Fragment-based modelling of protein-RNA complexes for protein design

Anna Kravchenko

UNIVERSITÉ DE LORRAINE

École doctorale IAEM Lorraine

Fragment-based modelling of protein-RNA complexes for protein design

THÈSE

présentée et soutenue publiquement le 20 December 2023

pour l'obtention du

Doctorat de l'Université de Lorraine

(mention informatique)

par

Anna Kravchenko

Composition du jury

Président :	Samuela Pasquali	Professeure, Université Paris Cité
Rapporteurs :	Juan Cortes	DR CNRS, LAAS, Toulouse
	Pablo Chacon	Chercheur, Institute of Physical Chemistry, Madrid
	Piotr Setny	Chercheur, University of Warsaw
Examinateur :	Martin Zacharias	Professeur, Technical University of Munich
Encadrants :	Isaure Chauvot de Beauchêne	CR CNRS (*HDR*), Nancy
	Malika Smaïl-Tabbone	MDC HC, Nancy

Laboratoire Lorrain de Recherche en Informatique et ses Applications — UMR 7503

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No. 813239

Experiments presented in this thesis were carried out using the Grid'5000 experimental testbed, being developed under the INRIA ALADDIN development action with support from CNRS, RENATER and several Universities as well as other funding bodies (see https://www.grid5000.fr)

Acknowledgements

I express my appreciation to the members of my thesis jury, **Pablo Chacon**, **Juan Cortes**, **Piotr Setny**, **Samuela Pasquali** and **Martin Zacharias** for the time and expertise you dedicated to reviewing my thesis.

As the pages of this thesis unfold, I find myself reflecting on the myriad of individuals who have accompanied me at different points during this academic endeavour. My gratitude knows no bounds for **Isaure Chauvot de Beauchêne,** to whom I owe these four exciting years. Thank you, Isaure, for entrusting me with this PhD project. Thank you for patiently elucidating the fundamentals of molecular biology and delving into the intricacies of protein-ssRNA docking. Your scientific enthusiasm, dedication, discernment, and intelligence provide an excellent example for an aspiring scientist like myself, and your ability to generate spot-on sentences in seconds is a skill I aim to have one day. I am thankful for the freedom and support you provided, as well as the generous amount of time invested in my project. Thank you for 'holding my hand' while I was submitting my first conference abstract, my first paper, and, well, this thesis. Thank you for tirelessly correcting all aforementioned texts, these efforts haven't gone unnoticed. My admiration for your supervision is only comparable to the pleasure I had in working with you. Finally, thank you for bringing rock climbing into my life. I am deeply grateful for the support my family and I received from your family during the darkest time of the Russian invasion of Ukraine. This thesis and the person I have become in the process owe a great deal to you.

I am profoundly grateful for the encouragement and wisdom of **Malika Smaïl-Tabbone**. Malika, thank you for taking care of my PhD at a time when I needed it. Your support, particularly but not limited to the navigation of numerous administrative questions, has been very helpful throughout my PhD years. Our collaborative efforts, marked by numerous meetings, were not only productive but also enlightening, thanks to the wealth of insightful suggestions and advice you generously provided. Your constructive feedback has been invaluable to my research growth and your meticulous attention to detail has not gone unseen. I extend my gratitude for your guidance and I very much appreciate the positive impact you had on my research skills and overall academic growth.

I would like to extend my sincere thanks to **Sjoerd Jacob de Vries**, who has been attentive to my research from its inception. Sjoerd, I extend my heartfelt thanks for your patience, even when faced with the need to repeat every second sentence approximately 42% of the time. Your crucial contributions to coding HIPPO have played a pivotal role in the success of this project. Our informal lunch discussions have been very enjoyable.

Special thanks are due to **Patrice Ringot**, the best technical support person I have encountered thus far. Thank you for always being kind to me and explaining things in an understandable way. Thank you for the introduction to the Grid'5000, and thank you for dealing with all the cluster memory issues I have caused during these 4 years. It was always a pleasure to communicate with you, I appreciate you and your support immensely.

Appreciation is owed to my co-bureau and 'work wife', **Antoine Moniot**, who has graciously responded to an incalculable number of questions and posed some queries that led to various insights. Antoine, I appreciate your guidance in familiarising me with Linux and your invaluable assistance in curating the benchmark. Thank you for your unwavering support and consistently positive attitude. My PhD years would have been significantly duller without you.

Many thanks to **Hrishikesh Dhondge**, with whom we navigated the PhD journey together, offering mutual support. Hrishi, it was a pleasure to explore new aspects together, both scientific and administrative. Although our main focus was often on individual projects, collaborating closely with

you on the 'RRM-RNA dock' was a delightful experience. I wish you success in your future research, and I hope we can collaborate on many more projects together.

I must acknowledge **Yann Guermeur**, even though our interactions were relatively brief. Yann, thank you for our discussions; they held significant importance for my mindset. I am sincerely grateful for the time and attention you devoted to me and my research.

I would like to thank the members of the CAPSID team, past and present. **Marie-Dominique**, **Emmanuel**, **Philippe**, **Claire**, **Kevin**, **Bishnu**, **Bernard**, **Sabeur**, **Yasaman** and **Hamed** – thank you for our engaging 'tea-time' discussions, valuable ideas, advice, and the annual outings. Your collective contributions have created and maintained a great atmosphere over the years.

I express my gratitude to the members of the RNAct consortium for the synergy, support, motivation and positivity. **Wim**, **Thomasso**, **Michael**, **Jos** and **Guillermo**, your wisdom and feedback have been invaluable. **Jose**, **Roswita**, **Guillermo**, and **Stefano**, thank you for the camaraderie that made this collaborative effort enjoyable. **Anna P.**, your high-quality theoretical biology support is greatly appreciated. **Joel**, thank you for all the versions of the RRMScorer. I offer special thanks to the members of Dynamic Biosensors, **Wolfgang** and **Anahi**, for hosting me during the secondment. I highly appreciate the efforts of Anahi, **Nikki**, and **Luca** in guiding me through wet experiments. The wet lab experience you created for me added a unique dimension to my PhD. Last but not least, **Aitor**, thank you for the positive attitude and your aid throughout the communication cycle.

I extend my gratitude to the personnel of the LORIA/INRIA lab, whose positive impact enriched my PhD journey. Special thanks to **Sylvie Hilbert** for providing quality support with administrative procedures. **Antoinette Courrier**, despite our language barrier, I appreciate your assistance with my frequent missions, especially your help with the finances for the mission in Munich. **Isabelle**, your welcoming demeanour, kindness, and infectious good humour always made a difference. **Caro**, thank you for supplying the essential brain fuel, 'double ristretto avec du lait', and for your morning smiles. **Florianne**, thank you for the positivity and our chats every so often, and **Marie Baron**, thank you for your help in organizing multiple collections for donations to Ukraine. **Aurore Tranchina**, thank you for being very patient with me during the thesis submission process.

On a more personal level, I want to express my heartfelt gratitude to the friends I made during this time. **Dom**, my appreciation for you is immense, and I'm thankful for a very long list of things. I hope the Frenglish Institute's legacy will endure indefinitely. **Lisa**, thank you for hosting delightful lunches and dinners in the Blue House, your carrot cake is truly out of this world. **Athénaïs**, you significantly enriched my social life, and I'm grateful for all the coffees we shared after lunch and the MacDo dinners we enjoyed after work (and so much more). **Hans**, thank you for the juggling classes and our insightful discussions on various topics. **Alina** and **Vlad**, I appreciate your company and the time you drove me around Munich. **Valik**, thank you for your kindness and for letting me crash on your couch in Munich. **Ira**, our countless hours spent in virtual meetings were incredibly beneficial for me mentally. **Andrii**, **Anastasiia**, **Vilalii**, and all the others whose names would take a long time to mention individually, please know that I haven't forgotten your impact on my journey.

Heartfelt appreciation goes to **Thomas**, who provided relentless support, warmth, and nourishment throughout the writing of this manuscript. Your boundless understanding and relentless support mean the world to me. I am eternally grateful to have you in my life.

I find it challenging to put into words the immense gratitude I feel towards **my family**. Thank you for being the pillars of my life and for sharing in both the triumphs and trials. My achievements are, in many ways, a reflection of the values and strength you instilled in me.

Finally, I would like to express my thankfulness to my university professor, **Ludmila Kovalchuk-Khymiuk**, who approached me about five and a half years ago and asked,

'*Would you like to go to France? You know, they have a great Data Science program in Nancy.*'

If I have seen further it is by standing on the shoulders of Giants.
– Isaac Newton

Contents

Part II Contributions

Appendices

General Introduction

Knowledge of three-dimensional (3D) structures of molecular complexes, particularly those involving proteins and RNA, is important for biological research. This knowledge not only aids in elucidating life processes at the cellular level but also has far-reaching practical applications, including the development of new and/or personalised treatments, genetic engineering, and protein design, offering a path to a better future.

Protein-RNA 3D structures can be obtained either via 'wet-lab' experiments or computational modelling. 'Wet' experiments generally are very reliable, but they come with a hefty price tag and require a team of specialists, specialised equipment, biomaterials etc. In contrast, computational modelling is often a more cost-effective and efficient approach. However, it's not without its challenges, as the reliability of computational models is lower compared to their 'wet' experimental counterparts. This thesis focuses on improving the reliability of computational modelling of 3D structures of single-stranded (ss) RNA bound to proteins.

Context and Motivation

My PhD project is a part of the Marie Skłodowska-Curie Innovative Training Network "RNAct project". The research aim of the RNAct project is to design novel RNA recognition motif (RRM) proteins for exploitation in synthetic biology and bio-analytics. Ten Early Stage Researchers (ESRs, PhD students) combined efforts in both computational and experimental biology to achieve this goal within the RNAct project (Fig. 1). My role (ESR4) in the frame of this project is to enhance the method to dock RRMs to ssRNA, i.e. model 3D structure of RRM-ssRNA complex.

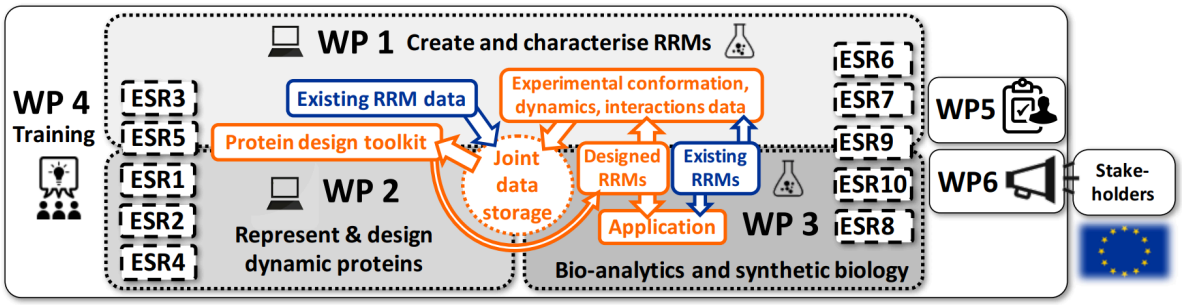


Figure 1 - Overview of the work packages (WP) and ESR involvement within the RNAct project. Image by Wim Vranken taken from the RNAct proposal.

Docking protein/RRM-ssRNA complexes is notoriously challenging. This complexity emerges from the inherent flexibility of ssRNA chains, leading to a very large conformational search space. Due to the disordered state of ssRNA, modelling protein-ssRNA complexes cannot be done by classical rigid or semi-flexible docking. A relatively small number of solved protein-ssRNA complexes impairs the application of machine/deep learning.

This problem can only be addressed using fragment-based docking approaches, with ssRNA'TTRACT being the current state-of-the-art method. It consists of docking and assembling various 3D conformations of RNA fragments on the protein, using a combinatorial approach. It is conducted under two fundamental assumptions: (i) the 3D structure of the protein is known, and (ii)

the sequence and secondary structure of the RNA chain interacting with the protein is known, i.e. only the single-stranded part of the chain is taken into consideration.

The current protein-RNA energy parameters of the ATTRACT scoring function (ASF) are not specific to ssRNA and could be optimised. In addition, in the specific case of RRM-ssRNA domains, consistent RRM topology allows for statistical analysis which can provide docking constraints for data-driven docking. My PhD project has 2 corresponding main axes.

Axis 1: A New Approach to Develop Protein-ssRNA Scoring Functions

Fragment-based docking encounters scoring issues, partly due to variations in the binding strength of different RNA fragments. Some fragments bind more strongly to proteins than others, leading to issues with ranking near-native poses, especially ones with a weaker binding strength. We have developed a novel approach that addresses this issue by producing several scoring potentials ($\mathcal{H}$), capable of accounting for different binding modes. These scoring potentials are based on the frequency of bead-bead distances in near-native versus non-native poses. To streamline the process and to stay in the frame of the RNAct project, I focused on the RRM domain, deriving HIPPO (HIstogram-based Pseudo-POtential), comprising four distinct parameter sets ($\mathcal{H}$) for scoring RRM-ssRNA poses in ATTRACT coarse-grained representation. HIPPO enriches the number of near-natives in the top-ranked 20% of poses by 3 to 4-fold, outperforming ASF. HIPPO also outperforms ASF in the general case of protein-ssRNA docking, proving its generalisability. Lastly, this protocol is in principle applicable to any ligand type in (pseudo)atom-based representation.

Axis 2: Incorporating Database Knowledge as Constraints in RRM-ssRNA Docking

With my colleague Hrishikesh Dhondge (ESR3), we have developed an anchor-driven fragment-based docking pipeline for RRM-ssRNA docking, as an updated version of an already existing strategy. RRMs have two amino acids in conserved positions, each performing an aromatic stacking with a nucleotide of the bound ssRNA. Those amino acids can be used as anchor points in the docking. We collected all RRM-ssRNA experimental structures with such stacking contacts, extracted the local protein backbone and stacked nucleotide, and clustered them to obtain a set of prototype positions of a stacked nucleotide toward the local backbone of a stacking amino acid. I then set up an RRM-ssRNA docking pipeline using the ATTRACT docking engine, with the RRM sequence, RNA sequence, and identification of the stacked nucleotides as input. The pipeline retrieves the RRM structure from AlphaFoldDB, identifies possible 3D positions of the stacked nucleotides, and runs ATTRACT docking of RNA fragments with maximal distance restraints toward each possible position.

Outline of the Manuscript

Chapter 1 gives an introduction to the biological concepts that are prerequisites for the computational modelling of protein-ssRNA complexes. It covers basic aspects of protein and RNA molecules (sequence, structure, flexibility), delves into primary interactions between them, and explores potential conformational changes that may take place before or during the binding process. The chapter lists common RNA-binding protein domains, with a focus on RRMs and the specificities of their binding with ssRNA. This chapter concludes with a brief overview of structural experimental techniques for solving protein-ssRNA complexes.

Chapter 2 aims to cover parts of the field of computational structural modelling relevant to this thesis. It begins with a concise overview of databases containing pertinent data for protein-ssRNA modelling and outlines approaches useful for modelling single protein and RNA chains. The chapter

then proceeds to explain fundamental principles and categories of docking, i.e. modelling of complexes, with a particular emphasis on rigid-body docking. This includes a discussion on molecular representation, sampling approaches, and commonly employed scoring functions. Additionally, the chapter introduces several knowledge-based protein-RNA scoring functions, each constructed following a distinct approach. The final section delves into protein-ssRNA docking techniques, with a detailed description of the state-of-the-art method, ssRNA'TTRACT.

Chapter 3 begins the presentation of the original contribution of the thesis. The first part of this chapter details the optimisation of the original ASF parameters through a stochastic Monte Carlo Simulated Annealing approach. Although the resulting parameter sets did not surpass ASF in performance, key insights were gained, ultimately leading to fruitful developments (Chapter 4). The second part of Chapter 2 delves into the examination of the inadequate values of the ASF parameters for Tryptophan-Cytosine interactions. While manual fine-tuning of these values proved unsuccessful, this section allows for a discussion of the systematic evaluation of the parameter subsets and the stacking problem in the context of scoring.

Chapter 4 presents the pivotal contributions of the thesis, made in the frame of Axis 1. It begins with the concise presentation of a preliminary histogram-based approach and is followed by a detailed description of the original protocol used to derive HIPPO, which is presented as a stand-alone research paper. Thereafter, the performance of the scoring functions ASF, HIPPO and BP (usage of the best-performing $\mathcal{H}$ out of 4 for each fragment) on a new benchmark of experimentally solved protein-ssRNA complexes is discussed. Lastly, an incremental fragment assembly is performed on selected complexes using BP and ASF, accompanied by a performance evaluation.

Chapter 5 provides an overview of the contributions made in the frame of Axis 2. It begins with a presentation of the previously developed anchor-docking strategy and principles of the creation of anchoring patterns. Subsequently, the original RRM-ssRNA docking pipeline is introduced, followed by the evaluation of its performance against *ab initio* docking. The final section introduces experimental non-structural data extracted from literature as a potential source of docking restraints.

As a final point, **Chapter 6** summarises the contributions of this thesis, and presents several promising directions for further developments along with scientific prospects.

Part I

Context and State-of-the-Art

Chapter 1: Biological Background of Protein-RNA Complexes

1.1 Aims

In this chapter we provide a preliminary introduction to protein and ribonucleic acid (RNA) macromolecules and their interactions, focusing on the 3-dimensional (3D) structures and the significance of the knowledge of these structures in the context of modern biology. We begin with the basic features of these macromolecules, then we investigate the foundation of the protein-RNA interactions. Additionally, we provide a brief overview of the experimental techniques utilised to acquire structural data on these protein-RNA complexes. This allows us to transition from the concept of macromolecules to their computational representation and delve into the field of bioinformatics.

1.2 Macromolecules

Life is widely recognised to depend on nucleic acids and proteins, two essential macromolecules composed of long chains of covalently linked monomeric units. These long chains adopt intricate shapes governed by the principles of atomic physics and chemical interactions. The specific 3D structures of macromolecules dictate their functions, making the investigation of structure-function relationships a crucial endeavour. A multitude of interconnected fields, including structural biology, biochemistry, molecular biology, computational biology and bioinformatics collaboratively explore the intricate world of macromolecules to elucidate their roles in enzymatic activity, signal transduction, molecular recognition, and cellular regulation, and, in a more practical context, in disease mechanisms investigation, drug design and development, and rational design.

1.2.1 Proteins

Among organic macromolecules, proteins are one of the most abundant and functionally versatile. A protein molecule can be described in a hierarchical manner (Fig 1.1):

- The primary structure is simply a linear sequence of amino acids, that make up the long protein chain;
- The secondary structure is a 3D folding of the continuous segment of the protein chain into repetitive patterns (e.g. helix), occurring due to the interactions between the amino acids in the chain;
- The tertiary structure is a 3D organisation of the whole protein chain, formed by packing elements of the secondary structure into one or several compact units;
- The quaternary structure is a 3D organisation of proteins when they bind to each other or/and to other (macro) molecules, forming complexes of 2 or more members.

Amino acids, the constituents of the protein chain, are small organic molecules composed of a central carbon atom (alpha carbon), bonded to four different chemical groups: an amino group (-NH2), a carboxyl group (-COOH), a hydrogen atom (-H), and a side chain or R-group, which varies among the different amino acids (Fig 1.2 a). While chemically there are hundreds of possible amino acids, human proteins predominantly incorporate a set of 20 standard amino acids. Amino acids are linked in the chain by a highly stable peptide bond. This covalent bond is formed between the carboxyl group of one amino acid and the amino group of the adjacent amino acid and is accompanied by the release of a water molecule (Fig 1.2 b, c). The end of the chain, which contains an amino acid with a free amino group, is called the N-terminus and is often referred to as the beginning of the chain. The opposite end of the chain is called C-terminus. The chain itself consists of the repeating units, which are classified as the backbone (main chain), the same for all amino acids; and side chains, different for different types of amino acids [2].

Amino acids can be classified based on the chemical properties of their side chains into distinct groups, including hydrophobic, aromatic, charged (positively or negatively) and polar amino acids. This classification is of high importance as it aids in explaining the folding of protein chains and general molecular interactions. For instance, the “hydrophobic effect” in proteins illustrates the role of hydrophobic and hydrophilic amino acids in protein folding and assists the formation of the most common elements of the secondary structure. In water-based environments where most proteins exist, hydrophobic amino acids tend to be packed in the protein’s core, while hydrophilic ones form the protein's surface. This arrangement allows for the establishment of a stable network between the protein and its surroundings, with both side chains and the backbone of hydrophilic amino acids

engaging in hydrogen bonding with water molecules at the surface. However, the interior regions of the backbone lack this direct interaction with water. To compensate for this, the polar and so hydrophilic backbone adopts secondary structures such as alpha-helices or beta-sheets, characterised by local hydrogen bonding within the protein interior, stabilising the corresponding region of the chain [3].

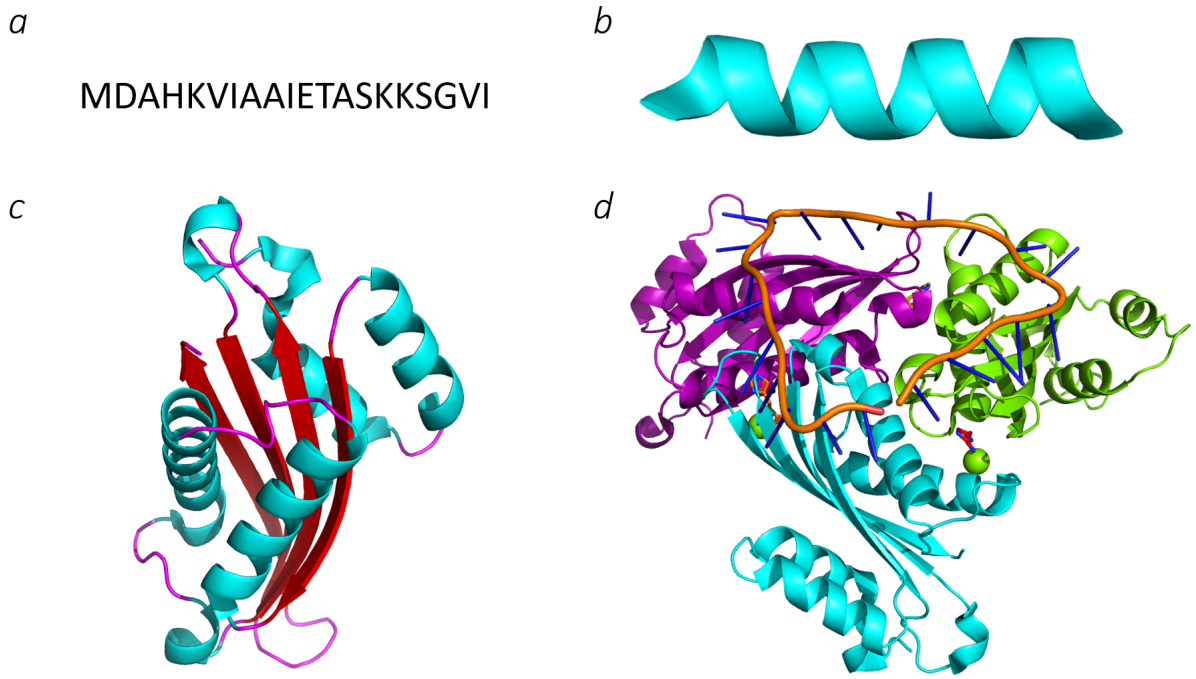


Figure 1.1 - Four levels of protein structure illustrated via the crystal structure of the HutP antitermination complex bound to the HUT mRNA, pdb code 3BOY. (a) Primary structure (sequence); (b) Secondary structure (alpha helix) of the given sequence; (c) Tertiary structure containing alpha helices (cyan), antiparallel beta sheets (red) and loops (magenta); (d) Quaternary structure containing 3 protein chains (magenta, green, cyan) and RNA (orange). Here and onward 3D models were made with PyMol [1] unless specified otherwise.

In an alpha-helix, the backbone forms a tightly coiled structure held in place by a hydrogen bond occurring between the peptide bond and the 4th amino acid of the consecutive chain. Such structure creates charged regions on the helix, a positive charge at one end and a negative charge at the other. This allows for the attraction of molecules with opposite charges, particularly those containing phosphate groups, which tend to bind near the positive end of the helix. In a beta-sheet, contrary to the alpha helix, the backbone forms a planar structure held in place by hydrogen bonds occurring between amino acids located in the adjacent strands of the chain. This type of secondary structure is more spacious, which increases accessibility for external molecules [4].

The elements of the secondary structure are interconnected by regions called loops. Unlike structured secondary elements, loops generally do not possess a defined secondary structure. Instead, they serve as flexible connectors between the secondary structure elements. Loops vary in length, but typically stay on the shorter side (3-20 amino acids) and have a higher propensity to interact with the surrounding environment compared to the amino acids within the protein chain.

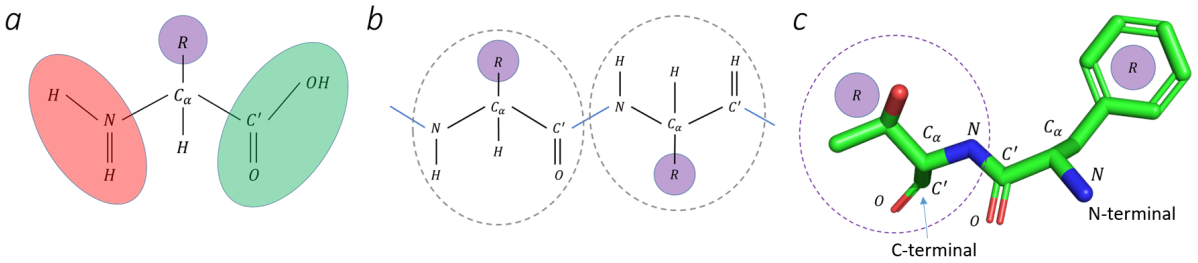


Figure 1.2 - Amino acid and chain of linked amino acids (a) Schematic diagram of amino acid its amino group (red), a carboxyl group (green), and side chain or R-group (purple); (b) Schematic diagram of a polypeptide chain made by two amino acids (highlighted by the dashed circles) connected via a peptide bond (blue); (c) 3D structure of a polypeptide chain made by amino acids phenylalanine and threonine. The latter one is highlighted by a dashed circle.

The tertiary structure - elements of the secondary structure connected via loops - is stabilised by various interactions, namely hydrophobic interactions, hydrogen bonds, ionic interactions, van der Waals interactions, disulphide bridges, stacking interactions and interactions with the surrounding solvent (see §1.3.1, as the mechanisms behind inter- and intra-molecular interactions are governed by the same physical principles). The resulting 3D structures tend to be energetically favourable. They can exhibit a wide range of characteristics, e.g. be symmetric or asymmetric, and have varying surface features such as smooth surfaces, deep cavities, or even open channels.

Some tertiary structures correspond to domains, which are compact (highly structured) and independently folded regions within proteins. Domains exhibit well-defined three-dimensional structures that align with known templates from a curated list of structured domains [5, 6]. Domains are considered functional units in the protein world, as they can perform specific enzymatic activities, interact with other molecules such as DNA or RNA (a concise list of the latter ones can be found in §1.3.4.1), participate in signalling pathways, or bind to other proteins. It is common for a single protein chain to contain multiple domains, which are interconnected by flexible regions known as linkers. Such a structure is known to increase stability and boost functionality [7].

The protein world is not limited to the known structured domains. Numerous structured domains are yet to be discovered and characterised, expanding the diversity of protein structures. Moreover, some proteins are characterised by remarkable conformational flexibility and lack a well-defined structure, yet still remain functional. They are known as intrinsically disordered proteins (IDPs). Intrinsically disordered regions (IDR) can be present in proteins that are otherwise structured, making “protein hybrids” [8]. The existence of structured domains, IDPs, and protein hybrids showcases the remarkable diversity within proteins accompanied by a wide range of abilities and functions (see section 1.3.4 RNA-Binding Proteins).

1.2.2 RNAs

RNA macromolecules, much like proteins, are fundamental components in cellular processes and a standalone subject in the field of molecular biology. Previously, RNAs were predominantly seen as minor characters in the flow of genetic information, primarily involved in facilitating protein synthesis through transcription and translation. Nowadays, we know that it was a cameo appearance — the versatility and diverse functionality of RNA are evident, despite the fact that the functions of many RNAs are still undiscovered. They act as regulatory molecules, exhibit catalytic activities as ribozymes, and play active roles in crucial cellular processes like RNA interference and gene expression regulation.

RNA consists of small organic molecules, called nucleotides, which are composed of ribose sugars attached to nitrogenous bases and phosphate groups. This group varies for different nucleotides. Among the many possible nucleotides, RNA is composed of 4 standard nucleotides: adenine (A), cytosine (C), guanine (G), and uracil (U) (Fig 1.3). Nucleotides are linked in a chain by a phosphodiester bond, a strong covalent bond between the 3'-carbon atom of one nucleotide's sugar and the 5'-carbon atom of the adjacent nucleotide's phosphate (Fig 1.4 a, b). This linkage creates a free 5'-position at one end of the chain and a free 3'-position at the other end. The 5'-end is considered the beginning of the chain [9].

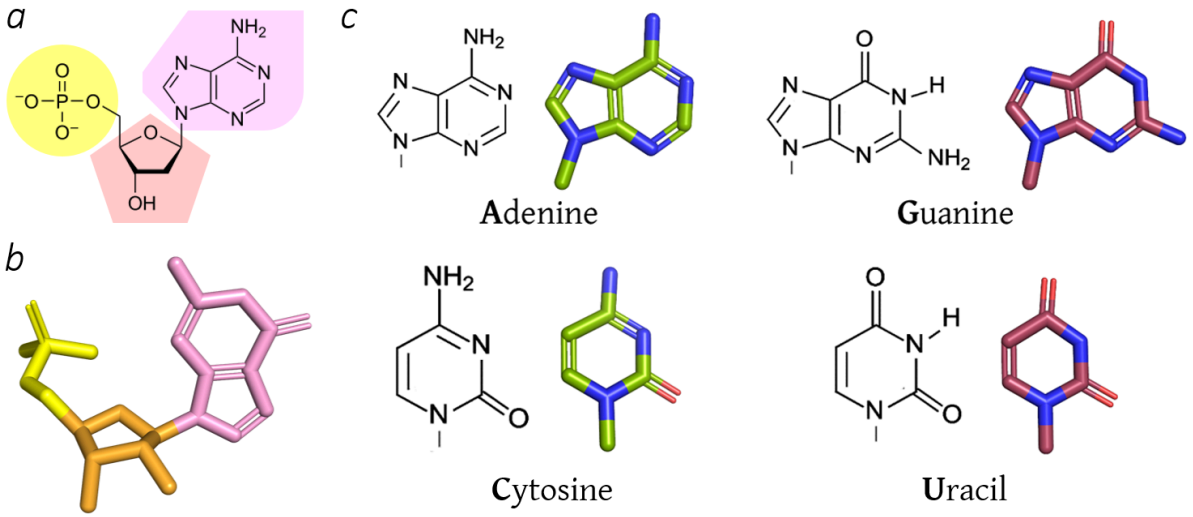


Figure 1.3 - Nucleotide and chain of nucleotides (a) Schematic diagram of a nucleotide with a phosphate group (yellow), ribose sugars (orange) and a nitrogenous base (pink); (b) 3D image of a nucleotide with a phosphate group (yellow), ribose sugars (orange) and a nitrogenous base (pink); (c) Nucleotide base types, schematically and in 3D

Nucleotides can be classified into 2 groups: purines (A, G) and pyrimidines (C, U), based on the types of nitrogenous bases they contain. Purines are larger and consist of two aromatic rings, while pyrimidines are smaller and consist of a single ring. This difference in size and structure leads to distinct binding preferences. For instance, purine-rich sequences have the ability to form specific secondary structures, such as G-quarters, which are formed by 4 G molecules interacting with each other, and subsequently, G-quadruplexes, formed by stacking of G-tetrads on top of each other [10, 11].

Similarly to proteins, RNA structure can be described at 4 levels of complexity (Fig 1.4 c, d, e and Fig 1.1 d). The primary structure of RNA is defined as a linear sequence of nucleotides, typically written from 5’-end to 3’-end. The secondary structure is defined by the local folding of the chain, occurring due to the interactions between nucleotides, primarily through base pairing - edge-to-edge hydrogen bonding interaction between two bases.

The most common and stable base pairs are Watson-Crick pairs, or canonical pairs, which involve geometric correspondence and the formation of 2 or 3 hydrogen bonds between bases A-U and C-G, respectively. A common non-canonical base pair is the "wobble" U-G pairing [12]. Other non-canonical base pairings do exist [13]. There are 12 classes (or families) of possible base pairs, constructed based on the geometry of the bonding patterns observed in RNA structures (Fig 1.5).

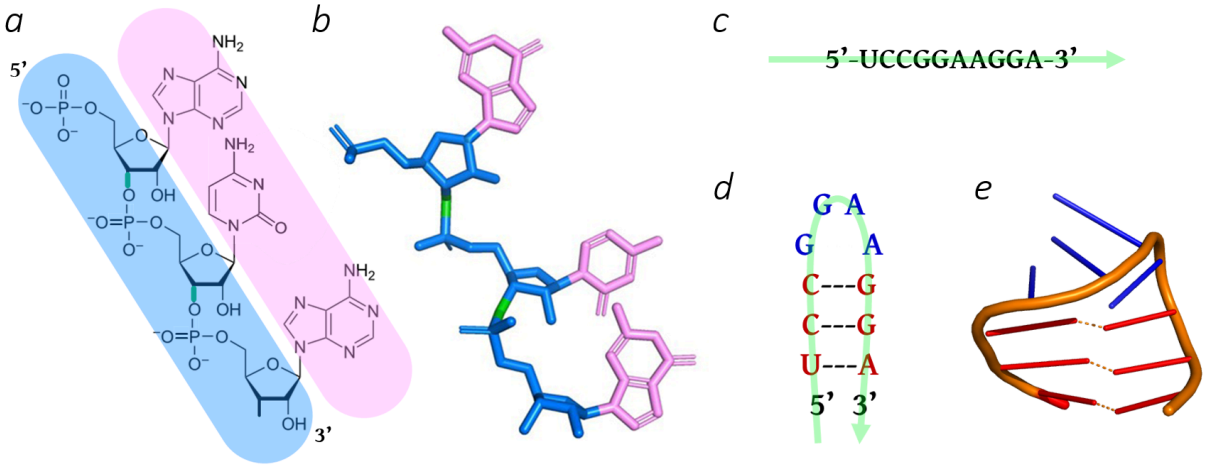


Figure 1.4 - RNA chain and 3 levels of its representation a) Schematic diagram of a chain of 3 nucleotides linked by phosphodiester bond (green). The backbone is displayed in blue, the side chain is in pink; b) a 3D image of a chain of 3 nucleotides linked by phosphodiester bonds (green). The backbone is displayed in blue and the side chain is in pink; c) primary structure of RNA; d) secondary structure of RNA in 2D showcasing single-stranded region (blue) and double-stranded region (red); e) tertiary structure of RNA in 3D showcasing single-stranded region (blue) and double-stranded region (red) (pdb code 28SP).

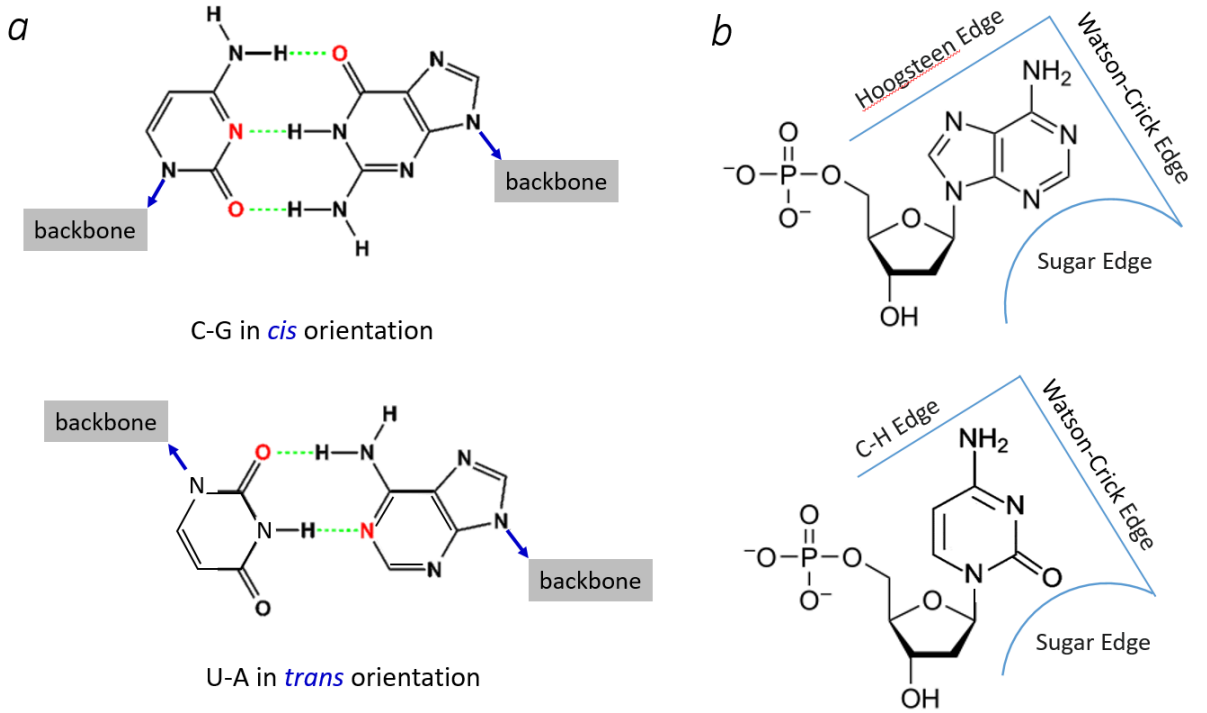


Figure - 1.5 RNA base pairings variety a) Watson-Crick base pairs between bases (top) C and G in *cis* orientation and (bottom) U and A in *trans* orientation. In the *cis*/*trans* orientation, the two bases involved in the pairing are located on the same/opposite side(s) of the backbone. Orientations are highlighted with blue arrows; b) Types of the interacting edges of the base, which differ for the purines (top) and pyrimidines (bottom). 12 families of possible base pairs are defined by the combination of the orientation of the bases (cis/trans) and the interacting edges (Watson-Crick/Sugar/Hoogsteen/C-H).

Base pairing, primarily canonical, leads to the formation of double-stranded (ds) structures called helices. Helices tend to be relatively short and consist of around 12 paired nucleotides. It is thought that longer consecutively paired regions are too stable and rigid for the majority of RNAs to function properly. Typically, helices alternate with unpaired, single-stranded (ss) regions, which are highly flexible. Together they form secondary structure elements such as stems, loops, bulges and hairpins. (Fig 1.6 a).

The next level - tertiary structure - is a 3D arrangement of the RNA chain. The formation and stabilisation of tertiary structures involve various interactions, including hydrogen bonds, stacking interactions, van der Waals interactions, hydrophobic surface burial, and sometimes the involvement of metal ions. Commonly observed motifs of the tertiary structures in RNA include junctions (Fig 1.6 b) pseudoknots (Fig 1.6 c, d), kink turns, triplexes, and quadruplexes.

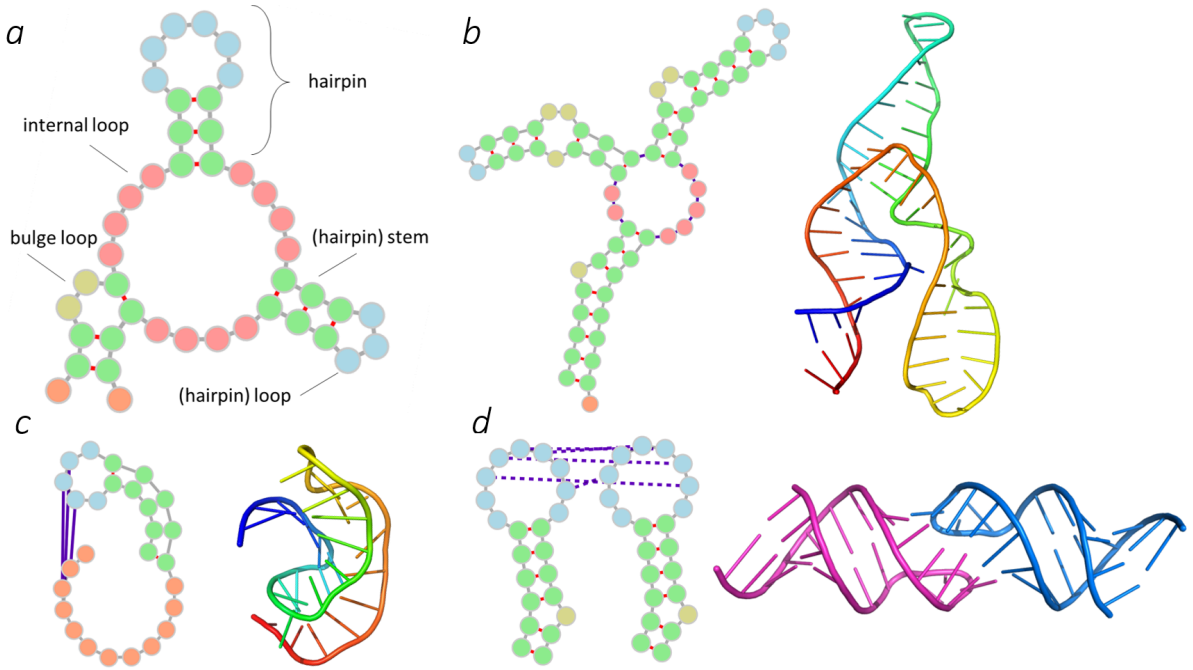


Figure 1.6 - Common elements and motifs of the RNA structure. On the 2D diagrams, phosphodiester bonds are represented as grey lines, base-pairings are represented as red lines, paired nucleotides are represented as green circles and unpaired nucleotides are represented by not-green circles. Finally, bonds that are defined for a particular motif are in purple. a) Basic elements of the secondary structure; b) 2D diagram (left) and corresponding 3D models (right) of the 3-way junction (pdb code 2N3Q); c) Pseudoknot (pdb code 1KPZ); d) Kissing loops (pdb code 2FCY). 2D images have been produced with the help of *forna* [20].

RNA structure, both at the level of base pairing and the overall 3D shape, can be characterised by the principle of isostericity [14,15]. Isostericity refers to the concept that different nucleotide sequences can lead to similar structural and functional properties. This principle enables different RNA sequences to adopt similar structures, allowing RNA molecules with distinct sequences to perform similar biological functions and interact with the same molecular partners, such as proteins. The isostericity of RNA, along with its flexibility, are crucial for the versatility and adaptability of RNA-based interactions in the cells.

1.2.3 The Source of Flexibility

Both protein and RNA molecules in their unbound state (when isolated from the other molecules except the solvent) possess a certain degree of flexibility, although the extent of flexibility varies among different components of the secondary structure. The overall flexibility of the protein is defined by the plasticity of the backbone within two amino acids. This plasticity is limited by two angles between the planes defined by adjacent atoms along the polypeptide backbone, known as the torsional angles. The flexibility of RNA also depends on the plasticity of the backbone, which is noticeably more flexible and defined by 6 dihedral angles describing the rotation around the bonds involving the sugar and phosphate moieties in the RNA backbone (Fig 1.7). RNA's backbone is capable of adapting around 50 distinct conformations [16, 17, 18]. For both molecules, their respective angles are constrained by steric clashes, which prevent atoms from overlapping, and the need to maintain favourable interactions within the molecular structure. For example, the flexibility of the RNA loops in hairpins is often constrained by the interactions of the nucleotides with each other or distal parts of the same RNA [19]. Both molecules can exhibit local flexibility, achieved by the rotations of the side chain units.

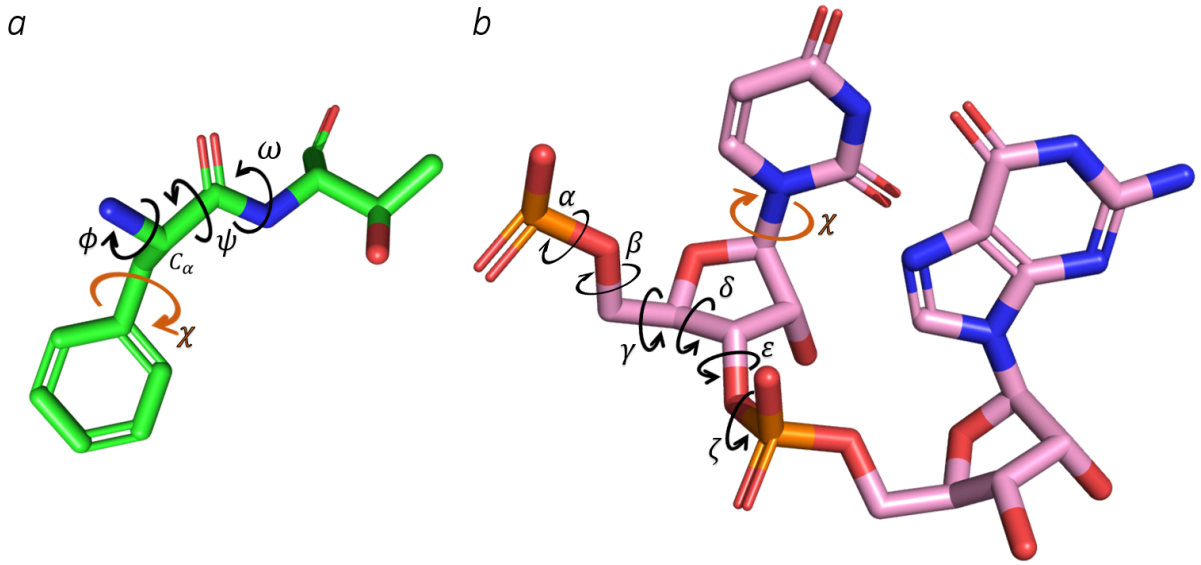


Figure 1.7 - Torsional angles in protein/RNA. a) Torsion angles in the protein backbone (ψ,φ) and of the peptine bond (ω) are shown in black, the angle of the side chain rotation (χ) is shown in orange; Most amino acids adopt *trans* peptide bond, but proline is capable to form both *cis* and *trans* isomers [273]; b) Torsional angles in the RNA backbone (α, β, γ, δ, ε, ζ) are shown in black, the angle of the side chain rotation (χ) shown in orange. Notably, ribose can adopt different conformations (so called sugar puckering) with the most common form called C3′-endo.

Unstructured protein regions possess greater flexibility compared to structured ones (Fig 1.8 a). Protein linkers generally exhibit more flexibility than loops, which, in turn, tend to be more flexible than elements of the secondary structure. For example, the linker between two domains is capable of changing the relative position and/or orientation of the domains, allowing for an adaptation to different binding partners (Fig 1.8 b) [21].

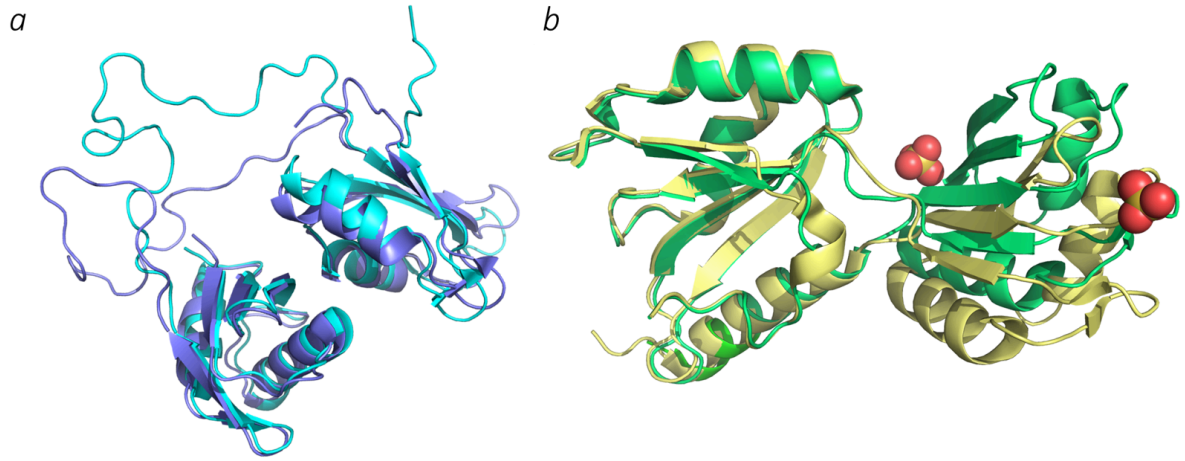


Figure 1.8 - Examples of protein flexibility. a) Two conformations (cyan, blue) of 2YH0 illustrate the high flexibility of the linker; b) Two orientations (yellow, green) of the RRM domain in raver1 (pdbs 3H2U, chain B and 3SMZ) showcasing spatial shift of the domain, located on the right.

Proteins and RNAs can undergo structural transitions between different conformations. Some are known to be stable in a single structure, some can switch between several conformations. For instance, hammerhead ribozymes are known to exist in active, with the binding site exposed and correctly aligned, and relaxed states with several distinct intermediate conformations [22]. Another example is an HIV-1 reverse transcriptase, which is a protein with 3 different structures [23]. Finally, unstructured protein regions can become structured, which is illustrated by SR proteins that are known to switch from a highly disordered state to a partially rigid, arch-like structure, under the influence of phosphorylation [24].

1.3 Macromolecular Interactions

Molecular interactions involve direct physical contact between molecules, ranging from weak and transient interactions to strong and very stable ones. These interactions impact the structure and the functions of the molecules involved, and certain interactions, described below, lead to the formation of macromolecular complexes. A macromolecular complex refers to the precisely arranged assembly of multiple macromolecules into a functional unit, capable of performing complex tasks. Up to 10% of the entire proteome may be bound to RNA [25]. Protein-RNA complexes are one of the fairly common macromolecular complexes, which are central to a plethora of vital processes (several examples are given in §1.3.4.2). Understanding the mechanisms behind such interactions provides insights into the mechanisms underlying these processes and their disruption, as well as allowing for rational design - the creation of the molecules with the desired functionality, tuned via their structural characteristics.

1.3.1 Protein-RNA Interaction Types

Proteins can interact with RNAs through the backbone of any amino acid and the side chain of most amino acids. The following noncovalent interactions are known to occur:

- **Aromatic interactions or pi-interactions.** Aromatic interactions involve the nitrogenous base of the nucleotide. The aromatic ring of the base contains pi electron clouds, which can interact with other pi systems or electron acceptors. Typically, interacting partners are located

within 2.7-4.3Å. The interaction is relatively strong with the contribution of approximately 2-6 kcal/mol per interaction with multiple interactions that can be present in one complex. These interactions are known to substantially contribute to the complex stability, with some examples where they are crucial to binding function [26 and references within]. Purines are considered to be better stacking partners compared to pyrimidines [27], but practically stacking of all 4 bases was observed with approximately similar frequency [28]. Several subtypes of aromatic interactions are recognised:

- *Pi-pi stacking interactions* (Fig 1.9 a) occur between the RNA bases and the amino acids containing aromatic rings, namely tryptophane, histidine, tyrosine and phenylalanine. Typically, aromatic rings are laying parallel, on top of each other, hence the term 'stacking'. Although the perfect geometry is not mandatory as interactions with both rings being angled towards each other, or even located perpendicular ('edge-to-face' or 'T-stack') have been observed. In parallel orientations, the pi-electrons from one ring interact with the pi-electrons of the other ring through attractive forces, such as van der Waals interactions. In the perpendicular orientation, the pi-electrons of one ring interact with protons of the other ring. Interestingly, 3 rings can be involved in stacking interaction. This happens when one ring is being "sandwiched" between another two. This interaction may involve two nucleotides with one amino acid and vice versa [29];
- *Pi-cation interactions* (Fig 1.9 b) occur between the RNA bases and the guanidinium group of arginine amino acid due to the attraction of the positively charged group toward the pi-electron cloud of the aromatic ring. All possible orientations between the ring and guanidinium group could be found. Studies hint at the preference of arginine to bind U, A, and C bases over G. Pi-cation interactions have been observed between RNA bases and lysine or histidine. Notably, pi-cation interactions are observed more often in protein-DNA complexes [30, 31];
- *Other pi-interactions* include contact of the aromatic ring with the amino group of the glutamine and asparagine, and interaction with the peptide bond, but they are observed less commonly. Finally, proline can form CH/π interaction with the aromatic ring, due to interaction between the pi aromatic, the polarised C-H bonds and the hydrophobic effect, yet CH/π interaction is rarely mentioned in the literature [32].

- **Hydrophobic interactions**. Hydrophobic interactions involve hydrophobic and sometimes nonpolar amino acids and RNA bases, all of which are nonpolar. This type of interaction does not involve direct chemical interactions between hydrophobic molecules, but rather their clustering in an attempt to minimise the surface, exposure to the water molecules and minimise disturbance of the latter. Hydroponic interactions can occur at 3.8-5.0Å, and contribute around 1-2 kcal/mol. In some cases, these interactions account for half of the binding within the protein-RNA complex, and they are highly important for its stabilisation [33].

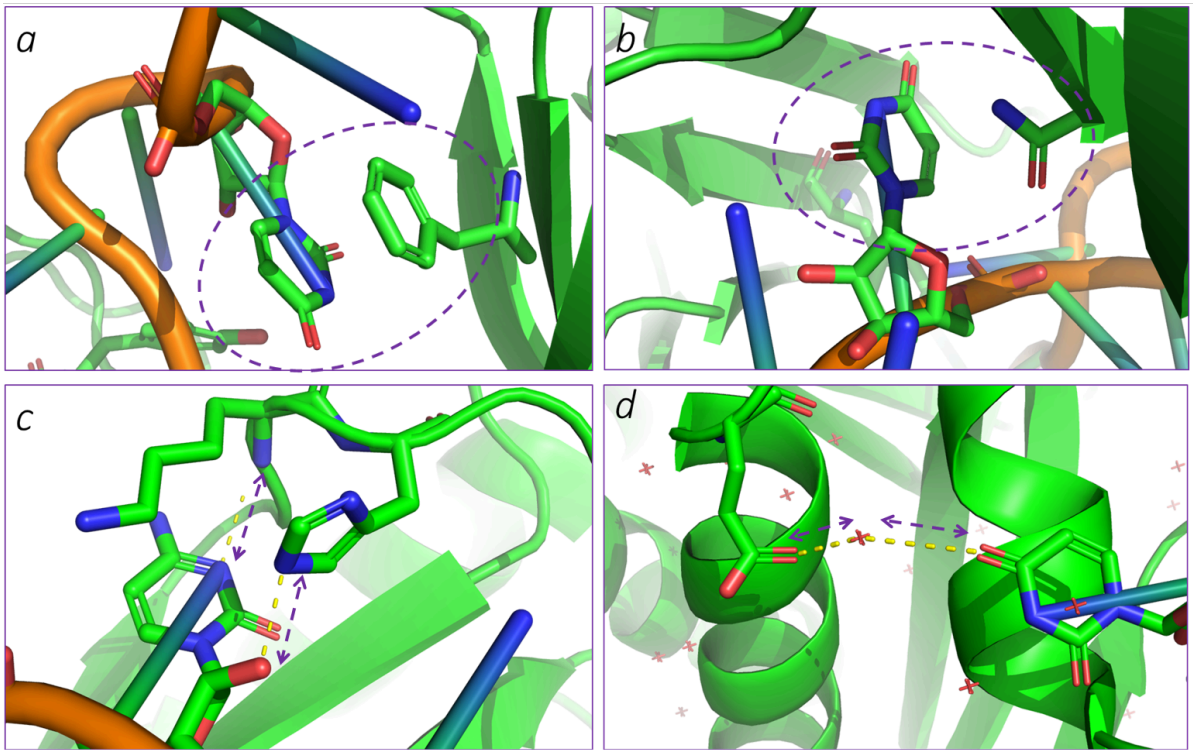


Figure 1.9 - Protein-RNA interaction examples, highlighted by purple dashed circles/lines. Protein and RNA are shown in the cartoon representation, and their interacting residues are shown in a stick representation. a) Pi-pi stacking between Phe_256 and U_5 (pdb code 1B7F); b) Pi-cation stacking between Tyr_214 and G_4 (pdb code 1B7F); c) Hydrogen bonds between His_411 and C_540; and Lys_410 and C_540 (pdb code 2ADC); d) Hydrogen bond bridged by the water molecule, shown as a red cross, between Glu_78 & U_213 (pdb code 2ARN).

- **Electrostatic or Coulombic interactions**. Electrostatic interactions involve the attraction between differently charged molecules, and the repulsion between the similarly charged ones. Nucleic acids are negatively charged, and the vast majority of the interactions that lead to protein-RNA binding are of an electrostatic nature. Several subtypes of electrostatic interactions are recognised:
    - *Ionic interactions* or *ionic bonds* occur between positively charged protein side chains and negatively charged phosphate groups of RNA. Ionic interactions are very strong as they involve fully charged molecules. They may occur at a wide range of distances, and close proximity of molecules does not necessarily lead to stronger binding and vice versa. Interactions with metal ions, such as magnesium or zinc ions, often aid the folding and stabilisation of RNA [34] or/and protein [35], as well as their complex [36];
    - *Hydrogen bonds* (*conventional)* (Fig 1.9 c) occur between two electronegative atoms, that share a proton, most commonly between nitrogen and oxygen via hydrogen that is covalently bound to one of these atoms (X-H…Y (X = O or N; Y = O or N). The hydrogen bond is formed through electrostatic attraction between the partially positive hydrogen atom and the partially negative atom with the lone pair of electrons. This bond can be bridged by the external water molecule (Fig 1.9 d). Hydrogen bonds are often presented as a stand-alone type of interaction, as they are stable interactions in protein-RNA bonding. All four bases and phosphodiester bonds can form hydrogen bonds with both the side chain and the backbone of the protein,

but interactions involving the RNA side chain are the most common. Typically, hydrogen bonds occur at 2.4-3.0Å and contribute 0.5-4.5 kcal/mol [26] Similarly to stacking interactions, the geometry of the interaction matters as research revealed that the precise energy contribution depends on the exact relative orientation of the interactors;

- *C-H…O hydrogen bonds* occur between carbon and oxygen via hydrogen that is covalently bound to a carbon. The importance of this interaction type was uncovered for nearly half a century [37]. The strength of the unconventional hydrogen bond depends on the acidity of the hydrogen and is at its strongest when the CH group is adjacent to N;
- *Van der Waals (VdW) interactions* can occur between any two or more molecules and are dependent on slight fluctuations of the electron densities. They consist of dipole-dipole forces, which occur between polar molecules, and (London) dispersion forces. VdW are weaker, approximately 0.5-1 kcal/mol, and occur at a distance exceeding 3Å. In protein-RNA complexes, VdW interactions are abundant, surpassing other types of interactions in overall contribution despite the weakness of a single VdW interaction [26].

1.3.2 Conformational Changes Induced by the Binding

Various rearrangements in RNA and/or protein structures may occur, before or during the binding, changing the structures from an unbound state to a bound one (Fig 1.10). These rearrangements are classified as *conformational selection* or *induced fit* respectively.

- Conformational selection, also known as conformational capture or tertiary structure capture, refers to the ability of the proteins and/or RNA, which have multiple unbound conformations, to select the most suitable one for binding. Incorrect conformations are not recognised by the potential binding partner. An example can be seen in MS2-RNA recognition, where RNA hairpin structure is required for the proper binding [38].
- Induced fit refers to the ability of the molecules, individually or together, to undergo conformational changes upon binding to ensure the complementarity of their shapes, which promotes stronger and more stable binding. Induced fit can involve both local changes such as backbone shifts or base-flipping, and more substantial changes, e.g. change in domain orientation, both of which can be observed in the ribosomes [39]. In protein-RNA binding, unstructured loops often adopt a structure upon binding [ref green review and references within]. An example of mutually induced fit can be seen in tRNA-MiaA recognition, where the protein forms a deep rift, which accommodates and partially unfolds the RNA anticodon loop [40].

These two mechanisms are interconnected: when the conformations of two molecules are sufficiently close to being complementary, they start to bind and one or both molecules undergo conformational changes to reinforce newly established interactions and bond, which in turn leads to a more optimal fit and stabilises the complex.

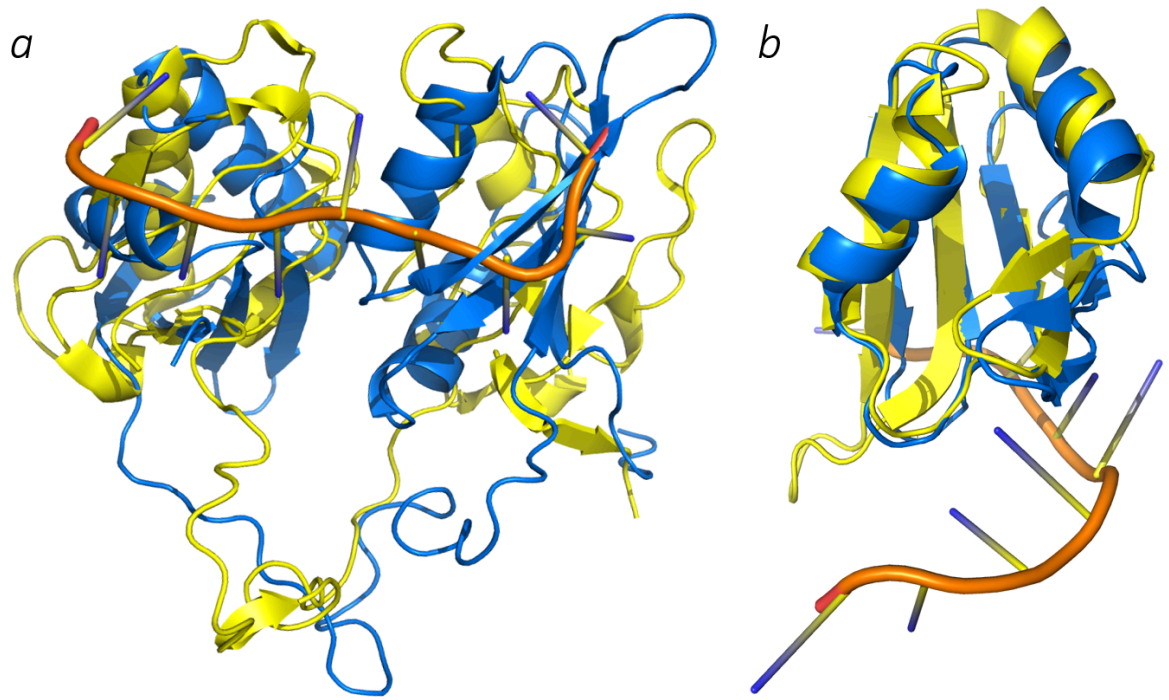


Figure 1.10 - Differences between the unbound (blue) and bound (yellow) conformations of the a) Human U2AF65 tandem RRM 1 and RRM2 (pdb codes 2YH0 and 2YH1 respectively); b) Human TDP-43, RRM1 (pdb codes 4SMZ and 4IUF respectively).

1.3.3 Binding Energy

The strength of molecular binding is often quantified by the binding free energy (ΔG), which represents the change in energy upon complex formation compared to the unbound states of the molecules involved. This binding free energy is influenced by several factors, including intermolecular forces, conformational dynamics, and solvent rearrangement [41]. For a complex to form, its energy must be lower than the sum of energies of the solvent-separated molecules. Stable conformations of a complex correspond to energy minima, which can be either global or local, although individual bound structures may not always correspond to the absolute energy minimum [42].

1.3.4 RNA-Binding Proteins

Not all proteins have the ability to interact with and bind RNA molecules. Proteins that possess this capability are referred to as RNA-binding proteins (RBPs) and often contain specific RNA-binding domains (RBDs) or motifs. RBDs exhibit unique structural features that facilitate binding with RNA molecules. While individual domains typically interact with RNA molecules with relatively low affinity, as only a few amino acids are directly involved in the interaction, often single RBP contains multiple copies of the same RBD, which enhances the strength of the binding and stability of the RNA-protein complex [26].

1.3.4.1 RNA-Binding Domains

Some commonly observed RBDs include RNA recognition motif (RRM), zinc-binding domain (ZBD), Pumilio homology domain (PUF), and K-homology domain (KH).

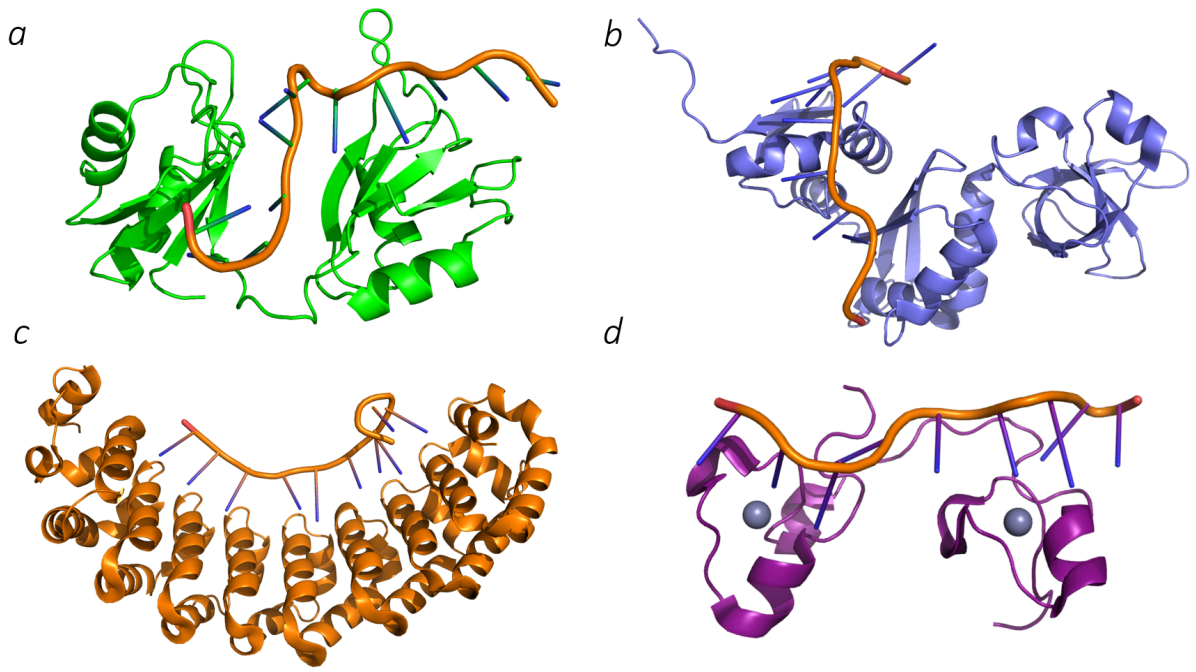


Figure 1.11 - 3D models of protein-RNA complexes in the cartoon representation. a) Sex-lethal protein demonstrates RRM tandem (pdb code 1B7F); b) NusA protein demonstrates 3 copies of KH domains, type II (pdb code 2ASB); c) Human Pumilio1 protein demonstrates 9 copies of PUF domain (pdb code 2YJY, chain A); d) Protein Tis11d demonstrates tandem zinc fingers (zinc ions are shown in grey).

