.org/10.1002/prot.10085.

(6) Berggård, T.; Linse, S.; James, P. Methods for the Detection and Analysis of Protein–Protein Interactions. *Proteomics* **2007**, *7* (16), 2833–2842. https://doi.org/10.1002/pmic.200700131.

(7) Durham, J.; Zhang, J.; Humphreys, I. R.; Pei, J.; Cong, Q. Recent Advances in Predicting and Modeling Protein–Protein Interactions. *Trends Biochem. Sci.* **2023**, *48* (6), 527–538. https://doi.org/10.1016/j.tibs.2023.03.003.

(8) Cheng, F.; Zhao, J.; Wang, Y.; Lu, W.; Liu, Z.; Zhou, Y.; Martin, W. R.; Wang, R.; Huang, J.; Hao, T.; Yue, H.; Ma, J.; Hou, Y.; Castrillon, J. A.; Fang, J.; Lathia, J. D.; Keri, R. A.; Lightstone, F. C.; Antman, E. M.; Rabadan, R.; Hill, D. E.; Eng, C.; Vidal, M.; Loscalzo, J. Comprehensive Characterization of Protein–Protein Interactions Perturbed by Disease Mutations. *Nat. Genet.* **2021**, *53* (3), 342–353. https://doi.org/10.1038/s41588-020-00774-

y.

(9) Kim, M.; Park, J.; Bouhaddou, M.; Kim, K.; Rojc, A.; Modak, M.; Soucheray, M.; McGregor, M. J.; O'Leary, P.; Wolf, D.; Stevenson, E.; Foo, T. K.; Mitchell, D.; Herrington, K. A.; Muñoz, D. P.; Tutuncuoglu, B.; Chen, K.-H.; Zheng, F.; Kreisberg, J. F.; Diolaiti, M. E.; Gordan, J. D.; Coppé, J.-P.; Swaney, D. L.; Xia, B.; van 't Veer, L.; Ashworth, A.; Ideker, T.; Krogan, N. J. A Protein Interaction Landscape of Breast Cancer. *Science (80-.).* **2021**, *374* (6563). https://doi.org/10.1126/science.abf3066.

(10) Brito, A. F.; Pinney, J. W. Protein–Protein Interactions in Virus–Host Systems. *Front. Microbiol.* **2017**, *8*, 1–11. https://doi.org/10.3389/fmicb.2017.01557.

(11) Casalino, L.; Gaieb, Z.; Goldsmith, J. A.; Hjorth, C. K.; Dommer, A. C.; Harbison, A. M.; Fogarty, C. A.; Barros, E. P.; Taylor, B. C.; McLellan, J. S.; Fadda, E.; Amaro, R. E. Beyond Shielding: The Roles of Glycans in the SARS-CoV-2 Spike Protein. *ACS Cent. Sci.* **2020**, *6* (10), 1722–1734. https://doi.org/10.1021/acscentsci.0c01056.

(12) Lu, H.; Zhou, Q.; He, J.; Jiang, Z.; Peng, C.; Tong, R.; Shi, J. Recent Advances in the Development of Protein–Protein Interactions Modulators: Mechanisms and Clinical Trials. *Signal Transduct. Target. Ther.* **2020**, *5* (1), 213. https://doi.org/10.1038/s41392-020-00315-3.

(13) Zhang, Q. C.; Petrey, D.; Deng, L.; Qiang, L.; Shi, Y.; Thu, C. A.; Bisikirska, B.; Lefebvre, C.; Accili, D.; Hunter, T.; Maniatis, T.; Califano, A.; Honig, B. Structure-Based Prediction of Protein–Protein Interactions on a Genome-Wide Scale. *Nature* **2012**, *490* (7421), 556–560. https://doi.org/10.1038/nature11503.

(14) Wodak, S. J.; Janin, J. Computer Analysis of Protein-Protein Interaction. *J. Mol. Biol.* **1978**, *124* (2), 323–342. https://doi.org/10.1016/0022-2836(78)90302-9.

(15) Artificial Intelligence in Structural Biology Is Here to Stay. *Nature* **2021**, *595* (7869), 625–626. https://doi.org/10.1038/d41586-021-02037-0.

(16) Berman, H. M. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28* (1), 235–242. https://doi.org/10.1093/nar/28.1.235.

(17) Crystallography: Protein Data Bank. *Nat. New Biol.* **1971**, *233* (42), 223–223. https://doi.org/10.1038/newbio233223b0.

(18) Yip, K. M.; Fischer, N.; Paknia, E.; Chari, A.; Stark, H. Atomic-Resolution Protein Structure Determination by Cryo-EM. *Nature* **2020**, *587* (7832), 157–161.

https://doi.org/10.1038/s41586-020-2833-4.

(19) Tunyasuvunakool, K.; Adler, J.; Wu, Z.; Green, T.; Zielinski, M.; Žídek, A.; Bridgland, A.; Cowie, A.; Meyer, C.; Laydon, A.; Velankar, S.; Kleywegt, G. J.; Bateman, A.; Evans, R.; Pritzel, A.; Figurnov, M.; Ronneberger, O.; Bates, R.; Kohl, S. A. A.; Potapenko, A.; Ballard, A. J.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Clancy, E.; Reiman, D.; Petersen, S.; Senior, A. W.; Kavukcuoglu, K.; Birney, E.; Kohli, P.; Jumper, J.; Hassabis, D. Highly Accurate Protein Structure Prediction for the Human Proteome. *Nature* **2021**, *596* (7873), 590–596. https://doi.org/10.1038/s41586-021-03828-1.

(20) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596* (7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2.

(21) Smyth, M. S. X Ray Crystallography. *Mol. Pathol.* **2000**, *53* (1), 8–14. https://doi.org/10.1136/mp.53.1.8.

(22) Hu, Y.; Cheng, K.; He, L.; Zhang, X.; Jiang, B.; Jiang, L.; Li, C.; Wang, G.; Yang, Y.; Liu, M. NMR-Based Methods for Protein Analysis. *Anal. Chem.* **2021**, *93* (4), 1866–1879. https://doi.org/10.1021/acs.analchem.0c03830.

(23) Luchinat, E.; Banci, L. In-Cell NMR: Recent Progresses and Future Challenges. *Rend. Lincei. Sci. Fis. e Nat.* **2023**, *34* (3), 653–661. https://doi.org/10.1007/s12210-023-01168-y.

(24) Bai, X.; Gonen, T.; Gronenborn, A. M.; Perrakis, A.; Thorn, A.; Yang, J. Challenges and Opportunities in Macromolecular Structure Determination. *Nat. Rev. Mol. Cell Biol.* **2024**, *25* (1), 7–12. https://doi.org/10.1038/s41580-023-00659-y.

(25) Cressey, D.; Callaway, E. Cryo-Electron Microscopy Wins Chemistry Nobel. *Nature* **2017**, *550* (7675), 167–167. https://doi.org/10.1038/nature.2017.22738.

(26) Nogales, E.; Scheres, S. H. W. Cryo-EM: A Unique Tool for the Visualization of Macromolecular Complexity. *Mol. Cell* **2015**, *58* (4), 677–689. https://doi.org/10.1016/j.molcel.2015.02.019.

(27) Nannenga, B. L.; Gonen, T. The Cryo-EM Method Microcrystal Electron Diffraction (MicroED). *Nat. Methods* **2019**, *16* (5), 369–379. https://doi.org/10.1038/s41592-019-0395-x.

(28) Marks, D. S.; Colwell, L. J.; Sheridan, R.; Hopf, T. A.; Pagnani, A.; Zecchina, R.; Sander, C. Protein 3D Structure Computed from Evolutionary Sequence Variation. *PLoS One* **2011**, *6* (12), e28766. https://doi.org/10.1371/journal.pone.0028766.

(29) Moult, J.; Fidelis, K.; Kryshtafovych, A.; Schwede, T.; Tramontano, A. Critical Assessment of Methods of Protein Structure Prediction: Progress and New Directions in Round XI. *Proteins Struct. Funct. Bioinforma.* **2016**, *84*, 4–14. https://doi.org/10.1002/prot.25064.

(30) Wang, S.; Sun, S.; Li, Z.; Zhang, R.; Xu, J. Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. *PLOS Comput. Biol.* **2017**, *13* (1), e1005324. https://doi.org/10.1371/journal.pcbi.1005324.

(31) Senior, A. W.; Evans, R.; Jumper, J.; Kirkpatrick, J.; Sifre, L.; Green, T.; Qin, C.; Žídek, A.; Nelson, A. W. R.; Bridgland, A.; Penedones, H.; Petersen, S.; Simonyan, K.; Crossan, S.; Kohli, P.; Jones, D. T.; Silver, D.; Kavukcuoglu, K.; Hassabis, D. Improved Protein Structure Prediction Using Potentials from Deep Learning. *Nature* **2020**, *577* (7792), 706–710. https://doi.org/10.1038/s41586-019-1923-7.

(32) Lensink, M. F.; Brysbaert, G.; Nadzirin, N.; Velankar, S.; Chaleil, R. A. G.; Gerguri, T.; Bates, P. A.; Laine, E.; Carbone, A.; Grudinin, S.; Kong, R.; Liu, R.; Xu, X.; Shi, H.; Chang, S.; Eisenstein, M.; Karczynska, A.; Czaplewski, C.; Lubecka, E.; Lipska, A.; Krupa, P.; Mozolewska, M.; Golon, Ł.; Samsonov, S.; Liwo, A.; Crivelli, S.; Pagès, G.; Karasikov, M.; Kadukova, M.; Yan, Y.; Huang, S.; Rosell, M.; Rodríguez-Lumbreras, L. A.; Romero-Durana, M.; Díaz-Bueno, L.; Fernandez-Recio, J.; Christoffer, C.; Terashi, G.; Shin, W.; Aderinwale, T.; Maddhuri Venkata Subraman, S. R.; Kihara, D.; Kozakov, D.; Vajda, S.; Porter, K.; Padhorny, D.; Desta, I.; Beglov, D.; Ignatov, M.; Kotelnikov, S.; Moal, I. H.; Ritchie, D. W.; Chauvot de Beauchêne, I.; Maigret, B.; Devignes, M.; Ruiz Echartea, M. E.; Barradas-Bautista, D.; Cao, Z.; Cavallo, L.; Oliva, R.; Cao, Y.; Shen, Y.; Baek, M.; Park, T.; Woo, H.; Seok, C.; Braitbard, M.; Bitton, L.; Scheidman-Duhovny, D.; Dapkūnas, J.; Olechnovič, K.; Venclovas, Č.; Kundrotas, P. J.; Belkin, S.; Chakravarty, D.; Badal, V. D.; Vakser, I. A.; Vreven, T.; Vangaveti, S.; Borrman, T.; Weng, Z.; Guest, J. D.; Gowthaman, R.; Pierce, B. G.; Xu, X.; Duan, R.; Qiu, L.; Hou, J.; Ryan Merideth, B.; Ma, Z.; Cheng, J.;

Zou, X.; Koukos, P. I.; Roel-Touris, J.; Ambrosetti, F.; Geng, C.; Schaarschmidt, J.; Trellet, M. E.; Melquiond, A. S. J.; Xue, L.; Jiménez-García, B.; van Noort, C. W.; Honorato, R. V.; Bonvin, A. M. J. J.; Wodak, S. J. Blind Prediction of Homo- and Hetero-protein Complexes: The CASP13-CAPRI Experiment. *Proteins Struct. Funct. Bioinforma.* **2019**, *87* (12), 1200–1221. https://doi.org/10.1002/prot.25838.

(33) Baek, M.; DiMaio, F.; Anishchenko, I.; Dauparas, J.; Ovchinnikov, S.; Lee, G. R.; Wang, J.; Cong, Q.; Kinch, L. N.; Schaeffer, R. D.; Millán, C.; Park, H.; Adams, C.; Glassman, C. R.; DeGiovanni, A.; Pereira, J. H.; Rodrigues, A. V.; van Dijk, A. A.; Ebrecht, A. C.; Opperman, D. J.; Sagmeister, T.; Buhlheller, C.; Pavkov-Keller, T.; Rathinaswamy, M. K.; Dalwadi, U.; Yip, C. K.; Burke, J. E.; Garcia, K. C.; Grishin, N. V.; Adams, P. D.; Read, R. J.; Baker, D. Accurate Prediction of Protein Structures and Interactions Using a Three-Track Neural Network. *Science (80-.).* **2021**, *373* (6557), 871–876. https://doi.org/10.1126/science.abj8754.

(34) Lensink, M. F.; Brysbaert, G.; Mauri, T.; Nadzirin, N.; Velankar, S.; Chaleil, R. A. G. G.; Clarence, T.; Bates, P. A.; Kong, R.; Liu, B.; Yang, G.; Liu, M.; Shi, H.; Lu, X.; Chang, S.; Roy, R. S.; Quadir, F.; Liu, J.; Cheng, J.; Antoniak, A.; Czaplewski, C.; Giełdoń, A.; Kogut, M.; Lipska, A. G.; Liwo, A.; Lubecka, E. A.; Maszota-Zieleniak, M.; Sieradzan, A. K.; Ślusarz, R.; Wesołowski, P. A.; Zięba, K.; Del Carpio Muñoz, C. A.; Ichiishi, E.; Harmalkar, A.; Gray, J. J.; Bonvin, A. M. J. J. J. J.; Ambrosetti, F.; Vargas Honorato, R.; Jandova, Z.; Jiménez-García, B.; Koukos, P. I.; Van Keulen, S.; Van Noort, C. W.; Réau, M.; Roel-Touris, J.; Kotelnikov, S.; Padhorny, D.; Porter, K. A.; Alekseenko, A.; Ignatov, M.; Desta, I.; Ashizawa, R.; Sun, Z.; Ghani, U.; Hashemi, N.; Vajda, S.; Kozakov, D.; Rosell, M.; Rodríguez-Lumbreras, L. A.; Fernandez-Recio, J.; Karczynska, A.; Grudinin, S.; Yan, Y.; Li, H.; Lin, P.; Huang, S.; Christoffer, C.; Terashi, G.; Verburgt, J.; Sarkar, D.; Aderinwale, T.; Wang, X.; Kihara, D.; Nakamura, T.; Hanazono, Y.; Gowthaman, R.; Guest, J. D.; Yin, R.; Taherzadeh, G.; Pierce, B. G.; Barradas-Bautista, D.; Cao, Z.; Cavallo, L.; Oliva, R.; Sun, Y.; Zhu, S.; Shen, Y.; Park, T.; Woo, H.; Yang, J.; Kwon, S.; Won, J.; Seok, C.; Kiyota, Y.; Kobayashi, S.; Harada, Y.; Takeda-Shitaka, M.; Kundrotas, P. J.; Singh, A.; Vakser, I. A.; Dapkūnas, J.; Olechnovič, K.; Venclovas, Č.; Duan, R.; Qiu, L.; Xu, X.; Zhang, S.; Zou, X.; Wodak, S. J. Prediction of Protein Assemblies, the next Frontier: The CASP14-CAPRI Experiment. *Proteins Struct. Funct. Bioinforma.* **2021**, *89* (12), 1800–

1823. https://doi.org/10.1002/prot.26222.

(35) Hekkelman, M. L.; de Vries, I.; Joosten, R. P.; Perrakis, A. AlphaFill: Enriching AlphaFold Models with Ligands and Cofactors. *Nat. Methods* **2023**, *20* (2), 205–213. https://doi.org/10.1038/s41592-022-01685-y.

(36) Tsuchiya, Y.; Yamamori, Y.; Tomii, K. Protein–Protein Interaction Prediction Methods: From Docking-Based to AI-Based Approaches. *Biophys. Rev.* **2022**, *14* (6), 1341–1348. https://doi.org/10.1007/s12551-022-01032-7.

(37) Geng, C.; Jung, Y.; Renaud, N.; Honavar, V.; Bonvin, A. M. J. J.; Xue, L. C. IScore: A Novel Graph Kernel-Based Function for Scoring Protein–Protein Docking Models. *Bioinformatics* **2020**, *36* (1), 112–121. https://doi.org/10.1093/bioinformatics/btz496.

(38) 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. https://doi.org/10.1021/ja026939x.

(39) Heo, L.; Arbour, C. F.; Feig, M. Driven to Near-experimental Accuracy by Refinement via Molecular Dynamics Simulations. *Proteins Struct. Funct. Bioinforma.* **2019**, *87* (12), 1263–1275. https://doi.org/10.1002/prot.25759.

(40) Heo, L.; Janson, G.; Feig, M. Physics-based Protein Structure Refinement in the Era of Artificial Intelligence. *Proteins Struct. Funct. Bioinforma.* **2021**, *89* (12), 1870–1887. https://doi.org/10.1002/prot.26161.

(41) Gronenborn, A. M. Integrated Multidisciplinarity in the Natural Sciences. *J. Biol. Chem.* **2019**, *294* (48), 18162–18167. https://doi.org/10.1074/jbc.AW119.008142.

Section 2 – Methods

This section covers the main physics-based computational approaches used in the projects discussed in this PhD thesis. They are all based on a classical description of the systems, an approximation needed to deal with proteins, which consists of thousands of atoms. Ther size makes them not suitable for a complete, quantum mechanics-based description, which would have to account for electrons explicitly. However, such an approximation is suitable when the object of the study is not directly affected by electrons behaviour, as it happens for example in chemical reactions or processes where electron transfers occur. The study of protein-protein or protein-glycans interactions, on the other hand, can be safely performed with approaches based on classical physics. Non-covalent interactions between (bio)molecules are mainly driven by van der Waals and electrostatics contributions, which, although governed by electrons, can be described with simple potentials.

Molecular Mechanics (MM) and Molecular Dynamics (MD), used to study the dynamic behaviour of molecular systems at the atomic level, are first discussed (**Section 2.1** and **Section 2.2**, respectively).

Molecular docking, a method widely used for the prediction of biomolecular complexes, such as protein-protein or protein-glycan complexes, is then introduced (**Section 2.3**).

2.1 – Molecular Mechanics

The Born-Oppenheimer approximation states that the time-dependent wave function of a molecular system can be treated separately for electrons and nuclei. This is a consequence of the much smaller masses of electrons compared to the masses of nuclei, which causes them to move on different time scales, with electrons obviously moving faster. It can therefore be assumed that the electrons follow the motion of nuclei instantaneously. With the Born-Oppenheimer approximation, only the nuclear motion can be considered, while electronic degrees of freedom influence the dynamics of nuclei in the form of a potential energy surface (PES).

In the Molecular Mechanics (MM) approach, the PES is described by means of a set of functions, empirically derived, that present mathematical forms typical of classical mechanics. The MM method is a natural development of the concepts and formalisms of vibrational spectroscopy. A molecule is considered as a set of bond lengths, bond angles, and torsional angles. The energy of the molecule is associated with the geometric deformations described in terms of these coordinates plus the contribution of the van der Waals forces acting between non-bonded atoms. The basic idea of MM is that there are "natural" values of the geometric parameters. In the absence of steric interactions between atoms, i.e., in an "ideal" molecular system, each geometric parameter will assume its "natural" value. In real molecular systems, each atom interacts with all other atoms in the molecule; the values of the geometric parameters will then be deformed compared to their natural values: it is said that the molecule has a *strain*. The MM method assumes that it is possible to calculate the energy associated with these deformations.

The set of functions used to describe the PES and the empirical parameters that appear in the functions are called force fields (f.f.). The parameters used in the functions to calculate potential energy are adjustable parameters; they are optimized to reproduce a range of experimental or calculated molecular properties, such as geometries, conformational energies, heats of formation, vibrational frequencies, and so on.

The basic assumption of MM is that the parameters optimized for a certain molecular fragment, for example, the parameter related to the C-C bond distance, are transferable, for the same fragment, from one molecule to another; in other words, the parameters are retrieved from a reduced set of simple molecules and then used in the calculation of more complex molecular systems.

The parameter transferability assumption cannot be proved a priori, but it finds its validity in the results obtained, generally in good agreement with experimental evidence, thus justifying a posteriori the hypothesis that individual molecular fragments have similar properties in both simple and complex molecular systems. Furthermore, parameter transferability allows the f.f. to be extended from a set of already optimized parameters to include new molecular fragments or new classes of molecules. A single standard transferable value, the natural value, corresponds to a certain type of bond or bond angle and the equilibrium geometry is found relaxing the molecule to its minimum energy value. In MM, all possible internal coordinates plus those concerning non-bonded interactions are used. Each coordinate (bond distance, bond angle, torsional angle and non-bonded distance) tries to assume its natural value: the equilibrium geometry derives from the balance of the forces associated with each coordinate.

Molecules are viewed as “mechanical models” in which atoms are represented by Newtonian point particles linked together by springs (bonds), described by generalized Hooke’s law.

When a molecule consisting of n atoms and defined in terms of 3n x_i coordinates is deformed with respect to its reference geometry, $\{x°_i\}$, corresponding to a minimum of potential energy, V_o, its potential energy can be written as Taylor series expansion:

$$V_{\text{pot}} = \mathrm{V}_0 + \sum_i \left(\frac{\partial V}{\partial x_i}\right)_0 \Delta x_i + \frac{1}{2}\sum_{i,j}\left(\frac{\partial^2 V}{\partial x_i \partial x_j}\right)_0 \Delta x_i \Delta x_j + \ldots.$$

The V_0 term is a constant for a particular molecule and can be considered as a reference value, which is equivalent to setting it equal to zero. The first derivative of V, calculated for the equilibrium geometry, is zero by definition. Moreover, considering small displacements, terms of higher than second order can also be neglected (harmonic approximation). In first approximation, therefore, the potential energy will depend only on the third term of the expansion, that is, on the second derivative of V with respect to the coordinates $\{x_i\}$.

Substituting the expression of the second derivatives, which are the force constants, with the symbol f_{ij}, we obtain the relationship corresponding to a simple harmonic force field:

$$V_{\text{pot}} = \frac{1}{2}\sum_{i,j}^{3n} f_{ij} \Delta x_i \Delta x_j$$

This equation exactly defines, within the harmonic approximation, a system of coupled harmonic oscillators. The force constants are typically represented as a matrix in which the diagonal terms correspond to i = j. If all off-diagonal terms are zero, that is, if the set of oscillators is totally decoupled, the relation simplifies to the Hooke's law:

$$V_{\text{pot}} = \frac{1}{2}\sum_{i}^{3n} f_{ii} \Delta x_i^2$$

However, to obtain a better description of nuclear motions, and consequently high-quality equilibrium molecular geometries, it is necessary to add a number of mixed terms to the harmonic equations; the inclusion of mixed terms is a necessary condition for the development of transferable force fields.

In the MM method, the energy of the molecule is defined as the steric energy, E_s, given by the sum of M different potential energy functions, V, each dependent on the value of N geometric coordinates of *i* type (bond, angle, torsion, …), q_{ik}:

$$E_s = \sum_{i}^{M}\sum_{k}^{N_i} V_i\,(q_{ik}) = \sum_{i=1}^{bond} V_{stretching}\,(q_i) + \sum_{j=1}^{angle} V_{bending}(q_j) + \sum_{k=1}^{torsion} V_{torsional}(q_k) +$$

$$+ V_{van\ der\ Waals} + V_{electrostatics} + V_{others}$$

Within the harmonic approximation, a generic potential function V is expressed by the generalized Hooke's law:

$$V_i(q_{ik}) = \sum_{k}^{N} \frac{1}{2} K_{ik}(q_{ik} - q_{ik}^{0})^2$$

where K_{ik} are the force constants, q_{ik} is the value of the k-th geometric coordinate, and $q°_{ik}$ is the "natural" value that the q_{ik} coordinate would take if strains were absent.

The interactions generally considered in the potential energy function are:

- 1-2 interactions (bond lengths, stretching)

- 1-3 interactions (bond angles, bending)

- 1-4 interactions (dihedral angles between pairs of bonds, twist)

- Interactions between non-bonded atoms, or between atoms separated by more than two bonds. Generally, they include a van der Waals contribution and an electrostatic contribution.

The steric energy is given by:

$$E_s = V_{stretching} + V_{bending} + V_{torsional} + V_{non\text{-}bonded} + \text{other terms}$$

Between the "other terms", the polar term E_{pol} and the mixed terms (such as stretch-bending term, $E_{s\text{-}b}$), can be particularly relevant.

The following paragraphs show some frequently used potential functions.

Bonded potentials

Stretching and bending potential functions

For these two potential functions, the harmonic approximation is generally assumed to be valid, and the oscillators are considered to be independent:

$$V_{stretching}(r) = \frac{1}{2}K_s(r - r_0)^2$$

$$V_{bending}(\theta) = \frac{1}{2}K_b(\theta - \theta_0)^2$$

In these two expressions r and θ are the values assumed by a particular bond distance or bond angle, while r ° and θ ° are the corresponding natural values in the absence of strains.

The harmonic approximation may not be sufficient for r and θ values far from natural values. Therefore, higher order corrective terms with third- and sixth-degree functions were introduced for stretching and bending, respectively.

To reproduce the stretching that occurs in bonds in response to an angle deformation, the introduction of a mixed term (the stretching-bending potential) is required:

$$V_{stretching-bending} = \frac{1}{2}K_{sb}(r - r_0)(\theta - \theta_0)$$

Torsional potential functions

The potential function must be periodic: after a 360° rotation of the torsional angle, the potential must return to its initial value. In addition, torsional motions require rather small energies, when compared to stretching and bending energies. Thus, molecules may present significant torsional distortions from the minimum. For this reason, it is not a good practice to use a Taylor series expansion of the torsional potential, but it is preferable to use a Fourier series expansion. For the ABCD torsional angle, the torsional potential is given by:

$$V_{torsional}^{ABCD} = \sum_{n=1} k_n^{ABCD}\, cos(n\omega)$$

Where n = 1 describes a periodic rotation wit period 360°; n = 2 is periodic with period 180°; n = 3 is periodic with period 120°; and so on. The value of the k_n constant determines the height of the barrier: depending on the case, some k_n constant can be equal to zero.

Non-bonded potentials

The interaction energy between non-bonded atoms is calculated as the sum of two contributions: the van der Waals term and the electrostatic term.

van der Waals potential function

The general form of any non-bonded potential function is given by the sum of two contributions: a repulsive one, acting at short range; an attractive one, acting at long distance and tending asymptotically to zero as the distance r increases. The first repulsive contribution arises from the repulsive force that is established between the electronic distribution of two close enough atoms. This force is also known as exchange force, or overlap force, since it is established between electrons with the same spin. The second attractive contribution results from molecular interactions between instantaneous dipoles, even though the interacting fragments do not possess a permanent dipole moment. These forces, which in quantum mechanical treatment derive from electron correlation, are called dispersive forces or London forces.

A function frequently used to describe the van der Waals potential is the Lennard-Jones pair potential (also called 6-12 potential):

$$V_{Lennard-Jones} = \varepsilon\left[\left(\frac{r*}{r}\right)^{12} - 2\left(\frac{r*}{r}\right)^{6}\right]$$

where ε defines the depth of the potential well, r* the minimum energy distance (associated with the van der Waals radii of the interacting pairs of atoms), and the exponent 12 defines the hardness of the potential, that is, its steepness for distances below equilibrium.

The use of a pair potential means that the interaction between an atom of type A and an atom of type B will be described using the parameters ε_{AB} and $r*_{AB}$ obtained, by appropriate mixing rules, from the parameters related to the AA and BB interactions. The rules generally adopted are as follows:

$$r*_{AB} = r*_{AA} + r*_{BB}$$

$$\varepsilon_{AB} = \sqrt{\varepsilon_{AA}\varepsilon_{BB}}$$

Electrostatic potential function

The electrostatic contribution is generally calculated by the interaction between the net atomic charges, using Coulomb's law:

$$V_{electrostatics} = \sum_{i,j} \frac{q_i q_j}{\varepsilon r_{ij}}$$

where ε is the dielectric constant. The atomic charges of the atoms in a molecule cannot be uniquely defined, nor can they be derived from experimental measurements, since they are not physical observables of the system. In MM, the atomic charges can be treated as parameters: their values can be determined by quantum mechanical calculations.

The classical electrostatic contribution contains only pair contributions. However, in polar species the three bodies contribution are not negligible: these contributions can be modeled by including a polarization term in the electrostatic potential.

Force fields

As mentioned above, force fields are the combination of the set of functions used to describe the PES with the empirical parameters that appear in those functions. Depending on the size of the molecular system and the level of detail one wishes to achieve in its study, the molecular system can be represented with one ore more of the following classes of force fields:

1) All-atom force fields. As the term suggests, molecules are treated at the atomic level, i.e., all the atoms are considered explicitly.
2) United-atom force fields. In these f.f. certain groups of atoms (usually non-polar hydrogen atoms attached to a heavy atom) are treated as single entities. This simplification reduces the number of particles and interactions, making calculations faster.
3) Coarse grained force fields. These are even more simplified models where several atoms or even entire functional groups are represented by single interaction sites. They are used for the study of very large systems and/or long processes.

Besides this classification based on the level of detail used to describe the systems, force field can also be distinguished depending on the systems they are designed to describe, i.e., proteins, nucleic acids, lipids, small molecules, or oligosaccharides.

2.2 – Molecular Dynamics

Molecular Dynamics (MD) simulations exploit the MM principles to predict the time-dependent behaviour of a molecular system. In a MD simulation, Newton's laws of motion are integrated over time to obtain trajectories, which are essentially “paths” through the phase space of the system, described below.

The phase space

Classical mechanics allows a complete description of a system consisting of N particles with 3N spatial coordinates $\mathbf{r}_i$ ($\mathbf{r}^N$) and 3N momentum coordinates $\mathbf{p}_i$ ($\mathbf{p}^N$). The space defined by this (3N + 3N) set of variables is called *phase space*. The state of the system is defined by the values assumed by these 6N coordinates.

In a classical description of the system, its energy is defined by the sum of the kinetic energy $K(\mathbf{p}^N)$, which depends on the momentum of the N particles, and of the potential energy $V(\mathbf{r}^N)$, which depends on their positions.

$$E\ (r_1, ..., r_N, p_1, ..., p_N) = E\ (\mathbf{r}^N, \mathbf{p}^N) = K(\mathbf{p}^N) + V(\mathbf{r}^N)$$

The energy assumes a continuous spectrum of values, for each fixed set of coordinate values (r_1, ..., r_N, p_1, ..., p_N), i.e., for each point in the phase space.

The canonical partition function Q for a system of indistinguishable particles takes the form:

$$Q = k \int\int \exp[-\beta E\ (\mathbf{r}^N, \mathbf{p}^N)]\ d\mathbf{r}^N\ d\mathbf{p}^N$$

where k is a normalization factor equal to $1/(N!\ h^{3N})$. The double integral is shown for convenience. In fact, there should be 6N signs of integration since the integration must be done with respect to 3N position and 3N momentum variables.

In the dominant configuration, the probability $P^*(\mathbf{r}^N, \mathbf{p}^N)$ of a state within the phase space is given by the relation:

$$P^*(\mathbf{r}^N, \mathbf{p}^N) = \exp[-\beta E\,(\mathbf{r}^N, \mathbf{p}^N)]\ /Q$$

The average value of a property A, denoted by <A>, is given by the average of the values the property takes in the various states of the system (i.e., in the phase space) weighted by the probability of the states in the dominant configuration:

$$<A> = \int\int A(\mathbf{r}^N, \mathbf{p}^N)\, P^*(\mathbf{r}^N, \mathbf{p}^N)\, d\mathbf{r}^N\, d\mathbf{p}^N =$$

$$= [\int\int A(\mathbf{r}^N, \mathbf{p}^N) \exp[-\beta E\,(\mathbf{r}^N, \mathbf{p}^N)]\, d\mathbf{r}^N\, d\mathbf{p}^N\ ]\ /\ [\int\int \exp[-\beta E\,(\mathbf{r}^N, \mathbf{p}^N)]\, d\mathbf{r}^N\, d\mathbf{p}^N]$$

The problem is to calculate the values of these integrals. In general, we should calculate the value of the energy of the system at each point (state) in the phase space. This procedure is not feasible, because the number of variables is too high.

Moreover, it is not an efficient procedure in the sense that the calculation of the integral at the denominator (corresponding to Q) involves the generation of a large number of states (in principle, all). Therefore, even all the high energy states, having consequently low probability and low weight in the calculation of <A>, would be calculated. In other words, all states would have equal weight, when in reality they don’t.

Molecular simulations methods, among which MD, were thus introduced for the generation of a set of states representative of the phase space accessible to the system, which are then exploited for the calculation of its energetic, structural, thermodynamic and dynamics properties.

General concepts of Molecular Dynamics

In the MD approach, the value of a desired property A is obtained as the average value that property A assumes in a single system that evolves for an ideally infinite time.

Suppose we want to determine the value of a macroscopic property of the system. Its value will depend on the position ($\mathbf{r}^N$) and momentum ($\mathbf{p}^N$) of the particles that constitute the system. The instantaneous value of the property will thus depend on the instantaneous values of these variables at time t, that is: $A(t) = A[\mathbf{r}^N(t), \mathbf{p}^N(t)]$. Over time, the instantaneous value of the property will undergo fluctuations due to the interactions between the particles. The experimentally measured value will be the time average of instantaneous values, denoted as $\{A\}_t$. As the "observation time" τ of the system increases, the value of the time average will approach the true value of the property A_{real}, becoming equal to A_{real} for an infinite observation time. The time average is given by the relation:

$$A_{real} = \{A\}_t = \lim_{\tau\to\infty} \frac{1}{\tau}\int_{t=0}^{t=\tau} A[\boldsymbol{r}^N(t), \boldsymbol{p}^N(t)]dt$$

In the *ergodic hypothesis*, time average and ensemble average coincide:

$$A_{real} = \lim_{\tau\to\infty} \frac{1}{\tau}\int_{t=0}^{t=\tau} A[\boldsymbol{r}^N(t), \boldsymbol{p}^N(t)]dt = \lim_{M\to\infty} \frac{1}{M}\sum_{i=1}^{M} A[\boldsymbol{r}_i^N(t), \boldsymbol{p}_i^N(t)]$$

and this holds true for any choice of initial conditions values.

To calculate the instantaneous values of $A[\mathbf{r}^N(t), \mathbf{p}^N(t)]$, it is necessary to simulate the dynamic behaviour of the system. For this purpose, the following operations must be performed:

1. Choice of a proper force field for the description of intra- and inter-molecular interactions of the system.
2. Calculation, using the selected force field, of the potential energy value $V(\mathbf{r}^N)$ for a given initial disposition of molecules in the space.

3. Calculation of the forces acting on each atom of the system by differentiating the potential energy expression $F_i = - [\partial V(\mathbf{r}^N)/\partial r_i]$
4. Calculation of the accelerations, once the forces acting on each atom are known, using Newton's equation of motion $F_i = ma_i$
5. Integration of the equation of motion for each particle to determine changes in position, velocity, and acceleration as a function of time. Integration of the equations is done by considering M time intervals, Δt, small enough to assume the acceleration acting on the particles in the interval as constant. In this way, the trajectory is calculated using the equation of uniformly accelerated motion.
6. Calculation of the property A of the system as time average of the values that A assumes in the considered M time intervals. As $t_{obs} = M\,\Delta t$, the equation

$$A_{real} = \{A\}_t = \lim_{\tau\to\infty} \frac{1}{\tau}\int_{t=0}^{t=\tau} A[\boldsymbol{r}^N(t), \boldsymbol{p}^N(t)]dt$$

becomes

$$\{A\}_t = \left(\frac{1}{M\Delta t}\right)\sum_{i=1}^{M} A_i\ (\boldsymbol{r}^N, \boldsymbol{p}^N)\Delta t = \left(\frac{1}{M}\right)\sum_{i=1}^{M} A_i\ (\boldsymbol{r}^N, \boldsymbol{p}^N)$$

The choice of the time interval Δt depends on the system under investigation. Generally, a time interval of one (or two) orders of magnitude less than the frequency of the fastest motion within the system should be chosen. If one wants to increase the Δt values to perform faster computations, the degrees of freedom with higher frequencies can be removed from the system by means of suitable algorithms, such as, for example, the SHAKE or the LINCS algorithms.

Setting up a Molecular Dynamics simulation

Boundary Conditions

During MD simulations, it is necessary to ensure that the molecules at the boundaries of the system are not affected by the so called "wall effect". For instance, it is necessary to prevent water molecules forming the solvation shell of a protein from "evaporating". To avoid these effects, the molecules forming the system are inserted in boxes surrounded by a periodic series of exact copies of the original box (Periodic Boundary Condition, PBC). In this way, the molecules at the outer boundary of the original box interact with the molecules in the adjacent box. Moreover, to prevent a molecule from "seeing" itself in a copy-box, the Minimum Image Convention applies: each atom sees no more than one image of every other atom of the system. To speed up the calculation of non-bonded interactions, a threshold value is introduced for their calculation. The interaction energy between two atoms is calculated if they are at a distance less than or equal to the threshold value. The threshold must be less than half the length of the box, so that a particle does not see itself or the same molecule twice.

Initial configuration.

In general, it is best to start a MD simulation from an initial configuration that is as close as possible to the equilibrium situation to be described. For this purpose, the potential energy of the system is initially minimized, so that no excessively "distorted" regions are present.

Initial velocities

To conduct the simulation, the particles must be assigned initial velocities compatible with the desired temperature. These are randomly assigned to the atoms, with the constraint that the total linear and angular momentum of the system be zero, or, in other words, that the system does not translate or rotate as a whole.

Thermodynamic ensemble

Performing a MD simulation requires the choice of a thermodynamic ensemble among the canonical (N, V, T), microcanonical (N, V, E), isothermal-isobaric (N, P, T), and grand canonical ensemble (μ, V, T). Depending on the selected ensemble, the physical quantities that must remain constant must be checked during the simulation. To this end, algorithms acting as computational "thermostats" and "barostats" allow for the appropriate scaling of temperature and pressure, respectively, in order to guarantee the predetermined conditions.

Equilibration and production phases

Starting from the initial conditions, the system evolves over time and reaches equilibrium. If the simulation is performed at constant T, velocities are scaled until the desired T is reached. During the equilibration, the values of various properties (E, K, V, T, P) that characterise the selected simulation ensemble are monitored. When these quantities assume constant values the production phase of the dynamics begins: data collected in this phase are then used to calculate the properties of interest. If the simulation sampled all points in the phase space (ergodic trajectory), the results would be independent of the initial configuration. However, due to the extremely large size of the phase space, it is not possible to obtain an ergodic trajectory from a single simulation. For this reason, simulations are usually repeated starting from different initial configurations. In order to improve phase space sampling, enhanced/biased MD approaches can also be applied.

2.3 – Molecular Docking

General concepts and workflow

Molecular docking is a widely used computational technique for the prediction of the three-dimensional structures of biomolecular assemblies, such as protein-protein and protein-ligands complexes. It consists of searching for the geometry of the complex, starting from the unbound forms of the two (or more) partners, by generating possible solutions (poses) and ranking the resulting poses using appropriate functions.

Docking was first introduced more than 40 years ago[1] and has since then impressively progressed, thanks to development of efficient algorithms and the availability of increasingly powerful computing resources. A variety of docking approaches are available at the present date[2]. All of them are still characterized by a nearly common workflow foreseeing the following steps[3].

1) Preparation of the three-dimensional input structures
2) Generation of the poses
3) Scoring
4) Refinement

An overview of this stages is given in the following lines.

Preparation of the three-dimensional input structures

The structures can be experimentally determined or, if not available, they can be predicted too, although this of course affects the reliability of the docking result. A choice also needs to be done on the representation of the unbound structures, that can be described with all-atom, united-atom, coarse-grained force fields, or even with a residue-based description, useful for dealing with particularly large systems. At this stage, it is important to consider whether the binding could result in a significant change in the conformation of the receptor (and/or ligand); if this could be the case, an ensemble of conformations could be used as input.

Generation of the poses

This step, also found in literature as “searching” or “sampling”, aims at producing the possible dispositions of a partner with respect to the other(s). Almost all approaches, for reasons of computational efficiency, keep the larger partner, called the 'receptor', fixed, while the smaller one, the 'ligand', is rotated and translated.

Sampling methods can be exhaustive/systematic or stochastic[4]. Systematic approaches will be reviewed in the following lines.

The development of the Katchalski-Katzir algorithm[5], based on the fast Fourier transform (FFT), really accelerated the computationally onerous systematic search stage. In FFT-based algorithms, the structures of the proteins are first represented on a 3D-cartesian grid, where discrete functions distinguishing between the surface and the interior of the proteins are used. The matching of the surfaces is then evaluated with correlation functions that assess the degree of overlap between the partners for every shift of the ligand with respect to the receptor. Some degree of penetration between the proteins is allowed to take into account small conformational changes. The calculation of correlation functions is performed with the FFT. Then, the angles defining the orientation of the ligand are varied at defined intervals; the correlation function is calculated again for all the relative orientations of the partners. The advantage of this method is that the use of correlation functions, calculated with the FFT, allows to evaluate all possible reciprocal dispositions of the two partners in a more rapid way then previous methods for exhaustive search in six dimensions.

In advanced FFT-based methods, terms representing electrostatic, hydrophobic, and solvation contributions are included too.

PIPER[6], an FFT-based algorithm implemented in the docking programme and web server ClusPro[7,8,9], includes pairwise structure-based interaction potentials with the aim of improving the systematic search.

Besides FFT-based methods, the generation of the poses can be performed in other ways[3], such as via geometric hashing docking or spherical harmonic-based docking, where the spherical polar Fourier correlations are used to accelerate the search.

HADDOCK[10] uses instead a different approach for the generation of the poses. A randomization stage is first performed, where the two partner proteins are positioned at 25 Å from each other in space and are randomly rotated around their centre of mass. A rigid body minimization with the OPLS force field[11] is then carried out in multiple steps that foresee: i) four cycles of rotational orientation in which each partner is allowed to rotate in order to minimise the intermolecular energy; ii) two cycles of rotational and translational rigid body minimization in which each partner is treated as a rigid body.

Scoring

The docking poses generated at the previous steep need then to be analysed in such a way to identify, among a large pool of models, the near-native ones, i.e., the ones probably closer to the true structure of the complex. This analysis is performed on the basis of the docking score of the pose, calculated by a suitable scoring function.