- **RRM** (Fig 1.11, a) are the most abundant domain out of all known, occurring in 1-2% of human proteins and known to bind 2-8 nucleotide-long ssRNAs. This domain consists of 90-100 amino acids, which form 4 antiparallel beta sheets packed against 2 alpha helices. Up to 6 copies of the domain can be found in one protein. The binding mode with ssRNA is diverse: the primary binding interface is located on the beta sheets 1 and 3, where 3 conserved aromatics are stacked with nucleotides; however, some RRMs lack these conserved aromatics (sometimes they are called quasi-RRMs). Loops 1, 3 and 5 are known to be highly important for binding. Linkers, if present, contribute to the binding and stability as well;
- **KH** domain (Fig 1.11, b) consists of ~70 amino acids with either beta-alpha-alpha-beta-beta-alpha (type I) or alpha-beta-beta-alpha-alpha-beta (type II) and conserved GXXG loop between alpha helices. This domain forms a hydrophobic binding pocket for both ssRNA and ssDNA and typically binds 4 nucleotide-long sequences. Notably, stacking interactions are rare within this domain [43];
- **PUF** domain (Fig 1.11, c) is 3 parallel alpha helices containing ~35 highly conserved amino acids. PUF domains, typically around 8, are packed in half-doughnut-shaped proteins which can bind 8-11 nucleotide-long ssRNA. RNA bases bind along proteins' inner surface via stacking and hydrogen bonds, and phosphates are solvent-exposed [44]. Wild-type domains do not recognize C, but due to the comprehensive understanding of the PUF-RNA binding, the artificially engineered domains can bind specific sequences of RNA, including C [45];
- **ZBD** (Fig 1.11, d) is a type of domain that binds zinc ions via cysteine and histidine residues. Different subtypes of ZBD are known, most of which bind DNA, but some can bind RNA.

Several subtypes of Zinc Fingers can bind the backbone of dsRNA or bulging nucleotides [46], while the other form bonds with G in ssRNAs [44];

- **Other domains**. It's important to note that there are additional RBDs beyond the described domains, including ordered domains like double-stranded RNA binding domain and helicase [26], as well as disordered domains like RG[G] repeats and low-complexity sequences [47].

1.3.4.2 RBPs' Functionality

RBPs are involved in a myriad of vital cellular processes and their interactions with RNAs influence RNA processing, such as splicing, alternative splicing, editing, decay and more [48, 49]. Disruption of RBP-RNA binding or aberrant protein-RNA interactions can lead to genetic diseases [48], including cancer [50, 51, 52], neurodegenerative diseases [53] and other disorders [54, 55]. For instance, mutations in the proline residue at position 95 of the RRM-containing protein SRSF2 resulting in an altered mRNA binding site, have been observed in cancer patients [56]. Another example is the TDP-43 protein, which contains 2 RRM domains and is involved in splicing [57]. Mutations in the alanine residue at position 315 of TDP-43 are associated with neurodegenerative diseases [58]. Moreover, viruses rely on cellular RBPs to replicate and spread [59].

When a certain protein is known to play a significant role in the development, progression or treatment of a disease, such protein is referred to as a therapeutic target. Targeting such protein and changing its behaviour or expression can be a promising strategy in the treatment or even prevention of disease. The specificities of RBPs often make them the key therapeutic targets in many diseases [60]. For instance, an elevated expression of the Musashi proteins has been observed in cancer patients [61]: overexpression of Musashi-2 potentially causes an aggressive form of leukaemia [62]. This is why the introduction of the ligand, such that binds exclusively to Musashi, blocks its binding site thus inhibiting its functions, could be a potential therapeutic strategy to treat cancer patients with elevated Musashi expression. This is a highly challenging task, however, there are known cases of the successful development of such a therapeutic. For example, Rbox proteins have been successfully engineered to bind specific RNAs, leading to increased expression of tumour suppressors [63].

To sum up, a comprehensive understanding of the binding mechanisms between proteins and RNAs is highly important. This understanding not only contributes to unravelling the functionality of protein-RNA interactions but also enables the design of artificial protein-RNA complexes with desired functionality.

1.3.5 Specificity of ssRNA Binding

The focal point of this work, single-stranded RNA binding, is characterised by the inherent high flexibility of the single-stranded chain. Unbound ssRNAs are commonly considered to be unstructured, as the information on their free form is scarce. Another possibility is that unbound RNA adopts a secondary/tertiary structure, and becomes single-stranded as a result of the binding (e.g. through interaction with helicase [64, 65]). The significant aspect of ssRNA binding is the accessibility of unpaired bases, which often leads to the formation of stacking interactions. Unlike double-stranded nucleic acids, ssRNA exhibits greater conformational flexibility and possesses different binding modes [66]. Unlike small ligands, ssRNA chains are typically longer. Additionally, research suggests that RNA is initially attracted to the protein via electrostatics, and then locked in a bound form via stacking and hydrogen bonds [67].

Proteins can bind RNAs both in specific and nonspecific ways. Nonspecific recognition involves more general interaction between the RNA and the protein, when a protein can bind multiple RNAs, for example in the RNA degradation process [68]. On the other hand, specific recognition implies that

a certain RNA sequence (or structure in the case of structured RNA) is required for the binding. For instance, zinc finger CCCH binds sequence-specifically to the 5'-UAUU sequence [68] and the RbFox binds RNAs with a GCAUG motif [70].

1.3.5.1 RRM-ssRNA Binding

Since RRMs account for nearly half of all RBPs [71], it is important to highlight the aspects of RRM-ssRNA interactions in detail. RRMs have been extensively studied and are remarkably versatile in their ability to recognize RNA. RRMs bind ssRNA with variable sequence specificity - from preference towards sequences rich in particular bases to fully specific sequences, and different affinity - from milli- to micromolar. Their recognition capability can be modulated allosterically [72].
Several RRM-ssRNA binding modes are known, including 1) canonical binding to the β-sheet surface, 2) canonical binding with involvement of N- and C-termini; 3) binding to conserved loops, and 4) binding to an α-helix. All 4 modes exhibit significant differences in the thermodynamics of the binding process, meaning that there are differences in the dynamics upon binding [73].

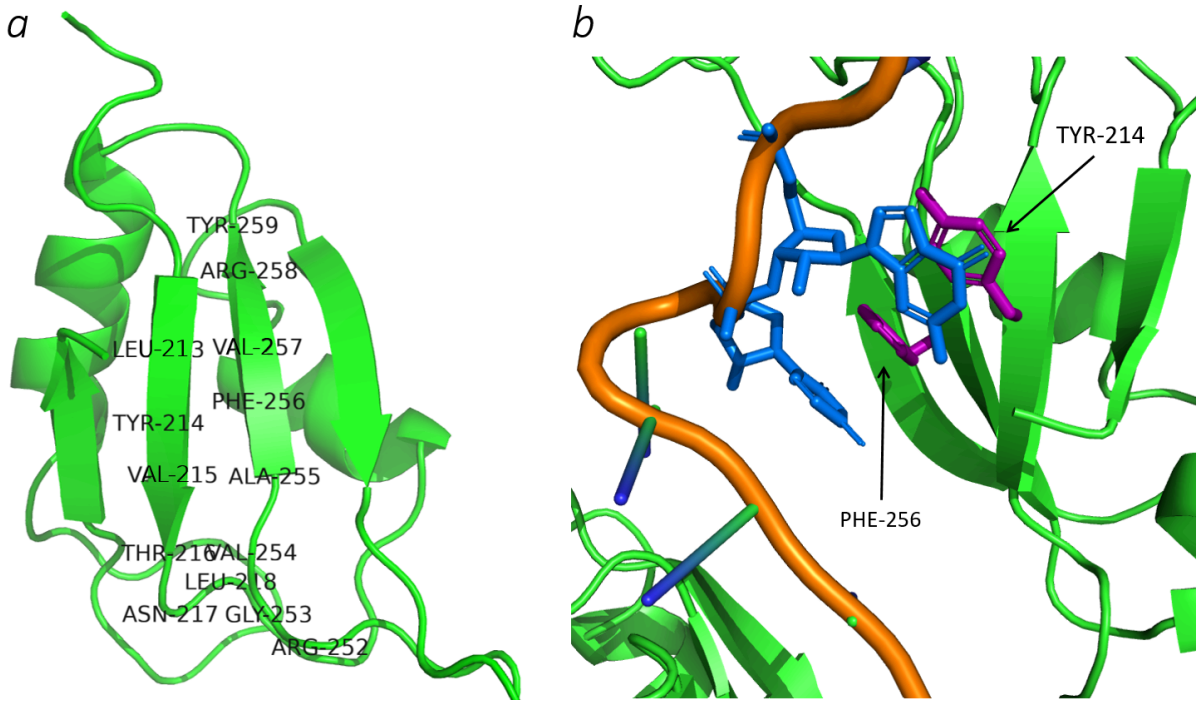


Figure 1.12 - Details of the RRM domain and its binding with RNA on the example of Sxl (pdb code 1B7F, RRM2). a) A typical structure of the RRM domain with conserved residues in RNP1 (right) and RNP2 (left); b) An example of binding between RRM and ssRNA showcasing stacking interactions between conserved amino acids in RNPs (purple) and nucleotide bases (blue), Typ_214 with G_4 and Phe_256 with U_5.

As mentioned before, the most characteristic feature of RRM-ssRNA binding is stacking interactions occurring with conserved aromatics in the beta sheets. These aromatics are located on conserved RNP1 and RNP2, with the sequences (R/K)-G-(F/Y)-(G/A)-(F/Y)-V-X-(F/Y) and (L/I)-(F/Y)-(V/I)-X-(N/G)-L, respectively [74], where amino acids in bold are responsible for the stacking interactions (Fig 1.12 a, b). The amino acid in the 2nd position in RNP2 is typically stacking with nucleotides around 5' (exact positions of the nucleotides can vary), and one in the 5th position,

RNP1, is stacking with the nucleotide around 3'. The amino acid in the 3rd position, RNP1, is inserted between two sugar rings of the adjacent nucleotides.

1.4 Experimental Structural Biology

The study of the protein-RNA complexes and their constituents begins with *in vitro* data (e.g. binding preferences, kinetic constants, etc.) followed by insights into their structural and functional features. Structural information can be obtained through complex experimental methods and techniques, each offering unique data with certain specificities and limitations. The determination of the detailed 3D structure of a complex, including precise information about the spatial arrangement and molecular interactions, is referred to as 'solving' the structure of a complex. This section presents a concise overview of the classical techniques associated with molecular complex solving.

1.4.1 X-Ray Crystallography

This technique produces a crystal structure of an object by analysing the series of diffraction patterns of the X-ray beam directed onto the crystalised object. The diffraction pattern consists of a series of spots, which correspond to the interference of X-rays scattered by the crystal. The intensity and position of these spots provide information about the arrangement of atoms within the crystal. Computational analysis of the diffraction patterns results in the creation of a 3D electron density map of the crystal, which represents the distribution of electrons within the crystal, and can be further interpreted to determine the positions of the atoms in the object. Some parts of the object before crystallisation may exhibit flexibility and thus have different positions in different crystals. To quantify the uncertainty of the atoms' positions, the B-factor for each atom is calculated. Higher values of the B-factor are indicative of the vibration of the atoms around their positions in the final 3D structure.

X-ray crystallography is capable of producing high-resolution images (<3Å, often ~1Å) of proteins and complexes with atomic-level detalization. It is suitable for both large and small objects. The scourge of this technique is the requirement for crystallisation. This is a tedious and complex process, which requires the optimisation of multiple parameters [75]. Prior to crystallisation, molecules have to be expressed and purified, which may be challenging on its own. Moreover, flexible molecules are not suitable for crystallisation. While it is possible to obtain their crystals, they do not provide an accurate representation of such molecules, for instance, highly flexible regions could be not visible on the electron density map if their positions in crystals are highly diverse. Thus, X-ray crystallography is not a suitable technique to study the dynamics of the molecules/complexes. It is especially difficult to apply X-ray crystallography to RNA molecules due to such issues as misfolding, rapid decay of the crystals, etc [271].

Crystallisation may lead to false positives - crystal structures that are non-natives, yet have been captured via X-ray crystallography. Such structures are known as crystal artifacts and typically occur due to the distortions of the native conformation caused by crystal packing or twinning, the presence of unknown lignans, radiation damage to the crystals, and human errors such as mislabeling of samples. Many cases of crystal artifacts are known, for example, a case of purification and crystallisation of YodA instead of GNAT [76]. Another issue is the possibility of obtaining a conformational snapshot of a complex when the most stable conformation out of several possible ones is obtained. In other words, the complex may have multiple stable conformations, but due to experimental conditions or other factors, only one of these conformations is captured in the crystal.

Crystal contacts often can present as an artefact as well - when 2 units of the crystal are in contact in the crystal packing, but such interface does not occur in the cell.

1.4.2 Nuclear Magnetic Resonance Spectroscopy

This technique, NMR, produces multiple types of data for a labelled object in the solution. Analysis of these data provides information on the structure, dynamics, and interactions of objects in solution, capturing their close-to-natural behaviour. Individual atoms, molecules, or whole complexes are labelled and placed in a strong magnetic field, then probed with radio waves. Under these conditions, certain atomic nuclei, such as hydrogen and atoms used for labelling (isotopes), absorb and emit energy at specific frequencies, which are captured during the experiment. These signals are changing depending on the chemical environment, which allows tracking of the information about the 3D structure of the molecules. Recorded signals can be transformed into distance maps, which, in turn, can be utilised to determine the spatial structures of the molecules using a combination of distance geometry approach and molecular dynamics simulations [77]. The resolution of the 3D structure is directly proportional to the duration of the captured signal. NMR is capable of capturing information about large-scale conformation dynamics.

However, the complexity of the interpretation limits the utilisation of NMR to relatively small molecules/complexes, with a maximum mass of ~100 kDa [78]. It is often challenging to determine the 3D structure due to the overlapping of the signals, or their insufficient intensity. This method requires a high concentration of the objects in the solution, which can be challenging for complexes with low solubility or low stability.

1.4.3 Cryo-Electron Microscopy

This technique, Cryo-EM, produces multiple 2D projections of a frozen object, which are then processed computationally into a 3D density map of the object. An object in the solution is placed on a supportive grid and rapidly plunged into cryogen (e.g. liquid nitrogen), which results in the rapid freezing of the object in its native state. The grid is then transferred to the electron microscope, where 2D images of the object are obtained at different orientations and angles. Processing of 2D images is the most challenging part of the process, which requires selections of the individual images, their alignment, 3D reconstruction, refinement, and final validation and interpretation.

Nowadays cryo-EM is suitable for large complexes (~200kDa) and is capable of capturing their structures with exceptionally high resolution, up to 1.22Å [79] (some years ago only near-atomic resolution, ~3Å, was achievable). It is possible to obtain multiple conformations for a single molecule/complex. This technique is currently in the state of active development - new image-processing algorithms and other innovations push the boundaries of the determination of high-resolution molecular structures [80].

The main disadvantages of cryo-EM include a limited resolution for small complexes (<100 kDa), as they have fewer structural features and may provide a low signal-to-noise ratio in the 2D images. The heterogeneity and flexibility of the complexes may lead to the blurring of the electron density.

1.4.4. Low-Resolution Techniques

There are countless other experimental techniques capable of providing insights into the shape, dynamics and behaviour (e.g. assembly and disassembly, ligand recognition, etc.) of the

molecular complexes and information on specific contacts, but typically are insufficient for solving a complex. A few examples are given below.

In small-angle X-ray scattering (SAXS) and small-angle neutron scattering (SANS), a beam of X-ray/neutrons is directed at a complex in the solution and the intensity and distribution of the scattered particles at different angles are recorded to derive information like compactness or the overall size of the complex (radius of gyration).

In Fluorescence resonance energy transfer (FRET), two molecules are labelled with fluorophores donor and acceptor. The energy transfer between fluorophores is observed when they are located in close proximity, and the change in this energy can be converted into the distance. This provides information about the binding kinetics, conformational changes, and dynamics of the complex.

In cross-linking, molecules are artificially linked via covalent bonds, which immobilises the complex in its spatial arrangement at the time of cross-linking, preserving the proximity and interaction information between the fragments of molecules that are in the proximity of the linking agents. Upon analysis, the binding site can be found along with approximate spatial constraints.

In order to apply these techniques effectively, it is important to have prior information about the molecules involved in the protein-RNA complex. The information about the binding partners can be obtained via such techniques as RNA electrophoretic mobility shift assay (EMSA) or cross-linking immunoprecipitation (CLIP). Finally, the measuring of the binding kinetics (e.g. with Surface Plasmon Resonance (SPR), Isothermal Titration Calorimetry (ITC) [81]) in combination with mutagenesis approaches helps to determine key residues involved in the binding.

The main advantages of low-resolution techniques over high-resolution ones are relative simplicity, a shorter length of the whole process, and a lower cost of a single experiment.

1.5 Conclusion

In this chapter, we have presented the biological foundation underlying our research, namely protein-RNA complexes and their constituents. We have illustrated the complexity of individual molecules, their 3D structures, and the diversity of intermolecular interactions within protein-RNA complexes. We have highlighted the origins and significance of molecular flexibility, which plays a crucial role in the binding by allowing a single molecule to adopt multiple conformations, and by allowing the existence of such phenomena as selection and induced fit. We have delved into a diverse universe of RNA-binding proteins, providing examples of their functions, emphasising the detrimental consequences of disruptions in their activity, and the ways to address these disruptions. An overview of common RNA-binding domains and their general binding preferences has been presented. Special attention has been given to RRM, the most common RNA-binding domain, along with the specificities associated with ssRNA binding. Finally, we have reviewed classical experimental techniques suitable for acquiring high-resolution 3D structures of protein-RNA complexes. In the next chapter, we will delve into the computational aspect of the research surrounding the structure of protein-RNA complexes. We will review computational methods employed for structure prediction and discuss associated challenges, including the simulation of molecular flexibility and conformational changes occurring upon binding in protein-RNA complexes.

Chapter 2: Structural Bioinformatics of the Protein-RNA Complexes

2.1 Aims

In this chapter, we provide a brief introduction to the field of bioinformatics, particularly its relevance to the 3D modelling of protein-RNA complexes. We offer a short description of the approaches developed for modelling single chains, focusing on the similarities and differences inherent in modelling protein chains compared to RNA chains. Following this, a concise overview of the current state of the expansive field of molecular complex modelling is provided, delving deeper into the domain of molecular docking. In this context, we focus on the aspects most suitable for protein-ssRNA docking. Finally, we discuss ssRNA'TTRACT, a state-of-the-art method in fragment-based protein-ssRNA docking, which serves as a cornerstone of the research conducted in the frame of this PhD project. This sets the stage for the contributions of this PhD project.

2.2 Structural Bioinformatics

Computer-aided computations have long been an integral and pivotal part of biological research, revolutionising the pace of scientific progress. This transformation became possible through the cumulative efforts of generations of researchers who designed and developed mathematical models, suitable for the investigation of various biological processes. Structural bioinformatics is a specialised field within the broader discipline of bioinformatics that focuses on the 3D structures of biological macromolecules, primarily proteins and nucleic acids. This field combines principles of structural biology, computer science and data analysis to decipher, model and analyse the structures of biomolecules and their complexes.

One of the main branches of structural bioinformatics is molecular modelling [84]. It involves the use of computational techniques to simulate and predict or refine the 3D shapes of molecules and their complexes. The focus of this project is on the modelling of 3D structures. Such modelling is important, because it helps to bridge the gap between the known sequences of molecules and their unknown structures. For example, a minuscule fraction, less than 0.03% of all the known protein sequences have experimentally resolved high-resolution structures [85]. Computational 3D models for such proteins can provide critical insights into their spatial arrangements, which in turn is often imperative for understanding functions and interactions of these molecules. Subsequently, understanding of molecular functions and interactions allows for understanding of the disease mechanisms, followed by the drug design, as discussed in §1.3.4.2.

Structural bioinformatics offers distinct advantages over experimental structural biology, including speed, efficiency, theoretical justifiability and the ability to investigate multifaceted biological interactions that may be highly challenging, time-consuming and expensive to study experimentally. Still, structural bioinformatics integrates with experimental biology, using experimentally obtained data to guide the modeling and create predictive models, which in turn require experimental validation. However, the effectiveness and robustness of the predictions and models heavily rely on data quality and availability, algorithm sophistication, and current biological knowledge. Additionally, research in bioinformatic often requires large amounts of computational resources (GPU or CPU, RAM, storage), which may be unavailable within academia. The synergy between structural bioinformatics and experimental biology enhances our understanding of biological systems and informs decision-making in various fields like biotechnology and (personalised) medicine.

2.2.1 Databases

Computational structural studies heavily rely on access to experimental structural data of biological molecules. The worldwide repository for such data is the Protein Data Bank (PDB) [86, 87], which holds an extensive collection of protein, DNA and RNA structures (Fig. 2.1). This repository includes various states of some molecules, such as their unbound forms and complexes with different ligands. As of September 3, 2023, the PDB contains over 1 terabyte of structural data, among which 12,076 protein-nucleic acid complexes. Ribosomes are known to account for approximately 20% of all these assemblies, leaving around 9,660 non-ribosomal protein-nucleic acid complexes in contact with a protein. However, this number is significantly lower than the number of protein-only structures, which is equal to 181,324 entries.

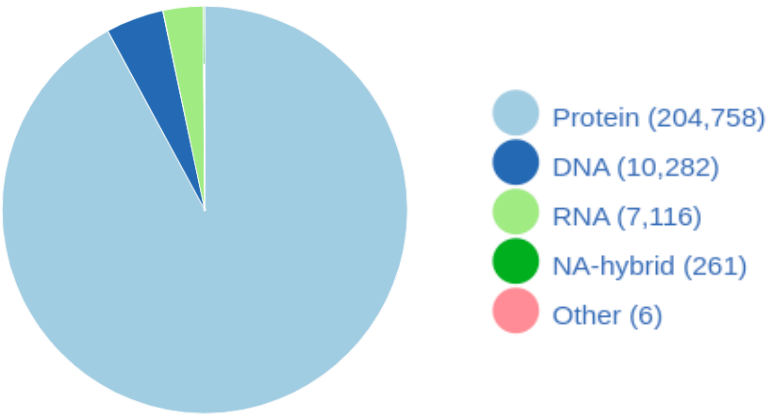


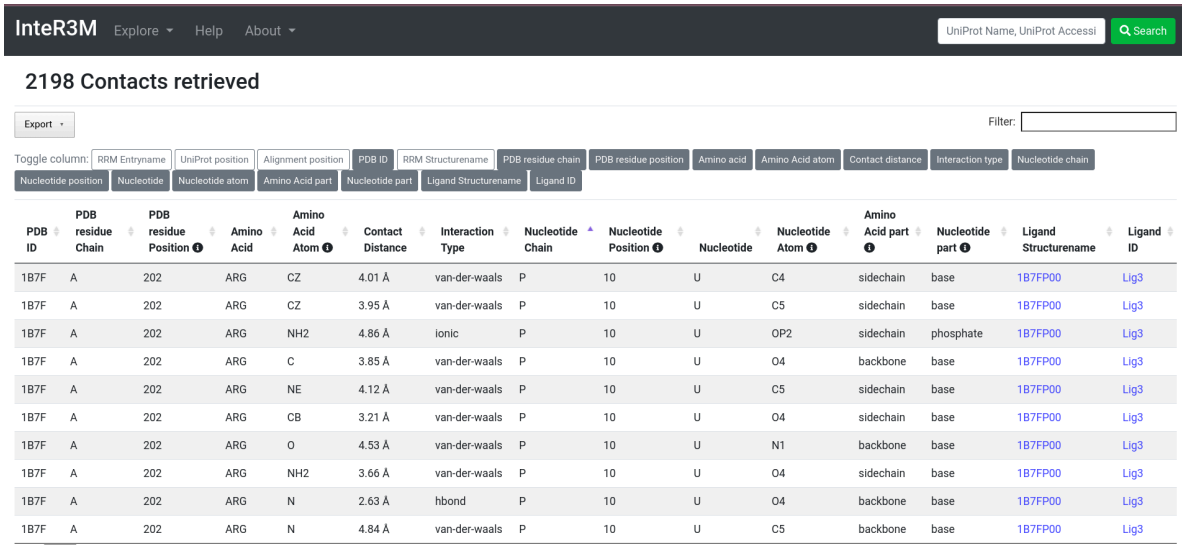
Figure 2.1 - Distribution of the data by polymer entity type in PDB as of September 2023. A single PDB structure may contain multiple entities of different macromolecular types [88].

Databases containing interface annotations of the molecular complexes play a pivotal role is the research. These databases hold information on individual contacts between protein and RNA residues that are essential for the binding. Such data is often required for statistical analyses and machine learning (ML) or deep learning (DL) applications.

Several of such databases are available online, but not all of them have been updated recently. To name a few, the 'Protein–RNA Interface Database' (PRIDB [91]), last updated in 2011, contains protein–RNA interfaces extracted from complexes in the PDB. The 'RNA-Binding Protein DataBase' (RBPDB [92]), last updated in 2012, contains experimental observations of RNA-binding sites. Other databases have also been developed [93].

Recently the Protein Data Bank in Europe (PDBe) announced a release of a new version of the 'Proteins, Interfaces, Structures and Assemblies' (PISA), a tool for analysis of intermolecular interactions within existing assemblies [94, 95]. This tool provides, among others, the interface details for a given assembly in the PDB, including the atom-atom interaction details for each contact.

A specialised database titled 'Interactions of RNA and RNA Recognition Motif Database' (**InteR3Mdb**) was created in the frame of the RNAct project [96]. As the name suggests, it contains data specifically on the RRM-RNA interactions, including the atom-atom distance, type of the interaction, etc (Fig. 2.2). Along with the database, a tool 'RRMScorer' for the prediction of RRM-ssRNA binding based on sequence [97]. Furthermore, relevant information may be published in scientific articles detailing experimental studies (mutagenesis, cross-linking, etc.), on specific complexes. Often, such information requires manual extraction and processing before it can be added to the database. This process complicates the maintenance of up-to-date databases. While some computational tools are available [89], they appear to not have gained widespread adoption. There is hope that recent advancements in Large Language Models (LLMs), applicable to computational biology [90], will be integrated into data management workflows in the near future to facilitate the collection and management of experimental data from scientific articles.



PDB ID	PDB residue Chain	PDB residue Position	Amino Acid	Amino Acid Atom	Contact Distance	Interaction Type	Nucleotide Chain	Nucleotide Position	Nucleotide	Nucleotide Atom	Amino Acid part	Nucleotide part	Ligand Structurename	Ligand ID
1B7F	A	202	ARG	CZ	4.01 Å	van-der-waals	P	10	U	C4	sidechain	base	1B7FP00	Lig3
1B7F	A	202	ARG	CZ	3.95 Å	van-der-waals	P	10	U	C5	sidechain	base	1B7FP00	Lig3
1B7F	A	202	ARG	NH2	4.86 Å	ionic	P	10	U	OP2	sidechain	phosphate	1B7FP00	Lig3
1B7F	A	202	ARG	C	3.85 Å	van-der-waals	P	10	U	O4	backbone	base	1B7FP00	Lig3
1B7F	A	202	ARG	NE	4.12 Å	van-der-waals	P	10	U	C5	sidechain	base	1B7FP00	Lig3
1B7F	A	202	ARG	CB	3.21 Å	van-der-waals	P	10	U	O4	sidechain	base	1B7FP00	Lig3
1B7F	A	202	ARG	O	4.53 Å	van-der-waals	P	10	U	N1	backbone	base	1B7FP00	Lig3
1B7F	A	202	ARG	NH2	3.66 Å	van-der-waals	P	10	U	O4	sidechain	base	1B7FP00	Lig3
1B7F	A	202	ARG	N	2.63 Å	hbond	P	10	U	O4	backbone	base	1B7FP00	Lig3
1B7F	A	202	ARG	N	4.84 Å	van-der-waals	P	10	U	C5	backbone	base	1B7FP00	Lig3

Figure 2.2 - An example of contact database InteR3M showing interface annotation between protein and ligand of the complex 1B7F [96].

2.3 Single-Chain Structural Modelling

Single-chain structural modelling based on sequence, also known as the ‘structure prediction problem’, has been researched for over 50 years [98, 99]. A multitude of tools has been developed to tackle this challenge. In this section, we will primarily focus on the similarities and distinctions between protein and RNA modelling. Subsequently, we will provide a brief overview of the common techniques and describe several popular tools for modelling protein and RNA chains respectively.

2.3.1 Similarities and Differences in Protein and RNA Modelling

The progress achieved in protein modelling compared to RNA modelling is drastic. **The protein structure prediction problem** is generally considered to be solved, minus the spesificities associated with the highly flexible/disordered regions. In contrast, the RNA folding problem, while experiencing recent advancements, remains largely unsolved. Such striking differences sprout from the inherent dissimilarities between the molecules characteristics, principles of their folding, evolution etc.

Nevertheless, despite the differences discussed thereafter, proteins and RNAs share several fundamental characteristics. Both are polymers, consisting of long chains created out of a limited set of basic building blocks. Both are subjected to so-called evolutionary conservation, wherein evolutionary related molecules maintain similar 3D folds despite potential divergence of their sequences. Furthermore, both types of molecules have 4 levels of structure, with traceable links between structures and functions. Folding, often spontaneous, of proteins and RNAs is governed by a number of somewhat common general principles [100].

Among the various differences, what stands out most is the striking distinction in the amount of experimental structural data available. This is particularly frequently mentioned in recent reviews in connection with the development of DL-based methods in the field of structure prediction [101]. At present (September 03, 2023), the total number of protein-only entries in the PDB is 100 times greater than the number of RNA-only entries. This is quite consistent with the structures deposited since the beginning of the year, as the number of protein structures has already reached 7,766, whereas the count for RNA structures is only 77.

The principles of folding, while somewhat common, i.e. the folded structure corresponds to the energy minimum, differ in details. First of all, the principles governing the formation of the secondary structure are different: while protein's secondary structure is stabilised by the hydrogen bonds between particles of the backbone, RNA's secondary structure is primarily defined via canonical base pairing, i.e. hydrogen bonds between side chain residues. The tertiary structure of the RNA is often stabilised by stacking interactions, which is not the case for the proteins [102].

The energy landscapes also differ. The energy landscape of proteins is based on the folding funnel hypothesis, which assumes that proteins fold into secondary and tertiary structures with the lowest possible free energy, typically displaying cooperative two-state thermodynamic transitions, lacking standing-out intermediates [103]. However, RNA folding may involve very rugged landscapes, frequently containing low-energy intermediate states [104].

Finally, RNA shows less dependence between sequence and structure, thus it is more challenging to model the structure of a given sequence relying on the structure of the similar sequences. The level of flexibility in RNA is typically higher than that in proteins.

Despite the differences, the general approaches to modelling protein and RNA structure are fairly similar, partially because both branches were historically developed in a close proximity, so the ideas, which were successful in one field, were applied to the other field.

2.3.2 Modelling Approaches

There are 3 groups of methods applicable to the structure prediction problem: template-based, template free and hybrid [105]. While protein modelling typically entails modelling of tertiary structures, for RNA this field lags behind and modelling of tertiary structures remains challenging. So often RNA modelling entails secondary structure prediction, and not tertiary [101].

Template-based methods are known to be the most accurate, as they leverage sequence-structure relationships, i.e. the idea that proteins/RNAs sharing similar sequences are likely to exhibit analogous structures. They work by aligning the sequence of the target to one or multiple templates with solved structures [106]. These methods are relatively efficient, thus suitable for modelling of large molecules. However, obviously such methods are limited by the number of known structures and are not capable of discovering novel folds. The other limitation is the selection of a correct template, which can be computationally expensive (but not as expensive as template-free modelling) and not accurate enough.

For the proteins, template-based methods encompass homology/comparative modelling and threading. Homology modelling typically involves 4 stages: (i) identification of the template; (ii) alignment of the target and template sequences; (iii) construction and refinement of the model, and (iv) assessment of the model's quality. The threading/fold recognition techniques are used when no closely related homolog with a known structure is available [107].

For RNAs, template-based modelling is represented by the comparative modelling/covariation-based methods and fragment assembly modelling. The first ones generally follow the same steps as protein comparative modelling. In fragment assembly, known RNA structures are divided into fragments, which form a fragment library. To predict an RNA structure, it searches for matching fragments in the library that closely align with the target sequence and assembles them into the final predicted RNA structure [108].

Template-free approaches predict 3D structure from first principles, based on the assumption that the folded states are likely to correspond to the energy minimum for the given sequences. These methods are typically based on various force fields, and they are very demanding computationally, as a large search space has to be covered to find the global minimum. The utilisation of restraints to prevent the exploration of unrealistic conformations is very useful. These methods are useful when no

suitable template can be found, as they are capable of discovering novel folds. However, the accuracy of the predictions is low compared to the template-based modelling approaches. This type of modelling may not be applicable to large molecules.

Hybrid approaches, which do not use templates directly, but still utilise some form of structural knowledge, such as structural patterns or statistical parameters, are a rapidly growing category of methods. Currently, ML and DL methods are known to perform very well for the protein structure prediction. For RNA, the performance lags behind, mainly due to the scarcity of available structural data.

This section would be incomplete without acknowledging AlphaFold2 [109] and its successors [110]. The outstanding performance of this method single-handedly revolutionised the field of both template-based and template-free protein modelling, garnering attention and headlines far beyond the boundaries of the scientific community. AlphaFold2 initiates its process with a multiple sequence alignment (MSA) as input, leveraging the co-evolution data within this MSA. From co-evolution data, the algorithm approximates an inter-residue distance map, which serves as the foundation for structure prediction. An enhancement in AlphaFold2 is the integration of an attention mechanism into its convolutional neural network, significantly improving its outcomes. At the core of AlphaFold2's methodology lies Evoformer, a pivotal component featuring two transformer blocks that collaboratively extract structural insights from the MSA. This extracted information is subsequently relayed to a structural module responsible for constructing a 3D representation of the protein's structure. Furthermore, AlphaFold2 incorporates a Recycling stage, wherein it iteratively refines its predictions using the generated 3D structural information. Shortly after, the AlphaFold DB emerged, covering protein-sequence space with more than 200 millions accurate models [111, 112]

Another great application is ESMfold [113], which employs a masked transformer-based protein language model. It outperforms AlphaFold2 if the prediction is done without MSA utilisation. Simultaneously, this elimination of MSA construction accelerates prediction significantly. Leveraging this methodology, the authors introduced the ESM Metagenomic Atlas, an open atlas of 617 million predicted metagenomic protein structures. Among these, 225 million structures received high-confidence predictions, encompassing novel ones [114].

However, a multitude of exciting challenges remains within the protein modelling domain, including modelling of disordered regions, loops over 20 amino acids and multiple conformers [110].

2.4 Modelling of Complexes

The modelling of a complex involves predicting the 3D structure of each component and their relative spatial orientation. This work is focused on the interaction between a single pair of molecules, rather than on multicomponent complexes. The smaller molecule in the complex is named ligand, and the larger one is the receptor. The main, but not the sole, challenge - molecular flexibility - is amplified when dealing with complexes, as both ligand’s flexibility and receptor’s flexibility should be considered. Different types of complexes possess different characteristics and present different challenges upon modelling. The most studied complex types are protein-small ligand (aka protein-ligand) and protein-protein [115].

Modelling **protein-ligand** complexes typically presents fewer challenges compared to the other complexes types. Here the small molecule often undergoes conformational changes, while the protein conformation remains relatively unchanged (often, but not always [274]). Additionally, information about the binding pocket is typically available. Thus, a high-resolution unbound receptor structure is sufficient for a successful modelling. As for the ligand, it is generally feasible to model all its possible conformations, thanks to its small size and small number of rotatable bonds. The number of resulting

conformations allows to dock each of those. Furthermore, the scoring (evaluation) functions are highly accurate - to the point that they can even solve the interaction problem i.e. predict the binding strength of a small molecule. Therefore protein-small ligand docking can be applied to the (drug) design problem of finding the best binder for a given protein. There are quite some cases of proteins interacting with small oligonucleotides, which bear similarities to the protein-small ligand modelling problem. SsRNA can be cut into short fragments to be treated as small ligands (more detail on fragment-based docking in §2.4.1 and §2.4.4.1).

Modelling **protein-protein** complexes is a more challenging task, especially in the case where the unbound-to-bound conformational changes are significant. The enumeration of all possible conformations is unfeasible, thus, the models are generated based on an unbound structure or, in some cases, using the sequences only. Nowadays, the main challenge of the protein-protein modelling lies in modelling of the highly flexible and disordered regions [116], and in exploring the dynamic of the complexes.

Modelling **protein-RNA** complexes is impeded by a blend of challenges typical of both protein-protein and protein-ligand modelling. RNAs are generally smaller and more flexible than proteins, but much larger compared to small molecules, which results in a very high number of possible conformations, typically too high to handle computationally. As mentioned before, there is less conservation between the sequence and 3D structure of the RNA chain compared to the protein chain, which makes template-based modelling less precise compared to protein-protein modelling. The key difference between protein-protein and protein-RNA interfaces are as follows [117]:

- The atom packing of protein-RNA interfaces is less dense compared to the protein-protein interfaces;
- The protein-RNA interfaces have smaller buried surface area;
- The protein-RNA interfaces typically have more positively charged amino acids which leads to a stronger electrostatic interaction compared to the protein-protein interfaces. The most preferred residues in the protein-RNA interfaces are Arg and Lys, and the least preferred ones are Asp and Glu;
- Stacking interactions, both pi and pi-pi, are a very important part of the protein-RNA interfaces.

Additionally, the number of available protein-RNA structures is much smaller compared to protein-protein and protein-ligand, which makes DL techniques less reliable compared to the protein-protein modelling.

Currently, numerous modelling approaches for each type of complex are available. Similarly to single-chain modelling, methods are classified as template-based, template-free and hybrid. If a suitable template is available, most types of complexes can be modelled via template-based modelling [118, 119]. In the absence of a template, template-free techniques are employed. Nowadays, the distinction between template-based and template-free modelling is gradually blurring as modern modelling tools increasingly embrace hybrid approaches, incorporating both template-based and template-free modelling methods [120, 121, 122, 123]. For example, the state-of-the-art in protein-protein modelling, AlphaFold-Multimer, which is an end-to-end deep learning-based tool, is known to produce more accurate results when combined with the MULTICOM3 package. This package, among other features, contains a template identification feature [124, 125], which makes this tool suitable for both template-based and template-free modelling.

In the following section, we introduce the most suitable approach for protein-RNA modeling, which is equally applicable to other types of complexes, known as docking.

2.4.1 Types of Docking

The terms ‘docking’ and ‘modelling of a 3D structure of a complex’ are often used interchangeably. For greater clarity within the context of this work, the term ‘docking’ specifically refers to ‘template-free modelling of a 3D structure of a macromolecular complex’. Macromolecular docking focuses on predicting the 3D structure of a complex from the input of 3D structures of its individual components. It aims to provide the most probable structures of the complex, assuming their binding, rather than solving the interaction problem. Several common ways to classify docking procedures based on the input and the procedure itself are detailed below.

Bound vs Unbound

Depending on the type of 3D structure used as input, the docking problem can be divided into two categories: bound docking and unbound docking. *Bound* docking consists in re-docking the 3D structures of the component extracted from a solved structure of their complex, and its just a test exercise with no practical utility in real cases. *Unbound* docking uses structures of the components either in an unbound state or bound to another molecule than in the complex to solve (the latter being also referred to as pseudo-native structures).

Ab initio vs Data-Driven

When no a priori information about the binding site is available, the docking problem is referred to as *ab initio (or* blind, free) docking. This contrasts with the *data-driven* or integrative docking methods that use explicit information about the receptor/ligand binding site to guide the docking process. This reduces the search space substantially, which in turn can accelerate and/or intensify the sampling procedure.

Information regarding the binding site can be classified into two categories: interface data and contact data. Interface data is the knowledge of receptor's/ligand’s residues that are directly engaged in the binding. This type of data can be acquired e.g. through mutagenesis. Contact data is the knowledge of pairs of residues in contact. The contact data can be acquired, e.g., through high-resolution cross-linking.

Rigid-Body vs Flexible

Docking algorithms can explore either (i) the mutual orientation of the partners, or (ii) the mutual orientation of the partners along with the flexibility (conformational changes) of one or both partners. Based on this characteristic, docking is categorized into rigid-body docking and flexible docking, respectively (Fig 2.3).

Rigid-body docking, the “lock-and-key” approach, treats both partners as rigid bodies. Typically, the receptor remains fixed in place, while the ligand explores the 3D space around the receptor by undergoing 3D rotations and translations. Rigid-body approaches played a pioneering role in the development of docking methodologies. Due to the lack of explicit consideration of molecular flexibility, the number of potential models is reduced substantially, which allows for an extremely fast sampling [126]. The relative simplicity of rigid-body docking facilitates its implementation. However, the rigidity of the partners significantly limits the accuracy of the docking models when one or both partners undergo conformational changes during or prior to binding. Consequently, the application of pure rigid-body docking is limited to the molecules whose conformational changes upon binding are insignificant.

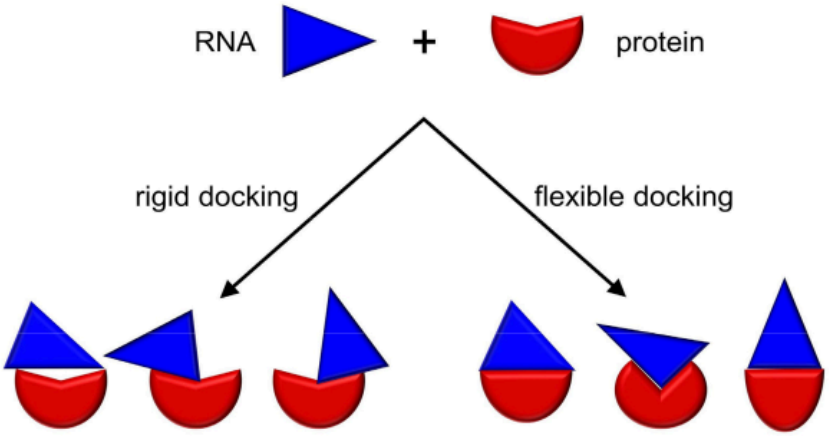


Figure 2.3 - Schematic illustration of the diffenrece between rigid-body and flexible docking approaches [127].

The necessity to account for molecular flexibility led to the development of a broad category of *flexible* docking methods, which aim to model induced fit and/or conformational selection phenomenons (discussed in §1.3.2). The degree of conformational change that occurs upon binding, i.e. the amount of difference between the bound and unbound structure of each molecule, can vary. Based on the magnitude of the changes, docking targets are classified as 'easy' (minor changes), 'medium' (moderate changes) and 'difficult' (large changes) [131]. Based on the approach to address flexibility, docking methods are classified as soft docking, ensemble docking and docking with explicit flexibility modelling [128].

Soft docking methods introduce a certain level of softness or 'fuzziness' into the molecular surface, usually via various types of the molecular representation [129, 130, 131, 132] (more in the next section). The soft surface allows otherwise rigid molecules to overlap, accounting implicitly for small conformational changes or inaccuracies in the side chain orientation. An advantage of these methods is that they are relatively inexpensive computationally and are typically based on rigid body sampling. They are suitable for easy targets but not for modelling large structural rearrangements, and their precision is limited according to the level of fuzziness [133].

The principle of *ensemble* docking methods is to obtain a set of discrete conformations, aka conformational ensemble, of one or both partners prior to docking and perform multiple docking runs with separate conformers [134] to model conformational selection [135]. To obtain the conformations, ensembles of experimental structures or stand-alone modelling can be used. In theory, it is possible to model large backbone movements (e.g. with Normal Mode Analysis [135]), which makes ensemble docking suitable for difficult targets [136, 132]. However, the challenge lies in the sampling and selection of the most representative structures as well as the increased difficulty of discrimination between correct vs incorrect models due to the large number of models [137].

The explicit *flexible* docking methods allow sidechain and/or backbone movement during docking or include post-docking refinement into the protocol [128, 138]. Since explicit modelling of the entire complex is often (but not always [139]) prohibitively expensive computationally, it is necessary to reduce the search space. This can be done by incorporating information about the binding site (data-driven docking) and/or flexible regions. To make the procedure less expensive, many docking procedures include a flexible refinement of only side chain position via energy minimisation, i.e. finding optimal set of torsion angles [140, 141, 142, 143, 144, 145, 146, 147]. Modelling the backbone, especially in difficult targets, presents a bigger challenge compared to side-chain modelling

[148, 149]. Some known methods are replica exchange [150, 151, 131], and backbone displacement followed by MD or Normal Mode Analysis [152, 153, 154, 143]. The primary drawback of explicit flexibility methods is their exceedingly high computational cost.

Additionally, ***fragment-based docking*** approaches are successful in dealing with the flexibility of ligands such as peptides [155], small molecules [156] and RNAs [157]. The principle is to subdivide the ligand into small fragments, dock them separately (typically with a rigid-body algorithm) onto the receptor, assemble them and refine the final model. This allows to bypass exhaustive sampling of all possible conformations of the ligand. The main limitation of this approach is tied to the concept of **hot-spot-binding** and **cold-spot-binding** [270]. The hot-spots (HS) are defined as amino acids within the binding interface that contribute more to the binding energy compared to the other residues in the interface [138]. In the context of the same binding interface, the poses located in the vicinity of the HS are expected to have lower energy/score values compared to the poses in the vicinity of the cold-spots (CS). Thus, within a list of protein-fragment docking poses, HS-bound poses would always be ranked above the CS-bound poses. This is illustrated well by the protein-ssRNA docking with ssRNA'TTRACT of the 'XXX' fragments that belong to poly-X chains. For such a complex, despite the actual number of distinct fragments, a single round of sampling and scoring is performed, and the generated pool of poses is used to represent all fragments. Typically, only one or two of these fragments, presumably related to hot spots, will well-docked, leading to issues during the assembly stage, as there are few or no "correct" poses for CS-bound fragments.

2.4.2 Molecular Representations

The experimentally solved structure of a complex is usually described as a set of points in the Cartesian coordinate system of 3D space, where each point corresponds to the position of an atom. Many docking tools and stand-alone sampling functions are designed to work with such all-atom representation directly [160, 161, 162]. On one side, such high-resolution representation is very straightforward and requires little to no preprocessing of the input 3D structure. The all-atom representation allows for a very detailed search space, which in principle enables the sampling of highly precise docking models and the selection of the most realistic models during the scoring. It also allows for the explicit consideration of both backbone and side-chain flexibility. All-atom representation of the molecules is frequently used in MD simulations, thus multiple generations of researchers have been developing and improving all-atom force fields, e.g. OPLS-aa [163], CHARMM36 [164], which can serve as a base for the docking [19]. However, the utilisation of all-atom representations requires considerable computational resources and thus may be not feasible for big systems. The all-atom energy landscapes tend to be rugged, with a multitude of local minima, which leads to the sampling getting trapped in these minima, preventing the discovery of a global minimum. There is no 'fuzziness' in this representation, so even small conformational changes are not accounted for during rigid-body docking. Consequently, even small inaccuracies in the side-chain position, which may occur, for example, due to the difference in bound vs unbound structure or experimental bias, can prevent the discovery of an accurate binding mode [165].

Reduced representation, obtained by removing some atoms, helps to avoid certain imprecisions, for example, the aforementioned inaccuracies in the side-chain position. This type of representation, with only 3 backbone atoms per amino acid, was used to develop a 6D potential to model proteins and their loops [165]. The same reduced protein representation was coupled with all-atom ligand representation in the development of a scoring function for protein-ligand complexes [166].

Another type of molecular representation is coarse-grained, where a group of atoms is typically represented as a single bead (also called pseudoatom or particle). For example, in ATTRACT coarse-grained protein representation the backbone is represented as two beads, located at N and O,

and the side chain is represented as one or two beads depending on its size, located at the geometric mean of side chain heavy atoms [167, 168]. Another model, CABS, also represents amino acids as 4 beads, located at the center of mass of Cα, Cβ, side chain, and virtual Cα–Cα bond [169]. MARTINI force field represents each amino acid by 1-5 beads (1 for the backbone, 0-4 for the side chain) [170]. The HiRe-RNA force field represents each nucleotide by 6-7 beads [171]. Coarse-grained representation addresses some of the issues presented by all-atom representation [172, 173]. It allows for a more robust sampling, as small inaccuracies in the side chain are accounted for. The energy landscape is smoother compared to an all-atom landscape, which facilitates global minimum discovery. The search space is smaller, so less computational resources are required, which makes multiple docking runs, for example in ensemble docking, less expensive computationally. However, a coarse-grained representation (in some cases a full coarse-grained force field) requires thorough parameterization, e.g. definition of the bead sizes, which is non-trivial and time-consuming.

Many docking methods use a grid-based molecule representation, when a 3D grid overlays the molecule, and each voxel receives a value, denoting if it's inside, outside, or on the surface of the molecule, sometimes with additional features like charge [174, 175, 171, 172]. Other types of representations involve tessellations, creating polyhedra around points (e.g., atom or residue centers) [176], or graph-based approaches, where nodes represent atoms (sometimes with physicochemical/empirical characteristics) and edges indicate connections between them [177].

2.4.3 Evaluation of Docking Models

In a docking test case, it is possible to evaluate the produced docking models by comparing their location with the native position of the ligand. It is done by calculating the ligand root mean squared deviation (LRMSD):

$$LRMSD = \sqrt{\frac{\sum_{i=1}^{n}\left(\left(x_i^a - x_i^b\right)^2 + \left(y_i^a - y_i^b\right)^2 + \left(z_i^a - z_i^b\right)^2\right)}{n}},$$

where n is a number of ligand-composing beads, (x,y,z) are the coordinates of a given bead, superscript *a* indicates a native model (real ligand structure), superscript *b* indicates a docking model.

A model can be labelled as ‘near-native’, if its LRMSD value is under a chosen threshold. The same evaluation approach is applicable for the individual docking poses (models of an isolated fragment). For the fragment-based docking, a pose (model of the fragment) can be evluated using the same metric.

Other existing methics, such as ‘Fraction of the Native Contacts’ (FNAT) [277], ‘Global Distance Test Total Score’ (GDT_TS) [275] or ‘Global Distance Test High Accuracy’ (GDT_HA) [276] are not used in the frame of this thesis.

2.4.4 Rigid-Body Docking

Traditionally, the core of the docking procedure consists of two subsequent stages: sampling and scoring. During the *sampling* stage, 3D models of the complex are generated. They are referred to as docking models. The goal of sampling is to explore the possible mutual orientation of the interacting partners under the guidance of the energy function (or simple shape complementarity function), which results in the generation of putative realistic models. Among these models, one or multiple are expected to closely resemble the 'true' shape of the complex within a given precision threshold.

The following stage, *scoring*, is the ranking (sometimes categorising) of the docking models under the guidance of a scoring function. The difference between a sampling function and a scoring function is detailed thereafter. The aim of the scoring is to distinguish models close to the 'true' complex arrangement (so-called near-natives) from the rest of the models. This process results in the selection of the final model or a relatively small set of highly plausible models.

2.4.4.1 Sampling

Sampling consists of a search algorithm exploring the solution space under the guidance of a sampling/energy function. A sampling/energy function typically aims to generate a wide range of possible binding poses, while a scoring function aims to rank generated poses based on their accuracy. Thus, while both functions are technically used to evaluate docking models, an energy function must compromise between the speed and accuracy of the sampling process [177]. However, it is not uncommon to use the same function for both sampling and scoring, as far as it complies with the requirements of the sampling search algorithm. Unfortunately, the terminology in the docking field has been somewhat jumbled in literature, and it is not uncommon to use 'scoring function' and 'energy function' terms interchangeably [178].

Energy Function Types

In the field of rigid-body docking, two principal approaches are applied to guide the sampling: geometric complementarity and (pseudo)energy minimization (sometimes called simply heuristic) [179, 180].

Geometric complementarity **(**or molecular shape complementarity-based) docking is founded on the idea that the shapes of the receptor and ligand should match to create a stable complex [181]. Thus shape complementarity can serve as a simple energy/filtering function [174] to guide the sampling of feasible models. It works by aligning geometric patches, including concave, convex, and flat surface components, on the molecular surfaces [182]. While pure shape-matching, which was widely used at the dawn of docking [183, 184, 185], does not take into account physical interaction between partners explicitly. But it is known to correlate with the energies of some interactions, such as van der Waals interactions [186]. Nowadays, global or local shape matching is used [187], as well as combination of the shape complementarity with other terms, such as physicochemical complementarity [188]. Geometric methods are typically very efficient compared to energy minimisation methods. They often sacrifice accuracy to achieve efficiency, but the limits are being pushed [189, 190].

Energy minimization-based docking is built around the assumption that the "true" bound structure of the complex corresponds to the energy minimum or at least a low-energy state. These docking algorithms search the energy landscape for the mutual orientation(s) with the most favourable (low) binding energy. This sampling relies on a force field, a mathematical model defining interatomic interactions within a molecular complex. A force field consists of equations and parameters to compute a molecule's potential energy primarily based on atomic coordinates [191]. Often, an approximation of the energy is used to reduce the computational cost.

Search Approaches

The exploration of the search space during the sampling stage can be roughly categorised into systematic search and stochastic search [192].

Systematic search involves the enumeration of all possible mutual orientations. While exhaustive enumeration might seem computationally prohibitive and impractical, some clever approaches more

efficient than brute-force exist to achieve this goal. One of the most common approaches is a discretisation of the search space into a grid. This allows the search space to be reduced to a finite number of mutual orientations, speeding up calculations and enabling sampling density to be controlled by modifying the grid granularity. But it leads to a loss of resolution, i.e. the grid may not be fine enough to capture the binding mode with the precision required for the scoring function to identifiy near-native models [193, 194]. This can become an issue if grid refinement (making grid more dense) is too costly. The most well-known grid-based method is the fast Fourier transform (FFT) correlation [195], which enables the simultaneous sampling of the entire 3D translational space in one step, for a single point in rotational space. FFT relies on the shape complementarity between ligand and receptor and requires a grid-based representation of their structures. The limitation of FFT is the requirement for the energy function to be represented as the sum of correlation functions [196]. Regardless, this search algorithm is highly popular and a great number of docking tools are based on it, for example, PIPER [197], GRAMM [198], ZDOCK [199], HEX (spherical polar Fourier correlations) [200], HDOCK [201], FMFT (Fourier transforms on 5D rotational manifolds) [202]. Systematic search is not limited to FFT. It can be achieved through geometric hashing [203, 204].

Another example of the systematic docking approach is ATTRACT [205, 168] which uses a Quasi-Newton energy minimizer, a gradient descent-based minimisation approach which requires the energy function to be differentiable [168]. ATTRACT reduces the search space by limiting the number of the starting points.

Stochastic search, typically used in energy minimisation-based algorithms, involves random movements within the energy landscape. Pure stochastic methods do not provide a guarantee to find a global minimum, however utilisation of grids, smoothing of the energy landscape and/or the limitation of the search to a specific region increases the chances to locate the global minimum within a given precision threshold [188]. Several methods are widely used, for example, Monte Carlo minimisation in Rosetta [206], Simulated Annealing and Steepest Descent in HADDOCK [142], Iterated Local Search global optimizer in AutoDock Vina [140]. Other popular choices are Swarm Particle Optimisation [208, 209], evolutionary algorithms (often genetic algorithms specifically) [210, 211], and Tabu Search. These algorithms are capable of jumping over certain energy barriers, which prevents them from being trapped in local energy minima, thus increasing chances of finding the global energy minimum [212]. The drawbacks are the computational cost and the aforementioned absence of a guarantee to locate the global minimum.

2.4.3.2 Scoring

Scoring Function Types

Scoring functions can be classified based on the principles of their work and construction. Nowadays, there are 4 distinct categories, along with the hybrid type: physics-based, knowledge-based, empirical and parametric [213].

Physics-based scoring functions, also known as force-field-based scoring functions, are founded on our understanding of the fundamental principles governing the physics and chemistry of molecular interactions [214, 215]. These scoring functions seek to approximate a system's potential energy by directly modelling the physical effects of intermolecular forces. In protein-ligand docking, they aim for an accurate calculation of the binding affinities.

Typically, physics-based scoring functions are formulated as a sum of various energy terms primarily derived from the force field. These components encompass electrostatic forces, van der

Waals interactions, contributions from hydrogen bonding, desolvation effects, occasionally torsional combinations, etc. The typical functional form is:

$$\Delta G_{binding} = \Delta E_{VdW} + \Delta E_{electorostatic} + \Delta E_{H-bond} + \Delta G_{solvation}$$

Physics-based scoring functions are considered 'white-box' models, meaning their behavior is intuitively clear, and the parameter values can be logically explained. However, these models often come with a computational cost, as they frequently require all-atom representations and exhaustive sampling. They have a tendency to provide unrealistic energy values, in part due to the cumulative errors that arise from the aggregation of individual energy terms. Moreover, they may not be well-suited for fragment-based docking since individual fragments may be subjected to different force combinations compared to the entire ligand.

Knowledge-based scoring functions, or potentials, sometimes referred to as potential mean force scoring functions, are based on statistical observations within the multiple solved structures [216, 217, 218, 219, 220, 221]. They are designed to identify local pairwise geometries that are statistically characteristic of the experimentally solved structures, and they are thus expected to be able to distinguish correct models from incorrect ones. Such functions usually map pairwise distances to score (pseudo-energy) values.

Traditionally, knowledge-based functions are presented as a sum of pairwise terms:

$$S = \sum_{i}^{receptor} \sum_{j}^{ligand} w_{ij}(r),$$

where the distance-dependent pairwise potential $w_{ij}(r)$ is derived from Boltzmann statistical distributions of distances [223]:

$$w_{ij}(r) = -k_B T \cdot ln(\frac{\rho_{ij}(r)}{\rho^*_{ij}}),$$

where $\rho_{ij}(r)$ is the numeric density of atom pair (i, j) at distance r and ρ^*_{ij} is the numeric density of the same atom pair in a reference state where atoms are assumed to not interact with each other.

These functions typically are less expensive computationally compared to physics-based functions [213]. They are known to account for entropy and solvation implicitly, however, there are examples of explicit additional terms [223]. The main disadvantage associated with these functions is the sparse data problem [234], meaning that the parameters associated with interactions (geometries, contacts, etc.) underrepresented in the training set are often imprecise or uncertain, inaccurate or undefined. Another challenge is the problem of defining the reference state [235]. An ideal reference state is defined as infinite separation where the particle interaction is zero. To our knowledge, such state is not achievable [226]. Instead, different approximations are used successfully. Several examples are given in the following subsection. Knowledge-based functions are also less interpretable compared to physics-based functions.

Empirical scoring functions, also called regression-based, are a linear combination of individual terms aiming to capture the fitness of the model based on an affinity/energy approximation, regardless of underlying physical/statistical properties [227, 228, 229]. These scoring functions differ from physics-based, as each term has its own weight, optimised via supervised learning. Often, these scoring functions are a sum of penalty and reward terms. A widely used example is ChemScore [230]:

$$ChemScore = S_{H-bond} + S_{metal} + S_{lipophilic} + P_{rotor} + P_{strain} + P_{clash} + [P_{covalent} + P_{constraint}]$$

where S-terms are rewarding scores for hydrogen bonding, bonds with metal ions and lipophilic contacts, and P-terms are penalty scores for frozen rotatable bonds, internal strain energy of the ligand and steric clashes between ligand and receptor.

Empirical functions are less computationally-heavy compared to physics-based functions and require less training data compared to knowledge-based functions (but the data must be labelled). They are highly customisable, as terms can be easily added/removed [231]. Conversely, functions of this type are unlikely to account explicitly for factors uninterpretable by humans. They are highly sensitive to data inaccuracies and likely to not capture underrepresented data (e.g. pi-cation interactions) [213].

Descriptor-based scoring functions, often referred to as ML/DL-based, draw inspiration from quantitative structure-activity relationship (QSAR) based techniques. They are based on a set of various descriptors, which can account for specific interactions, molecular geometry, etc. Compared to the other types of scoring functions, descriptor-based ones contain a considerably higher number of parameters. Popular choices for the creation of these scoring functions are convolutional neural networks [232], graph neural networks [233], graph convolutional networks [234] etc. To date, descriptor-based scoring functions have carved a prominent niche for themselves, as they are often very efficient and accurate [235, 223, 236, 222]. However, many researchers are concerned by their black-box nature [237]. Training of these functions requires large datasets and is very costly computationally. In addition, these functions are very sensitive to the smallest data leaks.

Nowadays, many scoring functions can be classified as hybrid (e.g. [238]), as a combination of multiple approaches can aid in capturing the most significant/relevant terms of each approach for the input data. Multiple papers on enhanced scoring functions are being published, pushing the limits of the accuracy and efficiency of the docking tools. According to the reviews, the application of multiple independent scoring functions, and, if possible, multiple sampling approaches, can improve the accuracy of the docking results [127, 239]

Knowledge-Based Scoring Functions: Construction and Performance

A wide variety of protein-RNA scoring functions have been developed [240, 241, 242, 158, 243, 161, 117, 244, 245]. In the context of this thesis, scoring functions ITScore-RP, DECK-RP and 3dRPC-Score are of particular interest, as they define the reference state differently.

ITScore-PR is an interatomic pairwise distance-dependent potential [162]. It is obtained by a statistical mechanics-based method that circumvents the reference state problem [225]. The main idea behind it is to use the comparison between the predicted atomic pair distribution functions $g_{ij}^{(k)}(r)$ of the protein-RNA complexes (calculated using both native structure and structures generated by docking) and the observed atomic pair distribution functions $g_{ij}^{obs}(r)$ of the native (crystal) structures to improve the scoring potential u_{ij} iteratively:

$$u_{ij}^{(n+1)}(r) = u_{ij}^{(n)}(r) + \Delta u_{ij}^{(n)}(r) = u_{ij}^{(n)}(r) + \lambda\left(g_{ij}^{(n)}(r) - g_{ij}^{obs}(r)\right), \quad (1)$$

where ij stand for different atom pairs in protein and RNA molecules respectively, λ is a convergence parameter, and n represents the number of the iteration.

The distribution function $g_{ij}^{(n)}(r)$ is calculated based on the $u_{ij}^{(n)}(r)$ as follows:

$$g_{ij}^{(n)}(r) = \frac{1}{K}\sum_{k=1}^{K}\sum_{l=0}^{L} P_k^l \cdot g_{ij}^{kl}(r), \quad (2)$$

where K is the total number of complexes in the training set, L is the set of docking models of each complex ($l = 0$ is indicative of the native structure), $g_{ij}^{kl}(r)$ is the distribution function for ij-pair observed in the l-structure of the k-complex; P_k^l is a score-dependant Boltzmann probability, obtained using current $u_{ij}^{(n)}(r)$.

The distribution function $g_{ij}^{obs}(r)$ is calculated based on the crystal structures:

$$g_{ij}^{obs}(r) = \frac{1}{K}\sum_{k=1}^{K} g_{ij}^{k*}(r)\,, \quad (3)$$

where $g_{ij}^{k*}(r)$ is calculated as the densities of the ij-pair non-neighbouring residues in the reference sphere.

This method converges when the absolute difference between $g_{ij}^{(k)}(r)$ and $g_{ij}^{obs}(r)$ is under a certain threshold. It is sensitive to the initial potential $u_{ij}^{(0)}$, as the iterative procedure may be trapped in local optima.

Recently, this method was used to develop ITScore-NL, a scoring function for nucleic acid-ligand interactions [246]. Interestingly, in this function the stacking and electrostatic interactions are included explicitly, on top of the usual for knowledge-based function pairwise potential. Since stacking interactions are of a high importance for the protein-ssRNA interactions, it could be interesting to incorporate a stacking term into protein-RNA scoring function as well.

DECK-RP is a Distance- and Environment-dependent, Coarse-grained and Knowledge-based function for protein-RNA complexes [117]. The reference state includes the mol-fraction corrected component, which takes into consideration the interface concentration (preferences of amino-acid residues and nucleotides to be in the interface), and the decoy-based component, which takes into consideration all structures generated by docking, along with the propensities of different types of amino acids and nucleotides. Then nucleotides and amino acids are clustered into several types of residues resulting in 168 pairs of interacting residues. The energy of each ij-pair is defined similarly to (6), replacing number of ij-contacts by the probability of ij-contact.

The observer probability (numerator) is obtained as fraction of near-native ij-contacts in the bin d over all near-native pairs in this bin. The expected probability (denominator) calculation involve mole fraction component:

$$P_{exp}(i,j,r) = \frac{N_d(i,j,r)}{\left(\frac{f_i(r_{cut})\cdot f_j(r_{cut})}{f_i(r)\cdot f_j(r)}\right)\sum_{ij}^{a} N_d(i,j,r)}, \quad (8)$$

where $N_d(i,j,r)$ is the number of ij-pairs at a distance bin r in all modelled structures, and $f_{i/j}(r)$ is the mol-fraction of the i/j residue, a is a convergence parameter.

3dRPC-Score is a pair-conformation-dependent scoring function [240], based on the concept of conformational similarity (i.e. the nucleotide-residue pairs should have the same energy if their conformations are similar [278]) rather than a more typical distance-dependance. All nucleotide-residue pairs were extracted from crystal structures (of ribosomes) and classified using k-means based on the relative RMSD between conformations into 10 classes per pair (800 classes in total, for 20 types of amino acids and 4 types of nucleotides). The pairs in each class have similar conformations, thus they are considered to have the same energy, expressed as follows:

$$E_{ij}(C) = -ln\left(\frac{P_{ij}(C)}{P_i P_j * P_v}\right), \quad (9)$$

where C is the class, $P_{ij}(C)$ is the occurrence probability of the ij-pair in class C, $P_{i/j}$ is the probability of the residue i/j in the interface based on the solvent accessible area (determined from the statistics of the selected training structures), and P_v is the probability of class C in the whole conformational space of nucleotide-residue pairs in the reference state.

In the reference state, each class has the same probability and the statistical potential (9) can be written as:

$$E_{ij}(C) = -ln\left(\frac{P_{ij}(C)}{P_i P_j}\right) + constant, \quad (10)$$

where the $constant$ is equal to $ln(P_v)$ and has negative value as $P_v < 1$.

The best values for P_v were determined by ranking the training set with different values and comparing the results.

Performance

The performances of ITScore-PR, 3dRPC-Score and DECK-RP were compared [240] on a full Huang and Zou benchmark (72 structures) [247], and 64 structures from Pérez-Cano et al benchmark [248]. Scoring functions were evaluated by their ability to select at least 1 near-native model (LRMSD<10Å) in top 1 or top 10 solutions. The results suggest a superior performance of ITScore-PR (Tab. 2.1). These papers do provide information regarding the performance spesifically on protein-ssRNA complexes.

Table 2.1 - Success rates (%) of scoring functions on different benchmark test sets.
HZR stands for the Huang and Zou benchmark with sampling performed using RPDOCK [117];
HZZ stands for the Huang and Zou benchmark with sampling performed using ZDOCK [199];
PCZ stands for the Pérez-Cano et al benchmark with sampling performed by using ZDOCK [199].

	Top 1			Top 10		
	HZR	**HZZ**	**PCZ**	**HZR**	**HZZ**	**PCZ**
ITScore-PR	46	41	12	64	58	28
3dRPC-Score	46	34	19	60	50	42
DECK-PR	36	28	21	54	45	44
Nb of complexes	72	72	64	72	72	64

2.4.4 Selected Protein-ssRNA Docking Tools

Protein–ssRNA complexes are notoriously complicated to dock. Classical docking approaches typically rely on having an unbound structure as a starting point. But when dealing with ssRNA, such a structure is not available because of its **disordered nature in its unbound state**. The inherent flexibility of ssRNA makes docking of systematically modelled conformations computationally impractical. Applying DL to protein-RNA docking is challenging due to the relatively **limited number of solved structures** available and also because, within each structure, the interactions between RNA and protein constitute only a small fraction of all atomic contacts.

Last point is illustrated by RoseTTAFoldNA, a DL-based end-to-end predictor designed to predict the structures of protein-nucleic acid complexes without the need for homologs [272]. This model, somewhat similar to AlphaFold, takes as input one or more aligned protein and nucleic acid sequences, and transforms this information in parallel 1D (sequence), 2D (residue-pair distances) and 3D (structure in cartesian coordinates) tracks, outputting 3D protein-nucleic acid structures. It is based on all-atom representation, and takes into account both atom-atom distances, as well as orientations of the nucleotide (a coordinate frame defined by atoms of the phosphate group, P, OP1 and OP2, is used).

Alas, all but one of them appear in the ‘Failure models’ category, with the fraction of native contacts (fnat) under 0.05 and interface RMSD (iRMSD) under 6.7Å (Tab 2.2). A single successfully modelled protein-ssRNA complex (PDB id: 4PWM with the fnat = 0.77 and iRMSD = 2.0Å) had a close homologue in the training set. The paper notes that such small single-stranded nucleic acids or slight deviations in monomer structures represent 20% of RoseTTAFoldNA failures, which leads to the conclusion that a main RoseTTAFoldNA limitation is related to single-stranded nucleic acids.

Table 2.2 - Statistics of the proten-ssRNA modelling achieved with RoseTTAFoldNA [272].

PDB id	**fnat**	**iRMSD, Å**
4PMW	0.77	2.0
6YYM	0	12

PDB id	fnat	iRMSD, Å
7A9W	0	12
7A9X	0.05	6.7
7M5O	0	50
7B0F	0.02	14
7OM3	0.04	10

Thus, fragment-based approaches are the only ones currently capable of handling docking of small ssRNAs to the protein to some extent. There are 4 existing approaches:

- RNA-LIM represents each nucleotide by one non-oriented bead and could only predict their position at 15Å resolution when tested on only one example [249];
- FBDRNA uses mononucleotide fragments in all-atom representation, docked with MCSS on a pre-defined binding site. While showing discriminative power on nucleotides' positions, it could not provide accurate models for full oligonucleotides [250];
- RNP-denovo, a Rosetta method to simultaneously fold-and-dock RNA to a protein surface, uses the exact position of a few nucleotides [251], which would be unavailable for real-life docking cases;
- ssRNA'TTRACT, the state-of-the-art in the fragment-based protein-ssRNA docking, is the most accurate approach that uses only a protein structure and the RNA sequence as input. It uses trinucleotides as RNA fragments and an overlapping criterion based on LRMSD for assembly [252].

ssRNA'TTRACT is what lies in the core of this research, thus this approach is presented in detail in the following section.

2.4.4.1 ssRNA'TTRACT

The idea behind ssRNA'TTRACT is to handle ssRNA flexibility by subdividing its sequence into fragments that are small enough for their conformations to be exhaustively (including close-to-bound conformation) sampled within a given accuracy threshold. Each fragment is sampled and scored individually, generating a pool of docking poses - certain positions and orientations of particular conformations of the fragment with respect to the protein. The poses of the adjacent fragments are then assembled into the docking models, i.e. models of the full ssRNA chain.

Fragment Library

Prior to the docking, it is mandatory to have a fragment library of the trinucleotide conformations, which aims to represent all naturally possible conformations of a given trinucleotide sequence under a certain threshold. Tri-nucleotides possess distinct advantages when compared to di-nucleotides or mono-nucleotides, as they engage in a greater number of interactions with the protein, resulting in binding to more specific positions. Tri-nucleotides also offer an advantage over tetra-nucleotides, as they have a significantly smaller number of possible conformations, making them more manageable.

In this work, an in-house library is used [253, 254]. This library contains a set of fragment conformations, extracted from existing 3D structures and clustered by pairwise RMSD with a 1Å threshold. To increase the number of conformations for each motif (trinucleotide sequence), all combinations of artificial mutations were applied to each fragment to transform purines and pyrimidines (G⇔A and C⇔U). It was done under the assumption that removal/addition of a single heavy atom has an insignificant effect on the overall conformation of a trinucleotide, and because we aim at being as exhaustive as possible at the cost of allowing some potential inaccurate conformations. This procedure was carried out using ProtNAff [255].

Docking Initiation, Sampling, Scoring

The input for the docking is a rigid protein structure (coarse-grained) and an RNA sequence (the RNA chain is assumed to be single-stranded). The RNA sequence is split into trinucleotides overlapping by 2 nucleotides, so if, for example, the RNA sequence is n nucleotides long, it is split into n-2 fragments. Each unique sequence motif, present among all fragments, is docked on the protein via a rigid body docking procedure.

Two options for the initiation of the sampling are available: ‘systsearch’ and ‘randsearch’. In ‘systsearch’, each available conformation of a motif is placed in 228 orientations at predefined positions around the protein, at a certain distance from its surface. Each such conformation at predefined position and orinetation serves as a starting point for a sampling. In ‘randsearch’, a random position is chosen on a sphere centered around the protein's center of mass. A random conformation of the motif is selected from the library, places in the chosen position in a random orientation and used as a staring point for the sampling. In this work, ‘randsearch’ sampling strategy is used with 30 million starting points used by default.

Next, the poses are driven from their staring points towards the protein surface via score minimisation with gradient descent-based minimisation steps, until convergence. During the sampling stage, to enhance computational efficiency, pre-calculated score values on a grid for each ligand bead type are used. The resulting poses are clustered using an RMSD threshold of 0.2Å to remove redundant poses. During the scoring stage, each pose is assessed based on the interactions between pairs of beads that fall within a defined distance threshold. These poses are ranked from the lowest to the highest score, and typically, the top 10 million best-scored docking poses are retained for each motif. The goal is to maintain a minimal number of poses to prevent computational bottlenecks during the assembly, while simultaneously ensuring that at least 1 near-native pose per fragment is preserved. This is imperative to get at least 1 near-native chain.

Assembly Procedure

The final stage, assembly, involves the creation of a connectivity oriented graph. Each node in this graph represents a docking pose, retained after the scoring. An edge between two nodes is added if the two poses share compatible sequences and if the pairwise atomic distances between their common 2 nucleotides are all below a certain threshold (overlap criterion). The overlap is defined empirically using test cases. It should be as strict as possible to eliminate the maximum number of incompatible poses, thus reducing the size of the graph; but not so strict as to eliminate the near-native chains, i.e. chains containing only near-native poses. After this, to reduce the graph size, the poses that are not a part of any full chain are eliminated. Remaining connected poses are assembled in the chains by averaging the positions of the overlapping atoms.

Performance and Limitations

ssRNA'TTRACT is capable of sampling and scoring near-native poses (nns) and assembing them in near-native chains, as shown in [159] (Tab 2.3). However, **ssRNA'TTRACT suffers from 2 main main limitations - sampling problem and scoring problem.**

Table 2.3 - Comparison to the bound form of the poses obtained by (I) bound docking and (II) position-specific filtering of chain-forming poses [159].
Hits and near-hits correspond to poses with LRMSD<2Å and <5Å respectively. Chains are formed by 6 fragments. Position-specific filtering is a filtering based on the propensity of each pose to form chains.

		I. Docking poses (top 20%)				II. 6-pool filtered poses			
		Hits	Near-hits	Total	min RMSD	Hits[a]	Near-hits[a]	Total	min RMSD
1B7F	frag1	11	110	2566	0.7 Å	10	17	23	0.7 Å
	frag2	13	82	2409	1.0 Å	13	17	24	1.0 Å
	frag3	29	92	2568	0.9 Å	25	31	47	0.9 Å
	frag4	31	110	2526	0.5 Å	25	31	58	0.5 Å
	frag5	32	109	2471	0.5 Å	30	40	68	0.5 Å
	frag6	20	62	2517	0.5 Å	20	35	89	0.5 Å
	total	*136*	*565*	*15157*		*123*	*171*	*292*	
	chain					6.1×10^5	6.1×10^5	6.1×10^5	1.0 Å
1CVJ	frag1	9	10	2193	1.2 Å	9	10	22	1.2 Å
	frag2	21	77	2036	1.1 Å	21	42	42	1.1 Å
	frag3	17	56	1881	0.9 Å	16	19	19	0.9 Å
	frag4	13	27	2110	0.9 Å	13	13	13	0.9 Å
	frag5	3	6	2088	0.8 Å	3	3	3	0.8 Å
	frag6	5	10	2174	0.6 Å	5	5	5	0.6 Å
	total	*68*	*186*	*12482*		*67*	*92*	*104*	
	chain					3.2×10^5	3.4×10^5	3.4×10^5	0.9 Å

In the frame of this thesis, ssRNA'TTRACT docking was conducted on a benchmark of 410 experimentally solved protein-ssRNA complexes, collected from PDB by February 2021, including redundant cases, totaling 1998 fragments (Tab 2.4). It was observed that for 67% of these fragments, no near-native poses (nns) with an LRMSD under 2Å could be found within the pool of 10 million docking poses. This issue, known as a *sampling problem*, affected 88% of the complexes in total.

Table 2.4 - Statistics of the sampling and scoring of ATTRACT scoring function.

Sampling	LRMSD < 2Å		LRMSD < 5Å	
	< 1 nns	< 100 nns	< 1 nns	< 100 nns
% of failed fragments (< x nns sampled)	67	98	15	42
% of complexes with over 1 failed fragments	88	100	31	61
% of complexes with all fragments being failed	43	91	8	26
Avg. number of nns in 10 million sampled	10		4,223	

Scoring (sampling+scoring)	< 1 nns in top20%	
% of failed fragments	4 (71)	10 (25)
% of complexes with over 1 failed fragments	22 (100)	15 (46)
% of complexes with all fragments being failed	47 (90)	6 (14)
Avg. % of nns in top20% (out of all sampled nns)	0.1	33

Furthermore, out of the remaining 33% of fragments with at least 1 nn sampled, only 2% had over 100 nns. Although having fewer than 100 nns is not typically classified as a sampling problem, these fragments require a higher percentage of near-native poses to be top-ranked in to facilitate successful assembly. In instances where the LRMSD threshold is relaxed to LRMSD under 5Å, a sampling problem was encountered for 15% of the fragments and affected 31% of the complexes.

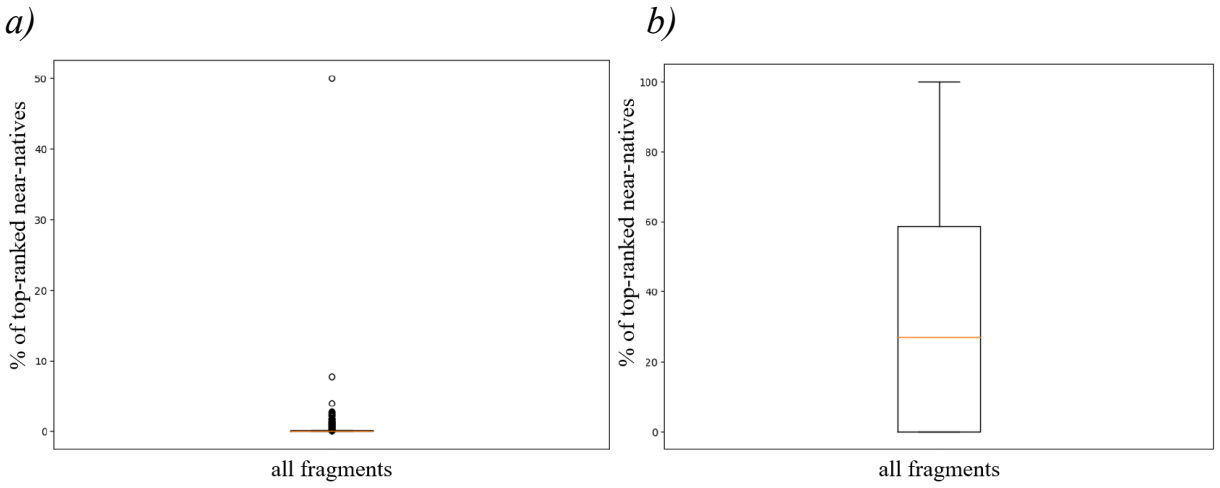


Figure 2.4 - Boxplot of the percentage of top-ranked near-native poses out of all sampled near-naive poses, calculated over 1998 fragments. a) LRMSD<2Å. Note that for this graph y-axis reaches a maximum of 50%; b) LRMSD<5Å.

For nns with LRMSD under 2Å, a mere 0.1% of all sampled poses fell within the 20% of top-ranked poses (top20%) (Fig 2.4 a). Given an average of 10 nns sampled per fragment, it is unlikely to create near-native chains using a pure *ab initio* docking procedure with such threshold. For the relaxed threshold (LRMSD < 5Å), on average, 33% of all near-native poses were within the top20%, offering a greater chance of successful assembly (Fig 2.4 b). However, for 10% of the fragments, no nns were present in the top 20%, which is known as a *sampling problem*. The sampling problem impacted 15% of the complexes, such that were not impacted by a scoring problem. If both sampling and scoring problems are considered, the total of 46% of the complexes are impacted.

Both sampling and scoring problems originate from the parameters of the ATTRACT scoring function. These parameters were derived from a benchmark of double-stranded RNA-protein structures back in 2010 [256], which offers potential for improvement in the context of protein-ssRNA docking.

2.5 Conclusion

In this chapter, we have introduced general approaches applicable to single-chain modelling and molecular docking. Our discussion has encompassed different types of docking, various sampling techniques, and distinct categories of scoring functions. Several noteworthy examples of knowledge-based protein-RNA scoring functions have been presented. To transition from a general concept of rigid-body docking to the specific problem of protein-ssRNA docking, we have highlighted the inherent complexity of modelling these complexes. This includes the absence of the unbound ssRNA structure, which inhibits the application of traditional rigid-body and semi-rigid docking methodologies; the high flexibility of ssRNA, which inhibits exhaustive modelling of possible conformations; and the low number of solved protein-ssRNA structures, which limits the application of DL-based methods. We have concluded that the current approach to docking protein-ssRNA complexes involves fragment-based methods, with the state-of-the-art being the ssRNA'TTRACT method. While this method demonstrates good overall performance, it is affected by both sampling and scoring problems, along with inherent limitations of fragment-based approaches in sampling and scoring HS-bound and CS-bound fragments.

In the next chapter, we will present our attempts to enhance the performance of ssRNA'TTRACT through the optimisation of docking parameters.

Part II

Contributions

Chapter 3: Protein-ssRNA Docking Parameters Optimisation

3.1 Aims

In the previous chapter, we introduced the state-of-the-art in fragment-based protein-ssRNA docking, ssRNA'TTRACT. As discussed, this method suffers from sampling and scoring problems, partly caused by the original parameters of the ATTRACT scoring function (ASF), which were not designed specifically for ssRNA. Therefore, these problems could be addressed by the optimisation of the ASF parameters. Despite the sampling problem being more critical, it is also more challenging to address since re-sampling is more computationally expensive than re-scoring. Thus, our focus here is on addressing the scoring problem, with the expectation that improved scoring parameters are likely to indirectly benefit the sampling aspect.

In this chapter, we delve into several unsuccessful attempts to optimise the parameter set for protein-ssRNA docking and propose a hypothesis on why these attempts fell short. Additionally, we introduce two potential avenues for further improvement – the use of all-atom force fields to estimate initial parameter values and a brute-force approach to evaluate the change in the docking performance caused by an update of a small subset of parameters. Lastly, we touch upon the idea of explicitly incorporating stacking interactions into the scoring process.

3.2 Introduction

3.2.1 Coarse-Grained Representation

ssRNA'TTRACT uses a coarse-grained representation and a knowledge-based scoring function [168, 257, 256]. Each amino acid is represented by 3 to 4 beads (pseudoatoms): 2 beads correspond to N and O in the backbone, and 1 or 2 beads, located at the geometric mean of side chain heavy atoms, describe short and long side chains, respectively. In total, there are 31 protein bead types, including 11 modified amino acids. Each nucleotide is represented by 6 to 7 beads: 3 beads describe the backbone (1 bead for the phosphate group, 2 for the sugar), and 3 to 4 beads describe the base, pyrimidines and purines, respectively. This coarse-grained representation allows for efficient calculations and maintains reasonable details of physicochemical features, e.g. retains information about the orientation of nitrogen bases in space. In total, there are 17 RNA bead types, describing exclusively standard nucleotides.

3.2.2 Scoring Function

The interactions between protein and RNA beads are described by a distance-dependent potential in two forms - attractive and repulsive (Fig 3.1).

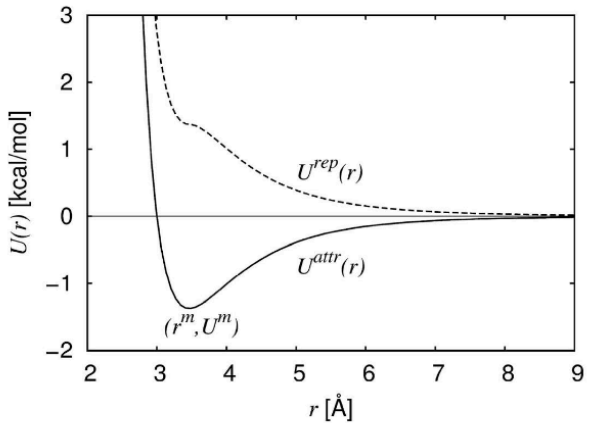


Figure 3.1 - Global shape of the scoring function for protein-ssRNA interactions, employed by the ATTRACT docking engine.

The attractive potential is a Lennard-Jones-like function with a soft repulsive term:

$$U_{ij}^{attr}(r) = \varepsilon_{ij}\left(\frac{\sigma_{ij}^{8}}{r^{8}} - \frac{\sigma_{ij}^{6}}{r^{6}}\right), \quad (1)$$

where r is the inter-bead distance, $i = \overline{1,31}$ is a protein bead type, $j = \overline{1,17}$ is an RNA bead type, the parameter σ_{ij} indicates the range of the ij interaction (to be precise, it is equal to the value of r at

the point, where the curve crosses the x-axis, i.e. where the score is equal to zero), and parameter ε_{ij} indicates the strength of the ij interaction.

The repulsive potential is defined as follows:

$$U_{ij}^{rep}(r) = \begin{cases} U_{ij}^{attr}(r) + 2U_{ij}^{m}(r), if\ r \leq r_{ij}^{m} \\ -U_{ij}^{attr}(r), if\ r > r_{ij}^{m} \end{cases} \quad (2)$$

where U_{ij}^{m} and r_{ij}^{m} corresponds to the minimum of $U_{ij}^{attr}(r)$.

The total score of a docking pose is defined as the sum of the individual bead-bead scores.

The optimisation process targeting the scoring function parameters σ_{ij} and ε_{ij} was performed via two distinctive stages [257]. Firstly, an initial parameter set was obtained by fitting equations (1) and (2) to a statistical potential derived from Boltzmann statistical distributions of distances in the PDB. Next, σ_{ij} were optimised through a Monte Carlo-like (MC) approach. A set of adjacent σ values - for a single parameter at a time - was explored by performing potential energy minimizations for a set of native complexes. The value resulting in the best score was kept. Secondly, ε_{ij} were tuned using a set of 200 decoys (LRMSD > 5Å) per complex. Similarly to the first stage, a set of adjacent ε values was explored. The best values were selected based on the ability to provide the lowest score for native-like complexes with respect to their corresponding decoys. The set of docking parameters obtained via the described process is referred to as the original parameter set in the text below. It consists of 1054 parameters. The values of σ are ranging from 2 to 6.4, and the values of ε are ranging from 0.02 to 20.

Equations (1) and (2) with the original parameter set are used for both sampling and scoring, as described in §2.4.4.1.

A perfect scoring function would be capable of establishing a linear dependency between the score-rank and the LRMSD-rank of the near-native poses, regardless of the ranking of the decoys. Given the current state of the protein-ssRNA docking field, it is acceptable for practical use to rank near-native poses by their score on top of the ranking list.

3.3 Optimisation of the Protein-ssRNA Parameter Set

As described in §2.4.4., the current ASF often faces the *scoring problem*, i.e. fails to assign high ranks to the near-native protein-ssRNA poses. Therefore, a large number of the docking poses, around 10 million, must be kept post-scoring to assemble near-native chains. Such a large number of poses per fragment hurdles the assembly as its computational cost is very high and the number of chains (mostly non-natives) is immense, which makes identification of the near-native chains out of the pool a complicated task with unreliable results.

The scoring problem can be addressed by the optimisation of the original docking parameters or by the development of a new scoring function. Since ASF is used for both sampling and scoring, both options could improve the quality of the sampling as well (with the limitation that the function must remain differentiable for the sampling stage, or that another minimisation process than gradient-descent is used). Both options are highly challenging and time-consuming tasks. The development of an optimisation protocol for the parameters of the current scoring function had been

initiated by a former CAPSID student, Agnibha Chandra. Thus, it was decided to commit to the parameters optimisation.

The original parameters ε_{ij} and σ_{ij} of the ASF were subjected to optimisation with the objective of enhancing fragment-based protein-ssRNA scoring. Our aim was to increase the number of near-native poses within the list of 10% of top-ranked poses, which can be achieved by minimising the average score of the near-native poses. To avoid the simultaneous minimisation of average scores for non-native poses, the focus is on the maximisation of the discrepancy between the average scores of near-native poses and non-native poses. To avoid the scaling problem, the average score difference is normalised by the average score of both types of poses.

Near-native poses are defined by LRMSD values below 3Å, while non-native poses have LRMSD values exceeding 5Å. Intermediate poses, defined by LRMSD values between 3Å and 5Å, were excluded to establish a distinct boundary between near-natives and non-natives. This separation theoretically assists in increasing the score difference assigned to these pose types.

Based on the description above, the objective function $f(\varepsilon, \sigma)$ was formulated as follows:

$$f(\varepsilon, \sigma) = \frac{\overline{S_{near-naive}} - \overline{S_{non-naive}}}{\bar{S}}, \quad (3)$$

where $\overline{S_{near-naive}}$ and $\overline{S_{non-naive}}$ are the average score of the near-natives and non-natives poses respectively and $\bar{S}$ is the average score of both types of poses.

For the maximisation of (3), it was decided to use a metaheuristic approach to approximate global optimisation, namely, the Monte Carlo Simulated Annealing (MCSA) algorithm [258, 259].

3.3.1 Monte Carlo Simulated Annealing

The MCSA algorithm combines the principles of both MC methods and Simulated Annealing (SA), yielding a stochastic optimisation technique capable of exploring complex search spaces. This approach combines random sampling, inspired by MC methodologies, and annealing-like temperature control to find near-optimal solutions. A well-known advantage of using SA is its capability of escaping local minima. This iterative algorithm starts with high randomness (random walk-like behaviour) that allows for such escape(s) and gradually becomes more deterministic (hill-climbing-like behaviour) to refine the solution.

Each iteration, known as an annealing step, requires the initialisation of the parameters undergoing optimisation, the initial temperature value that is a starting point of the annealing process, and the selection of a cooling schedule - a function describing the gradual reduction of temperature. During the annealing step, a set of neighbouring parameters is generated by a small random perturbation of the current parameter values. Two objective function values are calculated: one using the current parameter set and another using the neighbouring parameters. As the aim is to maximise the objective function, if the neighbouring set's value surpasses the current set's value, the new set is accepted. Otherwise, acceptance depends on temperature. High temperatures allow the acceptance of even worse solutions, while low temperatures only accept better or just slightly worse solutions.

3.3.2 Experiments

3.3.2.1 Dataset

The benchmark I used for the optimisation consisted of 42 protein-ssRNA complexes. These complexes were solved experimentally through NRM or X-RAY with a resolution higher than 4Å. Each complex contains RNA with a single-stranded region (at least 3 nucleotides long) that is bound to the protein, i.e. with at least one pair of heavy atoms located within 5Å per nucleotide. All complexes had been deposited on PDB by July 2018.

This benchmark can be decomposed into a dataset of 309 distinct protein-bound trinucleotides (fragments) referred to simply as data cases further in the text. This benchmark is heavily dominated by the fragments with UUU and AAA motifs, 29% and 15% of the size of the dataset respectively. Each data case was docked using ATTRACT and ASF with its original docking parameters (following the procedure described in §2.4.4.1). Each pose is labelled as near-native or non-native. As mentioned before, intermediate poses are removed. Note that if a given RNA chain contains several fragments with identical motifs, such cases require a single docking run, followed by an individual LRMSD computation for each case.

The number of sampled near-native poses (LRMSD<3Å) per case varies from 0 to 525. As anticipated, the majority of near-native poses are of UUU and AAA motifs. Unexpectedly, their percentages over all near-native poses for all cases are almost equal, 36% and 33% respectively, despite the UUU fragments being twice as frequent as AAA. The final training set contains 36 complexes, and the validation set contains 6 complexes. Comprehensive details of the benchmark and its docking statistics can be found in Appendix A.1.1.

3.3.2.2 Protocol

The protocol comprises the following stages: initiation, optimisation, testing and validation. The flowchart for each stage is provided in Appendix A.1.2.

The initiation stage necessitates the selection of the following elements:

- The initial temperature T_0;
- The cooling schedule among the following options: logarithmic, exponential, linear, zigzag, wall climber, and constant (implemented by Agnibha Chandra);
- The initial parameter set, which can be the original parameter set, one of the sub-optimal sets obtained during previous optimisation rounds, or a set of random numbers adhering to the domain of the function;
- The step size to sample neighbouring parameters, representing the maximum interval by which each parameter can be changed;
- The strategy for sampling neighbouring parameters among the following options: 'normal' (where parameters are changed with unscaled random values) or 'adaptive' (where the magnitude of change is scaled at each step by a coefficient $\frac{current\ T}{T_0}$);
- The stopping criterion defined as a number of consecutive steps where the change in the objective function value remains below a specified threshold. If convergence of a scoring function is not observed, a maximum number of iterations after which the optimisation process should be terminated is specified.

The optimisation process involves a grid-accelerated calculation [260] of the scores for the docking poses of each fragment of each training complex. Scores are calculated for all sampled near-natives and only the first 1 million non-natives. The average score difference is computed across all training complexes. Once the stopping criterion is met, the new parameter set is preserved. Additionally, the option exists to save multiple intermediate parameter sets, which can serve as starting points for subsequent annealing steps.

The testing consists of rescoring the entire training set with a new parameter set. The overall improvement I is calculated as follows:

$$I = \sum_{tCmp} \sum_{fr} \frac{n_{TR} - n_{TS}}{n} \cdot 100, \quad (4)$$

where $tCmp$ are the complexes of the training set; fr are the fragments of a given complex; n_{TR} is the total number of near-native poses in the 10% top-rescored poses, i.e. scored and ranked by the new parameter set; n_{TS} is the total number of near-native pose in the 20% top-scored poses, i.e. scored and ranked by the original parameter set; and n is the total number of sampled near-natives.

If for a given new parameter set $I \geq 20$, then it is labelled as satisfactory and validation is carried out. *The validation* process is the application of the new parameters to the validation set of complexes for scoring.

3.3.3 Results and Discussion

I performed multiple rounds of optimisation and tested multiple combinations of hyperparameters. Those tests suggest that an exponential cooling schedule along with an adaptive approach are best suited for the used benchmark. The distinct setups that were used to create 4 parameter sets are presented in Tab. 3.1.

Table 3.1 - Hyperparameters for the training of the scoring parameters with MCSA.

	Hpar1	**Hpar2**	**Hpar3**	**Hpar4**
T_0	1000	1000	1000	1000
Cooling schedule	Exponential	Linear	Linear	Exponential
Initial par set	ATTRACT	Random	Intermediate	Intermediate
Step size	0.01	0.05	0.025	0.02
Sampling strategy	Adaptive	Adaptive	Adaptive	Adaptive
Max steps	5000	5000	50,000	50,000
Min Δ score	1e-13	1e-13	1e-13	1e-13
Stopping criterion	10	10	10	10

During the testing stage, each of the optimised parameter sets was deemed satisfactory and validation was performed. The numbers of near-natives present in the top10% were compared. While

some optimised parameters outperform the original ones for certain fragments, the overall performance of all sets is very similar. The results of the validation are detailed below (Fig 3.2). Because MCSA is a stochastic algorithm, the possibility exists that the optimal solution might not have been reached. More detailed results for each distinct fragment can be found in Appendix A.1.3.

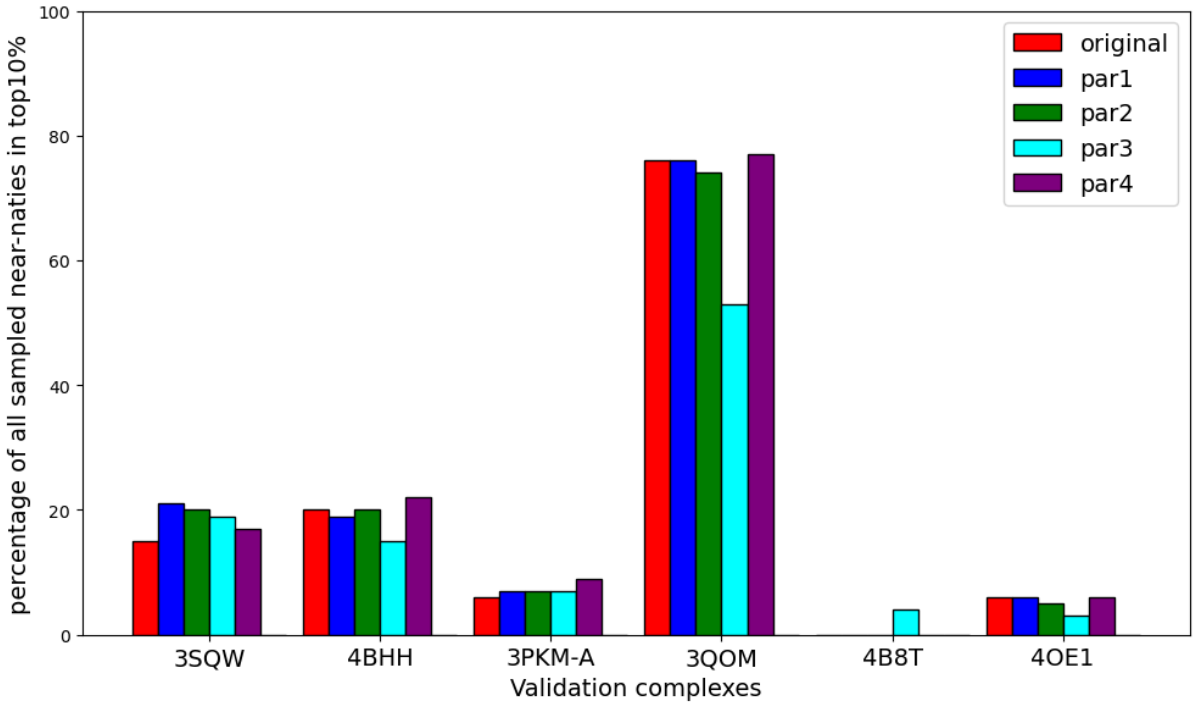


Figure 3.2 - Comparison of the performance of different parameter sets on the validation set. Each cluster of adjacent bars shows the percentage of the near-natives located in the top10% of poses, ranked by one of the 5 different parameter sets, listed in the legend.

The diversity of the binding modes (discussed in §1.3.5), and the results of MCSA optimisation led to the formulation of the hypothesis that **a singular parameter set might be insufficient for accurate scoring of protein-ssRNA fragments**. This statement is the main conclusion regarding the MCSA optimisation project.

If this hypothesis is disproven, the further optimisation of the docking parameter set could be carried out by replicating the stochastic optimisation protocol developed by Piotr Setny for the creation of the original parameter set, but with a larger benchmark containing only protein-ssRNA structures, to create a parameter set specific to protein-ssRNA. Working on the larger, exclusively protein-ssRNA dataset could result in a set of more accurate parameters. However, there is no guarantee of a solution to the challenge of parameter sensitivity to the order of pairwise potentials adjustments, mentioned in the original paper. Thus, it could be more promising to explore the field of non-stochastic optimisation techniques, suitable for global non-convex optimisation, such as fuzzy logic [261] or Bayesian optimisation [262].

Creation of the Distinct Parameter Sets for the Most Common Motifs

Considering the distribution of generated near-native poses across different motifs, one might argue that the abundance of near-native poses for 'AAA' and 'UUU' motifs should be taken into account. Without such consideration, optimizing the parameters for the A-X pair of beads, where X represents any protein bead type, may result in overfitting on the distances specific to 'AAA'

fragments and underfitting on the distances typical of fragments where only one out of three nucleotides is A.

Normalisation of the number of near-native poses per motif used in the optimisation process could address this concern. An alternative could involve creating three separate parameter sets: one for 'AAA', one for 'UUU', and one for other motifs. This approach would enable parameter calibration for interactions between beads in fragments where only one out of three nucleotides is A, as well as ensuring that the parameters for 'AAA' and 'UUU' motifs are optimal, which may not be the case if normalisation is used.

However, it's essential to ensure that a sufficient volume of data is available for optimising parameters not associated with 'AAA' and 'UUU'. For motifs with at least a few non-redundant solved structures available, the sampling problem can be overcome either by increasing the LRMSD threshold for near-natives or by employing data-driven sampling to obtain suitable near-natives for optimisation.

Approximation of Initial Parameter Values Using an All-Atom Force Field

The outcome of the MCSA optimisation suggests the original parameter set could reside within a deep suboptimal region, such that is very complicated to escape from. A fresh parameter set might offer a more promising starting point for optimisation. Furthermore, the creation of a new set has the potential to eliminate the inter-parameter dependencies within the original set, created as a consequence of the ordered optimisation of the initial parameters.

An OPLS all-atom force field, specifically given by OPLS binding energies, can be leveraged to derive a new set of coarse-grained docking parameters. The idea is to fit the ATTRACT parameters to the OPLS binding energies. First, equation (1) is to be rewritten in a functional form suitable for the grid-accelerated computation of ATTRACT scores [260] as follows:

$$\sum_{i,j}(\alpha_{ij}R_{ij}^{-8} - \beta_{ij}R_{ij}^{-6}) = U_{ij}^{attr}(r), \tag{5}$$

where i, j represent the protein and RNA bead type respectively, R_{ij} is the sum of the distances between all (i, j) pairs of beads in a particular complex, and:

$$\begin{aligned} \alpha_{ij} &= \varepsilon_{ij}^{8}\sigma_{ij}, \\ \beta_{ij} &= \varepsilon_{ij}^{6}\sigma_{ij}, \end{aligned} \tag{6}$$

where ε and σ are the docking parameters.

Next, set equation (5) equal to the OPLS binding energy for the corresponding set of atoms:

$$\sum_{i,j}(\alpha_{ij}R_{ij}^{-8} - \beta_{ij}R_{ij}^{-6}) = E_{OPLS}, \tag{7}$$

where E_{OPLS} is the all-atom binding energy of a complex.

A benchmark of n complexes, or protein-fragment cases, allows to build the next system of linear equations:

$$\begin{cases} \sum_{i,j} \left[\alpha_{ij} R_{AB}^{-8} - \beta_{ij} R_{AB}^{-6}\right] = E_{OPLS}^{complex\ 1} \\ \sum_{i,j} \left[\alpha_{ij} R_{AB}^{-8} - \beta_{ij} R_{AB}^{-6}\right] = E_{OPLS}^{complex\ 2} \\ \dots \\ \sum_{i,j} \left[\alpha_{ij} R_{AB}^{-8} - \beta_{ij} R_{AB}^{-6}\right] = E_{OPLS}^{complex\ n} \end{cases} \quad (8)$$

where α_{ij} and β_{ij} are the set of unknown variables. The values of the new initial parameter set ε_{ij} and σ_{ij} can be easily calculated from the values of α_{ij} and β_{ij}.

According to the Rouche-Fontene theorem (also known as Rouche-Capelli and Kronecker-Capelli theorem), a linear system has a unique solution only if the rank of its coefficient A is equal to the rank of its augmented matrix [A|b]. Thus, at least 1054 non-redundant complexes are required for the system to have a unique solution. However, an approximate solution can be found using numerical methods.

The usage of a whole complex to define each line in the system would likely lead to an underdetermined system. However, splitting the complexes into separate protein-fragment cases may provide a system with enough lines for it to have an analytical or numerical solution.

If the system (8) has a solution, it can be treated as the initial parameter set. Such a set will reflect the physical properties of molecular interactions, and the parameters can be subjected to further optimisation. Due to time constraints, this project was not carried on, and the focus was shifted toward the creation of a novel knowledge-based scoring function (detailed in Chapter 4).

3.4 Fine-Tuning Tryptophan-Cytosine Parameters

3.4.1 Problem Statement

In ATTRACT docking, poses are ranked by increasing scores: negative scores are indicative of favourable interactions and positive scores are indicative of unfavourable interactions, e.g. clashing. Consequently, native structures are anticipated to be assigned negative scores. This expectation holds true at the fragment level, as well as at the level of individual pairwise contacts. However, for certain pairs of bead types, the distances observed in native structures result in positive scores. I have identified 3 cases:

- Tryptophan (TRP) side chain - C base;
- TRP side chain - U base;
- Phenylalanine (PHE) side chain - G base.

PHE and TRP side chains contain an aromatic ring, so all listed pairs can form pi-pi stacking interactions, which are crucial for protein-ssRNA binding. The contacts themselves could be underrepresented or missing entirely from the benchmark that was used for the optimisation process,

as this was composed of dsRNA. A coarse-grained representation of the side chains of TRP and C, as well as their indexation within the ATTRACT parameters system, are presented in Fig. 3.3.

I focused on the exploration of the TRP-C pair to understand the issue more deeply, assess if it affects the docking of the fragments containing the TRP-C pair, and, if it is the case, formulate possible ways to address it. This exploration consisted of the identification of the structures containing TRP-C pair at a close distance, assessment of the scores given to such native structures, examination of the docking statistics for such structures and, finally, manual fine-tuning of the relevant docking parameters to enhance docking results.

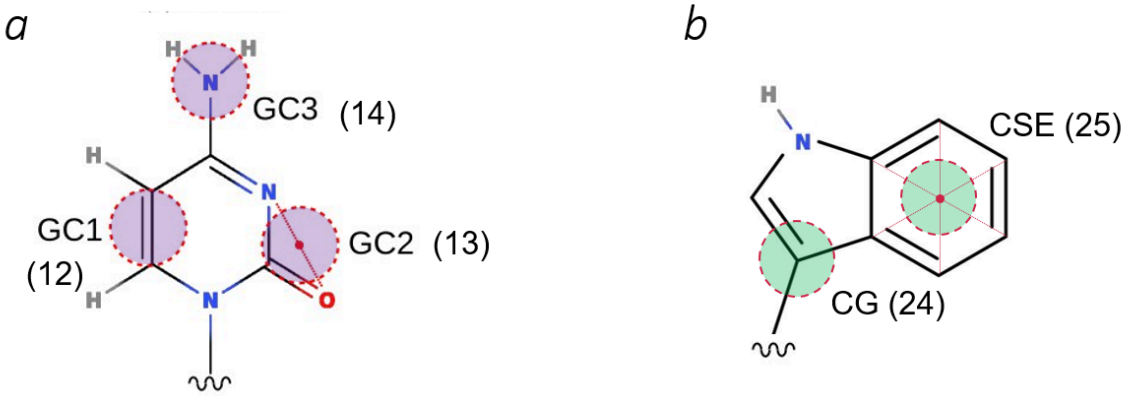


Figure 3.3 - Schematic representation of the beads of interest. The beads of the side chain of C are in purple, and the beads of the side chain of TRP are in green.

3.4.2 Experiments

3.4.2.1 Preliminaries

An example of TRP-C interaction was taken from a native protein-ssRNA complex (PDB 5YTX, protein 65 and RNA 5, YB1 cold-shock domain in complex with CAAC, Fig 3.4). The score for each pair of beads was calculated using ASF with original parameters (Appendix A.2). The results, in a form $S(i, j, r)$ with r the inter-bead distance, are as follows:

$$S(25,\ 12,\ 4.5) = +0.86$$
$$S(25,\ 13,\ 4.2) = +106.62$$
$$S(25,\ 14,\ 3.5) = +3.11$$
$$S(26,\ 12,\ 3.8) = +0.84$$
$$S(26,\ 13,\ 4.1) = -1.45$$
$$S(26,\ 14,\ 4.8) = -1.39$$

Theses numbers indicate that the most unfit parameter pair is the one describing interactions between beads 25 and 13, i.e. $\varepsilon_{25\,13}$ and $\sigma_{25\,13}$.

The docking results of both fragments of 5YTX are given in Tab 3.2. The RNA sequence is CAAC. For fragment 1 the distance from TRP to C is ~12Å, and for fragment 2 it is ~4Å. Additionally, fragment 1 has 13% fewer bead-bead contacts under 8Å compared to fragment 2, which typically is a disadvantage in fragment-based docking.

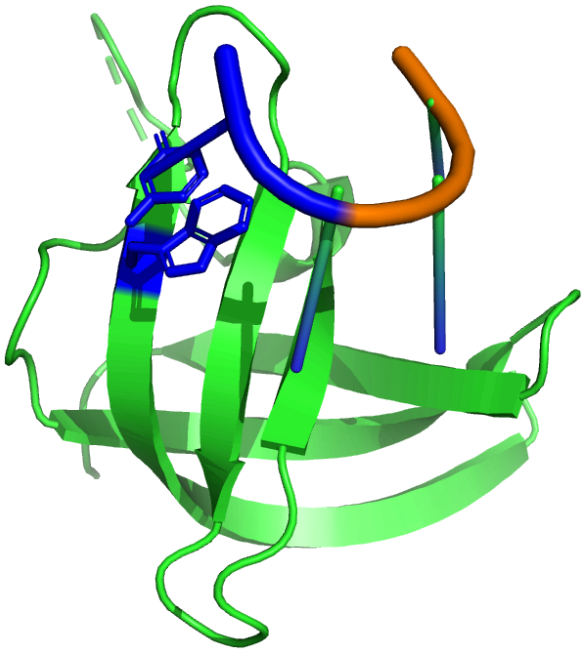


Figure 3.4 - An example of TRP-C interaction (blue), which is, in this case, a pi-pi stacking interaction (pdb code 5YTX: YB1 cold-shock domain in complex with 3'-CAA**C**-5', where the last nucleotide (5') is interacting with TRP).

Fragment 1 exhibits a higher number of sampled near-natives for LRMSD under 2Å and 3Å compared to fragment 2, despite having 7% fewer contacts under 8Å in the native structure. The percentages of top-ranked near-natives for both fragments show minimal differences. The observed dissimilarity in sampling these relatively similar fragments suggests an influence of the distance between TRP-C residues on this process, indicating that close TRP-C contact hinders the sampling.

Table 3.2 - Sampling and scoring results of two fragments of 5YTX complex. The native structure of fragment 2 contains TRP-C interaction.

	LRMSD<2Å		**LRMSD<3Å**		**LRMSD<5Å**		**Native score**	**TRP-C distance**	**TRP-C score**
	all	top20%	all	top20%	all	top20%			
Frag1	8	88%	306	78%	9505	53%	-21	12Å	0
Frag2	0	0%	158	87%	16243	59%	+76	4Å	+97

3.4.2.2 Dataset and Initial Analysis

A benchmark of complexes where a bead pair (25,13) is present and the distance between these beads is under 7Å (empirical threshold) was derived from our high-resolution experimentally solved protein-ssRNA benchmark (see Chapter 4). This TRP-C containing benchmark comprises 15 complexes, released before February 2021. These complexes were manually examined to determine if the (25,13) interaction resembles pi-pi stacking (Tab 3.4), to further assess if the presence of the stacking influences the docking performance. Stacking interactions were found in 9 complexes.

Initial analysis

The scores for the TRP-C pair were calculated for each structure, as well as the score of the native fragment containing the target C (nucleotide C which is 7Å or closer to TRP residue) (Tab 3.3). For all 15 structures, the target pair of residues has a positive score. For 4 structures, the fragment containing target C has negative scores, despite the positive score of the TRP-C pair (in green). For 4 structures, the fragment score is abnormally high, but not due to the TRP-C pair (in grey). For the remaining 6 structures, the fragment score is high due to the score of the target pair. Interestingly, for all these 6 structures, and only for those, the target pair forms a stacking interaction.

Table 3.3 - Distances between beads (25, 13) in the benchmark, and corresponding docking results.
The scores of a TRP-C pair of residues are indicated as 'TRP-C score'. the score of a native fragment containing C (in the middle of the fragment, if possible, on the edge otherwise) is indicated as "Native fragment score".

Complex, PDB_id	Distance, Å	Stacking	TRP-C score	Native fragment score	LRMSD<3Å		LRMSD<5Å	
					all	top20%	all	top20%
1F7U	4.68	yes	31.81	37.86	0	-	25	12%
3ADC	4.07	yes	137.33	132.53	0	-	138	17%
5YTS	4.45	yes	61.43	56.21	260	92%	25661	70%
5YTV	4.45	yes	61.21	41.23	-	-	-	-
5YTX	4.23	yes	97.14	75.92	158	87%	16243	59%
6A6J	4.58	yes	45.23	36.02	52	75%	4181	74%
2CSX	6.63	yes	3.28	-14.19	2	0%	43	9%
2CT8	6.59	yes	3.17	-14.71	1	0%	29	3%
4Z0C	5.51	yes	4.25	56.47	0	-	77	1%
2HGH	5.46	no	5.15	49.89	7	0%	1506	12%
3TS2	6.91	no	0.95	72.43	0	-	25	76%
6KTC	5.93	no	0.32	198.49	51	88%	4697	80%
6KUG	5.79	no	1.14	274.16	58	91%	4582	86%
2FMT	6.25; 5.71	no	2.49	-6.77	176	17%	5036	16%
6SQN	5.89	no	0.57	-11.64	22	100%	428	81%

Over 100 near-natives with LRMSD<3A were sampled for 4/15 cases. If the threshold is relaxed to 5A, the number of cases increases to 10/15. I found a positive correlation (Pearson correlation, $r = 0.39$) between the TRP-C score and a number of sampled poses with LRMSD<5Å. For this thresholds, in approximately ½ of the cases, the percentage of near-natives in the top20% is low (under 20%), while on a general benchmark (1640 fragments without target TRP-C pair and with at

least 1 near-native sampled) this occurs only for ¼ of the cases, which indicates that there is some room for improvement.

Interestingly, for 4 structures in the TRP-C benchmark, both sampling and scoring are very successful, despite the unfavourable score of the native fragment. This could indicate some kind of compensation, i.e., a very low score assigned to some adjacent pair of residues.

3.4.2.3 Protocol

In an attempt to improve the scoring of the target fragments, i.e. fragments containing a target nucleotide, as well as to examine how the change of a small subset of the docking parameters affects the scoring, TRP-C parameters were tuned manually. The following protocol was implemented:

- The TRP-C parameter values were adjusted empirically, as shown in Tab 3.4; the displayed value $\varepsilon_{25\,13}$ was chosen to move the point, where $U_{25\,13} = 0$, closer to the beginning of the coordinates, i.e. shorten the clash distance, and $\varepsilon_{25\,13}$ was chosen to move the point $U^{m}_{25\,13}$ down the y-axis, i.e. assign lower energy to given pair of beads (Fig. 3.5);
- The native poses of the target fragments were scored using the updated parameter set, with an expectation of obtaining lower scores compared to those given by the original ASF parameters;
- The target fragments were subjected to re-docking using the updated parameter set;
- A comparison was conducted between the number of the near-native poses sampled using the updated parameter set and the original set.

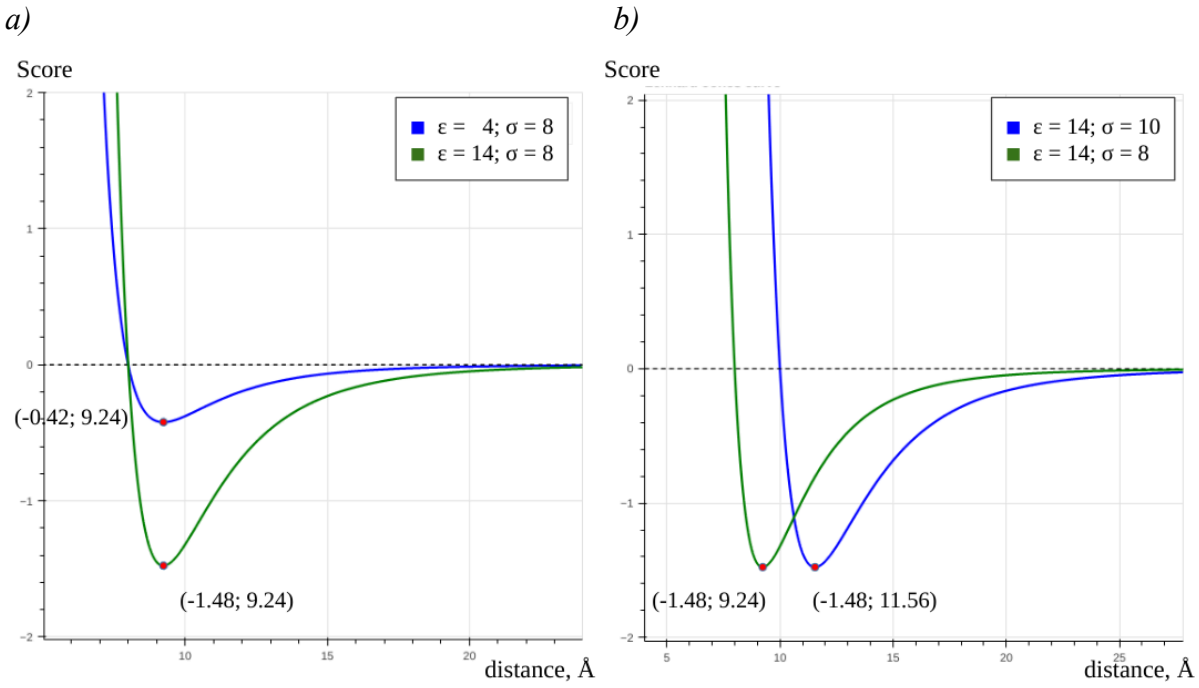


Figure 3.5 - Influence of changes of the values (ε, σ) on the ASF. The curve given by the values (14; 8) is shown in green. The curve given by updated parameter values is in blue. The point (r^{m}; U^{m}) for each curve is shown in red and its coordinates are displayed on the graph.

Table 3.4 - The principle of updating target parameter values.

	ε	σ
Par1	$\varepsilon_{25\,13} = 3.72$, the other such that $U_{ij}^{m} = avg(U^{m})$ over parameters set excluding target parameters	$\sigma_{25\,13} = 15.00$, the other such that $r_{ij}^{m} = avg(r^{m})$ over parameters set excluding target parameters
Par2	$\varepsilon_{25\,13} = 3.72$; the other unchanged	$\sigma_{25\,13} = 15.00$, the other unchanged
Par3	Same as PHE-C	Same as PHE-C
Par4	Such that the curve is flat	Such that the curve is flat

3.4.3 Results and Discussion

The application of the updated parameter set Par1 has shown a deterioration in the performance compared to the original set - the number of sampled near-natives diminished. Par2 and Par3 performed similarly but did not surpass the performance of the ASF parameters. Finally, Par4, "the flat parameters", were tested to see the consequences of complete removal of the TRP-C interactions from the docking process, and yet again, the results were not much different from the ASF (Appendix A.2). Beyond this point, the experiments were terminated, as the re-docking procedure is computationally expensive and it is clear that a simple manual tuning is not effective.

These observations, as well as the initial analysis of assigned scores vs sampling of the benchmark of 15 structures with TRP-C, suggest that the original parameter set could compensate for unfavourable TRP-C scores by assigning highly favourable scores to the other pairs of residues. Thus there is a possibility that it is not feasible to update a part of the parameter set in isolation and that in order to improve the scores for TRP-C, the whole set should be re-optimised. However, this hypothesis needs to be tested. It is possible to perform such a test for a subset of the target parameters using a systematic approach and a coarse grid, as described below.

Systematic change of a subset of the target parameters

Let us estimate the computational time required to evaluate the performance of parameter sets that differ in a subset of the parameters (ε_{xy}; σ_{xy}), where $x \in \{25, 26\}$, $y \in \{12, 13, 14\}$, so 6 pairs of (ε; σ) in total. Let us also assume that the values of these (ε_{xy}; σ_{xy}) should remain discrete and lay within [min; max] boundaries of the original parameter set (including all 1054 parameters). However, the value $max(\sigma) = 6.4$ appears to be too high, as many of the native poses have contact with a distance under 6.4Å, and the $min(\sigma) = 2$ appears to be too small as the minimum distance of those contacts is 4.07Å (Tab 3.4). Thus, let's consider the new range for sigmas $\sigma \in [3; 5]$ and keep the existing range for epsilons, $\varepsilon \in [0.02; 20]$. Finally, to obtain new (ε_{xy}; σ_{xy}) values, let's discretise σ-range with step size 1, and ε-range with step size 5. This gives 3 distinct values for σ_{xy} and 4 distinct values for ε_{xy}. In this case, the total number of possible (ε_{xy}; σ_{xy}) tuples is equal to 12, and subsequently, the number of possible parameter sets is equal to $12^{6} = 2,985,984$ distinct sets.

Straightforward scoring of a single pose using 1 CPU takes in the order of 1 second. CAPIDS's cluster consists of 300 CPUs. Assuming the test set consists of 10 fragments, scoring the native poses will take an absolute maximum of:

$$\frac{num\_parameter\_sets \cdot num\_fragments \cdot time\_per\_tack}{number\_of\_CPUs} = \frac{2{,}985{,}984 \cdot 10 \cdot 1}{300} = 99532.8\,sec \approx 27.7\,hours$$

It is likely that some sets will assign a positive score to the native poses. Such sets are to be removed. Since the estimation of this number without additional computational experiments is not possible, let's carry on the calculations over all possible sets. Scoring a pool of 10 million poses (1 CPU) takes in the order of 1 hour. Thus, rescoring 10 pools of poses with all sets will take $\frac{2{,}985{,}984 \cdot 10 \cdot 60}{300} = 5{,}971{,}968\,min \approx 11.4\,years$, which is positively unfeasible. However, the bead-bead distances can be precomputed, as well as the scores for all other pairs of bead types, which could accelerate computations significantly.

Alternatively, one could also keep only x sets that give the best average score for the native poses. The value of x can be adjusted to the amount of time dedicated to this experiment. During 2 months of computations, one could test 43800 sets (~1.47% of all sets, possible under given conditions); during 3 months - 65700 sets (2.2%).

Another alternative is to test only 4 bead pairs out of 6, keeping the pairs with indexes $(26,\ 13)$ and $(26,\ 14)$ unchanged. Such choices lead to $12^4 = 20{,}736$ distinct sets, which can be tested (applied to score 10 pools of 10 million poses) in $\frac{20{,}736 \cdot 10 \cdot 60}{300} = 41{,}472\,min \approx 5.5\,months$. This is still a hefty chunk of time, thus, it is appealing to streamline the process and test 50% of sets that give the best average score for the native poses or develop a faster way to score poses. There is a risk of facing the memory issues, which may force one to limit the percentage of sets to undergo testing even further.

After testing the sets via rescoring, one can select several of the most promising sets to test their sampling performance via re-docking. Docking with 30 million starting points and 30 CPUs takes at least 6 hours depending on the size of the protein, so only a handful of the sets can be tested via re-docking. Generally speaking, it is possible to apply positional restraints to reduce the search space and lower the number of starting positions. This will lower the time of each docking run significantly, allowing to test more sets via sampling. For example, lowering the number of starting positions from 30 million to 100, which is acceptable for data-driven docking, decreases execution time from 6 hours to 10+/-5 minutes (on 30 CPU).

Stacking Problem

We define a stacking problem as an inadequate representation of the stacking interactions within the ATTRACT force field. Out of 6 non-redundant fragment TRP-C containing structures, only one had a negative native score (Tab 3.5). Among these 6 cases, the TRP-C pair within ~4.5Å causes such a high score in 4 cases (1F7U, 5YT*, 6A6J, 3ADC). For the last case (4Z0C) the main reason is the proximity of the GLU-C pair (C bead GC1_12 is 3.88Å from GLU bead CB_10 giving a score of +30, as $\sigma_{10\,12} = 5.1$). Stacking interaction plays an important role in RNA-protein interactions, and more particularly in the case of RRMs, thus addressing these high scores and possibly incorporating explicit score rewards for stacking could benefit the docking.

Explicit score rewards for stacking require the identification of both stacking interactions in the pose and an appropriate score value for the reward. Given that the ATTRACT coarse-grained model has three to four per RNA base, determining the spatial orientation of the aromatic ring is feasible.

However, amino acid rings only contain two beads, limiting the orientation determination. Regardless of this limitation, one may add a score reward for the poses in which the distances between RNA beads to protein beads are in a certain range and are similar under a certain threshold. As an alternative approach for a more accurate stacking identification, one could either introduce a new ghost bead for amino acids (such that would only affect the pose's score upon the detection of a stacking interaction) or transform poses with potential stacking into an all-atom representation. The value of the score reward could be determined empirically, or using binding energies as reference values (see the following section).

Table - 3.5 Structures with stacking.
For the redundant structures (5YTX, 5YTS and 2CSX, 2CT8) the avg values are given.

Complex PDB_id	Distance, Å	TRP-C score	Native fragment score	LRMSD<3Å		LRMSD<5Å	
				all	top20%	all	top20%
1F7U	4.68	31.81	37.86	0	-	25	12%
5YTX/S	4.37	73.26	57.78	209	89%	20779	65%
6A6J	4.58	45.23	36.02	52	75%	4181	74%
3ADC	4.07	137.33	132.53	0	-	138	17%
2CSX/T8	6.61	3.23	-14.45	1	0%	37	6%
4Z0C	5.51	4.25	56.47	0	-	77	1%

For testing each possible solution, a larger and more diverse benchmark of structures with and without stacking is required (beyond TRP-C targets). Resources like InteR3M or PISA-lite offer a means to identify stacking interactions without manual assessment.

Measuring binding energies associated with stacking interactions

One may use the measurement of binding energies to estimate a score reward assigned to the residues in the stacking orientation. A project for the measurement of such energies was initiated within the RNAct project. Collaborating with a group of both computational and experimental PhD students (Joel Roca Martínez, Hrishikesh Dhondge, Niki Messini, Rosa Anahí Higuera) our objective was to determine the energies of the interaction between the sex-lethal protein and a poly-U ssRNA, for which the PDB X-ray structure 1B7F of the complex is available.

The concept was to first measure values for the wild type, then introduce amino acid mutations that would change the stacking to another type of interaction, while preserving binding, followed by measuring new affinity values. This approach in theory allows us to estimate the difference in the affinities associated with stacking versus non-stacking interactions.

The candidates for mutations were identified (I128Y and D172R). The essential computational checks were completed to ensure that the mutations would not change the 3D structure of the protein, by Joel Roca Martínez. Regrettably, the experimental aspect encountered delays beyond our expectations, ultimately leading to the premature termination of this project.

3.6 Conclusion

In this chapter, we have delved into the specifics regarding the coarse-grained representation used in ATTRACT and the original set of protein-ssRNA docking parameters of the ASF, which in turn is used simultaneously for sampling and scoring. As ssRNA'TTRACT suffers from the scoring problem due to the parameters of ASF, we have applied the MCSA optimisation protocol to address this problem. Despite extensive testing involving multiple hyperparameter sets, none of the resulting parameter sets has outperformed the original one. This has prompted us to explore potential directions for further improvement of the ASF parameters. These directions have included the generation of a new initial parameter set based on binding energies derived from the OPLS all-atom force field and the development of three distinct parameter sets - two for the most common RNA motifs, 'AAA' and 'UUU', and one for all remaining motifs. More importantly, our findings have led us to a hypothesis that a single parameter set may be insufficient for scoring protein-ssRNA fragment-based poses. This idea will be explored in the following chapter.

In the second part of this chapter, we have focused on the analysis of a small subset of parameters known to assign unfavourable scores to native poses, using TRP-C parameters as an example. We have established a small benchmark of experimental structures with pertinent interactions. Notably, we have discovered that all native TRP-C geometries receive unfavourable scores. Intriguingly, despite this, both sampling and scoring are relatively successful in approximately half of the protein-fragment cases. We have conducted several attempts to empirically refine the target parameter values to enhance sampling and investigate the impact of these updates on the overall ASF performance. The results, which have revealed performance highly similar to ASF despite the use of different target parameter values, have raised questions about the feasibility of optimising the parameter subset in isolation. This feasibility may be tested via a systematic approach, proposed in this chapter.

In the next chapter, we will outline the development of a novel knowledge-based scoring function designed to replace ASF in the ssRNA'TTRACT method, intending to address the scoring problem.

Chapter 4: HIPPO

HIstogram-based Pseudo-POtential for the scoring of protein-ssRNA fragment-based docking poses

4.1 Aims

In the previous chapter, we discussed several attempts to enhance ssRNA'TTRACT performance and address the scoring problem by optimising parameters of ASF. The results have been unsuccessful and have led us to the conclusion that more than one parameter set is required for more accurate sampling, as protein-ssRNA binding modes are very diverse.

In this chapter, we introduce HIstogram-based Pseudo-Potentials (HIPPO), a novel scoring potential designed for protein-ssRNA fragment-based docking poses. HIPPO is based on the analysis of relative frequencies of bead-bead distances in near-native and non-native poses. While the comprehensive protocol and the results of the application are provided in [263 or Appendix B.1], this chapter offers a concise summary of the preliminary experiments that have led to HIPPO's development and primarily focuses on the outcomes of applying HIPPO to new protein-ssRNA complexes, which were neither part of the training nor testing of the scoring potential.

4.2 Preliminaries to the Histogram-Based Approach

4.2.1 Method

To achieve our goal of enhancing the protein-ssRNA scoring, we developed a novel histogram-based approach, which iteratively adjusts the parameters of the Lennard-Jones energy curve to the frequency of occurrences of bead-bead pairs at certain distances in the near-native poses vs non-native poses in ATTRACT coarse-grained representation.

For each ij-pair of interacting bead types (i is for the bead of the protein, j is for the bead of the RNA), we:

- convert the current energy function $E_{ij}^{attr}(r) = \varepsilon_{ij}\left(\frac{\sigma_{ij}^{8}}{r^{8}} - \frac{\sigma_{ij}^{6}}{r^{6}}\right)$ into a **log-odds histogram** (Fig 4.1, a) of the expected occurrences of bead-bead distances (discretized into bins r) in native/non-native poses, using the Boltzmann equation $log(P(r_{ij})) = -kTE_{ij}$, where k is the Boltzmann constant, T is absolute temperature, r_{ij} is the distance between bead in the ij-pair, $E_{ij}(r)$ is the energy approximation for the ij-pair given by the ASF;
- obtain the corresponding histogram on observed occurrences of bead-bead distances in native/non-native poses by counting the number of the corresponding ij-pair within each distance bin, over the training set of the docking poses. This histogram corresponds to the residual error of the energy function and is titled **residual histogram** (Fig 4.1, b). In a residual histogram, the bars in quadrant I signify the enrichment in near-natives, while the bars in quadrant IV signify the depletion in near-natives.
- sum up the log-odds and residual histograms (Fig 4.1, c) and fit the energy parameters to the **resulting histograms** (Fig 4.1, d). At the end of this process, a new pair of docking parameters (ε'_{ij}, σ'_{ij}) is obtained for the current ij-pair of bead types.

When the procedure is finished for all pairs of bead types, the benchmark is re-docked with the new parameter set, and a new iteration of the procedure begins. This procedure is repeated until convergence - until the residual histogram is flat. After convergence, this procedure should generate equal distributions of bead-bead distances in near-native poses and non-native poses, which are thus indistinguishable by bead-bead distance criteria. The number of near-native poses will then be completely optimized based on the bead-bead distances.