Scoring functions are historically mainly divided in energy-based and knowledge-based.

Energy-based scoring functions contain several properly weighted terms that account for the van der Waals and electrostatics contributions of the interaction, for the desolvation energy and for other empirical parameters such as the buried surface area (BSA), and the shape complementarity. Energy-based scoring functions are used in both ClusPro and HADDOCK. They are shown in **Section 3.2** and **Section 4**, respectively.

Knowledge-based scoring functions, as the term suggests, use information derived from experimentally known protein-protein complexes. This information is "converted" in potentials that derive from the statistical occurrences observed in the known complexes, by means of an inverse Boltzmann equation.

However, scoring functions can combine the different approaches at the same time. They can also exploit machine learning techniques for identifying the set of coefficients that leads to a better discrimination among the docking models.

Ideally, a scoring function should perfectly correlate with model closeness to the experimental structure. However, although scoring functions are improving, as the CAPRI rounds (see below) have shown over the years, this is still not the case.

It is also important to note that native models are not isolated in the global energy landscape; they instead are expected to form "funnels", i.e., groups of docking solutions characterised by similar, low energy. The pool of the obtained models can thus be clustered.

Clustering of the models

Clustering, for instance, can be based on the root-mean-square deviations (RMSD) between the models, or using the fraction of common contacts (FCC)[12]; these two approaches are both implemented in HADDOCK[10]. In HADDOCK3[13], they can be introduced at the desired stage of the workflow, e.g. both after the rigid body stage, for the selection of a subset of models to be refined, and after the (semi-flexible) refinement stage (see **Section 4.1**).

In ClusPro[8], the 1000 lowest-energy rigid body models are clustered based on the interface-RMSD (I-RMSD, see below). I-RMSD values are calculated among all the structures, and the one having the highest number of neighbours within a 9 Å cut off is selected as the centre of the first cluster. The structures within the given cutoff from the centre constitute the first clusters and are thus removed from the initial pool. The process is then repeated until a maximum of 30 clusters are produced, which are then ranked based on their populations. This procedure is applied in **Section 3.2**.

Refinement

A force-field based refinement of the best scoring models can be included too. For example, ClusPro performs an energy minimization of the clustered models, where backbones are kept fixed and only the van der Waals term of the CHARMM potential[14] is used.

HADDOCK uses instead the OPLS force field[11]. Besides energy minimization, a flexible refinement stage via MD simulations can be performed too, which is of essential importance when partners characterized by a non-negligible conformational variability are involved (see **Section 4.1**).

Post-docking procedures

As already mentioned, the scoring functions do not allow a single correct docking model to be identified with absolute certainty. Instead, similar scores are often obtained for several models.

From this arises the need to use "post-docking" procedures aimed at the reranking of the (best scoring) docking models. Ideally, such procedures should be based on different assumptions than those already included in the scoring phase of the docking programme used, thus ensuring greater reliability. Several web servers[3] are available for this purpose. The algorithms they use can be energy-, knowledge-, evolution-, or consensus-based, like in CONSRANK[15].

The evaluation and reranking of the docking models can also be carried out by performing MD simulations. An example is the approach proposed by Jandova et al.[16], which is based on the idea that the mutual positions of the partners in near-native models should not change significantly during a MD simulation, i.e., the predicted complex should show a certain degree of stability. On the other hand, non-native poses should change along the MD trajectories, possibly leading to the (partial) dissociation of the complex.

Finally, the rescoring of the docking models can also be carried out by comparison with the interface residues[17], which can be predicted on the basis of evolutionary, geometric, physico-chemical and interface propensity features.

One of these methods is the Matrix of Local Coupling Energies method (MLCE)[18], which combines both energetic and structural considerations for the prediction of the interface residues; it is covered in detail in **Section 3.2.1**.

Data-driven docking

Docking is a complex issue. Therefore, the inclusion of information - preferably experimentally derived - about the interface area of the partners helps to find more reliable models.

HADDOCK, starting from the sampling stage, can incorporate information derived from NMR chemical shift perturbation data, mutagenesis data or any kind of data providing information on the interface, in the form of Ambiguous Interaction Restraints (AIRs), which are used as an additional restraint energy term. A more detailed description is given in **Section 4.1**.

In ClusPro the docking process can be guided by assigning attraction or repulsion to residues that are thought to be, or not to be, part of the interface. An attractive force during the docking process is applied to the binding-involved residues of one or both sides of the interface. On the contrary, repulsive forces are applied to the residues that are expected not to be at the interface. Such an approach is used in **Section 3.2**.

CAPRI evaluation

In 2001, the European Bioinformatics Institute, part of the European Molecular Biology Laboratory (EMBL-EBI) started the Critical Assessment of PRediction of Interactions (CAPRI), a community wide experiment aimed at the monitoring the progresses in protein-protein docking approaches. There have now been 54 CAPRI rounds, where the participants have been asked to predict the structure of complexes, whose experimental structures are not released, of increasing difficulty.

A joint initiative with the Critical Assessment of Structure Prediction (CASP) community started in 2014, with the aim of addressing the problem of predicting single protein structures and protein-protein interactions at the same time, starting uniquely from the amino acid sequences.

The first paper of the joint initiative was published in 2016[19], where CASP 11 season was carried out along with CAPRI round 30. Since then, there have been four more joint experiments: CASP12-CAPRI[20] (2018), CASP13-CAPRI[21] (2019), CASP14-CAPRI[22] (2021), and CASP15-CAPRI (2022).

The ability of the predictors in producing good quality models, also for the most difficult targets, has of course improved during the years. A comparison between CASP13-CAPRI[21] and CASP14-CAPRI[22] rounds is shown in *Figure 2.1*.

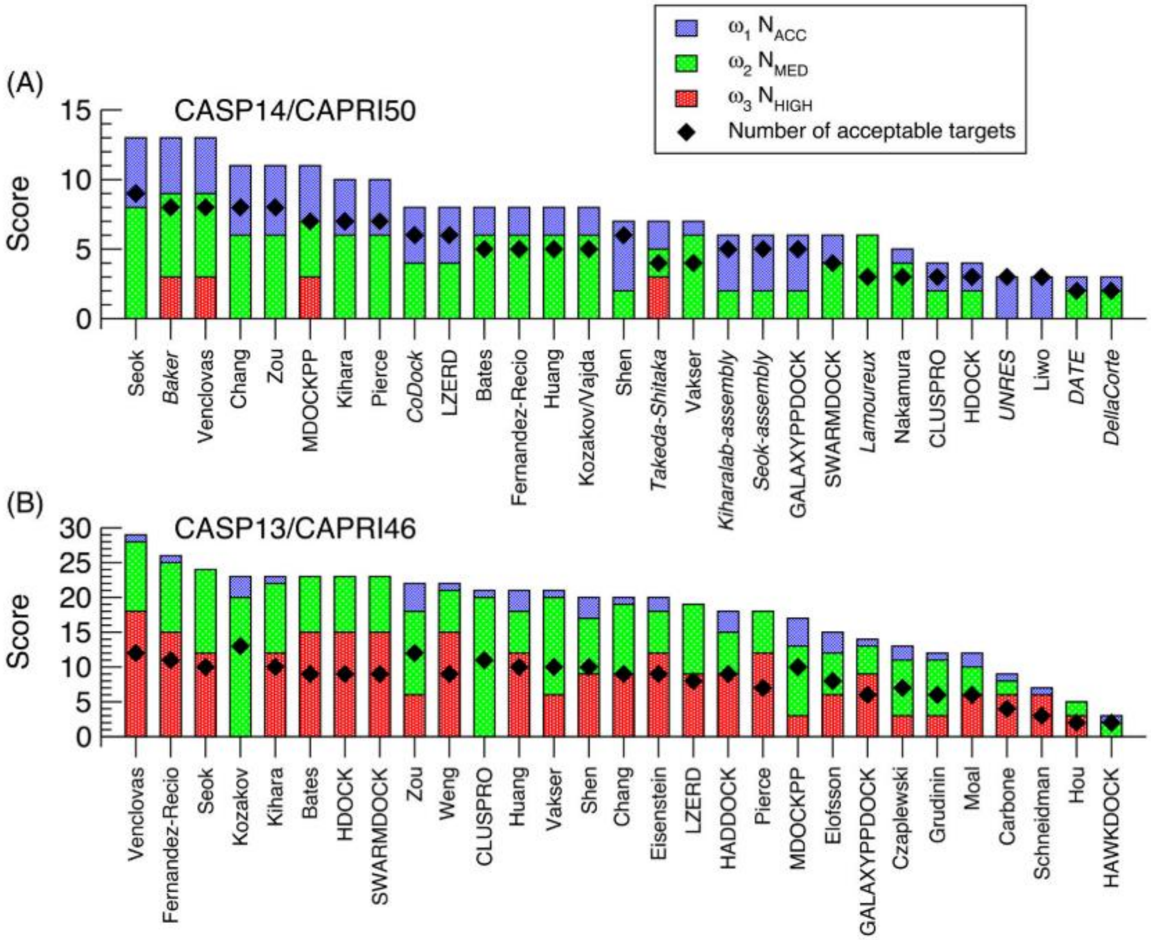


Figure 2.1 – Figure reproduced from reference[22]. Panel (A) shows the performance score of the top 29 ranking predictor and server groups (both CAPRI and CASP-only groups; server groups are listed in capital letters). The height of the bar is the Score$_G$, calculated as a weighted sum of the number of targets of high-, medium., or acceptable-quality models. The total number of targets for which at least an acceptable quality model was produced is indicated in the graph by a diamond. Panel (B) shows the same data from the previous CASP13-CAPRI Round.

The criteria for the assessment of the quality of the structures generated by different docking tools are still mainly the original CAPRI evaluation criteria[23,24]: ligand-RMSD (L-RMSD), interface-RMSD (I-RMSD), and the fraction of native contacts (Fnat). Following the definition given in reference[24], they are calculated as follows.

L-RMSD. The RMSD of the ligand (the smaller of the two partners) in the predicted versus target complexes after superposition of the receptors. Calculation of the RMSD and superpositions are both computed on backbone atoms (N, C_α, C, O).

I-RMSD. The RMSD of backbone atoms is calculated on the interface residues only, defined as those having at least one atom within 10 Å of an atom on the other molecule.

Fnat. The number of native (correct) residue–residue contacts in the predicted complex divided by the number of contacts in the target complex. A pair of residues on different sides of the interface are considered to be in contact if any of their atoms are within 5 Å.

Starting from CASP13-CAPRI[21], the continuous parameter DockQ[25] has been routinely used too for an overall evaluation of the predictions. This parameter encompasses L-RMSD, I-RMSD, and Fnat, thus providing a single measure of the docking performance.

DockQ is calculated as follows, yielding to a score in the range [0,1].

$$\text{DockQ} = (\text{Fnat} + \text{L-RMSD}_{\text{scaled}}\ (\text{L-RMSD}, d_i) + \text{I-RMSD}_{\text{scaled}}\ (\text{L-RMSD}, d_i))\ /\ 3$$

L-RMSD and I-RMSD are scaled as in the following equation:

$$\text{RMSD}_{\text{scaled}} = 1\ /\ [1 + (\text{RMSD}\ /\ d_i\ )^2]$$

Where d_i = 8.5 Å for L-RMSD and d_i = 1.5 Å for I-RMSD. d_i values were optimized for obtaining DockQ values in the range [0,1] and also to ensure that RMSD values that should be considered equally bad, e.g., I-RMSD of 7 Å or 14 Å both obtain the same low I-$\text{RMSD}_{\text{scaled}}$ score.

References

(1) Wodak, S. J.; Janin, J. Computer Analysis of Protein-Protein Interaction. *J. Mol. Biol.* **1978**, *124* (2), 323–342. https://doi.org/10.1016/0022-2836(78)90302-9.

(2) Pagadala, N. S.; Syed, K.; Tuszynski, J. Software for Molecular Docking: A Review. *Biophys. Rev.* **2017**, *9* (2), 91–102. https://doi.org/10.1007/s12551-016-0247-1.

(3) Vangone, A.; Oliva, R.; Cavallo, L.; Bonvin, A. M. J. J. Prediction of Biomolecular Complexes. In *From Protein Structure to Function with Bioinformatics*; Springer Netherlands: Dordrecht, 2017; pp 265–292. https://doi.org/10.1007/978-94-024-1069-3_8.

(4) Sunny, S.; Jayaraj, P. B. Protein–Protein Docking: Past, Present, and Future. *Protein J.* **2022**, *41* (1), 1–26. https://doi.org/10.1007/s10930-021-10031-8.

(5) Katchalski-Katzir, E.; Shariv, I.; Eisenstein, M.; Friesem, A. A.; Aflalo, C.; Vakser, I. A. Molecular Surface Recognition: Determination of Geometric Fit between Proteins and Their Ligands by Correlation Techniques. *Proc. Natl. Acad. Sci.* **1992**, *89* (6), 2195–2199. https://doi.org/10.1073/pnas.89.6.2195.

(6) Kozakov, D.; Brenke, R.; Comeau, S. R.; Vajda, S. PIPER: An FFT-Based Protein Docking Program with Pairwise Potentials. *Proteins Struct. Funct. Bioinforma.* **2006**, *65* (2), 392–406. https://doi.org/10.1002/prot.21117.

(7) Comeau, S. R.; Gatchell, D. W.; Vajda, S.; Camacho, C. J. ClusPro: An Automated Docking and Discrimination Method for the Prediction of Protein Complexes. *Bioinformatics* **2004**, *20* (1), 45–50. https://doi.org/10.1093/bioinformatics/btg371.

(8) Kozakov, D.; Hall, D. R.; Xia, B.; Porter, K. A.; Padhorny, D.; Yueh, C.; Beglov, D.; Vajda, S. The ClusPro Web Server for Protein–Protein Docking. *Nat. Protoc.* **2017**, *12* (2), 255–278. https://doi.org/10.1038/nprot.2016.169.

(9) Vajda, S.; Yueh, C.; Beglov, D.; Bohnuud, T.; Mottarella, S. E.; Xia, B.; Hall, D. R.; Kozakov, D. New Additions to the ClusPro Server Motivated by CAPRI. *Proteins Struct. Funct. Bioinforma.* **2017**, *85* (3), 435–444. https://doi.org/10.1002/prot.25219.

(10) 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. https://doi.org/10.1021/ja026939x.

(11) 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. https://doi.org/10.1021/ja00214a001.

(12) Rodrigues, J. P. G. L. M.; Trellet, M.; Schmitz, C.; Kastritis, P.; Karaca, E.; Melquiond, A. S. J.; Bonvin, A. M. J. J. Clustering Biomolecular Complexes by Residue Contacts Similarity. *Proteins Struct. Funct. Bioinforma.* **2012**, *80* (7), 1810–1817. https://doi.org/10.1002/prot.24078.

(13) Bonvin's Lab. HADDOCK3. 2022.

(14) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *J. Comput. Chem.* **1983**, *4* (2), 187–217. https://doi.org/10.1002/jcc.540040211.

(15) Chermak, E.; Petta, A.; Serra, L.; Vangone, A.; Scarano, V.; Cavallo, L.; Oliva, R. CONSRANK: A Server for the Analysis, Comparison and Ranking of Docking Models Based on Inter-Residue Contacts. *Bioinformatics* **2015**, *31* (9), 1481–1483. https://doi.org/10.1093/bioinformatics/btu837.

(16) Jandova, Z.; Vargiu, A. V.; Bonvin, A. M. J. J. Native or Non-Native Protein–Protein Docking Models? Molecular Dynamics to the Rescue. *J. Chem. Theory Comput.* **2021**, *17* (9), 5944–5954. https://doi.org/10.1021/acs.jctc.1c00336.

(17) Pozzati, G.; Kundrotas, P.; Elofsson, A. Scoring of Protein–Protein Docking Models Utilizing Predicted Interface Residues. *Proteins Struct. Funct. Bioinforma.* **2022**, *90* (7), 1493–1505. https://doi.org/10.1002/prot.26330.

(18) Scarabelli, G.; Morra, G.; Colombo, G. Predicting Interaction Sites from the Energetics of Isolated Proteins: A New Approach to Epitope Mapping. *Biophys. J.* **2010**, *98* (9), 1966–1975. https://doi.org/10.1016/j.bpj.2010.01.014.

(19) Lensink, M. F.; Velankar, S.; Kryshtafovych, A.; Huang, S.; Schneidman-Duhovny, D.; Sali, A.; Segura, J.; Fernandez-Fuentes, N.; Viswanath, S.; Elber, R.; Grudinin, S.; Popov, P.; Neveu, E.; Lee, H.; Baek, M.; Park, S.; Heo, L.; Rie Lee, G.; Seok, C.; Qin, S.; Zhou, H.; Ritchie, D. W.; Maigret, B.; Devignes, M.; Ghoorah, A.; Torchala, M.; Chaleil, R. A. G.; Bates, P. A.; Ben-Zeev, E.; Eisenstein, M.; Negi, S. S.; Weng, Z.; Vreven, T.; Pierce, B. G.; Borrman, T. M.; Yu, J.; Ochsenbein, F.; Guerois, R.; Vangone, A.; Rodrigues, J. P. G. L. M.; van Zundert, G.; Nellen, M.; Xue, L.; Karaca, E.; Melquiond, A. S. J.; Visscher,

K.; Kastritis, P. L.; Bonvin, A. M. J. J.; Xu, X.; Qiu, L.; Yan, C.; Li, J.; Ma, Z.; Cheng, J.; Zou, X.; Shen, Y.; Peterson, L. X.; Kim, H.; Roy, A.; Han, X.; Esquivel-Rodriguez, J.; Kihara, D.; Yu, X.; Bruce, N. J.; Fuller, J. C.; Wade, R. C.; Anishchenko, I.; Kundrotas, P. J.; Vakser, I. A.; Imai, K.; Yamada, K.; Oda, T.; Nakamura, T.; Tomii, K.; Pallara, C.; Romero-Durana, M.; Jiménez-García, B.; Moal, I. H.; Férnandez-Recio, J.; Joung, J. Y.; Kim, J. Y.; Joo, K.; Lee, J.; Kozakov, D.; Vajda, S.; Mottarella, S.; Hall, D. R.; Beglov, D.; Mamonov, A.; Xia, B.; Bohnuud, T.; Del Carpio, C. A.; Ichiishi, E.; Marze, N.; Kuroda, D.; Roy Burman, S. S.; Gray, J. J.; Chermak, E.; Cavallo, L.; Oliva, R.; Tovchigrechko, A.; Wodak, S. J. Prediction of Homoprotein and Heteroprotein Complexes by Protein Docking and Template-based Modeling: A CASP-CAPRI Experiment. *Proteins Struct. Funct. Bioinforma.* **2016**, *84* (S1), 323–348. https://doi.org/10.1002/prot.25007.

(20) Lensink, M. F.; Velankar, S.; Baek, M.; Heo, L.; Seok, C.; Wodak, S. J. The Challenge of Modeling Protein Assemblies: The CASP12-CAPRI Experiment. *Proteins Struct. Funct. Bioinforma.* **2018**, *86* (S1), 257–273. https://doi.org/10.1002/prot.25419.

(21) Lensink, M. F.; Brysbaert, G.; Nadzirin, N.; Velankar, S.; Chaleil, R. A. G.; Gerguri, T.; Bates, P. A.; Laine, E.; Carbone, A.; Grudinin, S.; Kong, R.; Liu, R. R.; Xu, X. M.; Shi, H.; Chang, S.; Eisenstein, M.; Karczynska, A.; Czaplewski, C.; Lubecka, E.; Lipska, A.; Krupa, P.; Mozolewska, M.; Golon, Ł.; Samsonov, S.; Liwo, A.; Crivelli, S.; Pagès, G.; Karasikov, M.; Kadukova, M.; Yan, Y.; Huang, S. Y.; Rosell, M.; Rodríguez-Lumbreras, L. A.; Romero-Durana, M.; Díaz-Bueno, L.; Fernandez-Recio, J.; Christoffer, C.; Terashi, G.; Shin, W. H.; Aderinwale, T.; Maddhuri Venkata Subraman, S. R.; Kihara, D.; Kozakov, D.; Vajda, S.; Porter, K.; Padhorny, D.; Desta, I.; Beglov, D.; Ignatov, M.; Kotelnikov, S.; Moal, I. H.; Ritchie, D. W.; Chauvot de Beauchêne, I.; Maigret, B.; Devignes, M. D.; Ruiz Echartea, M. E.; Barradas-Bautista, D.; Cao, Z.; Cavallo, L.; Oliva, R.; Cao, Y.; Shen, Y.; Baek, M.; Park, T.; Woo, H.; Seok, C.; Braitbard, M.; Bitton, L.; Scheidman-Duhovny, D.; Dapkūnas, J.; Olechnovič, K.; Venclovas, Č.; Kundrotas, P. J.; Belkin, S.; Chakravarty, D.; Badal, V. D.; Vakser, I. A.; Vreven, T.; Vangaveti, S.; Borrman, T.; Weng, Z.; Guest, J. D.; Gowthaman, R.; Pierce, B. G.; Xu, X.; Duan, R.; Qiu, L.; Hou, J.; Ryan Merideth, B.; Ma, Z.; Cheng, J.; Zou, X.; Koukos, P. I.; Roel-Touris, J.; Ambrosetti, F.; Geng, C.; Schaarschmidt, J.; Trellet, M. E.; Melquiond, A. S. J.; Xue, L.; Jiménez-García, B.; van Noort, C. W.; Honorato, R. V.; Bonvin, A. M. J. J.; Wodak, S. J. Blind Prediction of Homo-

and Hetero-Protein Complexes: The CASP13-CAPRI Experiment. *Proteins Struct. Funct. Bioinforma.* **2019**, *87* (12), 1200–1221. https://doi.org/10.1002/prot.25838.

(22) Lensink, M. F.; Brysbaert, G.; Mauri, T.; Nadzirin, N.; Velankar, S.; Chaleil, R. A. G. G.; Clarence, T.; Bates, P. A.; Kong, R.; Liu, B.; Yang, G.; Liu, M.; Shi, H.; Lu, X.; Chang, S.; Roy, R. S.; Quadir, F.; Liu, J.; Cheng, J.; Antoniak, A.; Czaplewski, C.; Giełdoń, A.; Kogut, M.; Lipska, A. G.; Liwo, A.; Lubecka, E. A.; Maszota-Zieleniak, M.; Sieradzan, A. K.; Ślusarz, R.; Wesołowski, P. A.; Zięba, K.; Del Carpio Muñoz, C. A.; Ichiishi, E.; Harmalkar, A.; Gray, J. J.; Bonvin, A. M. J. J. J. J.; Ambrosetti, F.; Vargas Honorato, R.; Jandova, Z.; Jiménez-García, B.; Koukos, P. I.; Van Keulen, S.; Van Noort, C. W.; Réau, M.; Roel-Touris, J.; Kotelnikov, S.; Padhorny, D.; Porter, K. A.; Alekseenko, A.; Ignatov, M.; Desta, I.; Ashizawa, R.; Sun, Z.; Ghani, U.; Hashemi, N.; Vajda, S.; Kozakov, D.; Rosell, M.; Rodríguez-Lumbreras, L. A.; Fernandez-Recio, J.; Karczynska, A.; Grudinin, S.; Yan, Y.; Li, H.; Lin, P.; Huang, S.; Christoffer, C.; Terashi, G.; Verburgt, J.; Sarkar, D.; Aderinwale, T.; Wang, X.; Kihara, D.; Nakamura, T.; Hanazono, Y.; Gowthaman, R.; Guest, J. D.; Yin, R.; Taherzadeh, G.; Pierce, B. G.; Barradas-Bautista, D.; Cao, Z.; Cavallo, L.; Oliva, R.; Sun, Y.; Zhu, S.; Shen, Y.; Park, T.; Woo, H.; Yang, J.; Kwon, S.; Won, J.; Seok, C.; Kiyota, Y.; Kobayashi, S.; Harada, Y.; Takeda-Shitaka, M.; Kundrotas, P. J.; Singh, A.; Vakser, I. A.; Dapkūnas, J.; Olechnovič, K.; Venclovas, Č.; Duan, R.; Qiu, L.; Xu, X.; Zhang, S.; Zou, X.; Wodak, S. J. Prediction of Protein Assemblies, the next Frontier: The CASP14-CAPRI Experiment. *Proteins Struct. Funct. Bioinforma.* **2021**, *89* (12), 1800–1823. https://doi.org/10.1002/prot.26222.

(23) Janin, J.; Henrick, K.; Moult, J.; Eyck, L. Ten; Sternberg, M. J. E.; Vajda, S.; Vakser, I.; Wodak, S. J. CAPRI: A Critical Assessment of PRedicted Interactions. *Proteins Struct. Funct. Genet.* **2003**, *52* (1), 2–9. https://doi.org/10.1002/prot.10381.

(24) Méndez, R.; Leplae, R.; De Maria, L.; Wodak, S. J. Assessment of Blind Predictions of Protein–Protein Interactions: Current Status of Docking Methods. *Proteins Struct. Funct. Bioinforma.* **2003**, *52* (1), 51–67. https://doi.org/10.1002/prot.10393.

(25) Basu, S.; Wallner, B. DockQ: A Quality Measure for Protein-Protein Docking Models. *PLoS One* **2016**, *11* (8), 1–9. https://doi.org/10.1371/journal.pone.0161879.

Section 3 – Employing affitins as molecular probes towards the receptor HER2

Introduction

This section covers a project carried out in collaboration with Bracco S.p.A and Istituto di Scienze e Tecnologie Chimiche "Giulio Natta" (SCITEC) - Italian National Research Council (CNR). The project is closely linked to the publication of two patents[1,2] concerning the application of two small proteins, namely affitins, as molecular probes for targeting the human epidermal growth factor receptor 2 (HER2).

Molecular probes are tools used, within a molecular imaging technique, for the visualization, characterization, and quantification of biological processes at the molecular and cellular level in humans and other living systems[3]. Although the definition given by Mankoff[3] mentions the term process, the target of an imaging technique can correspond to simple molecular entities, such as proteins. Molecular probes consist of a targeting moiety, i.e., a moiety that specifically recognizes the target, and a signal agent, the nature of which depends on the molecular imaging technique; a linker connecting the targeting moiety to the signal agent can be present too[4]. Among molecular imaging techniques, it is worth mentioning positron emission tomography (PET), single photon emission computed tomography (SPECT), fluorescence imaging, and molecular magnetic resonance imaging (mMRI), where the signal agents are radionuclides, fluorescent molecules, and magnetic molecules, respectively.

Molecular imaging techniques are widely employed for the detection of biomarkers associated with cancer diseases both at the diagnostic stage and during the therapy, to monitor its effectiveness. Among these pathologies, breast cancer is of serious concern, considering the estimated 2.26 million new cases worldwide in 2020[5]. Studies aimed at understanding the molecular biology of breast cancer have allowed to identify appropriate targets for the development of specific targeting moieties to be included in molecular probes. Examples of these targets include: i) hormone receptors, i.e., progesterone and estrogen; ii) angiogenic factors such as vascular endothelial growth factor receptors (VEGFR); iii) growth factor receptors, such as the type 1 insulin-like growth factor receptor (IGF-1R), the human epidermal growth factor receptor 1 (HER1 or EGFR), and the human epidermal growth factor receptor 2 (HER2)[6].

The patents[1,2] owned by Bracco S.p.A., from which this study originates, focus on HER2, which is overexpressed in 20-25% of breast cancer cases[7] and is therefore a well-assessed target for both cancer diagnostics and treatment. HER2-targeted therapies involve the use of monoclonal antibodies (mAbs), such as Trastuzumab[8] and Pertuzumab[9], which have been approved for over twenty and ten years respectively, and are also used in combination[10]. The structures of the complexes that the antigen-binding fragments (Fab) of Trastuzumab and Pertuzumab form with the extracellular domain (ECD) of HER2 have been determined with X-ray diffraction[11,12]. They are deposited in the Protein Data Bank (PDB)[13] with the PDB IDs 1N8Z (https://www.rcsb.org/structure/1n8z) and 1S78 (https://www.rcsb.org/structure/1s78), respectively. The cryogenic electron microscopy (cryo-EM) structure of HER2-trastuzumab-pertuzumab complex is available too[14], under the PDB ID 6OGE (https://www.rcsb.org/structure/6OGE).

With the aim of monitoring the efficacy of a therapy based on Trastuzumab and Pertuzumab, there is a clear need for a molecular probe that includes a targeting moiety able to recognize HER2 epitopes other than those recognized by the two mAbs. The answer could lie yet in another mAb. However, it is known that mAbs are characterized by several drawbacks[15], mostly related to their large size (~150 kDa), which causes difficult penetration into tissues and also has an impact on their production, due to their multi-domain structure that makes them rather unstable. The limitations associated with the use of mAbs led researchers to consider alternatives. First, the focus has been on the use of mAb fragments such as antigen-binding fragments (Fab), single-chain variable fragments (scFv), diabodies, triabodies, minibodies and single domain antibodies (sdAb)[16]; however, their use still has some limitations[15]. For this reason, antibody mimetics, i.e., small-sized, stable, synthetic proteins that have nothing in common with mAbs except the ability to specifically bind a partner, have been considered[15,16].

These alternatives to mAbs, and to their fragments, have also been considered for HER2 targeting. Several examples can be found in the recent literature, including an affibody[17], a designed ankyrin repeat protein (DARPin)[18], and a repebody[19].

In the domain of the antibody mimetics, affitins, small (7 kDa, around 66 amino acids), single-chain affinity proteins engineered from the naturally occurring DNA-binding protein family Sul7d[20], can also be included. Proteins from this family, such as Sac7d and Sso7d, are expressed by extremophile organisms *Sulfolobus acidocaldarius* and *Sulfolobus solfataricus*, respectively, and act to prevent DNA denaturation thanks to their stability over a wide temperature (up to 100°C) and pH (0-12) range. The general topology of Sac7d is that of the OB-fold family. Its tertiary structure (*Figure 3.1*) consists of a five-stranded incomplete β-barrel (β1=residues 3-8, β2=11-16, β3= 20-26, β4 = 29-36, β5= 39-46), capped at the opening by a three-turn C-terminal α-helix (residues 53-63). The triple-stranded β-sheet (β3-β4-β5) has been identified as the DNA binding surface[21].

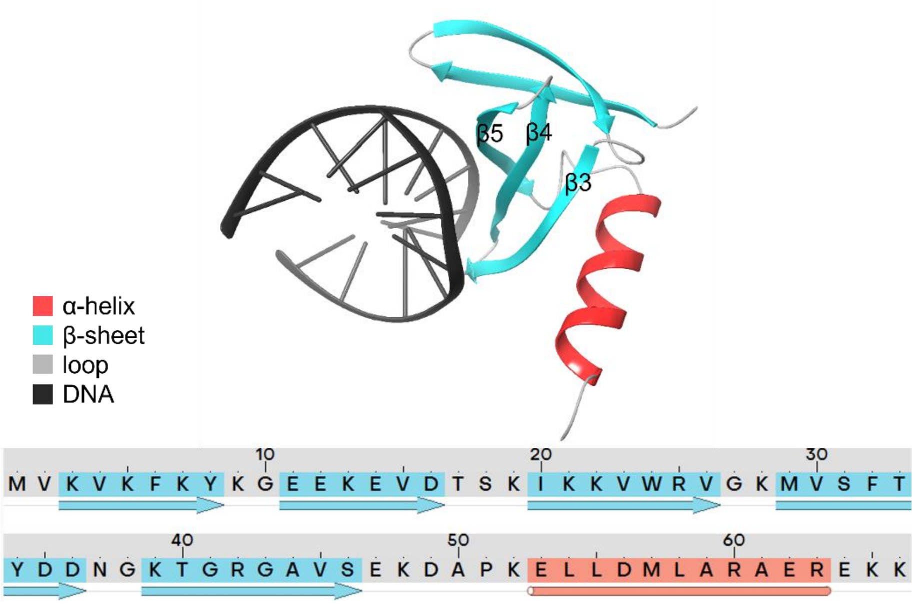

β5
β4
β3
α-helix
β-sheet
loop
DNA
10
20
30
M V K V K F K Y K G E E K E V D T S K I K K V W R V G K M V S F T
40
50
60
Y D D N G K T G R G A V S E K D A P K E L L D M L A R A E R E K K

Figure 3.1 – Representation of the structure of the wild type affitin Sac7d bound to DNA duplex d(GTAATTTAC)2 (PDB ID: 1AZQ). Affitin residues are coloured following the secondary structure assignment, as shown in the legend. DNA is shown in black.

Affitins have been studied for their interesting properties such as high tissue penetration potential and preservation of the exceptional biophysical features of the wild type, i.e., resistance to temperature and pH. Engineered affitins produced by Affilogic S.A.S, which are commercially known with the name Nanofitins®, have already been designed for the targeting of EGFR[22,23].

The same philosophy has been embraced by Bracco S.p.A. In a study conducted in collaboration with Affilogic S.A.S, affitins were engineered with several rounds of ribosome display by the full randomization of the 10-14 Sac7d residues responsible for DNA binding, in order to achieve a high binding affinity towards HER2. Two mutated affitins, among many considered, were identified as the most suitable for HER2 recognition and became the subject of two patents[1,2]. In this thesis, the two affitins will be called Affitin_1 (patent number: WO/2021/122726[1]) and Affitin_2 (patent number: WO/2021/122729[2]).

Competitive binding assays, which are also discussed in the patents, showed that both Affitin_1 and Affitin_2 bind different HER2 epitopes than those involved in the binding of Trastuzumab and Pertuzumab. Furthermore, the two affitins compete for the same binding site (Bracco S.p.A. internal communication). This implies that with proper functionalization, which is also the subjects of the patents, the two affitins could act as molecular probes for HER2 detection during Trastuzumab- and/ or Pertuzumab-based treatments. This would allow continuous monitoring of HER2 levels, enabling real-time assessment of the efficacy of the therapy. Although both affitins have been shown to bind HER2, the structures of the complexes remain undetermined; however, establishing them is crucial to optimize binding affinity, among other things.

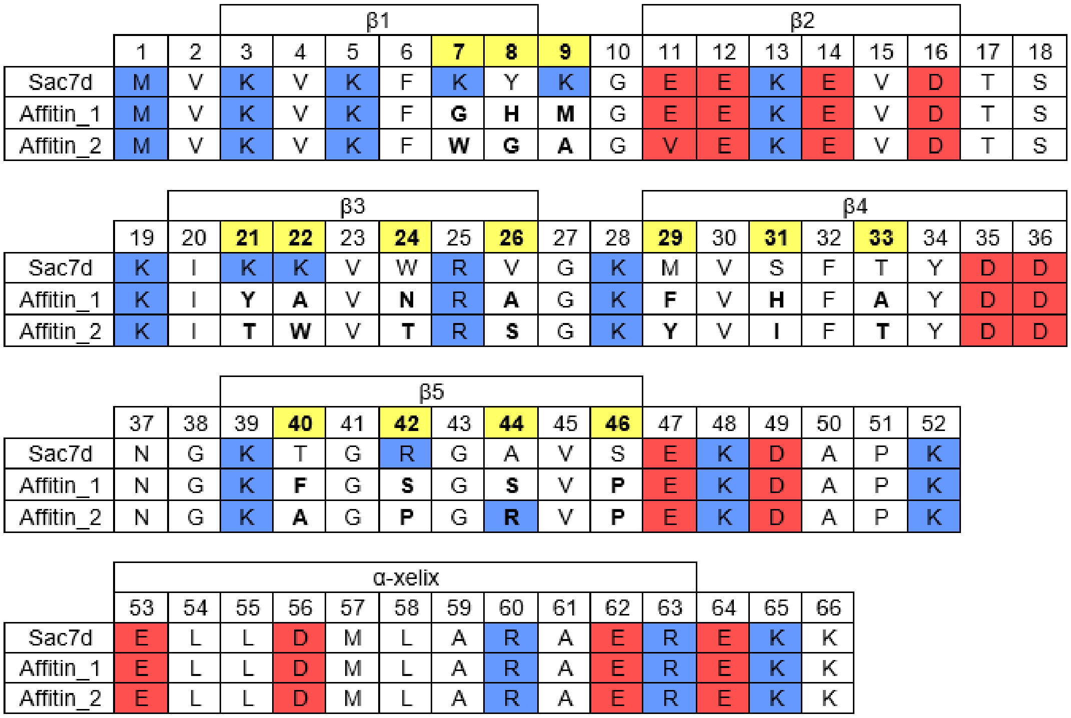
The main aim of this project is therefore to predict, by means of several computational approaches, the structure of HER2-Affitin_1 and HER2-Affitin_2 complexes. It will be shown how these predictions can be exploited to guide further experimental tests, which are necessary for an unambiguous determination of the three-dimensional structures of the complexes. A study is also conducted to understand how the fold of affitins is influenced by the introduction of mutations.

The study is presented in two main sections. In **Section 3.1**, the fold of affitins is studied as a function of the mutations introduced in the sequence of the wild type affitin. **Section 3.2** is in turn divided into two parts. **Section 3.2.1** proposes a procedure for the evaluation of the docking models, which has been published[24] and will be used for the affitins-HER2 use case. **Section 3.2.2** concerns the prediction of affitins-HER2 complexes, driven by the available experimental information, and illustrates how the prediction can be exploited to guide competitive binding assays.

3.1 – A study of the fold of affitins following the introduction of mutations

The sequence of a protein determines its fold, i.e., its three-dimensional structure[25]. The structure is strictly connected to the biological function of the protein in a living organism[26], and thus also to its ability to bind an eventual biomolecular partner. It is known that the introduction of mutations in the amino acid sequence of a wild type protein allows to modulate its binding affinity toward partners other than those occurring in nature. This aspect is exploited for obtaining artificial proteins that specifically recognize protein partners of interest.

Such an approach has been adopted in the study conducted by Bracco S.p.A. aimed at obtaining affitins with high binding affinity towards the HER2 receptor. The sequences of two affitins, Affitin_1 and Affitin_2, have been published in patents WO/2021/122726[1] and WO/2021/122729[2], respectively, and are shown in *Figure 3.2*.



			β1								β2							
	1	2	3	4	5	6	**7**	**8**	**9**	10	11	12	13	14	15	16	17	18
Sac7d	M	V	K	V	K	F	K	Y	K	G	E	E	K	E	V	D	T	S
Affitin_1	M	V	K	V	K	F	**G**	**H**	**M**	G	E	E	K	E	V	D	T	S
Affitin_2	M	V	K	V	K	F	**W**	**G**	**A**	G	V	E	K	E	V	D	T	S

		β3									β4							
	19	20	**21**	**22**	23	**24**	25	**26**	27	28	**29**	30	**31**	32	**33**	34	35	36
Sac7d	K	I	K	K	V	W	R	V	G	K	M	V	S	F	T	Y	D	D
Affitin_1	K	I	**Y**	**A**	V	**N**	R	**A**	G	K	**F**	V	**H**	F	**A**	Y	D	D
Affitin_2	K	I	**T**	**W**	V	**T**	R	**S**	G	K	**Y**	V	**I**	F	**T**	Y	D	D

			β5													
	37	38	39	**40**	41	**42**	43	**44**	45	**46**	47	48	49	50	51	52
Sac7d	N	G	K	T	G	R	G	A	V	S	E	K	D	A	P	K
Affitin_1	N	G	K	**F**	G	**S**	G	**S**	V	**P**	E	K	D	A	P	K
Affitin_2	N	G	K	**A**	G	**P**	G	**R**	V	**P**	E	K	D	A	P	K

	α-xelix													
	53	54	55	56	57	58	59	60	61	62	63	64	65	66
Sac7d	E	L	L	D	M	L	A	R	A	E	R	E	K	K
Affitin_1	E	L	L	D	M	L	A	R	A	E	R	E	K	K
Affitin_2	E	L	L	D	M	L	A	R	A	E	R	E	K	K

Figure 3.2 – Alignment of the sequences of the wild type affitin Sac7d and of Affitin_1 and Affitin_2, object of the patents. Assignment of the putative secondary structure is based on Sac7d structure. The positions of the mutations are shown in yellow and are located at β-strands β1, β3, β4 and β5. The residues carrying a positive and negative charge are highlighted in blue and red, respectively.

However, the experimental three-dimensional structures of Affitin_1 and Affitin_2 are not available.