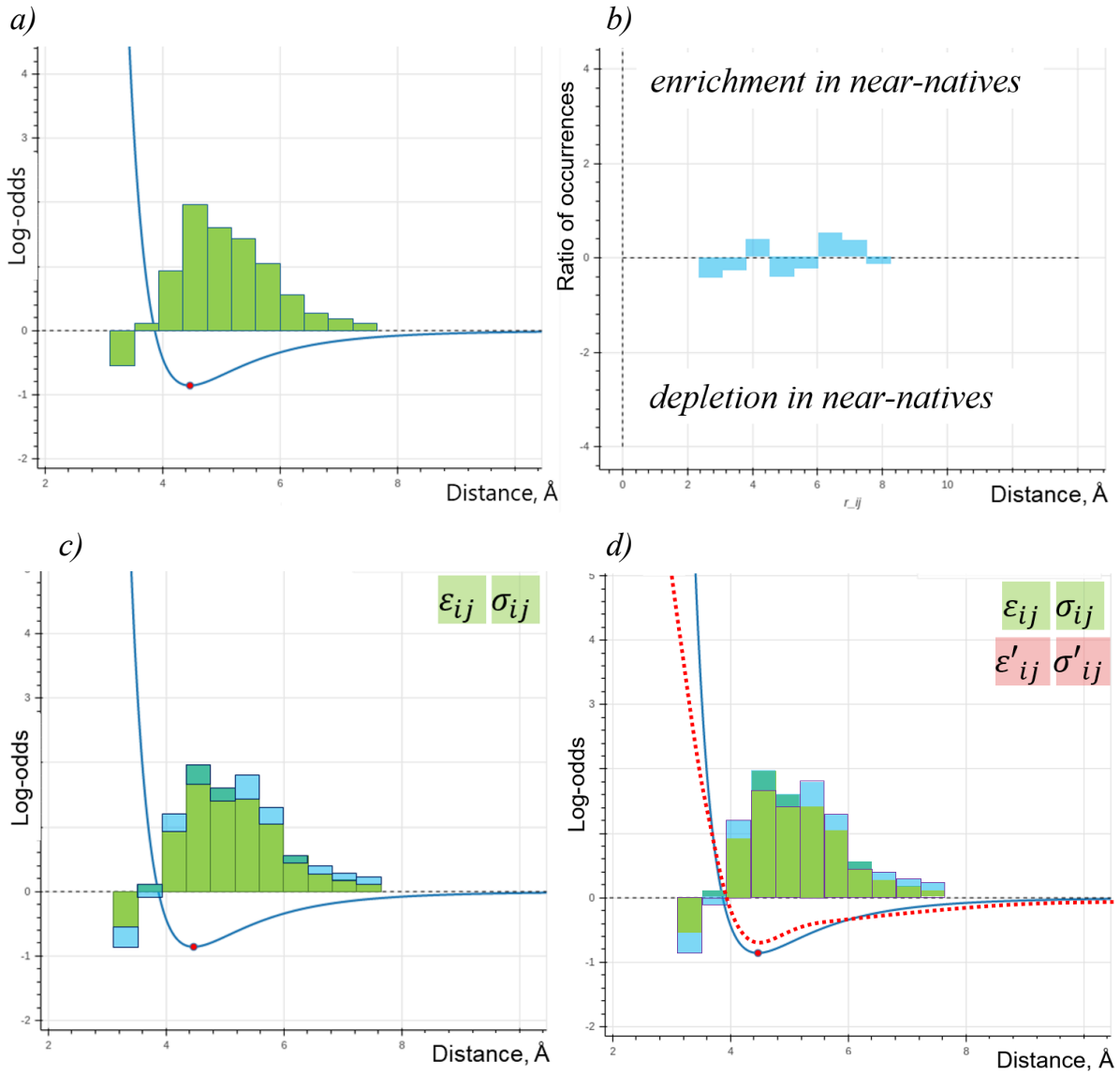


Figure 4.1 - A graphical example of the a) Conversion of the energy curve, displayed in dark blue, into a log-odds histogram, displayed in green. Each histogram bar is defined for the respective distance bin; b) Residual histogram; c) Summation of the residual (blue) and log-odds (green) histograms, which results in the so-called resulting histogram; d) Fitting of a curve onto the resulting histogram. The old curve is displayed in dark blue (solid line), and the fitted curve is displayed in red (dotted line). The shape of the curve is a limiting factor.

4.2.2 Results

We executed one iteration of the histogram-based approach on a toy example. The training set consisted of 3 protein-ssRNA complexes (1M5K C_1_92 B_35_44; 1B7F A_1_167 P_3_12 and 1WMQ A_1_B_143 C_1_7), and the test set of 3 different complexes (1VBX A_1_95 B_49_58, 1DRZ A_1_91 B_48_57 and 2ANN A_1_148 B_8_16).

Docking poses for each training and test case were generated using the ATTRACT docking engine following the process described in [263]. The threshold for near-natives was set at 3Å and for non-natives at 5Å. Then, all sampled poses in the training set were pooled together and one iteration

of the histogram-based approach was executed resulting in a set of histograms for each pair of bead types. As a first test, this set of histograms was used to score the docking poses of the test cases.

The scoring of a docking pose with a histogram set is done by retrieving a histogram-based score $h(i,j,r)$ for each ij-pair of beads located at a distance r, and summing up all these scores:

$$Score_{histogram-based} = \sum_{i,j,r} h(i,j,r) \quad (1)$$

Ranking the docking poses by a histogram set appears to be promisingly more accurate compared to the ranking by the ASF (Fig 4.2).

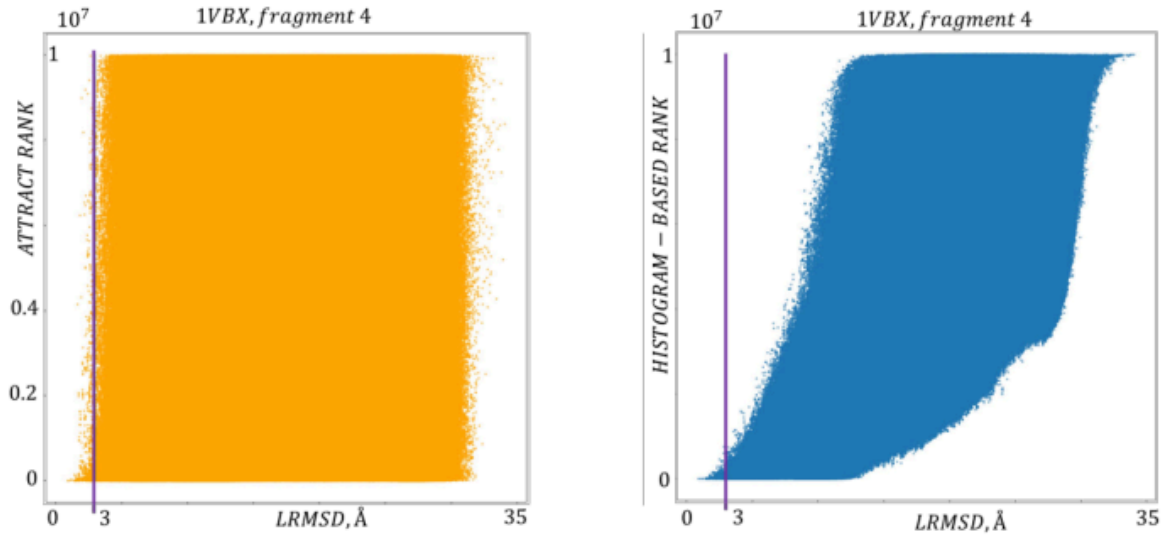


Figure 4.2 - An example of the rankings of the docking poses generated for the test case 1VBX_GCA. LRMSD is shown on the x-axis, and rank on the y-axis is given by the ASF (orange) or by the histogram (blue). The purple vertical line (x = 3) separates the near-native poses from the others (non-natives and intermediate).

However, the process of fitting the energy curves to the corresponding histograms produced notably poorer results compared to the ASF, which suggests that the Lennard-Jones curve is not the most efficient shape for fragment-based scoring. Therefore, we shifted our focus to developing a histogram-based scoring approach without the fitting step, eventually creating the HIPPO, as briefly outlined in the following section and detaild in the original paper. In the realm of the knowledge-based scoring functions, HIPPO is closest to the DECK-RP [117], as it is also based on the docking poses.

4.3 HIPPO protocol

As mentioned before, HIPPO is based on the analysis of the relative frequencies of bead-bead distances in near-native and non-native poses. HIPPO is a composite function, consisting of four distinct scoring potentials. Here, we provide intuition behind the process of deriving a single scoring potential $\mathcal{H}$, followed by the protocol to derive HIPPO (four scoring potentials) (Fig 4.3).

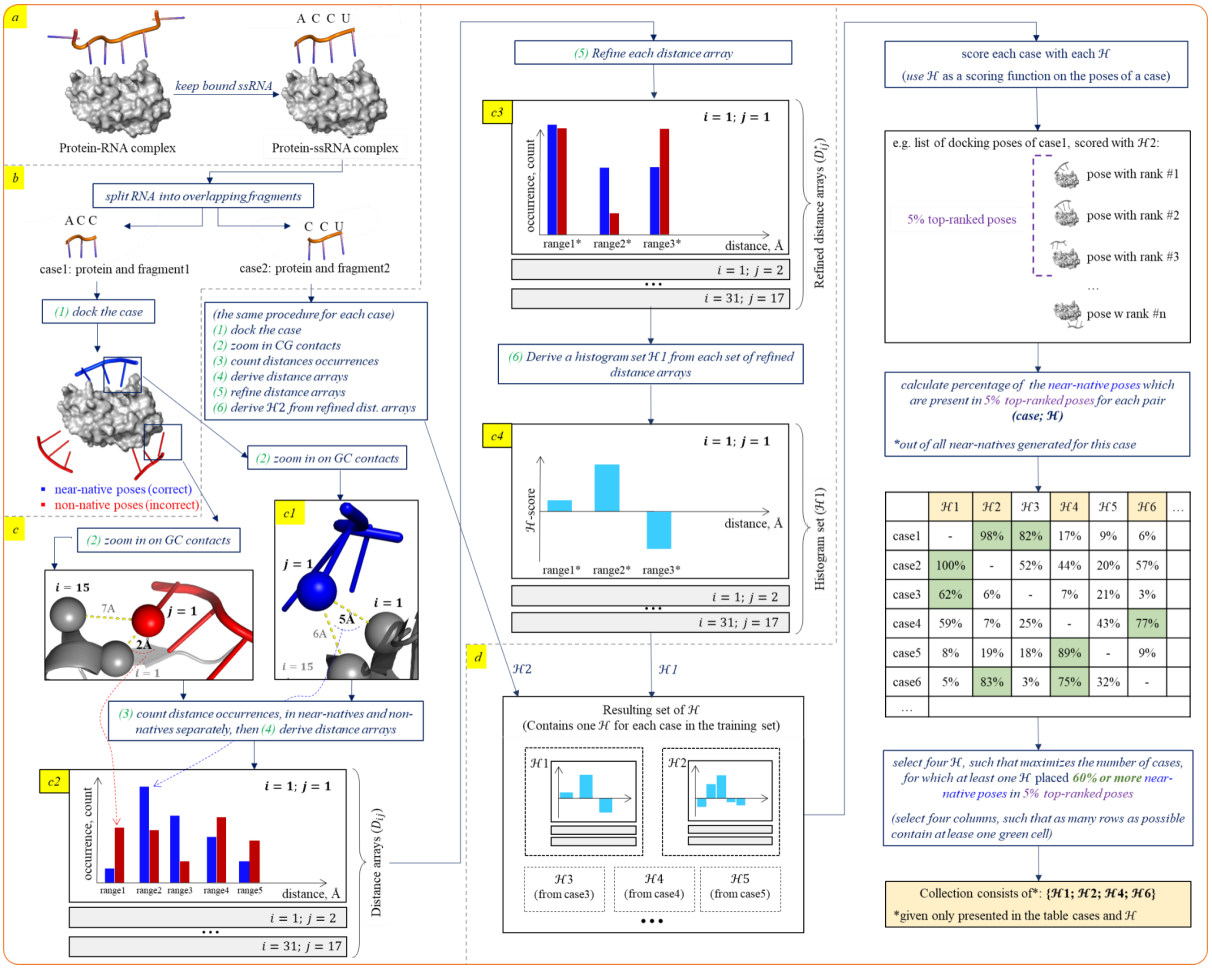


Figure 4.3 - Graphical pipeline for building HIPPO as a collection of four histogram sets ($\mathcal{H}$). **a)** Transition from a protein-RNA complex to a protein-ssRNA complex with ssRNA that is at least 3 protein-bound nucleotides long. This step was achieved using ProtNAff. **b)** Creation of a pool of labelled docking pose using ATTRACT. Each protein-fragment case of protein-ssRNA complex is docked and each docking pose is labelled as near-native or non-native. **c)** Construction of the distance arrays, refinement of the distance arrays and derivation of the histogram set $\mathcal{H}$ from refined distance arrays. The frequency of occurrences of individual bead-bead distances within a single pool of docking poses are captured within distance arrays, one array per each pair of bead types. **c1)** Close-up of contacts between RNA bead j=1 and protein bead $i = 1$=1 and $j = 15$. **c2)** An intuitive schema of the distance array for the pair of bead types $(i = 1; j = 1)$ is shown as an expanded plot. The distance ranges are shown on the x-axis, and the numbers of occurrences of the distances are shown on the y-axis. For each distance range, the number of occurrences for the near-native poses is displayed as a blue bar, and for the non-natives as a red bar. The blue dashed line from c1 to c2 shows the contribution of the contact to the near-native distance array, range2. The other distance arrays (for other pairs of bead types) are not shown (collapsed). **c3)** An intuitive schema of the refined distance array for the pair of bead types $(i = 1; j = 1)$ is shown as an expanded plot. Due to the relatively low number of near-native contacts in range1, it is merged with $range2$, forming a new $range1*$. The following $range3$, which contains a sufficient number of near-native contacts, remains unchanged and is renamed as $range2*$ to preserve the range order. Finally, $range4$, which also contains an insufficient number of near-native contacts, is merged with $range5$ to form a new $range3*$. **c4)** An intuitive schema of the histogram set $\mathcal{H}$, derived from the refined distances arrays, which are in turn built from the pool of the docking poses of the case1. A histogram for the pair of bead types $(i = 1; j = 1)$ is shown as an expanded plot, other histograms are collapsed. **d)** Schematic pipeline of the partitioning algorithm, employed to derive a collection of four histogram sets out of all sets.

A single scoring potential $\mathcal{H}$ is derived from a pool of docking poses of a single data case (protein-bound ssRNA fragment and corresponding protein). First, the frequencies of occurrence of each bead-bead distance, discretised into distance ranges, are counted separately for near-native and non-native poses (Fig 4.3, c). This is done individually for each pair of bead types, leading to a set of the distance arrays (i.e. contact frequency arrays), as many as pairs of beads in the pool of poses, typically around 500 arrays. As a reminder, there are 17 RNA bead types and 31 protein bead types in ATTRACT coarse grained representation.

Each distance array is converted into a histogram, capturing the propensity of each contact distance to occur in the near-native vs non-native poses (Fig 4.3, c4). If a certain contact is overrepresented in the near-native poses (and correspondingly underrepresented in the non-natives), this contact is rewarded with a positive score. Otherwise, if a certain contact is overrepresented in the non-native poses (and underrepresented in the near-natives), this contact is penalised with a negative score. The absolute value of these scores is determined based on the fraction of the given distance among all observed distances for this pair of bead types.

Upon converting all distance arrays into histograms, a single scoring potential $\mathcal{H}$ is obtained, with each histogram corresponding to each pair of bead types present in the pool of poses.

Applying $\mathcal{H}$ to a data case involves scoring and ranking its docking poses. The process assigns a histogram-based score to each pair of beads within a pose, aggregating these scores. The individual score for a pair of beads is determined by the height of the bin corresponding to the distance between the beads. This bin should be taken from the histogram with corresponding bead types.

HIPPO is derived using all data cases of the training set, with subsequent application to the data cases of the test set to assess HIPPO's performance. This protocol consists of the following steps (Fig 4.3, d):

1. Derive scoring potential $\mathcal{H}$ from each training data case;
2. Apply each $\mathcal{H}$ to all training data case, excluding the case from which $\mathcal{H}$ was derived. Calculate percentage of the near-native poses (out of all sampled near-natives) within top5% (5% of top ranked poses);
3. Label each pair (data case, $\mathcal{H}$) as successful if the percentage of the near-native poses in the top5% is 60% or higher;
4. Select four scoring potentials $\mathcal{H}$ that maximise the number of cases for which at least one pair (data case, $\mathcal{H}$) is labelled as successful.

Choosing four as the number of scoring potentials was determined through testing. Fewer potentials resulted in fewer successful pairs, while more potentials did not increase success rate.

To apply HIPPO to a data case, its docking poses are scored with each $\mathcal{H}$ separately. Approximately the top 5% of poses from each ranking are merged to achieve a list corresponding to the top20%. If redundant poses are detected, only the top-ranked is retained, and additional poses are added until the top 20% is reached.

While working on the described protocol, we experimented with additional ideas, but unfortunately, the results were unsatisfactory. These ideas are neither described in the paper nor this chapter but can be found in Appendix B.2 Several potential approaches for tuning HIPPO are described there as well (Appendix B.3).

4.4 Application to New Complexes

The final HIPPO consists of 4 distinct potentials, derived from the following protein-trinucleotide cases: 1M5K-GCA (protein: C_1_92; trinucleotide: B_38_40), 5MPG-UAG (A_1_97; B_2_4),

4N0T-AGA (A_1_363; B_20_22), 6DCL-UUA (A_1_B_171; C_7_9). For brevity, we refer to these potentials as $\mathcal{H}_{\mathrm{I}}$, $\mathcal{H}_{\mathrm{II}}$, $\mathcal{H}_{\mathrm{III}}$, and $\mathcal{H}_{\mathrm{IV}}$, respectively, throughout the text. To assess the generalisability of HIPPO, we applied it to a benchmark of protein-ssRNA complexes not used during the creation and testing of HIPPO.

Furthermore, we investigate the applicability of the best-performing potentials (BP) for incremental docking. In this scenario, a relatively low number of top-ranked poses is retained for one fragment, and a higher number of top-ranked poses is kept for adjacent fragments. The retained poses are assembled into models of the full ssRNA chain. The concept of BP involves using a single, best-performing potential out of $\mathcal{H}_{\mathrm{I}}$, $\mathcal{H}_{\mathrm{II}}$, $\mathcal{H}_{\mathrm{III}}$, and $\mathcal{H}_{\mathrm{IV}}$ for each protein-trinucleotide case. This approach, although currently only possible in a test case, eliminates false positive poses, given by the less suitable potentials, and typically results in a greater number of near-native poses among the top-ranked ones.

To evaluate the suitability of the BP for incremental docking, we assembled the best-docked fragment with the adjacent fragments on each side. The results of both experiments, scoring and assembly, are compared with the performance of the ASF below.

4.4.1 Scoring

4.4.1.1 Data

The benchmark consists of the experimentally solved protein-ssRNA structures that (i) are solved with NMR or X-RAY with resolution 3Å or higher, and (ii) contain a 5-nucleotide or longer ssRNA sub-chain, bound to the protein (i.e. at least 5 pairs of protein-RNA heavy atoms were located within 6Å from each other). It can be separated into two subsets:

- Subset ‘newRRM’ consists of 6 RRM-ssRNA complexes that were deposited to PDB after the date the benchmark for HIPPO was collected (after February 2021). This set consists of 29 distinct data cases (1 case is 1 protein-trinucleotide structure);
- Subset ‘nonRRM’ consists of 150 protein-ssRNA complexes (519 cases). All proteins in this set do not contain an RRM domain, which was verified using InteR3Mdb. These complexes were deposited to PDB before February 2021. All cases of this subset are non-redundant on the bead-bead contact level (see description in the paper, §2.1.4) with each other and with cases used for the training and testing of HIPPO.

In this section, we refer to a subset of all cases that share the same protein as a ‘complex’.

4.4.1.2 Protocol

All complexes were docked using ATTRACT, with the same setting used for HIPPO derivation. For each protein-trinucleotide case, the 10 million docking poses, top-scored by the ASF, were retained. Next, the poses were scored and ranked by each of the 4 potentials comprising HIPPO separately, and then approximately the top5% of each ranking were pooled together, removing redundant poses until the total number of 20% (2,000,000 poses) was reached.

4.4.1.3 Results and Discussion

The scoring is considered successful / very successful if the top20% contains at least 60% / 80% of all sampled near-native poses (LRMSD<5Å). The evaluation revealed that (Tab 4.1):

- On ‘newRRM’, HIPPO’s success rate (over all 29 cases) is 55%, while it is only 14% for the ASF. HIPPO is very successful for 28% of cases, while the ASF is for none of the cases;
- On ‘nonRRM’, HIPPO’s success rate (over all 519 cases) is 40%, while it is 34% for the ASF. HIPPO is very successful for 28% of cases, and the ASF is very successful for 12% of cases.

On ‘newRRM’, HIPPO ranks on average an additional 15% of the near-natives in the top20% compared to the ASF. For ‘nonRRM’ this number is equal to additional 2% of the near-natives in the top20% (Tab 4.2, columns ‘per case’).

Table 4.1 - Comparison of the ASF, HIPPO and BP success rates (%) over the cases and over the complexes.
The scoring is considered successful / very successful if the top20% contains at least 60% / 80% of all sampled near-native poses.

	newRRM				**nonRRM**			
	Per case		**Best case per complex**		**Per case**		**Best case per complex**	
	Over 60%	Over 80%	Over 60%	Over 80%	Over 60%	Over 80%	Over 60%	Over 80%
ASF	14	0	50	0	34	12	47	20
HIPPO	55	28	100	66	40	28	53	41
BP	90	70	100	100	72	54	85	69

For incremental docking, at least one fragment per complex should be well-docked. In terms of the scoring, this translates into the requirement for at least one fragment per complex to have a high percentage of the near-natives in the top-ranked poses. Thus, we also measured the success rate per complex, i.e. the number of complexes where, for at least one fragment, 60% / 80% of the near-natives were in the top20%. The results are the next (Tab 4.1, columns “best case per complex”):

- On ‘newRRM’, HIPPO’s success rate per complex is 100%, while for the ASF it’s 50%. HIPPO is very successful for 66% of complexes, while the ASF is very successful for none of the complexes;
- On ‘nonRRM’, HIPPO’s success rate per complex is 53%, while for the ASF it’s 47%. HIPPO is very successful for 41% of complexes, while the ASF is very successful for 20% of complexes.

Table 4.2 - Comparison of the average percentages of the near-natives in the top20% ranked poses by the ASF, HIPPO and BP.

	newRRM		**nonRRM**	
	Per case	**Best case per complex**	**Per case**	**Best case per complex**
ASF	38	58	45	55
HIPPO	53	86	47	62
BP	83	99	73	85

For each complex, we calculated the percentage of the near-natives ranked in the top20% for the best-scored fragment (i.e. the maximum value among all fragments). Then we averaged these values over all complexes. HIPPO ranks on average an additional 28% of the near-natives in the top20% compared to the ASF on ‘newRRM’, and additional 7% on ‘nonRRM’ (Tab 4.2, columns ‘best case per complex’).

To sum up, HIPPO outperforms the ASF both on ‘newRRM’ and on ‘nonRRM’. Even better results are achieved if the BP is used. The BP achieved ~25% higher success rate both per case and per complexes compared to HIPPO (Tab. 4.1). For ‘newRRM’, the best potential ranked on average 99% of the near-natives in top20% for the best-ranked fragment per complex. For ‘nonRRM’ this number is 85%. Similarly to the results obtained during the initial cross-validation of HIPPO, each of the 4 potentials takes the role of the best potential for approximately a quarter of the cases: 32%, 27% 23% and 18% on ‘nonRRM’; and 31%, 35%, 10% and 24% on ‘newRRM’ for $\mathcal{H}1$, $\mathcal{H}2$, $\mathcal{H}3$, and $\mathcal{H}4$ respectively.

Although HIPPO demonstrates improved performance compared to the ASF, there is still room for enhancement. For instance, when considering only the top5% of poses, BP, HIPPO and the ASF were successful in 15/519, 7/519, and 1/519 cases for 'nonRRM', respectively. There are two promising directions for further HIPPO development. First, creating a model capable of identifying the best potential for a given case could significantly improve its performance, as discussed in the original paper and now confirmed by the new results. Second, an iterative optimisation of HIPPO, possibly by the histogram-based method, is worth exploring. Preliminary results suggest that converting a scoring potential into an energy curve leads to a significant loss of signal. However, HIPPO could be employed as a sampling function without this conversion, using the MC-based ATTRACT sampling procedure. If HIPPO's sampling performance matches or surpasses that of the ASF, it would be an indication to continue exploring the histogram-based approach, including updating a set of potentials based on docking poses obtained through HIPPO. As it is possible to identify the best potential for each case during training, the histogram-based approach could be applied separately for each potential, along with the corresponding dataset (set of cases for which the given potential is best-performing).

HIPPO demonstrates improved performance on 'nonRRM' complexes, despite being exclusively trained on RRM-ssRNA complexes, suggesting its potential for generalisation. These results further reinforce the prospect of applying the protocol to derive scoring potentials for other types of complexes.

4.4.2 Fragment Assembly

As previously mentioned, to compare the effectiveness of BP ranking against the ASF ranking in incremental docking, we conducted fragment assembly of the best-docked fragment with the adjacent fragments on each side for several complexes. The assembly procedure has been detailed in §2.4.4.1, subsection "Assembly Procedure”. The results are detailed below.

4.4.2.1 Data

From the benchmark described in the previous section, we selected the most promising candidates for the assembly procedure. These are the complexes with 3 adjacent fragments, such that the middle fragment (hot-spot) has at least 460 near-native poses in the top5% and the remaining 2 fragments (side-fragment) have at least 65 near-native poses in the top5% each (all scored with BP). In total, 14 complexes were selected, 2 from ‘newRRM’ and 12 from ‘nonRRM’ (Appendix B.4).

4.4.2.2 Protocol

First, we defined the assembly hyperparameters, which include:

- The overlap in angstroms between the last 2 nucleotides of the *i*-fragment and the first 2 nucleotides of the (*i*+1)-fragment;
- The number of poses of hot-spot used for assembly;
- The number of poses of side-fragments used for assembly (the same for both side-fragments).

We determined suitable hyperparameters for generating at least one near-native chain, i.e. a chain composed of near-native poses exclusively, by conducting the assembly using solely the near-native poses. For each overlap value leading to the generation of at least one near-native chain, we determined the minimum rank of the pose required for each fragment, starting from the hot-spot, to generate near-native chains. By aggregating these values across the entire dataset, we estimated the number of poses needed for hot-spots and side-fragments during assembly, thus completing the entire hyperparameter set. Afterwards, we handpicked several of those hyperparameter sets to perform the assembly of all poses (not restricted to near-natives).

Each complex in the dataset was assembled using both the BP ranking and the ASF ranking, with the same hyperparameter set. We compared the percentages of near-native chains produced among all chains, as presented below.

4.4.2.3 Results and Discussion

The near-native assembly was carried out with the overlap values ranging from 0.5Å to 2.0Å with 0.1Å increment, and 4 distinct hyperparameter sets were determined (Tab 4.3). All-poses assembly was conducted using these hyperparameter sets.

Table 4.3 - Hyperparameters for the assembly.

	Overlap	**Poses per hot-spot**	**Poses per side-fragment**
Hpar1	0.9Å	100k (top1%)	2mil (top20%)
Hpar2	0.9Å	100k (top1%)	500k (top5%)
Hpar3	1.4Å	20k (top0.2%)	500k (top5%)
Hpar4	1.4Å	10k (top0.1%)	1mil (top10%)

All-poses assembly is considered successful if at least 1 near-native chain has been produced. BP consistently outperforms the ASF, giving on average 16% higher success rate over all hyperparameter sets (Tab 4.4). The percentage of near-native chains for each complex for each hyperparameter set can be found in Appendix B.5.

Table 4.4 - Comparison of the all-poses assembly with AFS vs BP.

	Percentage of complexes with at least 1 near-native chain	
	ASF	**BP**
Hpar1	64% (9/14)	79% (11/14)
Hpar2	50% (7/14)	71% (10/14)
Hpar3	79% (11/14)	100% (14/14)
Hpar4	71% (10/14)	79% (11/14)

For all hyperparameters, I found a consistent positive correlation between the number of near-native poses of hot-spot and the percentage of near-native chains (Tab 4.5, 'num _poses_hs to %_nn_chains'). There is also a consistent correlation between the difference in the number of near-native poses of hot-spot used for assembly in the BP vs the ASF ranking and the difference in the percentage of near-native chains (Tab 4.5, Δposes_hs to Δ%_nn_chains).

Table 4.5 - Correlations across each hyperparameter set.

	num _poses_hs to %_nn_chains		**Δposes_hs to Δ%_nn_chains**
	ASF	**BP**	
Hpar1	0.591	0.563	0.415
Hpar2	0.669	0.480	0.340
Hpar3	0.624	0.495	0.490
Hpar4	0.624	0.742	0.400

These preliminary results lead to the conclusion that BP is a suitable scoring function for incremental fragment-based docking and highlight the need to build a model to identify the best-performing potential for each fragment. Moreover, BP should be tested as a sampling function for fragment-based docking through MC-based sampling.

It would be interesting to obtain a more detailed overview of the BP performance. For this, one could conduct assembly on the extended benchmark, e.g. one that was used for the scoring and even the RRM-ssRNA benchmark used to derive HIPPO. The selection of the hot-spots could be made based on the number of near-native poses in the top1%, as this threshold produced the highest percentage of near-native chains during the preliminary assembly (hpar1). Subsequently, different hyperparameter sets should be tested to identify the most optimal set.

4.5 Conclusions

In this chapter, we have introduced a novel histogram-based optimisation approach and HIPPO, a novel scoring potential designed for protein-ssRNA fragment-based docking poses in ATTRACT's coarse-grained representation. The main specificity of HIPPO is that it is a composite function comprising four distinct scoring potentials, each capable of covering specific protein-ssRNA binding modes. The application of these potentials, followed by the aggregation of their results, has yielded a

more precise ranking compared to the state-of-the-art ASF. HIPPO has notably enhanced the scoring of the best-docked fragment within each complex, enabling the use of this fragment as an anchor for incremental docking. We have also introduced the concept of BP, best-performing potential, which is currently limited to the test case. In this approach, the best-performing potential out of the four is identified and used in isolation for scoring.

Subsequently, we have presented the results of applying HIPPO and BP to a benchmark of complexes that were not used during HIPPO's development. Notably, HIPPO has outperformed the ASF in scoring these complexes, and BP has drastically outperformed both HIPPO and the ASF.

Moreover, we have employed the ASF ranking and BP ranking to assemble poses for a small subset of 14 complexes (3 fragments per complex). The preliminary results are highly promising, as BP has surpassed the ASF's performance across all four tested hyperparameter sets. This success provides strong motivation for the development of a model to derive BP from HIPPO for a given case.

It's worth noting that the protocol used to develop HIPPO (and potentially the model to go from HIPPO to BP) is a priori applicable to other types of complexes. Moreover, it could be used to address an inherent limitation of the fragment-based docking approach, hot-spot- and cold-spot-binding. We explore this topic further in Chapter 6.

The next chapter is dedicated to our work on data-driven RRM-ssRNA docking, with the primary aim to enhance the sampling.

Chapter 5: Data-Driven Docking for RRM-ssRNA Complexes

5.1 Aims

In the previous chapter, we have introduced a novel scoring function that enhances the scoring and might improve ab initio sampling for protein-ssRNA complexes. Here, we delve into an alternative approach to tackle the sampling problem for a specific subset of RRM-ssRNA complexes. This approach lies in data-driven docking guided by conserved stacking interactions. In this chapter, we present 'RRM-RNA-dock', a docking pipeline tailored to address the intricacies of RRM-ssRNA complexes. Additionally, we explore a potential source of experimental restraints that could serve as guidance for the broader field of protein-ssRNA data-driven docking.

5.2 Introduction

RRMs are known to have a conserved β1α1β2β3α2β4 structural topology and two consensus sequences called RNP1 and RNP2, located on β3 and β1 respectively. RNP1 and RNP2 often play key roles in RNA-binding, as the residue F from the RNP1 position 5 and the residue F/Y from the RNP2 position 2 often form pi-pi stacking interactions with ssRNA nucleotides (Fig. 5.1). This is known as a canonical binding mode (Fig. 5.2). A more detailed introduction of RRMs and specificities of RRM-ssRNA binding has been given in §1.3.4.1 and §1.3.5.1 respectively.

a)

b)

Figure 5.1 - Sequence motifs from the alignment of RRM domains for a) RNP1; b) RNP2. The positions of stacking residues are indicated by the red rectangles. These images were produced with the help of *WebLogo* [264].

5.2.1 Anchored Docking Methodology

Conserved stacking interactions can be leveraged for data-driven docking. A method for iterative data-driven fragment-based docking has been previously developed and validated on several RRM-RNA complexes [252]. In the context of this work, the term ‘anchor’ refers to a nucleotide (or bead) involved in a conserved stacking interaction, and the ‘anchoring amino acid’ refers to the amino acid involved in the stacking. ‘Anchored fragment’ refers to a fragment which contains one or several anchors (Fig. 5.2).

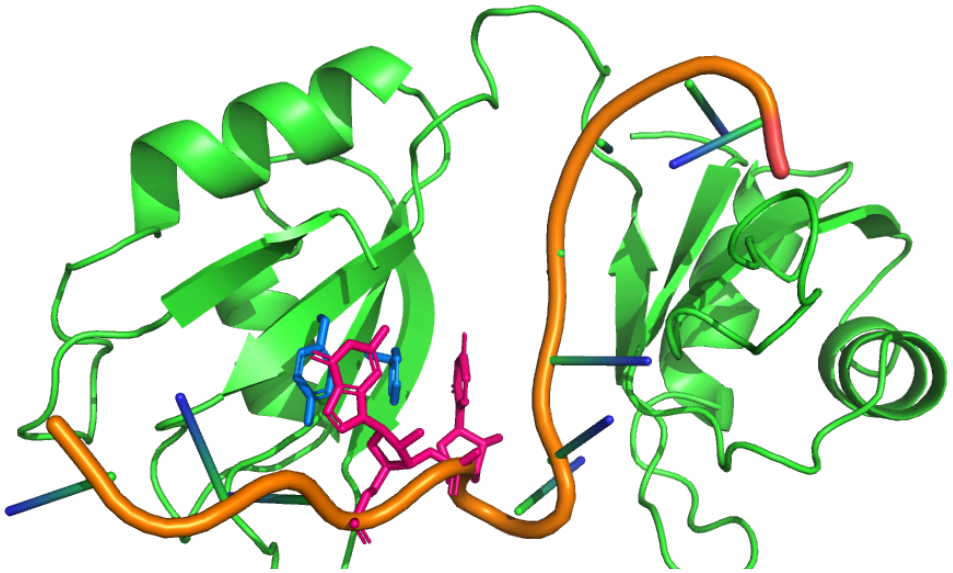


Figure 5.2 - Visualisation of the RRM-ssRNA complex (PDB_ID 1B7F) in canonical binding mode. Anchors (G_4 and U_5) are displayed in pink, and anchoring amino acids of RRM2 (Y_214 and F_256 respectively) are in blue. For both are shown in a stick representation.

The previously developed anchored docking protocol involves (i) docking of one or several anchored fragments with the help of the positional restraints between anchor(s) and its predicted position(s), followed by (ii) docking of the fragments adjacent to the anchored fragment with the help of positional restraints between the ends of the anchored fragment and the adjacent-to-it fragment (Fig. 5.3).

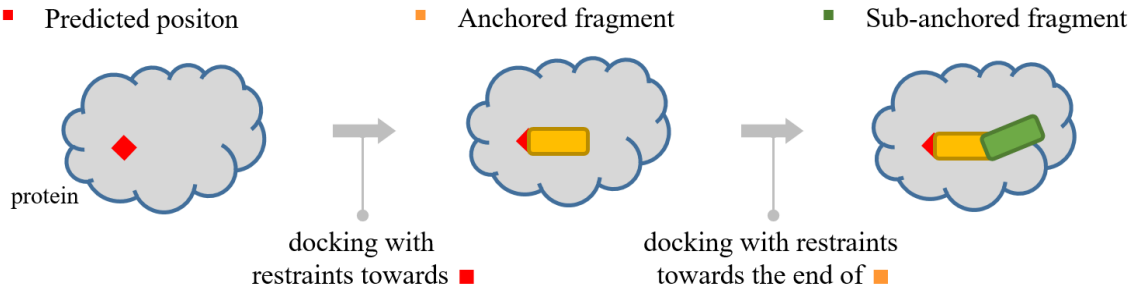


Figure 5.3 - Schematic image of the anchored docking of the anchored fragment (in yellow) and one adjacent fragment (in green). Predicted anchor position shown in red.

The position of each anchor has been predicted using the positions of the corresponding nucleotide in known complexes as reference positions. Application of the positional restraints essentially creates a spherical region around the chosen reference bead (or set of beads) position. If a pose's corresponding bead(s) is outside this region, an attractive force is applied between the pose and the selected reference bead(s). This force increases linearly with the distance between the bead and its reference position. Positional restraints are defined by two parameters, specifically the radius of the spherical region and the energy penalty.

In case the RNA chain is docked onto a tandem RRM, it is possible to dock anchored fragments on RRM1 and RRM2, and then iteratively grow the RNA chain in both 5' and 3' directions from these anchored fragments. In this case, along with short-range positional restraints, long-range distance

restraints can be implemented to guide the fragment's orientation towards the anchor located on the opposite RRM (Fig. 5.4).

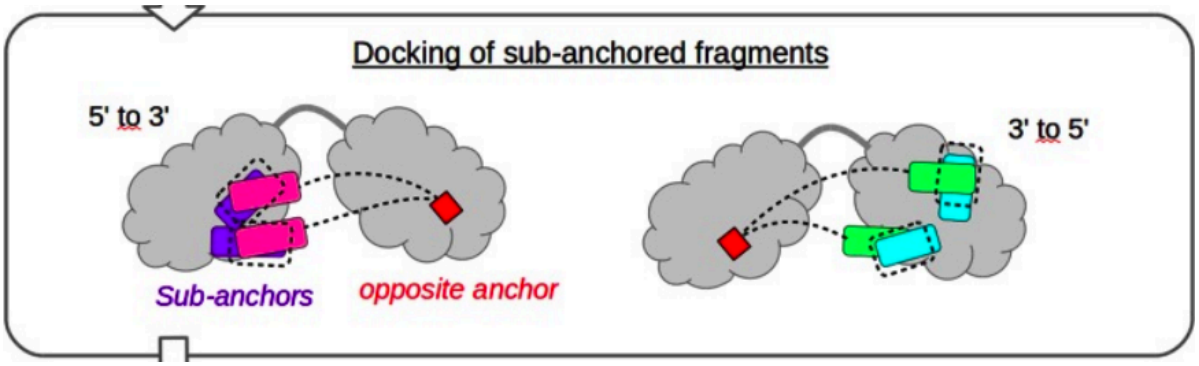


Figure 5.4 - Schematic image of the anchored docking of ssRNA on tandem RRM [271]

It is important to note that the accuracy of the docking results is directly influenced by the predicted positions of the anchor beads used for the restraints. When these positions closely match the actual positions of the beads in the bound nucleotide (within approximately 1Å), we can reasonably expect the best docking poses to be located within 1-2Å of the actual nucleotide positions. As the distances between the real bead positions and the positions used for restraints increase, the quality of the docking results decreases.

5.2.2 Anchoring Patterns

To improve the accuracy of the anchored docking method, a new set of predicted RRM-ssRNA anchor positions, called anchoring patterns, was created (Fig. 5.5). An anchoring pattern consists of a fitting region and an anchor. The fitting region is a stretch of 3 amino acids with the anchoring amino acid in the middle. When the fitting region of the anchoring pattern is aligned with the target RRM structure, the position of the anchor in the pattern can be used to drive the docking of an anchored fragment.

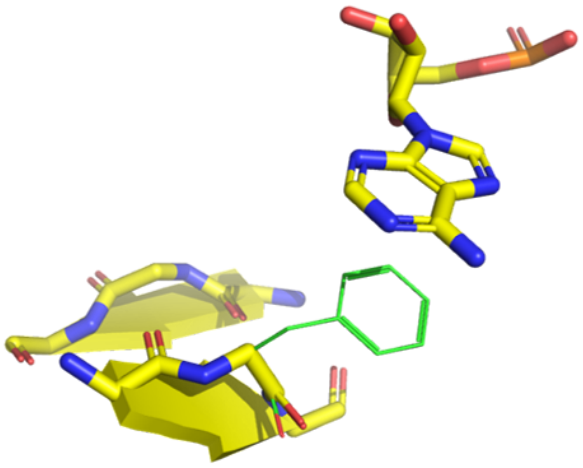


Figure 5.5 - Visualisation of the anchoring pattern. The fitting regions and anchor are in yellow, the anchoring amino acid's side chain is in green.

To ensure the success of the docking process, it's crucial that the anchoring patterns closely resemble the actual target position. The most straightforward way to achieve such close resemblance is through the generation of a set of anchoring patterns by clustering the known positions of the anchors, relative to the fitting region. In this context of clustering, each cluster representative constitutes one anchoring pattern.

Currently, there is no method for the selection of the best pattern for a given docking case, therefore the docking procedure must be conducted for each pattern. This results in a substantial pool of docking poses if a large number of patterns are used. The substantial pool of docking poses leads to a greater challenge in the identification of the near-natives. Additionally, multiple docking runs become computationally expensive quickly. Thus, a trade-off must be found when determining the clustering cutoff. It should be set at a level that yields a limited number of patterns (achieved with a loose clustering cutoff) while ensuring that each pattern remains sufficiently proximate to its nearest centre of a cluster (achieved with a tight clustering cutoff).

To obtain anchoring patterns, a novel hierarchical agglomerative clustering (HAC) was employed. This method is designed to minimise the number of clusters while ensuring that each initial element remains within a specified distance threshold from at least one of the final patterns [254]. Notably, the representatives of each cluster are not the same as one of the initial elements; instead, they are centroids of their respective clusters. To assess the similarity between 3D structures and positions of the same residue, the RMSD was used.

The novelty of this method lies in its combination of hierarchical agglomerative clustering and the computation of minimum enclosing balls to derive ε-nets of finite sets in a reproducing kernel Hilbert space. It produces ε-nets with smaller cardinalities compared to state-of-the-art methods. Unlike classical hierarchical agglomerative clustering, this method stops the algorithm as soon as the candidate merging produces a set with a minimum enclosing ball radius greater than or equal to ε. Additionally, the prototypes generated by this method reside in a Reproducing Kernel Hilbert Space, which can be infinite dimensional, posing no difficulties due to the kernel trick [254].

A set of anchoring patterns was created by Hrishikesh Dhondge by clustering 257 RRM-RNA structures with conserved stacking. These structures were collected and aligned using InteR3M [96]. The structures with stacking occurring within RNP2 are referred to as the ‘Beta1’ group, and the structures with stacking in RNP1 are referred to as the ‘Beta3’ group.

The fitting regions along with the stacking nucleotides were extracted from each structure and superimposed onto the reference structure. The crystal structure of the sex-lethal protein, specifically RRM1 (chain A) from 1B7F, was used as a reference. Next, all 4 residues (3 amino acids plus 1 nucleotide) were converted from all-atom to the ATTRACT coarse-grained representation. For the clustering, the same number of points is required, however, the bases of purines are represented by 4 beads, while the ones of pyrimidines are only by 3. Thus, to cluster nucleotides of both types simultaneously, the 4th bead (created from the N7 atom) was removed. Its position can be calculated using the coordinates of the remaining 3 beads.

Application of HAC to the available data with the clustering thresholds 3.0Å and 3.5Å for ‘Beta1’ and ‘Beta3’ respectively resulted in a set of 4 patterns for ‘Beta1’ and 5 patterns for ‘Beta3’ (Fig 5.6).

a) *b)*

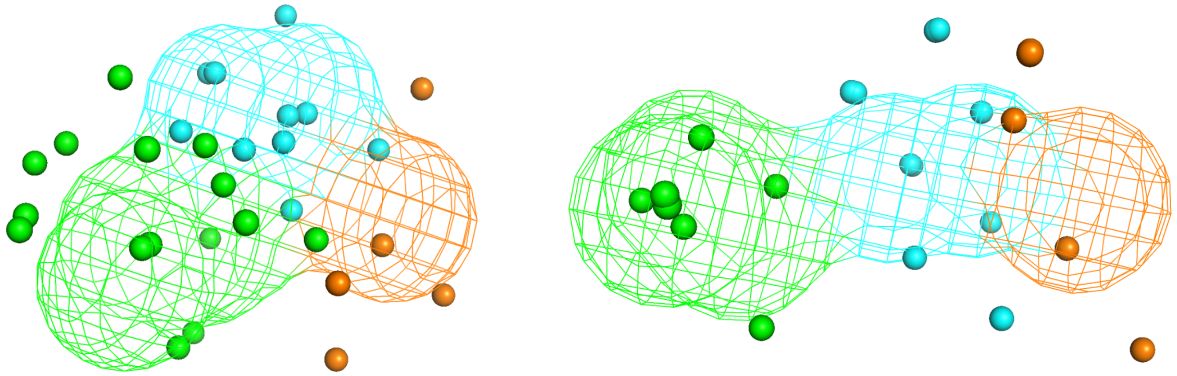


Figure 5.6 - Prototype and members (only nucleotides are displayed) for a) 'Beta1', clustered at 3.5Å; b) 'Beta3', clustered at 3.0Å. The prototype is displayed as mesh, all members are displayed as spheres. Phosphate spheres/mesh are shown in orange, sugar spheres/mesh are shown in cyan and base spheres/mesh are shown in green [Image by Hrishikesh Dhondge].

5.2.3 RRM Structure Modelling

The fragment-based docking protocol, including its anchored version, necessitates a 3D structure of the receptor. This process often involves manual preparation of the structure, including domain selection, unwanted cofactors and modifications removal, and potential protein region remodelling. For RRMs, an added complexity arises from the necessity to identify anchoring amino acids for the further utilisation of the anchoring patterns. Those amino acids are typically associated with specific RRM sequence positions, so they can be identified either via sequence alignment or detailed manual structure inspection. In essence, these requirements made baseline docking protocols more suited for skilled structural bioinformaticians rather than researchers with limited 3D structure expertise.

However, these challenges can be addressed by the utilisation of AlphaFold DB [112] and InteR3M DB [96]. The robustness, accessibility, and extensive coverage of computational models available through AlphaFold DB allow to obtain an accurate model of the RRM structure. On the other hand, accessibility and extensive coverage of RRMs provided by InteR3M allow for the identification of the anchoring amino acids. Together, these databases allow us to streamline the preparation necessary for the anchored docking.

5.3 RRM-RNA dock

5.3.1 Pipeline

Motivation

Execution of the anchored docking strategy with the integration of anchoring patterns requires (Appendix C.1):

- Preparation of the receptor structure;
- Identification of the reference anchor position using anchoring patterns;

- Creation of the positional restraints file;
- Running of the docking with restraints.

All these steps are necessary for the initial docking of the anchored fragment. Completion of these steps requires skills both in structural biology and programming, and is generally time-consuming, especially for novice users. This severely limits the accessibility of the anchored docking. This issue can be solved by creating an automated pipeline capable of performing the aforementioned steps based on AlphaFold DB, InteR3M DB and the ATTRACT docking engine. This pipeline also simplifies the process of testing the anchoring patterns.

Please note that this pipeline was developed before the final version of HIPPO, thus within the pipeline an original ATTRACT scoring function is used.

Overview

I designed and implemented 'RRM-RNA dock,' an accessible anchored docking pipeline tailored for RRM-ssRNA complexes. This user-friendly tool aims to broaden the availability of anchored docking. The current iteration of this pipeline focuses on docking a single anchored fragment (trinucleotide) of ssRNA. Within this fragment, two nucleotides interact with distinct amino acids from RRM, specifically 'Beta1' and 'Beta3.

This pipeline is implemented in Python (version 3.9.13) and Bash (version 5.1). It is compatible with Linux and can be executed from the command line. For the proper functioning of the pipeline, the ATTRACT docking engine should be installed on the machine.

The following user input is required to execute the pipeline (Fig. 5.7):

- Protein identifier (UniProtKB accession number);
- RRM domain index;
- RNA sequence;
- Position of the ‘Beta1’ anchor in the RNA sequence;
- Position of the ‘Beta3’ anchor in the RNA sequence.

Initially, the pipeline verifies the proper set-up of the environment and the user’s input (user manual is available in Appendix C.2.) Then the structure of the given protein is downloaded from the AlphaFold DB. Complementary information about the protein (indexes of the given domain and its fitting regions) is taken from InteR3M DB and is used to extract the given RRM domain and to identify the 2 fitting regions within. Next RRM structure is converted to ATTRACT coarse-grained representation and the anchoring prototypes (their fitting regions) are superimposed onto the corresponding RRM’s fitting regions. The restraint files are created based on the position of the anchors in the patterns and types of user-given anchored nucleotides.

Since the current pipeline version handles fragments containing 2 anchors, all possible combinations of 2 anchoring patterns are used, resulting in 20 distinct docking settings. The corresponding folder structure is created with a folder dedicated to each combination of patterns (Fig. 5.8).

```
(attract) akravche@garimpeiro:/data1/akravche/scripts/anchoring/git$ python3 $SCR/anchoring/git/pip.py --help
usage: pip.py [-h] -wdir WORK_DIRECTORY -id UNIPROTID -rrm RRM_DOMAIN_ID -seq
              SS_RNA_SEQUENCE -ancNucB1 ANCHORING_NUCLEOTIDE_ID_BETA1
              -ancNucB3 ANCHORING_NUCLEOTIDE_ID_BETA3
              [-conf [CONFIGURAION_FILE]]

optional arguments:
  -h, --help            show this help message and exit
  -wdir WORK_DIRECTORY, --work_directory WORK_DIRECTORY
                        Path to directory where RRM folder for the docking
                        will be made.
  -id UNIPROTID, --uniProtID UNIPROTID
                        UniProt ID for the RRM, e.g. 'P31946' or 'P62258' etc.
  -rrm RRM_DOMAIN_ID, --rrm_domain_id RRM_DOMAIN_ID
                        RRM domain index e.g. '1' or '2' etc.
  -seq SS_RNA_SEQUENCE, --ss_rna_sequence SS_RNA_SEQUENCE
                        Single-stranded RNA sequence to be docked onto the
                        RRM, at least 3 nucleotides long, e.g. 'CAC' or 'GCAC'
                        etc.
  -ancNucB1 ANCHORING_NUCLEOTIDE_ID_BETA1, --anchoring_nucleotide_id_beta1 ANCHORING_NUCLEOTIDE_ID_BETA1
                        Anchoring nucleotide index for betasheet 1, e.g. '1'
                        or '2' etc.
  -ancNucB3 ANCHORING_NUCLEOTIDE_ID_BETA3, --anchoring_nucleotide_id_beta3 ANCHORING_NUCLEOTIDE_ID_BETA3
                        Anchoring nucleotide index for betasheet 3, e.g. '2'
                        or '3' etc.
  -conf [CONFIGURAION_FILE], --configuraion_file [CONFIGURAION_FILE]
                        Configuration file with advanced parameters. Leave
                        empty to use default file.
(attract) akravche@garimpeiro:/data1/akravche/scripts/anchoring/git$ _
```

Figure 5.7 - Help message of the main script of the pipeline.

```
(attract) akravche@garimpeiro:/data1/akravche/dataset/data21/anchoring/data/P26368/rrm1$ ls *
boundfrag.list        fitted-protoB3-1.pdb  proteinAFold.pdb        protoB3-1.pdb
domain.pdb            fitted-protoB3-2.pdb  protoB1-1-renumber.pdb  protoB3-2.pdb
domainr.pdb           fitted-protoB3-3.pdb  protoB1-1.pdb           protoB3-3.pdb
extract_domain.pdb    fitted-protoB3-4.pdb  protoB1-2.pdb           protoB3-4.pdb
fitted-protoB1-1.pdb  fitted-protoB3-5.pdb  protoB1-3.pdb           protoB3-5.pdb
fitted-protoB1-2.pdb  frag.info             protoB1-4.pdb           protoB3.pdb
fitted-protoB1-3.pdb  motif.list            protoB1.pdb             restraints.txt
fitted-protoB1-4.pdb  proteinAFold.fasta    protoB3-1-renumber.pdb

b1_1:
b3_1  b3_2  b3_3  b3_4  b3_5

b1_2:
b3_1  b3_2  b3_3  b3_4  b3_5

b1_3:
b3_1  b3_2  b3_3  b3_4  b3_5

b1_4:
b3_1  b3_2  b3_3  b3_4  b3_5
(attract) akravche@garimpeiro:/data1/akravche/dataset/data21/anchoring/data/P26368/rrm1$ _
```

Figure 5.8 - An example of the output folder structure.

The default parameter settings for the positional restraints are:

- For ‘Beta1’:
  - The radius of the spherical region is 4.5Å;
  - The energy penalty is 3.5 units.
- For ‘Beta3’:
  - The radius of the spherical region is 4.0Å;
  - The energy penalty is 3.5 units.

Such choice of the radiuses is dictated by the clustering thresholds, 3Å and 3.5Å respectively. Smaller distances could introduce overfitting to the patterns, while larger distances could lead to underfitting. The suitable values of energy penalty were determined empirically.

This is followed by the 20 subsequential docking runs, each with a distinct pattern combination. Each docking run utilises 8 CPUs by default. The default parameters (radius, penalty and CPU) are declared in a configuration file and can be modified.

The output of the pipeline is a set of files containing docking poses generated for each pattern combination. In a test case, when the native structures of the fragment are available, files containing the LRMSDs are generated as well.

5.3.2 Results and Discussion

The performance of the anchored docking was tested by selecting a test protein which is known to have 2 conserved stacking interactions in at least one of its complexes with RNA that are experimentally solved. These complexes were excluded from the data set and the anchoring patterns were re-generated with the same clustering settings. Then, the pipeline was executed for this test protein using the knowledge of the sequence of the anchored fragment and the position of the anchors in sequence. The docking results given by each pattern combination were pooled together, and redundant poses were removed. Further in the text, this is referred to as 'all-patterns'.

Table 5.1 - Test set and docking results for the protein

PDB.ID_chain	'Beta1' anchor index_chain	'Beta3' anchor index_chain	Total number of poses with LRMSD<2Å	Rank of the first pose with LRMSD<2Å
1AUD_A	43_B	44_B	603	7,174
1DRZ_A	152_B	153_B	974	30,706
1DZ5_A	43_D	44_D	1114	12,156
1DZ5_B	17_D	18_D	852	8,892
1M5K_C	40_B	41_B	1051	35,000
1M5K_F	40_E	41_E	1074	34,035
1M5O_C	40_B	41_B	986	37,681
1M5O_F	40_E	41_E	986	37,681
1SJ3_P	152_R	153_R	951	30,706
1U6B_A	1007_B	1008_B	825	56,079
1URN_A	10_P	11_P	931	30,706
1VBX_A	152_B	153_B	908	12,156

This protocol was run for the U1 small nuclear ribonucleoprotein A (UniProtKB accession number P09012), RRM1, and a selected set of 12 structures showing stacking interactions with the fragment 5'-CAC-3', specifically C1 in 'Beta1' and A2 in 'Beta3' (Tab. 5.1). The pipeline was run with the parameters [P09012; 1; CAC; 1; 2]. The all-patterns results were compared to each of the selected structures. The average rank of the first docking pose similar to the experimental structure (LRMSD < 2Å) was equal to ~28,000. Most importantly, more than 600 near-native poses among the

1.075.454 docking poses were found for each reference structure, with at least one in the 40,000 top-ranked poses for 11/12 structures.

Using the best-fitted pattern combination would produce a much better ranking than using all pattern combinations:

- for the experimental structure 1DZ5_B, the best docking pose (LRMSD = 0.7Å) was ranked 3,184/48,717 in the list of poses given by the combination of the second pattern for 'Beta 1' and the first pattern for 'Beta 3', versus 61,037/1,075,454 in the pool from all pattern combinations;
- for experimental structure 1VBX_A, the best docking pose (LRMSD = 1.0Å) was ranked 10,156/59,498 in the list of poses given by the patterns combination of the second pattern for 'Beta 1' and the second pattern for 'Beta 3', versus 282,998/1,075,454 in the pool from all pattern combinations.

We also compared the quality of anchor-driven sampling (both usage of 20 distinct pattern combinations and 1 all-patterns) against *ab initio* sampling for the case 1DRZ-CAC (Tab 5.2). Notably, 30 million starting points were used for the *ab initio* sampling, while the use of the patterns allowed to reduce this number to only 100 starting points. This reduced the number of sampled poses from 10 million to the range of [33,628; 69,929] respectively. Such a reduced number of poses is much easier to process computationally. For example, on 30 CPUs it takes ~6h to sample one fragment *ab initio* and only 10+/- 5 minutes with anchored sampling. In case if all pattern combinations are used, the number of the starting points technically increases to 2 million in total (100 starting points multiplied by 20 separate docking runs), which is still less than 30 million.

Only one anchor-driven sampling (1/21), precisely one with the pattern combination [4;5], resulted in a lower percentage of near-natives (LRMSD<2Å) than *ab initio* sampling. For the distinct pattern combinations, the min value is 0% (second lowest is 0.0022%), the max is 0.3319%, and the average is 0.0726%, versus 0.0906% for all-patterns and 0.0004% for *ab initio*. If the LRMSD threshold is relaxed to 3Å, then all anchor-driven sampling runs (21/21) produce a higher percentage of near-natives than *ab initio* sampling. These results suggest that anchor-driven sampling outperforms *ab initio* sampling (Appendix C.3).

Table 5.2 - Comparison of different types of samplings, namely (i) *ab initio*, (ii) anchored with each pattern combination, and (iii) anchored all-patterns, for case 1DRZ-CAC.

The titles of the columns indicate the type of sampling or the used pattern combination (the first number indicates the number of the pattern for 'Beta1', and the second number - the pattern for 'Beta3'). The values in the cells indicate the percentage of near-native poses under a given threshold among all sampled poses and the total number of sampled poses.

Among the values given by distinct patterns, 4 per row with the highest percentages of near-natives are underlined, and the values higher than average (among the 20 distinct patterns) are highlighted in yellow. A single value lower than *ab initio* is highlighted in red.

LRMSD	***Ab initio***	**All-patterns**	**1 1**	**1 2**	**1 3**	**1 4**	**1 5**	**2 1**	**2 2**	**2 3**	**2 4**
Under 2Å	0.0004	0.0906	0.0461	0.0290	0.0205	0.2900	0.0051	0.0807	0.0319	0.1128	0.0358
2Å to 3Å	0.0122	2.5095	5.9883	2.5540	2.7058	5.7699	1.7521	5.1094	1.6522	3.4892	5.6271
All poses	10 mil	1,075,454	58,614	48,717	59,785	68,834	55,254	63,958	59,498	62,926	69,929

Table 5.2 - Continuation

LRMSD	2 5	3 1	3 2	3 3	3 4	3 5	4 1	4 2	4 3	4 4	4 5
Under 2Å	0.0218	0.0822	0.2644	0.0048	0.0451	0.0065	0.0264	0.0154	0.0022	0.3319	0
2Å to 3Å	1.3967	2.3555	4.0387	0.1774	0.8996	0.6159	1.6332	0.4440	0.4330	2.9235	0.6572
All poses	63,375	46,901	38,111	62,025	55,976	52,464	47,032	33,628	50,487	45,493	49,072

In terms of scoring, for this particular case, the poses sampled with the best-fitted pattern combination generally scored better compared to the *ab initio* poses. Specifically, for poses with LRMSD<2Å, both in the top5% and top20%, the scoring of the *ab initio* poses is inferior to that of the poses obtained with the best-fitting pattern combination. For poses with LRMSD<3Å, the *ab initio* poses scoring in the top5% is also less effective, whereas, in the top20%, it surpasses the scoring of the poses obtained with the best-fitting pattern combination (Tab 5.3 and see Appendix C.3 for details).

Table 5.3 - Comparison of the scoring of the *ab initio* poses and poses sampled with the use of the best-fitted pattern combination, for case 1DRZ-CAC.
Here the best-fitted pattern combination is determined based on the scoring results.

LRMSD	*Ab initio*	3 4	LRMSD	*Ab initio*	3 3
<2Å in top5%	45	56	**<3Å in top5%**	44	28
<2Å in top20%	78	96	**<3Å in top20%**	73	75

Several additional features can be implemented to enhance the quality of the docking results for a single fragment. One potential improvement is in customising the positional restraints based on different bead types. For example, for the bead representing the Phosphate group (GP1), using the maximum distance observed during clustering, which is 6Å, for the restraints could lead to more accurate docking models by reducing overfitting to the pattern.

Furthermore, it would be immensely advantageous to predict the most appropriate patterns (or discard the least appropriate ones) for each docking case beforehand. Since the positioning of the stacking nucleotide relies on the side-chain position of the amino acid involved in stacking, a manual examination of side-chain positions in the clustered structures has shown that most positions are unique to a single cluster. Only a few positions are shared across multiple clusters. Consequently, it would be feasible to predict the ideal pattern for docking on a specific target protein if one could predict the bound position of the side chain of the amino acid (Fig 5.9). This prediction could be based on either sequence information or an unbound structure of the protein.

As mentioned previously, the current pipeline is limited to the docking of a single fragment with 2 anchored nucleotides to a single RRM. From the technical viewpoint, its functionality can be easily expanded to cover the anchored docking of a single fragment with a single anchored nucleotide. The subsequent stage involves integrating a tandem RRM as a receptor and introducing long-range distance restraints between the fragment and the anchoring amino acid(s) located on the opposite RMM.

Ultimately, the goal would be to implement a comprehensive data-driven fragment-based protocol based on the described anchoring methodology, capable of docking adjacent fragments within the RNA sequence, starting with the fragment containing 1 or 2 anchored nucleotides. This protocol would then select chains of compatible poses (one pose per fragment), thus discarding, for

each fragment, most of the poses that cannot be connected to any pose of the adjacent fragment(s). The resulting models of RRM-ssRNA chains could subsequently undergo testing through molecular dynamics simulations to identify the most stable model.

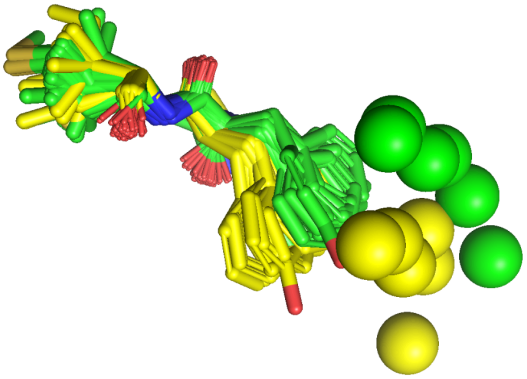


Figure 5.9 - The prototypes from clusters 1, shown in green, and cluster 4, displayed in yellow, of the 'Beta3' along with the amino-acid residues used for fitting.

5.4 Additional Experimental Restraints

RRM-ssRNA docking can be guided by experimental data of a non-structural origin [265]. These data can be integrated into the docking protocol as a set of interface or contact restraints. As part of the RNAct project, I have led a short-term project intending to collect experimental data from the scientific literature. The primary goal of this project is to curate a set of non-structural experimental data for RRM-ssRNA complexes for which a 3D structure is also available. This data collection is essential for the establishment of a data-driven docking protocol, which includes (i) the assessment of the quality of the collected data itself and (ii) the assessment of the quality of the docking results through docking test cases.

To assess the quality of the collected data, each entry should be validated by the experimental 3D structure. From all those results together, we can infer a correspondence between the type of experimental data and:

- the distance of the restraint to apply;
- the confidence in this restraint (to be translated into a probability of fulfilling the restraint).

Once these results are ready, inferred restraints should be tested via data-driven docking. To evaluate the improvement of the docking performance, data-driven docking results are to be compared with ab initio docking results.

Several types of data to be collected are:

- Positive interface data: the knowledge of a set of RRM's residues that are directly involved in the binding of RNA chain with a certain probability;
- Negative interface data: the knowledge of a set of RRM's residues which are proven to not influence the binding;
- Contact data: the knowledge of the direct binding between an RRM's residue and an RNA's residue with a certain probability and at a certain distance interval.

The sources of such data are results obtained by experimental techniques, including but not limited to the following methods:

- NMR chemical shift (interface data);
- FRET (contact data with broad distance);
- Mutagenesis (interface data);
- High-resolution cross-linking (contact data);
- H/D-exchange on RRM with/without RNA (interface data);
- Sequence Variations in RNA (interface data).

To limit the scope of the task, 49 solved RRM-ssRNA complexes were targeted (Appendix C.4). Around 100 articles, published by 27/04/2020, were collected. These articles contained data regarding 46 proteins. Experimental data was extracted and organised in a spreadsheet (Tab 5.4). Initially, 105 entities were collected, out of which 82 entries were validated by Isaure Chauvot de Beauchene and added to the resulting spreadsheet. Two examples are presented in Tab 5.5 and a full spreadsheet is available online [266].

Data-driven docking is to be performed for entries with '1' in the 'Does_it_bind' field and '3' in the 'Binding_probability' field, gradually moving to the values '2' and '1'. The quality of the docking poses will allow us to assess the quality of the collected data. If the results outperform *ab initio* docking consistently, similar data could be collected for the unsolved RRM-ssRNA and used for the data-driven docking.

At present, a similar type of data can be found on UniProt (Fig 5.10), which could expand the collected data.

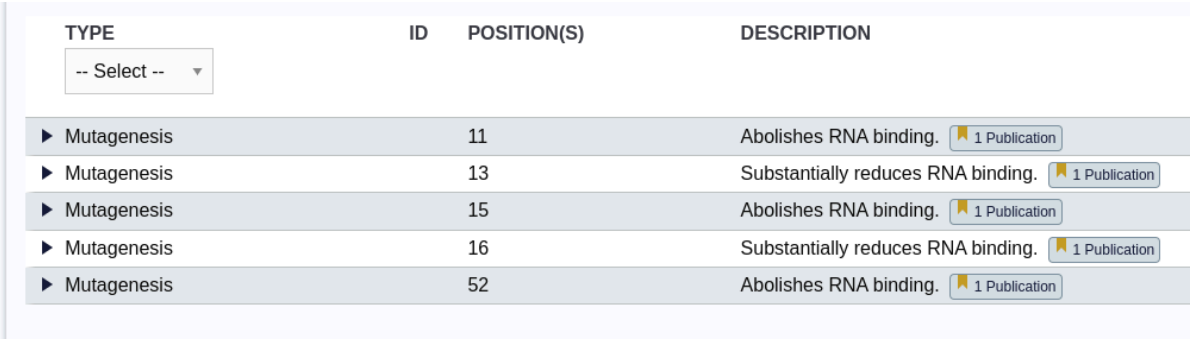


Figure 5.10 - Experimental data for U1 small nuclear ribonucleoprotein A, available via UniProt [267].

Table 5.4 - Organisation of the spreadsheet for the experimental data.

Column title	Description	Column title	Description
#1 Protein_name	The name of the protein	#9 Stacking	'1' if interaction is identified as stacking (for contact data only)
#2 PDB_ID	PDB_ID of the bound complex	#10 Min/Max distance	Minimal or maximal distance, at which interaction is possible (self assessed)
#3 Unbount PDB_ID	PDB_ID of the unbound protein	#11 Does_it_bind	'1' if the data is positive '0' if the data is negative (for interface data only)
#4 RRM's_Residue_ID	Index and type of the residue If residue was mutated - both wild type and mutant	#12 Binding_probability	(self assessed based on described impact of the experiment) '1' for low probability '2' for medium probability '3' for high probability
#5 Backbone/Side_chain	'sc' if interaction occurs with the side chain 'bb' if interaction occurs with the side chain	#13 Method_Name	The name of the used method(s)
#6 RNA_Sequence	5' to 3'	#14 Article_Link	Link or DOI
#7 Nucleotide_ID	Index and type of the residue	#15 Quote	Quote from the article describing experimental result(s)
#8 Phosphate/sugar/base	Identification of the interacting part of the nucleotide	#16 Comments	Additional comments, if needed

Table 5.5 - Examples of collected data.

Limited number of columns is shown. The numbers in the title row corresponding to the numbers given in the Tab 5.4.

#1	#2	#4	#5	#6	#8	#9	#10	#11	#12	#13	#15
TDB-43	4IUF, 4Y0F, 4BS2, 4Y05	W113A	sc	GUGUGAAUGAAU	base	1	5	1	2	Mutagenesis	The Trp113 mutation results in a modest three-fold loss in binding affinity compared to that of wild-type TDP-43 RBD.
TDB-43	4IUF, 4Y0F, 4BS2, 4Y05	R151A	sc	GUGUGAAUGAAU	-	-1	5	0	2	Mutagenesis	R151A did not impact sequence recognition of the RRM1-RRM2 tandem construct.

5.5 Conclusion

In this chapter, a data-driven approach compatible with fragment-based docking has been presented. It uses so-called anchoring patterns, which represent the average positions of the anchor (stacking nucleotide), to drive the docking. These patterns have been derived through the clustering of RRM-RNA structures that feature conserved stacking interactions. My contribution lies in the development of the docking pipeline designed to facilitate the data-driven docking of fragments containing two anchors onto RRM. This pipeline has a simple command-line interface. It uses AlphaFold DB to obtain a model of the RRM, InteR3Mdb to identify anchoring amino acids, anchoring patterns to identify an approximate location of the anchors relative to anchoring amino acids, and the ATTRACT docking engine to perform the docking with restraints following a previously established anchor-driven protocol. Despite the limited functionality, this pipeline significantly streamlines the data-driven docking process by eliminating the necessity for manual receptor structure preparation and the manual creation of restraint files. As a result, it enhances the accessibility of data-driven docking, particularly for users with limited experience in structural biology. As expected, the sampling performance of this docking is better than that of ab initio docking.

The latter part of the chapter has presented a set of non-structural experimental data, collected manually from the literature. This dataset could be used as a source of additional docking restraints to ultimately expand 'RRM-RNA dock' to a general protein-ssRNA docking pipeline. We will explore this, along with other potential prospects for further research in the next, and final, chapter.

Chapter 6: Conclusions and Perspectives

6.1 Aims

In this chapter, we conclude the thesis by summarising the proposed contributions and by highlighting future work for both HIPPO (Histogram-based Pseudo Potential) and 'RRM-RNA dock' pipeline.

6.2 Summary of Contributions

The primary objective of this thesis was to improve protein-ssRNA docking by addressing the scoring problem. Our approach to this problem involved the development of a novel scoring function, HIPPO, tailored specifically to protein-ssRNA interactions. Additionally, we developed a user-friendly docking pipeline 'RRM-RNA dock', tailored to RRM-ssRNA complexes. In this section, we will summarise both contributions and highlight the key outcomes, as well as briefly touch upon some smaller projects undertaken in the frame of this doctoral research.

6.2.1 HIPPO

The foundation for HIPPO was an unsuccessful Monte Carlo Simulated Annealing optimisation of the whole original docking parameters set of the ATTRACT scoring function (ASF) followed by the design of the histogram-based optimisation approach. The first project led us to the hypothesis that a singular parameter set is insufficient for the accurate scoring of protein-ssRNA fragments. The second project was a prerequisite for the implementation of the protocol to derive HIPPO. Unlike existing scoring functions, HIPPO is composed of 4 distinct scoring potentials, capable of accounting for different binding modes.

Experimental evaluation of HIPPO was performed by scoring a set of protein-ssRNA complexes. HIPPO outperformed ASF, state-of-the-art in protein-ssRNA fragment-based docking in coarse-grained representation. Moreover, these results proved HIPPO's generalisability, as it was derived from RRM-ssRNA complexes exclusively. Furthermore, the use of the best-performing potential (BP) for each scoring case yielded a better ranking compared to both ASF and HIPPO. Preliminary results of the assembly of 3 fragments of several complexes suggest that BP would be a suitable scoring function for incremental docking.

6.2.2. RRM-RNA dock

Stacking interactions between amino acids in conserved positions and unpaired nucleotides can serve as anchors and drive protein-ssRNA docking. This pre-existing anchoring approach requires information about the possible positions of the stacking nucleotide (anchor) with respect to the stacking amino acid, i.e. anchored patterns. A set of anchoring patterns was generated by Hrishikesh Dhonge via clustering of the experimentally determined 3D structures of RRM-RNA complexes. By uniting these anchoring patterns with the anchoring-docking methodology, we created an ATTRACT-based pipeline for RRM-ssRNA fragment docking. This pipeline sources a model of RRM from AlphaFoldDB and runs ATTRACT docking for a fragment with two stacking nucleotides, with maximal distance restraints toward each possible anchor position.

As anticipated, this pipeline provides a better sampling compared to *ab initio* docking. Its notable advantage lies in its accessibility to non-experts in computational structural biology. Users are exempt from the tasks of preparing the receptor's 3D structure or identifying amino acid positions, building restraints, etc.

6.2.3 Other Contributions

Smaller in scale but still noteworthy were two projects involving the examination of TRP-C docking parameters and the collection of experimental data for data-driven docking.

The examination of TRP-C docking parameters was initiated due to the original ASF parameters assigning unfavourable scores to native structures containing TRP-C residues. This was counterintuitive as one would expect somewhat favourable scores for these structures, especially when the residues were in a stacking-like orientation. To address this inconsistency, we attempted to fine-tune the TRP-C parameters manually. While these adjustments did render negative scores for native poses, subsequent sampling and scoring using this updated parameter set showed a decline in performance. Further fine-tuning might be achieved through a systematic brute-force approach. Alternatively, a parameter set optimisation could be carried out using an approximation of initial parameter values using an all-atom force field, rather than the original scoring parameters.

A collection of experimental data from the literature was conducted during a short-term collaboration with 9 other PhD students within the RNAct project, with the aim to ultimately enhance data-driven docking. Non-structural data for solved protein-ssRNA complexes were gathered from the literature. These data will be tested to assess its suitability for data-driven docking. If found suitable, similar data for unsolved complexes (to be gathered) could be employed in actual docking cases.

6.3 Perspectives

6.3.1 An Incremental Approach vs. Dual Potentials for Hot-Spot and Cold-Spot Binding

A challenge, associated with fragment-based docking, is related to the concept of hot-spot (HS) binding and cold-spot (CS) binding. This issue could be handled by HIPPO in combination with an *incremental docking* strategy, where a single HS-bound fragment is docked with high accuracy, and the rest of the chain is modelled fragment-by-fragment from the poses of the first fragment. HIPPO considerably increases the part of the near-native poses among the top-ranked poses for the best-docked fragment of a complex, allowing for such incremental modelling. In this scenario, the bottleneck lies in the identification of the HS-bound fragment prior to docking (at least prior to assembly). When the identification is not possible, one could consider each fragment as HS-bound, iterate over all fragments, and pool together the top-ranked results of each iteration (Fig. 6.1).

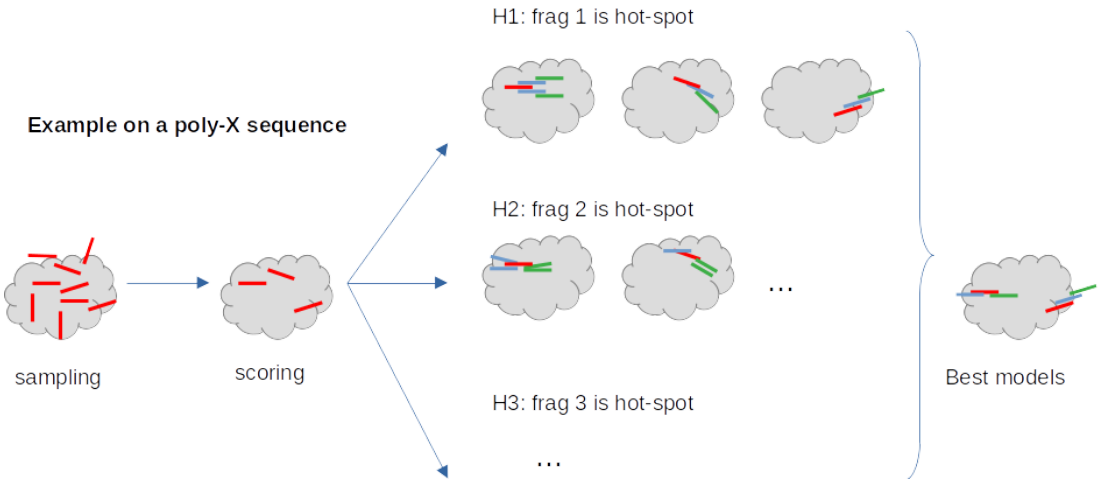


Figure 6.1 Simplified illustration of iterative docking using each fragment as an HS-bound.

Alternatively, one could investigate the possibility of the development of *dual scoring potentials*, capable of an accurate scoring of the docking poses of either HS-bound fragments or CS-bound fragments separately, i.e. using a classical fragment-based approach. To obtain such dual potential, training data cases (benchmark fragments to derive scoring potentials) should be labelled as HS-bound and CS-bound prior to the training. There are two approaches to doing so:

- Approach 1. Let us define the HS-bound fragment as the one for which the current HIPPO/BP is successful;
- Approach 2. Let us define the HS-bound fragment as in close proximity to the HS amino acids. HS amino acids could be identified using a specialised approach [268, 269, 270] (initial testing for protein-ssRNA may be required) .

Based on these two training sets, a double scoring potential, HIPPO-HS and HIPPO-CS can be derived. It is possible that fewer than four scoring parameter sets will be required for each training set, potentially simplifying the application of the resulting potentials. If, at this point, there is no model capable of identifying the fragment type (HS- or CS-bound) prior to docking, each fragment would undergo two rounds of scoring - once with HIPPO-HS and once with HIPPO-CS. Multiple assemblies (k assemblies), where k represents the total number of fragments in the complex, would be performed.

The development of a dual scoring potential would yield eight distinct sets of parameters (four for HIPPO-HS and four for HIPPO-CS). While it may be a distant possibility, there could be an avenue to transition from HIPPO-HS and HIPPO-CS to the concept of the best-performing potential (BP). This transition could be achieved by training a classifier to identify the appropriate BP parameters for each fragment based on the docking input or directly from the pool of docking poses.

6.3.2 Characterisation of the Protein–ssRNA Binding Modes using BP

HIPPO is capable of accounting for the different protein-ssRNA, and especially RRM-ssRNA binding modes by using 4 distinct scoring potentials. The concept of BP consists in identifying, for a given complex and fragment, the single $\mathcal{H}$ that outperforms the other three $\mathcal{H}$ in ranking near-native poses in the top5%. This approach may enable the characterisation of different protein-fragment binding modes - sets of distinct protein-ssRNA interactions with a biological meaning behind them - essentially distinguishing one mode from the others, based on the BP.

The BP can be easily identified for each test case. Consequently, all cases (the entire benchmark) can be categorised into five classes, with four classes corresponding to each $\mathcal{H}$, and the remaining class representing the outliers, where none of the $\mathcal{H}$ is successful (success criterion is to be determined). Ideally, there will be distinct sets of bead-bead distances for each class, or higher-level features (e.g. distances and angles). Initial identification of those features could be achieved through the manual examination, followed by e.g. application of the pattern mining approaches.

Additionally, if a clear distinction is observed between the class of outliers and the other classes, it can serve as the foundation for a classifier to identify CS-bound fragments (i.e. HIPPO outliers) This classifier can be applied when using the dual scoring potential, particularly if Approach 1 is employed.

6.3.3 Pipeline for Iterative Docking

The current version of 'RRM-RNA dock' is tailored for the user-friendly docking of a single ssRNA fragment with two stacking nucleotides to a model of RRM. While this pipeline already outperforms *ab initio* docking, there are opportunities for significant expansion. First, it should be tested for fragments with a single stacking.

Another prospect involves leveraging further the knowledge that the anchored fragment forms a stacking interaction. One could introduce a stacking reward for poses exhibiting stacking (see §3.4.3, section Stacking Problem), thereby elevating their positions in the scoring list.

Subsequently, the pipeline should be expanded to an iterative docking of the full RNA. This should include restraints between the anchored fragment and its adjacent fragments, building upon the existing methodology.

A comparative analysis between scoring with ASF and HIPPO within the pipeline is essential. If HIPPO outperforms ASF, as anticipated, it should be implemented as the default scoring function in the pipeline. Additionally, leveraging RRMScorer to propose RNA sequences and identify potential nucleotides for stacking would enhance the pipeline. While several attempts for the latter have been made (see Appendix C.5), conclusive results are pending.

Expanding the pipeline to a more general data-driven protein-ssRNA tool, empowering users to integrate custom restraints based on experimental data, is a logical progression. However, it is imperative to keep the pipeline user-friendly, in this context through the assisted translation of the experimental data into docking restraints. To achieve this, thorough testing of non-structural data (as described in §5.4) has to be carried out.

Maintaining the pipeline's user-friendliness while allowing customisation for advanced users is crucial. Following this paradigm throughout development could enable the pipeline to encompass various aspects of ATTRACT functionality, making protein-ssRNA docking more accessible to the scientific community.

6.3.4 Other Perspectives

Several additional ideas to explore in the future are introduced in this subsection.

Expanding the Protein-ssRNA Benchmark: The protein-ssRNA benchmark, curated in the course of this work, encompassed only the first ssRNA sub-chain (comprising three or more consecutive protein-bound nucleotides). However, one, two, or even three suitable sub-chains are present in 117 complexes. All these sub-chains could be included in the benchmark as stand-alone complexes, or at least as data cases. The availability of such complexes, comprising several single-stranded and one or several double-stranded sub-chains, will allow to test docking of all sub-chains in isolation and perform simultaneous assembly of all poses, which may lead to more accurate docking results.

Additionally, around 100 protein-ssRNA complexes (the first ssRNA sub-chain only) within the benchmark, out of a total of 527 complexes, remain undocked due to time constraints. In both scenarios, the structures are prepared for docking, and the necessary pipeline is in place; thus, the only limitation is computational time. While redundancy may exist among some of these structures, they have the potential to significantly enrich the benchmark, with an anticipated increase of approximately 55% (the most optimistic estimate).

Stringent Cross-Validation for HIPPO: Initially, HIPPO underwent cross-validation using 29 test sets. These test sets were defined based on the similarity of RRM sequences in the data cases, with a threshold of over 40% sequence similarity. Nevertheless, RRMs are notorious for having highly similar structures, even when their sequence similarity is as low as 20%. Consequently, a more

stringent form of cross-validation, one that relies on the structural shape of the RRMs rather than their sequence identity, could be introduced. It is also possible to apply a leave-one-out procedure, however, in this case, a stricter criterion for the redundancy on the contact level should be introduced.

Extended HIPPO Application:

- The main work direction regarding HIPPO is training of a classifier to enable the use of BP beyond the training set case. Such a model could be obtained based on the sequence of the fragment and the sequence or/and structure of the protein, and/or on the docking poses. Enabling BP would enhance sampling significantly, as shown in §4.4;
- As mentioned before, HIPPO could be used for sampling using the ATTRACT docking engine and Monte-Carlo minimisation procedure. This is one of the most promising perspectives, as it might mitigate the sampling problem;
- During the derivation of HIPPO, the LRMSD threshold for near-native poses was relaxed from 3Å to 5Å to obtain more data cases with a higher count of near-natives. It is worth exploring whether a lower number of near-natives is sufficient for the derivation of an effective HIPPO;
- It would be interesting to test HIPPO's accuracy on the full ssRNA chains;
- Lastly, the protocol can be applied to the other types of complexes, such as protein-ssRNA beyond the RRM domain, protein-ssDNA and protein-peptides.

Flexibility Investigation: In the context of fragment-based docking, the flexibility of the ssRNA is taken into account via the fragment library and the coarse-grained representation; and the flexibility of the protein is taken into account only via the coarse-grained representation. It would be interesting to investigate which degree of flexibility on the protein side our fragment-based docking approach can handle. A possible approach would be to perform several independent rounds of docking using different bound protein structures, for instance given by NMR, and to determine if the docking performance is similar for the different structures. For such a comparison, one should use the metric 'fnat', i.e. fraction of the native contacts, rather than LRMSD.

This concludes the work presented in this thesis. The contributions and perspectives outlined herein, hopefully, will contribute to advancing the field of protein-ssRNA docking.

Appendix A

ATTRACT parameters optimisation

A.1 Monte Carlo Simulated Annealing Optimisation

A.1.1 Benchmark

42 complexes used for the MCSA optimisation:
1A9N 1B7F 1CVJ 2IX1 2L5D 2LI8 2XFM 2XZL 3BOY 3HSB 3PKM 3Q0M 3QJJ 3QJL 3R1H 3R2C 3SQW 3V6Y 3X1L 3ZLA 4B3G 4B8T 4BHH 4EI1 4F1N 4F3T 4H5P 4HT9 4IFD 4JNG 4JVH 4KRE 4KRF 4KXT 4M59 4MDX 4OE1 4OE1 4PJO 4PMW 5GAO 5JEA 5W1H

These 42 protein-ssRNA complexes constitute 308 fragments.

For this benchmark, a total of 7,435 near-native poses with LRMSD<3Å have been generated using the ATTRACT docking engine (‘randsearch’ with 30 million starting points) (Tab A.1).

As noted in §3.3.2.1 Dataset, this benchmark is unbalanced, with motifs UUU and AAA being over-represented (Fig. A.1). This pattern is also observed in the generated near-native poses (see Fig A.2).

Table A.1 - Benchmark and sampling statistics.

Motif	Number of fragments in the benchmark	Number of near-native poses (LRMSD<3Å)
UUU	88	2985
AAA	46	529
AUU	10	141
UAA	10	113
UUA	10	48
AUA	9	195
AAU	6	192
AGA	6	65
GUA	6	106
UAU	6	486
AGU	5	168
CUU	5	26
UUG	5	108
AAG	4	30
AGG	4	16
CAG	4	6
CAU	4	285
GAU	4	45
GUG	4	45
UAG	4	365
UGC	4	188
UGU	4	218
UUC	4	7

Motif	Number of fragments in the benchmark	Number of near-native poses (LRMSD<3Å)
AAC	3	3
ACA	3	27
ACU	3	25
GAG	3	9
GCA	3	21
GGU	3	46
GUU	3	15
UGA	3	0
ACC	2	0
CCA	2	118
CUC	2	6
CUG	2	4
GAA	2	3
GCU	2	17
UAC	2	1
UCA	2	10
UCU	2	3
AUC	1	121
CAA	1	0
CAC	1	42
CCG	1	0
CCU	1	11
CGA	1	1
CUA	1	10
GAC	1	0
GCC	1	447
GGA	1	2
GGC	1	4
GGG	1	16
UCC	1	106
UGG	1	-
AUG	0	-
ACG	0	-
AGC	0	-
UCG	0	-
CCC	0	-
CGU	0	-
CGC	0	-
CGG	0	-
GUC	0	-
GCG	0	-

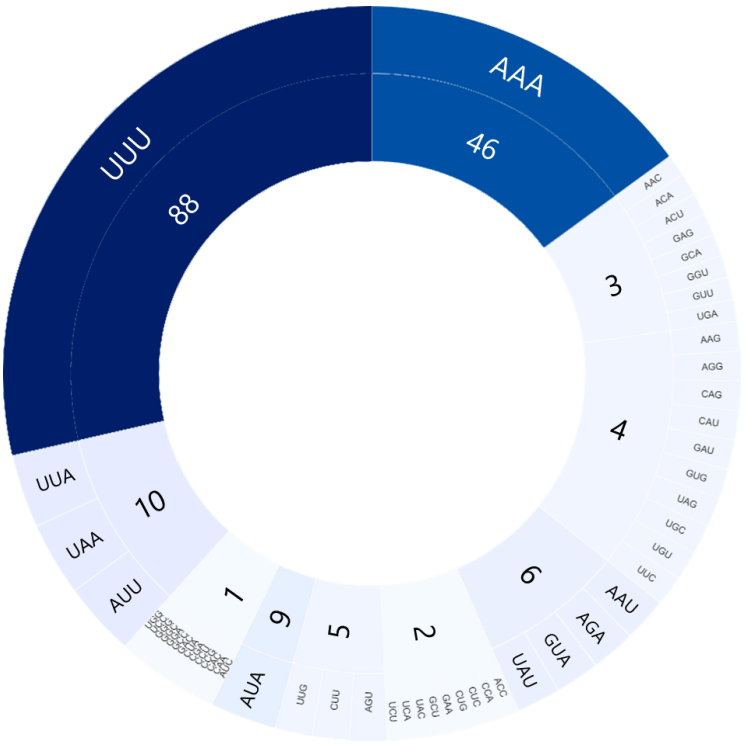


Figure A.1 - Visualisation of the number fragments of each distinct motif in the benchmark

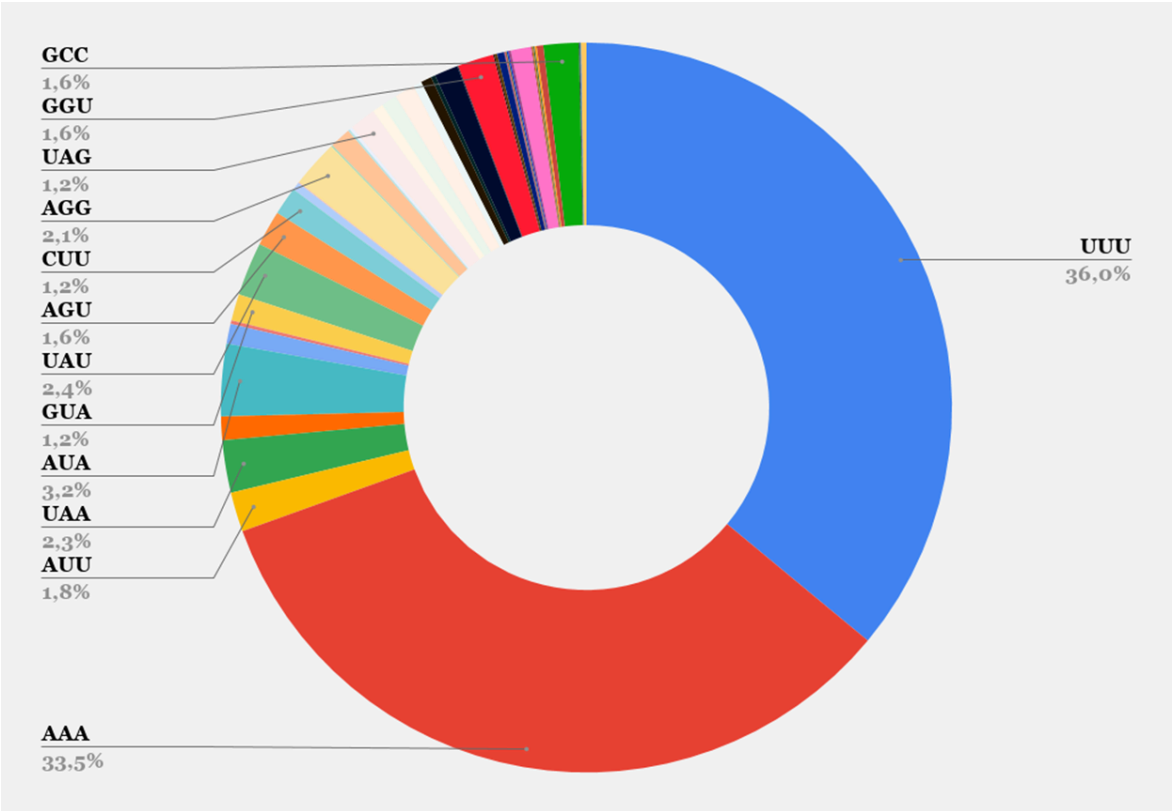
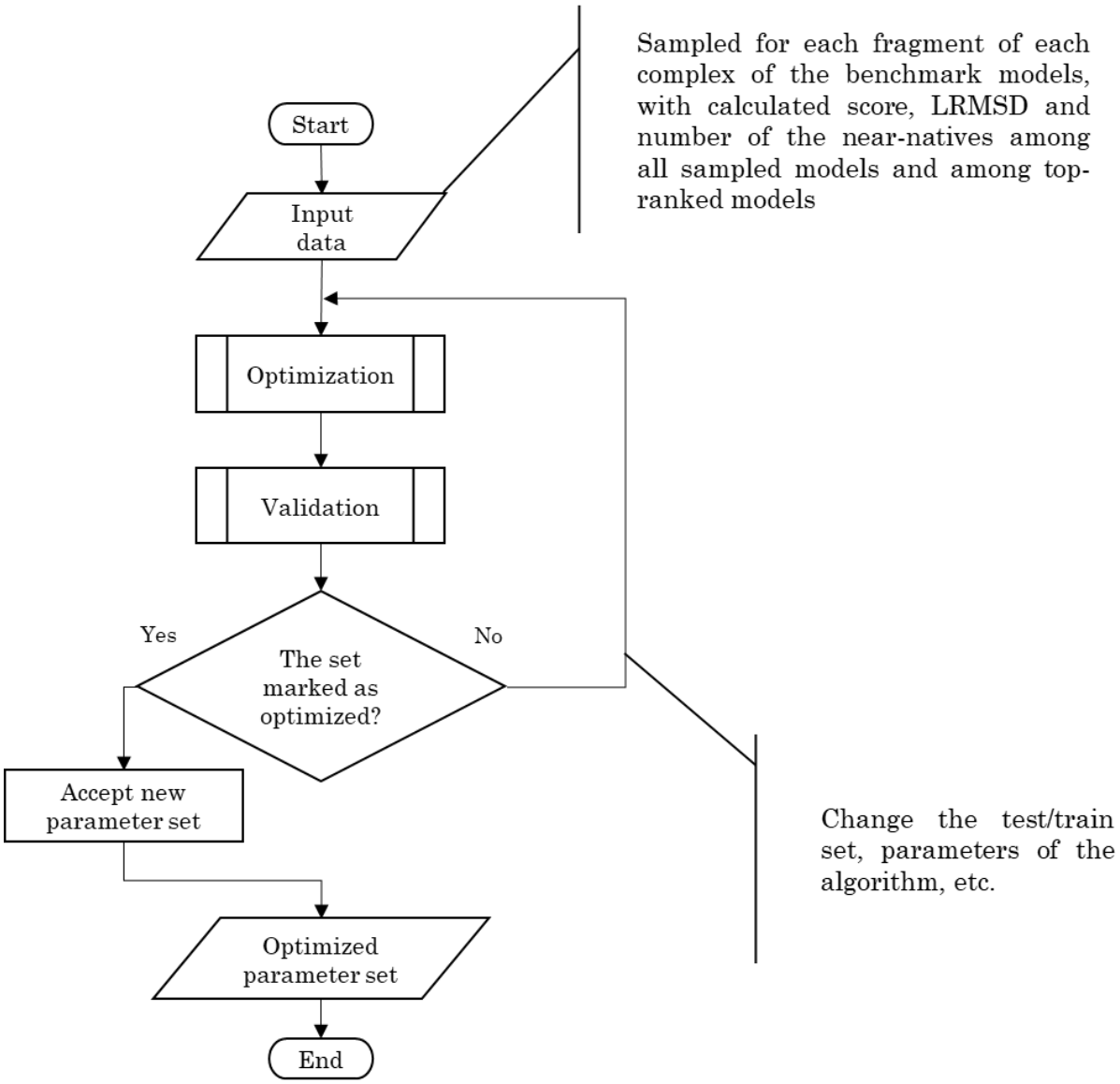


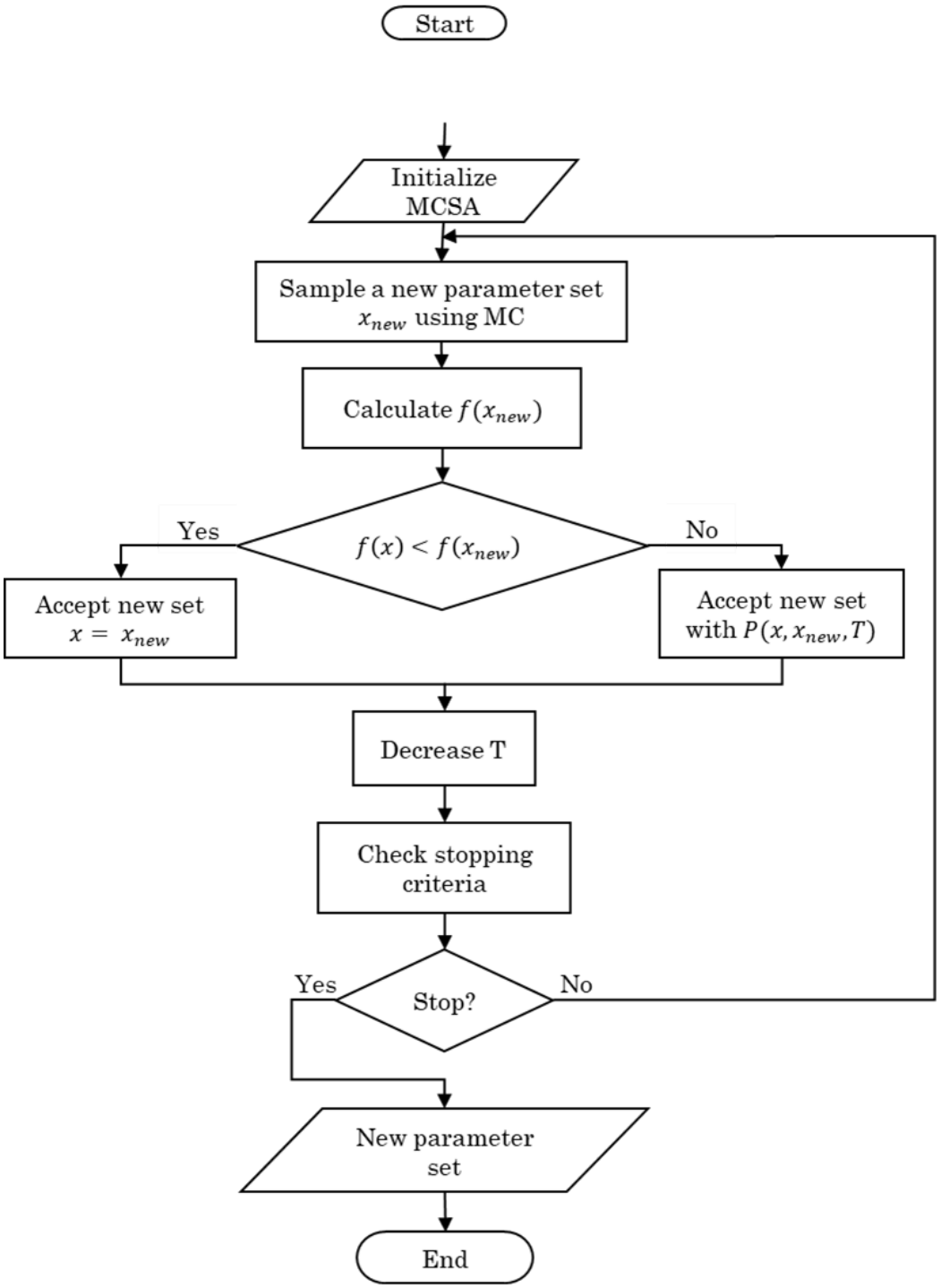
Figure A.2 - Visualisation of the number of near-native poses for each distinct motif in the benchmark

A.1.2 MCSA Flowcharts

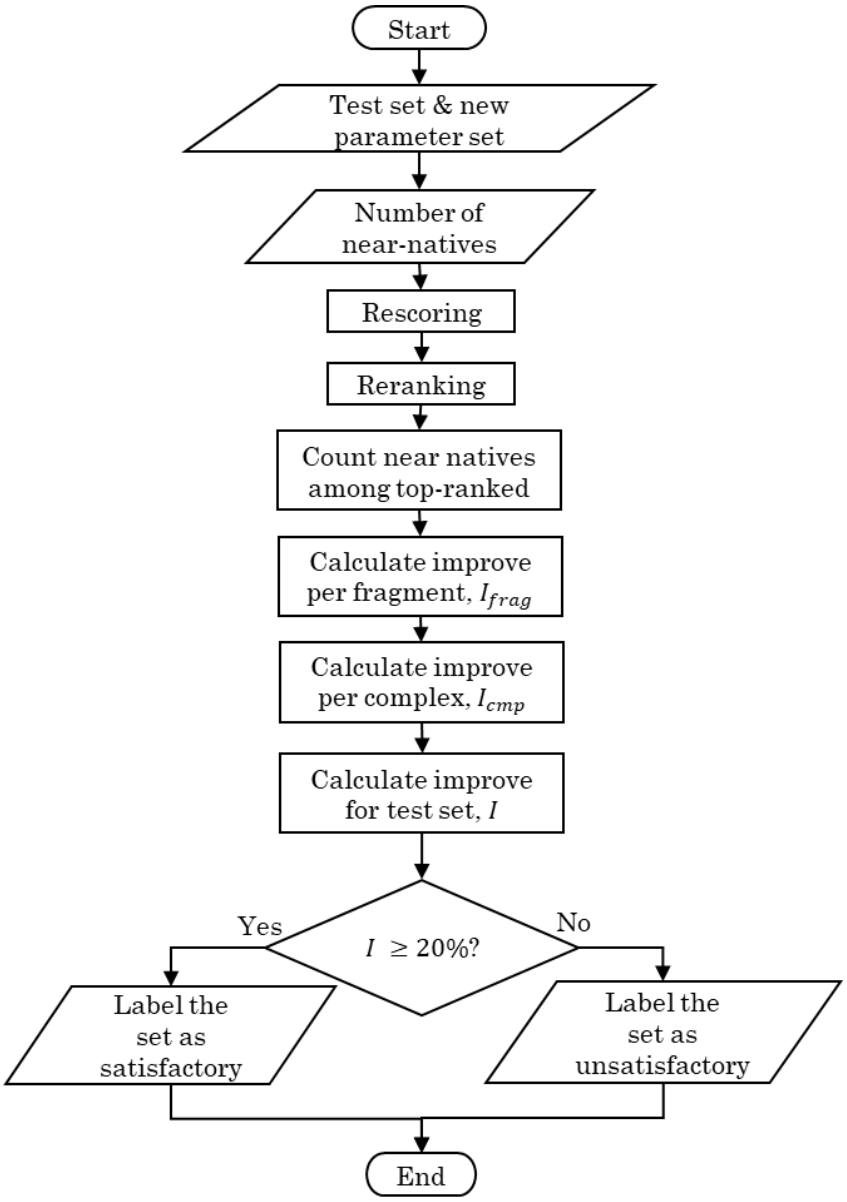
Main algorithm



Optimisation algorithm



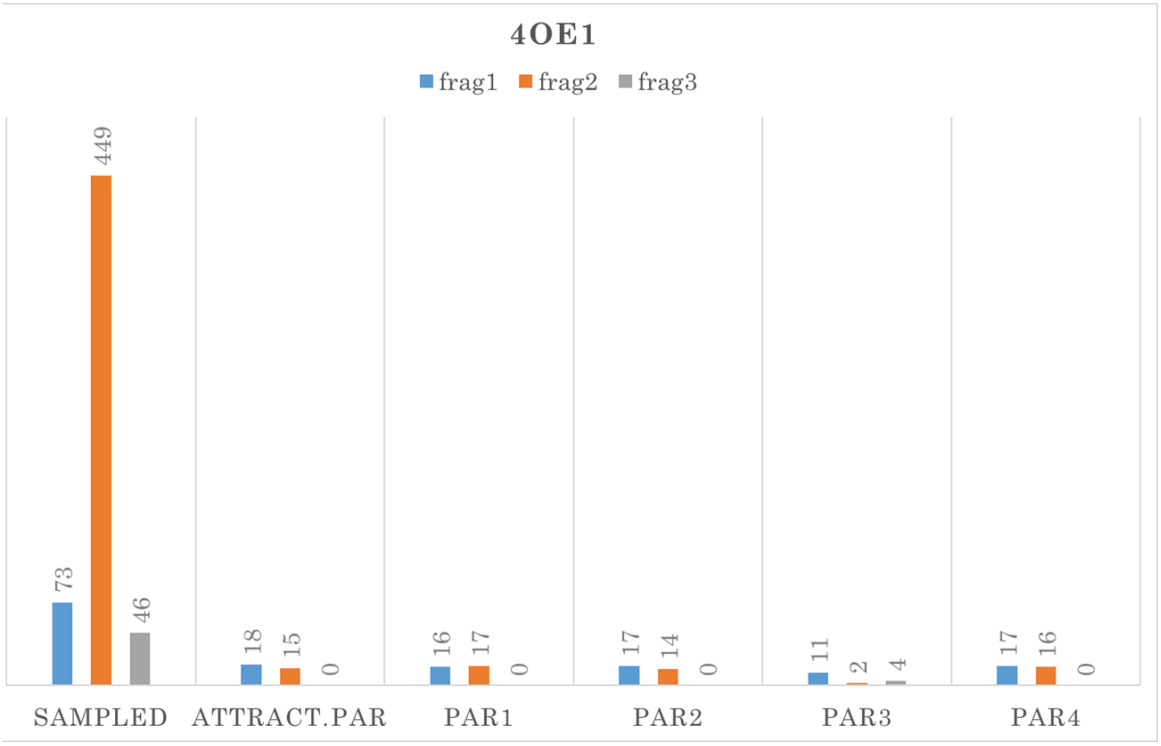
Validation algorithm

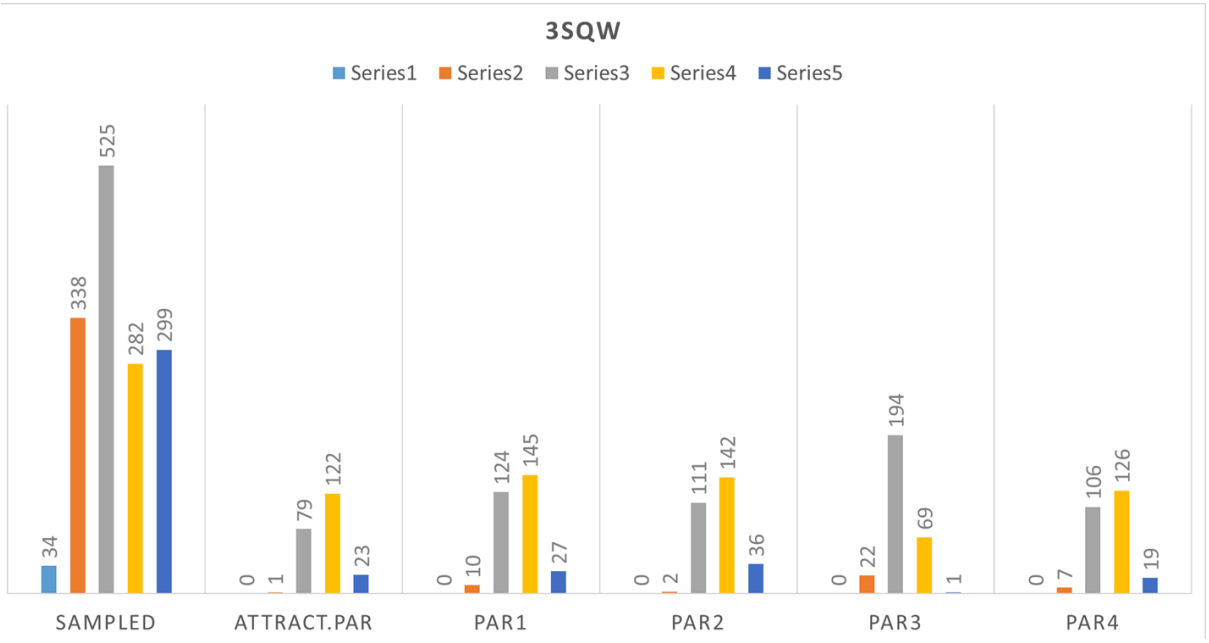


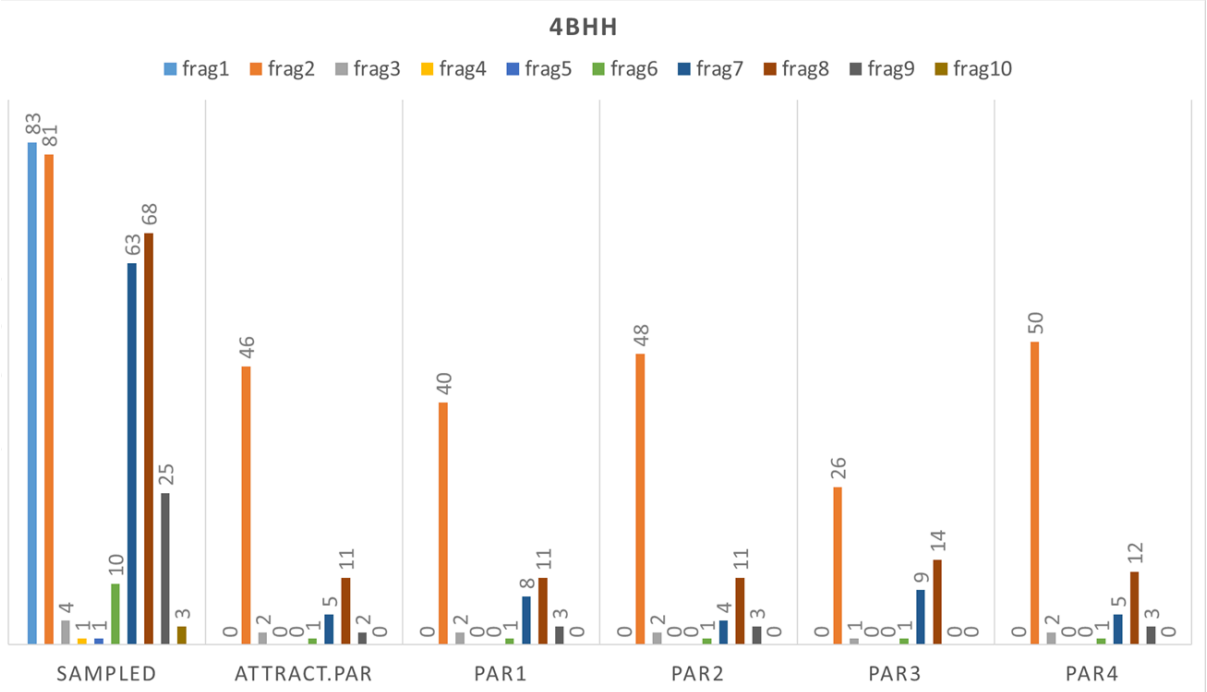
A.1.3 MCSA Results

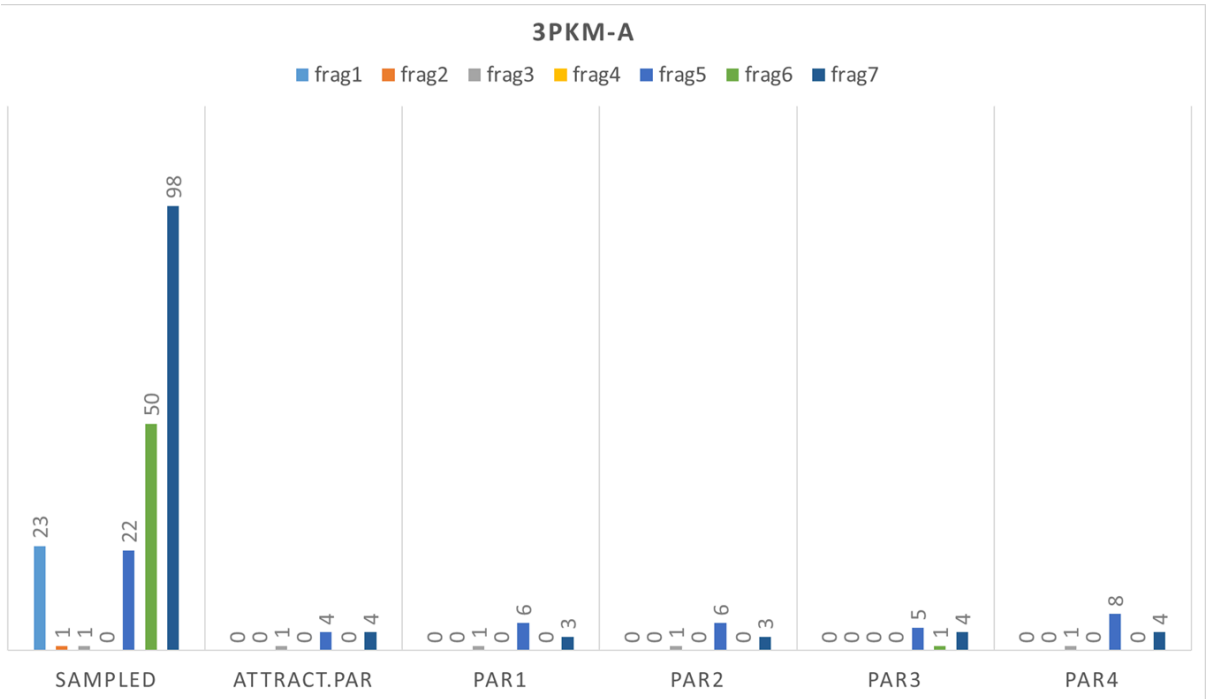
Each graph displays the number of near-natives within a complex. The first group of bars on each graph displays the total number of sampled near-natives per fragment, and each subsequent group of bars displays the number of near-natives in the 10% of top-ranked poses, i.e. in 1 million top-ranked.

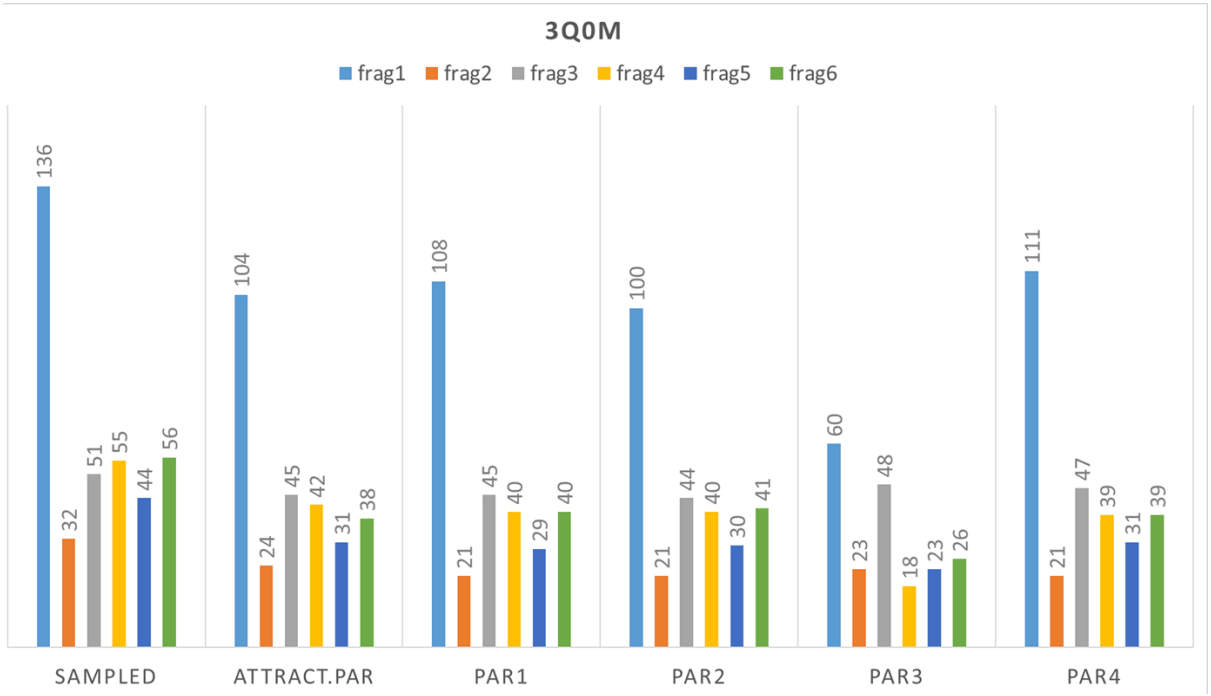
PAR1, PAR2, PAR3 and PAR4 are the sets of the docking parameters obtained using MCSA with the corresponding hyperparameters, presented in §3.3.3, Tab. A.3.

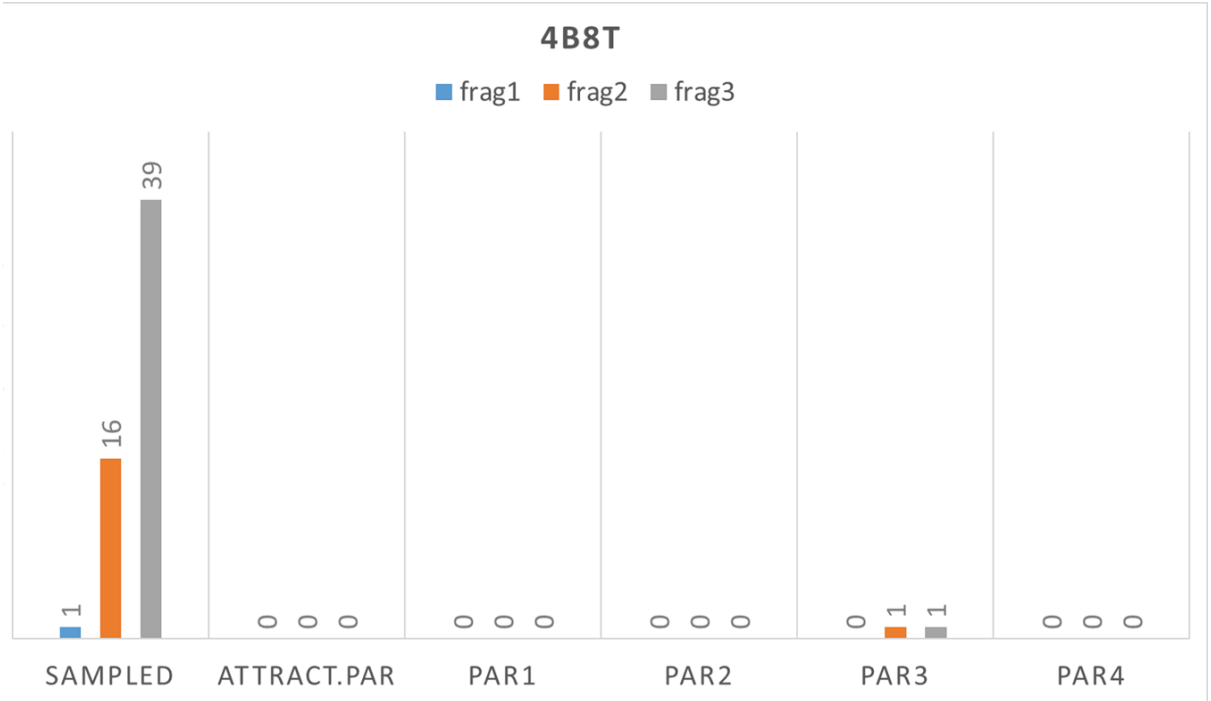





4BHH
frag1 frag2 frag3 frag4 frag5 frag6 frag7 frag8 frag9 frag10
83 81 4 1 1 10 63 68 25 3
0 46 2 0 0 1 5 11 2 0
0 40 2 0 0 1 8 11 3 0
0 48 2 0 0 1 4 11 3 0
0 26 1 0 0 1 9 14 0 0
0 50 2 0 0 1 5 12 3 0
SAMPLED
ATTRACT.PAR
PAR1
PAR2
PAR3
PAR4


3PKM-A
frag1 frag2 frag3 frag4 frag5 frag6 frag7
23 1 1 0 22 50 98
0 0 0 1 0 4 0 4
0 0 0 1 0 6 0 3
0 0 0 1 0 6 0 3
0 0 0 0 0 5 1 4
0 0 0 1 0 8 0 4
SAMPLED
ATTRACT.PAR
PAR1
PAR2
PAR3
PAR4


3Q0M
frag1
frag2
frag3
frag4
frag5
frag6
136
32
51
55
44
56
104
24
45
42
31
38
108
21
45
40
29
40
100
21
44
40
30
41
60
23
48
18
23
26
111
21
47
39
31
39
SAMPLED
ATTRACT.PAR
PAR1
PAR2
PAR3
PAR4


4B8T
frag1
frag2
frag3
1
16
39
0
0
0
0
0
0
0
0
0
0
1
1
0
0
0
SAMPLED
ATTRACT.PAR
PAR1
PAR2
PAR3
PAR4

A.2 TRP-C Fine-Tuning

This section holds the original ATTRACT parameter values for TRP-C side chain beads, as well as 4 updated TRP-C subsets (Tab A.2). The set of 7 complexes was initially docked with original ATTRACT parameters and subsequently re-docked with ‘par1’ and ‘par2’. The results are in Tab A.3.

Table A.2 - Values of the original and updated parameters for TRP-C side chain beads.

$i\ j$	**Param. name**	$\sigma_{ij}\ \varepsilon_{ij}$
25 12	**original**	**3.87 8.13**
	par1	3.87 13.46
	par2	3.87 8.13
	par3	4.63 3.25
	par4	0.01 0.01
25 13	**original**	**6.40 6.28**
	par1	3.72 15.00
	par2	3.72 15.00
	par3	3.80 4.70
	par4	0.01 0.01
25 14	**original**	**3.77 12.33**
	par1	3.08 18.8
	par2	3.77 12.33
	par3	4.64 3.07
	par4	0.01 0.01
26 12	**original**	**3.00 8.97**
	par1	3.07 13.16
	par2	3.0 8.97
	par3	3.70 2.03
	par4	0.01 0.01
26 13	**original**	**3.75 15.43**
	par1	3.32 15.43
	par2	3.75 15.43
	par3	4.83 1.68

$i\ j$	**Param. name**	$\sigma_{ij}\ \varepsilon_{ij}$
	par4	0.01 0.01
26 14	**original**	**3.80 15.27**
	par1	3.80 15.27
	par2	3.80 15.27
	par3	4.03 3.35
	par4	0.01 0.01

Table A.3 Docking results for a small subset of fragments, obtained with original attract parameters, 'par2' and 'par1' (see Tab. A.2).
The target fragment (with TRP-C pair being closest to each other in the given complex) is marked with *.

	attract	**par2**	**par1**	**attract**	**par2**	**par1**	**attract**	**par2**	**par1**
2CSX									
	fragment 1			fragment 2*			fragment 3		
<2Å	0	0	0	0	0	0	15	16	14
<3Å	0	0	0	2	2	1	77	69	77
<4Å	2	0	0	8	7	5	336	363	347
2CT8									
	fragment 1			fragment 2*			fragment 3		
<2Å	0	0	0	0	0	0	15	15	14
<3Å	0	0	0	1	0	0	71	70	76
<4Å	1	3	4	10	10	7	267	275	276
5YTS									
	fragment 1			fragment 2*					
<2Å	74	74	73	7	5	10			
<3Å	734	815	808	253	262	291			
<4Å	4560	5401	5404	4082	4344	4654			
6A6J									
	fragment 1			fragment 2			fragment 3*		
<2Å	2	2	3	0	0	0	0	1	0

	attract	**par2**	**par1**	**attract**	**par2**	**par1**	**attract**	**par2**	**par1**
<3Å	157	158	160	70	69	93	169	178	280
<4Å	1233	1248	1411	1746	1746	1941	1420	1535	1808
3ADC									
	fragment 1			fragment 2*					
<2Å	0	0	0	0	0	0			
<3Å	1	1	0	0	0	0			
<4Å	3	2	0	6	5	0			
2HGH									
	fragment 1*								
<2Å	0	0	0						
<3Å	8	8	6						
<4Å	47	44	51						
2FMT									
	fragment 1			fragment 2*			fragment 3		
<2Å	0	0	0	7	8	8	237	263	225
<3Å	19	24	23	143	180	157	2089	2245	2368
<4Å	143	157	180	981	1011	1169	9420	9038	11914

Appendix B

Histogram-based Pseudo Potential Related

B.1 Original HIPPO Paper

HIPPO: HIstogram-based Pseudo-POtential for scoring protein-ssRNA fragment-based docking poses

Anna Kravchenko[1], Sjoerd Jacob de Vries[1], Malika Smaïl-Tabbone[1] and Isaure Chauvot de Beauchene[1,*]

[1] Université de Lorraine, CNRS, Inria, LORIA F-54000 Nancy, France

Abstract

Motivation: The RNA-Recognition motif (RRM) is a protein domain that binds single-stranded RNA (ssRNA) and is present in as much as 2% of the human genome. Despite this important role in biology, RRM-ssRNA interactions are very challenging to study on the structural level because of the remarkable flexibility of ssRNA. In the absence of atomic-level experimental data, the only method able to predict the 3D structure of protein-ssRNA complexes with any degree of accuracy is ssRNA'TTRACT, an ssRNA fragment-based docking approach using ATTRACT. However, this approach has limitations, such as the production of only a handful of near-native poses amid many non-natives, and the frequent failure of the ATTRACT scoring function (ASF) to recognize these near-natives. Nevertheless, since ASF parameters are not ssRNA-specific and were determined in 2010, there is substantial opportunity for enhancement.

Results: Here we present HIPPO, a composite RRM-ssRNA scoring potential derived analytically from contact frequencies in near-native versus non-native docking models. Validated on a fragment-based docking benchmark of 57 experimentally solved RRM-ssRNA complexes, HIPPO achieved a 3-fold or higher enrichment for half of the fragments, versus only a quarter with ASF. In particular, HIPPO drastically improved the chance of very high enrichment (12-fold or higher), a scenario where the incremental modelling of entire ssRNA chains from fragments becomes viable. However, for the latter result, more research is needed to make it directly practically applicable. Regardless, our approach already improves upon the state of the art in RRM-ssRNA modelling and is in principle extendable to other types of protein-nucleic acid interactions.

Keywords: scoring function, protein-ssRNA docking, RRM-ssRNA docking, fragment-based docking

1 Introduction

Protein-RNA complexes play an immensely important role in many cellular processes, including translation, transcription, and post-transcriptional gene expression [1]. The disruption of the binding can lead to tremendous cellular malfunctions [2]. A large part of these protein-RNA interactions involves one of the few conserved RNA-binding domains. In particular, over 50% of all RNA-binding proteins in humans contain an RNA recognition motif (RRM) [3]. This motif is critical for binding to RNA molecules, and to single-stranded RNAs (ssRNA) specifically, making RRM-ssRNA interactions crucial for understanding the underlying mechanisms of various cellular processes.

Although the 3D structure of these complexes provides valuable insights into their functions, the experimental resolution of such structures is a non-trivial task. Computational modelling of the 3D structure of a

protein-RNA complex, also known as protein-RNA docking, can facilitate experimental research, by proposing probable 3D structures to be experimentally tested.

Unfortunately, protein-ssRNA docking is a challenging task by itself as well. The classical docking approaches [4] require an unbound structure as a starting point, but no such structure is available for ssRNA due to its disorder in the unbound state. On the one hand, one may try to model all possible ssRNA conformations using its sequence, and then dock them. However, ssRNA's flexibility (~8 DOF per nucleotide [5]) makes systematic modelling of ssRNA conformations extremely demanding computationally and borderline impossible for long chains. On the other hand, in recent years, various powerful deep learning techniques ([6,7,8]) brought breakthroughs to protein-protein [9] and protein-ligand [10,11] docking. However, deep learning approaches are more challenging to apply to protein-RNA docking, not only due to the relatively low number of solved structures (about $1.16 \cdot 10^4$ protein-RNA structures compared to about $1.776 \cdot 10^5$ protein chains) but also because among all atomic contacts within each structure, the interaction between RNA and protein represents only a tiny fraction. This is even more true for ssRNA, which is only a small subset of RNA, and whose binding modes to proteins have some particularities compared to double-stranded (ds) RNA [12].

Fragment-based docking handles ssRNA flexibility by subdividing its sequence into fragments that are small enough for their conformations to be exhaustively (including close-to-bound conformation) sampled within a given accuracy threshold. The docking procedure consists of sampling and scoring. Sampling refers to the generation of docking *poses* - certain positions and orientations of particular conformations of the fragment with respect to the protein. A pool of docking poses is sampled for each fragment independently. Scoring is the evaluation of the probability of each pose being a near-native, followed by ranking. Finally, the presumably best poses of adjacent fragments are assembled into complete structures called docking *models*. In a test case, when the native structure of a complex is experimentally determined, both docking poses and models can be assessed based on their similarity to the corresponding parts of a native structure, and this similarity can be quantified by their ligand root mean squared deviation (LRMSD). The distinction is made between near-native (correct), non-native (incorrect), and intermediate poses/models based on LRMSD thresholds.

The main limitation of the fragment-based strategy stems from the concept of hot- [13] and coldspot binding. A fragment by itself (taken in isolation) may have much stronger binding and hence lower real interaction energy in a region of the protein that is different from the binding region of that fragment when it is in the chain. This is a case of coldspot binding. The term "coldspot" refers to an area of the protein surface that can bind fragments relatively weakly. The opposite term, "hotspot", refers to the part of the protein surface that binds fragments relatively strongly. Essentially, fragments that bind to the coldspots are only there because the adjacent fragments are tightly bound to the hotspots. From an energy perspective, binding to the coldspot leads to a shallow local energy minimum, whereas binding to the hotspot leads to a deeper (and possibly global) energy minimum. A mononucleotide tandem repeat sequence, such as the poly-U chain, provides a very intuitive example. For such an ssRNA, there are multiple overlapping native solutions for the same fragment sequence UUU that "compete" to be sampled and scored during the docking of UUU. As a consequence, there are usually one or two well-docked fragments, i.e. fragments with a lot of correctly ranked near-native poses, while the docking results for the remaining fragments are much worse.

The described hot/coldspot limitation directly contributes to the so-called sampling problem. The sampling problem lies in the fact that often not a single near-native pose is generated during the docking run. The sampling problem is critical because it has a high impact on the whole docking procedure: for successful docking of the whole RNA chain, at least one near-native pose must be sampled for each of the fragments. Otherwise, the docking for a given complex will certainly fail at the assembly step.

Another limitation is the scoring problem, which arises when none of the sampled near-natives is selected in the list of top-ranked poses. In this case, more poses per fragment must be retained to have a good chance to keep a near-native, which quickly becomes very expensive computationally in the assembly step. In turn, as there are more docking models, identification of the near-native model also becomes more challenging.

There are four existing fragment-based approaches for protein-ssRNA docking: RNA-LIM, FBDRNA, RNP-denovo, and ssRNA'TTRACT. RNA-LIM represents each nucleotide by one non-oriented bead and could only predict their position at 15Å resolution for one example [14]. FBDRNA uses mononucleotide fragments in all-atom representation, docked with MCSS on a pre-defined binding site. While showing discriminative power on nucleotides' positions, it could not provide accurate models for full oligonucleotides [15]. RNP-denovo, a

Rosetta method to simultaneously fold-and-dock RNA to a protein surface, uses the exact position of a few nucleotides [16], which would be unavailable for real-life docking cases. On the other hand, **ssRNA'TTRACT**, the state of the art, is the most accurate approach that uses only a protein structure and the RNA sequence as input. It uses trinucleotides as RNA fragments and an overlapping criterion based on LRMSD for assembly. Furthermore, when information about conserved protein-RNA contacts are available, ssRNA'TTRACT employs an anchored docking strategy to build the RNA chain incrementally by docking one fragment with contact restraints and using each of its top-ranked poses as an anchor to superimpose subsequent fragments [17]. This strategy tackles the sampling problem for the fragments.

ssRNA'TTRACT uses the **ATTRACT** docking engine and a library of RNA trinucleotide conformations developed in our research group [18,19]. A coarse-grained force field with Lennard-Jones type energy function with soft potential [20] is used for both sampling and scoring. In the coarse-grained representation, the RNA fragments and the protein are represented as sets of pseudo-atoms, called *beads*, each of which stands for a small group of real atoms. Coarse-grained representation provides several advantages compared to all-atom representations. First, it accounts for inaccuracies in atomic positions coming either from bound/unbound conformational differences or experimental biases and resolution; second, it smoothes the energy landscape, which prevents the poses from getting stuck in shallow local minima; and third, it reduces the computation time.

Despite its capabilities, ssRNA'TTRACT is still constrained by the aforementioned limitations. As the current ATTRACT protein-RNA scoring function was not designed to tackle ssRNAs specifically and its parameters were optimised back in 2010 on dsRNA alone, there is considerable potential for enhancement. Here we present HIstogram-based Pseudo-POtential (HIPPO), which aims to distinguish between near-native and non-native protein-ssRNA docking poses. HIPPO is based on the hypothesis that there exists a collection of scoring parameter sets (as opposed to a single parameter set) that can be used to effectively rank near-native protein-ssRNA docking solutions. HIPPO's parameters are derived analytically from contact frequencies in near-native versus non-native docking poses. These contact frequencies, derived from 4 different sets of docking poses, are discretised by a particular set of cutoffs into histograms, leading to a collection of 4 histogram sets $\mathcal{H}$ that together form the HIPPO scoring potential. Thus, HIPPO is a composite protein-ssRNA scoring potential: typically, the top 5% of the poses according to each histogram set are combined, selecting 20% of all docking poses in total. To streamline the process from dataset construction to the generation of final scoring parameters, we decided to focus exclusively on the RRMs, as this domain of the protein is particularly important for studying protein-ssRNA interactions and is present in many (approximately 65%) of the available protein-ssRNA structures. This allows us to provide proof of principle that the scoring function can indeed be improved using our method. However, the developed method and protocol can be applied to a wider benchmark, and more importantly, to other types of protein-nucleic acid interactions in the future.

HIPPO was derived from a fragment-based docking benchmark of 57 experimentally solved RRM-ssRNA complexes, corresponding to 217 overlapping ssRNA trinucleotide fragments in complexes with an RRM. Using cross-validation, HIPPO achieved a 3-fold enrichment (60% of all near-native poses in the 20% top-ranked poses) for 53% of the fragments, versus only 26% with the current state-of-the-art ATTRACT scoring function (ASF). In addition, these near-native poses were often selected mostly by a single $\mathcal{H}$ of the 4 histogram sets. Consequently, using the hypothetical knowledge of the best HIPPO histogram yielded a 12-fold enrichment for nearly 40% of the test fragments - something which is achieved with ASF in only 4% of the cases. Most importantly, 61% of the complexes show such a 12-fold enrichment for at least one fragment. Under these conditions, the incremental modelling of entire ssRNA chains from best-docked fragments becomes viable. However, the problems of blindly identifying the best HIPPO histogram set and selecting the best-docked fragments need to be solved first before this can become practical. Nevertheless, as it is, HIPPO already improves upon the state of the art in RRM-ssRNA modelling.