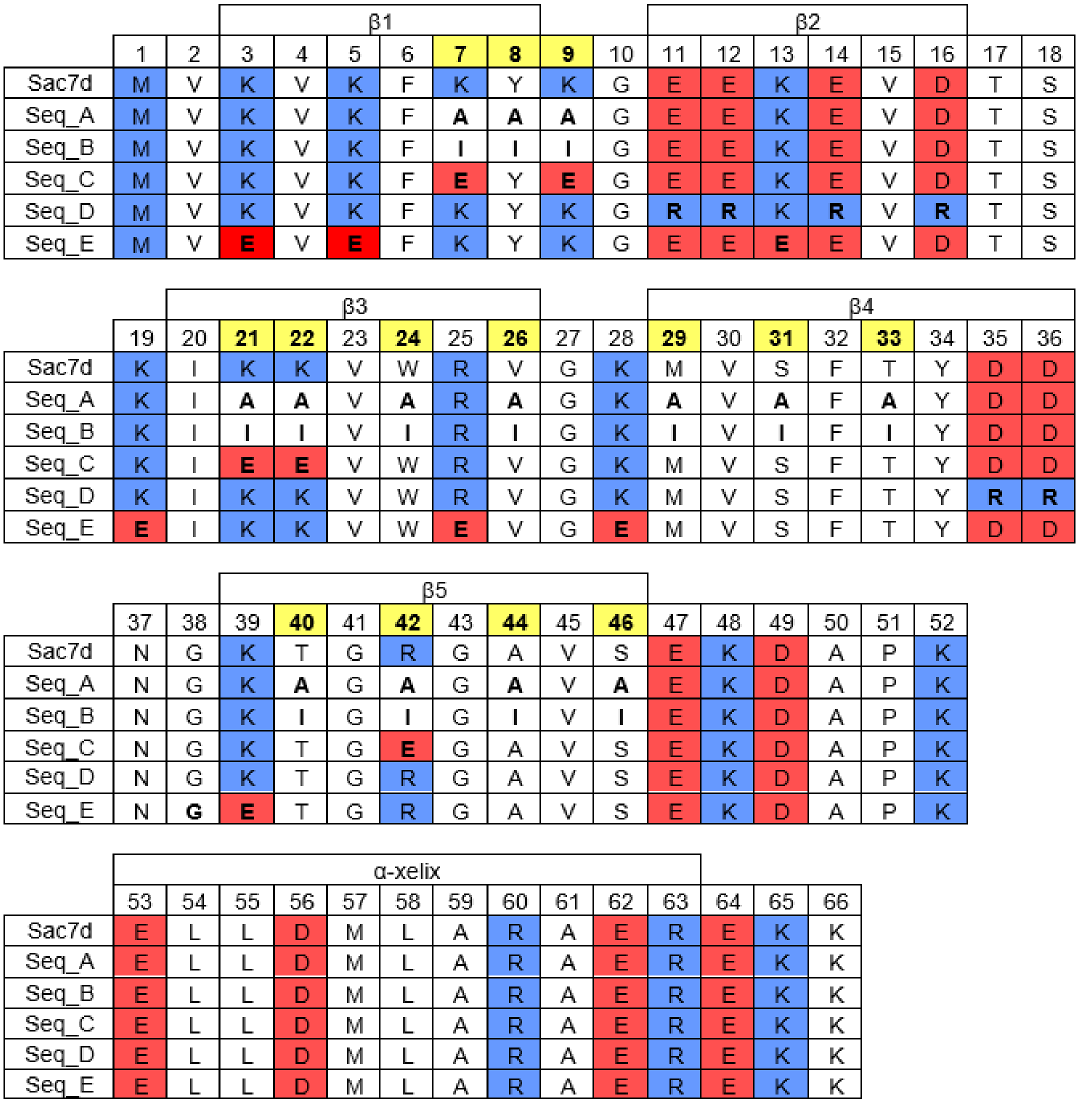
The aim of the first part of the study is thus to assess, by means of homology modelling and Molecular Dynamics (MD) simulations, whether the fold of the mutated affitins changes compared to the fold of the wild type affitin Sac7d. In addition to Affitin_1 and Affitin_2, mutated affitins available in the Protein Data Bank (PDB)[13] and other affitins purposely designed *in silico* by our research group are also considered.

Protocol

Affitins object of the study

The aim of this first part of the study is to have a view as complete as possible of the three-dimensional structures that mutated affitins can assume. Thus, in addition to the affitins covered by the patents (Affitin_1 and Affitin_2, whose sequences are shown in *Figure 3.2*), other affitins are considered.

With the theoretical purpose of designing a large library of affitins characterized by a variety of mutations, it was deemed necessary to assess the stability of Sac7d following the introduction of "extreme" changes in the amino acid sequence. Five sequences (*Figure 3.3*) were thus designed, with mutations mainly involving the 14 residues that interact with DNA. The emphasis was placed on amino acid sequences that allowed for the investigation of the structural role of electrostatic interactions. A mutant, named Seq_A, was thus created where the 14 DNA-binding residues were replaced with alanines. A different one, named Seq_B, foresaw the introduction of 14 isoleucine. These two were conceived to study the effects of small-sized (Seq_A) and slightly larger (Seq_B) apolar side chains. Subsequently, the study shifted to Seq_D, a mutant with arginine (whose side chain contains a guanidinium group, protonated at physiological pH) replacing all residues, among the 14, that had a negative charge in the side chain. Additionally, two mutants named Seq_C and Seq_E were studied, with glutamates replacing some residues with positively charged side chains in the wild type. The sequences of the five affitins Seq_A, Seq_B, Seq_D, Seq_C, and Seq_E are shown in *Figure 3.3*.



			β1								β2							
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Sac7d	M	V	K	V	K	F	K	Y	K	G	E	E	K	E	V	D	T	S
Seq_A	M	V	K	V	K	F	**A**	**A**	**A**	G	E	E	K	E	V	D	T	S
Seq_B	M	V	K	V	K	F	**I**	**I**	**I**	G	E	E	K	E	V	D	T	S
Seq_C	M	V	K	V	K	F	**E**	Y	**E**	G	E	E	K	E	V	D	T	S
Seq_D	M	V	K	V	K	F	K	Y	K	G	**R**	**R**	K	**R**	V	**R**	T	S
Seq_E	M	V	**E**	V	**E**	F	K	Y	K	G	E	E	**E**	E	V	D	T	S

		β3									β4							
	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36
Sac7d	K	I	K	K	V	W	R	V	G	K	M	V	S	F	T	Y	D	D
Seq_A	K	I	**A**	**A**	V	**A**	R	**A**	G	K	**A**	V	**A**	F	**A**	Y	D	D
Seq_B	K	I	**I**	**I**	V	**I**	R	**I**	G	K	**I**	V	**I**	F	**I**	Y	D	D
Seq_C	K	I	**E**	**E**	V	W	R	V	G	K	M	V	S	F	T	Y	D	D
Seq_D	K	I	K	K	V	W	R	V	G	K	M	V	S	F	T	Y	**R**	**R**
Seq_E	**E**	I	K	K	V	W	**E**	V	G	**E**	M	V	S	F	T	Y	D	D

			β5													
	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52
Sac7d	N	G	K	T	G	R	G	A	V	S	E	K	D	A	P	K
Seq_A	N	G	K	**A**	G	**A**	G	**A**	V	**A**	E	K	D	A	P	K
Seq_B	N	G	K	**I**	G	**I**	G	**I**	V	**I**	E	K	D	A	P	K
Seq_C	N	G	K	T	G	**E**	G	A	V	S	E	K	D	A	P	K
Seq_D	N	G	K	T	G	R	G	A	V	S	E	K	D	A	P	K
Seq_E	N	**G**	**E**	T	G	R	G	A	V	S	E	K	D	A	P	K

	α-xelix													
	53	54	55	56	57	58	59	60	61	62	63	64	65	66
Sac7d	E	L	L	D	M	L	A	R	A	E	R	E	K	K
Seq_A	E	L	L	D	M	L	A	R	A	E	R	E	K	K
Seq_B	E	L	L	D	M	L	A	R	A	E	R	E	K	K
Seq_C	E	L	L	D	M	L	A	R	A	E	R	E	K	K
Seq_D	E	L	L	D	M	L	A	R	A	E	R	E	K	K
Seq_E	E	L	L	D	M	L	A	R	A	E	R	E	K	K

Figure 3.3 - Alignment of the sequences of the wild type affitin Sac7d and of the affitins designed in silico (Seq_A, Seq_B, Seq_D, Seq_C, and Seq_E). Assignment of the secondary structure is based on Sac7d structure. The positions of the mutations are shown in yellow and are located at β-strands β1, β3, β4 and β5. The residues carrying a positive and negative charge are highlighted in blue and red, respectively.

Moreover, starting from the sequence of the wild type affitin Sac7d (PDB ID: 1AZQ, https://www.rcsb.org/structure/1azq), a search was performed in the PDB for proteins that show a sequence similar to that of the wild type. Excluding PDB entries in which the affitins are complexed with DNA or are alone, six affitins with mutations in the residues responsible for DNA binding in the wild type forms were retrieved. Each of these affitin binds a different protein partner. The PDB IDs of the affitin-protein complexes are shown in *Table 3.1*, together with the name of the protein partners, the sequence identity with respect to the wild type, the reference to the corresponding paper, and the year in which the structures were released in the PDB. As the three-dimensional structures of these affitins in complex with their protein partners are known, the aim is to verify whether the fold remains the same in solution. The affitins will be referred to with their respective PDB IDs.

PDB ID	Partner of the affitin	Sequence identity to 1AZQ (%)	Reference	Year
1AZQ	DNA	100	[21]	1999
4CJ1	Endoglucanase D	80	[27]	2014
4CJ0	Endoglucanase D	78	[27]	2014
4CJ2	Lysozyme C	78	[27]	2014
5UFE	GTPase KRas	64	[28]	2017
5UFQ	Mutated GTPase KRas	64	[28]	2017
5ZAU	Tyrosine-protein kinase Fyn	63	[29]	2019
6QBA	Retinol-binding protein 4	63	[30]	2020

Table 3.1 – List of the PDB IDs that identify the complexes affitins-partners. The partner of the affitins, the sequence identity with respect to the wild type, the reference, and the year of release of the structures in PDB are also shown.

Preparation of the three-dimensional structures of the affitins

The same procedure was applied for both the affitins object of the patents, i.e., Affitin_1 and Affitin_2, and the affitins designed *in silico*, i.e., Seq_A, Seq_B, Seq_D, Seq_C, and Seq_E. The three-dimensional structures were built, starting from their sequences (*Figure 3.2* and *Figure 3.3*), exploiting the availability of the three-dimensional structure of the wild type affitin Sac7d. A homology modelling procedure was employed, using a tool included in Bioluminate® (Schrödinger Release 2023-3: BioLuminate, Schrödinger, LLC, New York, NY, 2023) and setting the three-dimensional structure of Sac7d (PDB ID: 1AZQ) as template. The models were then refined via the Protein Preparation Wizard tool[31] (Schrödinger Release 2023-3: Protein Preparation Wizard; Prime, Schrödinger, LLC, New York, NY, 2023).

The structures of the affitins retrieved from the PDB (*Table 3.1*) were imported in Bioluminate® and processed with the Protein Preparation Wizard tool as well. Water molecules and counterions were removed, hydrogen atoms and other possibly missing atoms were added, missing side chains and loops were rebuilt, the hydrogen bonding network was optimized, and an energy minimization of hydrogen atoms was performed.

From this point onwards, all the affitins considered were treated with the same protocol.

Molecular Dynamics simulations

Set-up

The structures of the affitins were subjected to Molecular Dynamics (MD) simulations with the aim of evaluating whether the fold of the wild type affitin is preserved following the introduction of the mutations in the amino acid sequence. The same MD protocol was applied for all the affitins.

MD simulations were performed with the open-source software Gromacs[32] (2020.6 release). To speed up the calculations, a united atom force field, namely the Gromos 53A6 force field[33] was used. The SPC water model[34] was used, a choice derived from the fact that SPC is the water model used for both parameterization[33] and, of course, validation[35] of the Gromos 53A6 force field.

Affitins were centred in dodecahedral boxes, keeping a minimum distance of 1 nm from the edges, and solvated with water molecules. Chloride and sodium ions were added to achieve the electroneutrality of the systems. Periodic Boundary Conditions (PBC) were applied in the three dimensions. The systems were minimized with the steepest descent and conjugate gradient algorithms until a convergence criterium of 100 kJ mol^{-1} nm^{-1} was reached. Bonds involving hydrogen atoms were constrained with LINCS algorithm[36] at all stages of the simulations. The equations of motion of atoms were integrated with the leap-frog algorithm every 2 fs. A 1.4 nm cut-off was applied to van der Waals and electrostatics interactions, beyond which the latter have been treated with the Particle Mesh Ewald (PME) algorithm[37]. Initial velocities were generated from a Maxwell distribution at 300 K with a random seed. Solvent was equilibrated at constant temperature (300 K) for 1 ns and at constant temperature (300 K) and pressure (1 bar) for an additional 1 ns. Protein and solvent (including ions) were coupled to two velocity-rescaling thermostats[38,39] every 0.1 ps and to a Parrinello-Rahman barostat[40,41] every 2 ps. During this stage, position restraints (1000 kJ mol^{-1} nm^{-2}) were applied to the heavy atoms of the proteins. Three independent 300 ns production runs were performed at constant temperature (300 K) and pressure (1 bar), using the thermostats and the barostat mentioned in the previous lines. MD simulations were also carried out with the all-atom AMBER99SB-ILDN force field[42], to check whether the result was dependent on the force field used. These simulations were conducted with the same set-up used for those with the Gromos 53A6 force field. However, the TIP3P water model[43] was used, as it is the one used in the development of the AMBER99SB-ILDN force field.

Analysis

The MD trajectories were visually inspected with the Virtual Molecular Dynamics software[44]. Prior to this, the PBC were properly treated, and the trajectories were fitted on the first frame (*gmx trjconv* module). The analysis included: calculation of the root-mean-square deviation (RMSD) of backbone atoms, clustering of the sampled conformations, calculation of the root-mean-square fluctuations (RMSF) of backbone atoms, and calculation of the fraction of secondary structure. The RMSD calculations of backbone atoms were performed on the three trajectories of each affitin after fitting on the first frame (*gmx rms* module), mainly to verify the convergence of the MD simulations. For each affitin, the three 300 ns long trajectories were then concatenated, and the subsequent analysis were carried out on the cumulative trajectories every 500 ps, for a total of 1800 frames. RMSF of backbone atoms of each residue were calculated with the *gmx rmsf* module. The cumulative trajectories were then fitted again on the backbone atoms of the less movable residues (RMSF < 0.3 nm). Cluster analysis was performed as follows: i) RMSD matrix of backbone atoms was calculated (*gmx rms*); ii) clustering was done on the basis of RMSD matrix of backbone atoms with *gmx cluster* module, gromos algorithm[45] and a 0.4 nm cut-off. The central frames (from now on, centrotypes) of the most populated cluster were superimposed to the structure of the wild type affitin. The Define Secondary Structure of Proteins (DSSP) algorithm[46], implemented in *gmx do_dssp* module, was used to analyse the secondary structure of affitins during the MD simulations.

Results

In this section, the analysis carried out on the MD simulations of the affitins are shown. Most of the section will be devoted to the discussion of what was observed for Affitin_1 and Affitin_2, in relation to Sac7d, whose sequences were shown in *Figure 3.2*, as they are the main focus of this part of the study. The results for the affitins retrieved from the PDB (*Table 3.1*) and those designed *in silico* (*Figure 3.3*) will be shown here only briefly and reported more extensively in **Appendix 2.2**.

MD simulations analysis of Sac7d, Affitin_1 and Affitin_2

RMSD of backbone atoms

The RMSD of backbone atoms were calculated to verify the convergence of the MD simulations. *Figure 3.4* shows the trends of this value for the wild type affitin Sac7d and for the affitins object of the patents[1,2], Affitin_1 and Affitin_2.

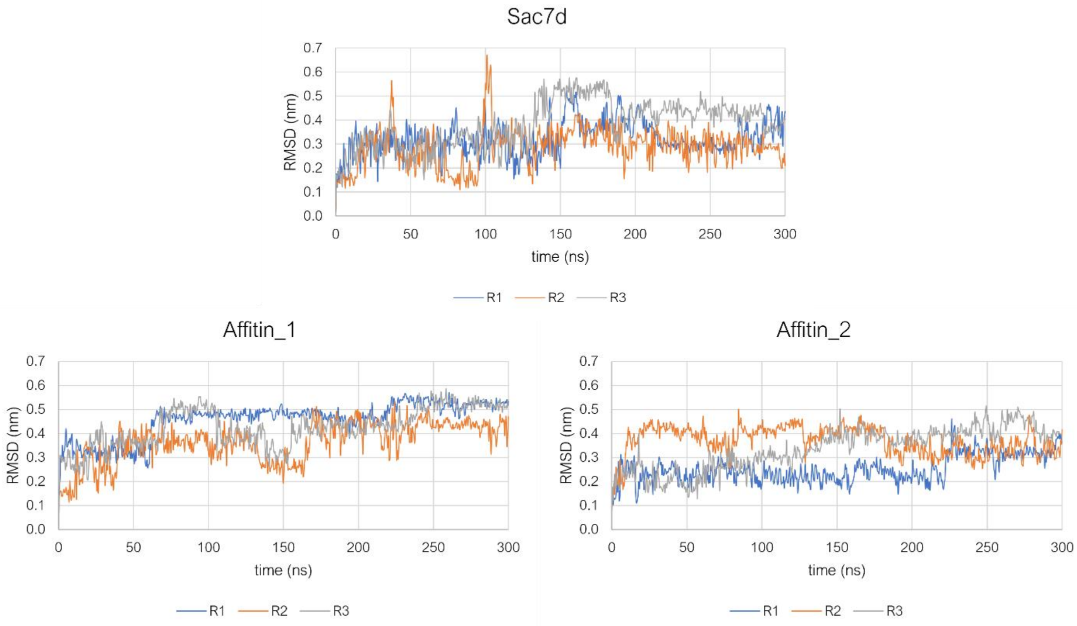


Figure 3.4 – Root-mean-square deviations (RMSD) of the backbone atoms of the wild type affitin Sac7d and of Affitin_1 and Affitin_2 during the three 300 ns long replica (R1-R2-R3).

Most of the increase in these values occurs in the first part of the simulations. Indeed, by calculating the standard deviation (SD) of RMSD in three time intervals of the simulations, i.e., 0-100 ns, 100-200 ns and 200-300 ns, the highest SD values are generally obtained for the first two time span. Considering the three replicas, the RMSD of Sac7d and Affitin_2 stabilizes at around 0.35 nm with respect to the starting frames, while it stabilizes at around 0.45 for Affitin_1. Overall, the simulations can be said to be converged.

RMSF of backbone atoms

The RMSF of backbone atoms was calculated to assess which residues show a greater mobility during the simulations and, on the other side, which retain their position and thus contribute to the stabilization of the structures. *Figure 3.5* shows the comparison of RMSF for the three affitins. Sac7d, Affitin_1, and Affitin_2 behave in a very similar way, showing a greater mobility in the residues (see assignment of the secondary structure of the wild type in *Figure 3.2*) belonging to the terminals (residues 1-2, and 64-66), α-helix (residues 53-63), bend or turn motifs in between the strands (residues 9-10, 17-19, 27-28, and 37-38) and between the strand β5 and the α-helix (residues 47-52).

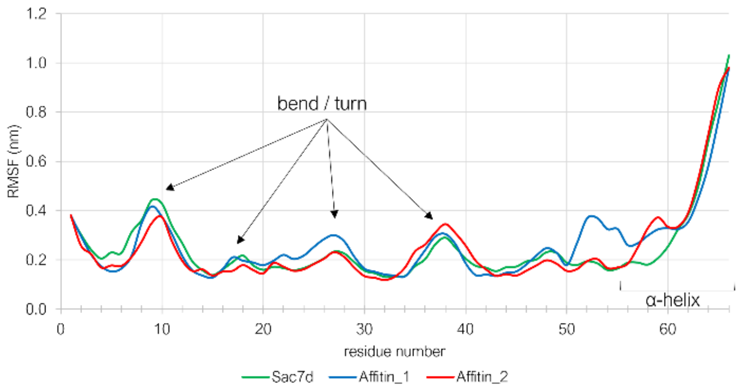


Figure 3.5 – Root-mean-square fluctuations (RMSF) of backbone atoms of each residue calculated for affitins Sac7d, Affitin_1, and Affitin_2 on the cumulative trajectories.

Cluster analysis

Cluster analysis of the sampled conformations was performed on the cumulative trajectories with the gromos algorithm and a 0.4 nm cut-off with the aim of identifying the most representative conformations of the affitins in aqueous solution, i.e., those most frequently sampled during the simulations. The population of the clusters is shown in *Table 3.2*.

Sac7d			Affitin_1			Affitin_2		
#cluster	pop	pop %	#cluster	pop	pop %	#cluster	pop	pop %
1	1355	75%	1	1416	79%	1	1558	87%
2	224	12%	2	243	13%	2	162	9%
3	80	4%	3	105	6%	3	29	2%
4	70	4%	4	23	1%	4	28	2%
5	40	2%	5	14	1%	5	13	1%
6	16	1%				6	7	0%
7	10	1%				7	3	0%
8	4	0%				8	1	0%
9	1	0%						
10	1	0%						

Table 3.2 – Population of the clusters calculated on the cumulative MD trajectories of Sac7d, Affitin_1 and Affitin_2

The centrotypes of the clusters were superimposed to the crystallographic structure of the wild type affitin Sac7d and visually inspected. All affitins present only one mostly populated cluster (representative of the 75%, 79%, and 87% of the overall sampling of Sac7d, Affitin_1, and Affitin_2, respectively); their centrotypes are shown in *Figure 3.6*.

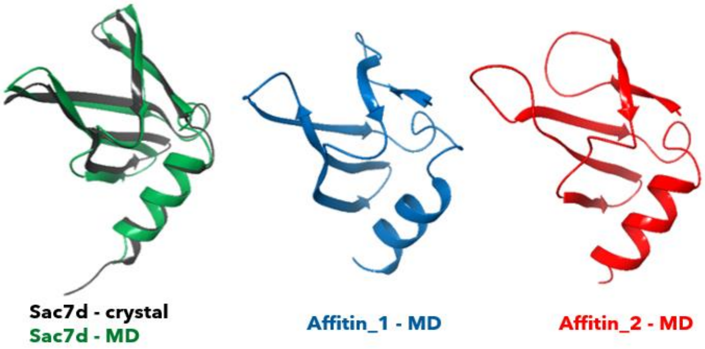


Figure 3.6 - Left: superposition of Sac7d crystallographic structure (PDB ID: 1AZQ, in black) and centrotype of the most populated cluster (75%, green). Centre: Affitin_1 centrotype of the most populated cluster (79%, blue). Right: Affitin_2 centrotype of the most populated cluster (87%, red).

It can be observed that: i) the fold of Sac7d in aqueous solution is almost entirely conserved with respect to the crystallographic structure; ii) the fold of the two engineered affitins is mostly identical to that of the wild type affitin.

Secondary structure

The Define Secondary Structure of Proteins (DSSP) algorithm[46] was employed for the calculation of the fractions of structured and not structured elements of the affitins observed along the MD simulations. Based on this approach, the sum of α-helix, β-sheet, β-bridge and turn elements constitutes the structured part of a protein. *Table 3.3* shows the time-averaged fraction of these elements.

	Structure	Coil	**β-Sheet**	**β-Bridge**	Bend	**Turn**	**α-Helix**	5-Helix	3-Helix
Sac7d	0.73	0.16	0.46	0.00	0.09	0.13	0.14	0.00	0.01
Affitin_1	0.67	0.18	0.47	0.01	0.11	0.12	0.07	0.02	0.02
Affitin_2	0.65	0.19	0.43	0.01	0.12	0.11	0.10	0.01	0.03

*Table 3.3 - Secondary structure elements of wild type affitin Sac7d and of the Affitin_1 and Affitin_2 calculated on the three concatenated replica (total simulation time is 3 * 300 ns for each structure) with the Define Secondary Structure of Proteins (DSSP) algorithm. The term "Structure" refers to the sum of α-helix, β-sheet, β-bridge and turn elements.*

As reported in *Table 3.3*, Sac7d shows the highest fraction of structured elements: 0.73 overall, to be compared with 0.67 (Affitin_1) and 0.65 (Affitin_2). Concerning the β-sheet fraction, a large part of which is involved in the mutations, it can be stated that it is totally conserved in Affitin_1 (0.47 to be compared with 0.46 of Sac7d) and mostly conserved in Affitin_2 (0.43). Part of the α-helix is lost (0.07 and 0.10 for Affitin_1 and Affitin_2, respectively, to be compared with 0.14 in the wild type) but this should be of no concern as it does not belong to an area involved in the binding of a partner.

The fraction of secondary structure elements can also be analysed as a function of the simulation time. As an example, these trends are shown in *Figure 3.7* for the three affitins, one replica each.

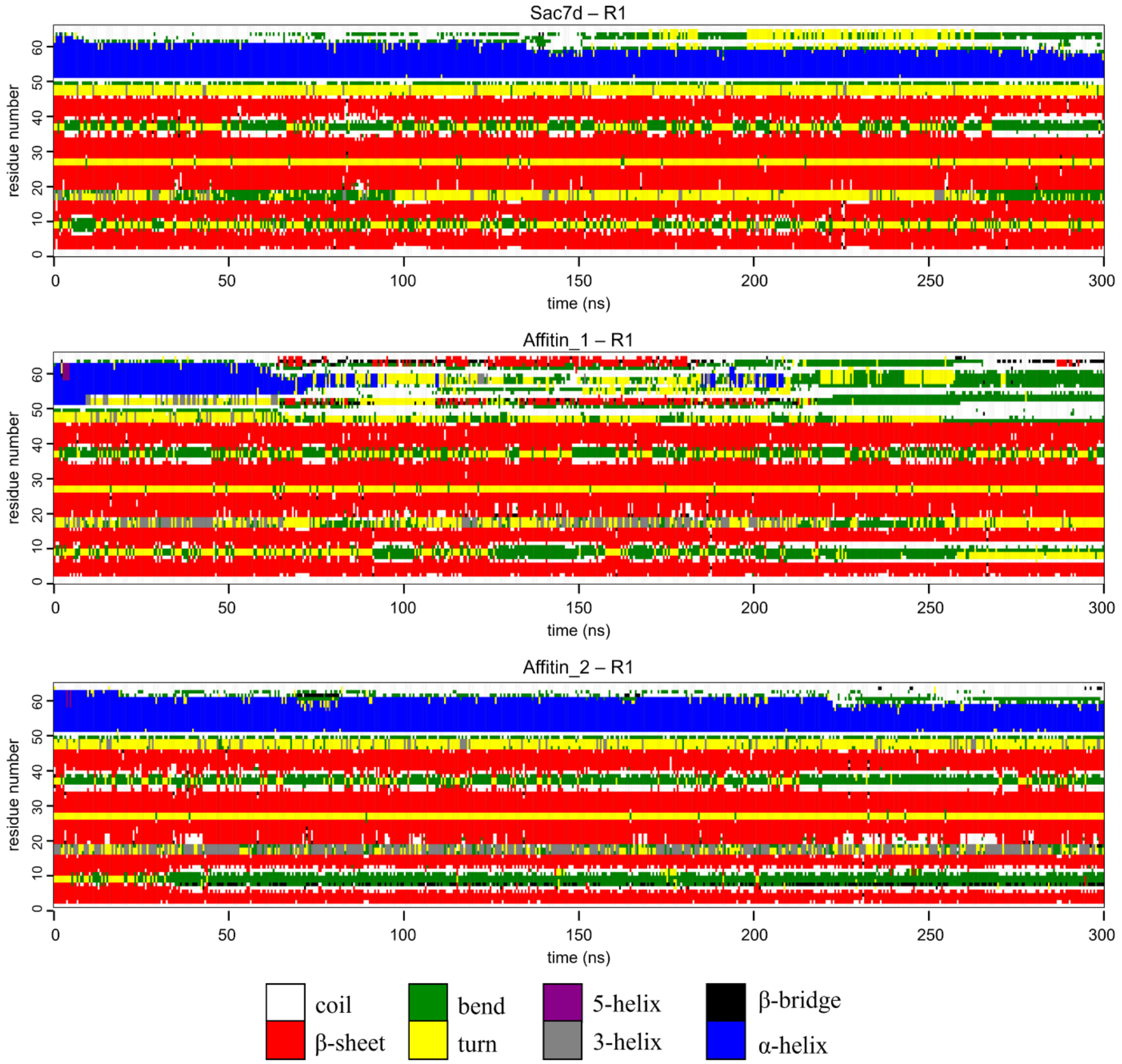


Figure 3.7 – DSSP analysis along the MD trajectories shown for replica 1 (R1) of Sac7d (top panel), Affitin_1 (middle panel), and Affitin_2 (bottom panel).

Overall, the analysis performed showed that the fold of these affitins is stable in aqueous solution and that Affitin_1 and Affitin_2 behave in a very similar way with compared to the wild type.

MD simulations of Sac7d, Affitin_1 and Affitin_2 were also performed with the all-atom AMBER99SB-ILDN force field[42], in order to check if what was observed was not totally dependent on the parameters of the Gromos 53A6 force fied[33]. Analysis of these trajectories showed an even higher stability. The data are shown in **Appendix 2.1**.

Considering that the two force fields led to similar results, it was decided to employ the Gromos 53A6 force field as it reduces the computational cost, being a united-atom force field.

MD simulations analysis of other affitins

MD simulations with the Gromos 53A6 force field were carried out also for the affitins shown in *Table 3.1* and *Figure 3.3*, revealing an overall similar behaviour in aqueous solution. The centrotypes of the most populated clusters are shown in *Figure 3.8*. The other analysis carried out on the trajectories are shown in **Appendix 2.2** and all contribute to the depiction of very stable structures.

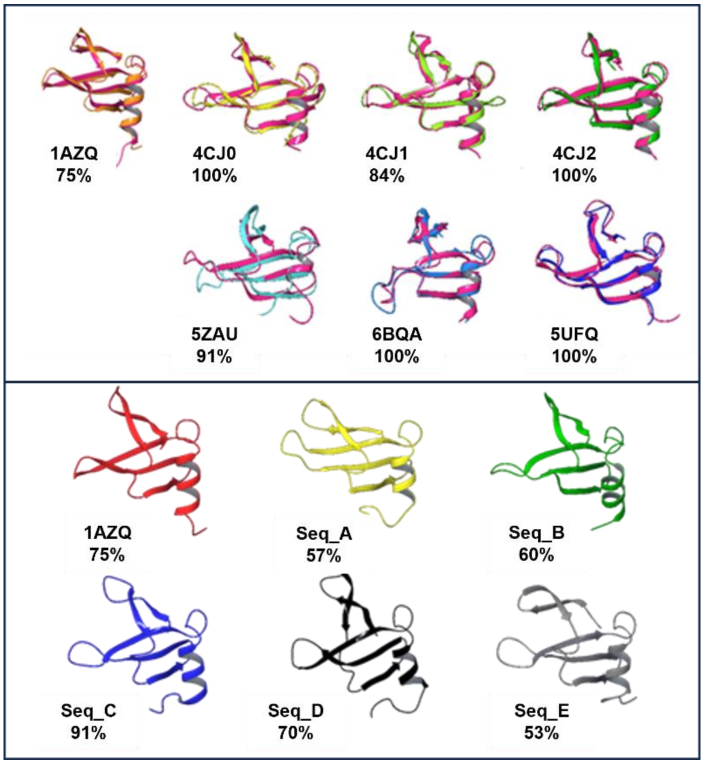


*Figure 3.8 – **Top panel**: superimposition of crystallographic structures of affitins shown in Table 3.1 (in pink) to the centrotypes of the most populated clusters, together with the percentage of their representativeness of the total sampling. **Bottom panel**: centrotypes of the most populated clusters of affitins shown in Figure 3.3, together with the percentage of their representativeness of the total sampling.*

Discussion

The aim of this part of the study was to assess whether and how the structure of mutated affitins changes compared to that of the wild type affitin Sac7d.

Homology models were therefore built for the affitins object of the patents (Affitin_1 and Affitin_2), and for five affitins designed *in silico* with the aim of having the broadest possible picture of the mutations that can be introduced without altering the fold. Among the 55 available in the PDB, the six mutated affitins in complex with a protein partner were also retrieved.

MD simulations were carried out with two different force fields, i.e., the united-atom Gromos 53A6 and the all-atom AMBER99SB-ILDN force fields, to analyse the behaviour of affitins in aqueous solution and to check the eventual dependency on the force field. The clustering of the sampled conformations, the calculation of the RMSF of backbone atoms, and the analysis of the secondary structure via the DSSP algorithm, all showed that the introduction of mutations in the DNA-binding region does not significantly affect the fold of any of the affitins considered here. The strands forming the β-sheets, where the mutations were introduced, are overall conserved. A partial unfold of the α-helix is sometimes observed during the simulations, but this is of less importance since this region is not involved in partner binding. Very similar behaviour was observed with both the Gromos 53A6 and the AMBER99SB-ILDN force fields.

In conclusion, it can be stated that it is possible to introduce any mutation in the β-sheet region of Sac7d without observing significant changes in the composition of the secondary structure, at least up to a sequence identity of around 60%. This result illustrates, in principle, the possibility of designing affitins with any sequence that could be used for targeting other protein partners of biological interest, thus confirming that affitins are useful antibody mimetics.

The main focus remains on Affitin_1 and Affitin_2, objects of the patents owned by Bracco S.p.A. The structures obtained by homology modelling followed by MD simulations, and the centres of the most populated clusters in particular, will be used for the prediction of the structures of the complexes they form with the HER2 receptor (**Section 3.2**).

3.2 – Guiding competitive binding assays using Affitins-HER2 interaction prediction

In the present section, the problem of predicting the structure of the complexes HER2-Affitin_1 and HER2-Affitin_2 is addressed.

As mentioned in **Section 2**, molecular docking approaches are widely employed for a rapid, preliminary prediction of the three-dimensional structure of a biomolecular complex, such as a protein-protein complex. However, the accuracy of the scoring functions that describe the likelihood of the many docking poses that are usually obtained, does not allow for the determination of a unique structure of a protein-protein complex.

Two consequences arise from this.

The first concerns the coupling of the modelling procedure with experimental tests, which provide unequivocable information on the binding interface. In fact, any available experimental evidence on the binding mode should be incorporated into the docking calculation, in order to "restrain" the results of the prediction to solutions that fit what is known with certainty. Then, what results from an *in silico* prediction cannot be taken for granted: instead, docking models should be used to perform targeted experimental tests that may or may not confirm what has been predicted with the model.

The second consequence concerns the need to assess the reliability of the obtained docking poses (see **Section 2.3** for a brief summary of possible "post-docking" procedures). This would also reduce the number of models to be considered for experimental testing.

Summing up, a correct procedure for the prediction of a protein-protein complex should include the following points:

1) The setting up of a docking calculation that takes into account all available information about the binding interface.
2) The evaluation of the docking models obtained with procedures based on different approaches, as the docking scoring function alone cannot unequivocally determine the most likely structure of the complex.
3) The comparison of (a subset of) the docking models with available experimental information, e.g., other eventual partners of one of the two docked proteins.

In the present section, a procedure for the evaluation of the docking models is first presented, based on a dataset of known complexes involving mutated affitins and protein partners (**Section 3.2.1**). The focus is then shifted to the central aim of the study, i.e., the prediction of the complexes HER2-Affitin_1 and HER2-Affitin_2 (**Section 3.2.2**). Experimental information to be exploited to guide the docking calculation is indeed available. In fact, competitive binding assays showed[1,2] that both Affitin_1 and Affitin_2 bind HER2 epitopes different from those involved in the binding of mAbs Trastuzumab and Pertuzumab: this will be used as an input in the docking prediction. Furthermore, it is known that Affitin_1 and Affitin_2 compete for the same binding site (Bracco S.p.A. internal communication); this evidence will be evaluated *a posteriori*. To conclude, HER2 has several known protein partners, and the structures of these complexes are available in the PDB. This information will be combined with a subset of docking models, identified based on the considerations made in **Section 3.2.1**, to drive experimental tests.

3.2.1 – Set-up of a docking and post-docking procedure based on a dataset of known affitin-protein complexes

Protocol

Dataset preparation

The affitin-protein complexes considered are those identified in **Section 3.1**. The PDB IDs are shown in *Table 3.4*, together with the chains selected for docking calculations.

PDB ID	Partner of the affitin: name; chain used	Affitin: name; chain used
4CJ1	Endoglucanase D; chain A	E12 affitin; chain B
4CJ0	Endoglucanase D; chain A	H3 affitin; chain B
4CJ2	Lysozyme C; chain B	H4 affitin; chain D
5UFE	GTPase KRas; chain A	R11.1.6; chain D (5UFQ)
5UFQ	Mutated GTPase KRas; chain A	
5ZAU	Tyrosine-protein kinase Fyn; chain A	Monobody binder; chain B
6QBA	Retinol-binding protein 4; chain A	DNA-binding protein 7a; chain A

Table 3.4 – PDB IDs of the complexes selected for the docking calculations. Proteins names and chains selected are stated; where multiples chains were available, the one more structurally complete was chosen.

The structures of the proteins partners in the complexes were superimposed to their respective unbound forms, when available, to check whether a significant change in their fold occurred upon binding. This does not appear to be the case: therefore, the bound structures of these proteins can be confidently used in the docking prediction.

The protein structures retrieved from the PDB were processed using the tool Protein Preparation Wizard[31] included in the Schrödinger package (Schrödinger Release 2023-3: Protein Preparation Wizard; Prime, Schrödinger, LLC, New York, NY, 2023). The structures were subjected to: 1) removal of water molecules and counterions, if present; 2) addition of hydrogens and other eventually missing atoms; iii) rebuilding of possibly missing side chains and loops with Prime[47,48]; iv) optimization of hydrogen bonding network at neutral pH; v) minimization of hydrogen atoms.

Docking calculations

Docking calculations were performed with the web server ClusPro[49,50,51,52]. The structures of affitins and their partners were uploaded as "ligand" and "receptor", respectively.

The ClusPro web server foresees the following three steps: i) rigid-body docking by sampling billions of conformations; (ii) root-mean-square deviation (RMSD)-based clustering of the 1000 lowest-energy structures generated; iii) refinement of the selected structures using energy minimization.

The interaction energy between two proteins is calculated with the following expression.

$$E = w_1E_{rep} + w_2E_{attr} + w_3E_{elec} + w_4E_{DARS}$$

where: E_{rep} and E_{attr} account for the repulsive and attractive contributions of the van der Waals interaction, respectively, E_{elec} accounts for the electrostatic interactions, and E_{DARS} is related to desolvation contributions. Four scoring schemes ("balanced", "electrostatic-favoured", "hydrophobic-favoured" and "van der Waals + electrostatics") are available, in which different weights (w_1, w_2, w_3, w_4) are assigned to the terms in the expression of the interaction energy[50].

The performance of the four scoring schemes was evaluated by comparing the obtained models with the reference crystallographic structures. The comparison was performed by calculating a parameter, henceforth called crystal_RMSD, which describes the distance of a docking model from the reference structure. crystal_RMSD is calculated as follows: 1) the C_α atoms of the "receptor" (the larger protein) of the docking models are fitted on the same atoms of the crystallographic structure; 2) the RMSD of the C_α atoms of the "ligand" (the smaller protein) of the docking models is calculated with respect to the crystallographic structure.

Docking poses having crystal_RMSD ≤ 5 Å are considered native, while those with crystal_RMSD > 5 Å are considered non-native.

The reranking of the top ten docking poses according to crystal_RMSD values highlights that, for these complexes, the "balanced" scoring scheme performs better than the others, in the sense that it usually ranks the docking poses with the lowest crystal_RMSD values in the first positions. More precisely, the "balanced" scheme ranks the pose with the lowest crystal_RMSD value in the first position for six out of the seven complexes in the dataset. This number is 5, 4 and 1 for the "electrostatics", "hydrophobic" and " van der Waals + electrostatics" scoring schemes, respectively (see **Appendix 2.3**). Therefore, the "balanced" scoring scheme was chosen for the following part of the study.

Evaluation of the docking models: DockQ

In **Section 2.3**, the limitations of docking scoring functions were illustrated. This implies the need to evaluate the reliability of docking poses with additional approaches. One of these focuses on assessing the stability of docking models during MD simulations[53] and is based on the idea that native models, i.e., models closer to the true structure of the complex, should be more stable during the simulations than non-native models. In other words, the mutual position of the two partners should not change significantly in native models.

The stability, and thus the quality of the models, is evaluated through the calculation of the parameter DockQ[54] along the MD trajectories. DockQ originates from the three CAPRI parameters[55] interface-RMSD (I-RMSD), ligand-RMSD (L-RMSD), and fraction of native contacts (Fnat), that were presented in **Section 2.3** together with the DockQ parameter. DockQ ranges from 0 to 1: high-quality models are defined by DockQ ≥ 0.80, medium-quality for 0.80 > DockQ ≥ 0.49, acceptable-quality for 0.49 > DockQ ≥ 0.23, and models are incorrect if DockQ < 0.23.

For all the complexes shown in *Table 3.4*, MD simulations were performed for the reference structure, for the two docking poses showing the two lowest crystal_RMSD values, and for the two docking poses showing the two highest crystal_RMSD values. The pose with the lowest crystal_RMSD value is labelled with A, the one with the second lowest value is labelled with B. The poses having the second highest and the first highest crystal_RMSD values, are labelled C and D, respectively.

MD simulations were performed with Gromacs[32] (release 2020.6) and the trajectories were visualized with Virtual Molecular Dynamics[44]. The united-atom Gromos 53A6 force field[33] was used together with the SPC water model[34]. Proteins were centred in cubic or dodecahedral boxes, keeping a minimum distance of 1 nm from the edges, and solvated with water molecules. Chloride and sodium ions were added to reach electroneutrality. Periodic Boundary Conditions (PBC) were applied in the three dimensions. The systems were minimized with steepest descent and conjugate gradient algorithms until a convergence criterium of 100 kJ mol^{-1} nm^{-1} was reached. The equation of motion of atoms were integrated with the leap-frog algorithm every 2 fs. A 1.4 nm cut-off was applied to van der Waals and electrostatics interactions, beyond which the latter were treated with PME[37]. The set-up of equilibration and production runs followed the work by Jandova et al.[53]: after energy minimization, initial velocities were generated from a Maxwell distribution at 50 K with a random seed. Then, systems were progressively heated up (50, 150, 300 K) while the heavy atoms were positionally restrained with decreasing force constants (1000, 100, 10 kJ mol^{-1} nm^{-2}).

Production runs were performed in NPT ensemble at 1 bar and 300 K, by coupling proteins and solvent to two velocity-rescaling thermostats[38,39] every 0.1 ps and to a Berendsen barostat[38] every 1 ps. All bonds were constrained with LINCS[36]. Analyses were performed every 500 ps. Two replicas (100 ns each) were carried out for the crystal structures and for each of the four docking poses considered, for a total simulation time of 1 μs for each complex.

In addition to the CAPRI parameters (I-RMSD, L-RMSD, Fnat) and the overall DockQ parameter, other parameters were monitored during the MD simulations. The buried surface area (BSA) was calculated with the gromacs module *gmx sasa*, the number of hydrogen bonds (HB) with the module *gmx hbond*, and the protein-protein interaction energy (EPP) with the module *gmx energy*.

Evaluation of the docking models: MLCE

The docking models were also evaluated through a totally different approach.