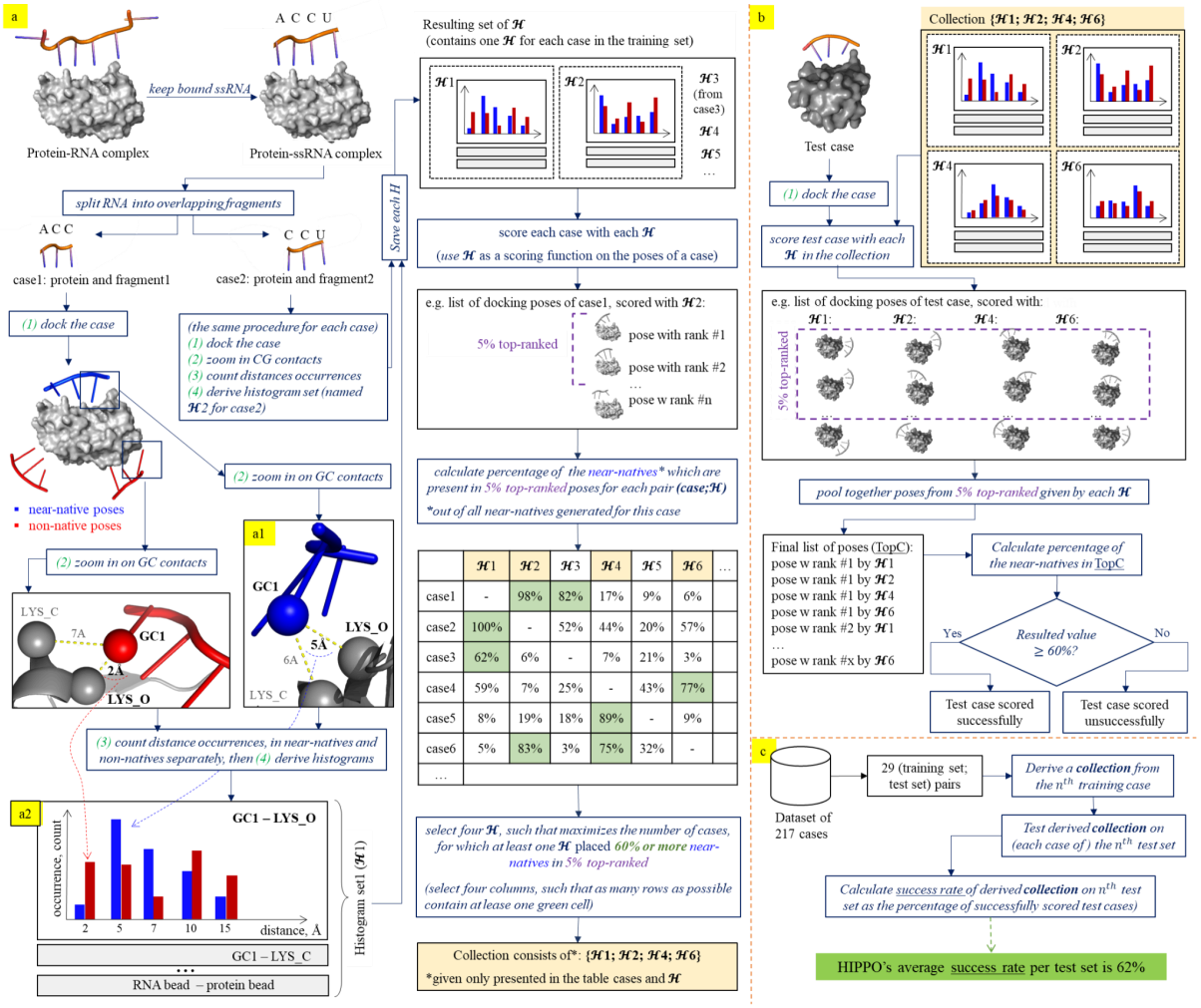


Fig 1. a) Graphical pipeline for building HIPPO as a collection of four histogram sets ($\mathcal{H}$). a1) Contacts between bead GC1 (in Cytosine side-chain) and bead LYS_O and LYS_C (Lysine backbone). a2) An intuitive schema of $\mathcal{H}$. The histogram for beads (GC1; LYS_O) is shown as an expanded plot. The blue dashed lines from a1 to a2 show the contribution of the contact to the histogram. The blue/red bars show the count of occurrences of distances in all near-native/non-native poses. The other histograms in this set $\mathcal{H}$, for other pairs of beads, are not shown (collapsed). b) Graphical pipeline for testing a collection on a test case. c) Graphical pipeline for the complete workflow. The creation of pairs of training and test sets is based on the protein's sequence similarity: proteins with sequence similarity of 40% or higher are never present in both training and test sets.

2. System and methods

Here we first present the dataset that we built and used for the training and validation of HIPPO. Next, we present step-by-step the process of constructing a set of scoring parameters in the form of a histogram set $\mathcal{H}$ and the process of building the final collection of several $\mathcal{H}$ (Fig 1).

2.1 Data

2.1.1 RRM-ssRNA benchmark

The number of experimentally solved protein-ssRNA structures is considerably low compared to protein-protein structures. We gathered all available data and built an up-to-date benchmark of experimental 3D structures of RRM-ssRNA complexes from the Protein Data Bank (PDB) by (i) downloading all experimentally solved (either NMR or X-RAY with resolution 3Å or higher) protein-RNA complexes and (ii) applying ProtNAff in order to retrieve complexes with 3 or more consecutive protein-bound single-stranded nucleotides.

We considered a nucleotide to be protein-bound if at least 5 pairs of RRM-RNA heavy atoms were located within 6Å from each other. Lastly, we filtered out complexes whose protein does not contain any RRM domain, according to the InteR3M database [21]. The resulting benchmark consists of 81 RRM-ssRNA complexes, released before February 2021.

2.1.2 Dataset of docking poses

From the benchmark, we created a dataset of labelled docking poses. We used the ATTRACT docking engine and library of RNA trinucleotide conformations [22] to dock each entry (each RRM-ssRNA complex) of the benchmark, by docking each overlapping trinucleotide fragment (e.g. chain AUCG => fragment AUC and fragment UCG), following the procedure described in [23]. For each fragment, a randomly selected conformation from ProtNAff was placed at each of $3 \cdot 10^7$ predefined starting points located within 30Å from the center of mass of the bound and rigid protein, with a random 3D rotation. Then the position of each starting pose was minimised using gradient descent. Redundant poses (RMSD<0.2Å) were filtered out of the resulting pool before scoring. The remaining docking poses were scored, and the 10^7 top-ranked poses were retained. Each pose was labelled as near-native if its LRMSD was under 5Å; as non-native if its LRMSD was over 7Å; as intermediate otherwise.

We used such relatively soft thresholds to lower the number of cases for which the sampling problem (zero near-native poses sampled) has arisen. For example, the more strict thresholds [3Å;5Å] resulted in 41% of cases with the sampling problem, versus just 8% with [5Å; 7Å]. To minimise the noise in the dataset, 60 cases where the number of sampled near-natives was less than 100 were excluded. This led to a set of 419 RRM-trinucleotide fragment docking cases. Note that in the case of multiple fragments with the same sequence bound to the same RRM, only a single docking is necessary.

2.1.3. Coarse-grained representation

As mentioned before, in the coarse-grained representation, groups of atoms are represented by beads. In the used representation, 31 bead types are used to represent proteins (2 for backbone and 0-2 for side chain) and 17 bead types are used to represent RNA (1 for phosphate group, 2 for sugar and 3-4 for base), leading to a maximum of 527 pairs of bead types [20]. Protein beads are denoted by index i and RNA beads are denoted by index j.

2.1.4 Redundancy

In order to eliminate possible dataset bias, we performed a redundancy check at the contact level, by comparing i-bead to j-bead distances within 6Å in the native poses of the protein-fragment cases. If such distance sets were very similar for two cases, these cases were considered redundant, and one of them was removed from the dataset. The final dataset consists of **217 RRM-fragment cases**, with 10^7 labelled docking poses per case. Its corresponding benchmark consists of **57 RRM-ssRNA complexes** and can be found in Additional file 1: Table S1.

2.1.5 Training and test sets

We separated the dataset into pairs of training and test sets based on protein sequence similarity, in a leave-homology-out procedure. Our sequence similarity threshold was 40%. We selected a random protein-ssRNA complex from the benchmark along with all other complexes whose protein sequence similarity was greater than 40%. All data cases derived from these complexes (protein-fragment structures along with their

docking poses) became the test set. The remaining data cases formed the corresponding training set. We repeated this procedure iteratively until each of the benchmark complexes was in one of the test sets. To prohibit repetitive and near-repetitive (training; test) pairs, we ensured that the first randomly selected case in each iteration did not belong to any of the previous test sets. All statistics reported in this paper correspond to the evaluation of HIPPO on the test sets, where for each test set the four histogram sets $\mathcal{H}$ derived from the corresponding training set were used. The final collection consists of 29 (training; test) pairs and can be found in Additional file 1: Table S2.

2.2 Creation of histogram set $\mathcal{H}$

The main steps - detailed thereafter - to obtain a scoring histogram set $\mathcal{H}$ are as follows:

1) construction of the *distance arrays* containing the number of occurrences of each bead-bead distance, in near-native vs in non-native poses (ignoring intermediate ones), for each pair of bead types $(i; j)$ independently;

2) refinement of the distance arrays to ensure that each of them provides sufficient signal;

3) derivation of $\mathcal{H}$ from the distance arrays, one histogram per distance array.

2.2.1 Histogram definition

Let's denote the bead types representing the protein by index $i \in \{1, 2, \dots 31\}$, and the bead types representing the RNA by index $j \in \{1, \dots 17\}$. Also let's define initial distance ranges by applying discretisations of 0.25Å and 1.5Å to the intervals [2Å; 7Å] and [7Å; 14.5Å] respectively. Such design of distance ranges allows to capture close-range interactions with high precision and to generalise long-range interactions. The resulting set contains 27 ranges: {(0, 2], (2, 2.25], …,(14.5, 999)}.

A distance array D_{ij} with the dimension 27x2 is designed to capture the number of occurrences of all $(i; j)$ distances within a pool of docking poses. The rows $d_{k'}, k = 1 \dots 27$, of D_{ij} correspond to the distance ranges. Each element of D_{ij} contains the count of distances within the indicated range. Elements d_{k1} in the first column account for the distances in near-native poses only, while elements d_{k2} from the second column capture distances in non-native poses.

To ensure that in each D_{ij} there are enough examples coming from near-native poses in each distance range to provide a sufficient signal, we set a threshold w for a minimum number of occurrences in near-natives d_{k1}. The threshold value is empirical and is determined individually for each $(i; j)$ pair as 1/60 of all distances counted in near-native poses:

$$w_{ij} = A_{ij}/60,$$

$$where\ A_{ij} = \sum_k d_{k1},\ \forall d_{k1} \in D_{ij}.$$

For each D_{ij}, if $d_{k1} < w_{ij}$, then the rows starting from k^{th} and beneath are summed until their sum exceeds the threshold. The new row resulting from the summation replaces the original row. This process is repeated until all values in the first column of the resulting array exceed the threshold. The resulting *refined distance array* D^*_{ij} has dimension qx2, where $q \leq 27$, and may vary for different $(i; j)$ pairs. Note that for each $(i; j)$ we must save the resulting set of refined distance ranges for further application of the histogram.

Finally, the following formula, inspired by the logarithm of the odds ratio, is used to obtain individual histograms H_{ij} from the corresponding D^*_{ij}:

$$H_{ij} = \left[ln\, d^*_{x1} - ln\, d^*_{x2} - (ln\, A_{ij} - ln\, B_{ij}) \right],$$

where $x = 1 \dots q,\ \forall x\ [d^*_{x1},\ d^*_{x2}\ ] \in D^*_{ij},\ B_{ij} = \sum_k d_{k2}, \forall d_{k2} \in D_{ij}.$

The dimension of H_{ij} is qx1. We define $\mathcal{H}$ as the set of individual histograms H_{ij} for all $(i; j)$ pairs, which are present in at least one pose out of the input pool of the docking poses.

Since 10^7 poses is a rather large pool, poses with vastly different ranks could possess different features. To account for this possibility, we divided the initial pool of poses into 3 sub-pools according to the rank of the

poses: [0, 99999], [10^5, 999999], [10^6,10^7]. Each D_{ij} and subsequently each H_{ij} consists of three parts, built on poses from the corresponding rank-based sub-pool.

2.2.2. Scoring with $\mathcal{H}$ and scoring assessment

To score a pose using $\mathcal{H}$, we count the occurrences of distances for each $(i; j)$ pair within each of the refined ranges, within each rank-based sub-pool. This information is stored in a qx1 array R_{ij}. The histogram-based score of a pose is calculated using the following formula:

$$S_{pose} = \sum_i \sum_j R_{ij} \cdot H_{ij}^T \quad (1)$$

In simpler terms, for every bead-bead distance in a pose that falls in one of the refined ranges, a corresponding sub-score is assigned. This process is repeated for each rank-based sub-pool separately. The sum of all sub-scores is the final histogram-based score of a pose.

To evaluate the performance of $\mathcal{H}$ for a data case, we score all docking poses from the pool of 10^7 poses using formula (1) and rank the poses by their score in descending order. Then we select the 5% of top-ranked poses and calculate the fraction of all near-native poses that are present in this selection. An $\mathcal{H}$ is labelled as successful for a given data case if this value exceeds 60%. Likewise, we can say that a given case is successfully scored by current $\mathcal{H}$.

2.3 Collection of $\mathcal{H}$

Initial analysis revealed that a single $\mathcal{H}$ was not sufficient to account for the diverse protein-ssRNA binding modes (Fig 2). Therefore, we opted for the creation of a small collection of $\mathcal{H}$, where each $\mathcal{H}$ is successful on a subset of the cases. When applied simultaneously, the collection should cover the majority of cases, except for a few outliers. The collection is created by selecting several best-performing $\mathcal{H}$, such that maximising the number of successfully scored cases in the training set. The full procedure is detailed in the next section (2.3.1).

Because in a real-life docking case, there will be no indication of which $\mathcal{H}$ from the collection is best suited for scoring, the case must be scored by all $\mathcal{H}$ and results must be pooled together (see 2.3.2). As the collection size increases, so does the chance of overfitting. For this reason, we have empirically limited the number of $\mathcal{H}$ to 4 per collection. Increasing this number to 5 or 6 had only limited influence (result not shown).

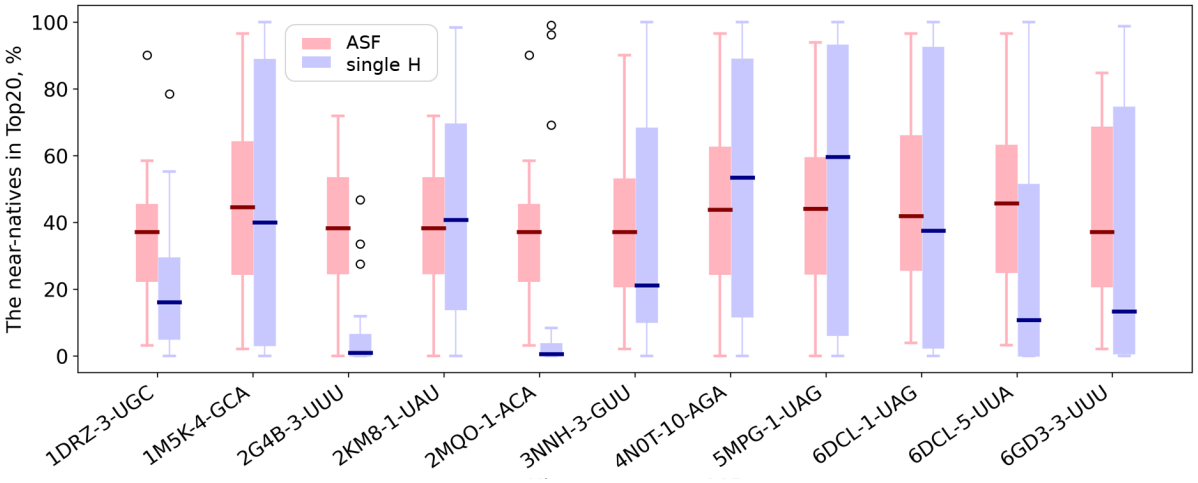


Fig 2. Comparison of the percentage of near-natives selected by a single $\mathcal{H}$ vs ASF. Each pair of adjacent boxes shows the distribution of the results produced by a corresponding $\mathcal{H}$ (purple) and ASF (pink) on the relevant for a given $\mathcal{H}$ test set(s) (sets used for the collection to which given $\mathcal{H}$ belongs), for a range from 0% to 100% of the near-natives in the 20% top-ranked poses.

2.3.1 Partitioning algorithm

While deriving a collection of 4 $\mathcal{H}$ - $\mathcal{H}_1$, $\mathcal{H}_2$, $\mathcal{H}_3$ and $\mathcal{H}_4$ - we partition the training cases into four subsets, plus a subset of outliers. This procedure is implemented as follows:

1) Derive $\mathcal{H}$ for each case individually;

2) Score each case with each $\mathcal{H}$;
3) For each pair (case; $\mathcal{H}$), calculate the percentage of the near-natives that end up in the 5% of top-ranked poses. If the calculated value is over 60%, then label this case as successfully scored by the given $\mathcal{H}$;
4) Select the four $\mathcal{H}$ that maximise the total number of successfully scored cases. This is the resulting collection.

Now, each training case either is associated with its best-performing $\mathcal{H}$ in the resulting collection or ends up in the set of outliers.

2.3.2. Scoring with collection and evaluation strategy

To score a case with a collection, we score its docking poses with $\mathcal{H}_1$, $\mathcal{H}_2$, $\mathcal{H}_3$ and $\mathcal{H}_4$ separately using (1). Then, for each $\mathcal{H}$, around 5% of its top-ranked poses are selected and pooled together in TopC (where "C" stands for a collection). If the same pose is present in several scorings, only its highest rank is kept. The size of the TopC should be equal to 20% of all sampled poses. The resulting set of poses TopC is expected to contain the best ones (the poses outside of TopC are dismissed).

To evaluate the performance of the collection for a case, the fraction of all near-native poses that end up in TopC is calculated. If this value exceeds 60%, then the collection is successful for a given data case.

3 Results

In this study, we developed a new protocol for deriving scoring parameters for molecular docking poses, based on distances between RNA and protein beads, in the form of a collection of 4 histogram sets ($\mathcal{H}$). We applied it to create HIPPO, a novel scoring function specifically for RRM-ssRNA fragment-based docking. To achieve this goal, we split every available RRM-ssRNA structure into RRM-fragment cases (fragments of 3 consecutive bound nucleotides), for each of which 10^7 docking poses were generated using the ATTRACT docking engine. Our initial benchmark consisted of 479 fragments from 81 complexes. Out of these, 262 fragments were unusable for training because of a sampling problem (less than 100 near-native poses sampled) or because of redundancy between fragments on the contact level (6Å), resulting in a dataset of 217 well-sampled non-redundant cases, coming from 57 RRM-ssRNA complexes. Within the resulting dataset, the average number of sampled near-native poses is 9112 and the median is 3145. To assess how HIPPO performance would generalise to new data cases, we used the leave-homology-out cross-validation strategy: 29 pairs of training and test sets were formed based on RRM sequence similarity. The size of the test set depended on the number of cases derived from each RRM-ssRNA complex of a given RRM and varied from 1 to 33 cases per set.

For a given pair of test and training sets, for each case in the training set, we derived an $\mathcal{H}$ by analysing the frequencies of bead-bead distances in the near-native (LRMSD<5Å) vs non-native (LRMSD>7Å) docking poses, and we applied it to each of the other cases in the training set. We selected the collection of 4 $\mathcal{H}$ sets that maximised the number of training cases for which at least one $\mathcal{H}$ ranks 60% of all near-native poses in the 5% top-ranked poses. Then, the collection was applied to the test cases, and the best of the 4 ranks for each pose was retained to obtain the 20% top-ranked poses (TopC). The collection was considered to be successful on a test case if at least 60% of all near-native poses were in TopC.

3.1 General performance

We applied the described protocol to each of the 29 training sets and derived 29 collections of 4 $\mathcal{H}$. We then applied these collections to the cases in the corresponding test sets and compared the percentages of near-natives selected in TopC with HIPPO and in the 20% top-ranked with ASF (Tab. 1, Fig 3). Further in the text, we refer to the percentage of near-natives present in TopC or 20% top-ranked as 'selected'. At least 60% of all near-natives selected (a 3-fold enrichment compared to random scoring) for more than half of the RRM-fragment test cases with HIPPO, versus a quarter with ASF (53% vs 26% of the test cases respectively). In one-third of the test cases, we even observed a 4-fold enrichment (80% of near-natives selected) with HIPPO, something which is rarely achieved by ASF (38% vs 7% of the test cases respectively). To ensure that our results were not skewed by cases coming from one or a few largest test sets, we compared the average success rates over the test sets and found 62% and 34% respectively (Fig. 4, a).

Tab. 1 Comparison of the performance of HIPPO vs ASF on the 217 cases (29 test sets, 57 complexes)

	ASF	HIPPO
% of near-natives in TopC/Top20, averaged over all test cases	43	55
Success rate (%) over all cases	26	53
Average highest % of near-natives inTopC/Top20 among the cases of a complex, over all test cases	60	72
Nb of complexes with the > 80% of near-natives in TopC/Top20 for at least one fragment	9	33
Nb of cases with > 80% of near-natives in TopC/Top20	15	75

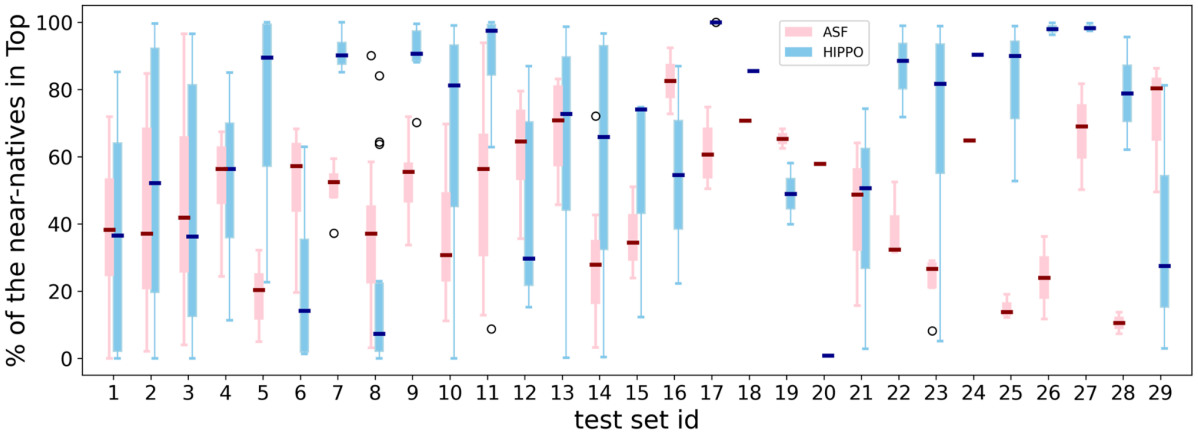


Fig 3. Comparison of the percentage of selected near-natives by collections vs ASF on the test sets. Each pair of adjacent boxes shows the distribution of the results produced by a corresponding collection (blue) and ASF (pink) on one of the 29 test sets, for a range from 0% to 100% of the near-natives in the corresponding Top (TopC/Top20 respectively).

3.2.1 Best-scored fragment per complex

We found a positive correlation (Pearson correlation, $r = 0.43$, Fig. 4, b) between the number of protein-fragment contacts under 5Å and the percentage of near-natives in TopC, which complies with the cold/hotspot theory. To perform anchored fragment-based docking, at least one fragment per complex must be well-docked. We thus analysed the distribution of successes among the complexes, with HIPPO and ASF. The number of complexes with at least one successfully scored fragment increased from 54% with ASF to 75% with HIPPO. With the success criterion raised to 80% of the near-natives selected (a 4-fold enrichment), the compared success rate percentages still increased from 16% with ASF to 58% with HIPPO. Moreover, the enrichment for the best-scored fragment per complex was increased with HIPPO compared to ASF in 68% of complexes. On average, for the best-scored fragment of each complex, HIPPO selects an additional 19% of all near-natives compared to ASF.

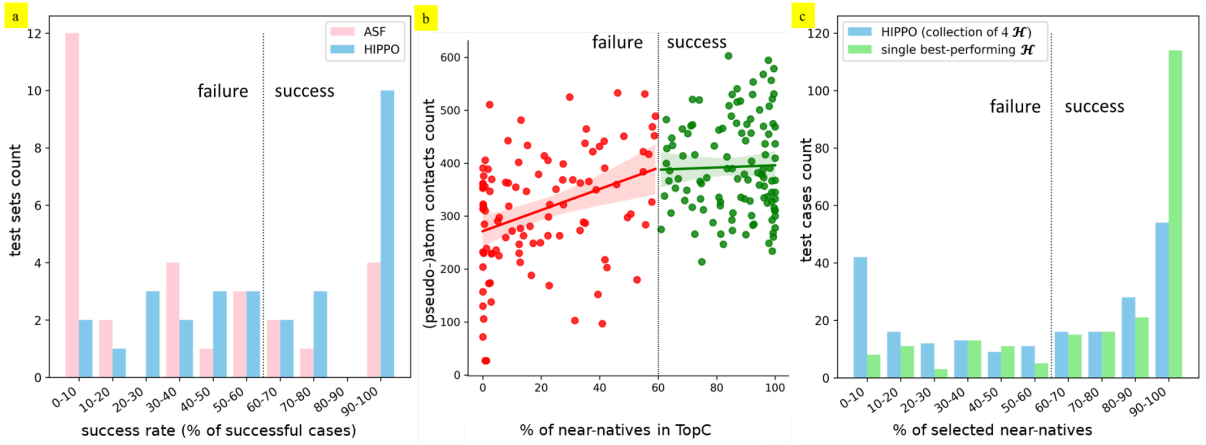


Fig 4. a) Distribution of the success rate per test set, achieved with ASF (pink) and HIPPO (blue). The black dotted line indicates the threshold of a 3-fold enrichment compared to random sampling. b) Relation between the number of contacts in a protein-fragment structure vs the percentage of near-natives in TopC achieved by HIPPO. c) Distribution per test case of the percentage of near-natives selected by a collection of 4 $\mathcal{H}$ (blue) versus by a single best-performing $\mathcal{H}$ (green).

3.3 Analysis of the collections

To assess the gains of using a collection (4 $\mathcal{H}$) instead of 1 $\mathcal{H}$, we evaluated if the 4 $\mathcal{H}$ bring complementary information, either for each test case (by selecting different near-native poses) or for each test set (by performing well on different test cases).

3.3.1. Complementarity of the 4 $\mathcal{H}$ in a collection

Out of 29 collections, the ones derived from the training sets 1, 2, 3, 4 and 8 are distinct (see Additional file 1: Table S3). The remaining collections are identical to the collection from training set 4. On the test set level, we can see that each single $\mathcal{H}$ is the best-performing (selects the highest number of near-natives) of the collection for 0% to 48% of the cases. In other words, there is never one $\mathcal{H}$ that is the best suited for half or more of the cases in a given test set. This complies with the hypothesis that several different $\mathcal{H}$ are required to account for different binding modes (Fig. 5, Additional file 1: Table S4), and that a few potentials better represent the diversity of RRM-ssRNA binding modes than one $\mathcal{H}$, by providing at least one well-suited $\mathcal{H}$ per case for most cases.

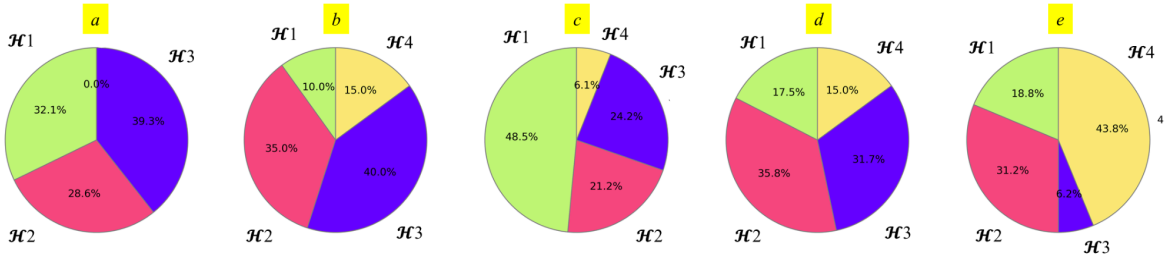


Fig 5. The percentage of cases within a test set, for which each of the 4 $\mathcal{H}$ in the collection is the best-performing one. a) For collection 1 on test set 1. b) For collection 2 on test set 2. c) For collection 3 on test set 3. d) For collection 4 on the united test set, suitable for validation of this collection's performance. This set consists of the test cases belonging to all test sets, excluding sets 1, 2, 3 and 8. e) For the collection 8 on test set 8.

3.3.2 Best-performing $\mathcal{H}$ per case or per complex

For half of the cases, most of the near-natives in the TopC were selected by a single $\mathcal{H}$ out of 4. If for each test case, we could use its best-performing $\mathcal{H}$ instead of the collection (and count near-natives in 20% top-ranked instead of pooling in the TopC), such modified application of HIPPO would reach a 3-fold enrichment for 77% cases (instead of 53% with the collection and 26% with ASF) and a 4-fold enrichment for 62% cases (instead of 38% with the collection and 7% with ASF) (Supplementary Section 4, Tab 4, Fig 2). Furthermore, selecting only the 5% top-ranked poses would show a 12-fold enrichment for 39% cases (vs 4% cases with ASF). For the

best-scored fragment per complex, a 12-fold enrichment was observed in 61% of complexes with HIPPO, while this is almost never achieved with ASF (7% of complexes). These numbers point toward the advantage of applying a single best-performing $\mathcal{H}$ per case rather than a collection if one could predict which $\mathcal{H}$ to apply to which case (Fig 6).

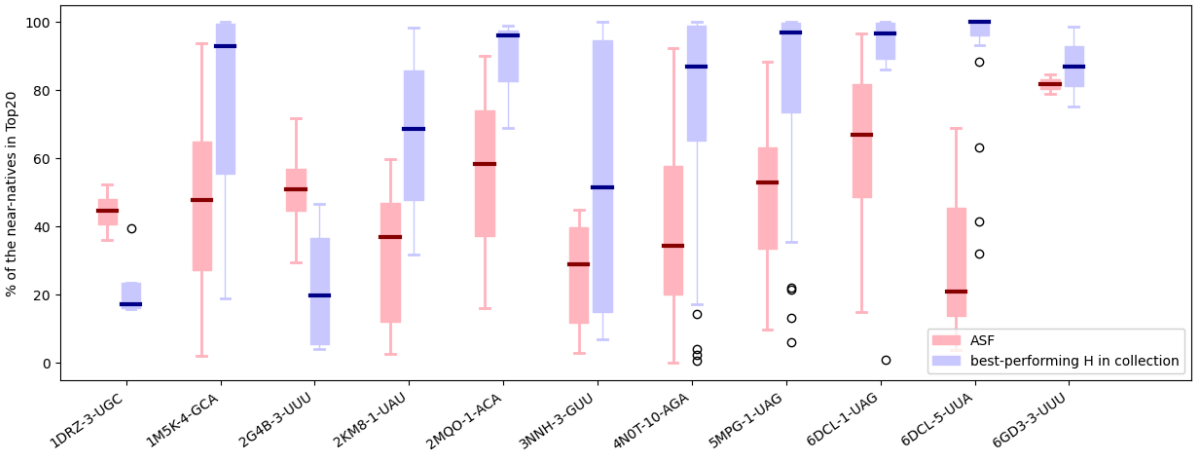


Fig 6. Comparison of the percentage of selected near-natives by ASF vs the best-performing $\mathcal{H}$. Each pair of adjacent boxes shows the distribution of the results produced by each best-performing $\mathcal{H}$ (purple) or ASF (pink) on the relative test cases for a range from 0% to 100% of all near-natives ranked in the 20% top-ranked poses.

4 Discussion

Despite the numerous biological roles of ssRNA-protein binding processes, there is still a lack of methods capable of addressing the dual challenges of the very high flexibility of ssRNA and the scarcity of its experimental structures. We previously developed a unique approach capable of modelling protein-bound ssRNA, by coarse-grained docking of ssRNA fragments with the ATTRACT docking software, followed by combinatorial assembly of geometrically compatible poses. This approach is successful in modelling the full ssRNA chain at high accuracy when conserved stacking contacts are known: the docking search space is reduced by constraints forcing the stacking of certain nucleotides on the conserved residues. In the absence of conserved contacts, this approach is limited by the poor sampling and low discriminatory power of the protein-RNA energy function of ATTRACT when applied to ssRNA fragments. With typically a few thousand near-native poses sampled out of 10^7 poses, the percentage of near-natives is less than 0.1%. In general, during assembly, low percentages of near-natives at the fragment level increase the probability of compatible non-native poses, leading to a prohibitive number of full-chain RNA models with an infinitesimally low percentage of quasi-native models. For direct applicability in the absence of conserved contacts, a very high enrichment is needed, followed by clustering and possibly refinement/rescoring with molecular dynamics, to arrive at an ensemble of perhaps a few hundred poses of which at least one is near-native.

In order to achieve such a high enrichment, we developed a new analytic approach for creating a scoring function for docking poses of coarse-grained ssRNA fragments, based on the frequencies of contact distances in near-native versus non-native poses. A specificity of our approach is to derive and combine a small set of potentials to better cover the diversity of ssRNA binding modes. We applied it to create HIPPO, a novel scoring function specifically for coarse-grained RRM-ssRNA fragment-based docking. On a benchmark of 57 RRM-ssRNA complexes.

HIPPO demonstrates a better discriminatory power for near-native poses than the state-of-the-art ATTRACT scoring function (ASF), making it the best coarse-grained scoring function tested for protein-ssRNA complexes to date.

The successfully and unsuccessfully scored cases are rather evenly distributed among the complexes (result not shown). HIPPO's strengths and weaknesses are thus not likely to be attached to any specific type of

complex, but rather to hot- and coldspots binding, meaning RNA fragments of a complex that are tightly and loosely attached to the protein respectively. This variability of docking performance over fragments is a difficulty inherent in a classical fragment-based docking approach, where each fragment must be docked (sampled and scored) within an accuracy threshold before the assembly. A way to tackle this is to ensure that at least one fragment per complex is very well docked and use each of its top-ranked poses as anchors to build a full RNA model by direct poses superposition followed by scoring. In the absence of evidence to identify the well-docked fragment from RNA sequence and protein structure, one would iteratively consider each fragment as such. We had previously applied a similar anchored docking of ssRNA on RRMs by using conserved stacking interactions between RRM aromatic residues and a nucleotide base as anchors [15]. Yet nearly half of RRM structures lack those conserved aromatics [21], and such a new hotspot approach would overcome this limitation. HIPPO will be better suited than ASF for this approach, since (i) more complexes have at least one successfully docked fragment compared to ASF, and (ii) the best-scored fragment in each complex has a higher enrichment for most complexes compared to ASF.

We have seen that for most cases (95%) the best-performing $\mathcal{H}$ of the collection performed better than the whole collection (Fig 4. c). A way to improve HIPPO's performance would be to determine which $\mathcal{H}$ from the collection will perform the best on a given protein-fragment case. This would allow us to apply only this one $\mathcal{H}$ and avoid retaining false positives returned by the other three $\mathcal{H}$. This may be achieved with the help of the supervised machine learning techniques based on the sequence of the fragment and the sequence or/and structure of the protein, and/or on the docking poses. Such a pre-trained classifier not only would drastically improve the performance of the scoring but could also give valuable insight into the most prevalent protein-ssRNA binding modes. More importantly, since scoring with the best performing $\mathcal{H}$ achieved 60% of near-natives in 5% top-ranked for the best-scored fragment in a complex for 61% of complexes, there is a great perspective in clustering these top-ranked poses and using the obtained prototypes as anchors.

We see several tuning possibilities that might yield improved HIPPO performance. In particular, we will try to apply a stricter threshold for near-native poses, and see if, despite the increased sampling difficulties encountered, there would still be enough signal for HIPPO to succeed for high-accuracy poses.

As mentioned earlier, we face not only scoring but also, primarily, a sampling problem in ssRNA docking. HIPPO can be considered as a pseudo-energy function, and as such, it is suitable for a sampling procedure based on energy minimisation that would not require derivability of the energy, such as a Monte Carlo approach [24]. We plan to test it against the current ATTRACT sampling procedure that uses ASF with gradient minimisation. Another possible way to apply HIPPO for the sampling is to convert each histogram into a differentiable function to be used directly in ATTRACT's gradient minimisation protocol.

To further evaluate the generalisability of our approach for deriving scoring potentials, we plan to expand our benchmark from only RRM-ssRNA structures to a more general protein-ssRNA benchmark, as well as to our benchmark of protein-ssDNA structures [25].

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The source code of HIPPO along with the final scoring parameter set are available via https://github.com/sjdv1982/histograms. The data set used for the study is available in the Supplementary Materials.

Competing interests

The authors declare that they have no competing interests.

Funding

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 813239.

Authors' contributions

AK assembled and processed the data (with input from ICB), participated in the design of the work and the creation of the software, and drafted the manuscript (with input from ICB). SJV contributed substantially to the conception and design of the work and to the creation of the software and revised the manuscript. MST contributed to the design of the work and revised the manuscript. ICB contributed substantially to the conception and design of the work and revised the manuscript. All authors read and approved the final manuscript.

Acknowledgements

Experiments presented in this paper were carried out using the **Grid'5000** testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations (see https://www.grid5000.fr).

References

1. Cléry, A., Blatter, M., & Allain, F. H. (2008). RNA recognition motifs: boring? Not quite. *Current opinion in structural biology*, *18*(3), 290–298
2. Choi, P. S., & Thomas-Tikhonenko, A. (2021). RNA-binding proteins of COSMIC importance in cancer. *The Journal of clinical investigation*, *131*(18), e151627.
3. Tsai, Y. S., Gomez, S. M., & Wang, Z. (2014). Prevalent RNA recognition motif duplication in the human genome. *RNA (New York, N.Y.)*, *20*(5), 702–712.
4. Bheemireddy, S., Sandhya, S., Srinivasan, N., & Sowdhamini, R. (2022). Computational tools to study RNA-protein complexes. *Frontiers in molecular biosciences*, *9*, 954926.
5. Chen S. J. (2008). RNA folding: conformational statistics, folding kinetics, and ion electrostatics. *Annual review of biophysics*, *37*, 197–214.
6. 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. (2020). Improved protein structure prediction using potentials from deep learning. *Nature*, *577*(7792), 706–710.
7. 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 SAA, Ballard AJ, 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 AW, Kavukcuoglu K, Kohli P, Hassabis D. (2021). Highly accurate protein structure prediction with AlphaFold. Nature, 596(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2
8. 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., … Baker, D. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. *Science (New York, N.Y.)*, *373*(6557), 871–876.
9. Bryant, P., Pozzati, G., & Elofsson, A. (2022). Improved prediction of protein-protein interactions using AlphaTest set2. *Nature communications*, *13*(1), 1265.
10. Yang, C., Chen, E. A., & Zhang, Y. (2022). Protein-Ligand Docking in the Machine-Learning Era. *Molecules (Basel, Switzerland)*, *27*(14), 4568.
11. Meli, R., Morris, G. M., & Biggin, P. C. (2022). Scoring Functions for Protein-Ligand Binding Affinity Prediction using Structure-Based Deep Learning: A Review. *Frontiers in bioinformatics*, *2*, 885983..
12. Pal, A., & Levy, Y. (2019). Structure, stability and specificity of the binding of ssDNA and ssRNA with proteins. *PLoS computational biology*, *15*(4), e1006768.
13. Mei, L. C., Hao, G. F., & Yang, G. F. (2022). Computational methods for predicting hotspots at protein-RNA interfaces. Wiley interdisciplinary reviews. RNA, 13(2), e1675. https://doi.org/10.1002/wrna.1675

14. Hall, D., Li, S., Yamashita, K., Azuma, R., Carver, J. A., & Standley, D. M. (2015). RNA-LIM: a novel procedure for analyzing protein/single-stranded RNA propensity data with concomitant estimation of interface structure. *Analytical biochemistry*, *472*, 52–61.
15. González-Alemán, R., Chevrollier, N., Simoes, M., Montero-Cabrera, L., & Leclerc, F. (2021). MCSS-Based Predictions of Binding Mode and Selectivity of Nucleotide Ligands. *Journal of chemical theory and computation*, *17*(4), 2599–2618.
16. Kappel, K., & Das, R. (2019). Sampling Native-like Structures of RNA-Protein Complexes through Rosetta Test seting and Docking. *Structure (London, England : 1993)*, *27*(1), 140–151.e5.
17. Isaure Chauvot de Beauchene, Sjoerd J. de Vries, Martin Zacharias (2016) Fragment-based modelling of single stranded RNA bound to RNA recognition motif containing proteins, *Nucleic Acids Research*, 44(10) 4565–4580,
18. Moniot, A., Guermeur, Y., de Vries, S. J., & Chauvot de Beauchene, I. (2022). ProtNAff: protein-bound Nucleic Acid filters and fragment libraries. *Bioinformatics (Oxford, England)*, *38*(16), 3911–3917.
19. Moniot, A., Chauvot de Beauchêne, I., Guermeur, Y. (2022). Inferring ε -nets of Finite Sets in a RKHS. In: Faigl, J., Olteanu, M., Drchal, J. (eds) Advances in Self-Organizing Maps, Learning Vector Quantization, Clustering and Data Visualization. WSOM+ 2022. Lecture Notes in Networks and Systems, vol 533. Springer, Cham.
20. Setny, P., & Zacharias, M. (2011). A coarse-grained force field for Protein-RNA docking. *Nucleic acids research*, *39*(21), 9118–9129.
21. InteR3M database https://inter3mdb.loria.fr/. Accessed 4 May 2023.
22. Moniot, A., Guermeur, Y., De Vries, S. J., & Chauvot de Beauchene, I. (2022). ProtNAff: Protein-bound Nucleic Acid filters and fragment libraries [Data set]. *Zenodo*.
23. Chauvot de Beauchene, I., de Vries, S. J., & Zacharias, M. (2016). Binding Site Identification and Flexible Docking of Single Stranded RNA to Proteins Using a Fragment-Based Approach. *PLoS computational biology*, *12*(1), e1004697.
24. Glashagen, G., de Vries, S., Uciechowska-Kaczmarzyk, U., Samsonov, S. A., Murail, S., Tuffery, P., & Zacharias, M. (2020). Coarse-grained and atomic resolution biomolecular docking with the ATTRACT approach. Proteins, 88(8), 1018–1028.
25. Mias-Lucquin, D., & Chauvot de Beauchene, I. (2022). Conformational variability in proteins bound to single-stranded DNA: A new benchmark for new docking perspectives. *Proteins*, *90*(3), 625–631.

Additional file 1

Table S1. Benchmark of solved RRM-ssRNA structures
pdb_id: proteinChain1_firstAtom_lastAtom-proteinChain2_firstAtom_lastAtom -...
rnaChain_firstAtom_lastAtom (numbers of the fragments used)

1A9N: A_1_162-B_1_94 Q_6_18 (2 3 4 5 6 7 11)
1B7F: A_1_167 P_3_12 (1 2 3 4 5 6 7 8)
1CVJ: A_1_169 M_2_9 (1 2 3 4 5 6)
1DRZ: A_1_91 B_48_57 (1 2 3 4 5 6 7)
1FJE: B_1_175 A_7_16 (1 2 3 4 5 7 8)
1FXL: A_1_167 B_1_8 (1 2 3 4 5 6)
1G2E: A_1_167 B_1_9 (1 2)
1M5K: C_1_92 B_35_44 (3 4 5 6)
1RKJ: A_1_175 B_8_15 (1 2 3 5)
1URN: A_1_96 P_6_12 (3 4 5)
1ZH5: A_1_181 D_2_9 (4 5 6)
2CJK: A_1_167 B_2_8 (1 2 3 4 5)
2G4B: A_1_172 B_3_7 (1 2 3)
2HYI: C_1_392-D_1_56-B_1_91-A_1_144 F_2_6 (1 2 3)

2J0S: A_1_391-C_1_143-D_1_89-T_1_44 E_1_6 (1)
2KG0: A_1_92 B_2_6 (1 2 3)
2KM8: B_1_84-C_1_167 A_1_13 (1 2 3 4 5 6 7 8 9 10)
2KXN: B_1_95 A_2_6 (1 2 3)
2M8D: B_1_91 A_2_8 (2 3 4 5)
2MGZ: A_1_94-B_1_105 C_2_12 (1 2 3 4 5 6 7 8 9)
2MKI: A_1_203 B_2_5 (1 2)
2MQO: A_1_105 B_2_6 (1 2 3)
2MQP: A_1_118 B_2_6 (1 2 3)
2MXY: A_1_105 B_2_7 (1 2 3 4)
2N3O: A_1_123 B_10_14 (1 2 3)
2RRA: A_1_99 B_2_6 (1 2 3)
2VOD: A_1_187 C_2_7 (4)
2VON: A_1_187 C_2_7 (3)
2XS7: A_1_86 B_2_4 (1)
3MOJ: B_1_75 A_46_49 (1 2)
3NNH: A_1_86-C_1_85 E_1_10 (2 3 4 5 6 7 8)
3RW6: A_1_245 H_11_15 (1 2 3)
4BS2: A_1_174 B_2_11 (1 2 3 4 5)
4CIO: A_1_97 B_2_7 (1 2 3 4)
4ED5: A_1_168 D_2_8 (1 2 3 4 5)
4F02: A_1_175-C_1_20 B_2_9 (6)
4N0T: A_1_363 B_11_29 (3 5 6 7 8 10 11 12 13 14 17)
4QQB: A_1_169-X_1_72 P_1_17 (1 2 3 4 5 6 7 8 10 11 12 13 14)
4YB1: P_1_92 R_8_11 (1)
5DET: A_1_91-B_1_94 Q_1_4 (1 2)
5HO4: A_1_179 B_2_5 (1 2)
5MPG: A_1_97 B_2_7 (1 2 3 4)
5MPL: A_1_102 B_2_6 (1 2 3)
5O1Y: A_1_163 B_2_4 (1)
5TF6: A_1_367 B_11_29 (11 12 13)
5WWE: A_1_174 B_2_5 (1 2)
5WWG: A_1_184 B_2_6 (1 2)
6ASO: A_1_369-B_1_95-C_1_79-D_1_59-E_1_79-F_1_75-G_1_67-H_1_83 I_8_26 (13 14)
6DCL: A_1_182-B_1_171 C_3_11 (1 2 3 4 5 6 7)
6F4G: A_1_175-B_1_95 C_9_19 (2 3 4 5 6)
6F4H: A_1_90 B_7_16 (5 6)
6G2K: A_1_80-B_1_80 R_1_6 (1 2 3 4)
6GBM: B_1_102 A_12_15 (1 2)
6GC5: A_1_79-B_1_76 E_1_5 (3)
6GD2: A_1_84-B_1_81 D_1_7 (2 3 4)
6GD3: B_1_83-C_1_84 P_1_6 (1 2 3 4)
6GX6: A_1_159 B_2_4 (1)

Table S2. Test sets

Test set 1 (28 cases) : 1A9N 1DRZ 1M5K 1URN 4YB1 6F4G 6F4H
Test set 2 (20 cases) : 5HO4 5MPG 5MPL 5WWE 5WWG 6DCL
Test set 3 (33 cases): 1B7F 1FXL 1G2E 4ED5 4QQB
Test set 4 (12 cases): 6G2K 6GC5 6GD2 6GD3
Test set 5 (7 cases): 3NNH
Test set 6 (4 cases): 1ZH5 2VOD 2VON
Test set 7 (4 cases): 2HYI 2J0S

Test set 8 (16 cases): 4N0T 5TF6 6ASO
Test set 9 (6 cases): 2KXN 2RRA
Test set 10 (15 cases): 2CJK 2KM8
Test set 11 (13 cases): 2MGZ 4CIO
Test set 12 (7 cases): 1CVJ 4F02
Test set 13 (4 cases): 2MXY
Test set 14 (11 cases): 1FJE 1RKJ
Test set 15 (3 cases): 2N3O
Test set 16 (2 cases): 6GBM
Test set 17 (4 cases): 4BS2
Test set 18 (1 cases): 5O1Y
Test set 19 (2 cases): 2MKI
Test set 20 (1 cases): 6GX6
Test set 21 (3 cases): 2MQO
Test set 22 (3 cases): 2MQP
Test set 23 (4 cases): 2M8D
Test set 24 (1 cases): 2XS7
Test set 25 (3 cases): 2KG0
Test set 26 (2 cases): 5DET
Test set 27 (3 cases): 2G4B
Test set 28 (2 cases): 3MOJ
Test set 29 (3 cases): 3RW6

Table S3. Composition of the distinct HIPPO collections in terms of $\mathcal{H}$

Collection id	1	2	3	4	8
Histogram sets $\mathcal{H}$	2G4B-3-UUU 2KM8-1-UAU 4N0T-10-AGA 5MPG-1-UAG	1M5K-4-GCA 3NNH-3-GUU 4N0T-10-AGA 6GD3-3-UUU	1M5K-4-GCA 4N0T-10-AGA 6DCL-1-UAG 6DCL-5-UUA	1M5K-4-GCA 4N0T-10-AGA 5MPG-1-UAG 6DCL-5-UUA	1DRZ-3-UGC 2MQO-1-ACA 3NNH-3-GUU 5MPG-1-UAG

Table S4. Performance of ASF versus each unique $\mathcal{H}$ on the test cases, where given $\mathcal{H}$ is best-performing. Numbers in bold indicates better performance between ASF vs best-performing $\mathcal{H}$.

test case* where $\mathcal{H}$ is best-performing	count of successfully scored cases		count of best-scored cases		avg % of selected near-natives per case	
	ASF	H	ASF	H	ASF	H
2G4B-3-UUU	**1**	0	**3**	1	**51**	23
2KM8-1-UAU	0	**7**	1	**9**	32	**67**
4N0T-10-AGA	15	**47**	6	**55**	39	**75**
5MPG-1-UAG	18	**47**	5	**50**	50	**82**
1M5K-4-GCA	12	**26**	3	**33**	48	**79**
3NNH-3-GUU	0	**5**	3	**7**	27	**54**
6GD3-3-UUU	2	2	1	1	82	**87**
6DCL-1-UAG	4	**6**	1	**6**	63	**82**

6DCL-5-UUA	3	**23**	0	**25**	31	**92**
1DRZ-3-UGC	0	0	**4**	0	**44**	22
2MQO-1-ACA	1	**3**	0	**3**	55	**88**
superior performance count	1	6	2	7	2	8

B.2 Attempted Approaches

Fitting Parameters to the Histograms

Before developing a current version of HIPPO's protocol, we implemented a histogram-based approach (§4.2), which involves summing up the log-odds and residual histograms and fitting the energy parameters to the resulting histograms. This part of the procedure was not included in the HIPPO's protocol, as the resulting parameters performed notably poorer than ASF (results not shown).

The fitting was done as follows:

- The input is a set of original ATTRACT energy parameters and the set of scoring histograms (one scoring potential);
- Obtain *xSet*, a set of distance values corresponding to the distance bins (distance ranges) of the scoring potential;
- Obtain *ySet*, a set of energy values for each distance in *xSet*, and for each (ε_{ij}; σ_{ij}) pair;
- Calculate an average over all rank_chunks (see HIPPO paper, the last paragraph of §2.2.1) value for each distance bin of each histogram (*finalBars*);
- Obtain energy values of the resulting histogram *yNewSet* by *ySet-coef·finalBars*. Different values of the *coef*, namely 1, 0.1 and 0.01, were tested separately;
- Each energy curve was fitted to the corresponding set of points (*xSet*, *yNewSet*);
- Updated parameter set was tested via re-scoring the training set and comparing results with ASF.

For the fitting a *scipy.optimize.curve_fit*, a non-linear least squares method, was used (documentation is available at https://docs.scipy.org/doc/scipy/reference/generated/scipy.optimize.curve_fit.html). The maximum of interactions was set up to 100,000, and the boundaries for σ and ε were set up to $[0; 25]$.

As this is a least squares method, the nature of the energy curve (i.e. steep slope and high absolute energy values of the points above the x-axis and very gentle slope and low absolute energy values of the points below the x-axis) impacts the fitting process. Therefore, each curve was fitted 4 times (Fig. B.1):

- Starting with the second point above the x-axis;
- Starting with the first point above the x-axis;
- Starting with the first point below the x-axis;
- Starting with the second point below the x-axis.

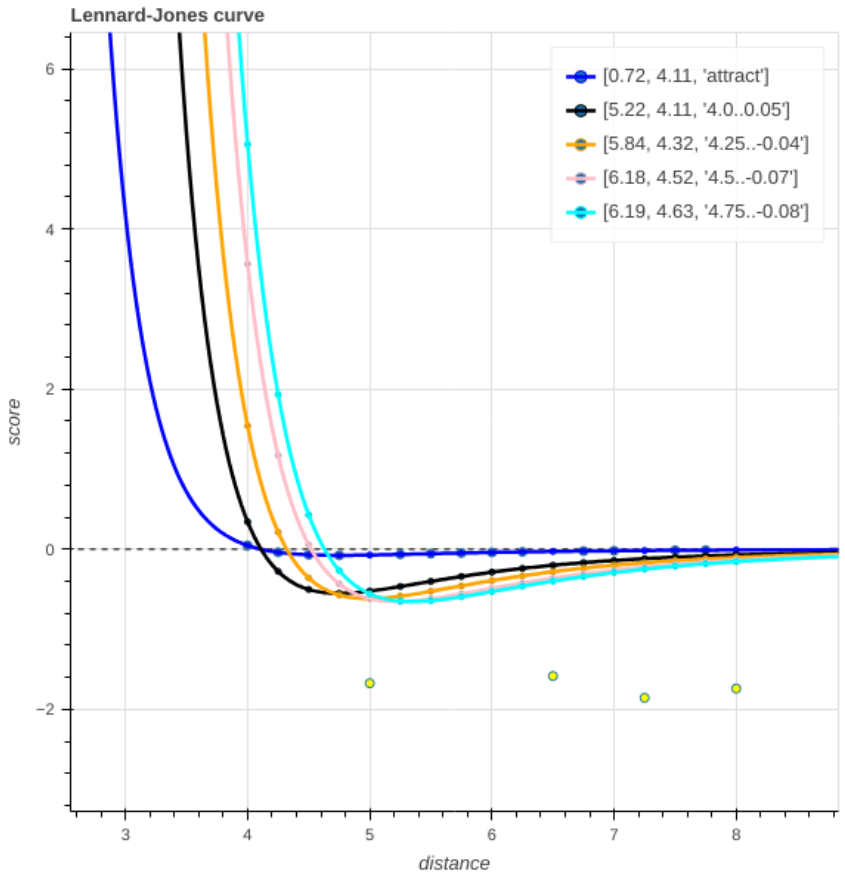


Figure B.1 - Fitting of the attract curve (blue) to the *yNewSet* (yellow points, the first displayed point is the 3rd point below the x-axis). Resulting curves are obtained by fitting the attract curve to the second point above the x-axis (black curve); the first point above the x-axis (orange); the first point below the x-axis (pink) and the second point below the x-axis (cyan).

Merging Several Pools of Docking Poses to Derive One Scoring Potential

In the current HIPPO protocol, a scoring potential is derived for a single data case, from its pool of the labelled docking poses. We have attempted to merge two or more pools of the docking poses (of the different data cases) and derive a scoring potential from such a merged pool.

Practically, not the pools of the near-natives are merged, but their distance arrays (see HIPPO paper, §2.2.1) are weighted and summed up. The weights are calculated based on the number of near-native poses in each pool:

$$weight_i = 1/(0.1 + N_i/\sum_{k=0}^{m} N_k),$$

where N_i is a number of the near-natives in the i_{th} pool of the docking poses, and m is a total number of the pools to merge.

The $weight_i$ is applied to each value in the distance array obtained from the i_{th} pool of the docking poses.

This procedure, while having some preliminary success [ref], is less effective compared to the current version of HIPPO.

Fill in Missing Histograms

As mentioned before, currently a scoring potential is derived for a single data case, which means that histograms for some pairs of beads are missing within a single potential. For example, current $\mathcal{H}1$ have been derived from data case 1M5K-GCA and it contains no histograms for the beads representing U side chain (GU1, GU2, GU3). In an attempt to fill up these empty spaces, we tried taking corresponding histograms from the first best-performing histogram containing missing beads. However, this resulted in a slightly worse performance of the scoring potential.

B.3 Possible Future Tuning

Minimum Number of Near-Natives in Each Distance Range

Currently, the minimum number of near-natives in each distance range is determined individually for each ij-pair as 1/60 of all near-native poses (see HIPPO paper, §2.2.1). This constant is tailored to the typical number of sampled near-natives with LRMSD < 5Å (note that cases with less than 99 near-natives are rejected). To apply HIPPO's protocol to other types of data (different LRMSD thresholds and/or different types of complexes), this threshold is likely to require manual fitting. Instead of such manual adjustments, it would be interesting to develop an automatic tuning protocol capable of selecting a constant (or an adaptive value) based on, for example, the median number of sampled near-natives.

Usage of the Rank Chunks

Each distance range consists of three rank chunks (i.e. rank-based sub-bins, see HIPPO paper, the last paragraph of §2.2.1). The average across these three rank chunks is used for scoring. However, initially, we intended to fit a set of coefficients for each rank chunk so that poses with a higher rank have more impact and vice versa. This idea was not explored due to time constraints, but it is interesting to explore.

Smoothing Individual Histograms

A subset of the histograms of $\mathcal{H}1$ is shown below (Fig. B.2). While fitting the energy curves to these histograms results in a loss of signal, it is possible that smoothing the histograms could yield better performance. This is because the score difference between poses differing by a small distance would be less pronounced.

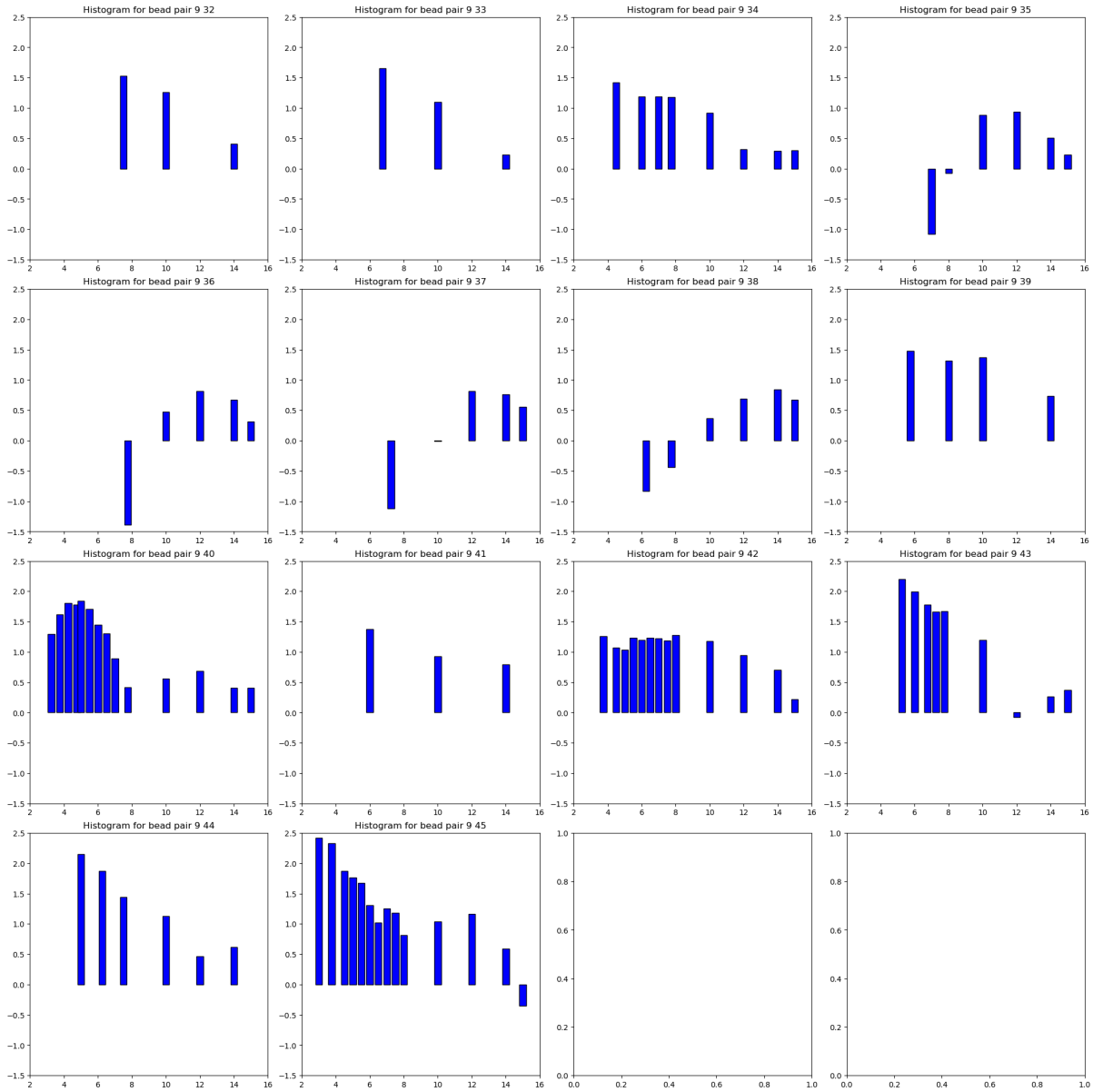


Figure B.2 - Subset of histograms for bead 9 (amino acid GLN, atoms CN1, CD OE1 NE2) and each RNA bead (32 to 45, starting from 1 bead for phosphate, then 2 beads for sugar, followed by 4 or 3 beads per bases A, G, C, U). The x-axis shows distances (in Å) discretised into distance bins/ranges, and the y-axis displays histogram-based scores (no units).

B.4 Benchmark for Scoring

This section details the benchmark of solved protein-ssRNA structures. The data is displayed in the following format:

PDB_ID proteinChain_firstAtom-proteinChain_lastAtom rnaChain_firstAtom-rnaChain_lastAtom

newRRM

6JVX	A_29-A_115	B_1-B_7
6YYM	A_580-A_726	B_1-B_12
7QDD	B_114-B_201	A_108-A_113
7VRL	B_107-B_208	A_1-A_7
7ZAP	A_141-B_237	B_101-B_107
7ZEW	A_1-A_114	B_1-B_6

nonRRM

1ASY	A_1-B_490	S_72-S_75	3BOY	A_1-C_147	D_2-D_21
1B23	P_1-P_405	R_71-R_74	3BSB	B_1-B_341	C_2-C_9
1BMV	1_1-2_374	M_1-M_5	3BSX	A_1-A_341	C_2-C_10
1DDL	A_1-C_175	D_1-D_7	3BT7	A_1-A_369	C_6-C_14
1EC6	A_1-A_87	D_5-D_16	3BX2	A_1-A_328	C_2-C_9
1ETF	B_1-B_23	A_23-A_27	3BX3	A_1-A_325	C_2-C_8
1F7V	A_1-A_606	B_13-B_22	3CUL	A_1-A_88	C_29-C_38
1HJI	B_1-B_26	A_6-A_9	3DD2	L_1-H_258	B_13-B_16
1I9F	B_1-B_19	A_23-A_27	3EX7	A_1-D_57	F_2-F_6
1J1U	A_1-A_299	B_35-B_38	3FHT	A_1-A_392	C_1-C_6
1JBT	B_1-B_149	D_14-D_17	3G9YA	A_1-A_29	C_2-C_6
1JID	A_1-A_114	B_13-B_16	3GIB	A_1-C_62	H_1-H_9
1K8W	A_1-A_304	B_10-B_14	3HSB	A_1-F_67	X_2-X_7
1KQ2	A_1-M_61	R_2-R_7	3I5X	A_1-A_509	B_3-B_10
1L9A	A_1-A_87	B_36-B_39	3IE1	A_1-A_431	E_1-E_4
1LNG	A_1-A_87	B_24-B_27	3IEVA	A_1-A_302	D_1-D_9
1M8V	A_1-N_71	S_1-S_6	3K49	A_1-A_353	B_2-B_10
1M8X	A_1-A_341	C_2-C_8	3K5Q	A_1-A_400	B_1-B_9
1M8Y	A_1-A_341	C_1-C_10	3K5Y	A_1-A_400	B_2-B_9
1N1H	A_1-A1_264	B_1-B_4	3K5Z	A_1-A_394	B_2-B_9
1NYB	A_1-A_22	B_10-B_14	3K61	A_1-A_393	B_2-B_9
1Q2R	A_1-A_376	E_11-E_14	3K62	A_1-A_400	B_2-B_9
1QFQ	B_1-B_35	A_6-A_10	3K64	A_1-A_400	B_2-B_9
1R3E	A_1-A_305	C_7-C_11	3L26	A_1-B_123	C_4-C_8
1RGO	A_1-A_70	D_2-D_9	3M7N	A_1-I_258	Y_1-Y_5
1SI3	A_1-A_117	B_2-B_9	3M85	A_1-I_259	X_1-X_4
1SJ3	P_1-P_95	R_49-R_58	3MDG	A_1-B_210	C_2-C_5
1TTT	A_1-A_405	D_73-D_76	3NVI	A_1-D_121	E_9-E_24
1WMQ	A_1-B_143	C_1-C_7	3O3I	X_1-X_108	A_11-A_14
1WNE	A_1-A_476	B_2-B_6	3O8C	A_1-B_645	C_1-C_6
1WPU	A_1-B_147	C_2-C_7	3OIJ	A_1-B_218	C_5-C_10
1ZBH	A_1-D_289	E_7-E_10	3PEW	A_1-A_391	B_1-B_5

1ZBN	B_1-B_17	A_14-A_17
1ZE2	A_1-A_300	C_10-C_14
1ZL3	A_1-A_302	B_9-B_15
2ANN	A_1-A_148	B_8-B_16
2ASB	A_1-A_226	B_2-B_11
2AZX	A_1-A_377	C_32-C_37
2BH2	A_1-A_418	C_2-C_12
2BQ5	A_1-C_129	R_6-R_12
2BU1	A_1-C_129	R_5-R_11
2C06	A_1-B_110	C_2-C_5
2CSX	A_1-A_464	C_33-C_39
2DB3	A_1-A_420	E_2-E_5
2DLC	X_1-X_339	Y_32-Y_35
2DRB	A_1-A_437	B_32-B_35
2FMT	A_1-A_314	C_73-C_77
2HGH	A_1-A_87	B_5-B_8
2I91	A_1-A_520	D_8-D_14
2IX1	A_1-A_643	B_1-B_13
2JLW	A_1-A_450	C_2-C_6
2JLX	A_1-A_451	C_2-C_7
2JPP	A_1-B_53	C_8-C_13
2KFY	A_1-A_102	B_2-B_6
2LI8	A_1-A_63	B_2-B_7
2MS1	A_1-A_55	B_6-B_9
2N8L	A_1-A_191	B_2-B_7
2N8M	A_1-A_191	B_1-B_7
2PLY	A_1-B_198	C_11-C_14
2PY9	A_1-B_66	E_6-E_12
2Q66	A_1-A_519	X_2-X_5
2QUX	A_1-B_121	C_11-C_16
2R7T	A_1-A1_073	X_2-X_7
2R7V	A_1-A1_073	X_2-X_5
2R7W	A_1-A1_073	X_2-X_7
2R8S	L_1-H_219	R_66-R_78
2RSK	C_1-D_12	A_2-A_12
2RU7	C_1-D_12	A_2-A_12
2VNU	D_1-D_676	B_2-B_9
2XGJ	A_1-A_964	C_2-C_5
2XZL	A_1-A_756	B_1-B_8
2YJY	A_1-A_337	C_1-C_10
2ZZM	A_1-A_329	B_35-B_41
3AEV	A_1-B_177	C_2-C_11
3AMT	A_1-A_405	B_33-B_41
3PF4A	B_1-B_66	R_2-R_5
3Q0M	A_1-A_337	C_2-C_8
3Q0Q	A_1-A_343	B_2-B_8
3Q0S	A_1-A_343	B_2-B_8
3QGB	A_1-A_400	B_2-B_9
3QGC	A_1-A_400	B_2-B_9
3QJJ	A_1-A_243	Q_2-Q_12
3QJL	A_1-B_240	X_2-X_12
3R2C	A_1-J_79	R_2-R_8
3R9W	A_1-A_302	B_24-B_34
3RC8	A_1-A_610	E_1-E_5
3RER	A_1-F_61	K_2-K_8
3T3O	A_1-A_553	B_1-B_5
3T5Q	A_1-A_306	C_1-C_8
3V71	A_1-A_364	B_1-B_7
484D	A_1-A_17	B_8-B_18
4ATO	A_1-A_168	G_2-G_5
4B8T	A_1-A_106	B_1-B_5
4BA2	A_1-I_213	R_1-R_4
4D25	A_1-A_427	D_1-D_6
4H5P	A_1-B_244	E_2-E_14
4I67	A_1-A_76	B_1-B_4
4J1G	A_1-B_227	E_2-E_44
4J7L	A_1-A_358	B_1-B_5
4JK0	D_1-D_120	B_1-B_5
4JNG	A_1-D_223	L_1-L_42
4JNXA	A_1-D_124	B_8-B_15
4JVYA	A_1-B_190	D_1-D_6
4K4U	A_1-A_462	B_1-B_5
4K4W	A_1-A_462	B_1-B_6
4KRE	A_1-A_799	R_1-R_9
4KRF	A_1-A_817	R_1-R_12
4Z0C	A_1-D_709	C_1-C_13
5YTS	A_1-A_74	B_1-B_4
5YTX	A_1-A_74	B_1-B_4
6KTC	A_1-A_74	V_1-V_4
6KUG	A_1-A_73	B_1-B_4
6RA4	A_1-B_130	M_2-M_9
6SQN	A_1-C_95	Z_5-Z_12
6UV1	A_1-A_438	C_2-C_7
6UV2	A_1-A_431	C_2-C_7
6UV4	A_1-A_426	C_1-C_7
6X5M	H_1-L_215	R_15-R_19

B.5 Chain Assembly

Benchmark

In the following table (Tab B.1) one can find a list of fragments which undergo assembly. This table also shows the number of near-native poses in the top5%, obtained with ASF and BP (columns 'num_nn top5_asf' and 'num_nn top5_bp' respectively), as well as which percentage of all sampled near-native this is (columns '%_nn top5_asf' and '%_nn top5_bp' respectively). The difference between these numbers and percentages is shown in the columns 'delta num' and 'delta %' respectively. Finally, hor the host-spot (middle) fragment, the rank first near-native pose, which is present in the near-native chain, is shown for two overlap values.