The Matrix of Local Coupling Energies (MLCE)[56,57,58,59] is a method that can be used for identifying areas (from now on, patches) of an isolate protein that are more likely to interact with a partner, by combining energetic and structural considerations. MLCE is based on the hypothesis that residues playing an important role in the stabilization of the protein folding are not the same that could bind a partner. The analysis of the interaction energy that each residue establishes with all other residues of the protein accounts for these different roles. Residues which strongly interact with the rest of the protein are related to the stabilization of the folding core. The recognition sites, instead, may have weaker pair interactions, as in this way they can easily undergo conformational changes which can make the protein able to recognize and bind a partner. The analysis of the interaction energies of all the amino acids in a protein consists in calculating for each residue the non-bonded part of the potential energy (van der Waals, electrostatic interactions, solvent effects) via a MM-GBSA calculation. The resulting symmetric N×N interaction matrix M_ij (where N is the number of residues of the protein) is then diagonalized and decomposed in eigenvalues and eigenvectors. The first eigenvector is then used to rebuild the energy matrix and multiplied with the contact matrix, which is built from the protein structure, through the Hadamard product, obtaining the MLCE matrix. This matrix is then used to rank spatially contiguous residue pairs with respect to the strengths of their energetic interactions (weakest to strongest). Potential interacting zones are then selected based on the spatial proximity of residues pairs showing the lowest energetic coupling with the rest of the protein, usually selecting the top 15% (but this cut-off can be varied) spatially contiguous residue pairs with the lowest-energy interactions.

The MLCE approach was thus used for predicting the binding sites, i.e., the patches, on the protein partners of the affitins (see *Table 3.4*) with the idea of exploiting the patches to evaluate the quality of docking models. If a model overlaps with a patch, then it can be expected to be more likely than poses that do not show a match with any of the patches.

Calculations were performed with the REBELOT program, version 1.3.2 (https://github.com/colombolab/MLCE). Calculations were performed on the centrotypes of a number of clusters covering at least 90% of the conformation variability sampled during three independent MD simulations (100 ns each, ran as in paragraph "Evaluation of the docking models: DockQ"). The patches were predicted on the centrotype of the most populated cluster considering the top 15% or top 10% of spatially contiguous residue pairs with the lowest-energy interactions.

The patches predicted on the affitins partners were then compared with the residues of the same proteins that are shown to interact with the affitins in the crystallographic structures and in the four docking poses.

Results

Docking calculations

Docking calculations between affitins and their partners were performed with the "balanced" scoring scheme in ClusPro[50,51]. The crystal_RMSD values were calculated for the docking poses of all the complexes considered. The crystal_RMSD values of the first ten poses are shown in *Table 3.5*. For all complexes, one or more native poses (crystal_RMSD ≤ 5 Å) are found among the top ten in the ClusPro ranking. In particular, for complexes 4CJ0, 4CJ1, 6QBA, and 5UFE, two native poses were produced by ClusPro, while only one was produced for 4CJ2, 5ZAU and 5UFQ.

	crystal_RMSD (Å)						
ClusPro ranking	4CJ1	4CJ0	4CJ2	5UFE	5UFQ	5ZAU	6QBA
#0	1.3	2.2	1.6	2.0	2.1	2.8	3.3
#1	3.8	8.0	25.1	14.0	10.0	14.6	8.1
#2	4.1	5.8	22.7	8.9	12.7	8.9	2.6
#3	7.1	5.9	11.2	4.5	11.2	6.6	10.6
#4	4.2	3.0	7.2	12.9	5.5	20.3	9.7
#5	5.4	7.9	26.3	11.9	6.9	17.6	11.3
#6	6.6	5.8	12.0	15.4	17.3	13.0	19.4
#7	8.5	4.0	24.5	5.2	9.7	18.6	12.5
#8	12.2	7.7	27.3	25.6	6.5	26.3	9.4
#9	9.8	3.0	24.5	26.8	9.9	19.7	9.4

Table 3.5 – Crystal_RMSD values are shown for the top ten poses ranked according to the ClusPro score (pose #0 is considered the best) for the complexes under study. Crystal_RMSD values are coloured from dark green (lowest value for the specific complex) to white (highest value).

It is evident that ClusPro ranking does not always correlate with the actual quality of the models, which is here quantified with crystal_RMSD. It sometimes happens that models closer to the reference structure, i.e., those with a lower crystal_RMSD value, are ranked after models of lower quality. This is for example the case for the complex 6QBA, whose best model, for which crystal_RMSD = 2.6 Å, is ranked third (#2 in *Table 3.5*), after a model (#1) with crystal_RMSD = 8.1 Å.

This highlights the need, for a realistic docking scenario where the reference structure is not available, to employ a procedure capable of distinguishing among correct / native and incorrect / non-native models.

For each complex, four docking poses were selected for further analysis aimed at finding out a possible way to properly discriminate among them. More specifically, the chosen poses are:

- the poses showing the lowest (pose A) and the second lowest (pose B) crystal_RMSD value;
- the poses showing the highest (pose D) and the second highest (pose C) crystal_RMSD value.

Poses A are native (crystal_RMSD ≤ 5 Å) for all the complexes, poses B are native (4CJ1, 4CJ0, 5UFE, and 6QBA) or non-native (4CJ2, 5UFQ, and 5ZAU), poses C and D are always non-native. These poses are summarized in *Table 3.6* and shown in *Figure 3.9*, superimposed on the reference structures of the complexes.

	Pose A	Pose B	Pose C	Pose D
4CJ1	#0 - 1.3 - N	#1 - 3.8 – N	#9 - 9.8 – NN	#8 - 12.2 - NN
4CJ0	#0 - 2.2 – N	#4 - 3.0 – N	#5 - 7.9 – NN	#1 - 8.0 – NN
4CJ2	#0 - 1.6 – N	#4 - 7.2 – NN	#5 - 26.3 - NN	#8 - 27.3 – NN
5UFE	#0 - 2.0 – N	#3 - 4.5 – N	#8 - 25.6 – NN	#9 - 26.8 – NN
5UFQ	#0 - 2.1 – N	#4 - 5.5 – NN	#2 - 12.7 – NN	#6 - 17.3 – NN
5ZAU	#0 - 2.8 – N	#3 - 6.6 – NN	#4 - 20.3 – NN	#8 - 26.3 – NN
6QBA	#2 - 2.6 – N	#0 - 3.3 – N	#7 - 12.5 – NN	#6 - 19.4 – NN

Table 3.6 - Docking poses selected for the further evaluations. Native poses (crystal_RMSD ≤ 5 Å) are highlighted in green, while non-native (crystal_RMSD > 5 Å) poses are in red. Labels N or NN are used to indicate native and non-native poses, respectively. The crystal_RMSD values are shown too.

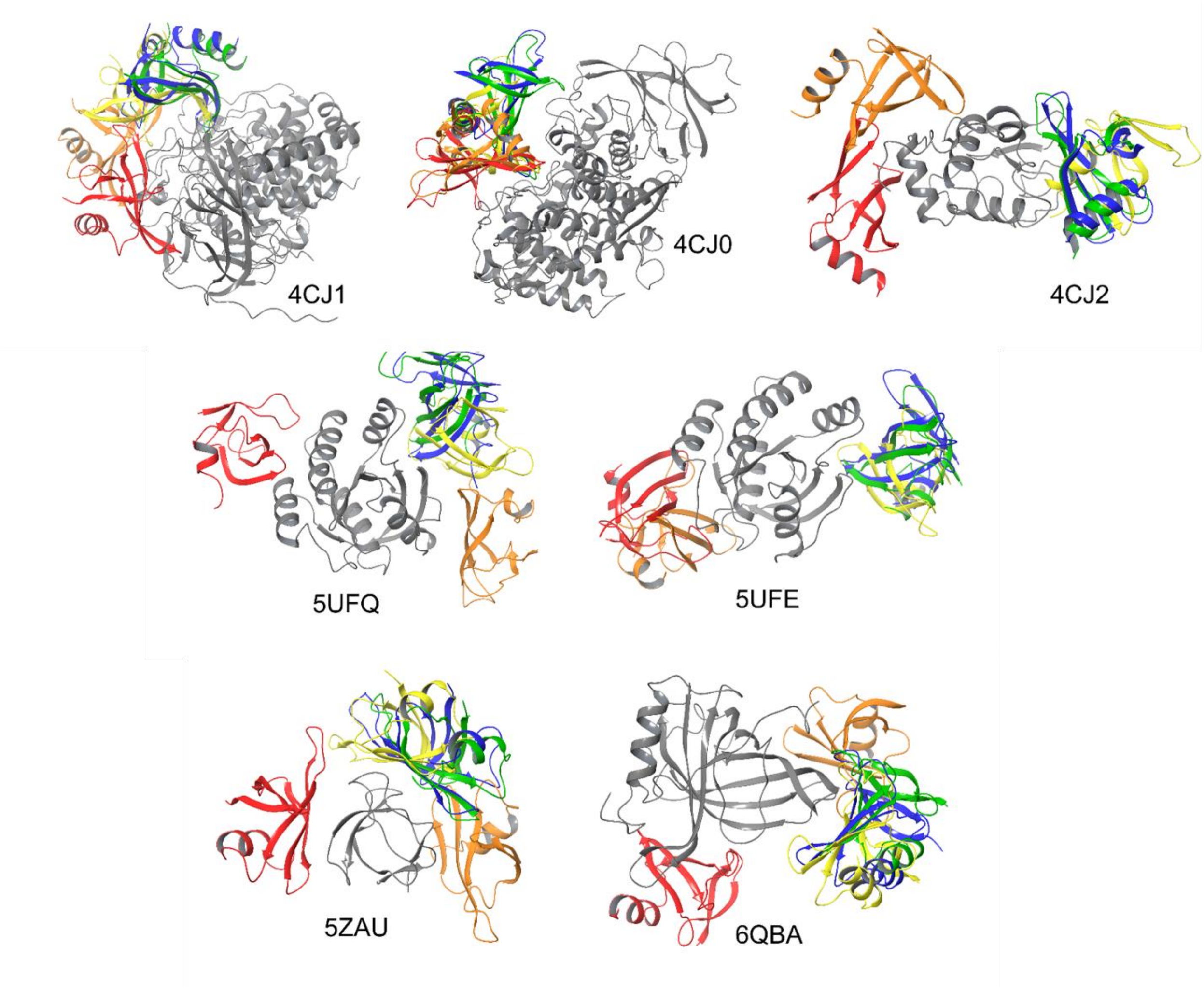


Figure 3.9 – Superimposition, on the structures of the affitins partners – shown in grey, of the four docking poses to the reference structures. The affitins in the reference structures are shown in blue. The affitins in poses A, B, C, and D are shown in green, yellow, orange, and red, respectively.

Evaluation of the docking models: DockQ

The crystallographic structures and the docking poses A, B, C, and D were subjected to MD simulations performed as explained in Methods.

The average values of the three parameters I-RMSD, L-RMSD and Fnat were calculated along the MD trajectories every 500 ps, together with the resulting DockQ values. *Table 3.7* shows the average values and standard deviations of DockQ of the two replicas of each system, obtained for the crystallographic structure of the complex and for poses A (native for all complexes), B (native for 4CJ1, 4CJ0, 5UFE, and 6QBA, and non-native for 4CJ2, 5ZAU and 5UFQ), C and D (always non-native).

	DockQ values and quality of the models				
	Crystal	Pose **A**	Pose **B**	Pose **C**	Pose **D**
4CJ1	0.43 (0.07) – A	0.28 (0.03) – N/A	0.39 (0.03) – N/A	0.48 (0.04) – NN/A	0.23 (0.04) – NN/A
4CJ0	0.34 (0.07) – A	0.50 (0.03) – N/M	0.31 (0.04) – N/A	0.33 (0.04) – NN/A	0.39 (0.03) – NN/A
4CJ2	0.64 (0.03) – M	0.52 (0.04) – N/M	0.27 (0.03) – NN/A	0.19 (0.03) – NN/I	0.39 (0.07) – NN/A
5UFE	0.60 (0.04) – M	0.49 (0.04) – N/M	0.25 (0.05) – N/A	0.31 (0.04) – NN/A	0.29 (0.05) – NN/A
5UFQ	0.43 (0.09) – A	0.45 (0.03) – N/A	0.30 (0.04) – NN/A	0.20 (0.05) – NN/I	0.23 (0.03) – NN/A
5ZAU	0.39 (0.05) – A	0.33 (0.04) – N/A	0.27 (0.04) – NN/A	0.32 (0.03) – NN/A	0.34 (0.04) – NN/A
6QBA	0.48 (0.07) – A	0.44 (0.04) – N/A	0.30 (0.05) – N/A	0.34 (0.04) – NN/A	0.44 (0.04) – NN/A

Table 3.7 – DockQ average values, derived from the two replicas of each system, and standard deviations in parenthesis. Labels N or NN are used to indicate native and non-native poses, respectively. Labels M, A, and I are used to indicate medium, acceptable, and incorrect models respectively, based on the DockQ values.

The quality of a model based on DockQ is high/ medium/ acceptable/ incorrect if the DockQ value is ≥ 0.80/ ≥ 0.49/ ≥ 0.23/ < 0.23[54].

Looking at the values in *Table 3.7*, it is essential to remark that none of the simulations performed on the crystallographic structures, which should show the highest DockQ values overall, led to DockQ values ≥ 0.80 (high quality), whereas only two out of seven (4CJ2 and 5UFE) fall in the medium-quality area (DockQ ≥ 0.49). For this reason, the discussion of the results obtained on the docking poses will not focus on absolute DockQ values; instead, the ability of the DockQ parameter to correlate with the crystal_RMSD will be analysed.

For what concerns poses A, three out of seven (4CJ0, 4CJ2 and 5UFE) are classified as medium quality ones (DockQ ≥ 0.49). As all of them are very close to the crystallographic structure of the complex (see *Figure 3.9* and the crystal_RMSD values in *Table 3.6*), similar DockQ values could be expected between poses A and the corresponding crystallographic structures. Instead, in the case of 4CJ0, the value obtained for the crystallographic structure is lower, whereas it is slightly higher for the other two. For the other four complexes (4CJ1, 5ZAU, 6QBA and 5UFQ) DockQ values of the poses A are in the acceptable quality range (DockQ ≥ 0.23). Moreover, 4CJ1 shows a 0.28 DockQ value, and it is thus almost classified as an incorrect model (DockQ < 0.23) despite its optimal superimposition to the crystallographic structure (crystal_RMSD = 0.13). On the other hand, almost all the poses C and D (non-native) have DockQ values falling in the acceptable range. The lowest DockQ values were obtained for non-native poses (C or D) only for three out of four complexes (pose D for 4CJ1, and C for 4CJ2 and 5UFQ). As for the other complexes, low DockQ values correspond to poses B, with some of them being good models. A specific result that is worth to mention concerns the complex 6QBA: the same DockQ value (0.44) was calculated both for pose A and pose D; in this case a discrimination based on DockQ cannot be done.

In conclusion, only four out of seven poses A (4CJ0, 4CJ2, 5UFE, and 5UFQ) were identified based on DockQ; the doubtful case of 6QBA must also be considered.

For this reason, it was deemed interesting to monitor other parameters along the MD trajectories, namely the buried surface area (BSA), the number of hydrogen bonds (HB) and the protein-protein interaction energy (EPP). The focus was addressed to the relative standard deviations (rSD) of BSA, HB, and EPP, rather than their average values, as the rSD directly reflects the changes that occur in the docking models during the MD simulations.

To better understand the potential usefulness of these parameters in defining the quality of a docking model, a Principal Component Analysis (PCA) was carried out on the correlation matrix of Spearman coefficients of the average values of CAPRI parameters, and of the rSD of BSA, HB, and EPP. This analysis was performed on the parameters obtained from the entire trajectories.

The first two principal components (PCs) account for nearly 85% of the whole data set variability. *Figure 3.10* shows the combined loadings and scores plots (biplot) derived from the PCA.

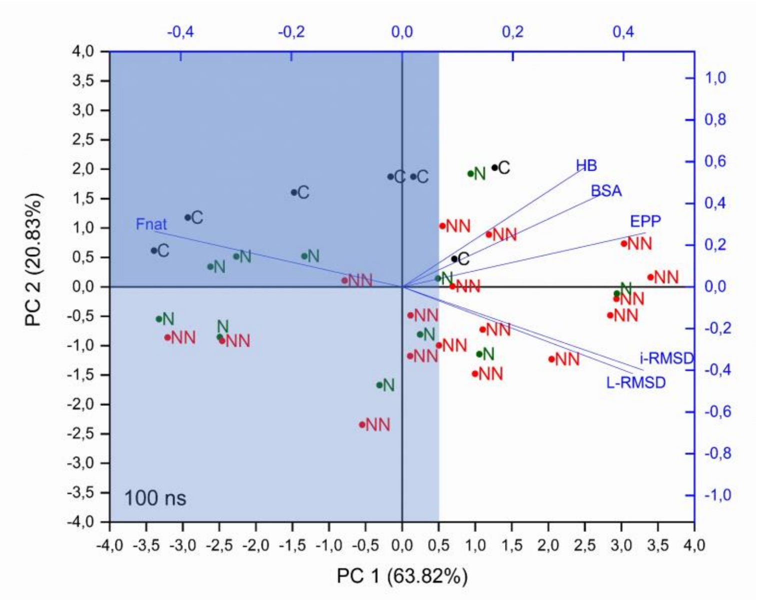


Figure 3.10 - Biplot of a PCA performed on the correlation matrix of Spearman coefficients of the average values of CAPRI parameters and of the rSD of BSA, HB and EPP, calculated on the whole trajectories. C, N, and NN labels indicate crystallographic structures, native and non-native poses respectively. Light blue area indicates PC1 < 0.5. Dark blue identifies the intersection between areas defined by PC1 < 0.5 and PC2 > 0.

The loadings of the principal components PC1 and PC2 are shown in *Table 3.8*.

	PC1	PC2
L-RMSD	0.4164	-0.4157
I-RMSD	0.4362	-0.3995
Fnat	-0.4483	0.2665
BSA	0.3652	0.4504
HB	0.3294	0.5714
EPP	0.4399	0.2593

Table 3.8 - Loadings of principal components PC1 and PC2, obtained from the PCA performed on the correlation matrix of Spearman coefficients of the average values of CAPRI parameters and of the rSD of BSA, HB and EPP, calculated on the whole trajectories.

PCA analysis shows that I-RMSD, L-RMSD and Fnat are highly correlated, with Fnat showing loadings (*Table 3.8*) of opposite sign. This is expected as the conservation of a high fraction of the contacts (a high value of Fnat) is accompanied by limited changes in the position of one protein partner relative to the other (low values of I-RMSD and L-RMSD). HB, BSA and EPP are highly correlated with each other and almost orthogonal to the CAPRI parameters, thus indicating a lack of a correlation with the latter.

Looking at the components of each object (labelled in *Figure 3.10* with C, N, and NN for crystallographic structure, native and non-native poses respectively), i.e., their position in the plane defined by PC1 and PC2, the following can be stated.

PC1 is partially able to discriminate the poses: over 70% of Cs and Ns fall within PC1 values < 0.5 (light blue area in *Figure 3.10*), while 65% of NNs are above this value. Along PC2, a clear distinction is only visible for Cs, as they all lie at PC2 > 0, while Ns and NNs are almost equally scattered. The dark blue area in *Figure 3.10* identifies the intersection between the areas defined by PC1 < 0.5 and PC2 > 0.

In summary, PCA analysis can discriminate around 70% of native and non-native poses. It could therefore be a useful tool for determining the quality of docking poses in a realistic docking scenario, where the crystallographic structure of the complex is not available.

Evaluation of the docking models: MLCE

The MLCE approach was used to predict the binding sites (patches) on the protein partners of the affitins. The aim was to exploit the predicted patches for the evaluation of the quality of the docking models: models in which the affitin overlaps with a patch can be considered more likely than those that do not show a match with any of the patches.

It should be remarked that MLCE predicts protein areas that can be recognized by any potential binding partner. In this case, this means that not all the predicted patches are regions that are actually involved in the binding of affitins.

Calculations were performed as in Methods. *Figure 3.11* shows the crystallographic structures of the complexes, with the patches on the protein partners of the affitins highlighted in different colours to make visible whether they correspond to the actual binding sites of the affitins. The reliability of the MLCE prediction varies depending on the protein considered.

For the partner binding the affitin in the complex 6QBA, a large patch involving all the interacting residues is predicted (*Figure 3.11*). For five out of the seven analysed complexes (4CJ0, 4CJ1, 4CJ2 and 5ZAU), a partial overlap is observed between the MLCE-based patches and the partners residues that actually interact with the affitins. Finally, the predictions performed on the partners of the affitins in the complexes 5UFE and 5UFQ do not include residues responsible for the binding of the affitins.

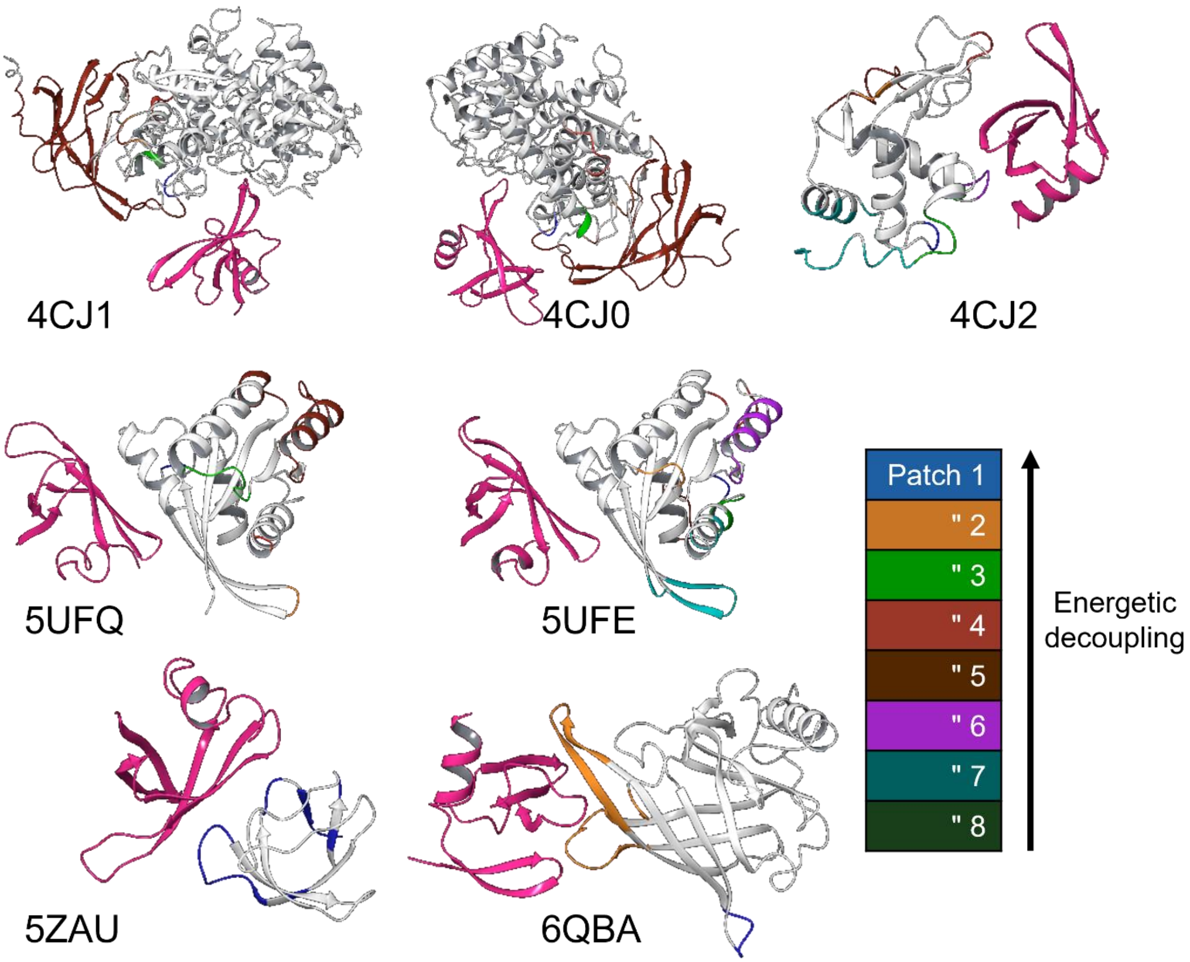


Figure 3.11 – Crystallographic structures of the complexes object of the study. The affitins are shown in pink. The partners of the affitins are shown in light grey. The MLCE-predicted patches on the partners of the affitins are shown as in the legend in the figure, according to their degree of coupling with the protein itself.

Given these results, it can be concluded that in most of the situations analysed here, MLCE was able to identify residues involved in the binding of the affitins. Therefore, MLCE could be used as a tool for the reranking of docking poses.

The number of residues of the affitins in the docking poses A, B, C, and D that interact with residues belonging to the predicted patches on the protein partners was calculated; these are shown in *Table 3.9*.

	4CJ1	4CJ0	4CJ2	5UFE	5UFQ	5ZAU	6QBA
Pose A	10	**7**	**15**	1	0	**14**	**23**
Pose B	3	4	5	1	1	13	**23**
Pose C	0	3	9	**15**	2	**14**	18
Pose D	**12**	1	12	13	**9**	2	1

Table 3.9 - Number of residues of the affitins in the docking poses A, B, C, and D that interact with residues belonging to the patches predicted on the protein partners. The highest number of residues for each complex is highlighted in green.

For complexes 4CJ0 and 4CJ2, the highest number of affitin residues interacting with a patch is observed for poses A, i.e., those closest to the reference structures. Complexes 5ZAU and 6QBA show the highest number of affitin residues interacting with a patch for two poses at the same time, namely poses A and C in 5ZAU, and poses A and B for 6QBA. Finally, for complexes 4CJ1, 5UFE, and 5UFQ, the highest number of affitin residues interacting with a patch is found for docking poses C or D, which are non-native.

Overall, two poses A (4CJ0 and 4CJ2) can be identified on the basis of MLCE analysis. Of the two doubtful cases (5ZAU, 6QBA) where poses A could not be distinguished from others, it is worth noting that, concerning 6QBA, pose B is close to pose A, having crystal_RMSD values of 3.3 Å and 2.6 Å, respectively (see also *Figure 3.9*).

These results indicates that the quality of a docking model cannot be defined solely on the basis of MLCE prediction only; however, at the same time, the comparison with the patches can be useful when considered in conjunction with a different approach aimed assessing the quality of the poses.

DockQ-MLCE consensus approach

In the paragraphs above, two totally different approaches, namely DockQ and MLCE, were used in the attempt to identify the docking models closest to the reference structures. DockQ[54] aims to measure the stability of the docking poses during MD simulations. MLCE[58] instead predicts the interacting residues (patches) of one of the two protein partners through the analysis of its energetic and structural-dynamical properties; the docking poses are then compared with the patches.

It was shown that these methods sometimes lead to incorrect evaluations as they are not always able to point out the best poses (named poses A here). More specifically, among the 7 complexes analysed in the study, four poses A (4CJ0, 4CJ2, 5UFE, and 5UFQ) were correctly identified on the basis of DockQ; there was also one doubtful case (6QBA), in which pose A could not be distinguished from pose D. Considering the MLCE results instead, two poses A (4CJ0 and 4CJ2) were identified; in two cases it was not possible to distinguish poses A from pose B (6QBA) or C (5ZAU).

Despite sometimes misleading results, the two approaches were thought to deserve a second chance, as they proved to be able to correctly discriminate among docking models. Moreover, as mentioned above, they rely on totally different assumptions, which is a plus.

It was therefore decided that it was worthwhile to make an attempt in which the quality assessment of the models was based on DockQ and MLCE at the same time. In other words, the decision on which model is best was made by seeking a consensus between the two approaches.

Table 3.10 shows the DockQ values and the number of affitins residues in the docking models interacting with the patches predicted based on MLCE.

Pose	Parameter	4CJ1	4CJ0	4CJ2	5UFE	5UFQ	5ZAU	6QBA
Pose A	DockQ	**0.28**	**0.50**	**0.52**	0.49	0.45	**0.33**	**0.44**
	MLCE	**10**	**7**	**15**	1	0	**14**	**23**
Pose B	DockQ	**0.39**	0.31	0.27	0.25	**0.30**	0.27	0.30
	MLCE	**3**	4	5	1	**1**	13	23
Pose C	DockQ	0.48	0.33	0.19	**0.31**	0.20	0.32	0.34
	MLCE	0	3	9	**15**	2	14	18
Pose D	DockQ	0.23	0.39	0.39	0.29	**0.23**	0.34	0.44
	MLCE	12	1	12	13	**9**	2	1

Selected model	DockQ	C	A	A	A	A	D	A/D
	MLCE	D	A	A	C	D	A/C	A/B
	DockQ + MLCE	**A/B**	**A**	**A**	**C**	**B/D**	**A**	**A**

Table 3.10 - Selection of the models based on DockQ and MLCE. For each complex and each pose, the DockQ values and the number of residues of the affitins interacting with a patch (labelled with MLCE) are reported. The selected models identified on the basis of the two approaches together are shown in bold.

The evaluation of both parameters at the same time was done by applying the following procedure for each complex:

i) selection of the pose presenting the highest values of both parameters. In this way, poses A of complexes 4CJ0 and 4CJ2 are selected as the most probable.

ii) if the previous point does not apply, exclusion of the poses characterized by one best- and one worst-scoring parameter at the same time. The following poses are excluded: for complex 4CJ1, pose C and D, for complexes 5ZAU and 6QBA, poses D, for complexes 5UFE and 5UFQ, poses A.

iii) selection of the pose presenting the highest values of both parameters, i.e., repeat the first step. In this way, poses A of complexes 5ZAU and 6QBA, and pose C of complex 5UFE are selected.

iv) if the previous point does not apply, selection of the poses presenting the highest value of each parameter, respectively. For complex 4CJ1, poses A and B are selected, for complex 5UFQ, poses B and D.

Overall, five unique poses were identified as the most likely ones: poses A for complexes 4CJ0, 4CJ2, 5ZAU and 6QBA, and pose C for complex 5UFE. For complexes 4CJ1 and 5UFQ the identification of a unique docking pose was not possible: for the former complex, the criteria adopted led to the selection of poses A and B, while for complex 5UFQ, to the selection of poses B and D.

Considering that poses A and B of complex 4CJ1 partially overlap (see *Figure 3.9*) the combination of the two approaches made it possible to identify five correct models (four poses A, and one A/B), improving the prediction obtained by applying the two approaches individually.

Discussion

In the present section, a dataset consisting of seven complexes composed of affitins and protein partners was exploited to propose an approach aimed at evaluating docking poses. The focus has been on the need for a procedure that, given a set of docking poses, is able to distinguish between good / native models, i.e., models close to the true structure of the complex, and incorrect / non-native models. This stage is an essential step in predicting the structure of a complex, which cannot rely on docking scoring functions alone.

The docking poses obtained with the ClusPro web server were compared with the reference structures, through the calculation of the crystal_RMSD parameter. For each complex, the two closest and the two furthest poses from the reference structure, among the top ten ranked by ClusPro, were selected for two further analyses, based on completely different assumptions.

The first analysis relates the quality of the docking poses to the mutual stability of the two partners, quantified by calculating the DockQ parameter along the MD trajectories. The higher the DockQ, the higher the stability and thus the plausibility of the pose.

The second approach involves the MLCE-based prediction of the binding sites (patches) on the isolated structure of one of the two protein; in this case, the calculations were done on the partners of the affitins. The existence of an eventual match between the patches and the binding areas involved in the docking models is then checked: if this match does indeed exist, the pose is considered more likely.

Both the DockQ and MLCE approaches proved to be rather reliable in identifying the correct docking models, although they sometimes produced misleading results. Of particular importance is the fact that the DockQ values obtained were all within a narrow range, thus not always reflecting the differences between native and non-native docking models. Therefore, the decision as to which docking models are the most probable was made on the basis of the two approaches simultaneously, i.e., seeking a consensus between the two. The combined use of DockQ and MLCE proved to be more effective, allowing more docking models to be retrieved than those identified on the basis of the two approaches considered separately.

Finally, although the ClusPro scoring function alone is not entirely reliable, as has been shown, it should also not be totally ignored when deciding which docking model is most likely. Instead, the ClusPro score can be included as a third criterium for determining the most probable docking poses, especially in cases where a univocal decision cannot be made on the basis of DockQ-MLCE.

The DockQ-MLCE approach presented here will be used to address the specific use case of determining the structure of the complexes HER2-Affitin_1 and HER2-Affitin_2 (**Section 3.2.2**).

3.2.2 – Application of the procedure to the Affitins-HER2 use case

Protocol

Preparation of the input files

Affitins. The structures of Affitin_1 and Affitin_2 were prepared as in **Section 3.1**. A homology modelling procedure was employed, exploiting the three-dimensional structure of the wild type affitin Sac7d as a template. Molecular Dynamics (MD) simulations were carried out (3 replicas, 300 ns each), and a cluster analysis was performed on the cumulative trajectories. The centrotypes of the most populated clusters, being the most representative structures of the overall sampling, were used for the docking calculations discussed in the present section.

HER2. The three-dimensional structure of the receptor was retrieved from the PDB (PDB ID: 6OGE). The structure file was processed with Protein Preparation Wizard[31] (Schrödinger Release 2023-3: Protein Preparation Wizard; Prime, Schrödinger, LLC, New York, NY, 2023.) to remove water molecules and counterions, add hydrogen and other eventually missing atoms, rebuild missing side chains and loops, optimize the hydrogen bonding network, and perform an energy minimization of hydrogen atoms.

Docking calculations

HER2-affitins docking calculations were carried out with the web server ClusPro[50,51]. The structures of the affitins and HER2 were uploaded as "ligand" and "receptor", respectively. The "balanced" scoring scheme was used, which had proven to be the best performing for complexes that include affitins (see **Section 3.2.1**).

The available experimental information was exploited as follows. An attractive potential was applied to the 14 mutated residues of the affitins, as shown in *Figure 3.12*, in order to drive the docking prediction towards models in which these residues are part of the affitin-HER2 interface.

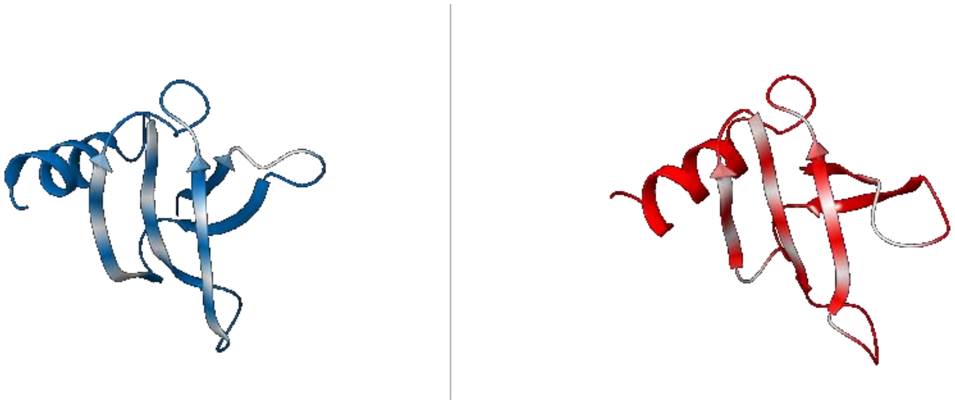


Figure 3.12 - Representation of Affitin_1 (left, blue) and Affitin_2 (right, red). The 14 mutated residues, to which the attractive potential was applied to guide docking calculations, are shown in white.

Concerning HER2, residues within 10 Å from the mAbs Trastuzumab and Pertuzumab were masked during the calculation (see *Figure 3.13*), in order to satisfy the experimental evidence according to which the affitins bind HER2 on different epitopes.

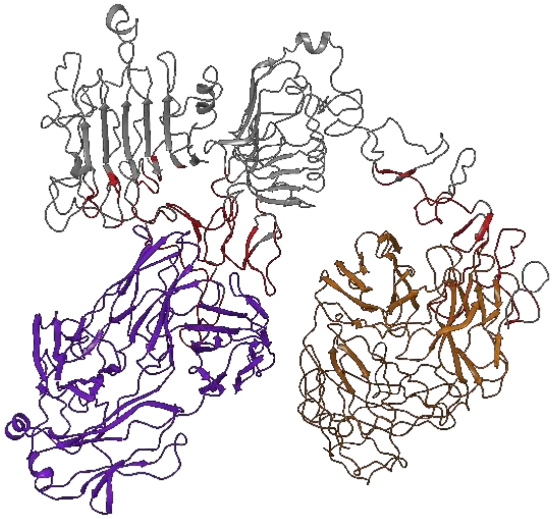


Figure 3.13 – Representation of the protein complex (PDB ID 6OGE) formed by HER2 (grey) and the mAbs Trastuzumab (orange) and Pertuzumab (violet). The HER2 residues within 10 Å from the two mAbs, which were masked during the docking calculations to satisfy the experimental evidence according to which the affitins bind HER2 on different epitopes, are shown in red.

The resulting docking models were visually inspected and, as explained in the Results, a subset of them was selected for the DockQ-MLCE analysis.

DockQ-MLCE evaluation of the docking models

MD simulations were conducted on the chosen docking models, following an approach[53] based on the idea that models that are more similar to the true structure of the complex will maintain stability throughout the simulation compared to less accurate or incorrect models. The set-up of MD simulations was the same as that employed for the complexes affitins-other protein partners (see **Section 3.2.1**). The stability and, consequently, the quality of the models was assessed by calculating the DockQ parameter[54] during the MD trajectories.

In parallel, the Matrix of Local Coupling Energies (MLCE) method[56,57,58,59], introduced in **Section 3.2.1**, was used to predict HER2 areas (patches) that are most likely to bind a partner, and thus also the affitins. The calculations were carried out with the program REBELOT, version 1.3.2 (https://github.com/colombolab/MLCE) on the centrotypes of 4 clusters which cover around 90% of HER2 conformation variability sampled during three MD simulations (100 ns each). The simulations were performed in presence of the mAb Trastuzumab to avoid a large displacement of the domain IV of the receptor, which is unlikely to occur when the mAb is bound, but that was observed for MD simulations carried out on the receptor alone.

The patches were predicted on the crystal structure of HER2 (PDB ID: 6OGE, the same used for docking calculations) considering the top 15% of spatially contiguous residue pairs with the lowest-energy interactions.

Results

Docking calculations

Docking calculations carried out with ClusPro, driven by the experimental information available, resulted in 27 and 28 models for the partners HER2-Affitin_1 and HER2-Affitin_2 respectively, out of a maximum of 30 that ClusPro can provide. They were visually examined by superimposing them on each other, on the receptor structure. Both Affitin_1 and Affitin_2 appear to bind only four different areas of the HER2 receptor. These areas are highlighted by colouring the affitins in red, yellow, orange, and green (*Figure 3.14*).

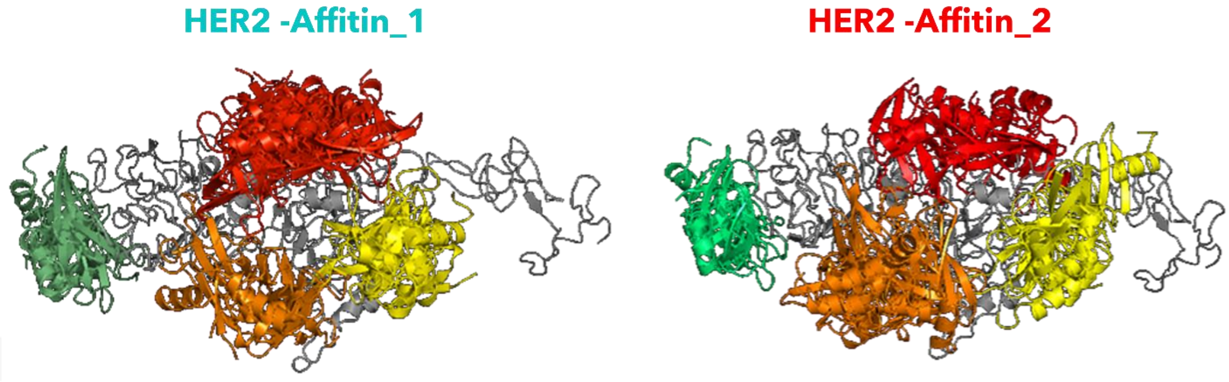


Figure 3.14 - Superimposition of all the docking models obtained for HER2-Affitin_1 (27 models, at left) and HER2-Affitin_2 (28 models, at right). HER2 is shown in grey. Affitins are coloured based on the four different areas: red, yellow, orange, and green.

Table 3.11 shows for the first ten models, ranked by ClusPro according to the number of cluster members, the binding area, indicated by the colours. Also indicated is the number of mutated affitin residues, to which an attractive potential was applied in the docking predictions, that are in contact with HER2 (within a cut-off of 5.5 Å).

ClusPro ranking		Num. of clusters members		Mutated Affitins residues in contact with HER2*	
Affitin_1	Affitin_2	Affitin_1	Affitin_2	Affitin_1	Affitin_2
🟥 #0	🟥 #0	131	92	**14**	**14**
🟨 #1	🟨 #1	110	85	12	**14**
🟥 #2	🟧 #2	68	66	13	**14**
🟥 #3	🟩 #3	64	54	**14**	11
🟩 #4	🟨 #4	52	54	10	12
🟥 #5	🟧 #5	49	53	**14**	9
🟨 #6	🟥 #6	47	53	13	13
🟧 #7	🟧 #7	46	50	12	13
🟥 #8	🟨 #8	45	42	12	7
🟩 #9	🟩 #9	42	41	12	9

Table 3.11 – List of the first ten docking models for Affitin_1 and Affitin_2. The models are ranked based on the ClusPro score, which consists of the number of structures in the clusters. The number of mutated Affitins residues in contact with HER2 (within a 5.5 Å cut-off) is shown too.

For the evaluation based on DockQ and MLCE, among all the models shown in *Figure 3.14* and *Table 3.11*, the best scoring models of each area for Affitin_1 and Affitin_2 were selected. The models are labelled by their ClusPro ranking and the colour indicating the binding area. For Affitin_1, the selected models are: #0-red, #1-yellow, #4-green, #7-orange. For Affitin_2, the selected models are: #0-red, #1-yellow, #2-orange, #3-green. They are shown in *Figure 3.15*.

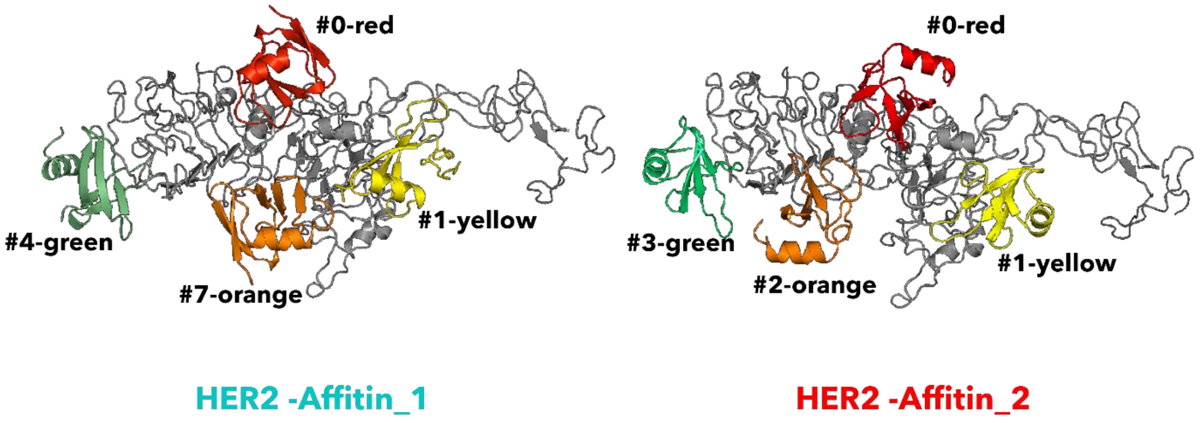


Figure 3.15 - Superimposition of the four HER2-Affitin_1 (left) and HER2-Affitin_2 (right) docking models selected for the DockQ and MLCE evaluations. HER2 is shown in grey. Affitins are coloured based on the four different areas: red, yellow, orange, and green. Affitins are also labelled with their ClusPro ranking, and the colour.

DockQ-MLCE evaluation of the docking models

The four models of each affitin (see *Figure 3.15*) were subjected to MD simulations performed as in **Section 3.2.1**, and the DockQ parameter was calculated along the trajectories.

In parallel, a MLCE calculation was performed on representative conformations of HER2. The docking models were compared with the patches by visual inspection and calculation of the number of HER2 residues belonging to a patch and involved in the docking solutions.

As pointed out in **Section 3.2.1**, the decision on which docking models are most likely should be made on the basis of the two combined approaches, i.e., possibly seeking a consensus between the two. It is important to remark that since DockQ and MLCE rely on totally different assumptions, their combined use can strengthen the conclusions that can be drawn. Moreover, it is worth nothing that the ClusPro score can be included as a third criterium to determine the most probable poses.

Table 3.12 shows, for the four models HER2-Affitin_1 and the four models HER2-Affitin_2, the DockQ values and the number of HER2 residues belonging to a patch that are at the same time involved in the binding of the affitin in the docking model.

Docking area	ClusPro ranking		DockQ		MLCE	
	Affitin_1	Affitin_2	Affitin_1	Affitin_2	Affitin_1	Affitin_2
Red	#0	#0	0.45	0.37	3	9
Yellow	#1	#1	0.27	0.29	27	27
Green	#4	#3	0.31	0.19	0	0
Orange	#7	#2	0.33	0.38	4	0

Table 3.12 – ClusPro ranking, DockQ values and number of HER2 residues belonging to a patch that are at the same time involved in the binding of the affitin in the docking model, for the four models HER2-Affitin_1 and HER2-Affitin_2.

Considering the HER2-Affitin_1 models, it can be seen that for three of the four poses, there is little or no overlap with the MLCE patches: model #0-red and model #7-orange match only 3 and

4 residues belonging to a patch, respectively; model #4-green has no match at all. Model #1-yellow totally overlaps to the patch (27 residues, in blue in *Figure 3.16*) instead, but it also has the lowest DockQ value overall (0.27). However, based on previous results[24], small differences (< 0.1) in DockQ values are not considered significant, as these values lie often in a narrow range, thus not being always remarkably useful for assessing the actual quality of the models. It was therefore concluded that model #1-yellow is the most likely, followed by model #0-red, which has the highest DockQ value (0.45) and is the first in the ClusPro ranking.

Similar considerations can be made for the HER2-Affitin_2 models. Models #2-orange and #3-green have no overlap with the patches; model #0-red shows only a partial match with a patch (9 residues). Model #1-yellow totally overlaps with a patch (27 residues, in blue in *Figure 3.16*), although it has the second lowest DockQ value in the series (0.29). As was stated for the HER2-Affitin_1 models, it can be concluded that also for the HER2-Affitin_2 pair, model #1-yellow might be the most probable, followed by model #0-red, which has the second highest DockQ value (0.37) and is the first one in the ClusPro ranking.

The most likely models for the HER2-Affitin_1 and HER2-Affitin_2 pairs, i.e., the #1-yellow models in both cases, show a high degree of overlap between the two affitins. The same is observed for the two second most likely models, i.e., models #0-red. The superimposition of these models is shown in *Figure 3.16*.

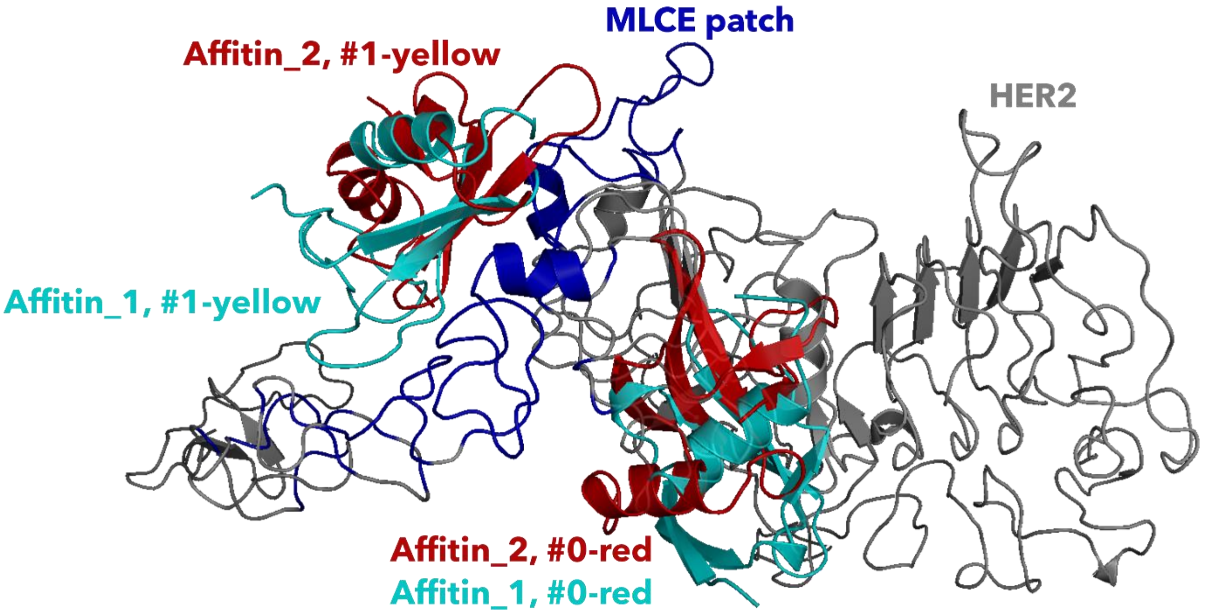


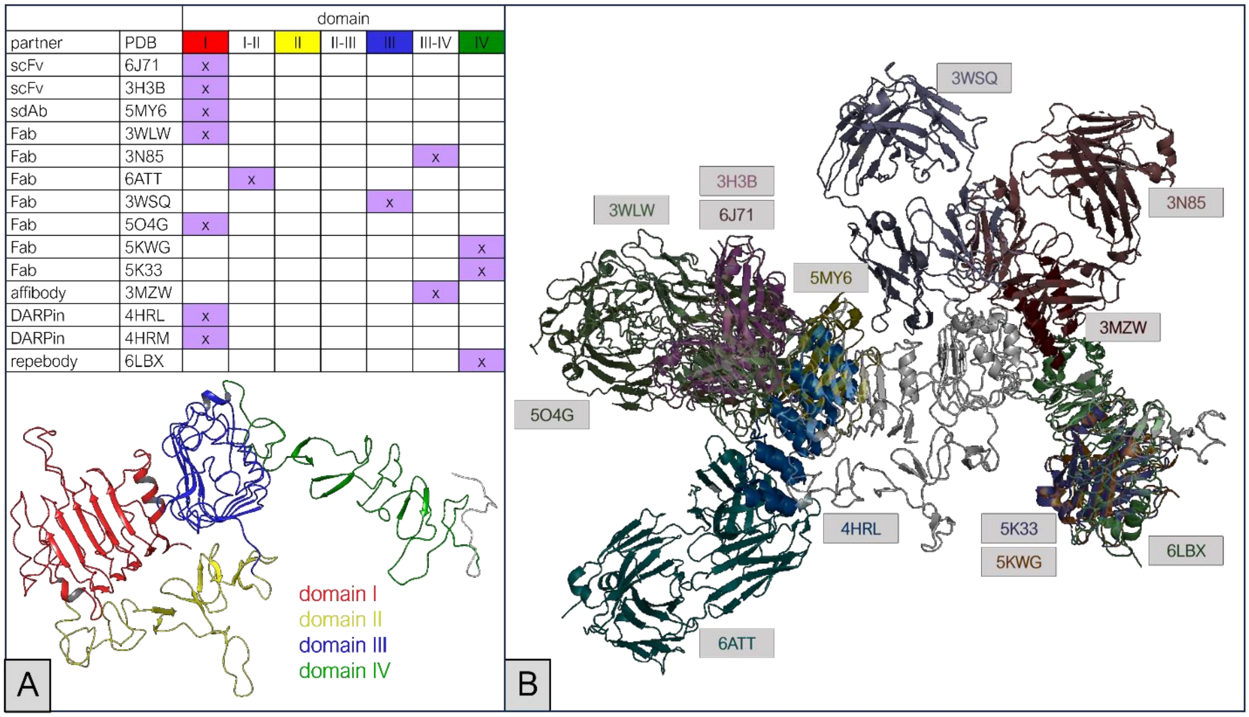
Figure 3.16 - Superimposition of HER2-Affitin_1 and HER2- Affitin_2 docking models #1-yellow and #0-red. HER2 is shown in grey, the MLCE patch in blue, Affitin_1 and Affitin_2 in cyan and red, respectively.

The overlap between the two #1-yellow models, and between the two #0-red models, is in accordance with the experimental evidence that Affitin_1 and Affitin_2 compete for the same binding site, i.e., epitope, on HER2 surface (Bracco S.p.A. internal communication).

Comparison of the docking models with the map of HER2 interactors

The result of a modelling procedure cannot be taken for granted. It can, however, be helpful to guide experimental tests, which are necessary to validate what is obtained through the *in silico* procedure. Among the experimental tests, competitive binding assays measure the binding affinity of a ligand towards a target, in presence of a different ligand whose binding mode to the same target is known.

In this perspective, the structures of complexes available in the PDB involving HER2 and different protein partners were collected; these were compared by superimposing the HER2 chains in the complexes considered onto each other. The resulting "map" of HER2 interactors is shown in *Figure 3.17*: 14 protein partners, including mAbs fragments (Fab, scFv, sdAb) and antibody mimetics bind HER2 to different, yet sometimes overlapping epitopes, mainly located in domains I, II, and IV.



		domain						
partner	PDB	I	I-II	II	II-III	III	III-IV	IV
scFv	6J71	x						
scFv	3H3B	x						
sdAb	5MY6	x						
Fab	3WLW	x						
Fab	3N85						x	
Fab	6ATT		x					
Fab	3WSQ					x		
Fab	5O4G	x						
Fab	5KWG							x
Fab	5K33							x
affibody	3MZW						x	
DARPin	4HRL	x						
DARPin	4HRM	x						
repebody	6LBX							x

*Figure 3.17 - **Panel A-top**: a table is shown listing the PDB IDs of complexes including HER2 and protein partners (complexes with mAbs Trastuzumab and Pertuzumab and their mutants are not considered), the category to which the protein partner belongs (Fab = antigen-binding fragments, scFv = single-chain variable fragments, sdAb = single domain antibodies), and the HER2 domains where the interactions occur. **Panel A-bottom**: HER2 domains are shown with different colours, as in the legend. **Panel B**: superimposition of the complexes listed in Panel A. HER2 is shown in grey, the protein partners with different colours.*

The aim was then to identify a subset of these protein partners to be exploited for competitive binding assays with Affitin_1 and Affitin_2.

HER2 structures in complexes with the protein partners were superimposed on HER2 structures in the most likely docking models, i.e., the #1-yellow and #0-red models. The eventual overlap between Affitin_1 and Affitin_2, and the known protein partners was then visually inspected.

Figure 3.18 shows that an overlap does exists between the docking models #1-yellow and the HER2 partners included in PDB entries 3MZW[17] and 3N85[60], of which the former is an affibody, i.e. , an antibody mimetic, and the latter is a Fab.

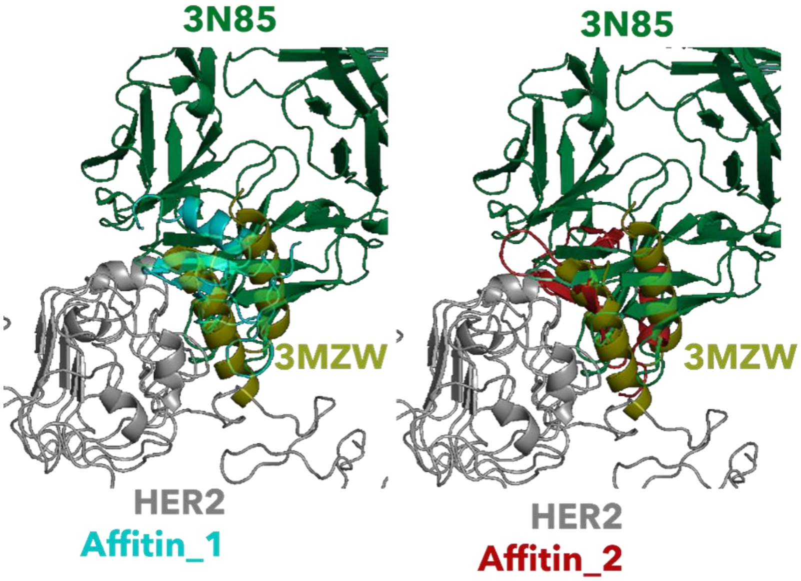


Figure 3.18 - Superimposition of the crystal structures 3N85 and 3MZW to the Afftin_1-HER2 docking model #1-yellow (left) and to the Afftin_2-HER2 docking model #1-yellow (right). HER2 is shown in grey, the affibody (HER2 partner in 3MZW) in yellow, the Fab (HER2 partner in 3N85) in green. Affitin_1 is shown in cyan and Affitin_2 in red.

The same comparison was carried out for #0-red models, as they were identified as the second most likely docking models. However, as shown in *Figure 3.19*, there is no overlap with the known HER2 partners. This would imply the need to use a different experimental approach to test whether these models represent the true binding modes, e.g., by introducing mutations, that would alter the binding, at the protein-protein interface.

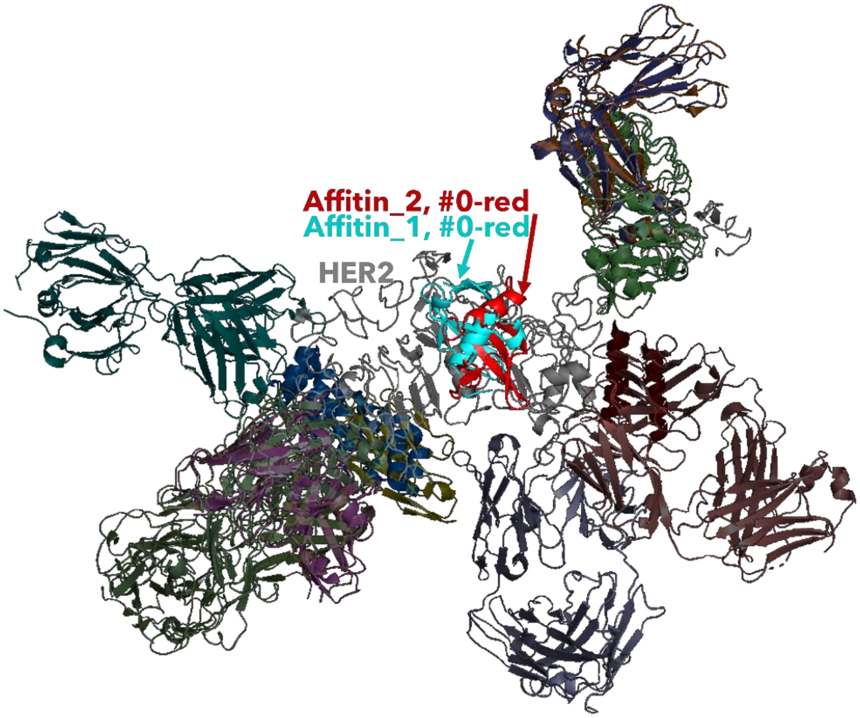


Figure 3.19 - Superposition of the crystal structures shown in Figure 3.17 onto the Afftin_1-HER2 docking model #0-red and the Afftin_2-HER2 docking model #0-red. HER2 is shown in grey, HER2 proteins partners with the remaining colours. Affitin_1 is shown in cyan and Affitin_2 in red.

Discussion

The main objective of the study was to make a reliable prediction of the structure of HER2-Affitin_1 and HER2-Affitin_2 complexes.

The preliminary study conducted on the dataset of known affitin-protein complexes (**Section 3.2.1**) served to define a procedure for the evaluation of docking models. This procedure combines DockQ and MLCE and was used in the present section to deal with the Affitins-HER2 use case.

Docking calculations were performed for the pairs of partners Affitin_1-HER2 and Affitin_2-HER2. First, the available experimental information on the binding interface was exploited. For both the affitins, the best scoring models of each of the four possible binding areas on the receptor were subjected to the combined DockQ-MLCE evaluation. The ClusPro score was also considered in the determination of the most likely docking models. The overall analysis led to a result that is in agreement with the available experimental information: Affitin_1 and Affitin_2 compete for the same epitope on HER2 receptor.

Conclusions

This section mainly focused on the problem of the prediction of the interaction between the HER2 receptor and two affitins, Affitin_1 and Affitin_2, which are the subject of two patents owned by Bracco S.p.A. These patents concern the use of the two affitins as molecular probes for the detection of HER2 during a therapy based on the two most currently used mAbs, i.e., Trastuzumab and Pertuzumab.

Affitins are antibody mimetics that, upon appropriate engineering of their sequence, can potentially recognize other receptors as well. For this reason, a study was conducted to find out whether the fold of the affitins depends on their sequence. Molecular dynamics simulations were performed for the two affitins covered by the patents, six mutated affitins retrieved from the Protein Data Bank, and five affitins whose sequences were designed by our research group. All the analysis showed that no significant changes in the overall fold occur and, most importantly, the β-sheet region involved in the mutations is not affected. This part of the study therefore showed that, in principle, it is possible to design affitins with any sequence to specifically recognize any protein partner.

The focus then shifted on the prediction of the structures of Affitin_1-HER2 and Affitin_2-HER2 complexes.

Prior to this, the need for a procedure capable of identifying the correct docking models in the pool of solutions provided by a docking programme was highlighted. A small dataset of complexes consisting of affitins and other protein partners whose three-dimensional structures are available in the PDB was used to test two approaches for the evaluation of docking models.

The first approach assesses the quality of a model based on its stability, which is quantified by the DockQ parameter, calculated along the MD trajectories. A higher DockQ indicates greater stability and thus a higher probability that the pose is correct. The second method uses MLCE to predict the binding sites (patches) on the isolated structure of proteins, in this case the partners of the affitins. A match between these MLCE-predicted patches and the binding sites present in the docking models is then checked. If a match is found, the pose is considered more likely. The combined use of DockQ and MLCE proved to be more effective in identifying the correct docking models, with respect to the use of the two approaches one at a time.

The prediction of the structures of HER2-Affitin_1 and HER2-Affitin_2 complexes was then addressed. The available experimental information was exploited to guide the docking calculations. The DockQ-MLCE approach, without neglecting the indications provided by the ClusPro scoring at the same time, was applied to determine the most likely docking models.

The overall modelling procedure revealed that Affitin_1 and Affitin_2 compete for the same HER2 epitope, which is in agreement with the experimental data. The most likely docking poses were compared with the experimental structures of complexes involving HER2 and protein partners available in the PDB. In this way, two known protein partners of the receptor, an affibody and a Fab, were identified as possible candidates for competitive binding assays. These experimental tests could assess whether the affitins actually bind HER2 on the predicted epitopes, which would also serve to validate the modelling procedure.

References

(1) Reitano, E.; Maiocchi, A.; Poggi, L.; Crivellin, F.; Huet, S.; Cinier, M.; Kitten, O. WO2021122726 - ANTI-HER2 POLYPEPTIDES DERIVATIVES AS NEW DIAGNOSTIC MOLECULAR PROBES. WO/2021/122726, **2021**.

(2) Reitano, E.; Maiocchi, A.; Poggi, L.; Crivellin, F.; Huet, S.; Cinier, M.; Kitten, O. WO2021122729 - ANTI-HER2 POLYPEPTIDES DERIVATIVES AS NEW DIAGNOSTIC MOLECULAR PROBES. WO/2021/122729, **2021**.

(3) Mankoff, D. A. A Definition of Molecular Imaging. *J. Nucl. Med.* **2007**, *48* (6), 18N, 21N.

(4) Chen, K.; Chen, X. Design and Development of Molecular Imaging Probes. *Curr. Top. Med. Chem.* **2010**, *10* (12), 1227–1236. https://doi.org/10.2174/156802610791384225.

(5) Ferlay, J.; Colombet, M.; Soerjomataram, I.; Parkin, D. M.; Piñeros, M.; Znaor, A.; Bray, F. Cancer Statistics for the Year 2020: An Overview. *Int. J. Cancer* **2021**, *149* (4), 778–789. https://doi.org/10.1002/ijc.33588.

(6) Meng, Q.; Li, Z. Molecular Imaging Probes for Diagnosis and Therapy Evaluation of Breast Cancer. *Int. J. Biomed. Imaging* **2013**, *2013*. https://doi.org/10.1155/2013/230487.

(7) Wang, J.; Xu, B. Targeted Therapeutic Options and Future Perspectives for HER2-Positive Breast Cancer. *Signal Transduct. Target. Ther.* **2019**, *4* (1), 34. https://doi.org/10.1038/s41392-019-0069-2.

(8) Slamon, D. J.; Leyland-Jones, B.; Shak, S.; Fuchs, H.; Paton, V.; Bajamonde, A.; Fleming, T.; Eiermann, W.; Wolter, J.; Pegram, M.; Baselga, J.; Norton, L. Use of Chemotherapy plus a Monoclonal Antibody against HER2 for Metastatic Breast Cancer That Overexpresses HER2. *N. Engl. J. Med.* **2001**, *344* (11), 783–792. https://doi.org/10.1056/NEJM200103153441101.

(9) Barthélémy, P.; Leblanc, J.; Goldbarg, V.; Wendling, F.; Kurtz, J.-E. Pertuzumab: Development beyond Breast Cancer. *Anticancer Res.* **2014**, *34* (4), 1483–1491.

(10) Nami, B.; Maadi, H.; Wang, Z. Mechanisms Underlying the Action and Synergism of Trastuzumab and Pertuzumab in Targeting HER2-Positive Breast Cancer. *Cancers (Basel).* **2018**, *10* (10), 342. https://doi.org/10.3390/cancers10100342.

(11) Cho, H. S.; Mason, K.; Ramyar, K. X.; Stanley, A. M.; Gabelli, S. B.; Denney, D. W.; Leahy, D. J. Structure of the Extracellular Region of HER2 Alone and in Complex with the Herceptin Fab. *Nature* **2003**, *421* (6924), 756–760. https://doi.org/10.1038/nature01392.

(12) Franklin, M. C.; Carey, K. D.; Vajdos, F. F.; Leahy, D. J.; de Vos, A. M.; Sliwkowski, M. X. Insights into ErbB Signaling from the Structure of the ErbB2-Pertuzumab Complex. *Cancer Cell* **2004**, *5* (4), 317–328. https://doi.org/10.1016/S1535-6108(04)00083-2.

(13) Berman, H. M. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28* (1), 235–242. https://doi.org/10.1093/nar/28.1.235.

(14) Hao, Y.; Yu, X.; Bai, Y.; McBride, H. J.; Huang, X. Cryo-EM Structure of HER2-Trastuzumab-Pertuzumab Complex. *PLoS One* **2019**, *14* (5), e0216095. https://doi.org/10.1371/journal.pone.0216095.

(15) Yu, X.; Yang, Y. P.; Dikici, E.; Deo, S. K.; Daunert, S. Beyond Antibodies as Binding Partners: The Role of Antibody Mimetics in Bioanalysis. *Annu. Rev. Anal. Chem.* **2017**, *10*, 293–320. https://doi.org/10.1146/annurev-anchem-061516-045205.

(16) Akbari, V.; Chou, C. P.; Abedi, D. New Insights into Affinity Proteins for HER2-Targeted Therapy: Beyond Trastuzumab. *Biochim. Biophys. Acta - Rev. Cancer* **2020**, *1874* (2), 188448. https://doi.org/10.1016/j.bbcan.2020.188448.

(17) Eigenbrot, C.; Ultsch, M.; Dubnovitsky, A.; Abrahmsen, L.; Hard, T. Structural Basis for High-Affinity HER2 Receptor Binding by an Engineered Protein. *Proc. Natl. Acad. Sci.* **2010**, *107* (34), 15039–15044. https://doi.org/10.1073/pnas.1005025107.

(18) Jost, C.; Schilling, J.; Tamaskovic, R.; Schwill, M.; Honegger, A.; Plückthun, A. Structural Basis for Eliciting a Cytotoxic Effect in HER2-Overexpressing Cancer Cells via Binding to the Extracellular Domain of HER2. *Structure* **2013**, *21* (11), 1979–1991. https://doi.org/10.1016/j.str.2013.08.020.

(19) Kim, T. Y.; Cha, J. S.; Kim, H.; Choi, Y.; Cho, H. S.; Kim, H. S. Computationally-Guided Design and Affinity Improvement of a Protein Binder Targeting a Specific Site on HER2. *Comput. Struct. Biotechnol. J.* **2021**, *19*, 1325–1334. https://doi.org/10.1016/j.csbj.2021.02.013.

(20) Kalichuk, V.; Béhar, G.; Renodon-Cornière, A.; Danovski, G.; Obal, G.; Barbet, J.; Mouratou, B.; Pecorari, F. The Archaeal "7 KDa DNA-Binding" Proteins: Extended Characterization of an Old Gifted Family. *Sci. Rep.* **2016**, *6* (1), 37274. https://doi.org/10.1038/srep37274.

(21) Robinson, H.; Gao, Y.-G.; McCrary, B. S.; Edmondson, S. P.; Shriver, J. W.; Wang, A. H.-J. The Hyperthermophile Chromosomal Protein Sac7d Sharply Kinks DNA. *Nature* **1998**,

392 (6672), 202–205. https://doi.org/10.1038/32455.

(22) Goux, M.; Becker, G.; Gorré, H.; Dammicco, S.; Desselle, A.; Egrise, D.; Leroi, N.; Lallemand, F.; Bahri, M. A.; Doumont, G.; Plenevaux, A.; Cinier, M.; Luxen, A. Nanofitin as a New Molecular-Imaging Agent for the Diagnosis of Epidermal Growth Factor Receptor Over-Expressing Tumors. *Bioconjug. Chem.* **2017**, *28* (9), 2361–2371. https://doi.org/10.1021/acs.bioconjchem.7b00374.

(23) Jacquot, P.; Muñoz-Garcia, J.; Fleury, M.; Cochonneau, D.; Gaussin, R.; Enouf, E.; Roze, C.; Ollivier, E.; Cinier, M.; Heymann, D. Engineering of a Bispecific Nanofitin with Immune Checkpoint Inhibitory Activity Conditioned by the Cross-Arm Binding to EGFR and PDL1. *Biomolecules* **2023**, *13* (4). https://doi.org/10.3390/biom13040636.

(24) Ranaudo, A.; Cosentino, U.; Greco, C.; Moro, G.; Bonardi, A.; Maiocchi, A.; Moroni, E. Evaluation of Docking Procedures Reliability in Affitins-Partners Interactions. *Front. Chem.* **2022**, *10* (December), 1–11. https://doi.org/10.3389/fchem.2022.1074249.

(25) Anfinsen, C. B. Principles That Govern the Folding of Protein Chains. *Science (80-.).* **1973**, *181* (4096), 223–230. https://doi.org/10.1126/science.181.4096.223.

(26) Dill, K. A.; MacCallum, J. L. The Protein-Folding Problem, 50 Years On. *Science (80-.).* **2012**, *338* (6110), 1042–1046. https://doi.org/10.1126/science.1219021.

(27) Correa, A.; Pacheco, S.; Mechaly, A. E.; Obal, G.; Béhar, G.; Mouratou, B.; Oppezzo, P.; Alzari, P. M.; Pecorari, F. Potent and Specific Inhibition of Glycosidases by Small Artificial Binding Proteins (Affitins). *PLoS One* **2014**, *9* (5), e97438. https://doi.org/10.1371/journal.pone.0097438.

(28) Kauke, M. J.; Traxlmayr, M. W.; Parker, J. A.; Kiefer, J. D.; Knihtila, R.; McGee, J.; Verdine, G.; Mattos, C.; Dane Wittrup, K. An Engineered Protein Antagonist of K-Ras/B-Raf Interaction. *Sci. Rep.* **2017**, *7* (1), 1–9. https://doi.org/10.1038/s41598-017-05889-7.

(29) Mukherjee, A.; Singh, R.; Udayan, S.; Biswas, S.; Reddy, P. P.; Manmadhan, S.; George, G.; Kumar, S.; Das, R.; Rao, B. M.; Gulyani, A. A Fyn Biosensor Reveals Pulsatile, Spatially Localized Kinase Activity and Signaling Crosstalk in Live Mammalian Cells. *Elife* **2020**, *9*, 1–48. https://doi.org/10.7554/eLife.50571.

(30) Zajc, C. U.; Dobersberger, M.; Schaffner, I.; Mlynek, G.; Pühringer, D.; Salzer, B.; Djinović-Carugo, K.; Steinberger, P.; De Sousa Linhares, A.; Yang, N. J.; Obinger, C.; Holter, W.; Traxlmayr, M. W.; Lehner, M. A Conformation-Specific ON-Switch for

Controlling CAR T Cells with an Orally Available Drug. *Proc. Natl. Acad. Sci.* **2020**, *117* (26), 14926–14935. https://doi.org/10.1073/pnas.1911154117.

(31) Madhavi Sastry, G.; Adzhigirey, M.; Day, T.; Annabhimoju, R.; Sherman, W. Protein and Ligand Preparation: Parameters, Protocols, and Influence on Virtual Screening Enrichments. *J. Comput. Aided. Mol. Des.* **2013**, *27* (3), 221–234. https://doi.org/10.1007/s10822-013-9644-8.

(32) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High Performance Molecular Simulations through Multi-Level Parallelism from Laptops to Supercomputers. *SoftwareX* **2015**, *1–2*, 19–25. https://doi.org/10.1016/j.softx.2015.06.001.

(33) Oostenbrink, C.; Villa, A.; Mark, A. E.; Van Gunsteren, W. F. A Biomolecular Force Field Based on the Free Enthalpy of Hydration and Solvation: The GROMOS Force-Field Parameter Sets 53A5 and 53A6. *J. Comput. Chem.* **2004**, *25* (13), 1656–1676. https://doi.org/10.1002/jcc.20090.

(34) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Hermans, J. Interaction Models for Water in Relation to Protein Hydration BT - Intermolecular Forces: Proceedings of the Fourteenth Jerusalem Symposium on Quantum Chemistry and Biochemistry Held in Jerusalem, Israel, April 13–16, 1981; Pullman, B., Ed.; Springer Netherlands: Dordrecht, 1981; pp 331–342. https://doi.org/10.1007/978-94-015-7658-1_21.

(35) Oostenbrink, C.; Soares, T. A.; van der Vegt, N. F. A.; van Gunsteren, W. F. Validation of the 53A6 GROMOS Force Field. *Eur. Biophys. J.* **2005**, *34* (4), 273–284. https://doi.org/10.1007/s00249-004-0448-6.

(36) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. LINCS: A Linear Constraint Solver for Molecular Simulations. *J. Comput. Chem.* **1997**, *18* (12), 1463–1472. https://doi.org/10.1002/(SICI)1096-987X(199709)18:12<1463::AID-JCC4>3.0.CO;2-H.

(37) Darden, T.; York, D.; Pedersen, L. Particle Mesh Ewald: An N·log(N) Method for Ewald Sums in Large Systems. *J. Chem. Phys.* **1993**, *98* (12), 10089–10092. https://doi.org/10.1063/1.464397.

(38) Berendsen, H. J. C.; Postma, J. P. M.; Van Gunsteren, W. F.; Dinola, A.; Haak, J. R. Molecular Dynamics with Coupling to an External Bath. *J. Chem. Phys.* **1984**, *81* (8), 3684–3690. https://doi.org/10.1063/1.448118.

(39) Bussi, G.; Donadio, D.; Parrinello, M. Canonical Sampling through Velocity Rescaling. *J. Chem. Phys.* **2007**, *126* (1). https://doi.org/10.1063/1.2408420.

(40) Parrinello, M.; Rahman, A. Polymorphic Transitions in Single Crystals: A New Molecular Dynamics Method. *J. Appl. Phys.* **1981**, *52* (12), 7182–7190. https://doi.org/10.1063/1.328693.

(41) Nosé, S.; Klein, M. L. Constant Pressure Molecular Dynamics for Molecular Systems. *Mol. Phys.* **1983**, *50* (5), 1055–1076. https://doi.org/10.1080/00268978300102851.

(42) Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E. Improved Side-Chain Torsion Potentials for the Amber Ff99SB Protein Force Field. *Proteins Struct. Funct. Bioinforma.* **2010**, *78* (8), 1950–1958. https://doi.org/10.1002/prot.22711.

(43) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983**, *79* (2), 926–935. https://doi.org/10.1063/1.445869.

(44) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual Molecular Dynamics. *J. Mol. Graph.* **1996**, *14* (1), 33–38. https://doi.org/10.1016/0263-7855(96)00018-5.

(45) Daura, X.; Gademann, K.; Jaun, B.; Seebach, D.; van Gunsteren, W. F.; Mark, A. E. Peptide Folding: When Simulation Meets Experiment. *Angew. Chemie Int. Ed.* **1999**, *38* (1–2), 236–240. https://doi.org/10.1002/(SICI)1521-3773(19990115)38:1/2<236::AID-ANIE236>3.0.CO;2-M.

(46) Kabsch, W.; Sander, C. Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features. *Biopolymers* **1983**, *22* (12), 2577–2637. https://doi.org/10.1002/bip.360221211.

(47) Jacobson, M. P.; Friesner, R. A.; Xiang, Z.; Honig, B. On the Role of the Crystal Environment in Determining Protein Side-Chain Conformations. *J. Mol. Biol.* **2002**, *320* (3), 597–608. https://doi.org/10.1016/S0022-2836(02)00470-9.

(48) Jacobson, M. P.; Pincus, D. L.; Rapp, C. S.; Day, T. J. F.; Honig, B.; Shaw, D. E.; Friesner, R. A. A Hierarchical Approach to All-Atom Protein Loop Prediction. *Proteins Struct. Funct. Bioinforma.* **2004**, *55* (2), 351–367. https://doi.org/10.1002/prot.10613.

(49) Kozakov, D.; Beglov, D.; Bohnuud, T.; Mottarella, S. E.; Xia, B.; Hall, D. R.; Vajda, S. How Good Is Automated Protein Docking? *Proteins Struct. Funct. Bioinforma.* **2013**, *81*

(12), 2159–2166. https://doi.org/10.1002/prot.24403.

(50) Kozakov, D.; Hall, D. R.; Xia, B.; Porter, K. A.; Padhorny, D.; Yueh, C.; Beglov, D.; Vajda, S. The ClusPro Web Server for Protein–Protein Docking. *Nat. Protoc.* **2017**, *12* (2), 255–278. https://doi.org/10.1038/nprot.2016.169.

(51) Vajda, S.; Yueh, C.; Beglov, D.; Bohnuud, T.; Mottarella, S. E.; Xia, B.; Hall, D. R.; Kozakov, D. New Additions to the ClusPro Server Motivated by CAPRI. *Proteins Struct. Funct. Bioinforma.* **2017**, *85* (3), 435–444. https://doi.org/10.1002/prot.25219.

(52) Desta, I. T.; Porter, K. A.; Xia, B.; Kozakov, D.; Vajda, S. Performance and Its Limits in Rigid Body Protein-Protein Docking. *Structure* **2020**, *28* (9), 1071-1081.e3. https://doi.org/10.1016/j.str.2020.06.006.

(53) Jandova, Z.; Vargiu, A. V.; Bonvin, A. M. J. J. Native or Non-Native Protein–Protein Docking Models? Molecular Dynamics to the Rescue. *J. Chem. Theory Comput.* **2021**, *17* (9), 5944–5954. https://doi.org/10.1021/acs.jctc.1c00336.

(54) Basu, S.; Wallner, B. DockQ: A Quality Measure for Protein-Protein Docking Models. *PLoS One* **2016**, *11* (8), 1–9. https://doi.org/10.1371/journal.pone.0161879.

(55) Méndez, R.; Leplae, R.; De Maria, L.; Wodak, S. J. Assessment of Blind Predictions of Protein–Protein Interactions: Current Status of Docking Methods. *Proteins Struct. Funct. Bioinforma.* **2003**, *52* (1), 51–67. https://doi.org/10.1002/prot.10393.

(56) Tiana, G. Understanding the Determinants of Stability and Folding of Small Globular Proteins from Their Energetics. *Protein Sci.* **2004**, *13* (1), 113–124. https://doi.org/10.1110/ps.03223804.

(57) Morra, G.; Colombo, G. Relationship between Energy Distribution and Fold Stability: Insights from Molecular Dynamics Simulations of Native and Mutant Proteins. *Proteins Struct. Funct. Genet.* **2008**, *72* (2), 660–672. https://doi.org/10.1002/prot.21963.

(58) Scarabelli, G.; Morra, G.; Colombo, G. Predicting Interaction Sites from the Energetics of Isolated Proteins: A New Approach to Epitope Mapping. *Biophys. J.* **2010**, *98* (9), 1966–1975. https://doi.org/10.1016/j.bpj.2010.01.014.

(59) Genoni, A.; Morra, G.; Colombo, G. Identification of Domains in Protein Structures from the Analysis of Intramolecular Interactions. *J. Phys. Chem. B* **2012**, *116* (10), 3331–3343. https://doi.org/10.1021/jp210568a.

(60) Fisher, R. D.; Ultsch, M.; Lingel, A.; Schaefer, G.; Shao, L.; Birtalan, S.; Sidhu, S. S.;

Eigenbrot, C. Structure of the Complex between HER2 and an Antibody Paratope Formed by Side Chains from Tryptophan and Serine. *J. Mol. Biol.* **2010**, *402* (1), 217–229. https://doi.org/10.1016/j.jmb.2010.07.027.

Section 4 – Protein-glycan docking protocol with HADDOCK3

The present section covers a project I have worked on when I was a visiting PhD student at the Computational Structural Biology group (Bijvoet Centre for Biomolecular Research, Universiteit Utrecht), under the supervision of Prof. Alexandre Bonvin and Dr. Marco Giulini.

Briefly, the present study aims to build a reliable protocol, based on the HADDOCK3 docking programme, developed at the CSB group, for the prediction of the structure of protein-glycan complexes.

4.1 – Introduction

Glycans are complex organic molecules formed by monosaccharides, simple sugar units, connected by glycosidic bonds. Based on the number of monosaccharides, they can be named disaccharides (2 units), oligosaccharides (3-10 units), or polysaccharides (more than 10 units). In the following text, the term glycan will be used for all these compounds, regardless of the number of sugar units.

Glycans structural complexity arises not only from the diverse nature of monosaccharides themselves[1], but also from how they connect to each other. Each glycosidic bond can form two possible stereoisomers at the anomeric carbon of one sugar, i.e., the carbon whose asymmetric centre is formed upon the cyclization of the monosaccharide Additionally, because of the several hydroxyl groups in sugars, regioisomers can exist. With the capability of forming multiple glycosidic bonds, monosaccharides can lead to branched chains, a feature that differentiates glycans from the linear structures that characterize peptides and oligonucleotides[2].

Glycans also show a great conformational variability at room temperature, thanks to the low free energy barriers between torsional angles around glycosidic bonds[2].

Glycans are universally present in living organisms, where they can either be linked to proteins or lipids, thus forming glycoproteins and glycolipids, respectively, or they can exist independently. Their biological roles[3] are manifold and fall into three broad groups:

1. Structural roles, e.g., they can contribute to the creation of external scaffolds like cell walls or they can be involved in protein folding.
2. Metabolic roles, i.e., they act for instance as energy reserves.
3. Informational roles[4]: glycans may interact with Glycan Binding Proteins (GBPs) to trigger various biological processes, both in plants and animals.

One notable example of glycans' importance is their role in the SARS-CoV-2 spike protein. This protein, which enters host cells by connecting to the angiotensin-converting enzyme (ACE2), is surrounded by a layer of glycans to hide from the immune system. A study has shown that specific sugars play a crucial role in the movement and structure of the part of the spike proteins that binds to ACE2[5]. The removal of these sugars results in diminished binding to ACE2, highlighting potential targets on the spike protein for vaccine design.