Table B.1 - Benchmark and assembly data.

fragment number	num_nn top5_bp	%_nn top5_bp	num_nn top5_asf	%_nn top5_asf	delta num	delta (%)	min_rank overlap 0.9A	min_rank overlap 1.4A
2BH2								
6	885	12%	12	0%	+873	12%	-	-
7	4846	23%	1914	9%	+2932	14%	1419	4
8	887	5%	2779	14%	-1892	-9%	-	-
2JLX								
1	5330	24%	158	1%	+5172	23%	-	-
2	6243	8%	74	0%	+6169	8%	4510	474
3	427	1%	249	1%	+178	0%	-	-
2JPP								
1	104	5%	11	0%	+93	5%	-	-
2	1124	18%	86	1%	+1038	17%	26439	6699
3	620	16%	35	1%	+585	15%	-	-
3BT7								
1	269	2%	422	4%	-153	-2%	-	-
2	902	6%	978	7%	-76	-1%	44015	16211
3	146	7%	113	5%	+33	2%	-	-
3CUL								
4	269	86%	727	8%	+7242	78%	-	-
5	902	53%	282	2%	+16803	51%	19	16
6	146	1%	18	0%	+77	1%	-	-

fragment number	num_nn top5_bp	%_nn top5_bp	num_nn top5_asf	%_nn top5_asf	delta num	delta (%)	min_rank overlap 0.9A	min_rank overlap 1.4A
3K62								
2	105	1%	684	7%	-579	-6%	-	-
3	1365	3%	8	0%	1357	3%	20651	13069
4	99	0%	742	3%	-643	-3%	-	-
3O8C								
1	197	13%	1	0%	196	13%	-	-
2	550	31%	24	1%	526	30%	no chains	18609
3	446	26%	31	2%	415	24%	-	-
4H5P								
2	121	44%	11	4%	110	40%	-	-
3	1713	55%	135	4%	1578	51%	no chains	49
4	1542	32%	120	3%	1422	29%	-	-
4H5P								
5	1656	12%	319	2%	1337	10%	-	-
6	2060	16%	1423	7%	1537	9%	no chains	642
7	650	70%	156	17%	494	53%	-	-
1SJ3								
4	7375	95%	880	11%	6495	+84%	-	-
5	2716	8%	1472	5%	1244	+3%	188	188
6	1446	17%	161	2%	1285	+15%	-	-
6UV2								
1	91	1%	203	3%	-112	-2%	-	-
2	3089	20%	2137	14%	952	6%	75514	1975
3	2233	35%	1751	27%	482	8%	-	-
6UV4								
2	1391	12%	2071	18%	-680	-6%	-	-
3	1703	34%	1543	31%	160	3%	40850	6805
4	272	26%	370	36%	-98	-10%	-	-
6JVX								

fragment number	num_nn top5_bp	%_nn top5_bp	num_nn top5_asf	%_nn top5_asf	delta num	delta (%)	min_rank overlap 0.9A	min_rank overlap 1.4A
1	621	11%	49	1%	572	10%	-	-
2	3428	30%	724	6%	2704	24%	170	39
3	2608	46%	308	5%	2300	41%	-	-
7VRL								
2	66	6%	25	2%	41	4%	-	-
3	466	6%	120	2%	346	4%	111704	7550
4	456	5%	342	4%	114	1%	-	-

Persentages of the Near-Native Chains

The following table (Tab. B.2) displays the percentages of the near-native chains out of all assembled chains for each hyperparameter set.

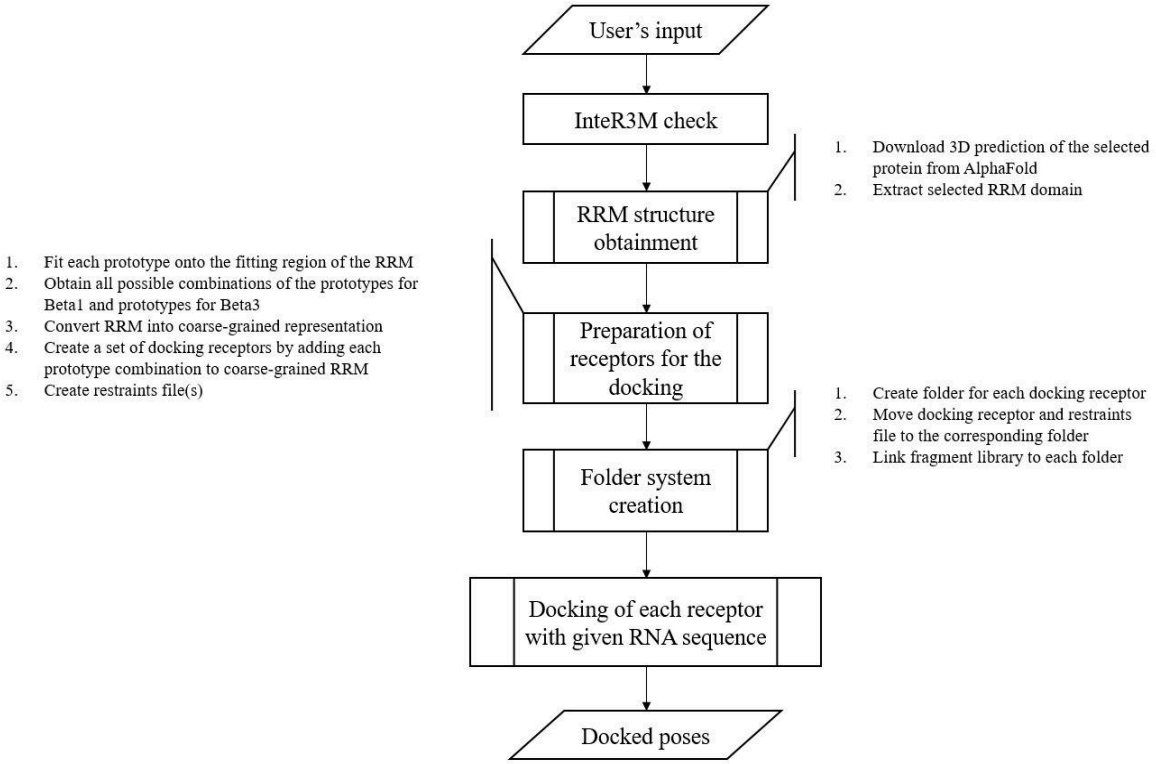
Table B.2 - Percentages of the near-native chains out of all assembled chains for each hyperparameter set, assembled using ASF or BP scoring.

	Hpar1		Hpar2		Hpar3		Hpar4	
	ASF, %	**BP, %**	**ASF, %**	**BP, %**	**ASF, %**	**BP, %**	**ASF, %**	**BP, %**
1SJ3	0.78	0.12	3.56	0.58	12.79	13.57	3.91	1.23
2BH2	2.92	11.81	5.59	24.82	1.15	20.86	4.49	12.13
2JLX	0.02	18.8	0.02	11.26	0.0	26.44	0	11.29
2JPP	0	0.01	0	0.0	0.25	0.0	0	0.01
3BT7	0.28	0.02	0.86	1.72	0.04	0.0	0.01	0
3CUL	0.53	0.65	0.47	0.72	0	1.44	0.57	2.18
3K62	0	10.43	0	6.78	0	0	0	0
3O8C	0	0	0	0	0	0	0	0
4H5P	0	0	0	0	0	1.49	0	0.28
4H5P	0	0	0	0	0.18	0.49	0.02	0.07
6JVX	0.02	0.04	0.03	0.43	0.01	1.02	0.01	0.25
6UV2	88.89	61.11	100.0	45.46	27.41	2.58	64.56	7.53
6UV4	13.64	10.0	0.0	10.35	41.89	1.86	16.52	1.13
7VRL	0	0.02	0	0.06	0	0	0	0

Appendix C

Data-driven docking

C.1 'RRM-RNA dock' Flowchart



C.2 'RRM-RNA dock' User Manual

To get started, visit https://github.com/AnnaKravchenko/RRM-RNA-dock/tree/main.

User Input

To execute the pipeline, run the *pip.py*.
The requirements for the input data are the next:

- The given combination of UniProt protein ID (-id/--uniProtId) and RRM index (-rrm/--rrm_domain_id) should belong to the InteR3M database;

- The given ligand sequence (-seq/--ss_rna_sequence) should contain nucleotides from RNA only. The nucleotides should be of a canonical type. The size of the sequence is unlimited, however, the current version of the pipeline docks a single fragment out of the whole chain;
- Both anchor nucleotides should be specified by their indexes in the RNA sequence (numbering starts from 1). These should be 2 distinct nucleotides within the given RNA sequence. They should be located within a single fragment, i.e. they should be separated by 0 to 1 non-anchor nucleotide). One nucleotide (-ancNucB1/----anchoring_nucleotide_id_beta1) should form a stacking interaction with the amino acid in position 5, in RNP1; and another nucleotide (-ancNucB3/--anchoring_nucleotide_id_beta3) should form stacking interaction with the amino acid in position 2, in RNP2.

Optionally, user can modify the next parameters of the configuration file (*config.ini*) (Fig. 1):

- dist_restraint_B1 and dist_restraint_B3 are the minimal distance for positional restraints in Angstrom for the docking of anchored nucleotides;
- Score_penalty is a score penalty applied if a positional restraint is violated;
- docking_cpu is the number of CPUs used during the docking run;
- Docking_tmp is the path to the folder where intermediate docking files are stored. The default value '0' redirects these files directly to the docking directory (e.g. /../rrm1/b1_1/b3_1/).

```
[DEFAULT]
mdir = ~/RRM-RNA-dock/git/mdir/
jsonFile = data.json
refB1 = /final_beta1_3.5_clust_proto_bb.pdb
refB3 = /final_beta3_3.0_clust_proto_bb.pdb
dist_restraint_B1 = 4.5
dist_restraint_B3 = 4
score_penalty = 3.5

[DOCKING]
docking_cpu = 8
docking_tmp = 0
```

Figure 1. Configuration file with the default setting.

In case the program fails to read or process the user's input, a corresponding error message will be displayed and the execution will be interrupted (Fig. 2).

```
(attract) akravche@garimpeiro:/data1/akravche/scripts/anchoring/git/mdir$ python3 /data1/akravche/scripts
/anchoring//pip.py -wdir /data1/akravche/dataset/data21//anchoring/data/ -id P26368 -rrm 1 -seq CAC -ancN
ucB1 1 -ancNucB3 8
**************************************
* Read user input...
**************************************
Valid UniProt index P26368
Valid RRM domain index 1
Valid ssRNA sequence CAC
*Invalid anchoring nucleotide index for Beta3: given -ancNucB3/--anchoring_nucleotide_id_beta1 value (8)
is out of range of given ssRNA sequence
Please re-run the script with valid anchoring nucleotide(s) index/indices
(attract) akravche@garimpeiro:/data1/akravche/scripts/anchoring/git/mdir$ _
```

Figure 2. An example of the error message.

If the folder on the given combination of UniProt protein ID (parameter -id/--uniProtId) and RRM (-rrm/--rrm_domain_id) already exists, it should be emptied before further execution on the pipeline. To prevent data loss, in such a case the user will be asked to confirm the action (Fig. 3).

```
(attract) akravche@garimpeiro:/data1/akravche/scripts/anchoring/git/mdir$ python3 /data1/akravche/scripts
/anchoring//pip.py -wdir /data1/akravche/dataset/data21//anchoring/data/ -id P26368 -rrm 1 -seq CAC -ancN
ucB1 1 -ancNucB3 2
**************************************
* Read user input...
**************************************
Valid UniProt index P26368
Valid RRM domain index 1
Valid ssRNA sequence CAC
Valid anchoring nucleotide index for Beta1 1
Valid anchoring nucleotide index for Beta3 2
********************************************************************************
* Load protein 3D prediction from AlphaFold
********************************************************************************
Directory /data1/akravche/dataset/data21//anchoring/data//P26368/rrm1 exists
In order to proceed, it has to be empty
Empty this directory?
[y/n]: _
```

Figure 3. An example of the command line interface if the path to an already existing folder is given.

To give the user more flexibility, after completion of the necessary docking preparations, the user is asked to confirm the beginning of the docking run (Fig. 4).

```
********************************************************************************
********************************************************************************
* Reduce domain
********************************************************************************
********************************************************************************
* Get just nucleotides from fitted prototypes
********************************************************************************
********************************************************************************
* Create dir structure, create reseptor file with 2 ghost nucleotides
********************************************************************************
********************************************************************************
* Work out which fragment will be docked
********************************************************************************
********************************************************************************
* Create restraint files
********************************************************************************
Would you like to start the docking?
[y/n]: _
```

Figure 4. An example of the command line interface after completion of the necessary docking preparations and before the beginning of the docking.

Output

Execution of the pipeline leads to the creation of the next file system (Fig. 5):

- For a given UniProt protein ID, a directory with the same name will be created in the working directory (-wdir/--work_directory);
- For a given RRM id, a directory with the name ‘rrmx’, where x is a given id, will be created in the protein directory. The related files created by the pipeline (before docking) are stored in this directory;
- For each prototype of a nucleotide in contact with beta1, a directory ‘b1_i’, where *i* is the index of the b1-prototype, will be made in ‘rrmx’ directory;
- For each prototype of a nucleotide in contact with beta3, a directory ‘b3_j’, where *j* is a number of b3-prototype, will be made in each of ‘b1_i’ directories. In these directories, the relevant docking output files are stored.

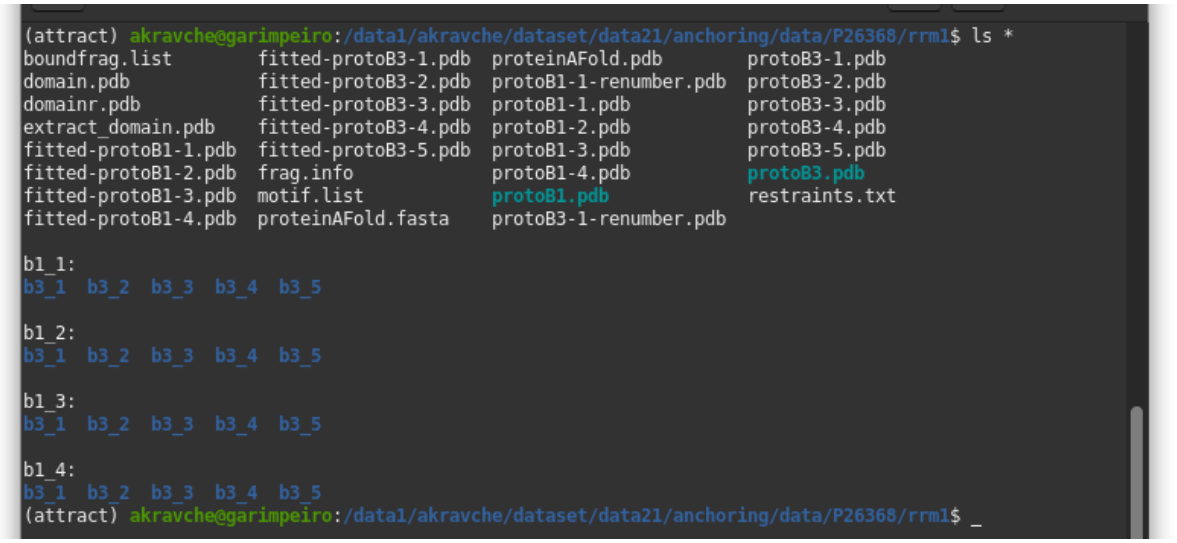


Figure 5. An example of the output folder structure.

C.3 Sampling and Scoring of the Poses Obtained with Anchoring Patterns vs *ab initio*

Data in Tab. C.1 an Tab. C.2 is provided for the case 1DRZ-CAC.

Table C.1 - Percentage of the near-native poses out of all sampled near-natives, which is precent within the list of top-ranked poses.

LRMSD	***Ab initio***	**All-patterns**	**1 1**	**1 2**	**1 3**	**1 4**	**1 5**	**2 1**	**2 2**	**2 3**	**2 4**
<2Å in top5%	45	4	15	6	0	0	0	33	11	25	0
<2Å in top20%	78	38	41	53	50	51	100	71	63	81	57

LRMSD	*Ab initio*	All-patterns	1 1	1 2	1 3	1 4	1 5	2 1	2 2	2 3	2 4
<3Å in top5%	44	5	7	3	6	0	1	12	10	11	0
<3Å in top20%	73	27	28	28	30	19	27	42	50	46	26

LRMSD	2 5	3 1	3 2	3 3	3 4	3 5	4 1	4 2	4 3	4 4	4 5
<2Å in top5%	15	4	7	0	**56**	25	0	0	0	0	-
<2Å in top20%	85	71	51	0	**96**	50	33	14	0	15	-
<3Å in top5%	6	8	6	**28**	11	9	6	5	7	0	2
<3Å in top20%	45	45	32	**75**	69	64	35	36	33	16	30

Table C.2 - The percentage of the poses, which have to be retained to keep a pose with the lowest LRMSD per list of all sampled poses.

	1AUD	1DRZ	1DZ5	1DZ5	1M5K	1M5K	1M5O	1M5O	1SJ3	1U6B	1URN	1VBX
1 1	29.04	27.01	7.97	13.32	36.79	36.79	36.79	36.79	6.31	10.22	81.58	6.31
1 2	10.43	10.43	9.59	16.03	44.27	44.27	44.27	44.27	7.59	10.43	98.15	7.59
1 3	5.30	20.96	7.82	13.06	36.07	36.07	36.07	36.07	6.18	14.65	79.98	6.18
1 4	16.25	15.71	6.79	11.34	31.33	31.33	31.33	31.33	5.37	45.05	69.46	5.37
1 5	6.49	16.47	8.46	14.13	39.03	39.03	39.03	39.03	6.69	16.47	86.54	6.69
2 1	9.44	9.44	7.31	12.21	33.72	33.72	33.72	33.72	5.78	20.41	74.76	5.78
2 2	17.07	17.07	7.86	13.12	36.25	36.25	36.25	36.25	6.21	17.07	80.36	6.21
2 3	36.85	11.75	7.43	12.41	34.27	34.27	34.27	34.27	5.88	11.75	75.99	5.88
2 4	12.00	19.52	6.68	11.16	30.84	30.84	30.84	30.84	5.29	19.52	68.38	5.29
2 5	20.95	16.50	7.38	12.32	34.03	34.03	34.03	34.03	5.83	60.29	75.45	5.83
3 1	0.46	7.97	9.97	16.65	45.98	45.98	45.98	45.98	7.88	50.69	11.98	7.88
3 2	11.10	3.03	12.26	20.48	56.59	56.59	56.59	56.59	9.70	3.03	5.00	9.70
3 3	2.59	66.05	7.54	12.59	34.77	34.77	34.77	34.77	5.96	66.05	25.46	5.96
3 4	2.38	2.42	8.35	13.95	38.53	38.53	38.53	38.53	6.60	2.42	77.09	6.60
3 5	29.12	29.12	8.91	14.88	41.11	41.11	41.11	41.11	7.05	40.01	85.42	7.05
4 1	2.02	41.66	9.94	16.60	45.86	45.86	45.86	45.86	7.86	41.66	91.14	7.86
4 2	42.80	29.54	13.90	23.22	64.13	64.13	64.13	64.13	10.99	29.54	21.66	10.99
4 3	54.77	50.77	9.26	15.46	42.72	42.72	42.72	42.72	7.32	50.77	42.1	7.32
4 4	7.36	18.08	10.27	17.16	47.41	47.41	47.41	47.41	8.13	18.08	94.71	8.13
4 5	2.37	7.53	9.52	15.91	43.95	43.95	43.95	43.95	7.53	97.44	51.01	7.53
All-patterns	32.61	26.31	5.80	5.68	22.77	24.98	22.77	24.98	26.31	50.66	26.31	26.31

Table C.3 - Comparison of the sampling using original anchoring docking pipeine [ref] to the updated pipeline (§5.3) for the fragment 2 (GUU, chain P, nucleotides 4-5-6, termed 'AMF' in NAR paper) of the complex 1B7F.
Presented numbers are not directly comparable, as different numbers of the starting points have been used and different numbers of top-ranked poses have been kept.

	Original Pipeline		**Anchoring Patterns Pipeline (all-patterns)**	
	bound	**unbound**	**unbound**	**unbound**
min LRMSD, Å	1	1	1.3	1.1
LRMSD<2Å	20	14	14	81
LRMSD<3Å	94	74	428	1481
nb starting points	1 million	1 million	10x20 docking runs	50x20 docking runs
nb. top-ranked poes kept	1000	1000	all sampled	all sampled

C.4 Collection of the Non-Structural Data

Proteins of Interest

The data was collected for the following proteins:

F1LQ48 G5ECJ4 G5EEW7 H2L051 O00425 O45189 O95319 P05455 P07910 P08199 P08579 P09012 P09651 P11940 P19339 P22626 P25299 P26368 P26378 P26599 P31483 P35637 P38159 P38996 P49960 P52597 P53617 P62995 P84103 Q01130 Q07955 Q13148 Q14103 Q15717 Q16630 Q17RY0 Q22039 Q389P7 Q61474 Q64368 Q8I3T5 Q921F2 Q92879 Q93062Q99383 Q9BZB8 Q9NWB1 Q9UHX1 Q9Y5S9

C.5 RRMScorer for Identification of Anchors Positions

RRMScorer

RRMScorer, a tool developed within RNAct, predicts a given RNA sequence's probability of binding a target RRM sequence in a canonical binding mode [ref]. It uses both the alignments of RRM and RNA to generate a scoring system and estimate RNA binding affinity. The alignments of multiple RRM-RNA structures, performed during the development of this tool, revealed 20 amino acids that are important for binding. These amino acids and 5 consecutive nucleotides form 30 contacts, frequently observed in the structures with the canonical binding mode (Fig. C.1).

RRMScorer scores these 30 contacts and outputs an overall binding probability. It also generates a table with scores, one table for each contact (Fig C.2). In this table, each row corresponds to an amino acid type seen in the specified alignment position. Each column corresponds to a nucleotide type, plus a column for the case of contact absence. The values in the table give the probability of each contact for each type of the residues. Positive values indicate that a specific amino

acid-nucleotide contact is likely to be encountered, while negative scores have the opposite meaning. Scores close to 0 indicate that there is no clear preference. Please note that RRMScorer scores are not connected in any way with the scoring stage of the docking.

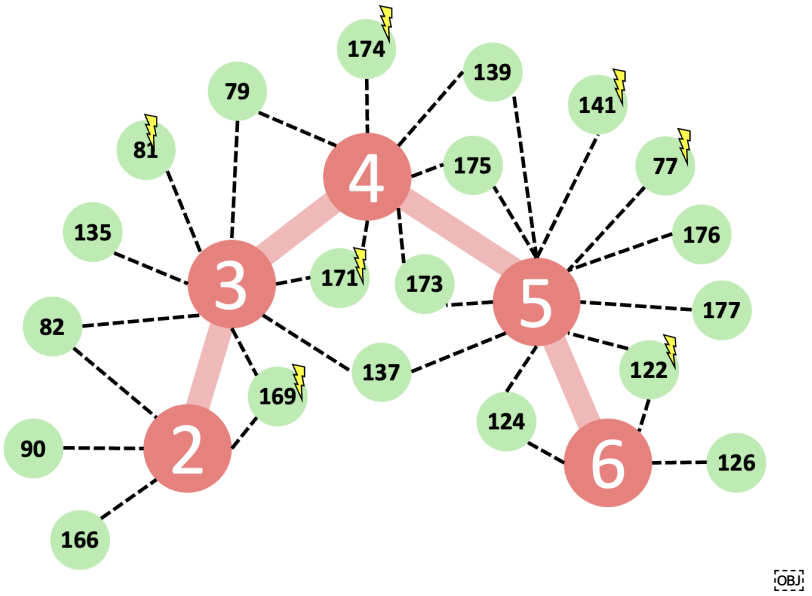


Figure C.1 - Typical for canonical RRM-RNA binding mode contacts (shown as black dashed lines). Amino acids are portrayed in green and nucleotides in pink. The indexes, both for RNA and RRM, correspond to the Master alignment [Image by Joel Roca Martínez].

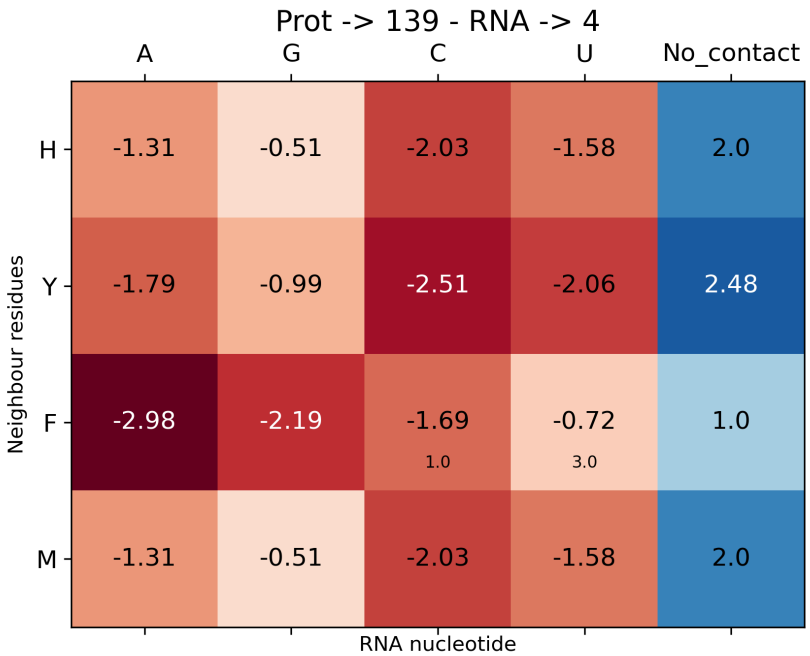


Figure C.2 - RRMScorer output table for the contact between amino acid in position 139 and nucleotide in position 4.

Problem Statement

When using the RRM-RNA dock, the user has to identify the sequence of a fragment to be docked, along with the positions of the anchors, which can present an unwanted challenge. It could be possible to use RRMScorer's output tables to predict the positions of anchors within a trinucleotide for a given RRM. If this approach proves viable, it can be incorporated into the pipeline to facilitate the RRM-RNA dock usage and to expand its usability.

The hypothesis is that certain scores are indicative of the stacking interactions exclusively and that these scores can be used to identify the most probable indexes of anchors given amino acids types and indexes. To test this hypothesis, first I need to obtain scores, given for the real stacking interactions and estimate the threshold value(s) that can distinguish between the stacking interaction and the other type of interactions for a given nucleotide index within alignment. Next, I will use these thresholds to predict the combination of nucleotides' indexes given the indexes and types of the anchoring amino acids, and subsequently calculate the accuracy of these predictions.

Since the current version of the RRM-RNA dock operates on a single fragment containing 2 anchors, I focus on the identification of the pair of stacking nucleotides, either adjacent or separated by a single nucleotide. All possible positions of the adjacent stacking nucleotides in terms of RRMScorer alignment are [2,3], [3,4], [4,5], [5,6]. For non-adjacent pairs [2,4], [3,5], [4,6].

Protocol & Data

A total of 155 RRM-RNA structures containing the 301 labelled stacking interactions were taken from InteR3M. This dataset comprised 36 proteins with distinct UniProt IDs. Out of these, 2 proteins did not exhibit stacking interactions and were used as a negative control.

First, the scores associated with all real stacking interactions (from the table for given amino_acid_index and nucleotide_index, row for given amino_acid_type, column 'no_contact') were obtained from all available complexes. The column 'no_contact' was used instead of the one corresponding to the given nucleotide_type, as the application of the protocol to the values given by nucleotide_type did not produce accurate predictions.

Next, 2 empirical approaches were applied to obtain the representative scores for each nucleotide_index:

- In 'mod1' an average score was taken (*avg_score*);
- In 'mod2' the scores with frequency over 25% were taken (*freq_score1, freq_score2, etc.*)

To predict anchor positions (stacking nucleotide indexes) for a particular RRM, the following protocol was applied:

1. For a given RRM sequence, get the positions of the target amino acids (Fig C.1) via the Master alignment;
2. Determine the type of each target amino acid;
3. Create a list of possible contacts (amino_acids_index; nucleotide_index) (following Fig C.1);
4. For each possible contact obtain a score given a corresponding RRMScorer table, row amino_acid_type, column 'no_contact';
5. For each possible contact, check if the obtained score value:
    - **is equal or less than *avg_score* for a given nucleotide_index in 'mod1'**;
    - **is equal to one of the *freq_score* in 'mod2'**.

    Create a list of probable contacts out of the contacts which meet the requirements;
6. Pair probable contacts into adjacent pairs and non-adjacent pairs.

While the application of this protocol with ‘mod1’ did not provide very accurate predictions, the ‘mod2’ produced promising preliminary results with an accuracy of 83% (Tab C.4)

Table C.4 - Confusion matrix of the predicted pairs of contacts with ‘mod2’.

	Real pairs of contacts		
Predicted pairs of contacts		**Positives**	**Negatives**
	Positives	251	30
	Negatives	20	NaN

Results and Discussion

To validate this approach, k-fold cross-validation was used and the dataset was divided into k=14 folds (sets), each fold containing from 1 to 6 proteins (Tab C.5).

Table C.5 - Set up for the cross-validation.

Name	**Proteins**	**Stacking count**
‘fold_1’	F1LQ48; P26378; P38159; P04147	8
‘fold_2’	Q61474; O95319; P07910	7
‘fold_3’	Q00916; P08621; P08579; Q93062	11
‘fold_4’	Q01130; Q64368; P19339	12
‘fold_5’	Q06AA4; P62995; P26368; P09651	13
‘fold_6’	Q9NWB1; P38996; O00425; P84103; Q7LL14; P26599	6
‘fold_7’	P0DJD3; O45189; P49960	11
‘fold_8’	P22626; G5EEW7	6
‘fold_9’	Q13148; Q99383; P43332	11
‘fold_10’	P11940	18
‘fold_11’	Q15717	31
‘fold_12’	P09012	108
‘fold_13’	Q92879	9
‘fold_14’	Q99181; G5ECJ4	0

As RRMScorer was trained on the same data, it has to be re-trained to avoid second-hand bias. RRMScorer was re-trained on each training set (k−1 folds). Then, the *freq_score*s were obtained using

re-trained RRMScorer tables and real stacking interactions from the training set (Tab. C.6). The described protocol with obtained *freq_score*s was applied to the test set (a fold left out of the training set).

The most accurate predictions with an average accuracy of 70% were obtained if the protocol (using 'mod2') was limited with amino_acis_type = {Y, W, F} and nucleotide_index = {3,4,5} (Fig C.3). The details per each test set, are given in Table C.7.

Table C.6 - The *freq_score* values for each training set.
In this table, each row corresponds to a training set obtained by excluding indicted in 'out fold' fold out of the dataset.

Out fold	Nucleotide in position 3	Nucleotide in position 4	Nucleotide in position 5
fold_1	0.25	-0.18, 0.47	0.05
fold_2	0.25	0, 0.31	0.05
fold_3	0.25	-0.14, 0.31	0.02
fold_4	0.25	-0.15, 0.31	0.06
fold_5	0.38	-0.11, 0.38	0.07
fold_6	0.38	-0.18, 0.47	0.06
fold_7	0.25	0.31, -0.13	0.01
fold_8	0.25	-0.19, 0.31	0
fold_9	0.38	-0.18, 0.38	0
fold_10	0.25	-0.18, 0.31	-0.01
fold_11	0.25	-0.15, 0.31	0.03
fold_12	0.37	-0.2	0
fold_13	0.25	-0.05, 0.31	0.02
fold_14	0.25	-0.2, 0.31	-0.02

Table C.7 - Evaluation of the prediction made using 'mod2' for each test set.

Test set	Accuracy, %	False positives count	False negatives count	True positives count
fold_1	88	0	1	7
fold_2	33	5	3	4
fold_3	92	1	0	11
fold_4	80	3	0	12
fold_5	65	4	3	13
fold_6	75	2	0	6

Test set	Accuracy, %	False positives count	False negatives count	True positives count
fold_7	73	0	3	8
fold_8	21	8	3	3
fold_9	85	2	0	11
fold_10	71	3	3	15
fold_11	70	13	0	31
fold_12	50	0	54	54
fold_13	100	0	0	9
fold_14	100	0	0	0

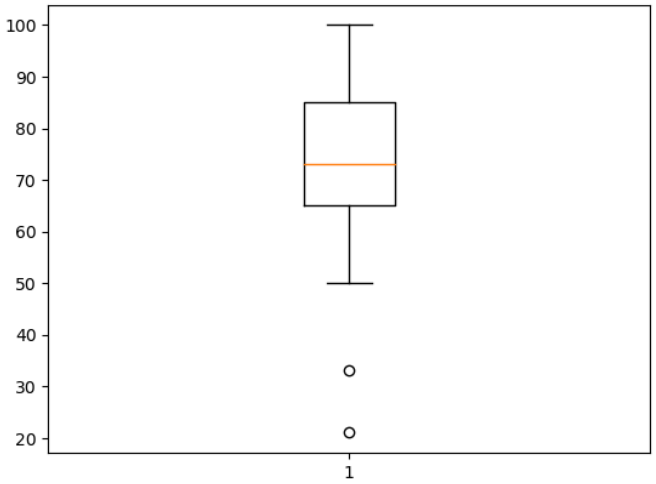


Figure C.3 - Boxplot describing the accuracy of the predictions made across each test set (in ‘mod2’). The minimal accuracy is 21%, the maximal accuracy is 100%. The average accuracy is ~70%.

While bringing optimistic results, ‘mod2’, i.e. using the most frequent RRMScore values as an exact match is not ideal. So we tested ‘mod1’ using the same set-up (amino_acis_type = {Y, W, F} and nucleotide_index = {3,4,5}) (Tab C.9). The average accuracy is 64%. The *avg_score* values for each training set can be found in Tab C.8.

Table C.8 - The *avg_score* values for each training set.

Out fold	Nucleotide in position 3	Nucleotide in position 4	Nucleotide in position 5
fold_1	0.3	0.22	0.06
fold_2	0.27	0.19	0.06
fold_3	0.29	0.15	0.03
fold_4	0.28	0,15	0.07
fold_5	0.4	0,2	0.08

fold_6	0.41	0.22	0.07
fold_7	0.28	0.16	0.07
fold_8	0.29	0.12	0.01
fold_9	0.41	0.16	0.01
fold_10	0.28	0.15	-0.01
fold_11	0.29	0.17	0.04
fold_12	0.51	-0.03	0.02
fold_13	0.28	0.18	0.03
fold_14	0.28	0.12	-0.01

Table C.9 - Evaluation of the prediction made using ‘mod1’ for each test set.

Test set	Accuracy, %	False positives count	False negatives count	True positives count
fold_1	62	0	3	5
fold_2	33	5	3	4
fold_3	75	1	2	9
fold_4	59	5	2	10
fold_5	58	3	5	11
fold_6	83	0	1	5
fold_7	73	0	3	8
fold_8	20	9	3	3
fold_9	46	2	5	6
fold_10	79	1	3	15
fold_11	91	3	0	31
fold_12	50	0	54	54
fold_13	100	0	0	9
fold_14	100	0	0	0

While optimistic, these results leave room for improvement. The use of the exact matches (‘mod2’ with *freq_score*s), while resulting in acceptable predictions for most test sets, is not an ideal approach methodologically. Simultaneously, the use of the average scores (‘mod1’) as a threshold lowers the accuracy. Thus, other approaches, empirical or otherwise, to obtain the threshold values can be explored as well. It could be possible to incorporate the statistics of the occurring stacking interactions to filter out false positives. Additionally, a version of RRMSorer trained on the interaction

of the target amino acids and Phosphate exists. It could be used to identify corresponding contacts in a given RRM-RNA complex.

Lastly, this protocol was created and evaluated for the prediction of the pair of stacking interactions within a fragment to match the current functionality of the RRM-RNA dock. However, it can be modified to predict a single most probable stacking interaction within a 5-nucleotide chain.

Bibliography

1. Schrödinger, L., & DeLano, W. (2020). PyMOL. Retrieved from http://www.pymol.org/pymol
2. Feenstra K.A., & Abeln S. (2023). Introduction to Protein Structural Bioinformatics (1st ed.) Biomolecules. https://doi.org/10.48550/arXiv.1801.09442
3. Branden, C.I., & Tooze, J. (1998). Introduction to Protein Structure (2nd ed.). Garland Science. https://doi.org/10.1201/9781136969898
4. Kloppmann, E., Reeb, J., Hönigschmid, P., Rost, B. (2019). Protein Secondary Structure Prediction in 2018. In: Roberts, G., Watts, A. (eds) Encyclopedia of Biophysics. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-35943-9_429-1
5. Letunic, I., Khedkar, S., Bork, P. (2021). SMART: recent updates, new developments and status in 2020, Nucleic Acids Research, 49(D1), D458–D460. https://doi.org/10.1093/nar/gkaa937
6. Paysan-Lafosse, T., Blum, M., Chuguransky, S., Grego, T., Pinto, B. L., Salazar, G. A., Bileschi, M. L., Bork, P., Bridge, A., Colwell, L., Gough, J., Haft, D. H., Letunić, I., Marchler-Bauer, A., Mi, H., Natale, D. A., Orengo, C. A., Pandurangan, A. P., Rivoire, C., Sigrist, C. J. A., … Bateman, A. (2023). InterPro in 2022. Nucleic acids research, 51(D1), D418–D427. https://doi.org/10.1093/nar/gkac993
7. Cléry A, H.-T. Allain F. FROM STRUCTURE TO FUNCTION OF RNA BINDING DOMAINS. In: Madame Curie Bioscience Database [Internet]. Austin (TX): Landes Bioscience; 2000-2013. Available from: https://www.ncbi.nlm.nih.gov/books/NBK63528/
8. Uversky, V.N. (2019). Intrinsically Disordered Proteins and Their "Mysterious" (Meta)Physics. Frontiers in Physics. (7) https://doi.org/10.3389/fphy.2019.00010
9. Wang, D., & Farhana, A. (2023). Biochemistry, RNA Structure. In StatPearls. StatPearls Publishing. PMID: 32644425 Bookshelf ID: NBK558999
10. Fay, M. M., Lyons, S. M., & Ivanov, P. (2017). RNA G-Quadruplexes in Biology: Principles and Molecular Mechanisms. Journal of molecular biology, 429(14), 2127–2147. https://doi.org/10.1016/j.jmb.2017.05.017
11. Kharel, P., Becker, G., Tsvetkov, V., & Ivanov, P. (2020). Properties and biological impact of RNA G-quadruplexes: from order to turmoil and back. Nucleic acids research, 48(22), 12534–12555. https://doi.org/10.1093/nar/gkaa1126
12. Lemieux, S., & Major, F. (2002). RNA canonical and non-canonical base pairing types: a recognition method and complete repertoire. Nucleic acids research, 30(19), 4250–4263. https://doi.org/10.1093/nar/gkf540
13. Leontis, N. B., Stombaugh, J., & Westhof, E. (2002). The non-Watson-Crick base pairs and their associated isostericity matrices. Nucleic acids research, 30(16), 3497–3531. https://doi.org/10.1093/nar/gkf481
14. Westhof E. (2014). Isostericity and tautomerism of base pairs in nucleic acids. FEBS letters, 588(15), 2464–2469. https://doi.org/10.1016/j.febslet.2014.06.031
15. Stombaugh, J., Zirbel, C. L., Westhof, E., & Leontis, N. B. (2009). Frequency and isostericity of RNA base pairs. Nucleic acids research, 37(7), 2294–2312. https://doi.org/10.1093/nar/gkp011

16. Icazatti, A. A., Loyola, J. M., Szleifer, I., Vila, J. A., & Martin, O. A. (2019). Classification of RNA backbone conformations into rotamers using 13C' chemical shifts: exploring how far we can go. PeerJ, 7, e7904. https://doi.org/10.7717/peerj.7904
17. Šponer, J., Mládek, A., Šponer, J. E., Svozil, D., Zgarbová, M., Banáš, P., … Otyepka, M. (2011). The DNA and RNA sugar–phosphate backbone emerges as the key player. An overview of quantum-chemical, structural biology and simulation studies. Physical Chemistry Chemical Physics, 14(44), 15257. https://doi.org/10.1039/c2cp41987d
18. Richardson, J. S., Schneider, B., Murray, L. W., Kapral, G. J., Immormino, R. M., Headd, J. J., Richardson, D. C., Ham, D., Hershkovits, E., Williams, L. D., Keating, K. S., Pyle, A. M., Micallef, D., Westbrook, J., Berman, H. M., & RNA Ontology Consortium (2008). RNA backbone: consensus all-angle conformers and modular string nomenclature (an RNA Ontology Consortium contribution). RNA (New York, N.Y.), 14(3), 465–481. https://doi.org/10.1261/rna.657708
19. Šponer, J., Bussi, G., Krepl, M., Banáš, P., Bottaro, S., Cunha, R. A., Gil-Ley, A., Pinamonti, G., Poblete, S., Jurečka, P., Walter, N. G., & Otyepka, M. (2018). RNA Structural Dynamics As Captured by Molecular Simulations: A Comprehensive Overview. Chemical Reviews, 118(8), 4177–4338. https://doi.org/10.1021/acs.chemrev.7b00427
20. ViennRNA Web Services. RNA Secondary Structure Visualization Using a Force Directed Graph Layout. Available from: http://rna.tbi.univie.ac.at/forna/
21. Rangarajan, E. S., Lee, J. H., & Izard, T. (2011). Apo raver1 structure reveals distinct RRM domain orientations. Protein Science: a publication of the Protein Society, 20(8), 1464–1470. https://doi.org/10.1002/pro.664
22. Menger, M., Eckstein, F., & Porschke, D. (2000). Multiple conformational states of the hammerhead ribozyme, broad time range of relaxation and topology of dynamics. Nucleic acids research, 28(22), 4428–4434. https://doi.org/10.1093/nar/28.22.4428
23. London R. E. (2019). HIV-1 Reverse Transcriptase: A Metamorphic Protein with Three Stable States. Structure (London, England: 1993), 27(3), 420–426. https://doi.org/10.1016/j.str.2018.11.011
24. Xiang, S., Gapsys, V., Kim, H. Y., Bessonov, S., Hsiao, H. H., Möhlmann, S., Klaukien, V., Ficner, R., Becker, S., Urlaub, H., Lührmann, R., de Groot, B., & Zweckstetter, M. (2013). Phosphorylation drives a dynamic switch in serine/arginine-rich proteins. Structure (London, England: 1993), 21(12), 2162–2174. https://doi.org/10.1016/j.str.2013.09.014
25. Bheemireddy, S., Sandhya, S., Srinivasan, N., & Sowdhamini, R. (2022). Computational tools to study RNA-protein complexes. Frontiers in molecular biosciences, 9, 954926. https://doi.org/10.3389/fmolb.2022.954926
26. Corley, M., Burns, M. C., & Yeo, G. W. (2020). How RNA-Binding Proteins Interact with RNA: Molecules and Mechanisms. Molecular cell, 78(1), 9–29. https://doi.org/10.1016/j.molcel.2020.03.011
27. Sigel, A., Operschall, B. P., & Sigel, H. (2014). Comparison of the π-stacking properties of purine versus pyrimidine residues. Some generalizations regarding selectivity. Journal of Biological Inorganic Chemistry: JBIC: a publication of the Society of Biological Inorganic Chemistry, 19(4-5), 691–703. https://doi.org/10.1007/s00775-013-1082-5
28. Wilson, K. A., Holland, D. J., & Wetmore, S. D. (2016). Topology of RNA-protein nucleobase-amino acid π-π interactions and comparison to analogous DNA-protein π-π contacts. RNA (New York, N.Y.), 22(5), 696–708. https://doi.org/10.1261/rna.054924.115
29. Morozova, N., Allers, J., Myers, J., & Shamoo, Y. (2006). Protein-RNA interactions: exploring binding patterns with a three-dimensional superposition analysis of high resolution

structures. Bioinformatics (Oxford, England), 22(22), 2746–2752. https://doi.org/10.1093/bioinformatics/btl470

30. Bergsma, S., Poulios, E., Charalampogiannis, N., Andraws, O., & Achinas, S. (2022). Cation-π Interaction as a Key Player in Healthcare: A Mini-Review. IntechOpen. http://dx.doi.org/10.5772/dmht.11.
31. Anbarasu, A., Anand, S., Mathew, L., & Sethumadhavan, R. (2007). Influence of cation-pi interactions on RNA-binding proteins. International journal of biological macromolecules, 40(5), 479–483. https://doi.org/10.1016/j.ijbiomac.2006.11.008
32. Zondlo N. J. (2013). Aromatic-proline interactions: electronically tunable CH/π interactions. Accounts of chemical research, 46(4), 1039–1049. https://doi.org/10.1021/ar300087y
33. Xiao, F., Chen, Z., Wei, Z., & Tian, L. (2020). Hydrophobic Interaction: A Promising Driving Force for the Biomedical Applications of Nucleic Acids. Advanced science (Weinheim, Baden-Wurttemberg, Germany), 7(16), 2001048. https://doi.org/10.1002/advs.202001048
34. Tan, Z. J., & Chen, S. J. (2011). Importance of diffuse metal ion binding to RNA. Metal ions in life sciences, 9, 101–124. PMID: 22010269 PMCID: PMC4883094.
35. Dudev, T., & Lim, C. (2014). Competition among metal ions for protein binding sites: determinants of metal ion selectivity in proteins. Chemical Reviews, 114(1), 538–556. https://doi.org/10.1021/cr4004665
36. Giacobelli, V. G., Fujishima, K., Lepšík, M., Tretyachenko, V., Kadavá, T., Makarov, M., Bednárová, L., Novák, P., & Hlouchová, K. (2022). In Vitro Evolution Reveals Noncationic Protein-RNA Interaction Mediated by Metal Ions. Molecular biology and evolution, 39(3), msac032. https://doi.org/10.1093/molbev/msac032
37. Itoh, Y., Nakashima, Y., Tsukamoto, S., Kurohara, T., Suzuki, M., Sakae, Y., Oda, M., Okamoto, Y., & Suzuki, T. (2019). N+-C-H···O Hydrogen bonds in protein-ligand complexes. Scientific reports, 9(1), 767. https://doi.org/10.1038/s41598-018-36987-9
38. Rolfsson, Ó., Toropova, K., Ranson, N. A., & Stockley, P. G. (2010). Mutually-induced conformational switching of RNA and coat protein underpins efficient assembly of a viral capsid. Journal of molecular biology, 401(2), 309–322. https://doi.org/10.1016/j.jmb.2010.05.058
39. Hassan, A., Byju, S., Freitas, F. C., Roc, C., Pender, N., Nguyen, K., Kimbrough, E. M., Mattingly, J. M., Gonzalez, R. L., Jr, de Oliveira, R. J., Dunham, C. M., & Whitford, P. C. (2023). Ratchet, swivel, tilt and roll: a complete description of subunit rotation in the ribosome. Nucleic acids research, 51(2), 919–934. https://doi.org/10.1093/nar/gkac1211
40. Seif, E., & Hallberg, B. M. (2009). RNA-protein mutually induced fit: structure of Escherichia coli isopentenyl-tRNA transferase in complex with tRNA(Phe). The Journal of Biological Chemistry, 284(11), 6600–6604. https://doi.org/10.1074/jbc.C800235200
41. Yu, B., Pettitt, B. M., & Iwahara, J. (2020). Dynamics of Ionic Interactions at Protein-Nucleic Acid Interfaces. Accounts of chemical research, 53(9), 1802–1810. https://doi.org/10.1021/acs.accounts.0c00212
42. Peach, M. L., Cachau, R. E., & Nicklaus, M. C. (2017). Conformational energy range of ligands in protein crystal structures: The difficult quest for accurate understanding. Journal of molecular recognition: JMR, 30(8), 10.1002/jmr.2618. https://doi.org/10.1002/jmr.2618
43. Valverde, R., Edwards, L., & Regan, L. (2008). Structure and function of KH domains. The FEBS journal, 275(11), 2712–2726. https://doi.org/10.1111/j.1742-4658.2008.06411.x
44. Wang, M., Ogé, L., Perez-Garcia, M. D., Hamama, L., & Sakr, S. (2018). The PUF Protein Family: Overview on PUF RNA Targets, Biological Functions, and Post Transcriptional Regulation. International journal of molecular sciences, 19(2), 410. https://doi.org/10.3390/ijms19020410

45. Chen, Y., & Varani, G. (2011). Finding the missing code of RNA recognition by PUF proteins. Chemistry & biology, 18(7), 821–823. https://doi.org/10.1016/j.chembiol.2011.07.001
46. Re, A., Joshi, T., Kulberkyte, E., Morris, Q., & Workman, C. T. (2014). RNA-protein interactions: an overview. Methods in molecular biology (Clifton, N.J.), 1097, 491–521. https://doi.org/10.1007/978-1-62703-709-9_23
47. Balcerak, A., Trebinska-Stryjewska, A., Konopinski, R., Wakula, M., & Grzybowska, E. A. (2019). RNA-protein interactions: disorder, moonlighting and junk contribute to eukaryotic complexity. *Open biology*, 9(6), 190096. https://doi.org/10.1098/rsob.190096
48. Cruz-Gallardo, I., Aroca, Á., Gunzburg, M. J., Sivakumaran, A., Yoon, J. H., Angulo, J., Persson, C., Gorospe, M., Karlsson, B. G., Wilce, J. A., & Díaz-Moreno, I. (2014). The binding of TIA-1 to RNA C-rich sequences is driven by its C-terminal RRM domain. *RNA biology*, 11(6), 766–776. https://doi.org/10.4161/rna.28801
49. Han, A., Stoilov, P., Linares, A. J., Zhou, Y., Fu, X. D., & Black, D. L. (2014). De novo prediction of PTBP1 binding and splicing targets reveals unexpected features of its RNA recognition and function. *PLoS Computational Biology*, 10(1), e1003442. https://doi.org/10.1371/journal.pcbi.1003442
50. Gebauer, F., Schwarzl, T., Valcárcel, J., & Hentze, M. W. (2021). RNA-binding proteins in human genetic disease. *Nature reviews. Genetics*, 22(3), 185–198. https://doi.org/10.1038/s41576-020-00302-y
51. Kelaini, S., Chan, C., Cornelius, V. A., & Margariti, A. (2021). RNA-Binding Proteins Hold Key Roles in Function, Dysfunction, and Disease. *Biology*, 10(5), 366. https://doi.org/10.3390/biology10050366
52. Qin, H., Ni, H., Liu, Y., Yuan, Y., Xi, T., Li, X., & Zheng, L. (2020). RNA-binding proteins in tumor progression. *Journal of hematology & oncology*, 13(1), 90. https://doi.org/10.1186/s13045-020-00927-w
53. Schultz, C. W., Preet, R., Dhir, T., Dixon, D. A., & Brody, J. R. (2020). Understanding and targeting the disease-related RNA binding protein human antigen R (HuR). *Wiley interdisciplinary reviews. RNA*, 11(3), e1581. https://doi.org/10.1002/wrna.1581
54. Hanson, K. A., Kim, S. H., & Tibbetts, R. S. (2012). RNA-binding proteins in neurodegenerative disease: TDP-43 and beyond. *Wiley interdisciplinary reviews. RNA*, 3(2), 265–285. https://doi.org/10.1002/wrna.111
55. Liu, J., & Cao, X. (2023). RBP-RNA interactions in the control of autoimmunity and autoinflammation. *Cell Research*, 33(2), 97–115. https://doi.org/10.1038/s41422-022-00752-5
56. de Bruin, R. G., Rabelink, T. J., van Zonneveld, A. J., & van der Veer, E. P. (2017). Emerging roles for RNA-binding proteins as effectors and regulators of cardiovascular disease. *European Heart Journal*, 38(18), 1380–1388. https://doi.org/10.1093/eurheartj/ehw567
57. Liu, W., Li, D., Lu, T., Zhang, H., Chen, Z., Ruan, Q., Zheng, Z., Chen, L., & Guo, J. (2023). Comprehensive analysis of RNA-binding protein SRSF2-dependent alternative splicing signature in malignant proliferation of colorectal carcinoma. *The Journal of Biological Chemistry*, 299(2), 102876. https://doi.org/10.1016/j.jbc.2023.102876
58. François-Moutal, L., Perez-Miller, S., Scott, D. D., Miranda, V. G., Mollasalehi, N., & Khanna, M. (2019). Structural Insights Into TDP-43 and Effects of Post-translational Modifications. *Frontiers in molecular neuroscience*, 12, 301. https://doi.org/10.3389/fnmol.2019.00301
59. Kamel, W., Noerenberg, M., Cerikan, B., … Castello, A. (2021). Global analysis of protein-RNA interactions in SARS-CoV-2-infected cells reveals key regulators of infection. *Molecular cell*, 81(13), 2851–2867.e7. https://doi.org/10.1016/j.molcel.2021.05.023

60. Cen, Y., Chen, L., Liu, Z., Lin, Q., Fang, X., Yao, H., & Gong, C. (2023). Novel roles of RNA-binding proteins in drug resistance of breast cancer: from molecular biology to targeting therapeutics. *Cell death discovery*, 9(1), 52. https://doi.org/10.1038/s41420-023-01352-x
61. Kudinov, A. E., Karanicolas, J., Golemis, E. A., & Boumber, Y. (2017). Musashi RNA-Binding Proteins as Cancer Drivers and Novel Therapeutic Targets. *Clinical Cancer Research : an official journal of the American Association for Cancer Research,* 23(9), 2143–2153. https://doi.org/10.1158/1078-0432.CCR-16-2728
62. Kazianka, L., & Staber, P. B. (2021). Blood cancer driver Musashi-2 as therapeutic target in chronic lymphocytic leukemia. *Leukemia*, 35(4), 982–983. https://doi.org/10.1038/s41375-021-01144-1
63. Chen, Y., Yang, F., Zubovic, L., Pavelitz, T., Yang, W., Godin, K., Walker, M., Zheng, S., Macchi, P., & Varani, G. (2016). Targeted inhibition of oncogenic miR-21 maturation with designed RNA-binding proteins. *Nature Chemical Biology*, 12(9), 717–723. https://doi.org/10.1038/nchembio.2128
64. Donsbach, P., & Klostermeier, D. (2021). Regulation of RNA helicase activity: principles and examples. *Biological Chemistry*, 402(5), 529–559. https://doi.org/10.1515/hsz-2020-0362
65. Li, P. T., Vieregg, J., & Tinoco, I., Jr (2008). How RNA unfolds and refolds. *Annual review of biochemistry*, 77, 77–100. https://doi.org/10.1146/annurev.biochem.77.061206.174353
66. Pal, A., & Levy, Y. (2019). Structure, stability and specificity of the binding of ssDNA and ssRNA with proteins. *PLoS Computational Biology*, 15(4), e1006768. https://doi.org/10.1371/journal.pcbi.1006768
67. Auweter, S. D., Oberstrass, F. C., & Allain, F. H. (2006). Sequence-specific binding of single-stranded RNA: is there a code for recognition? *Nucleic acids research*, 34(17), 4943–4959. https://doi.org/10.1093/nar/gkl620
68. Jankowsky, E., & Harris, M. E. (2015). Specificity and nonspecificity in RNA-protein interactions. *Nature reviews. Molecular cell biology*, 16(9), 533–544. https://doi.org/10.1038/nrm4032
69. Hudson, B. P., Martinez-Yamout, M. A., Dyson, H. J., & Wright, P. E. (2004). Recognition of the mRNA AU-rich element by the zinc finger domain of TIS11d. *Nature structural & molecular biology,* 11(3), 257–264. https://doi.org/10.1038/nsmb738
70. Ye, X., Yang, W., Yi, S., Zhao, Y., Varani, G., Jankowsky, E., & Yang, F. (2023). Two distinct binding modes provide the RNA-binding protein RbFox with extraordinary sequence specificity. *Nature communications*, 14(1), 701. https://doi.org/10.1038/s41467-023-36394-3
71. Maris, C., Dominguez, C., & Allain, F. H. (2005). The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression. *The FEBS journal*, 272(9), 2118–2131. https://doi.org/10.1111/j.1742-4658.2005.04653.x
72. Clingman, C. C., Deveau, L. M., Hay, S. A., Genga, R. M., Shandilya, S. M., Massi, F., & Ryder, S. P. (2014). Allosteric inhibition of a stem cell RNA-binding protein by an intermediary metabolite. *eLife*, 3, e02848. https://doi.org/10.7554/eLife.02848
73. Samatanga, B., Cléry, A., Barraud, P., Allain, F. H., & Jelesarov, I. (2017). Comparative analyses of the thermodynamic RNA binding signatures of different types of RNA recognition motifs. *Nucleic acids research*, 45(10), 6037–6050. https://doi.org/10.1093/nar/gkx136
74. SenGupta, D. (2013). RNA-binding domains in proteins. In: Brenner's encyclopedia of genetics, (2nd edn.) Elsevier, Amsterdam, pp 274–276 https://doi.org/10.1016/B978-0-12-374984-0.01356-5
75. Timofeev, V., & Samygina, V. (2023). Protein Crystallography: Achievements and Challenges. *Crystals*, 13(1), 71. https://doi.org/10.3390/cryst13010071

76. Niedzialkowska, E., Gasiorowska, O., Handing, K. B., Majorek, K. A., Porebski, P. J., Shabalin, I. G., Zasadzinska, E., Cymborowski, M., & Minor, W. (2016). Protein purification and crystallization artifacts: The tale usually not told. Protein Science: a publication of the Protein Society, 25(3), 720–733. https://doi.org/10.1002/pro.2861
77. Klebe, G. (2013). Experimental Methods of Structure Determination. In: Klebe, G. (eds) Drug Design. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-17907-5_13
78. Puthenveetil, R., & Vinogradova, O. (2019). Solution NMR: A powerful tool for structural and functional studies of membrane proteins in reconstituted environments. *The Journal of Biological Chemistry*, 294(44), 15914–15931. https://doi.org/10.1074/jbc.REV119.009178
79. Nakane, T., Kotecha, A., Sente, A., McMullan, G., … Scheres, S. H. W. (2020). Single-particle cryo-EM at atomic resolution. *Nature*, 587(7832), 152–156. https://doi.org/10.1038/s41586-020-2829-0
80. Guaita, M., Watters, S. C., & Loerch, S. (2022). Recent advances and current trends in cryo-electron microscopy. *Current opinion in structural biology*, 77, 102484. https://doi.org/10.1016/j.sbi.2022.102484
81. Krishnamoorthy, G. K., Alluvada, P., Hameed Mohammed Sherieff, S., Kwa, T., & Krishnamoorthy, J. (2019). Isothermal titration calorimetry and surface plasmon resonance analysis using the dynamic approach. *Biochemistry and biophysics reports*, 21, 100712. https://doi.org/10.1016/j.bbrep.2019.100712
82. López-Rubio, E., Ratti, E. (2021) Data science and molecular biology: prediction and mechanistic explanation. *Synthese* 198, 3131–3156. https://doi.org/10.1007/s11229-019-02271-0
83. Morrison-Smith, S., Boucher, C., Bunt, A., and Ruiz, J. (2015) Elucidating the role and use of bioinformatics software in life science research. In Proceedings of the 2015 British HCI Conference. ACM, New York, 230-238. https://doi.org/10.1145/2783446.2783581
84. Forster M. J. (2002). Molecular modelling in structural biology. *Micron (Oxford, England : 1993)*, *33*(4), 365–384. https://doi.org/10.1016/s0968-4328(01)00035-x
85. Varadi, M., Nair, S., Sillitoe, I., Tauriello, G., … Velankar, S. (2022). 3D-Beacons: decreasing the gap between protein sequences and structures through a federated network of protein structure data resources. *GigaScience*, *11*, giac118. https://doi.org/10.1093/gigascience/giac118
86. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., & Bourne, P. E. (2000). The Protein Data Bank. *Nucleic acids research*, *28*(1), 235–242. https://doi.org/10.1093/nar/28.1.235
87. Burley, S. K., Bhikadiya, C., Bi, C., Bittrich, S., Chao, H.,… Zardecki, C. (2023). RCSB Protein Data Bank (RCSB.org): delivery of experimentally-determined PDB structures alongside one million computed structure models of proteins from artificial intelligence/machine learning. *Nucleic acids research*, *51*(D1), D488–D508. https://doi.org/10.1093/nar/gkac1077
88. PDB Statistics (Website). Available at https://www.rcsb.org/stats/
89. Badal, V. D., Kundrotas, P. J., & Vakser, I. A. (2018). Natural language processing in text mining for structural modeling of protein complexes. *BMC bioinformatics*, *19*(1), 84. https://doi.org/10.1186/s12859-018-2079-4
90. Lubiana, T., Lopes, R., Medeiros, P., Silva, J. C., Goncalves, A. N. A., Maracaja-Coutinho, V., & Nakaya, H. I. (2023). Ten quick tips for harnessing the power of ChatGPT in computational biology. *PLoS Computational Biology*, *19*(8), e1011319. https://doi.org/10.1371/journal.pcbi.1011319

91. Lewis, B. A., Walia, R. R., Terribilini, M., Ferguson, J., Zheng, C., Honavar, V., & Dobbs, D. (2011). PRIDB: a Protein-RNA interface database. *Nucleic acids research*, *39*(Database issue), D277–D282. https://doi.org/10.1093/nar/gkq1108
92. Cook, K. B., Kazan, H., Zuberi, K., Morris, Q., & Hughes, T. R. (2011). RBPDB: a database of RNA-binding specificities. Nucleic acids research, 39(Database issue), D301–D308. https://doi.org/10.1093/nar/gkq1069
93. Yi, Y., Zhao, Y., Huang, Y., & Wang, D. (2017). A Brief Review of RNA-Protein Interaction Database Resources. Non-coding RNA, 3(1), 6. https://doi.org/10.3390/ncrna3010006
94. Improved macromolecular interactions data with PISA-lite (Website). https://www.ebi.ac.uk/pdbe/news/improved-macromolecular-interactions-data-pisa-lite
95. PDBePISA (Proteins, Interfaces, Structures and Assemblies) (Website). https://www.ebi.ac.uk/msd-srv/prot_int/cgi-bin/piserver
96. Interactions of RNA and RNA Recognition Motif (InteR3M) (Website). https://inter3mdb.loria.fr/
97. Roca-Martínez, J., Dhondge, H., Sattler, M., & Vranken, W. F. (2023). Deciphering the RRM-RNA recognition code: A computational analysis. PLoS computational biology, 19(1), e1010859. https://doi.org/10.1371/journal.pcbi.1010859
98. Anfinsen C. B. (1973). Principles that govern the folding of protein chains. Science (New York, N.Y.), 181(4096), 223–230. https://doi.org/10.1126/science.181.4096.223
99. Dill, K., Ozkan, S., Shell, M., & Weikl, T. (1993). The protein folding problem. Annual review of biophysics, 37, 289-316 . https://doi.org/10.1146/annurev.biophys.37.092707.153558
100. Rother, K., Rother, M., Boniecki, M., Puton, T., & Bujnicki, J. M. (2011). RNA and protein 3D structure modeling: similarities and differences. *Journal of molecular modeling*, *17*(9), 2325–2336. https://doi.org/10.1007/s00894-010-0951-x
101. Zhang, J., Fei, Y., Sun, L., & Zhang, Q. C. (2022). Advances and opportunities in RNA structure experimental determination and computational modeling. *Nature methods*, *19*(10), 1193–1207. https://doi.org/10.1038/s41592-022-01623-y
102. Jan Gorodkin and Walter L. Ruszzo (eds.) (2014). RNA Sequence, Structure, and Functions: Computational and Bioinformatic Methods, Methods in Molecular Biology, vol. 1097, Chapter 18 "Automated Modelling of RNA 3D structure" DOI https://doi.org/10.1007/978-1-62703-709-9_18
103. Onuchic, J. N., Luthey-Schulten, Z., & Wolynes, P. G. (1997). Theory of protein folding: the energy landscape perspective. *Annual review of physical chemistry*, *48*, 545–600. https://doi.org/10.1146/annurev.physchem.48.1.545
104. Chen, S. J., & Dill, K. A. (2000). RNA folding energy landscapes. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(2), 646–651. https://doi.org/10.1073/pnas.97.2.646
105. Kuhlman, B., & Bradley, P. (2019). Advances in protein structure prediction and design. *Nature reviews. Molecular cell biology*, *20*(11), 681–697. https://doi.org/10.1038/s41580-019-0163-x
106. Croll, T. I., Sammito, M. D., Kryshtafovych, A., & Read, R. J. (2019). Evaluation of template-based modeling in CASP13. *Proteins*, *87*(12), 1113–1127. https://doi.org/10.1002/prot.25800
107. Peng, J., & Xu, J. (2009). Boosting Protein Threading Accuracy. *Research in computational molecular biology : Annual International Conference, RECOMB: proceedings. RECOMB (Conference : 2005-)*, *5541*, 31–45. https://doi.org/10.1007/978-3-642-02008-7_3

108. Watkins, A. M., Rangan, R., & Das, R. (2020). FARFAR2: Improved De Novo Rosetta Prediction of Complex Global RNA Folds. *Structure (London, England : 1993)*, *28*(8), 963–976.e6. https://doi.org/10.1016/j.str.2020.05.011
109. Jumper, J., Evans, R., Pritzel, A., *et al.* (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2
110. Bertoline, L. M. F., Lima, A. N., Krieger, J. E., & Teixeira, S. K. (2023). Before and after AlphaFold2: An overview of protein structure prediction. *Frontiers in bioinformatics*, *3*, 1120370. https://doi.org/10.3389/fbinf.2023.1120370
111. Varadi, M., Anyango, S., Deshpande, *et al.* (2022). AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic acids research*, *50*(D1), D439–D444. https://doi.org/10.1093/nar/gkab1061
112. AlphaFold Protein Structure Database (Website). Available at https://alphafold.ebi.ac.uk/
113. Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., *et al.* (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science (New York, N.Y.)*, *379*(6637), 1123–1130. https://doi.org/10.1126/science.ade2574
114. ESM Metagenomic Atlas (Website). Avaliable at https://esmatlas.com/
115. Vakser I. A. (2020). Challenges in protein docking. *Current opinion in structural biology*, *64*, 160–165. https://doi.org/10.1016/j.sbi.2020.07.001
116. Christoffer, C., & Kihara, D. (2020). IDP-LZerD: Software for Modeling Disordered Protein Interactions. *Methods in molecular biology (Clifton, N.J.)*, *2165*, 231–244. https://doi.org/10.1007/978-1-0716-0708-4_13
117. Huang, Y., Liu, S., Guo, D., Li, L., & Xiao, Y. (2013). A novel protocol for three-dimensional structure prediction of RNA-protein complexes. *Scientific reports*, *3*, 1887. https://doi.org/10.1038/srep01887
118. Xu, X., Duan, R., & Zou, X. (2023). Template-guided method for protein-ligand complex structure prediction: Application to CASP15 protein-ligand studies. *Proteins*, 10.1002/prot.26535. Advance online publication. https://doi.org/10.1002/prot.26535
119. Chakravarty, D., McElfresh, G. W., Kundrotas, P. J., & Vakser, I. A. (2020). How to choose templates for modeling of protein complexes: Insights from benchmarking template-based docking. *Proteins*, *88*(8), 1070–1081. https://doi.org/10.1002/prot.25875
120. Wodak, S. J., Vajda, S., Lensink, M. F., Kozakov, D., & Bates, P. A. (2023). Critical Assessment of Methods for Predicting the 3D Structure of Proteins and Protein Complexes. *Annual review of biophysics*, *52*, 183–206. https://doi.org/10.1146/annurev-biophys-102622-084607
121. Sunny, S., & Jayaraj, P. B. (2022). Protein-Protein Docking: Past, Present, and Future. *The protein journal*, *41*(1), 1–26. https://doi.org/10.1007/s10930-021-10031-8
122. Padhorny, D., Porter, K. A., Ignatov, M., Alekseenko, A., Beglov, D., Kotelnikov, S., Ashizawa, R., Desta, I., Alam, N., Sun, Z., Brini, E., Dill, K., Schueler-Furman, O., Vajda, S., & Kozakov, D. (2020). ClusPro in rounds 38 to 45 of CAPRI: Toward combining template-based methods with free docking. *Proteins*, *88*(8), 1082–1090. https://doi.org/10.1002/prot.25887
123. Yan, Y., Wen, Z., Wang, X., & Huang, S. Y. (2017). Addressing recent docking challenges: A hybrid strategy to integrate template-based and free protein-protein docking. *Proteins*, *85*(3), 497–512. https://doi.org/10.1002/prot.25234
124. Liu, J., Guo, Z., Wu, T., Roy, R. S., Quadir, F., Chen, C., & Cheng, J. (2023). Enhancing AlphaFold-Multimer-based Protein Complex Structure Prediction with MULTICOM in CASP15. bioRxiv, 2023-05-16.