It is thus evident that understanding the way glycans interact with proteins is essential. As pointed out in **Section 1**, traditional experimental methods, like X-ray crystallography or Nuclear Magnetic Resonance, are not always feasible, besides being time-consuming and expensive. Computational techniques like molecular docking offer a more cost-effective and faster preliminary prediction of the three-dimensional structures of glycan-protein complexes. While progresses have been made[6], for example with the GlycanDock protocol developed within

Rosetta[7], glycan-protein docking isn't as advanced as other protocols aimed at the prediction of other biomolecular complexes. In fact, state-of-art protein-ligand docking software cannot properly address the conformational variability of glycans, as they are usually developed to deal with small, rigid molecules.

In this study, HADDOCK3[8,9] was used for addressing the glycan-protein interaction prediction problem.

This section is structured as follows.

In **Section 4.1**, general HADDOCK concepts are first introduced.

Information about the datasets employed for the study and the set-up of the docking calculations are then given, followed by the explanation of how the HADDOCK3 performance was evaluated (**Section 4.2**). In the last part of the section, it is shown how the sampling of glycans conformations was carried out within HADDOCK3 itself.

In the first part of the Results, the performance of HADDOCK3 on a dataset composed by 89 high-resolution experimental complexes (bound dataset), available in the Protein Data Bank (PDB)[10], is shown. The impact of the rigid body scoring function on the HADDOCK3 performance is evaluated, along with the overall performance on the bound dataset (**Section 4.3 – R1**). The dependence of the performance on glycan structural features and on the Ambiguous Interaction Restraints (AIRs) is also discussed (**R2**).

A protocol is then proposed for dealing with a realistic scenario, where the bound structures of the partners are unknown. The GLYCAM-web webserver[11,12] was used for the generation of glycans unbound structures, while protein ones were retrieved from the PDB. The whole of these structures constitutes the unbound dataset. HADDOCK3 performance is thus evaluated when dealing with the unbound dataset. Special emphasis is placed on the best way to select rigid body models for the refinement stage (**R3**).

In the last part of the study (**R5**), HADDOCK3 performance is assessed following the introduction of an ensemble of glycan structures generated through a short conformational sampling carried out within HADDOCK3 prior to the docking calculations (**R4**).

Finally, conclusions are drawn along with possible future developments (**Section 4.4**).

4.2 – Protocol

Docking with HADDOCK3 – general concepts

HADDOCK3 is the new, modular version of the well-established HADDOCK2.X software[13]. The original version of the program foresees three parameterizable steps: i) full randomization of the orientations of the two partners and rigid-body minimization ([rigidbody] module in HADDOCK3); ii) semi-flexible simulated annealing in torsion angle space ([flexref] module); iii) refinement in explicit solvent ([mdref] module). HADDOCK3 overcomes this rigid workflow structure as its constituent modules can be freely interchanged by the user. This allows to design protocols specific to the problem to be addressed.

HADDOCK scoring functions include terms accounting for electrostatic (E_{el}) and van der Waals (E_{vdW}) interactions (calculated with the OPLS force field[14]), for the energy associated with desolvation (E_{desolv})[15], for changes in buried surface area (E_{BSA}), and for the Ambiguous Interactions Restraints (E_{air}), which will be covered later. The coefficients with which these terms are weighted depend on the stage of the protocol.

In the present study, the scoring functions with the default coefficients were used (*Eq. 1*, *Eq. 3*). Moreover, a rigid body scoring function with an upweighted van der Waals energy term (1.0 instead of 0.01) was used too (*Eq. 2*), as it already proved to give better performance for small molecules (https://www.bonvinlab.org/software/haddock2.4/scoring/); it will be referred to with the label *vdW*, to be distinguished from the *default* one.

$$H\,(\text{rigidbody\_}\textit{default}) = 0.01E_{vdw} + 1.0E_{el} + 1.0E_{desolv} + 0.01E_{air} - 0.01E_{BSA} \qquad \text{(Eq. 1)}$$

$$H\,(\text{rigidbody\_}\textit{vdW}) = 1.0E_{vdw} + 1.0E_{el} + 1.0E_{desolv} + 0.01\ E_{air} - 0.01E_{BSA} \qquad \text{(Eq. 2)}$$

$$H\,(\text{flexref}) = 1.0E_{vdw} + 1.0E_{el} + 1.0E_{desolv} + 0.1E_{air} - 0.01E_{BSA} \qquad \text{(Eq. 3)}$$

A key feature of HADDOCK (2.X, 3) is the possibility of incorporating experimental data as restraints which are included in the energy function used for guiding the docking process. Ambiguous Interaction Restraints (AIRs) consist in a list of residues divided in two groups: active and passive. Active residues are of central importance for the interaction; they are thus restrained to be part of the interface throughout the docking and refinement processes, if possible, otherwise a scoring penalty is included. Passive residues could contribute to the interaction but are deemed of less importance; if such a residue does not belong to the interface there is no scoring penalty.

In this study, two scenarios in terms of AIRs were considered for each protein – glycan complex: i) true-interface scenario (**ti-aa**), where active residues, corresponding to the interface residues within 3.9 Å[16,17] from the partner, are defined for both the protein and the glycan; ii) true-interface-protein – full glycan passive scenario (**tip-ap**), where active residues are still defined for the protein interface, but all residues of the glycan are considered passive.

Dataset

HADDOCK3[8,9] performance in reproducing the binding geometries of glycan–protein complexes was evaluated by exploiting an adapted version of the dataset provided in GlycanDock[7]. This dataset is composed by 109 experimentally determined high-resolution (< 2.0 Å) protein-glycan complexes collected from the PDB.

The entries containing glycans not yet supported by HADDOCK had to be discarded. A dataset of 89 complexes was thus obtained, which will be referred to as *bound dataset* henceforth. The protein receptors in this dataset include 8 antibodies, 21 carbohydrate-binding modules, 18 enzymes, 27 lectins or glycan binding proteins (GBP) and 15 viral glycan binders. The length of the glycans ranges from 2 to 7 monosaccharides units; moreover, 72 of them are linear and 17 branched. This structural diversity will be considered in the analysis of the docking performance. With the aim of evaluating HADDOCK3 performance on unbound partners, the 55 out of 89 protein unbound structures available in the PDB were selected too. The corresponding unbound conformations of the glycans were generated with the GLYCAM-web webserver[11,12], using an in-house script to automate the process. This script consisted in building a URI, downloading the structures from the server, and subsequently converting the GLYCAM code for the carbohydrate residues to the residue names recognized by HADDOCK. The dataset composed by the 55 unbound conformations of the proteins and the glycans generated with GLYCAM-web webserver will be referred to as *unbound dataset* from now on. It contains 47 linear and 8 branched glycans; 25 glycans are composed by three or less monosaccharides units while 30 structures have more than three units. For the evaluation of HADDOCK3 performance on the *unbound dataset*, all the complexes including glycans made up by three or less monosaccharide units are treated together and referred to with the label **SL-SB**; for the *bound dataset*, linear (**SL**) and branched (**SB**) glycans are treated separately too. The complexes including glycans composed by more than three units (**L**) are divided into linear (**LL**) and branched (**LB**), which are of size 23 and 7 in the *unbound dataset*, respectively. The list of the complexes included in the *bound* and *unbound dataset* is shown in *Table 4.1*.

dataset	PDB ID bound	PDB ID unbound	Protein family	N	L/B	group	ref_Glycan_RMSD (Å)
bound	3OAU	-	Antibody	2	L	**SL**	-
bound	1WU6	-	Enzyme	2	L	**SL**	-
bound	3N17	-	Enzyme	2	L	**SL**	-
bound	1W6P	-	GBP	2	L	**SL**	-

bound	2IT6	-	GBP	2	L	**SL**	-
bound	3VV1	-	GBP	2	L	**SL**	-
bound	4R9F	-	GBP	2	L	**SL**	-
bound	5T4Z	-	Antibody	3	L	**SL**	-
bound	2XOM	-	CBM	3	L	**SL**	-
bound	4QPW	-	CBM	3	L	**SL**	-
bound	4D5I	-	Enzyme	3	L	**SL**	-
bound	2G7C	-	Viral	3	L	**SL**	-
bound	2YP3	-	Viral	3	L	**SL**	-
bound	5HZB	-	Viral	3	L	**SL**	-
bound	1UZ8	-	Antibody	3	B	**SB**	-
bound	1JPC	-	GBP	3	B	**SB**	-
bound	2WRA	-	GBP	3	B	**SB**	-
bound	5HZA	-	Viral	3	B	**SB**	-
bound	5V6F	-	Viral	3	B	**SB**	-
bound	6R3M	-	CBM	4	L	**LL**	-
bound	1JDC	-	Enzyme	4	L	**LL**	-
bound	3WH1	-	Enzyme	4	L	**LL**	-
bound	4YG0	-	Viral	4	L	**LL**	-
bound	1S3K	-	Antibody	4	B	**LB**	-
bound	1SL5	-	GBP	4	B	**LB**	-
bound	2I74	-	GBP	4	B	**LB**	-
bound	2CHB	-	Viral	4	B	**LB**	-
bound	6BE4	-	Antibody	5	L	**LL**	-
bound	1W8U	-	CBM	5	L	**LL**	-
bound	2WAB	-	Enzyme	5	L	**LL**	-
bound	2YP4	-	Viral	5	L	**LL**	-
bound	1GUI	-	CBM	6	L	**LL**	-
bound	1GWL	-	CBM	6	L	**LL**	-
bound	1GWM	-	CBM	6	L	**LL**	-
bound-unbound	6N35	6N32	Antibody	2	L	**SL**	0.50
bound-unbound	1C1L	1C1F	GBP	2	L	**SL**	0.18
bound-unbound	1I3H	1NLS	GBP	2	L	**SL**	0.17
bound-unbound	1KJL	5NFC	GBP	2	L	**SL**	0.18
bound-unbound	1PWB	3DBZ	GBP	2	L	**SL**	0.16
bound-unbound	1SLT	3W58	GBP	2	L	**SL**	0.20
bound-unbound	2RDK	2Z21	GBP	2	L	**SL**	0.95
bound-unbound	2ZKN	3W58	GBP	2	L	**SL**	0.14
bound-unbound	3G83	3DBZ	GBP	2	L	**SL**	0.18
bound-unbound	3P5H	3C22	GBP	2	L	**SL**	0.54
bound-unbound	5GAL	1BKZ	GBP	2	L	**SL**	0.20

bound-unbound	5YRG	5YRE	GBP	2	L	**SL**	0.39
bound-unbound	6H9Y	6H9W	Viral	2	L	**SL**	0.40
bound-unbound	2J1V	2J1R	CBM	3	L	**SL**	0.27
bound-unbound	2Y6G	2Y6H	CBM	3	L	**SL**	0.46
bound-unbound	154L	153L	Enzyme	3	L	**SL**	0.58
bound-unbound	3AOF	3AMC	Enzyme	3	L	**SL**	0.39
bound-unbound	5AWQ	5AWO	Enzyme	3	L	**SL**	0.41
bound-unbound	5JU9	5JTS	Enzyme	3	L	**SL**	0.49
bound-unbound	1QFO	1QFP	GBP	3	L	**SL**	0.36
bound-unbound	2VXJ	1L7L	GBP	3	L	**SL**	0.35
bound-unbound	3NV4	3NV1	GBP	3	L	**SL**	1.29
bound-unbound	4MBY	4MBX	Viral	3	L	**SL**	1.34
bound-unbound	6HA0	6H9W	Viral	3	L	**SL**	0.37
bound-unbound	3P5G	3C22	GBP	3	B	**SB**	0.65
bound-unbound	6MSY	6N32	Antibody	4	L	**LL**	1.64
bound-unbound	2J72	2J71	CBM	4	L	**LL**	0.72
bound-unbound	2J73	2J71	CBM	4	L	**LL**	2.17
bound-unbound	3ACH	3ACF	CBM	4	L	**LL**	0.69
bound-unbound	4XUR	4XUN	CBM	4	L	**LL**	0.52
bound-unbound	1KQZ	2HVM	Enzyme	4	L	**LL**	0.89
bound-unbound	1LMQ	1LMN	Enzyme	4	L	**LL**	0.56
bound-unbound	1UU6	1OLR	Enzyme	4	L	**LL**	1.76
bound-unbound	2BOF	2BOE	Enzyme	4	L	**LL**	0.92
bound-unbound	4DQJ	4DQ7	Enzyme	4	L	**LL**	0.78
bound-unbound	5GY0	5GXX	Enzyme	4	L	**LL**	0.36
bound-unbound	4YFZ	4YFW	Viral	4	L	**LL**	0.82
bound-unbound	2J1T	2J1R	CBM	4	B	**LB**	0.28
bound-unbound	2XJR	2XJQ	GBP	4	B	**LB**	1.60
bound-unbound	3ZWE	3ZW0	GBP	4	B	**LB**	1.36
bound-unbound	2Z8L	1M4V	Viral	4	B	**LB**	1.02
bound-unbound	1GNY	1US3	CBM	5	L	**LL**	0.68
bound-unbound	1OF4	1OF3	CBM	5	L	**LL**	0.69
bound-unbound	1UXX	1GMM	CBM	5	L	**LL**	1.26
bound-unbound	2ZEX	2ZEW	CBM	5	L	**LL**	0.65
bound-unbound	3OEB	2ZEW	CBM	5	L	**LL**	0.55
bound-unbound	1KQY	2HVM	Enzyme	5	L	**LL**	1.10
bound-unbound	5VX5	5VX4	Viral	5	L	**LL**	1.34
bound-unbound	5VX9	5VX8	Viral	5	L	**LL**	1.05
bound-unbound	3AP9	3AP5	GBP	5	B	**LB**	0.87
bound-unbound	6UG7	6UGA	Antibody	6	L	**LL**	3.15
bound-unbound	1PMH	1PMJ	CBM	6	L	**LL**	1.21

bound-unbound	4HK8	4HKO	Enzyme	6	L	**LL**	1.55
bound-unbound	1OH4	1OF3	CBM	7	B	**LB**	1.93
bound-unbound	2VUZ	2VUV	GBP	7	B	**LB**	4.53

*Table 4.1 – Composition of the datasets. The following information are shown: whether the entry is part of the bound dataset only or of both bound and unbound, the PDB ID of the complex, the PDB ID of the unbound protein, the family to which the proteins belong, the number (N) of monosaccharide units the glycans are composed of, if the glycans are linear (**L**) or branched (**B**), the group (**SL**, **SB**, **LL**, **LB**) the complexes are assigned to for the analysis, the ref_Glycan_RMSD (see below).*

In *Figure 4.1*, three protein-glycan complexes included in the study are shown as example.

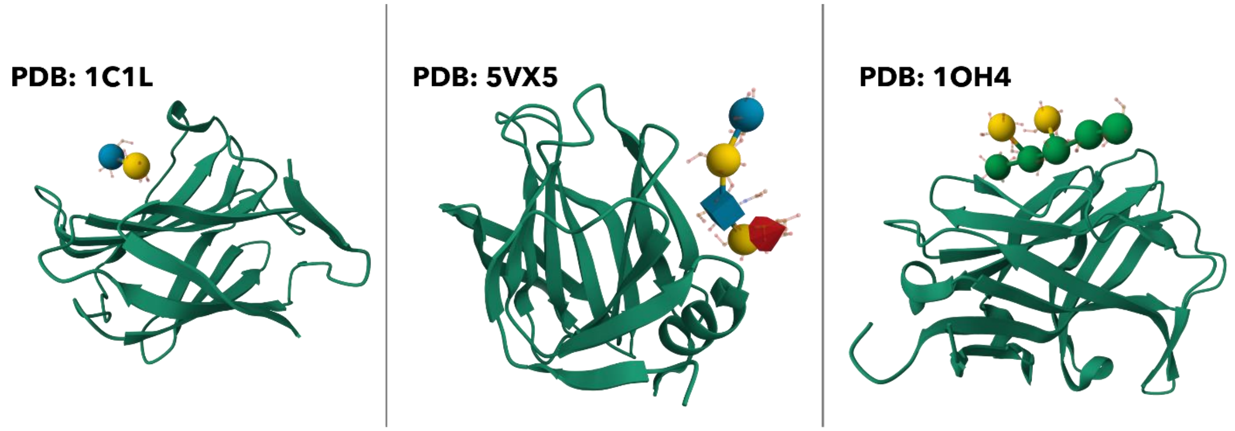


Figure 4.1 – Representation of three complexes object of the study. Proteins are shown as cartoon, glycans with the SNFG representation. Images from the Protein Data Bank.

Glycans and proteins structures were prepared using the programme pdb-tools[18]. Heteroatoms such as water molecules, cofactors and ions were removed, when not part of the protein-glycan interface. The residues were renumbered to start from 1, with pdb_reres, and the chains ID were modified to chain A and chain B for the receptor (protein) and the ligand (glycan), respectively, with pdb_chain. In some cases where the receptor consists of more than one chain, these were merged into a single chain, and the residue numbering was shifted if needed to avoid overlap in numbering.

Set-up of the calculations

Two different protocols were employed for the *bound* and *unbound datasets*.

For the *bound dataset* the following steps were performed (see *Table 4.2*): 1) creation of the topologies of the two partners; 2) generation of rigid body models, driven by AIRs; 3) evaluation of the quality of the models, through the calculation of the interface-ligand-root-mean-square deviation (IL-RMSD, see section "Evaluation of HADDOCK3 performance"), with respect to the reference structure; 4a) calculation of the RMSD matrix between all the models, based on either all the interface residues (when **ti-aa** AIRs are used) or the protein interface residues and the whole glycan (when **tip-ap** AIRs are used); 4b) clustering of the models based on the RMSD matrix; 5) cluster-based evaluation of the quality of the models.

Protocol bound dataset		
Step	Module	Parameters
1	[topoaa]	

2	[rigidbody]	sampling = 1000;
		w_vdw= 0.01 (default), 1.0 (vdW);
		ambig_fname = /path/to/tbl/file
		randremoval = false (**SL-SB** glycans only)

3	[caprieval]	reference_fname = /path/to/reference/pdb
4a	[rmsdmatrix]	resdic_A = [interface residues of the protein]; resdic_B = [interface (**ti-aa**) or all (**tip-ap**) residues of the glycan]
4b	[clustrmsd]	criterion = distance; linkage = average; threshold = 4; tolerance = 2.5
5	[caprieval]	reference_fname = /path/to/reference/pdb

Table 4.2 - Modules and parameters employed for the docking of the bound dataset.

Concerning the docking of the *unbound* partners, (see *Table 4.3*), the procedure was the same used for the *bound dataset* for steps 1) and 2) with the difference that a higher number of rigid body models is generated at step 2) when performing docking with an ensemble of glycans conformations. An evaluation of the quality of the rigid body models was then performed both on the 200 (400 in the ensemble scenario) best scoring models, ranked individually (step 3), and on the top 5 models of 50 (150 in the ensemble scenario) clusters (step 4 and 5). This double evaluation showed that a cluster-based selection of rigid body models allows to retain a higher number of good quality models, as it will be shown in section **R3**. This subset of models was then refined through short Molecular Dynamics (MD) simulations in explicit water (step 6), where all the residues except the ones at the interface are constrained to their initial coordinates. The quality of the refined models was then evaluated, both on single structures (step 7) and in a cluster-based manner (steps 8 and 9).

Protocol unbound dataset		
Step	Module	Parameters
1	[topoaa]	

2	[rigidbody]	sampling = 1000, 4000 (ensemble);
		w_vdw= 1.0 (vdW);
		ambig_fname = /path/to/tbl/file
		randremoval = false (**SL-SB** glycans only)

3	[caprieval]	reference_fname = /path/to/reference/pdb
4a	[rmsdmatrix]	resdic_A = [interface residues of the protein]; resdic_B = [interface // all residues of the glycan]
4b	[clustrmsd]	criterion = maxclust; tolerance = 50, 150 (ensemble)
4c	[seletopclusts]	top_models = 5
5	[caprieval]	reference_fname = /path/to/reference/pdb

6	[flexref]	tolerance = 5; nemsteps = 200, 400 (extended); mdsteps_rigid = 500, 1000 (extended); mdsteps_cool1 = 500, 1000 (extended); mdsteps_cool2 = 1000, 2000 (extended); mdsteps_cool3 = 1000, 2000 (extended);
		ambig_fname = /path/to/tbl/file

		randremoval = false (**SL-SB** glycans only)
7	[caprieval]	reference_fname = /path/to/reference/pdb
8a	[rmsdmatrix]	resdic_A = [interface residues of the protein]; resdic_B = [interface // all residues of the glycan]
8b	[clustrmsd]	criterion = distance; linkage = average; threshold = 4; tolerance = 2.5
9	[caprieval]	reference_fname = /path/to/reference/pdb

Table 4.3 - Modules and parameters employed for the docking of the unbound dataset.

Evaluation of HADDOCK3 performance

The quality of the models was evaluated with respect to the experimental structures through the calculation of the IL-RMSD, in which the model is first superimposed to the reference structure using the backbone atoms of the protein interface residues; the RMSD is then calculated only on the heavy atoms of the oligosaccharide. The choice of using this parameter is motivated by the fact that the protein interface is larger compared to the glycan interface. The calculation of IL-RMSD thus enables to identify the variations in the position of the ligand, as opposed to what happens e.g., with the calculation of the interface-RMSD (I-RMSD), one of the standard CAPRI parameters[19]. The cut-offs for IL-RMSD, inspired by the cut-offs for the I-RMSD and ligand-RMSD (L-RMSD) according to the CAPRI criteria for oligosaccharides[20], are as follows: high-quality models: IL-RMSD ≤ 1.0 Å; medium-quality models: IL-RMSD ≤ 3.0 Å; acceptable-quality models: IL-RMSD ≤ 6.0 Å.

HADDOCK3 performance is evaluated through the calculation of success rates (SR), that is, the fraction of complexes having at least one high-, medium-, or acceptable-quality model among a number of models, ranked according to the HADDOCK score. SR is also calculated on the clustered models, considering only the top scoring models within each cluster.

Glycans conformational sampling

Conformational sampling of the glycans was carried out with the HADDOCK3 water refinement module [mdref], starting from the structures generated with the GLYCAM-web webserver. For 11 out of 55 glycans more than one conformation was generated by the webserver; in such a case one of the conformations was randomly selected. The sampling and analysis protocol involved the use of the following HADDOCK3 modules: [topoaa], [mdref], [rmsdmatrix], [clustrmsd]. After the creation of the topology (module [topoaa]), conformational sampling in water was performed with the water refinement module ([mdref]), defining the glycans as fully flexible (parameters: nfle1 = 1; fle_sta_1_1 = 1; fle_end_1_1 = 7). At this stage, different scenarios were tested in terms of number of steps and number of models, as specified in *Table 4.4*. Three scenarios were run on 100 models with increasing the number of steps, thus the simulation time (sf100-x1, sf100-x8, sf100-x16). Then, the number of models (parameter sampling_factor) was increased to 400 while the simulation time was the same of two of the scenarios previously listed (sf400-x1 and sf400-x16). The overall time of the simulations ranges from 0.37 ns (sf100-x1) to 22.48 ns (sf400-x16).

scenario_name	Sampling_factor	waterheatsteps	watersteps	watercoolsteps	Total simulation time (ns)
Default [mdref] values	1	100	1250	500	
sf100-x1	100	100	1250	500	0.37
sf100-x8	100	100	10000	4000	2.82
sf100-x16	100	100	20000	8000	5.62
sf400-x1	400	100	1250	500	1.48
sf400-x16	400	100	20000	8000	22.48

*Table 4.4 - Glycans conformational sampling scenarios. In blue are highlighted the values that change with respect to the scenario sf100-x1, which is all default parameters but sampling_factor. Total simulation time (TSM) is calculated as follows. TSM (ns) = (sampling_factor * total n steps * timestep (ps))/1000. Where: total n steps = waterheatsteps + watersteps + watercoolsteps and timestep = 0.002 ps.*

For assessing the conformational variability over the sampled trajectory, the RMSD of atoms C1, O4, C4, C5 (*Figure 4.2*) was calculated for all the generated conformations after fitting on the same atoms on the bound conformations. This value will be called Glycan_RMSD from now on. The program ProFit v3.3[21,22] was used for this purpose.

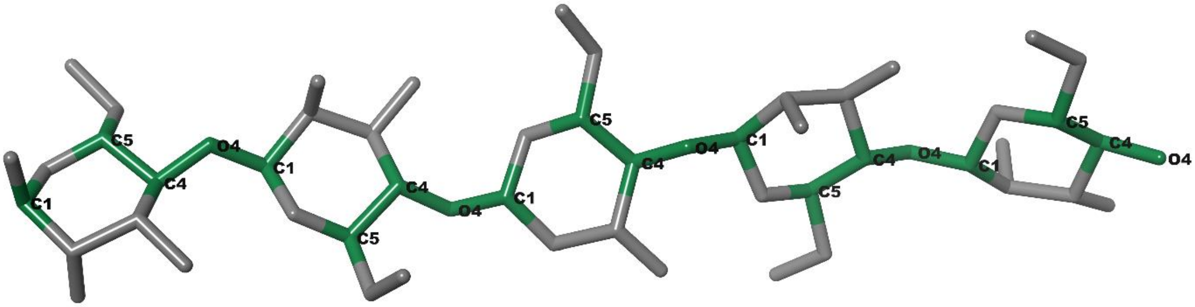


Figure 4.2 - Sticks representation of the glycan included in complex 3OEB. Atoms C1, O4, C4, C5 used for the fitting and the calculation of Glycan_RMSD are labelled and shown in green.

Distributions of the Glycan_RMSD values of the conformations obtained with the sampling procedures were plotted together with the Glycan_RMSD calculated for the webserver-generated conformations; the latter value will be called ref_Glycan_RMSD from now on.

The sampled conformations were clustered using RMSD-based hierarchical clustering[23,24]. The RMSD matrix between all the conformations generated was calculated with the module [rmsdmatrix], by specifying, through the parameter ‘resdic_’, the residues to be considered for the alignment and the RMSD calculation. The module [clustrmsd] was then exploited for clustering the conformations, with the following parameters: criterion = maxclust, linkage = average, tolerance = 10 (or 20). The 'maxclust' criterion clusters the structure in such a way to give a fixed number of clusters, defined by the parameter ‘tolerance’. The linkage governs the way clusters are merged in the creation of the dendrogram, i.e., it defines the method for calculating the distance between the newly formed cluster and each object which does not belong to a cluster yet. As it was done for all the sampled conformations, Glycan_RMSD values were calculated for the clusters centers, i.e. the points having the lower distance to all the other points in the cluster. Values of Glycan_RMSD corresponding to the cluster centers were plotted together with the overall sampling distribution to assess whether the clustering can capture the models that are closer to the glycan experimental structure.

The centres of the 20 clusters were then employed as an ensemble for new docking calculations.

4.3 – Results

This section is structured as follows. First, the impact of the rigid body scoring function on HADDOCK3 performance is discussed exploiting docking calculations performed on the *bound dataset* (paragraph **R1**). Considerations will be then made regarding the dependence of the performance on glycan structural features and on AIRs (paragraph **R2**).

Then, the performance on the *unbound dataset* is assessed, with a focus on which is the best way for selecting the rigid body models to be refined (paragraph **R3**). Conformational sampling of the glycans is then presented (paragraph **R4**). At last, the effect and the limitations of using the glycans ensemble is discussed (paragraph **R5**).

R1: Impact of the rigid body scoring function & performance on bound dataset

The first point that needed to be addressed was if the *vdW* rigid body scoring function performs better than the *default* one, as it happens for protein-small molecules docking. This was assessed by running docking calculations on the *bound dataset*, and with **ti-aa** AIRs. For complexes including glycans composed by three or less monosaccharide units, the AIRs were not randomly removed for each generated model (the parameter randremoval is set to false). A comparison of success rates (SR) obtained with the *default* (*Eq. 1*) and *vdW* (*Eq. 2*) scoring functions is shown in *Figure 4.3*.

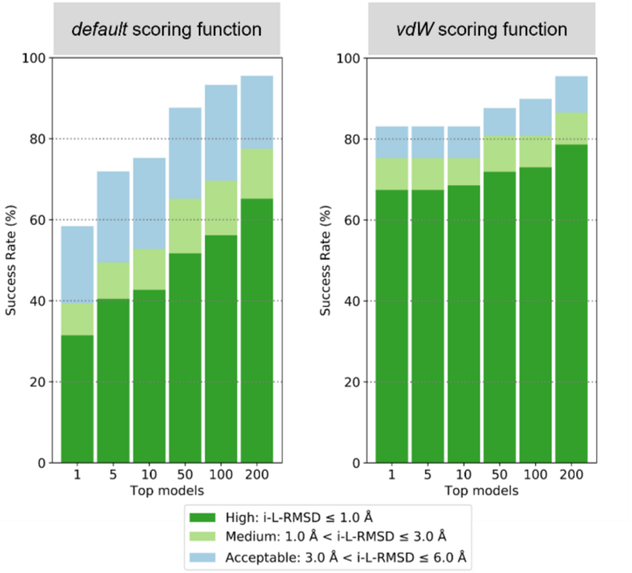


Figure 4.3 - Comparison of SR, calculated for the top (T) 1, 5, 10, 50, 100, and 200 models for the bound dataset, ***ti-aa*** *AIRs. Default and vdW (w_vdw = 1.0) scoring functions (see Eq. 1 and Eq. 2) are shown on the left and on the right, respectively.*

The *vdW* scoring function performs much better than the *default* one: SR is remarkably higher when the former is used, i.e., when a higher weight (1.00 instead of 0.01) is given to the van der Waals energy term. For example, considering top (T) 1 and T10 high-quality models, the difference SR_{vdw} – $SR_{default}$ is around 36% and 26%, respectively. The better performance given by *vdW* scoring function is due to the hydrophobicity of glycans; this was not unexpected as a similar behaviour was obtained in previous work involving smaller molecules than proteins, e.g., when docking cycling peptides[25] and small ligand in general.

Given these results, all the docking calculations discussed from now on are performed with the *vdW* scoring function.

R2: Dependence of the performance on glycans structural features and on AIRs

The dependency of the structural features of the glycans on the SR was then assessed. SR were thus calculated again, for the scenario **ti-aa** AIRs and with the *vdW* scoring function, grouping the complexes based on the size of the glycans (glycans composed by three or less units, labelled with **S**, or by more than three, labelled with **L**) and connectivity (linear, **L**, or branched, **B**). The SR, shown in *Figure 4.4*, indicate that HADDOCK3 performance was the best for **LL** glycans. For example, the SR for T1 high quality models is around 60%, 30%, 80%, and 70% for **SL**, **SB**, **LL**, and **LB** glycans, respectively. Considering a higher number of models (T50-T200), SR is almost the same for all the complexes but the ones involving **SB** glycans. The overall worse performance on those glycans, and, to a less extent on the **SL** ones, can be explained with two considerations. First, docking a smaller ligand could be more difficult with respect to a larger one, because the former must satisfy a lower number of spatial requirements. In other words, a smaller ligand could be docked to the protein partner with a greater variability of positions and orientations; the same could not hold true for a larger ligand, whose dimensions could force it to adopt a specific disposition with respect to the partner. The second consideration is that almost all (36/38) **SL** glycans bind the partner with all their residues, i.e., all the residues belong to the interface, while this happens for only half (3/6) among **SB** glycans. As the interface residues are restrained with AIRs to be part of the interface during the docking and refinement processes, this could explain why performance on the branched glycans is worse than the one on the linear of the same size.

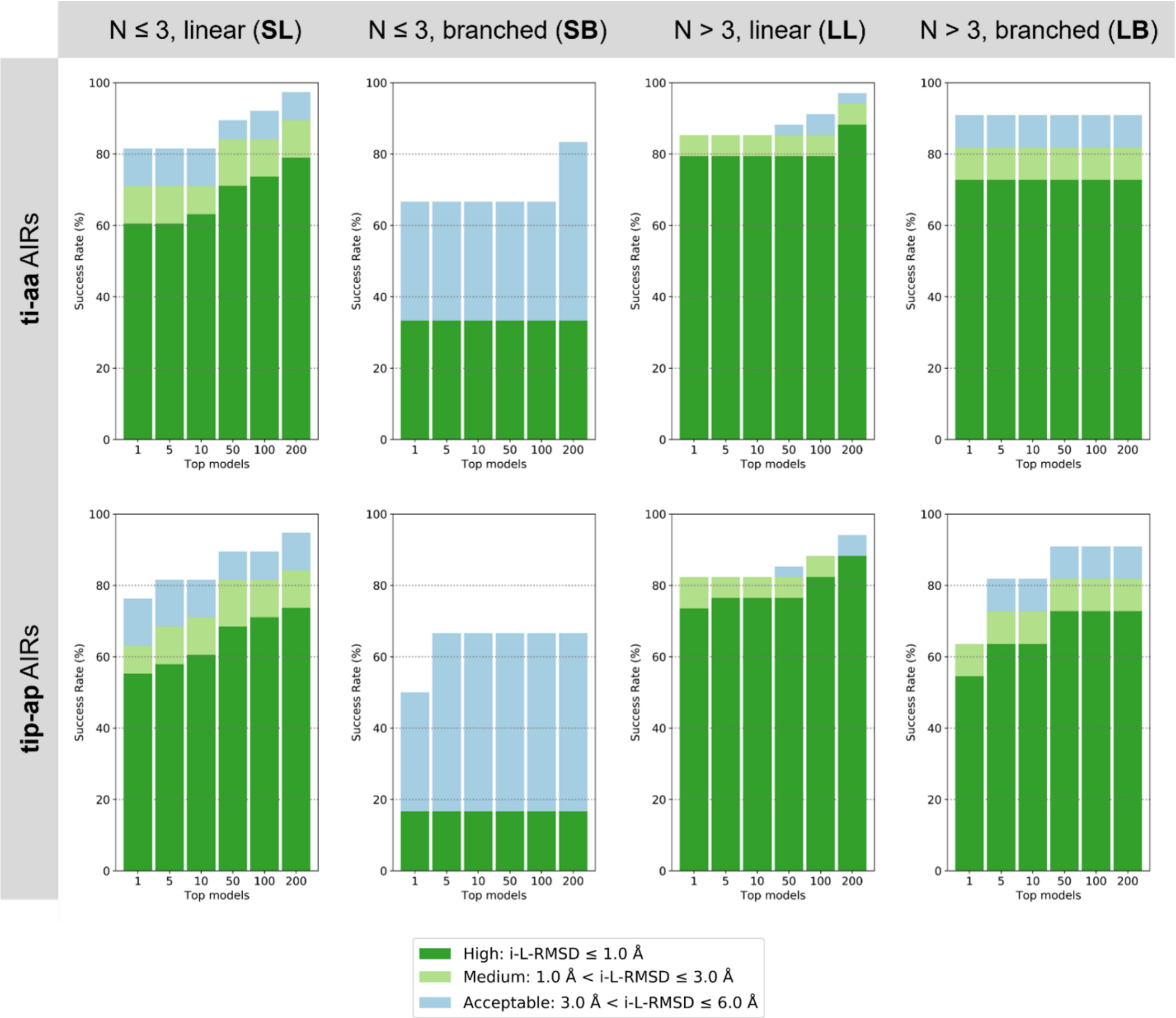


*Figure 4.4 - Performance of HADDOCK3 on bound dataset, based on glycans size (glycans composed by three or less monosaccharide units, **S**, or by more than three, **L**) and connectivity (linear, **L**, or branched, **B**). vdW scoring function is used. SR are also compared between docking runs performed with **ti-aa** and **tip-ap** AIRs, shown on the first and second row, respectively. SR are calculated for the top 1, 5, 10, 50, 100, and 200 models.*

Similar considerations can be helpful in discussing the performance obtained with **tip-ap** AIRs (second row of *Figure 4.4*), when no information is given about the glycans interface. The effect of using **tip-ap** AIRs is more pronounced for branched glycans than for linear ones. For example, concerning T1 high quality models, the difference SR_{tip-ap} - SR_{ti-aa} are around -5% and more than -15% for linear and branched glycans respectively, independently on their size. In the case of longer glycans also, most of the time (27/34) **LL** glycans have all the residues involved in the binding of the partner, whereas the same is true for only half (5/11) of the **LB** ones. This could be an explanation for the reason why docking branched glycans, independently on their size, is more difficult when no information about their interface is available.

However, HADDOCK3 performance on protein-glycan complexes is not dramatically affected using **tip-ap** AIRs. Considering that experimental interface information for the glycans is rarely available in a realistic scenario and that performance with the **ti-aa** scenario can only be better, in the next sections docking calculations with **tip-ap** AIRs will be mainly discussed.

R3: Performance on unbound dataset & how to select models for the refinement stage

In a realistic scenario the bound conformations of the docking partners are unknown and conformational rearrangements could occur in the binding process. For these reasons, HADDOCK3 performance was assessed also on the *unbound dataset* (see Methods). When dealing with unbound structures, the flexible refinement stage of the rigid body models plays a fundamental role, as it allows the partners to adapt to each other.

The problem that needed to be addressed before performing this stage was how to choose a subset of rigid body models to refine, namely whether to pick the first 200 models ranked by their HADDOCK score or the 5 best models of the first 50 clusters. These SR are shown, divided by glycan size and linearity, in *Figure 4.5*.

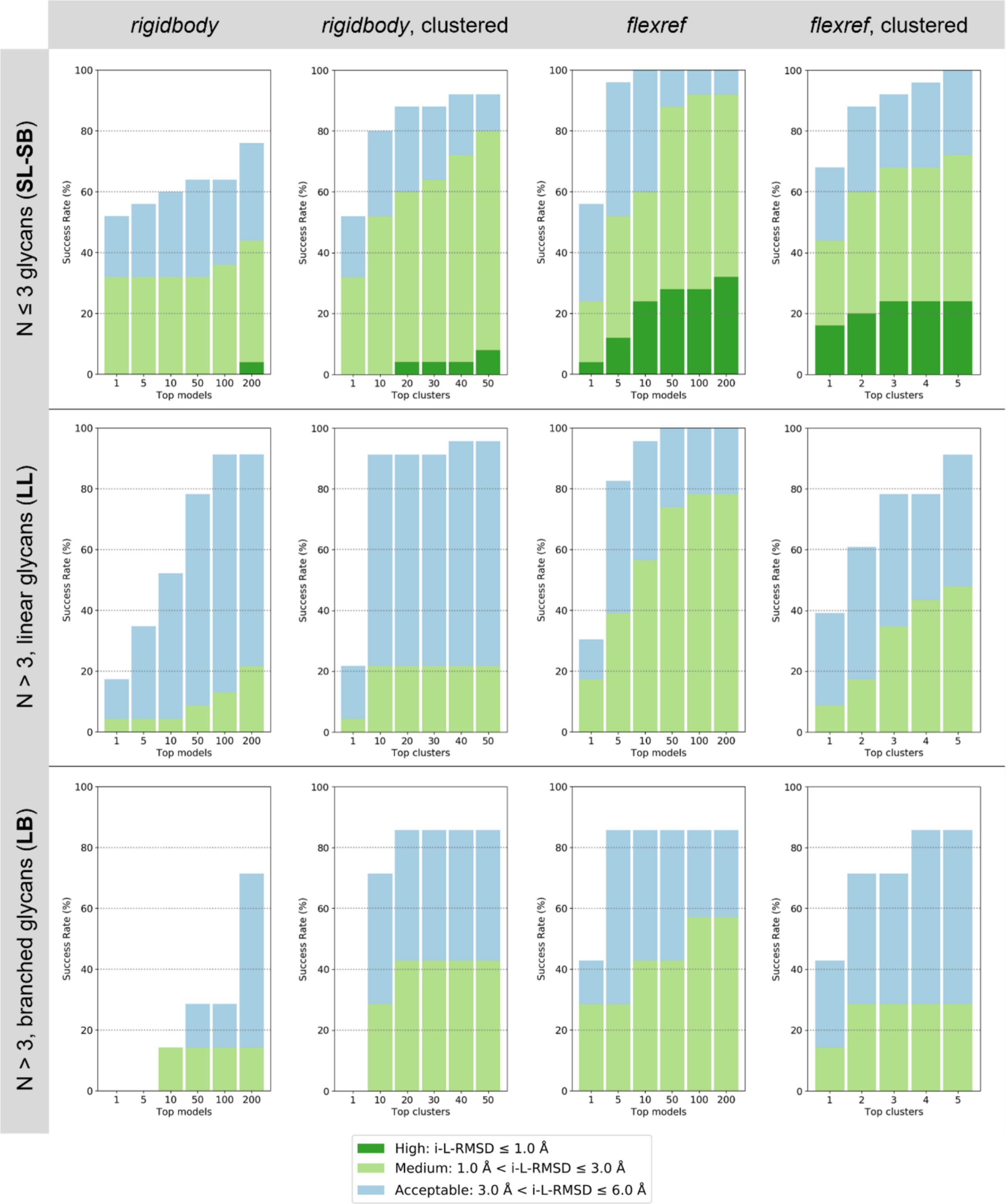


Figure 4.5 - HADDOCK3 performance on unbound dataset, using vdW scoring function and ***tip-ap*** *AIRs. SR are calculated: on the T1 to T200 rigidbody models (first column) and on T1 to T50 rigid body clusters, considering only the top 5 models of each cluster (second column); on the T1 to T200 refined models (third column) and on T1 to T5 refined clusters, considering only the top 4 models of each cluster (fourth column). Models are refined according to the cluster-based selection (second column). SR is calculated separately for the three categories of complexes grouped by glycans size and connectivity:* ***SL-SB*** *(top row),* ***LL*** *(middle row), and* ***LB*** *(bottom row).*

The comparison of the SR obtained on the single models (first column of *Figure 4.5*) with those obtained on the clustered models (second column of *Figure 4.5*), allows to observe that, overall, the selection of the clustered models is beneficial to the docking SR. More specifically, considering **SL** and **SB** glycans (*Figure 4.5*, first row), selecting the 200 best scoring rigid body models ranked singularly ("*rigidbody*" column) results in a 44% of medium-quality SR. On the other hand, if the top 5 models of the first 50 clusters are selected, medium-quality SR increases to 80% ("*rigidbody*, clustered" column). This results also in a slightly higher recovery of high-quality models (8% instead of 4%) and in almost the totality (92%) recovery of acceptable models. Therefore, for this group of glycans, selecting the rigid body models after clustering is by far the best way to choose the structures for the refinement stage. This is likely a consequence of the high similarity of the rigid body models ranked in the best positions: more diverse poses are recovered with the clustering.

A similar behaviour is observed for **LB** glycans (third row of *Figure 4.5*). Selecting the 200 best scoring rigid body models results in the retrieval of only 14% of medium quality models. If the top 5 models of the 50 clusters are selected instead, SR for medium quality models is around 43%. In this way, a few more acceptable-quality models are retrieved too.

Concerning the **LL** glycans (second row of *Figure 4.5*), clustering the rigid body models does not seem to be significantly useful. SR for medium quality models is around 20% both for T200 models and T50 clusters. The SR for acceptable quality models does not change significantly either. However, by comparing the SR for T50 medium-quality models (< 10%) with SR for T10 medium-quality clusters (> 40%), the latter consisting of around 50 models too, it can be seen how clustering is convenient, especially when limited computational resources limit the number of structures that can be refined.

Overall, selecting rigid body models after clustering is more convenient than using the standard, HADDOCK Score-based selection, as it allows the retrieval of a higher number of high-/ medium-quality models.

Flexible refinement was thus performed on 50 clusters, each of them composed by 5 or less models. SR was then calculated on the refined models, both on single structures (third column "*flexref*" of *Figure 4.5*) and cluster-based (fourth column, "*flexref*, clustered" of *Figure 4.5*).

The introduction of the flexibility at the interface region strongly improves the quality of the models. Considering all the refined models, i.e., comparing the SR of the T50 rigid body clusters (second column of *Figure 4.5*) with the SR of the T200 refined models (third column of *Figure 4.5*), we can observe how, for **SL-SB** glycans, high-quality SR increases from 8% to 32%, medium-quality SR increases from 80% to 92%, and the totality of the models falls within the acceptable quality cut-off. As for **LL** glycans, medium-quality SR increases from around 20% to almost 80%; all the models are of acceptable quality. Improvements in SR are still substantial for **LB** glycans, for which medium-quality SR increases from 43% to 57%.

The cluster-based SR (CB-SR) calculated on the refined models is consistently higher than the CB-SR calculated on the clustered rigid body models. As an example, the best scoring cluster (T1), medium quality CB-SR increases from 32% to 44%, from 4 to 9%, and from 0% to 14% for **SL-SB**, **LL**, and **LB** glycans, respectively.

In *Figure 4.6*, the best scoring refined models superimposed to the corresponding rigid body models and to the reference structures of three complexes, representative of the **SL-SB**, **LL**, and **LB** groups, are shown as an example. It can be seen that the refinement stage results in a better ranking of the models, whose quality improves too (lower IL-RMSD values are obtained following the refinement). The quality of the models decreases with the increasing complexity of the glycans: a high-quality model is produced for the 1C1L complex (**SL**, IL-RMSD = 0.52 Å after the refinement), a medium-quality model for 5VX5 (**LL**, IL-RMSD = 2.22 Å), an acceptable-quality model for 1OH4 (**LB**, IL-RMSD = 4.42 Å).

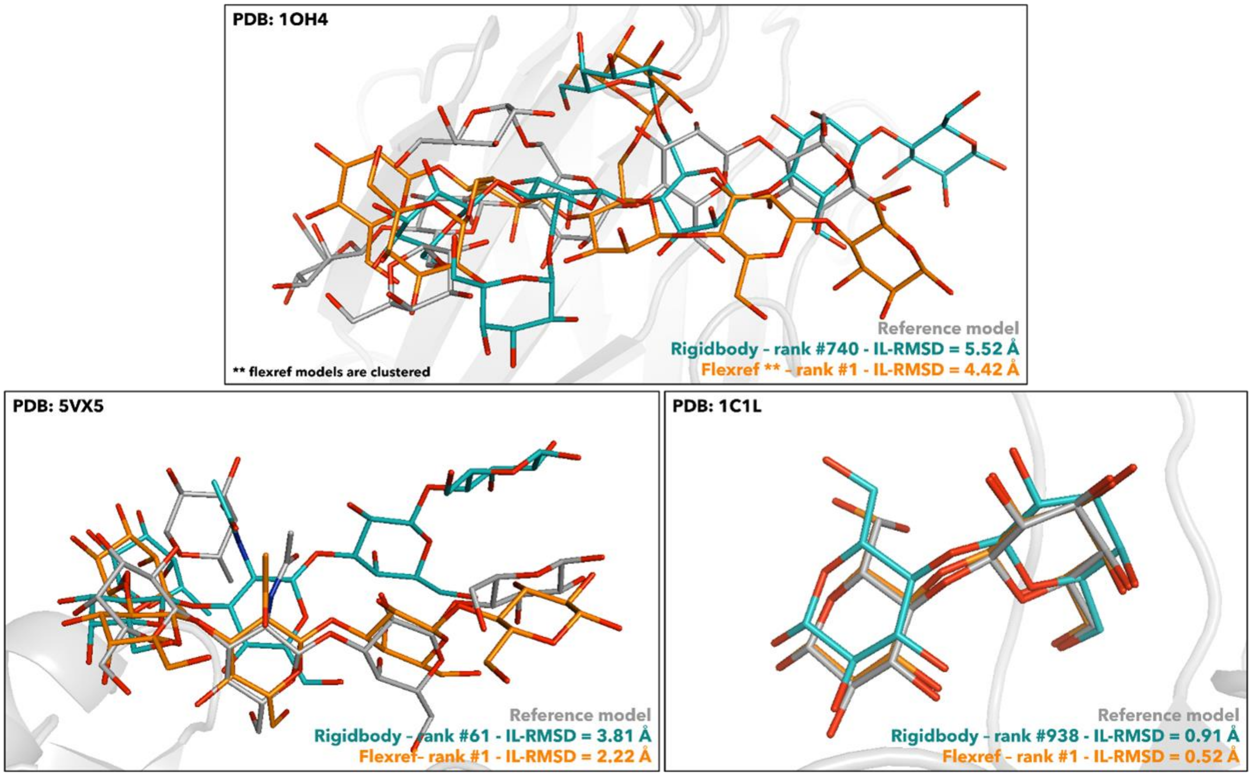

PDB: 1OH4
Reference model
Rigidbody - rank #740 - IL-RMSD = 5.52 Å
Flexref ** - rank #1 - IL-RMSD = 4.42 Å
** flexref models are clustered
PDB: 5VX5
Reference model
Rigidbody - rank #61 - IL-RMSD = 3.81 Å
Flexref- rank #1 - IL-RMSD = 2.22 Å
PDB: 1C1L
Reference model
Rigidbody - rank #938 - IL-RMSD = 0.91 Å
Flexref - rank #1 - IL-RMSD = 0.52 Å

*Figure 4.6 - Superimposition of the best scoring flexref models (orange) and the rigidbody models (teal) to the reference structures (grey) for the complexes 1OH4 (**LB**, top panel), 5VX5 (**LL**, bottom left panel), and 1C1L (**SL**, bottom right panel) and the unbound docking scenario carried out with vdW scoring potential and **tip-ap** AIRs. Oxygen atoms of the glycans are shown in red in all the structures, nitrogens in blue, hydrogens not shown. Ranking and IL-RMSD values with respect to the reference structures for the flexref and rigidbody models are shown too.*

HADDOCK3 performance on the *unbound dataset* was also evaluated using the **ti-aa** AIRs (*Figure 4.7*); considerations like the ones done for the **tip-ap** scenario can be done.

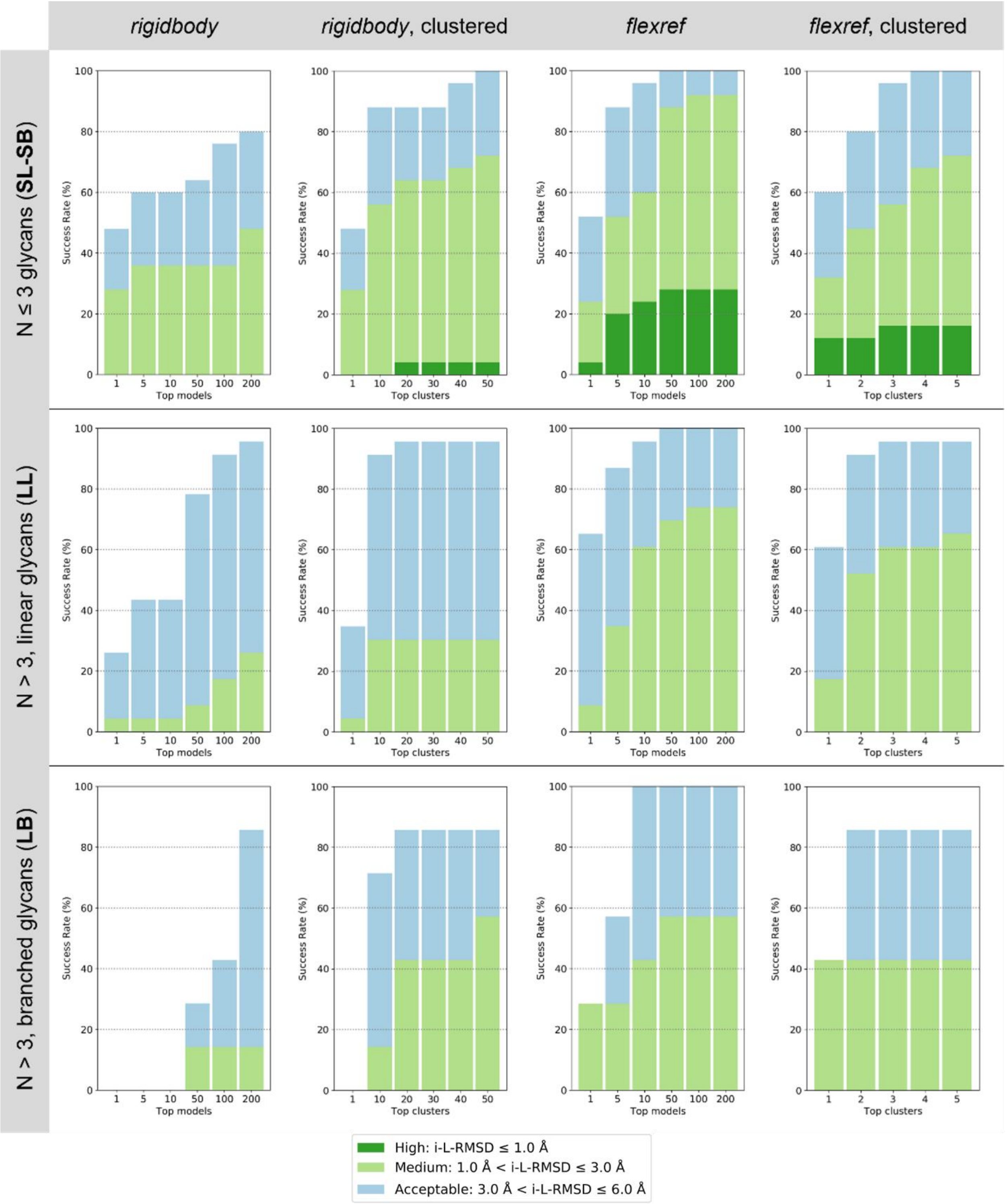


Figure 4.7 - HADDOCK3 performance on unbound dataset, using vdW scoring function and ***ti-aa*** *AIRs. SR are calculated: on the T1 to T200 rigidbody models (first column) and on T1 to T50 rigid body clusters, considering only the top 5 models of each cluster (second column); on the T1 to T200 refined models (third column) and on T1 to T5 refined clusters, considering only the top 4 models of each cluster (fourth column). Models are refined according to the cluster-based selection (second column). SR is calculated separately for the three categories of complexes grouped by glycans size and connectivity:* ***SL-SB*** *(top row),* ***LL*** *(middle row), and* ***LB*** *(bottom row).*

The SR obtained after the refinement stage are lower than the ones obtained on the *bound dataset*. This is because the refinement stage cannot consistently recover the bound conformation of the protein-glycan interface, especially if the glycan has a complex structure and its unbound conformation is far from the bound one. An attempt for improving the SR of unbound conformations was done by performing longer simulations during the refinement stage (see *Table 4.3* for the number of steps, labelled with "extended"). However, this extended refinement stage did not improve SR significantly nor in a uniform way, as it seemed to depend on the number of models and the group of glycans considered. As this behaviour could not be rationalized in a simple way, this approach was discarded.

The challenge of docking unbound conformations was then addressed by considering the glycans conformational variability prior to the docking process.

R4: Glycans conformational sampling

Conformational sampling of the glycans was carried out to perform docking calculations with an ensemble of structures. Results are discussed by considering separately the three categories of glycans: **SL-SB**, **LL**, and **LB**, given that ref_Glycan_RMSD (i.e., the Glycan_RMSD of the conformation generated by the GLYCAM-web webserver with respect to the bound one) has a strong dependency on glycans size and connectivity, as a consequence of the larger number of conformations that longer and more complex glycans can assume with respect to the shorter ones. The distributions of ref_Glycan_RMSD are shown in *Figure 4.8*.

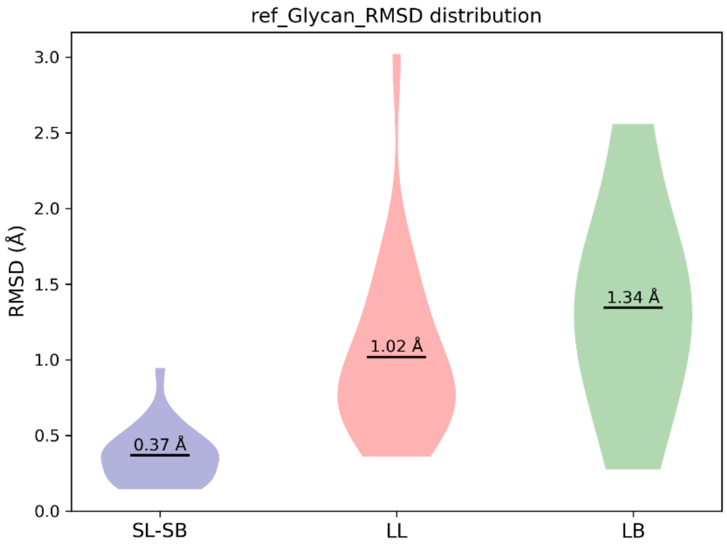


Figure 4.8 - Violin plot showing the distribution of ref_Glycan_RMSD for the three categories of glycans ***SL-SB****,* ***LL****, and* ***LB****. Median values are shown on top of the plot.*

The sampling was performed with five different protocols, where MD simulations of different length were carried out on different number of models. The parameters of these scenarios are shown in *Table 4.4*.

For each protocol, Glycan_RMSD values were calculated for all the generated conformations and compared with the ref_Glycan_RMSD. As an example, such plots are shown for glycans 1C1L (**SL),** 5VX5 (**LL**), 1OH4 (**LB**) in *Figure 4.9*.

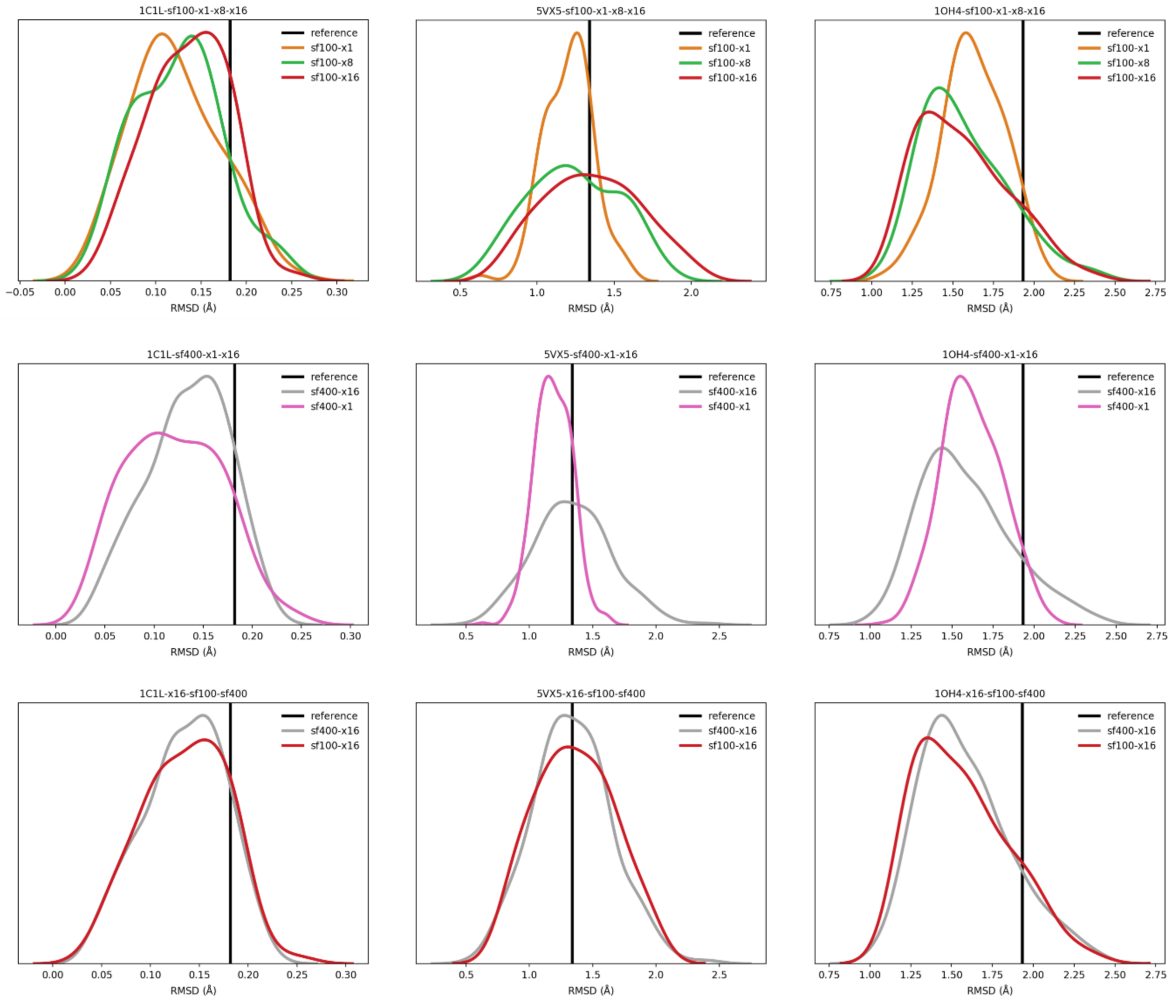


Figure 4.9 – Distribution of the Glycan_RMSD values obtained for glycans 1C1L (first column), 5VX5 (second column), and 1OH4 (third column) with the different sampling scenarios, with respect to the ref_Glycan_RMSD.

The best protocol is the one that allows to sample the largest pool of conformations and thus also the one showing the lowest Glycan_RMSD values, i.e., the closest to the bound structure. The distributions of the lowest Glycan_RMSD values, for the different sampling scenarios and for the three groups of glycans (**SL-SB**, **LL**, **LB**) are shown in the boxplots in *Figure 4.10*.

The protocol that allows to obtain the lowest median values (*Figure 4.10*) is sf400-x16. For **SL-SB** glycans, the median decreases from 0.36 Å (ref_Glycan_RMSD) to 0.14 Å (sf400-x16); for **LL** glycans, from 0.82 Å to 0.52 Å; for **LB** glycans, from 1.36 Å to 0.73 Å.

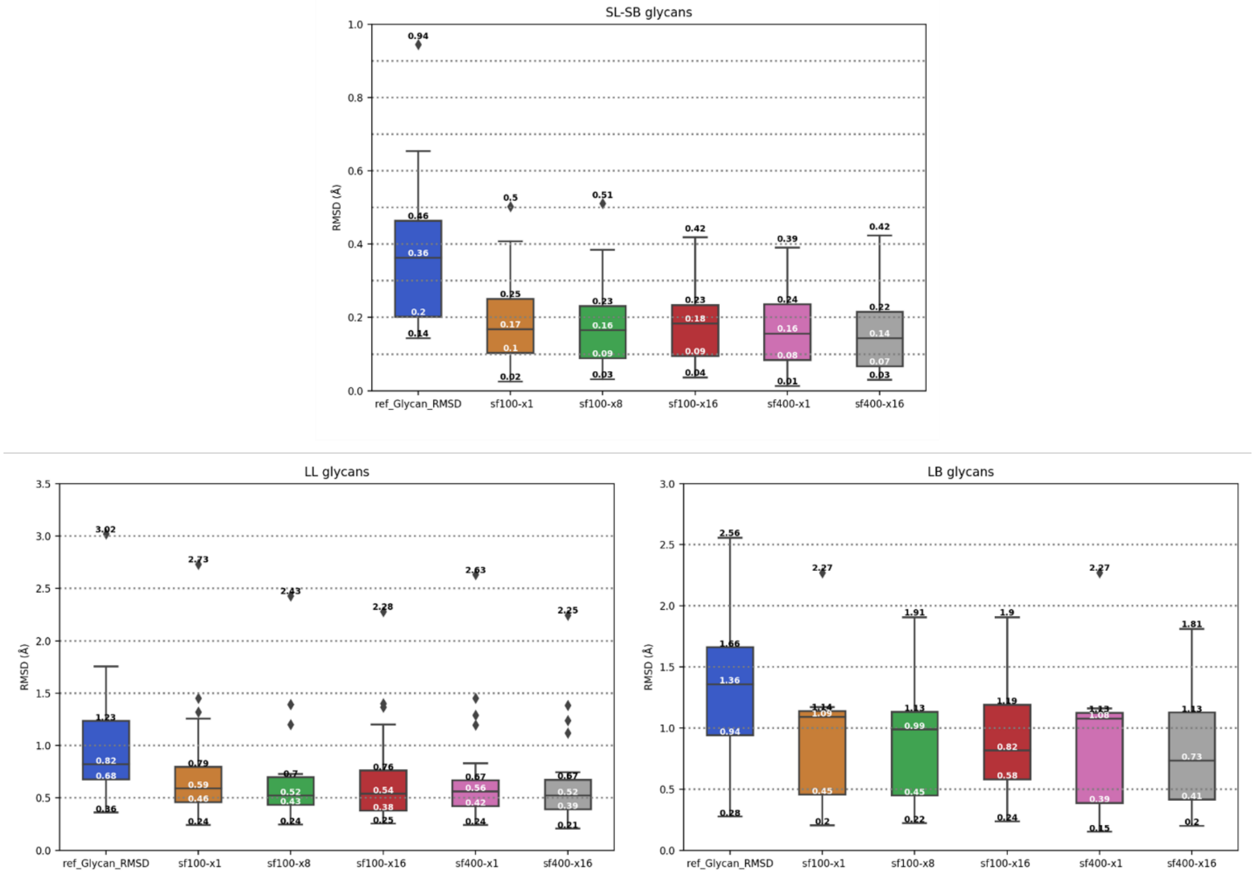


Figure 4.10 - Boxplots representing the comparison between ref_Glycan_RMSD and the lowest Glycan_RMSD obtained with the 5 sampling scenarios. The comparison is shown for the three groups of glycans: ***SL-SB*** *(top),* ***LL*** *(bottom left), and* ***LB*** *(bottom right).*

With the aim of obtaining limited ensembles of structures for further docking calculations, the [rmsdmatrix] and [clustrmsd] modules in HADDOCK3 were exploited, extracting either 10 or 20 clusters from the ensemble. Glycan_RMSD values of the centers of such clusters were calculated with ProFit v3.3[21,22] and compared to the overall sf400-x16 distribution, in order to understand which number of clusters is necessary to uniformly cover the whole distribution.

As an example, the plots showing the overall distribution and the values of 10 and 20 clusters centres are reported for glycans 1C1L (**SL),** 5VX5 (**LL**), 1OH4 (**LB**) in *Figure 4.11*.

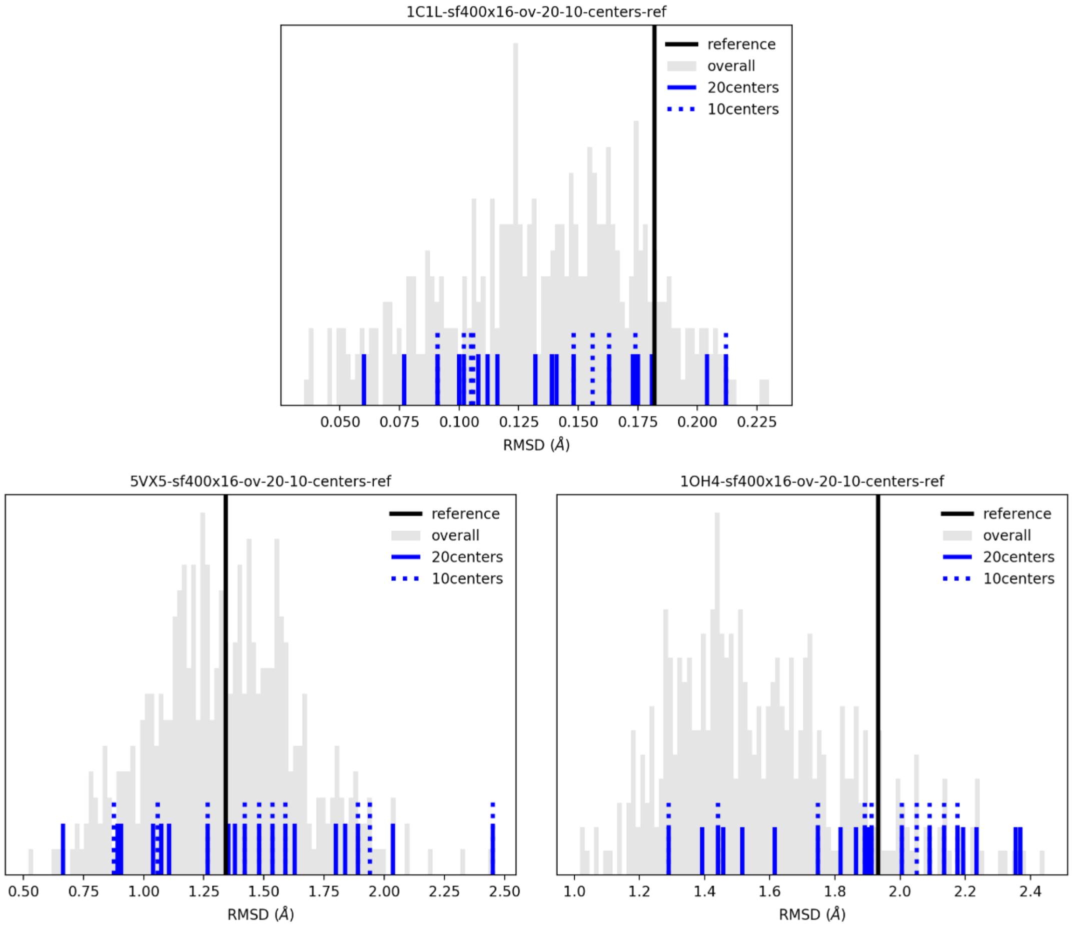


Figure 4.11 – Comparison of the overall sf400-x16 distribution with the Glycan_RMSD values of 10 and 20 clusters centres, for the glycans 1C1L (top), 5VX5 (bottom left), and 1OH4 (bottom right).

Overall, grouping the sampled models in 20 clusters was the more appropriate way for retaining low Glycan_RMSD conformations. However, the centers of the clusters characterized by the lowest Glycan_RMSD values rarely correspond to the sampled conformations closest to the bound form, i.e. the latter are typically "lost" in the clustering process. This happens for example for glycan 1OH4 (bottom right panel, *Figure 4.11*): conformations with a Glycan_RMSD around 1 Å are sampled, but the conformation closest to the bound structure which is retaind after the clustering process shows a Glycan_RMSD around 1.3 Å. For a complete view of the problem, for all the glycans, the Glycan_RMSD values of the clusters centers closed to the bound conformations are plotted against the same values from the overall sampling (*Figure 4.12*).

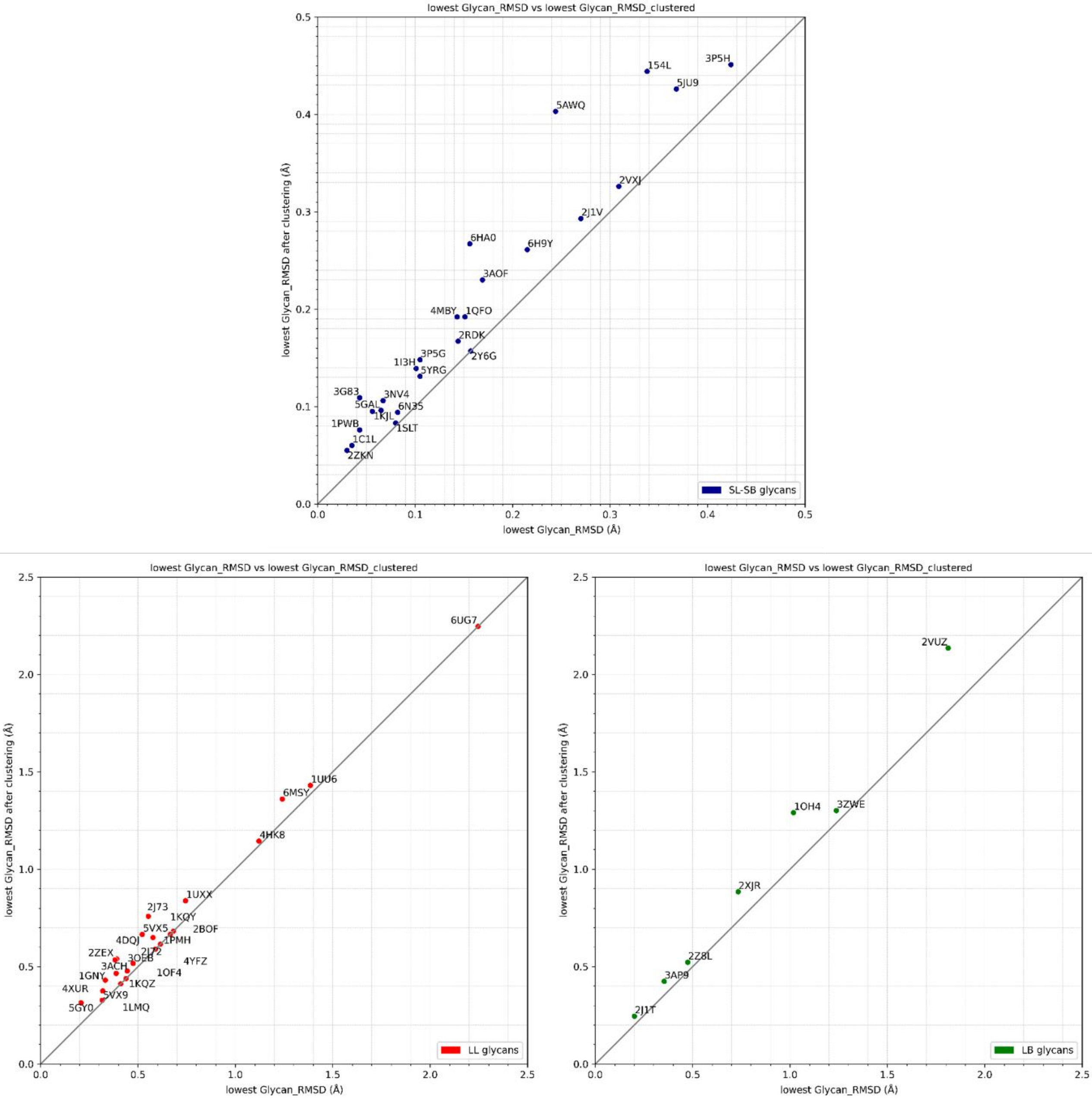


Figure 4.12 - Lowest Glycan_RMSD values after clustering are plotted against the lowest Glycan_RMSD from the overall sampling. The comparison is shown for the three groups of glycans: ***SL-SB*** *(top),* ***LL*** *(bottom left), and* ***LB*** *(bottom right).*

The centers of these 20 clusters were merged using pdb-tools[18] in an ensemble to be used in new docking calculations. Such structures, as an example, are shown in *Figure 4.13* for the three glycans 1C1L (**SL),** 5VX5 (**LL**), 1OH4 (**LB**).

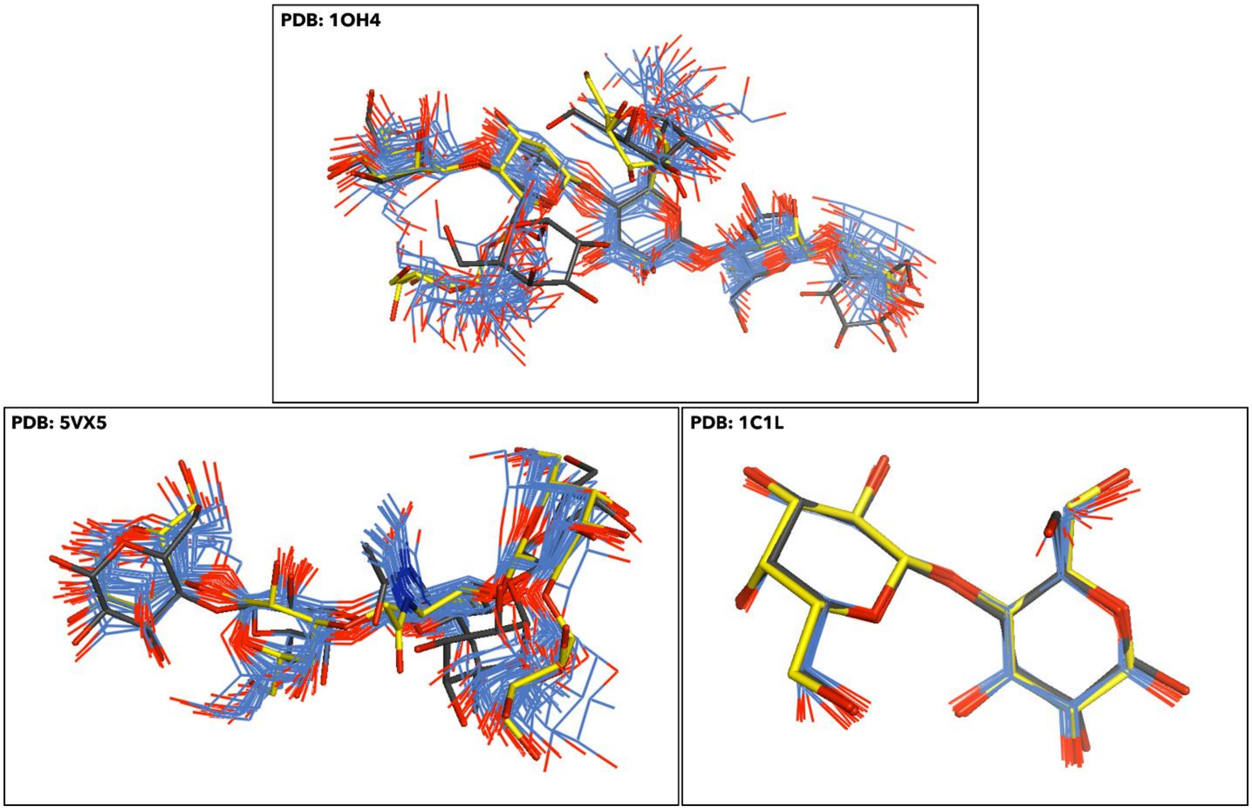


*Figure 4.13 - Superimposition of the centers of the 20 clusters (light blue) obtained from the sf400-x16 sampling scenario and of the unbound conformation generated by GLYCAM-Web webserver (yellow) to the bound conformations (black) for the complexes 1OH4 (**LB**, top panel), 5VX5 (**LL**, bottom left panel), and 1C1L (**SL**, bottom right panel). Oxygen atoms are shown in red in all the structures, nitrogens in blue, hydrogens not shown.*

R5: Using an ensemble slightly improves success rates in unbound docking

New docking calculations were then performed with the unbound conformations of the proteins (as in **R3**) and the ensemble of glycans to assess whether higher SR could be achieved by including conformations of glycans closer to the bound forms. In *Figure 4.14*, the corresponding SR calculated on the refined models is compared to the one obtained with single glycan conformations for the three categories of glycans. Improvements obtained with the latter are limited but present.

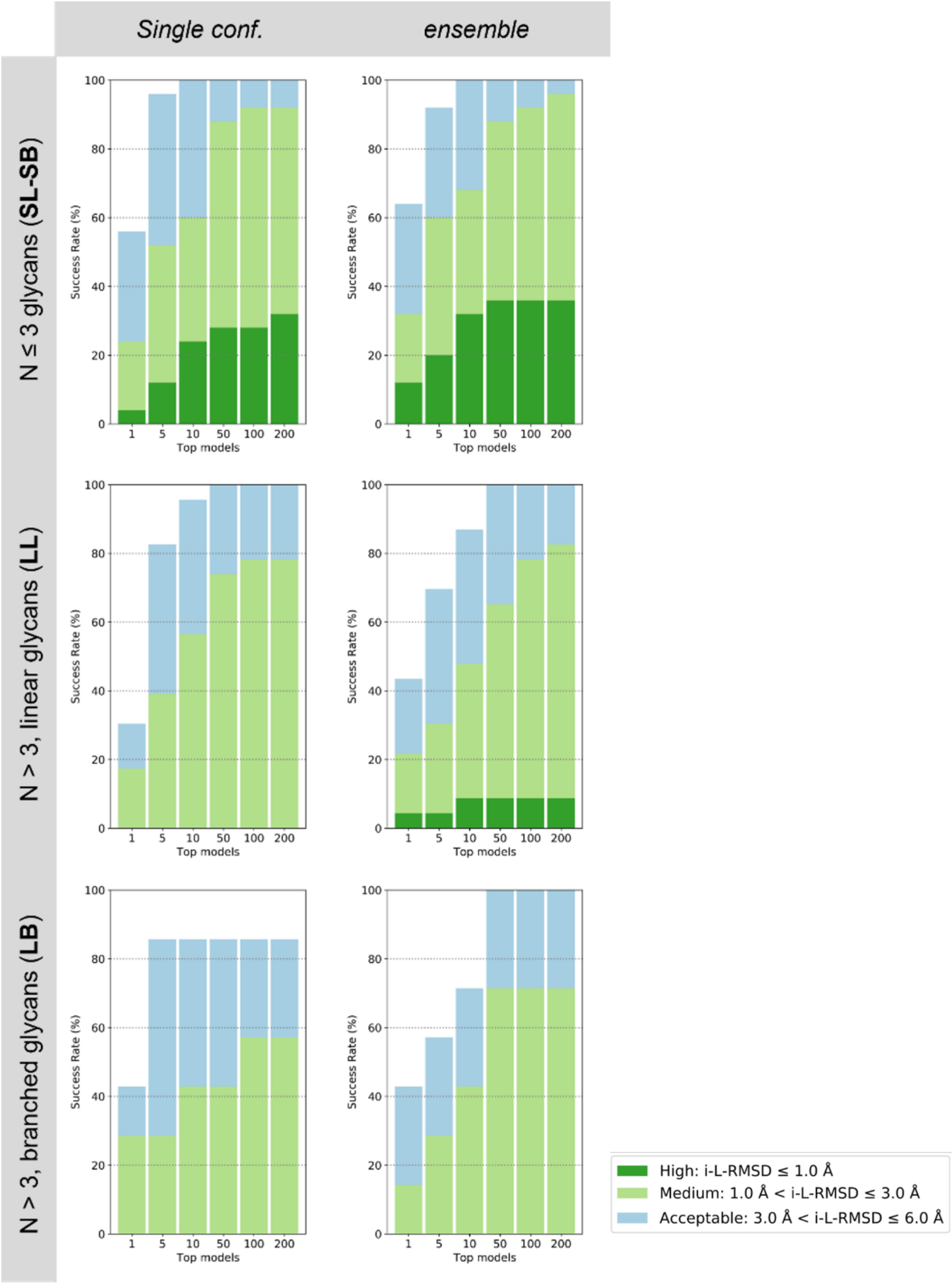


Figure 4.14 - HADDOCK3 performance on unbound dataset, using vdW scoring function and ***tip-ap*** *AIRs. SR, calculated on the top 1, 5, 10, 50, 100, and 200 refined models (flexref stage), are compared between single conformations runs (left column) and ensemble runs (right column). SR is calculated separately for the three categories of complexes grouped by glycans size and connectivity:* ***SL-SB*** *(top row),* ***LL*** *(middle row), and* ***LB*** *(bottom row).*

Considering complexes with **SL-SB** glycans (*Figure 4.14*, top row), high-quality SR increases from 4, 12, 24, and 32% to 12, 20, 32, and 36%, for T1, T5, T10, and T200 models respectively. The number of medium-quality models slightly increases too. As for complexes with **LL** glycans (*Figure 4.14*, middle row), some high-quality models are generated with the ensemble, with SR around 4% (T1 and T5 models) and around 9% (T10-T200); despite the medium-quality SR decreases with respect to the single conformations docking runs. If acceptable models are concerned, the T1 SR increases while the other do not change (T50-T200) or even decrease (T5-T10) with respect to the single conformations runs. Using the ensemble for complexes with **LB** glycans (*Figure 4.14*, bottom row) does not seem to significantly help as far as T1-T10 models are considered. Medium-quality SR decreases (T1) or does not vary (T5 and T10), while acceptable-quality SR remains constant for T1 and decreases for T5 and T10 models. Improvements following the introduction of the ensemble are only visible for T50-T200 models, where medium-quality SR increases from 42% (T50) and 57% (T100 and T200) to around 70% for all of them. Moreover, all the models from T50 onwards fall within the acceptable cut-off.

Docking runs with the ensemble were performed also with **ti-aa** AIRs, with similar results (*Figure 4.15*).

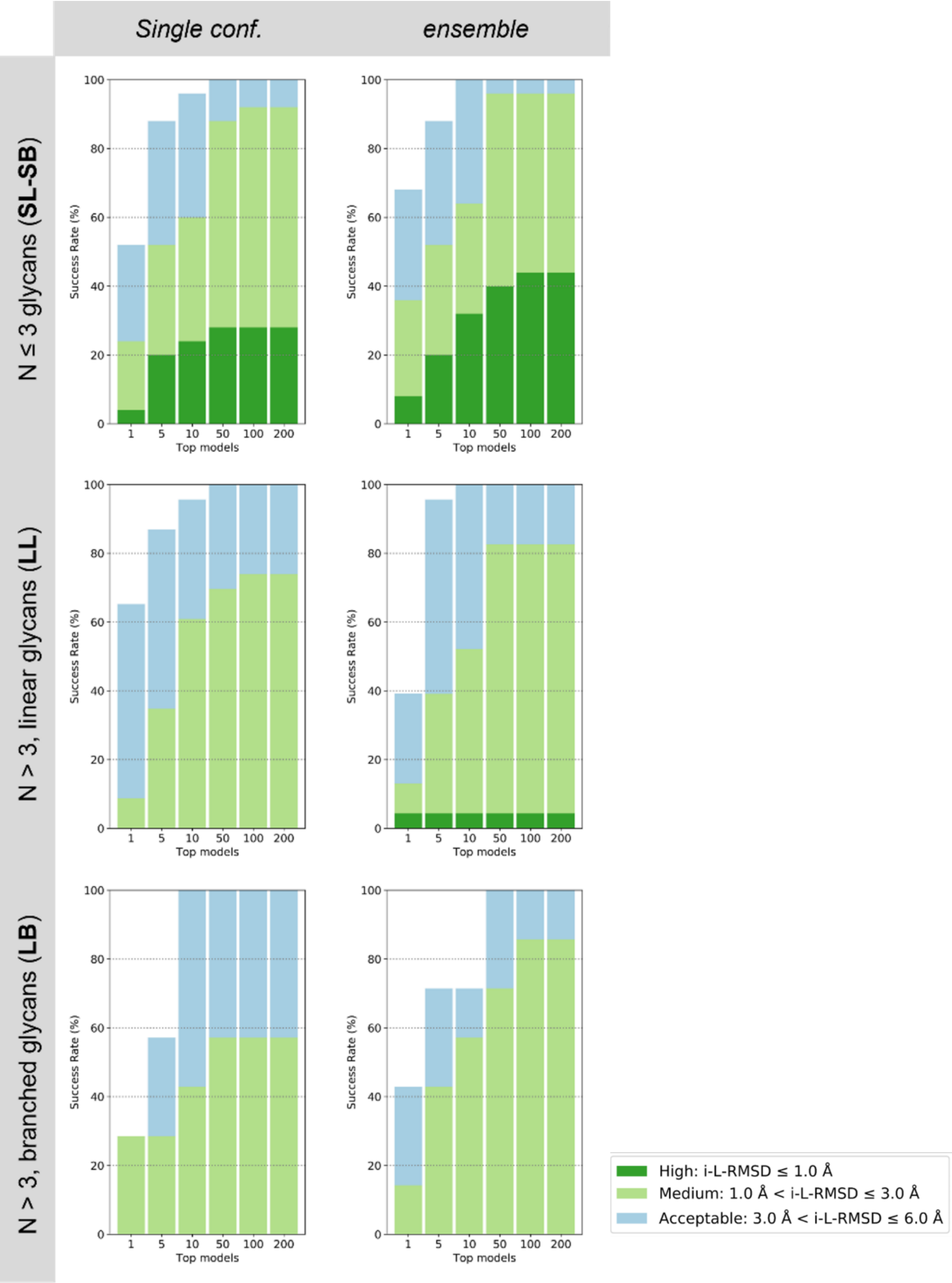


*Figure 4.15 - HADDOCK3 performance on unbound dataset, using vdW scoring function and **ti-aa** AIRs. SR, calculated on the top 1, 5, 10, 50, 100, and 200 refined models (flexref stage), are compared between single conformations runs (left column) and ensemble runs (right column). SR is calculated separately for the three categories of complexes grouped by glycans size and connectivity: **SL-SB** (top row), **LL** (middle row), and **LB** (bottom row).*

Overall, the rather limited improvements in SR following the introduction of the ensemble can be reconducted to the following interlinked reasons. The first one concerns the short, limited glycan conformational sampling performed with the HADDOCK3 water refinement module. At this stage, drawbacks could be due both to the relatively short MD simulations and to the force field employed in the program, i.e., the all-purpose OPLS-AA force field[14]. A second problem is the loss of some of the sampled conformations characterized by the lowest Glycan_RMSD values; this, as it was shown, is consequence of the unavoidable clustering of the whole pool of sampled conformations. At last, a consideration needs to be done on the docking stage. Following an increase in the number of input structures by a factor 20, the number of models to be generated and then refined was increased too. However, this was not done linearly with the number of structures to avoid exceedingly expensive calculations: 4000 models were generated instead of 1000 (single conformations runs); 150 cluster composed by at maximum 5 models were sent to the refinement stage, instead of the 50 of the single conformations runs.

4.4 – Conclusions

In the present study, the ability of HADDOCK3 in reproducing the tridimensional structures of glycan- protein complexes was assessed. First, the performance was evaluated on the rather simple, unrealistic scenario involving the partners in their bound conformations, giving full information about the interface (**ti-aa** AIRs). The best rigid body scoring function was shown to be the *vdW* scoring function, having an upweighted van der Waals energy term which better accounts for the glycans hydrophobicity. The SR were calculated on the *bound dataset* split in the four groups **SL**, **SB**, **LL**, and **LB** and it was shown that, overall, the prediction of the complexes is easier when **LL** and **LB** glycans are involved; this is probably a consequence of the lower number of possible dispositions that longer glycans can assume when binding the protein. Assuming no information on the glycan's interface (**tip-ap** AIRs) on the *bound dataset* causes decreases in the SR mainly for complexes involving branched glycans, independently on their size.

HADDOCK3 performance was then evaluated on the *unbound dataset*. Docking runs were performed on protein unbound structures retrieved from the PDB and on glycans conformations generated through the GLYCAM-web webserver. It was observed that a cluster-based selection of the rigid body models is beneficial for docking success, as it allows to pick more correct models for refinement.

Overall, the SR obtained on the *unbound dataset* are lower than the ones calculated on the *bound dataset*. Therefore, the glycan conformational variability was accounted for prior to the docking by means of MD simulations and clustering, thus producing an ensemble of glycan structures. It was observed that the introduction of a pool of glycans structures produces slight improvements, due to the not exhaustive glycan conformational sampling and to the clustering step, and to the "dilution problem" which may arise when having multiple conformations in input to the docking.

In conclusion, the present study showed that HADDOCK3 is indeed suitable for the prediction of glycan-proteins binding geometries, as it was seen with the evaluation of its performance on the *bound dataset*. However, docking unbound conformations remains a challenging open problem. This is mostly due to the large conformational variability of glycans, which grows with their complexity.

Future efforts could be oriented on the development of a more efficient protocol for the generation of glycans conformations to be incorporated in the docking calculations, possibly integrating different MD software such as GROMACS[29] or OpenMM[30] within the HADDOCK3 docking pipeline.

Besides the most challenging task of properly account for glycans conformational variability, there a few other aspects that need to be addressed. One of them concerns expanding the list of glycans supported by HADDOCK3.

References

(1) Seeberger, P. H. Monosaccharide Diversity. In *Essentials of Glycobiology*; Varki, A., Cummings, R. D., Esko, J. D., Eds.; Cold Spring Harbor Laboratory Press: Cold Spring Harbor (NY), Cold Spring Harbor (NY), **2022**; pp 21–32. https://doi.org/10.1101/glycobiology.4e.2.

(2) Lebrilla, C. B.; Liu, J.; Widmalm, G.; Prestegard, J. H. Oligosaccharides and Polysaccharides. In *Essentials of Glycobiology*; Varki, A., Cummings, R. D., Esko, J. D., Stanley, P., Hart, G. W., Aebi, M., Mohnen, D., Kinoshita, T., Packer, N. H., Prestegard, J. H., Schnaar, R. L., Seeberger, P. H., Eds.; Cold Spring Harbor (NY), **2022**; pp 33–42. https://doi.org/10.1101/glycobiology.4e.3.

(3) Gagneux, P.; Hennet, T.; Varki, A. Biological Functions of Glycans. In *Essentials of Glycobiology*; Varki, A., Cummings, R. D., Esko, J. D., Stanley, P., Hart, G. W., Aebi, M., Mohnen, D., Kinoshita, T., Packer, N. H., Prestegard, J. H., Schnaar, R. L., Seeberger, P. H., Eds.; Cold Spring Harbor (NY), **2022**; pp 79–92. https://doi.org/10.1101/glycobiology.4e.7.

(4) Molina, A.; O'Neill, M. A.; Darvill, A. G.; Etzler, M. E.; Mohnen, D.; Hahn, M. G.; Esko, J. D. Free Glycans as Bioactive Molecules. In *Essentials of Glycobiology*; Varki, A., Cummings, R. D., Esko, J. D., Stanley, P., Hart, G. W., Aebi, M., Mohnen, D., Kinoshita, T., Packer, N. H., Prestegard, J. H., Schnaar, R. L., Seeberger, P. H., Eds.; Cold Spring Harbor Laboratory Press: Cold Spring Harbor (NY), Cold Spring Harbor (NY), **2022**; pp 539–548. https://doi.org/10.1101/glycobiology.4e.40.

(5) Casalino, L.; Gaieb, Z.; Goldsmith, J. A.; Hjorth, C. K.; Dommer, A. C.; Harbison, A. M.; Fogarty, C. A.; Barros, E. P.; Taylor, B. C.; McLellan, J. S.; Fadda, E.; Amaro, R. E. Beyond Shielding: The Roles of Glycans in the SARS-CoV-2 Spike Protein. *ACS Cent. Sci.* **2020**, *6* (10), 1722–1734. https://doi.org/10.1021/acscentsci.0c01056.

(6) Perez, S.; Makshakova, O. Multifaceted Computational Modeling in Glycoscience. *Chem. Rev.* **2022**, *122* (20), 15914–15970. https://doi.org/10.1021/acs.chemrev.2c00060.

(7) Nance, M. L.; Labonte, J. W.; Adolf-Bryfogle, J.; Gray, J. J. Development and Evaluation of GlycanDock: A Protein–Glycoligand Docking Refinement Algorithm in Rosetta. *J. Phys. Chem. B* **2021**, *125* (25), 6807–6820. https://doi.org/10.1021/acs.jpcb.1c00910.

(8) Dominguez, C.; Boelens, R.; Bonvin, A. M. J. J. HADDOCK: A Protein−Protein Docking

Approach Based on Biochemical or Biophysical Information. *J. Am. Chem. Soc.* **2003**, *125* (7), 1731–1737. https://doi.org/10.1021/ja026939x.

(9) Bonvin's Lab. HADDOCK3. **2022**.

(10) Berman, H. M. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28* (1), 235–242. https://doi.org/10.1093/nar/28.1.235.

(11) Woods Group. Complex Carbohydrate Research Center. GLYCAM Web. University of Georgia, Athens, GA **2023**.

(12) Kirschner, K. N.; Yongye, A. B.; Tschampel, S. M.; González-Outeiriño, J.; Daniels, C. R.; Foley, B. L.; Woods, R. J. GLYCAM06: A Generalizable Biomolecular Force Field. Carbohydrates. *J. Comput. Chem.* **2008**, *29* (4), 622–655. https://doi.org/10.1002/jcc.20820.

(13) de Vries, S. J.; van Dijk, A. D. J.; Krzeminski, M.; van Dijk, M.; Thureau, A.; Hsu, V.; Wassenaar, T.; Bonvin, A. M. J. J. HADDOCK versus HADDOCK: New Features and Performance of HADDOCK2.0 on the CAPRI Targets. *Proteins Struct. Funct. Bioinforma.* **2007**, *69* (4), 726–733. https://doi.org/https://doi.org/10.1002/prot.21723.

(14) 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. https://doi.org/10.1021/ja00214a001.

(15) 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. https://doi.org/10.1016/j.jmb.2003.10.069.

(16) Wallace, A. C.; Laskowski, R. A.; Thornton, J. M. LIGPLOT: A Program to Generate Schematic Diagrams of Protein-Ligand Interactions. *"Protein Eng. Des. Sel.* **1995**, *8* (2), 127–134. https://doi.org/10.1093/protein/8.2.127.

(17) McDonald, I. K.; Thornton, J. M. Satisfying Hydrogen Bonding Potential in Proteins. *J. Mol. Biol.* **1994**, *238* (5), 777–793. https://doi.org/10.1006/jmbi.1994.1334.

(18) Rodrigues, J. P. G. L. M.; Teixeira, J. M. C.; Trellet, M.; Bonvin, A. M. J. J. Pdb-Tools: A Swiss Army Knife for Molecular Structures. *F1000Research* **2018**, *7*, 1961. https://doi.org/10.12688/f1000research.17456.1.

(19) Méndez, R.; Leplae, R.; De Maria, L.; Wodak, S. J. Assessment of Blind Predictions of Protein–Protein Interactions: Current Status of Docking Methods. *Proteins Struct. Funct.*

Bioinforma. **2003**, *52* (1), 51–67. https://doi.org/10.1002/prot.10393.

(20) Lensink, M. F.; Nadzirin, N.; Velankar, S.; Wodak, S. J. Modeling Protein-protein, Protein-peptide, and Protein-oligosaccharide Complexes: CAPRI 7th Edition. *Proteins Struct. Funct. Bioinforma.* **2020**, *88* (8), 916–938. https://doi.org/10.1002/prot.25870.

(21) Martin, A. C. R. ProFit.

(22) McLachlan, A. D. Rapid Comparison of Protein Structures. *Acta Crystallogr. Sect. A* **1982**, *38* (6), 871–873. https://doi.org/10.1107/S0567739482001806.

(23) Giulini, M.; Menichetti, R.; Shell, M. S.; Potestio, R. An Information-Theory-Based Approach for Optimal Model Reduction of Biomolecules. *J. Chem. Theory Comput.* **2020**, *16* (11), 6795–6813. https://doi.org/10.1021/acs.jctc.0c00676.

(24) Sokal, R. R.; Michener, C. D. A Statistical Method for Evaluating Systematic Relationships. *Univ. Kansas Sci. Bull.* **1958**, *38*, 1409–1438.

(25) Charitou, V.; van Keulen, S. C.; Bonvin, A. M. J. J. Cyclization and Docking Protocol for Cyclic Peptide–Protein Modeling Using HADDOCK2.4. *J. Chem. Theory Comput.* **2022**, *18* (6), 4027–4040. https://doi.org/10.1021/acs.jctc.2c00075.

(26) Xu, M.; Lill, M. A. Utilizing Experimental Data for Reducing Ensemble Size in Flexible-Protein Docking. *J. Chem. Inf. Model.* **2012**, *52* (1), 187–198. https://doi.org/10.1021/ci200428t.

(27) Korb, O.; Olsson, T. S. G.; Bowden, S. J.; Hall, R. J.; Verdonk, M. L.; Liebeschuetz, J. W.; Cole, J. C. Potential and Limitations of Ensemble Docking. *J. Chem. Inf. Model.* **2012**, *52* (5), 1262–1274. https://doi.org/10.1021/ci2005934.

(28) Amaro, R. E.; Baudry, J.; Chodera, J.; Demir, Ö.; McCammon, J. A.; Miao, Y.; Smith, J. C. Ensemble Docking in Drug Discovery. *Biophys. J.* **2018**, *114* (10), 2271–2278. https://doi.org/10.1016/j.bpj.2018.02.038.

(29) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High Performance Molecular Simulations through Multi-Level Parallelism from Laptops to Supercomputers. *SoftwareX* **2015**, *1–2*, 19–25. https://doi.org/10.1016/j.softx.2015.06.001.

(30) Eastman, P.; Swails, J.; Chodera, J. D.; McGibbon, R. T.; Zhao, Y.; Beauchamp, K. A.; Wang, L.-P.; Simmonett, A. C.; Harrigan, M. P.; Stern, C. D.; Wiewiora, R. P.; Brooks, B. R.; Pande, V. S. OpenMM 7: Rapid Development of High Performance Algorithms for

Molecular Dynamics. *PLOS Comput. Biol.* **2017**, *13* (7), e1005659. https://doi.org/10.1371/journal.pcbi.1005659.

Section 5 – Final remarks

This PhD thesis concerned the use of data-driven, physics-based computational methods for the determination of the three-dimensional structures of protein-protein and protein-glycans complexes.

The two projects presented here share some common points. Physics-based computational approaches, such as molecular dynamics simulations and molecular docking, were employed in both. In addition, the calculations performed for both projects are "data-driven", meaning that the available experimental data were exploited to increase the reliability of the modelling procedures.

In **Section 3**, a specific use case related to protein-protein interactions was presented. The project, which was carried out in collaboration with Dr. Alessandro Maiocchi (Bracco S.p.A) and Dr. Elisabetta Moroni (SCITEC-Italian National Research Council), followed the publication of two patents, owned by Bracco S.p.A. These patents concern the use of two small engineered proteins, namely affitins, for the recognition of a molecular target of undoubtable importance, i.e., the human epidermal growth factor receptor 2 (HER2). The aim was to predict the structure of the Affitin_1-HER2 and Affitin_2-HER2 complexes, knowledge of which is essential for the optimization of the binding affinity. This was achieved by, first of all, incorporating in the docking calculations the experimental evidence that the two affitins do not compete for the HER2 binding sites involved in the recognition of the monoclonal antibodies Trastuzumab and Pertuzumab. Then, the obtained docking models were evaluated with the aim of identifying the most likely ones.

This was done through a procedure previously developed on a dataset of experimentally known complexes consisting of affitins and other protein partners. The procedure is based on a consensus approach between DockQ and MLCE, two methods relying on different assumptions. DockQ assesses the quality of a model on the basis of its stability along MD trajectories. MLCE is used to predict binding sites the isolated structure of one of the two proteins. The presence of a match between these predicted sites and the binding areas in the docking models is then checked; if a match is found, the pose is considered more likely than the others.

The application of the DockQ-MLCE consensus approach, and the consideration of the ClusPro score at the same time, allowed to obtain the most likely docking poses for the complexes Affitin_1-HER2 and Affitin_2-HER2. These were compared with the structures of complexes, available in the Protein Data Bank, consisting of HER2 and different protein partners. In this way, two protein partners, namely an affibody and a Fab, were identified as possible candidates to perform competitive binding assays. These experimental tests could assess whether Affitin_1 and Affitin_2 actually bind HER2 on the predicted epitopes.

Section 4 reports the work done at the Computational Structural Biology (CSB) group (Bijvoet Centre for Biomolecular Research, Universiteit Utrecht), under the supervision of Prof. Alexandre Bonvin and Dr. Marco Giulini. The study aimed to build a reliable protocol, based on the HADDOCK3 docking programme, currently under development at the CSB group, for the prediction of protein-glycan complexes.

Docking calculations were driven by the information on the binding interface of the protein partners, which was included as a restraint in the docking calculations. This was possible as the building of the protocol of course implied the use of protein-glycan complexes whose structures are known and deposited in the Protein Data Bank. In a realistic scenario, where the structures would not be known, the information on the interface could be obtained, for example, by NMR analysis.

HADDOCK3 performance was evaluated on both a bound and an unbound dataset. The former dataset was used to evaluate the rigid body scoring function that best reproduces the structures of the complexes. The calculations carried out on the unbound dataset, instead, highlighted the importance of a finely tuned selection of the rigid body models to be used for the following MD-based refinement stage. It was found that a RMSD-based clustering of the rigid body models allows more good quality models to be selected for the refinement.

The success rates obtained with the unbound dataset were partially improved by including, in new docking calculations, an ensemble of glycans conformations generated within HADDOCK3. Given the results, there is still room for improvement. Therefore, future efforts could be directed toward developing a more efficient protocol for generating glycans conformations to be incorporated in the docking calculations.

Appendix 1 – Other activities

In this appendix, other activities carried out during the three years of the PhD studies are briefly covered. Both the two projects herein described mainly used unbiased (and sometimes enhanced) Molecular Dynamics (MD) simulations, and a little bit of quantum mechanical calculations.

Partition of organic and metalorganic compounds in the biphasic systems n-octanol/water and micelle/water

The aim of this project has been the study of the partition process of several molecules in two biphasic systems by using methods based on Molecular Dynamics (MD) simulations. All the species have been described with the same force field (2016H66), once its reliability has been tested comparing the n-octanol/water partition free energies calculated from the MD and Free Energy Perturbation (FEP) method with those obtained from the quantum-mechanical SMD method.

The protocol has then foreseen: i) steered MD (SMD) simulations for displacing the solutes from one phase to another; ii) umbrella sampling (US) simulations carried out on the configurations extracted from the SMD simulations; iii) use of the Weighted Histogram Analysis Method (WHAM) to obtain the Potential of Mean Force (PMF) profiles and thus the free energies of partition.

First, the biphasic system n-octanol/water has been considered together with molecules for which experimental values of partition coefficients are available. The aim of this first part has been to identify the set-up that would allow to best reproduce the experimental values. Different combinations of parameters in terms of dimensions of the simulations box, number of sampled conformations along the SMD trajectory and extension of the sampling of each umbrella window have been considered.

Then, attention has been given to a biphasic system micelle/water, which is made up by the self-organization of surfactant molecules dispersed in water. Such systems have gained importance because of their ability to behave as nanoreactors for carrying out organic reactions, including cross-coupling ones among the others.

The specific object of our study has concerned the non-ionic surfactant Kolliphor-EL (K-EL) and the Suzuki–Miyaura reaction (*Figure A1.1*) between an aryl halide **(1)** and a phenyl boronic acid **(2)** which, in the presence of a palladium(0) complex as catalyst **(3)**, give the cross-coupling product **(4)**. This reaction was performed by the research group of Prof. Luca Beverina (Department of Materials Science-UNIMIB).

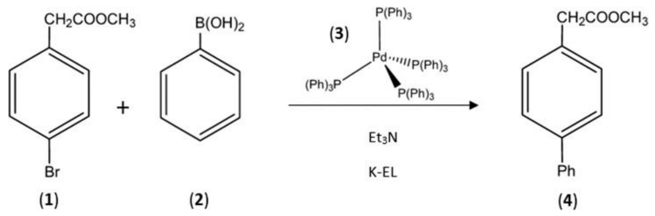


Figure A1.1 – Suzuki–Miyaura Cross-Coupling Reaction considered in the study. Figure reproduced from reference[1].

Starting from the K-EL molecules dispersed in water, a micelle model has been generated by MD simulations, adopting the 2016H66 force field. For each species involved in the reaction, six SMD simulations starting from different points around the micelle have been carried out, in order to account for anisotropy of the micelle. US simulations have then been performed on the configurations identified on these coordinates. Finally, an overall PMF profile for the transfer process between the two phases has been generated by using the WHAM.

The overall picture emerging from these results confirms that the molecular species involved in this reaction prefers the micellar environment and concentrates in different but close zones of the micelle, as it can be seen from the minima of the PMF profiles (*Figure A1.2*).

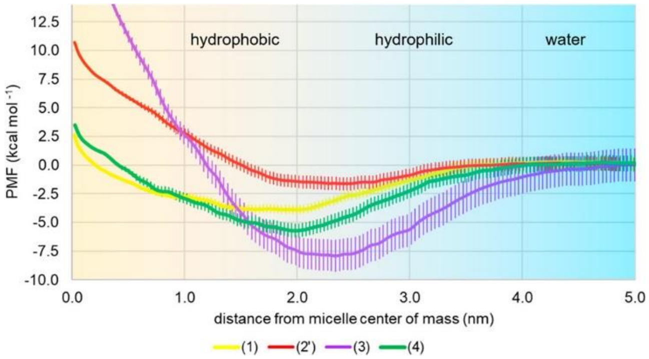


Figure A1.2 – PMF calculated for the four species by MD/US by six independent simulations starting from different positions of the solute with respect to the micelle. Standard deviation bars are reported. Figure reproduced from reference[1].

These results support the experimental evidence that the use of suitable surfactant agents promotes reactivity, allowing micelles to behave as nanoreactors in which reactive species are solubilized and enhance their local concentration.

Calix[4]arene-Based Sensitizers for Host-Guest Supramolecular Dyads for Solar Energy Conversion in Photoelectrochemical Cells

The photogeneration of electricity and solar fuels by solar irradiation in photoelectrochemical cells is one of the sectors with the highest growth potential in the decarbonised society. However, the use of different components, in particular photosensitizers and catalysts, can present problems of charge transfer efficiency at the interface, leading to lower final efficiencies. The research group of Prof. Alessandro Abbotto (Department of Materials Science-UNIMIB) designed novel integrated photosensitizer-catalyst dyads based on robust and flexible host-guest non-covalent interactions through the use of calix[4]arene cavities. A pictorial representation of the process is shown in *Figure A1.3*.

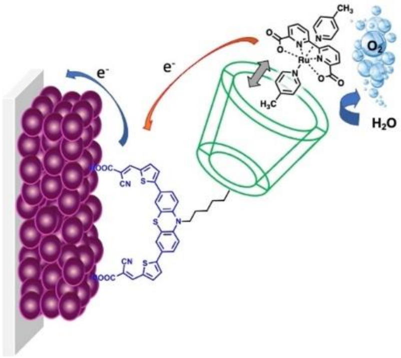


Figure A1.3 – Pictorial representation of the calix[4]arene moiety incorporated in the structure of a di-branched D-(π-A)2 metal-free organic sensitizer, anchored on a TiO_2 surface, to exploit host-guest interaction with a proper Ru(II) water oxidation catalyst. This should enhance the water oxidation in a photoactive device. Figure reproduced from reference[2].

Current photogeneration in photoelectrochemical cells showed greater efficiency in the integrated calixarene-based host-guest dyads compared to the traditional architecture based on the separate photosensitizer-catalyst pair. However, this efficiency depends on the photosensitizer used (Calix-PTZ and Calix-PTZ_2).

With the aim of rationalizing their different behaviour, Molecular Dynamics (MD) simulations of the two photosensitizers containing the catalyst in their cavities and anchored to the [101] anatase surface were performed.

Along the MD trajectories, a series of geometrical parameters able to affect the efficiency of the electron transfer processes were sampled: a) the distances between the centre of the thiophene rings (area of the LUMO, as identified from quantum mechanical calculations) and the TiO_2 surface; b) the angles between the planes defined by the thiophene rings and the TiO_2 surface, assumed as representative of the orientation of the LUMO with respect to the TiO_2 surface; and c) the distance between the centre of the phenothiazine ring (area of the HOMO) and the ruthenium atom of the catalyst. The main result obtained shows that the probability for the LUMO to approach the surface is higher for Calix-PTZ than for Calix-PTZ$_2$, thanks to the greater mobility of the former. This could be one of the reasons of the higher photocurrent density observed for Calix-PTZ.

References

(1) Ranaudo, A.; Greco, C.; Moro, G.; Zucchi, A.; Mattiello, S.; Beverina, L.; Cosentino, U. Partition of the Reactive Species of the Suzuki–Miyaura Reaction between Aqueous and Micellar Environments. *J. Phys. Chem. B* **2022**, *126* (45), 9408–9416. https://doi.org/10.1021/acs.jpcb.2c04591.

(2) Decavoli, C.; Boldrini, C. L.; Faroldi, F.; Baldini, L.; Sansone, F.; Ranaudo, A.; Greco, C.; Cosentino, U.; Moro, G.; Manfredi, N.; Abbotto, A. Calix[4]Arene-Based Sensitizers for Host-Guest Supramolecular Dyads for Solar Energy Conversion in Photoelectrochemical Cells. *European J. Org. Chem.* **2022**, *2022* (34). https://doi.org/10.1002/ejoc.202200649.

Appendix 2 – Additional data

Appendix 2.1 – MD simulations of affitins Sac7d, Affitin_1 and Affitin_2 with the AMBER99SB-ILDN force field.

The root-mean-square fluctuation (RMSF) of the backbone atoms (*Figure A2.1*) and the fraction of secondary structure elements (*Table A2.1* and *Figure A2.2*) shows that performing MD simulations with the AMBER99SB-ILDN force field leads to an even higher stability of all the affitins, with respect to what is observed with the Gromos 53A6 force field.

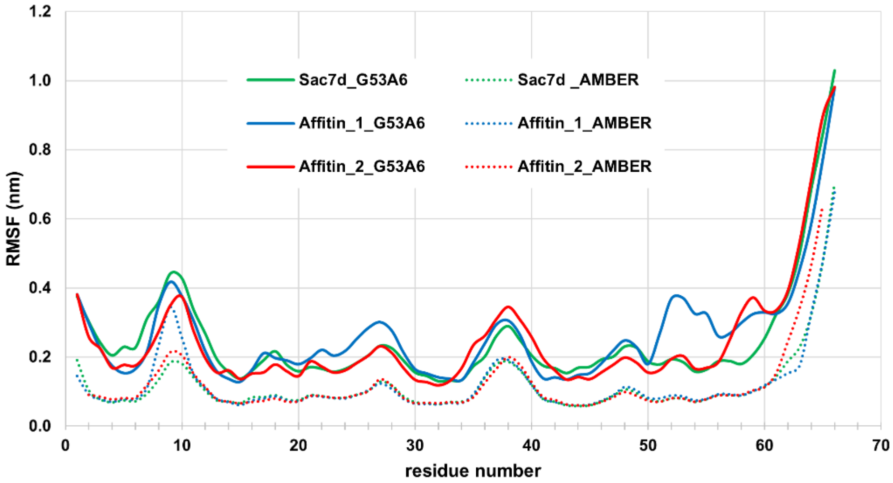


Figure A2.1 - Root mean square fluctuations (RMSF) of backbone atoms of each residue calculated for affitins Sac7d, Affitin_1, and Affitin_3 on the cumulative trajectories of the MD simulations carried out with Gromos 53A6 and AMBER99SB-ILDN force fields.

		Structure	Coil	**β-Sheet**	**β-Bridge**	Bend	**Turn**	**α-Helix**	5-Helix	3-Helix
Sac7d	**G53A6**	0.73	0.16	0.46	0	0.09	0.13	0.14	0	0.01
	AMBER	0.85	0.09	0.52	0	0.04	0.18	0.15	0	0.02
Affitin_1	**G53A6**	0.67	0.18	0.47	0.01	0.11	0.12	0.07	0.02	0.02
	AMBER	0.82	0.11	0.48	0	0.04	0.17	0.16	0	0.03
Affitin_2	**G53A6**	0.65	0.19	0.43	0.01	0.12	0.11	0.1	0.01	0.03
	AMBER	0.84	0.08	0.53	0	0.02	0.16	0.15	0	0.05

Table A2.1 - Secondary structure elements of wild type affitin Sac7d and of the Affitin_1 and Affitin_2 calculated with the Define Secondary Structure of Proteins (DSSP) algorithm on the cumulative trajectories of the MD simulations carried out with Gromos 53A6 and AMBER99SB-ILDN force fields. The term "Structure" refers to the sum of α-helix, β-sheet, β-bridge and turn elements.

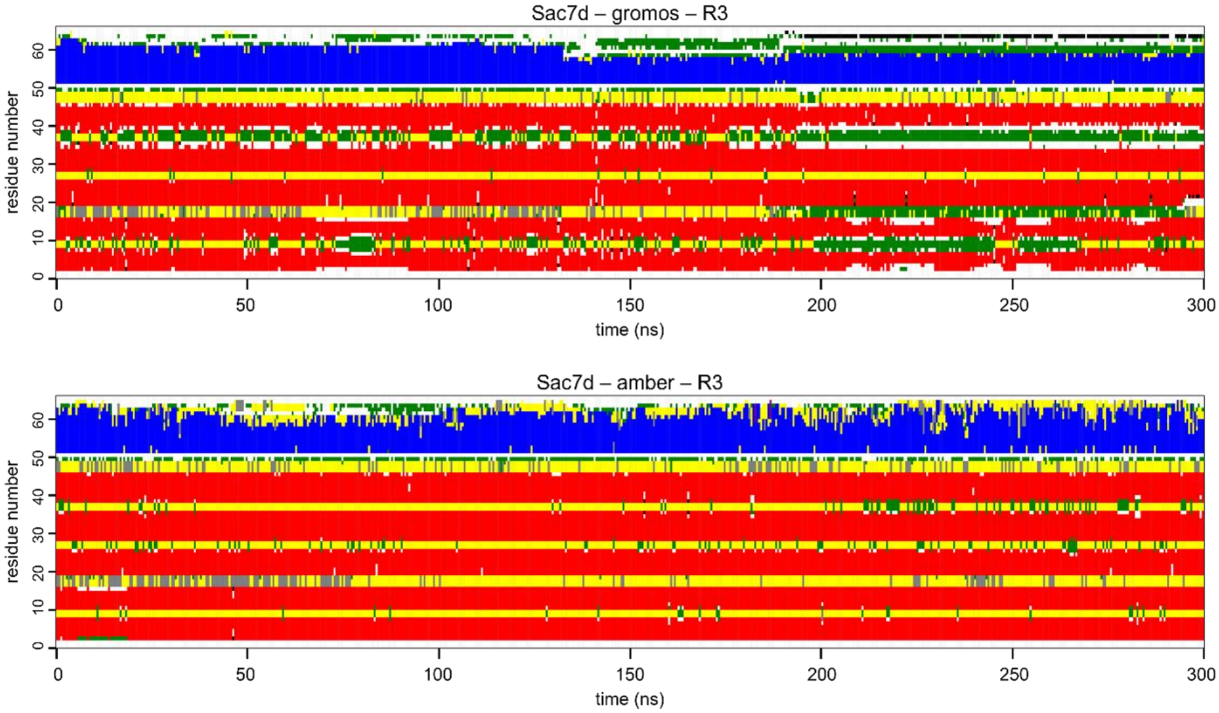


Figure A2.2 - DSSP analysis of affitin Sac7d along the MD simulation carried out with Gromos53A6 (top panel) and AMBER99SB-ILDN (bottom panel) force fields.

Appendix 2.2 – MD simulations of affitins retrieved from the PDB and of the ones designed *in silico.*

MD simulations with the Gromos 53A6 force field were carried out also for the affitins retrieved from the PDB (*Table 3.1*) and the ones designed *in silico* (*Figure 3.3*). The centrotype of the most populated clusters (*Figure 3.8*) show that all the affitins have a similar fold. The RMSF (*Figure A2.3* and *Figure A2.4*) and the fraction of secondary structure elements (*Table A2.2* and *Table A2.3*) confirm this behaviour.

Affitins retrieved from the PDB

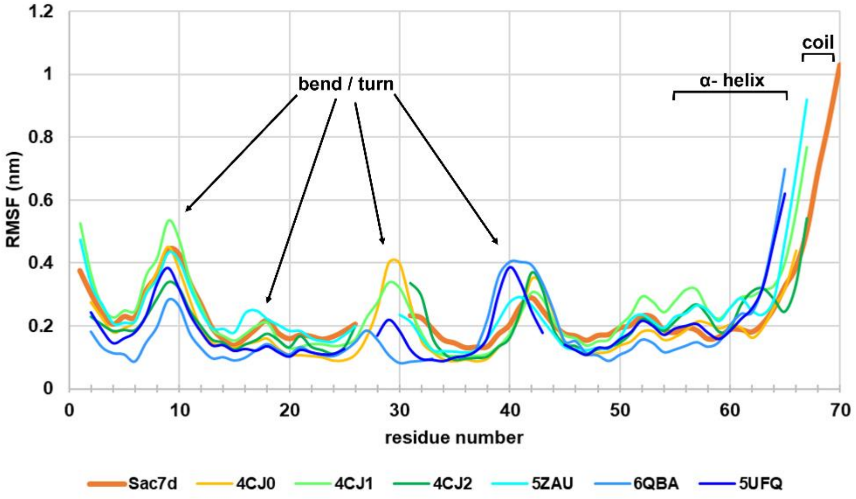


Figure A2.3 - Root mean square fluctuations (RMSF) of backbone atoms of each residue calculated, for the affitins retrieved from the PDB, on the cumulative trajectories of the MD simulations carried out with Gromos53A6 force field. Comparison is shown with the wild type Sac7d.

	Structure	Coil	β-sheet	Bend	Turn	α-helix
Sac7d	0.73	0.16	0.46	0.09	0.13	0.14
4CJ0	0.65	0.17	0.5	0.14	0.13	0.02
4CJ1	0.68	0.18	0.47	0.11	0.10	0.1
4CJ2	0.74	0.15	0.47	0.09	0.14	0.12
5ZAU	0.68	0.18	0.48	0.11	0.12	0.07
6QBA	0.63	0.21	0.40	0.13	0.15	0.08
5UFQ	0.65	0.20	0.43	0.11	0.13	0.08

Table A2.2 - Secondary structure elements of wild type affitin Sac7d and of the affitins retrieved from the PDB, calculated with the Define Secondary Structure of Proteins (DSSP) algorithm on the cumulative trajectories of the MD simulations carried out with the Gromos53A6 force field. The term "Structure" refers to the sum of α-helix, β-sheet, β-bridge and turn elements. β-bridge, 5-helix and 3-helix fractions are not shown as they are close to zero.

Affitins designed *in silico*

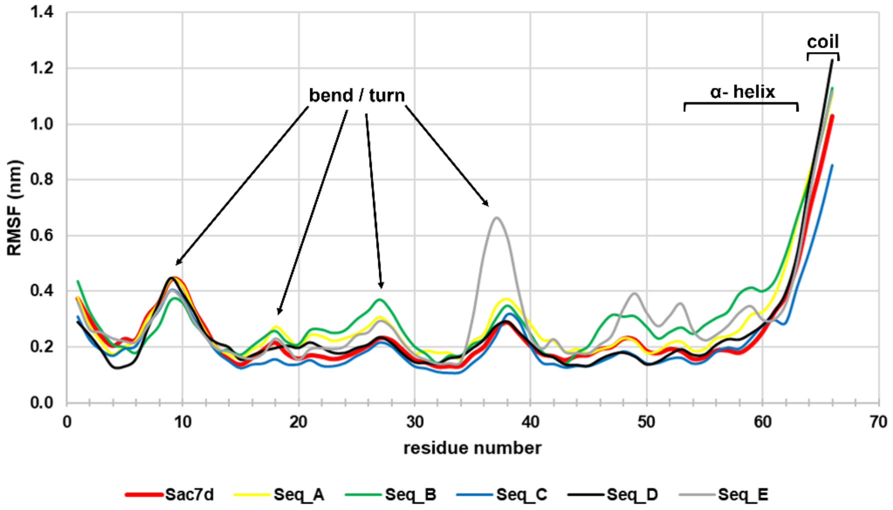


Figure A2.4 - Root mean square fluctuations (RMSF) of backbone atoms of each residue calculated, for the affitins designed in silico, on the cumulative trajectories of the MD simulations carried out with Gromos53A6 force field. Comparison is shown with the wild type Sac7d.

	Structure	Coil	**β-sheet**	Bend	**Turn**	**α-helix**
Sac7d	0.73	0.16	0.46	0.09	0.13	0.14
Seq_A	0.66	0.22	**0.42**	0.11	0.12	0.11
Seq_B	0.71	0.17	0.48	0.09	0.13	0.10
Seq_C	0.71	0.17	0.46	0.09	0.12	0.13
Seq_D	0.69	0.20	0.46	0.10	0.10	0.12
Seq_E	0.64	0.21	**0.38**	0.14	0.13	0.12

Table A2.3 - Secondary structure elements of wild type affitin Sac7d and of the affitins designed in silico, calculated with the Define Secondary Structure of Proteins (DSSP) algorithm on the cumulative trajectories of the MD simulations carried out with the Gromos53A6 force field. The term "Structure" refers to the sum of α-helix, β-sheet, β-bridge and turn elements. β-bridge, 5-helix and 3-helix fractions are not shown as they are close to zero.

Appendix 2.3 – Performance of the four ClusPro scoring schemes in the prediction of the complexes affitins-protein partners

Reranking of the first ten docking poses (labelled as in the docking server from best (0) to worst (9)) for the four scoring schemes on the basis of crystal_RMSD. The best performing scoring scheme is the “balanced” one, as for 6 out of 7 complexes the best docking pose found by the server is the closest to the crystallographic structure (lowest crystal_RMSD value). This number is 5, 4 and 1 for the "electrostatics", "hydrophobic" and "vdW+el" scoring schemes, respectively.

"balanced" scoring scheme

4CJ0	4CJ1	4CJ2	5ZAU	6QBA	5UFE	5UFQ
0	0	0	0	2	0	0
4	1	4	3	0	3	4
9	2	3	2	1	7	8
7	4	6	6	8	2	5
2	5	2	1	9	5	7
6	6	7	5	4	4	9
3	3	9	7	3	1	1
8	7	1	9	5	6	3
5	9	5	4	7	8	2
1	8	8	8	6	9	6

"electrostatics" scoring scheme

4CJ0	4CJ1	4CJ2	5ZAU	6QBA	5UFE	5UFQ
0	0	0	0	1	2	0
5	7	8	2	0	0	9
3	3	9	5	2	7	8
6	2	1	3	7	3	3
2	4	5	1	3	6	6
4	5	7	6	8	4	1
9	1	2	9	4	8	4
1	9	3	7	5	1	5
7	6	6	8	6	5	2
8	8	4	4	9	9	7

"hydrophobic" scoring scheme

4CJ0	4CJ1	4CJ2	5ZAU	6QBA	5UFE	5UFQ
0	0	0	1	3	0	2
1	1	2	4	0	3	0
7	4	3	2	6	8	5
2	2	6	6	1	2	9
5	3	1	9	4	6	7
3	5	8	7	2	5	4
4		4	3	5	1	1
6		7	0	7	4	6
9		5	5		7	3
8		9	8		9	8

"vdW+el" scoring scheme

4CJ0	4CJ1	4CJ2	5ZAU	6QBA	5UFE	5UFQ
4	0	2	9	5	1	3
0	3	5	1	8	6	0
9	7	1	8	9	8	6
3	8	3	5	6	7	5
7	9	9	7	1	3	4
5	4	0	2	3	9	7
6	6	8	3	0	0	2
2	1	4	0	2	5	8
8	2	6	4	7	4	1
1	5	7	6	4	2	9