125. MULTICOM3, github repository. Available at https://github.com/BioinfoMachineLearning/MULTICOM3
126. Matsuzaki, Y., Uchikoga, N., Ohue, M., Akiyama, Y. (2016). Rigid-Docking Approaches to Explore Protein–Protein Interaction Space. In: Nookaew, I. (eds) Network Biology. Advances in Biochemical Engineering/Biotechnology, vol 160. Springer, Cham. https://doi.org/10.1007/10_2016_41
127. Nithin, C., Ghosh, P., & Bujnicki, J. M. (2018). Bioinformatics Tools and Benchmarks for Computational Docking and 3D Structure Prediction of RNA-Protein Complexes. *Genes*, *9*(9), 432. https://doi.org/10.3390/genes9090432
128. Bonvin A. M. (2006). Flexible protein-protein docking. *Current opinion in structural biology*, *16*(2), 194–200. https://doi.org/10.1016/j.sbi.2006.02.002
129. Jiang, F., & Kim, S. H. (1991). "Soft docking": matching of molecular surface cubes. *Journal of molecular biology*, *219*(1), 79–102. https://doi.org/10.1016/0022-2836(91)90859-5
130. Zacharias M. (2010). Accounting for conformational changes during protein-protein docking. *Current opinion in structural biology*, *20*(2), 180–186. https://doi.org/10.1016/j.sbi.2010.02.001
131. Kurcinski, M., Kmiecik, S., Zalewski, M., & Kolinski, A. (2021). Protein-Protein Docking with Large-Scale Backbone Flexibility Using Coarse-Grained Monte-Carlo Simulations. *International journal of molecular sciences*, *22*(14), 7341. https://doi.org/10.3390/ijms22147341
132. Marze, N. A., Roy Burman, S. S., Sheffler, W., & Gray, J. J. (2018). Efficient flexible backbone protein-protein docking for challenging targets. *Bioinformatics (Oxford, England)*, *34*(20), 3461–3469. https://doi.org/10.1093/bioinformatics/bty355
133. Huang, S. Y., & Zou, X. (2010). Advances and challenges in protein-ligand docking. *International journal of molecular sciences*, *11*(8), 3016–3034. https://doi.org/10.3390/ijms11083016
134. Kurkcuoglu, Z., & Bonvin, A. M. J. J. (2020). Pre- and post-docking sampling of conformational changes using ClustENM and HADDOCK for protein-protein and protein-DNA systems. *Proteins*, *88*(2), 292–306. https://doi.org/10.1002/prot.25802
135. Andrusier, N., Mashiach, E., Nussinov, R., & Wolfson, H. J. (2008). Principles of flexible protein-protein docking. *Proteins*, *73*(2), 271–289. https://doi.org/10.1002/prot.22170
136. Antunes, D. A., Devaurs, D., & Kavraki, L. E. (2015). Understanding the challenges of protein flexibility in drug design. *Expert opinion on drug discovery*, *10*(12), 1301–1313. https://doi.org/10.1517/17460441.2015.1094458
137. Mohammadi, S., Narimani, Z., Ashouri, M., Firouzi, R., & Karimi-Jafari, M. H. (2022). Ensemble learning from ensemble docking: revisiting the optimum ensemble size problem. *Scientific reports*, *12*(1), 410. https://doi.org/10.1038/s41598-021-04448-5
138. Halperin, I., Ma, B., Wolfson, H., & Nussinov, R. (2002). Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins*, *47*(4), 409–443. https://doi.org/10.1002/prot.10115
139. Dauzhenka, T., Kundrotas, P. J., & Vakser, I. A. (2018). Computational Feasibility of an Exhaustive Search of Side-Chain Conformations in Protein-Protein Docking. *Journal of computational chemistry*, *39*(24), 2012–2021. https://doi.org/10.1002/jcc.25381
140. Trott, O., & Olson, A. J. (2010). AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry*, *31*(2), 455–461. https://doi.org/10.1002/jcc.21334

141. Jandova, Z., Vargiu, A. V., & Bonvin, A. M. J. J. (2021). Native or Non-Native Protein-Protein Docking Models? Molecular Dynamics to the Rescue. *Journal of chemical theory and computation*, *17*(9), 5944–5954. https://doi.org/10.1021/acs.jctc.1c00336
142. van Dijk, A. D., de Vries, S. J., Dominguez, C., Chen, H., Zhou, H. X., & Bonvin, A. M. (2005). Data-driven docking: HADDOCK's adventures in CAPRI. *Proteins*, *60*(2), 232–238. https://doi.org/10.1002/prot.20563
143. Dominguez, C., Boelens, R., & Bonvin, A. M. (2003). HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. *Journal of the American Chemical Society*, *125*(7), 1731–1737. https://doi.org/10.1021/ja026939x
144. Fernández-Recio, J., Totrov, M., & Abagyan, R. (2002). Soft protein-protein docking in internal coordinates. *Protein science : a publication of the Protein Society*, *11*(2), 280–291. https://doi.org/10.1110/ps.19202
145. Li, L., Chen, R., & Weng, Z. (2003). RDOCK: refinement of rigid-body protein docking predictions. *Proteins*, *53*(3), 693–707. https://doi.org/10.1002/prot.10460
146. Zhao, Y., & Sanner, M. F. (2008). Protein-ligand docking with multiple flexible side chains. *Journal of computer-aided molecular design*, *22*(9), 673–679. https://doi.org/10.1007/s10822-007-9148-5
147. Meiler, J., & Baker, D. (2006). ROSETTALIGAND: protein-small molecule docking with full side-chain flexibility. *Proteins*, *65*(3), 538–548. https://doi.org/10.1002/prot.21086
148. Harmalkar, A., & Gray, J. J. (2021). Advances to tackle backbone flexibility in protein docking. *Current opinion in structural biology*, *67*, 178–186. https://doi.org/10.1016/j.sbi.2020.11.011
149. Wang, C., Bradley, P., & Baker, D. (2007). Protein-protein docking with backbone flexibility. *Journal of molecular biology*, *373*(2), 503–519. https://doi.org/10.1016/j.jmb.2007.07.050
150. Harmalkar, A., Mahajan, S. P., & Gray, J. J. (2022). Induced fit with replica exchange improves protein complex structure prediction. *PLoS computational biology*, *18*(6), e1010124. https://doi.org/10.1371/journal.pcbi.1010124
151. Schindler, C. E., de Vries, S. J., & Zacharias, M. (2015). iATTRACT: simultaneous global and local interface optimization for protein-protein docking refinement. *Proteins*, *83*(2), 248–258. https://doi.org/10.1002/prot.24728
152. May, A., & Zacharias, M. (2008). Energy minimization in low-frequency normal modes to efficiently allow for global flexibility during systematic protein-protein docking. *Proteins*, *70*(3), 794–809. https://doi.org/10.1002/prot.21579
153. May, A., & Zacharias, M. (2005). Accounting for global protein deformability during protein-protein and protein-ligand docking. *Biochimica et biophysica acta*, *1754*(1-2), 225–231. https://doi.org/10.1016/j.bbapap.2005.07.045
154. Kuroda, D., & Gray, J. J. (2016). Pushing the Backbone in Protein-Protein Docking. *Structure (London, England: 1993)*, *24*(10), 1821–1829. https://doi.org/10.1016/j.str.2016.06.025
155. Alam, N., Goldstein, O., Xia, B., Porter, K. A., Kozakov, D., & Schueler-Furman, O. (2017). High-resolution global peptide-protein docking using fragments-based PIPER-FlexPepDock. *PLoS computational biology*, *13*(12), e1005905. https://doi.org/10.1371/journal.pcbi.1005905
156. Verdonk, M. L., Giangreco, I., Hall, R. J., Korb, O., Mortenson, P. N., & Murray, C. W. (2011). Docking performance of fragments and druglike compounds. *Journal of medicinal chemistry*, *54*(15), 5422–5431. https://doi.org/10.1021/jm200558u

157. Chauvot de Beauchene, I., de Vries, S. J., & Zacharias, M. (2016). Binding Site Identification and Flexible Docking of Single-Stranded RNA to Proteins Using a Fragment-Based Approach. *PLoS computational biology*, *12*(1), e1004697. https://doi.org/10.1371/journal.pcbi.1004697
158. Chen, Y., Kortemme, T., Robertson, T., Baker, D., & Varani, G. (2004). A new hydrogen-bonding potential for the design of protein-RNA interactions predicts specific contacts and discriminates decoys. *Nucleic acids research*, *32*(17), 5147–5162. https://doi.org/10.1093/nar/gkh785
159. Chauvot de Beauchene, I., de Vries, S. J., & Zacharias, M. (2016). Binding Site Identification and Flexible Docking of Single-Stranded RNA to Proteins Using a Fragment-Based Approach. *PLoS computational biology*, *12*(1), e1004697. https://doi.org/10.1371/journal.pcbi.1004697
160. Chen, Y., Kortemme, T., Robertson, T., Baker, D., & Varani, G. (2004). A new hydrogen-bonding potential for the design of protein-RNA interactions predicts specific contacts and discriminates decoys. *Nucleic acids research*, *32*(17), 5147–5162. https://doi.org/10.1093/nar/gkh785
161. Zheng, S., Robertson, T. A., & Varani, G. (2007). A knowledge-based potential function predicts the specificity and relative binding energy of RNA-binding proteins. *The FEBS journal*, *274*(24), 6378–6391. https://doi.org/10.1111/j.1742-4658.2007.06155.
162. Huang, Sheng-You & Zou, Xiaoqin. (2014). A knowledge-based scoring function for protein-RNA interactions derived from a statistical mechanics-based iterative method. Nucleic acids research. 42. DOI: 10.1093/nar/gku077
163. Jorgensen, William & Maxwell, David & Tirado-Rives, Julian. (1996). Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *Journal of the American Chemical Society*. 118. 11225-11236. 10.1021/ja9621760.
164. Huang, J., & MacKerell, A. D., Jr (2013). CHARMM36 all-atom additive protein force field: validation based on comparison to NMR data. *Journal of computational chemistry*, *34*(25), 2135–2145. https://doi.org/10.1002/jcc.23354
165. López-Blanco, J. R., & Chacón, P. (2019). KORP: knowledge-based 6D potential for fast protein and loop modeling. *Bioinformatics (Oxford, England)*, *35*(17), 3013–3019. https://doi.org/10.1093/bioinformatics/btz026
166. Kadukova, M., Machado, K. D. S., Chacón, P., & Grudinin, S. (2021). KORP-PL: a coarse-grained knowledge-based scoring function for protein-ligand interactions. *Bioinformatics (Oxford, England)*, *37*(7), 943–950. https://doi.org/10.1093/bioinformatics/btaa748
167. Glashagen, G., de Vries, S., Uciechowska-Kaczmarzyk, U., Samsonov, S. A., Murail, S., Tuffery, P., & Zacharias, M. (2020). Coarse-grained and atomic resolution biomolecular docking with the ATTRACT approach. *Proteins*, *88*(8), 1018–1028. https://doi.org/10.1002/prot.25860
168. Zacharias M. (2003). Protein-protein docking with a reduced protein model accounting for side-chain flexibility. *Protein science : a publication of the Protein Society*, *12*(6), 1271–1282. https://doi.org/10.1110/ps.0239303
169. Kolinski A. (2004). Protein modeling and structure prediction with a reduced representation. Acta biochimica Polonica, 51(2), 349–371.
170. Monticelli, L., Kandasamy, S. K., Periole, X., Larson, R. G., Tieleman, D. P., & Marrink, S. J. (2008). The MARTINI Coarse-Grained Force Field: Extension to Proteins. *Journal of chemical theory and computation*, *4*(5), 819–834. https://doi.org/10.1021/ct700324x

171. Pasquali, S., & Derreumaux, P. (2010). HiRE-RNA: a high resolution coarse-grained energy model for RNA. *The journal of physical chemistry. B*, *114*(37), 11957–11966. https://doi.org/10.1021/jp102497y
172. Jin, J., Pak, A. J., Durumeric, A. E. P., Loose, T. D., & Voth, G. A. (2022). Bottom-up Coarse-Graining: Principles and Perspectives. *Journal of chemical theory and computation*, *18*(10), 5759–5791. https://doi.org/10.1021/acs.jctc.2c00643
173. Noid W. G. (2023). Perspective: Advances, Challenges, and Insight for Predictive Coarse-Grained Models. *The journal of physical chemistry. B*, *127*(19), 4174–4207. https://doi.org/10.1021/acs.jpcb.2c08731
174. Bordner, A. J., & Gorin, A. A. (2007). Protein docking using surface matching and supervised machine learning. *Proteins*, *68*(2), 488–502. https://doi.org/10.1002/prot.21406
175. Venkatraman, V., Yang, Y. D., Sael, L., & Kihara, D. (2009). Protein-protein docking using region-based 3D Zernike descriptors. *BMC bioinformatics*, *10*, 407. https://doi.org/10.1186/1471-2105-10-407
176. Jafari, R., Sadeghi, M., & Mirzaie, M. (2016). Investigating the importance of Delaunay-based definition of atomic interactions in scoring of protein-protein docking results. Journal of molecular graphics & modelling, 66, 108–114. https://doi.org/10.1016/j.jmgm.2016.04.001
177. McNutt, A. T., Francoeur, P., Aggarwal, R., Masuda, T., Meli, R., Ragoza, M., Sunseri, J., & Koes, D. R. (2021). GNINA 1.0: molecular docking with deep learning. *Journal of cheminformatics*, *13*(1), 43. https://doi.org/10.1186/s13321-021-00522-2
178. Vajda, S., Hall, D. R., & Kozakov, D. (2013). Sampling and scoring: a marriage made in heaven. *Proteins*, *81*(11), 1874–1884. https://doi.org/10.1002/prot.24343
179. Eisenstein, M., & Katchalski-Katzir, E. (2004). On proteins, grids, correlations, and docking. *Comptes rendus biologies*, *327*(5), 409–420. https://doi.org/10.1016/j.crvi.2004.03.006
180. Trosset, J. Y., & Scheraga, H. A. (1998). Reaching the global minimum in docking simulations: a Monte Carlo energy minimization approach using Bezier splines. *Proceedings of the National Academy of Sciences of the United States of America*, *95*(14), 8011–8015. https://doi.org/10.1073/pnas.95.14.8011
181. Li, Y., Zhang, X., & Cao, D. (2013). The role of shape complementarity in the protein-protein interactions. *Scientific reports*, *3*, 3271. https://doi.org/10.1038/srep03271
182. Axenopoulos, A., Daras P., Papadopoulos G., & Houstis E. (2011). 3D protein-protein docking using shape complementarity and fast alignment. Proceedings - International Conference on Image Processing, ICIP. 1569-1572. 10.1109/ICIP.2011.6115747
183. Jiang, S., Tovchigrechko, A., & Vakser, I. A. (2003). The role of geometric complementarity in secondary structure packing: a systematic docking study. *Protein science : a publication of the Protein Society*, *12*(8), 1646–1651. https://doi.org/10.1110/ps.0304503
184. Norel, R., Petrey, D., Wolfson, H. J., & Nussinov, R. (1999). Examination of shape complementarity in docking of unbound proteins. *Proteins*, *36*(3), 307–317. PMID: 10409824
185. Kuntz, I. D., Blaney, J. M., Oatley, S. J., Langridge, R., & Ferrin, T. E. (1982). A geometric approach to macromolecule-ligand interactions. *Journal of molecular biology*, *161*(2), 269–288. https://doi.org/10.1016/0022-2836(82)90153-x
186. Zhang, Q., Sanner, M., & Olson, A. J. (2009). Shape complementarity of protein-protein complexes at multiple resolutions. *Proteins*, *75*(2), 453–467. https://doi.org/10.1002/prot.22256

187. Huang S. Y. (2014). Search strategies and evaluation in protein-protein docking: principles, advances and challenges. *Drug discovery today*, *19*(8), 1081–1096. https://doi.org/10.1016/j.drudis.2014.02.005
188. Axenopoulos, A., Daras, P., Papadopoulos, G. E., & Houstis, E. N. (2013). SP-dock: protein-protein docking using shape and physicochemical complementarity. *IEEE/ACM transactions on computational biology and bioinformatics*, *10*(1), 135–150. https://doi.org/10.1109/TCBB.2012.149
189. Yan, Y., & Huang, S. Y. (2019). Pushing the accuracy limit of shape complementarity for protein-protein docking. *BMC bioinformatics*, *20*(Suppl 25), 696. https://doi.org/10.1186/s12859-019-3270-y
190. Sunny S., & Jayaraj B. (2021). A Geometric Complementarity-Based Tool for Protein–Protein Docking. Journal of Computational Biophysics and Chemistry. 21. https://doi.org/10.1142/S273741652250003X
191. Hwang, S. B., Lee, C. J., Lee, S., … No, K. T. (2020). PMFF: Development of a Physics-Based Molecular Force Field for Protein Simulation and Ligand Docking. *The journal of physical chemistry. B*, *124*(6), 974–989. https://doi.org/10.1021/acs.jpcb.9b10339
192. Yadava, Umesh. (2018). Search algorithms and scoring methods in protein-ligand docking. Endocrinology&Metabolism International Journal. 6. 10.15406/emij.2018.06.00212.
193. Ban, T., Ohue, M., & Akiyama, Y. (2018). Multiple grid arrangement improves ligand docking with unknown binding sites: Application to the inverse docking problem. *Computational biology and chemistry*, *73*, 139–146. https://doi.org/10.1016/j.compbiolchem.2018.02.008
194. Wu, G., Robertson, D. H., Brooks, C. L., 3rd, & Vieth, M. (2003). Detailed analysis of grid-based molecular docking: A case study of CDOCKER-A CHARMm-based MD docking algorithm. *Journal of computational chemistry*, *24*(13), 1549–1562. https://doi.org/10.1002/jcc.10306
195. Katchalski-Katzir, E., Shariv, I., Eisenstein, M., Friesem, A. A., Aflalo, C., & Vakser, I. A. (1992). Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. *Proceedings of the National Academy of Sciences of the United States of America*, *89*(6), 2195–2199. https://doi.org/10.1073/pnas.89.6.2195
196. Desta, I. T., Porter, K. A., Xia, B., Kozakov, D., & Vajda, S. (2020). Performance and Its Limits in Rigid Body Protein-Protein Docking. *Structure (London, England : 1993)*, *28*(9), 1071–1081.e3. https://doi.org/10.1016/j.str.2020.06.006
197. Kozakov, D., Brenke, R., Comeau, S. R., & Vajda, S. (2006). PIPER: an FFT-based protein docking program with pairwise potentials. *Proteins*, *65*(2), 392–406. https://doi.org/10.1002/prot.21117
198. Vakser I. A. (1997). Evaluation of GRAMM low-resolution docking methodology on the hemagglutinin-antibody complex. *Proteins*, *Suppl 1*, 226–230.
199. Pierce, B. G., Hourai, Y., & Weng, Z. (2011). Accelerating protein docking in ZDOCK using an advanced 3D convolution library. *PloS one*, *6*(9), e24657. https://doi.org/10.1371/journal.pone.0024657
200. Ghoorah, A. W., Devignes, M. D., Smaïl-Tabbone, M., & Ritchie, D. W. (2013). Protein docking using case-based reasoning. *Proteins*, *81*(12), 2150–2158. https://doi.org/10.1002/prot.24433
201. Yan, Y., Zhang, D., Zhou, P., Li, B., & Huang, S. Y. (2017). HDOCK: a web server for protein-protein and protein-DNA/RNA docking based on a hybrid strategy. *Nucleic acids research*, *45*(W1), W365–W373. https://doi.org/10.1093/nar/gkx407

202. Padhorny, D., Kazennov, A., Zerbe, B. S., Porter, K. A., Xia, B., Mottarella, S. E., Kholodov, Y., Ritchie, D. W., Vajda, S., & Kozakov, D. (2016). Protein-protein docking by fast generalized Fourier transforms on 5D rotational manifolds. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(30), E4286–E4293. https://doi.org/10.1073/pnas.1603929113
203. Schneidman, Dina & Nussinov, Ruth & Wolfson, Haim. (2002). Efficient Unbound Docking of Rigid Molecules. Lecture Notes in Computer Science. 2452. 185-200. 10.1007/3-540-45784-4_14.
204. Goldman, B. B., & Wipke, W. T. (2000). QSD quadratic shape descriptors. 2. Molecular docking using quadratic shape descriptors (QSDock). Proteins, 38(1), 79–94.
205. Li, Y., Cortés, J., & Siméon, T. (2011). Enhancing systematic protein-protein docking methods using ray casting: application to ATTRACT. *Proteins*, *79*(11), 3037–3049. https://doi.org/10.1002/prot.23127
206. Zhang, Z., Schindler, C. E., Lange, O. F., & Zacharias, M. (2015). Application of Enhanced Sampling Monte Carlo Methods for High-Resolution Protein-Protein Docking in Rosetta. *PloS one*, *10*(6), e0125941. https://doi.org/10.1371/journal.pone.0125941
207. Gray, J. J., Moughon, S., Wang, C., Schueler-Furman, O., Kuhlman, B., Rohl, C. A., & Baker, D. (2003). Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *Journal of molecular biology*, *331*(1), 281–299. https://doi.org/10.1016/s0022-2836(03)00670-3
208. Torchala, M., Moal, I. H., Chaleil, R. A., Fernandez-Recio, J., & Bates, P. A. (2013). SwarmDock: a server for flexible protein-protein docking. *Bioinformatics (Oxford, England)*, *29*(6), 807–809. https://doi.org/10.1093/bioinformatics/btt038
209. Tai, H. K., Jusoh, S. A., & Siu, S. W. I. (2018). Chaos-embedded particle swarm optimization approach for protein-ligand docking and virtual screening. *Journal of cheminformatics*, *10*(1), 62. https://doi.org/10.1186/s13321-018-0320-9
210. Corbeil, C. R., Englebienne, P., & Moitessier, N. (2007). Docking ligands into flexible and solvated macromolecules. 1. Development and validation of FITTED 1.0. *Journal of chemical information and modeling*, *47*(2), 435–449. https://doi.org/10.1021/ci6002637
211. Li, C., Sun, J., & Palade, V. (2020). Diversity-guided Lamarckian random drift particle swarm optimization for flexible ligand docking. *BMC bioinformatics*, *21*(1), 286. https://doi.org/10.1186/s12859-020-03630-2
212. Ding, X., Wu, Y., Wang, Y., Vilseck, J. Z., & Brooks, C. L., 3rd (2020). Accelerated CDOCKER with GPUs, Parallel Simulated Annealing, and Fast Fourier Transforms. *Journal of chemical theory and computation*, *16*(6), 3910–3919. https://doi.org/10.1021/acs.jctc.0c00145
213. Liu, J., & Wang, R. (2015). Classification of current scoring functions. *Journal of chemical information and modeling*, *55*(3), 475–482. https://doi.org/10.1021/ci500731a
214. Ewing TJ, Makino, S., Skillman, A. G., & Kuntz, I. D. (2001). DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. Journal of computer-aided molecular design, 15(5), 411–428. https://doi.org/10.1023/a:1011115820450
215. Hess B, Kutzner, C., van der Spoel, D., & Lindahl, E. (2008). GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. Journal of chemical theory and computation, 4(3), 435–447. https://doi.org/10.1021/ct700301q
216. Huang, S. Y., & Zou, X. (2014). A knowledge-based scoring function for protein-RNA interactions derived from a statistical mechanics-based iterative method. *Nucleic acids research*, *42*(7), e55. https://doi.org/10.1093/nar/gku077

217. Huang, S. Y., & Zou, X. (2010) Mean-force scoring functions for protein-ligand binding. *Annu Rep Comput Chem* 6:280–296
218. Mooij W, & Verdonk, M. L. (2005). General and targeted statistical potentials for protein-ligand interactions. Proteins, 61(2), 272–287. https://doi.org/10.1002/prot.20588
219. Popov P, & Grudinin, S. (2015). Knowledge of Native Protein-Protein Interfaces Is Sufficient To Construct Predictive Models for the Selection of Binding Candidates. Journal of chemical information and modeling, 55(10), 2242–2255. https://doi.org/10.1021/acs.jcim.5b00372
220. Zhou H, & Skolnick, J. (2011). GOAP: a generalized orientation-dependent, all-atom statistical potential for protein structure prediction. Biophysical journal, 101(8), 2043–2052. https://doi.org/10.1016/j.bpj.2011.09.012
221. Chen, P., Ke, Y., Lu, Y., Du, Y., Li, J., Yan, H., Zhao, H., Zhou, Y., & Yang, Y. (2019). DLIGAND2: an improved knowledge-based energy function for protein-ligand interactions using the distance-scaled, finite, ideal-gas reference state. *Journal of cheminformatics*, *11*(1), 52. https://doi.org/10.1186/s13321-019-0373-4
222. Fujimoto, K. J., Minami, S., & Yanai, T. (2022). Machine-Learning- and Knowledge-Based Scoring Functions Incorporating Ligand and Protein Fingerprints. *ACS omega*, *7*(22), 19030–19039. https://doi.org/10.1021/acsomega.2c02822
223. Meli, R., Morris, G. M., & Biggin, P. C. (2022). Scoring Functions for Protein-Ligand Binding Affinity Prediction using Structure-Based Deep Learning: A Review. *Frontiers in bioinformatics*, *2*, 885983. https://doi.org/10.3389/fbinf.2022.885983
224. Grinter, S. Z., & Zou, X. (2014). Challenges, applications, and recent advances of protein-ligand docking in structure-based drug design. *Molecules (Basel, Switzerland)*, *19*(7), 10150–10176. https://doi.org/10.3390/molecules190710150
225. Huang, S. Y., & Zou, X. (2011). Statistical mechanics-based method to extract atomic distance-dependent potentials from protein structures. *Proteins*, *79*(9), 2648–2661. https://doi.org/10.1002/prot.23086
226. Thomas, P. D., & Dill, K. A. (1996). Statistical potentials extracted from protein structures: how accurate are they?. *Journal of molecular biology*, *257*(2), 457–469. https://doi.org/10.1006/jmbi.1996.0175
227. Friesner, R. A., Banks, J. L., Murphy, R. B., Halgren, T. A., Klicic, J. J., Mainz, D. T., Repasky, M. P., Knoll, E. H., Shelley, M., Perry, J. K., Shaw, D. E., Francis, P., & Shenkin, P. S. (2004). Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *Journal of medicinal chemistry*, *47*(7), 1739–1749. https://doi.org/10.1021/jm0306430
228. Wang, R., Lai, L., & Wang, S. (2002). Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *Journal of computer-aided molecular design*, *16*(1), 11–26. https://doi.org/10.1023/a:1016357811882
229. Korb, O., Stützle, T., & Exner, T. E. (2009). Empirical scoring functions for advanced protein-ligand docking with PLANTS. *Journal of chemical information and modeling*, *49*(1), 84–96. https://doi.org/10.1021/ci800298z
230. Verdonk, M. L., Cole, J. C., Hartshorn, M. J., Murray, C. W., & Taylor, R. D. (2003). Improved protein-ligand docking using GOLD. *Proteins*, *52*(4), 609–623. https://doi.org/10.1002/prot.10465
231. Pham, T. A., & Jain, A. N. (2008). Customizing scoring functions for docking. *Journal of computer-aided molecular design*, *22*(5), 269–286. https://doi.org/10.1007/s10822-008-9174-y

232. Kwon, Y., Shin, W. H., Ko, J., & Lee, J. (2020). AK-Score: Accurate Protein-Ligand Binding Affinity Prediction Using an Ensemble of 3D-Convolutional Neural Networks. *International journal of molecular sciences*, *21*(22), 8424. https://doi.org/10.3390/ijms21228424
233. Norberto, Sánchez-Cruz. (2023). Deep Graph Learning in Molecular Docking: Advances and Opportunities. Artificial Intelligence in the Life Sciences. 3. 100062. 10.1016/j.ailsci.2023.100062.
234. Cao, Y., & Shen, Y. (2020). Energy-based graph convolutional networks for scoring protein docking models. *Proteins*, *88*(8), 1091–1099. https://doi.org/10.1002/prot.25888
235. Li, H., Peng, J., Sidorov, P., Leung, Y., Leung, K. S., Wong, M. H., Lu, G., & Ballester, P. J. (2019). Classical scoring functions for docking are unable to exploit large volumes of structural and interaction data. *Bioinformatics (Oxford, England)*, *35*(20), 3989–3995. https://doi.org/10.1093/bioinformatics/btz183
236. Yang, L., Yang, G., Chen, X., Yang, Q., Yao, X., Bing, Z., Niu, Y., Huang, L., & Yang, L. (2021). Deep Scoring Neural Network Replacing the Scoring Function Components to Improve the Performance of Structure-Based Molecular Docking. *ACS chemical neuroscience*, *12*(12), 2133–2142. https://doi.org/10.1021/acschemneuro.1c00110
237. Gabel, J., Desaphy, J., & Rognan, D. (2014). Beware of machine learning-based scoring functions-on the danger of developing black boxes. *Journal of chemical information and modeling*, *54*(10), 2807–2815. https://doi.org/10.1021/ci500406k
238. Guedes, I. A., Barreto, A. M. S., Marinho, D., Krempser, E., Kuenemann, M. A., Sperandio, O., Dardenne, L. E., & Miteva, M. A. (2021). New machine learning and physics-based scoring functions for drug discovery. *Scientific reports*, *11*(1), 3198. https://doi.org/10.1038/s41598-021-82410-1
239. Kamal, I. M., & Chakrabarti, S. (2023). MetaDOCK: A Combinatorial Molecular Docking Approach. *ACS omega*, *8*(6), 5850–5860. https://doi.org/10.1021/acsomega.2c07619
240. Li, H., Huang, Y., & Xiao, Y. (2017). A pair-conformation-dependent scoring function for evaluating 3D RNA-protein complex structures. *PloS one*, *12*(3), e0174662. https://doi.org/10.1371/journal.pone.0174662
241. Li, C. H., Cao, L. B., Su, J. G., Yang, Y. X., & Wang, C. X. (2012). A new residue-nucleotide propensity potential with structural information considered for discriminating protein-RNA docking decoys. *Proteins*, *80*(1), 14–24. https://doi.org/10.1002/prot.23117
242. Guilhot-Gaudeffroy, A., Froidevaux, C., Azé, J., & Bernauer, J. (2014). Protein-RNA complexes and efficient automatic docking: expanding RosettaDock possibilities. *PloS one*, *9*(9), e108928. https://doi.org/10.1371/journal.pone.0108928
243. Pérez-Cano, L., Solernou, A., Pons, C., & Fernández-Recio, J. (2010). Structural prediction of protein-RNA interaction by computational docking with propensity-based statistical potentials. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 293–301. https://doi.org/10.1142/9789814295291_0031
244. Liu, S., & Vakser, I. A. (2011). DECK: Distance and environment-dependent, coarse-grained, knowledge-based potentials for protein-protein docking. *BMC bioinformatics*, *12*, 280. https://doi.org/10.1186/1471-2105-12-280
245. Tuszynska, I., & Bujnicki, J. M. (2011). DARS-RNP and QUASI-RNP: new statistical potentials for protein-RNA docking. *BMC bioinformatics*, *12*, 348. https://doi.org/10.1186/1471-2105-12-348

246. Feng, Y., & Huang, S. Y. (2020). ITScore-NL: An Iterative Knowledge-Based Scoring Function for Nucleic Acid-Ligand Interactions. *Journal of chemical information and modeling*, *60*(12), 6698–6708. https://doi.org/10.1021/acs.jcim.0c00974
247. Huang, S. Y., & Zou, X. (2013). A nonredundant structure dataset for benchmarking protein-RNA computational docking. *Journal of computational chemistry*, *34*(4), 311–318. https://doi.org/10.1002/jcc.23149
248. Pérez-Cano, L., Jiménez-García, B., & Fernández-Recio, J. (2012). A protein-RNA docking benchmark (II): extended set from experimental and homology modeling data. *Proteins*, *80*(7), 1872–1882. https://doi.org/10.1002/prot.24075
249. Hall, D., Li, S., Yamashita, K., Azuma, R., Carver, J. A., & Standley, D. M. (2015). RNA-LIM: a novel procedure for analyzing protein/single-stranded RNA propensity data with concomitant estimation of interface structure. Analytical biochemistry, 472, 52–61. https://doi.org/10.1016/j.ab.2014.11.004
250. González-Alemán, R., Chevrollier, N., Simoes, M., Montero-Cabrera, L., & Leclerc, F. (2021). MCSS-Based Predictions of Binding Mode and Selectivity of Nucleotide Ligands. Journal of chemical theory and computation, 17(4), 2599–2618. https://doi.org/10.1021/acs.jctc.0c01339
251. Kappel, K., & Das, R. (2019). Sampling Native-like Structures of RNA-Protein Complexes through Rosetta Folding and Docking. Structure (London, England : 1993), 27(1), 140–151.e5. https://doi.org/10.1016/j.str.2018.10.001
252. de Beauchene, I. C., de Vries, S. J., & Zacharias, M. (2016). Fragment-based modelling of single stranded RNA bound to RNA recognition motif containing proteins. *Nucleic acids research*, *44*(10), 4565–4580. https://doi.org/10.1093/nar/gkw328
253. Moniot, A., Guermeur, Y., de Vries, S. J., & Chauvot de Beauchene, I. (2022). ProtNAff: protein-bound Nucleic Acid filters and fragment libraries. Bioinformatics (Oxford, England), 38(16), 3911–3917. https://doi.org/10.1093/bioinformatics/btac430
254. Moniot, A., Chauvot de Beauchêne, I., Guermeur, Y. (2022). Inferring ε -nets of Finite Sets in a RKHS. In: Faigl, J., Olteanu, M., Drchal, J. (eds) Advances in Self-Organizing Maps, Learning Vector Quantization, Clustering and Data Visualization. WSOM+ 2022. Lecture Notes in Networks and Systems, vol 533. Springer, Cham.
255. Moniot, A., Guermeur, Y., de Vries, S. J., & Chauvot de Beauchene, I. (2022). ProtNAff: protein-bound Nucleic Acid filters and fragment libraries. Bioinformatics (Oxford, England), 38(16), 3911–3917. https://doi.org/10.1093/bioinformatics/btac430
256. Setny, P., & Zacharias, M. (2011). A coarse-grained force field for Protein-RNA docking. Nucleic acids research, 39(21), 9118–9129. https://doi.org/10.1093/nar/gkr636
257. Fiorucci, S., & Zacharias, M. (2010). Binding site prediction and improved scoring during flexible protein-protein docking with ATTRACT. Proteins, 78(15), 3131–3139. https://doi.org/10.1002/prot.22808
258. Vanderbilt, D., & Louie, S. G. (1984). A Monte carlo simulated annealing approach to optimization over continuous variables. Journal of Computational Physics, 56(2), 259–271. doi:10.1016/0021-9991(84)90095-0
259. Delahaye, D., Chaimatanan, S., Mongeau, M. (2019). Simulated Annealing: From Basics to Applications. In: Gendreau, M., Potvin, JY. (eds) Handbook of Metaheuristics. International Series in Operations Research & Management Science, vol 272. Springer, Cham. https://doi.org/10.1007/978-3-319-91086-4_1
260. Sasse, A., de Vries, S. J., Schindler, C. E., de Beauchêne, I. C., & Zacharias, M. (2017). Rapid Design of Knowledge-Based Scoring Potentials for Enrichment of Near-Native

Geometries in Protein-Protein Docking. PloS one, 12(1), e0170625. https://doi.org/10.1371/journal.pone.0170625

261. Cuevas, E., Gálvez, J., Avalos, O. (2020). Fuzzy Logic Based Optimization Algorithm. In: Recent Metaheuristics Algorithms for Parameter Identification. Studies in Computational Intelligence, vol 854. Springer, Cham. https://doi.org/10.1007/978-3-030-28917-1_6
262. Garrido-Merchán E. C., Hernández-Lobato D. (2016) Predictive Entropy Search for Multi-objective Bayesian Optimization with Constraints https://doi.org/10.1016/j.neucom.2019.06.025
263. Kravchenko A., De Vries S. J., Smaïl-Tabbone M. et al. HIPPO: HIstogram-based Pseudo-POtential for scoring protein-ssRNA fragment-based docking poses, 30 May 2023, PREPRINT (Version 1) available at Research Square https://doi.org/10.21203/rs.3.rs-2981840/v1
264. Schneider, T. D., & Stephens, R. M. (1990). Sequence logos: a new way to display consensus sequences. Nucleic acids research, 18(20), 6097–6100. https://doi.org/10.1093/nar/18.20.6097
265. Rodrigues, J. P., & Bonvin, A. M. (2014). Integrative computational modeling of protein interactions. The FEBS journal, 281(8), 1988–2003. https://doi.org/10.1111/febs.12771
266. Experimetal data Colection avaliable at https://docs.google.com/spreadsheets/d/1FL8H8uFEZRoHya5Oi20AMIlanMOAsk97/edit?usp=sharing&ouid=118289948335114655570&rtpof=true&sd=true
267. UniProt (Website) https://www.uniprot.org/
268. Deng, L., Sui, Y., & Zhang, J. (2019). XGBPRH: Prediction of Binding Hot Spots at Protein-RNA Interfaces Utilizing Extreme Gradient Boosting. Genes, 10(3), 242. https://doi.org/10.3390/genes10030242
269. Pan Y., Wang, Z., Zhan, W., & Deng, L. (2018). Computational identification of binding energy hot spots in protein-RNA complexes using an ensemble approach. Bioinformatics (Oxford, England), 34(9), 1473–1480. https://doi.org/10.1093/bioinformatics/btx822 .
270. Barik A., Nithin, C., Karampudi, N. B., Mukherjee, S., & Bahadur, R. P. (2016). Probing binding hot spots at protein-RNA recognition sites. Nucleic acids research, 44(2), e9. https://doi.org/10.1093/nar/gkv876
271. Jackson, R. W., Smathers, C. M., & Robart, A. R. (2023). General Strategies for RNA X-ray Crystallography. Molecules (Basel, Switzerland), 28(5), 2111. https://doi.org/10.3390/molecules28052111
272. Baek, M., McHugh, R., Anishchenko, I. et al. (2023) Accurate prediction of protein–nucleic acid complexes using RoseTTAFoldNA. Nat Methods. https://doi.org/10.1038/s41592-023-02086-5
273. Huanga M. C., Chen, W. H., Huang, C. W., Huang, K. Y., Horng, J. C., Hayashi, M., & Chen, I. C. (2020). Investigation of the cis-trans structures and isomerization of oligoprolines by using Raman spectroscopy and density functional theory calculations: solute-solvent interactions and effects of terminal positively charged amino acid residues. RSC advances, 10(57), 34493–34500. https://doi.org/10.1039/d0ra05746k
274. Ayaz, P., Lyczek, A., Paung, Y., Mingione, V. R., Iacob, R. E., de Waal, P. W., Engen, J. R., Seeliger, M. A., Shan, Y., & Shaw, D. E. (2023). Structural mechanism of a drug-binding process involving a large conformational change of the protein target. Nature communications, 14(1), 1885. https://doi.org/10.1038/s41467-023-36956-5
275. Li, W., Schaeffer, R. D., Otwinowski, Z., & Grishin, N. V. (2016). Estimation of Uncertainties in the Global Distance Test (GDT_TS) for CASP Models. PloS one, 11(5), e0154786. https://doi.org/10.1371/journal.pone.0154786

276. Zemla A. (2003). LGA: A method for finding 3D similarities in protein structures. Nucleic acids research, 31(13), 3370–3374. https://doi.org/10.1093/nar/gkg571

277. Best, R. B., Hummer, G., & Eaton, W. A. (2013). Native contacts determine protein folding mechanisms in atomistic simulations. Proceedings of the National Academy of Sciences of the United States of America, 110(44), 17874–17879. https://doi.org/10.1073/pnas.1311599110

278. Xiong, P., Wang, M., Zhou, X., Zhang, T., Zhang, J., Chen, Q., & Liu, H. (2014). Protein design with a comprehensive statistical energy function and boosted by experimental selection for foldability. Nature communications, 5, 5330. https://doi.org/10.1038/ncomms6330

Anna Kravchenko

Modélisation par fragment des complexes protéine-ARNsb pour la conception de protéines

Résumé étendu de la thèse en Français

Cette thèse a été réalisée dans le cadre d'un projet Européen plus vaste (ITN RNAct) dans lequel des approches informatiques et biologiques étaient combinées pour progresser vers la conception et la synthèse de nouveaux domaines protéiques appelés RNA-recognition Motifs (**RRM**) capables de se fixer sur des séquences spécifiques d'ARN (Acide RiboNucléique). Cet objectif global nécessitait le développement de méthodes capables de modéliser la structure 3D d'un complexe entre un RRM et un ARN donnés. Les RRM se lient principalement aux ARN simple-brin (**ARNsb**). Ces ARN n'ont pas de structuration 3D propre mais une multitude de conformations 3D possibles, et adaptent leur conformation à la protéine à laquelle ils se lient. La prise en compte de cette flexibilité nécessite des approches de modélisation spécifiques des complexes protéine-ARNsb, telles que l'amarrage par fragments. La thèse vise donc à améliorer les outils existants d'amarrage d'ARNsb par fragments et les adapter au problème spécifique des RRM.

La thèse comprend:

- une introduction générale donnant les clefs de lecture du manuscrit
- un chapitre présentant les connaissances en biologie nécessaires pour comprendre les enjeux et évaluer les contributions de la thèse (chapitre 1)
- un chapitre qui présente d'une part les notions de base nécessaires pour comprendre les méthodes et les ressources bioinformatiques employées dans la thèse, et d'autre part l'état de l'art dans lequel se placent les contributions (chapitre 2)
- trois chapitres de résultats originaux (Chapitres 3 à 5)
- une conclusion ouverte sur les perspectives de ce travail (Chapitre 6).

I. Contexte et état de l'art

Présentation des protéines RRM, des ARN et de leurs interactions

Les protéines se composent d'une séquence linéaire d'acides aminés (aa) qui constituent la chaîne protéique, parmi un vocabulaire de 20 aa différent. Ces aa interagissent entre eux pour former des éléments de structure "2D" locale (brins linéaires, hélices) et une structure 3D globale.

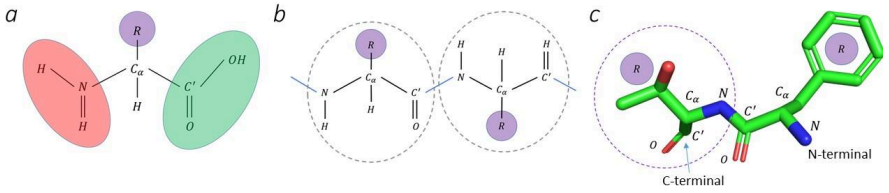


Fig 1: (a) Acide aminé composé d'un groupe amine (rouge), d'un groupe carboxyle (vert) et d'une chaîne latérale ou groupe R (violet) variable d'une type d'aa a l'autre; (b) chaîne polypeptidique composée de deux aa liés par une liaison peptidique (bleu); (c) Structure 3D d'une chaîne polypeptidique composée des acides aminés phénylalanine et thréonine. Ce dernier est mis en évidence par un cercle en pointillé.

L'ARN est constitué d'une séquence linéaire de nucléotides de 4 types différents: l'adénine (A), la cytosine (C), la guanine (G) et l'uracile (U). Chaque nucléotide comporte un groupement phosphate, un ribose, et une base de 1 ou 2 cycles qui varie en fonction du type de base. Les bases peuvent s'apparier deux a deux (paires A-U et C-G), formant la structure 2D de l'ARN. Les paires de bases s'empilent en hélices double-brin jointes par des boucles simple-brin, formant la structure 3D de l'ARN.

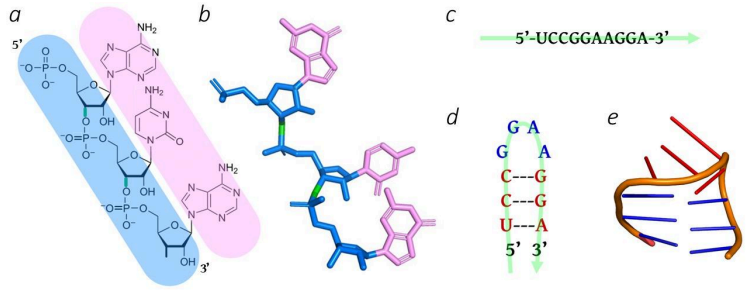


Fig 2: Représentations d'un ARN: (a, b) Formule chimique et représentation 3D d'une chaîne de 3 nucléotides (base en rose, sucre et phosphate en bleu); (c, d, e) séquence d'ARN, structure 2D (appariements) et structure 3D (représentation simplifié), avec parties double-brin en bleu et simple-brin en rouge.

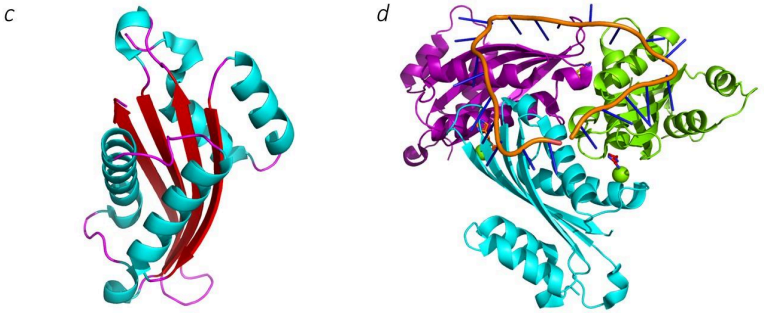


Fig. 3: Représentations simplifiées de structures 3D (c) d'une protéine (brins en rouge, hélices en cyan, boucles en magenta) et (d) d'un complexe entre 3 protéines (cyan, vert et magenta) et un ARNsb (orange et bleu).

Les **structures 3D des protéines, ARN et complexes protéine-ARN** peuvent être résolues grâce à des techniques de biophysique expérimentales. Mais la mise en place de ces techniques est très longue (mois ou années), coûteuse, et au succès aléatoire. Des méthodes computationnelles ont donc été développées pour modéliser ces molécules et complexes moléculaires à partir de leurs séquences, dont l'obtention est beaucoup plus facile. Ces méthodes s'appuient sur l'analyse des structures 3D expérimentalement connues, regroupées dans une base de données publique appelée PDB. Celle-ci contient quelques milliers de structures de protéines, d'ARN, et de complexes protéine-ARN. Le problème de la modélisation de la structure 3D des protéines - et des complexes de plusieurs protéines - a été en grande partie résolu ces quelques dernières années par AlphaFold, la méthode par apprentissage profond créée par DeepMind. Si des recherches sont en cours pour transférer ces résultats aux ARN (pour lesquels les méthodes classiques actuelles fonctionnent déjà dans la plupart des cas), leur transfert aux complexes protéine-ARN n'est à ce jour pas possible: les données d'apprentissage sont moins abondantes et plus hétérogènes. Cela est plus particulièrement vrai pour les ARNsb, qui nous intéressent dans cette thèse.

La plupart des protéines comportent des domaines, qui sont des régions structurées et indépendamment repliées. Chaque famille de domaine présente une structure tridimensionnelle majoritairement commune, avec quelques variations possibles d'une protéine à une autre. Les domaines sont des unités fonctionnelles des protéines, ils peuvent effectuer des activités enzymatiques spécifiques, interagir avec d'autres molécules telles que l'ADN ou l'ARN, etc. Parmi les familles de domaines protéiques, cette thèse se concentre sur les **RRM**. Ce sont les domaines les plus abondants, présents dans 1 à 2 % des protéines humaines, et connus

pour lier des ARNsb de 2 à 8 nucléotides de long. Un RRM se compose de 90 à 100 aa, qui forment 4 brins antiparallèles et 2 hélices. On peut trouver jusqu'à 6 RRM dans une protéine. L'interface de liaison principal est située sur les brins 1 et 3, sur lesquels 2 ou 3 aa aromatiques (ayant un cycle à doubles liaison) - situés à des positions conservées par l'évolution parmi les RRMs - peuvent lier chacun un nucléotide de l'ARN ; cependant, certains RRMs ne possèdent pas ces aromatiques conservés (ils sont parfois appelés quasi-RRMs). De plus, les boucles d'aa sans structure 2D propre - et par conséquent plus flexibles et de conformation 3D moins conservée parmi les RRMs - interviennent souvent également dans la liaison. Les liens entre RRMs d'une même protéine, s'ils sont présents, contribuent également à la liaison de l'ARN et à la stabilité du complexe ARN-RRM. Les domaines RRM partagent donc un mode de liaison (i) suffisamment commun pour que l'analyse des structures RRM-ARNsb puisse guider la modélisation d'autres de ces complexes, et (ii) suffisamment variable pour que leur modélisation nécessite le développement d'approches dédiées.

Modélisation des complexes protéine-ARNsb

La modélisation des complexes protéine-ARN repose sur l'hypothèse communément admise que la structure 3D portant la fonction biologique d'intérêt correspond à la structure 3D de plus basse énergie. L'approche classique consiste à utiliser une structure ou un modèle 3D de la protéine et de l'ARN respectivement, à échantillonner des milliers/millions de positionnements relatifs possibles, puis calculer un score (approximation d'énergie) de chaque modèle obtenu (appelé une **pose**) pour identifier les modèles de meilleurs scores, parmi lesquels doit se trouver la structure cible. La discrimination finale entre ces quelques modeles peut se faire par expérimentation in vitro. Ces méthodes dites d'**amarrage** sont applicables aux cas où la conformation 3D de la protéine et de l'ARN varie peu entre leur forme libre et leur forme dans le complexe. Cela est souvent vrai pour les protéines et ARN très structurés (compactes), mais ne s'applique pas aux ARN simple-brin. Ces derniers sont désordonnés, c'est-à- dire qu'ils n'ont pas de structuration 3D propre mais une multitude de conformations 3D possibles dans leur état libre. Leur conformation dans le complexe protéine-ARNsb dépend fortement de la protéine à laquelle ils se lient, ce qui empêche de la modéliser avant de modéliser le complexe, comme dans l'approche classique. Certaines approches échantillonnent les conformations les plus probables du ligand flexible pour ensuite les amarrer. Mais le nombre d'angles variables - 12 par nucléotides - entraîne une explosion combinatoire lorsque l'on essaye de modéliser toutes les conformations d'un ARN de plus de 4 ou 5 nucléotides. Les méthodes à base d'apprentissage profond, qui ont révolutionné l'amarrage protéine-protéine, ne sont pas non plus applicables aux complexes protéine-ARNsb en raison du nombre trop faible de structures connues pour l'apprentissage.

L'approche état-de-l'art pour les complexes protéine-ARNsb est **ssRNA'TTRACT**, approche d'**amarrage par fragments**, basée sur la suite logicielle d'amarrage ATTRACT, Elle consiste à découper la séquence d'ARN en triplets chevauchants, à échantillonner toutes les conformations possibles de chaque triplet, à les amarrer sur la protéine, puis à identifier les positions de triplets chevauchants qui peuvent être connectées en un ARN complet (Fig. 4).

Objectifs de la thèse

ssRNA'TTRACT a comme facteur limitant l'incapacité de ATTRACT a produire des poses quasi-natives pour certains fragments, et a mal discriminer les poses quasi-natives des poses incorrectes pour les autres fragments. Ceci est dû à deux problèmes cumulés. Le 1er problème, propre à ssRNA'TTRACT, est que les paramètres de la fonction de notation de ATTRACT ont été obtenus par entraînement sur seulement environ 500 structures expérimentales de complexes ARN-protéine disponibles en 2010, qui ne contenaient pas d'ARNsb. Le 2eme problème, intrinsèque à toute approche par fragment, est que certains fragments "hot-spot" d'un ARN complet se lient à la protéine de façon plus spécifique ou avec une plus forte énergie

d'interaction que les autres, et ces autres se placent à des positions sub-optimal autour du ou des fragments "hot-spot". Ceci est particulièrement vrai pour les cas où tous les fragments ont la même séquence. Un seul fragment de la chaîne d'ARN peut alors être à sa position optimale dans la structure réelle. Les autres fragments seront alors moins bien amarrés car biaisés vers la position "hot-spot".

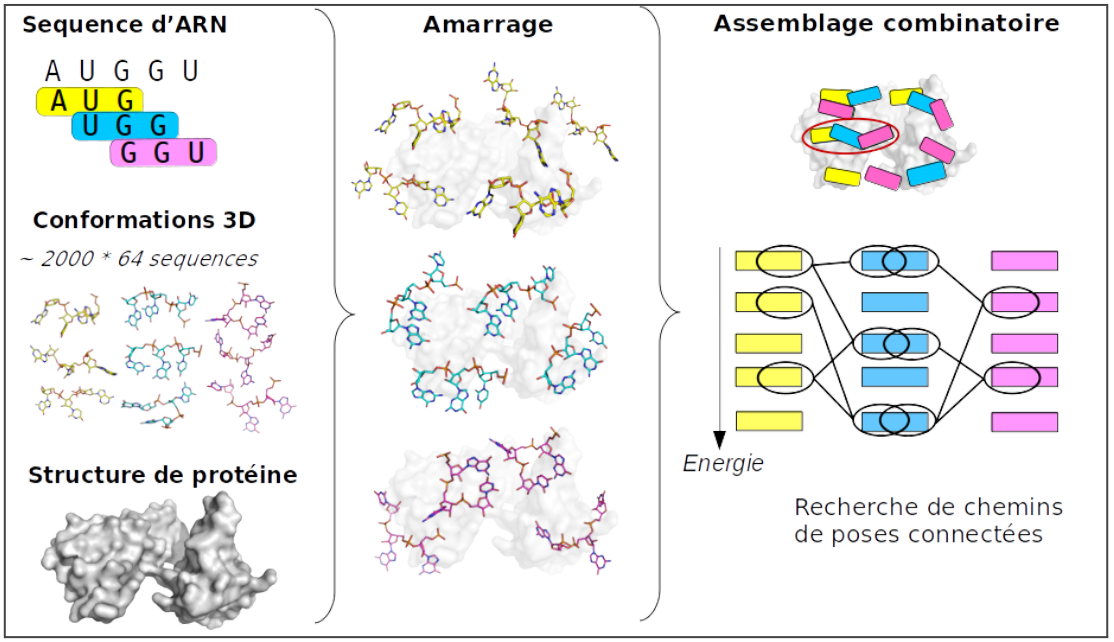


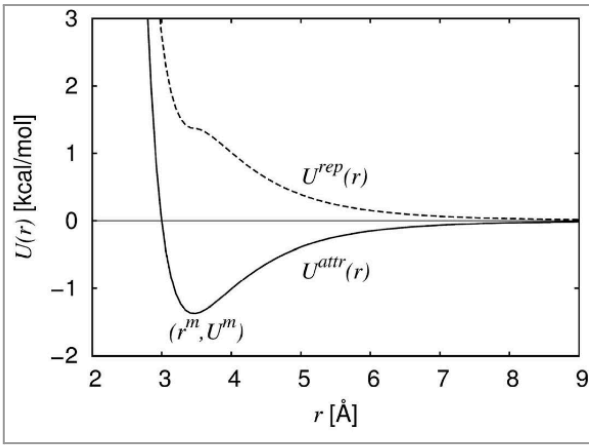
Fig.4 : Amarrage par fragments d'ARNsb.

L'objectif principal de la thèse était d'optimiser ou recalculer les paramètres de ATTRACT pour la notation des poses protéine-ARNsb, afin de résoudre le 1er problème. Un objectif secondaire était de palier au 2eme problème en restreignant l'espace de recherche des poses, par des contraintes d'amarrage correspondant à la caractéristique des interactions RRM-ARNsb qu'est l'empilement de nucléotides par les cycles aromatiques conservés du RRM.

II. Optimisation des paramètres de la fonction d'énergie de ATTRACT

Approche stochastique globale

ssRNA'TTRACT utilise une représentation gros-grain de l'ARN et de la protéine, dans laquelle plusieurs atomes sont remplacés par un pseudo-atome appelé bille. La note de chaque pose est calculée comme la somme des notes de chaque paire de bille ARN/protéine, qui elle-même dépend de la distance *d* entre ces 2 billes. Il existe 17 types de bille d'ARN et 31 de protéine. La fonction de notation entre deux billes de type i et j comporte deux paramètres, sigma_ij et epsilon_ij. La fonction de notation de ATTRACT, appelée ici ASF ("ATTRACT scoring function"), comporte donc 17*31*2 = 1054 paramètres d'interaction ARN-protéine. Ces paramètre avaient d'abord été obtenus à partir des structures protéine-ARN expérimentalement connues, en décrivant la probabilité de chaque distance pour chaque paire de type de bille a partir de sa fréquence dans ces structures connues, selon l'équation de Boltzmann. Ce 1er set avait été optimisé par une approche de Monte Carlo de façon à minimiser la note de la position native de l'ARN dans chaque complexe (pour sigma) puis pour minimiser le rang de la native parmi des poses leurre (pour epsilon).



$$U_{ij}^{attr}(r) = \varepsilon_{ij}\left(\frac{\sigma_{ij}^{8}}{r^{8}} - \frac{\sigma_{ij}^{6}}{r^{6}}\right)$$

$$U_{ij}^{rep}(r) = \begin{cases} U_{ij}^{attr}(r) + 2U_{ij}^{m}(r), if\ r \leq r_{ij}^{m} \\ -U_{ij}^{attr}(r), if\ r > r_{ij}^{m} \end{cases}$$

Fig 5: Forme et équations de ASF pour les paires de billes attractives ou répulsives. r est la distance interbilles, i dans [1,31] est un type de bille de protéine, j dans [1,17] est un type de bille d'ARN, le paramètre sigma_ij règle la distance de score minimal r_{ij}^{m} , et le paramètre epsilon_ij règle la valeur du score minimal U_{ij}^{m}. Plus le score est bas, mieux la paire de bille est notée.

Nous avons d'abord tenté d'optimiser ce set de 1054 paramètres par une approche de Monte-Carlo avec recuit simulé (approche **MSCA** pour *Monte Carlo Simulated Annealing*), un travail initié par Agniba Chandra, M2 dans l'équipe Capsid en 2019. Le but était d'augmenter la part de poses quasi-native dans les poses les mieux notées, lors d'expériences d'amarrage de fragments sur un ensemble test.

L'algorithme MCSA est une technique d'optimisation stochastique pour explorer des espaces de recherche complexes. Il commence par une marche aléatoire pour échapper aux minima locaux, puis devient progressivement plus déterministe pour affiner la solution. Il prend en entrée un set initial des paramètres à optimiser, la valeur initiale de la température et la sélection d'un programme de refroidissement pour réduire progressivement la température. A chaque itération, un nouvel ensemble de paramètres est généré en perturbant légèrement leurs valeurs actuelles. Si la valeur de de la fonction objectif obtenue avec le nouvel ensemble est meilleure qu'avec l'actuel, le nouveau est accepté. Sinon, l'acceptation dépend de la température. Les températures élevées permettent d'accepter aléatoirement des solutions moins bonnes, tandis que les températures basses n'acceptent que des solutions meilleures.

Nous avons extrait 42 structures RRM-ARNsb connues (totalisant 309 trinucleotides) de la base de données publique PDB, que nous avons divisé en un ensemble d'entraînement (36 complexes) et de validation (6 complexes). Nous avons empiriquement sélectionné 4 ensembles d'hyper-paramètres de MCsA (schéma de refroidissement, température initiale, etc), et créé 4 nouveaux ensembles de paramètres d'ASF, avec comme fonction objectif le nombre de quasi-native rangées dans les 10% de poses de meilleures notes.

Bien que certains des nouveaux ensembles soient plus performants que les paramètres originaux pour certains fragments, la performance globale de tous les ensembles s'est avérée très similaire. MCSA étant un algorithme stochastique, il est possible que la solution optimale n'ait pas été atteinte. Cependant, ces résultats suggèrent également qu'un seul ensemble de paramètres n'est pas capable de représenter la diversité des modes de liaison RRM-ARNsb, ce qui nous a amené à développer les paramètres HIPPO (voire section IV).

Analyse de quelques paramètres problématiques

Nous avons centré notre analyse sur un petit sous-ensemble de paramètres problématiques (reconnus pour attribuer des scores défavorables aux poses natives) en prenant les paramètres TRP-C (pour la chaîne latérale du tryptophane et la base C) comme exemple. Nous avons mis en place un petit ensemble de référence de structures expérimentales avec des interactions pertinentes et avons examiné les distances entre les billes, les scores des fragments natifs, le nombre de poses quasi-natives échantillonnées, etc. Nos investigations ont révélé que toutes les géométries TRP-C natives obtiennent des scores défavorables, malgré la formation de nombreuses interactions d'empilement pi-pi. De manière intrigante, malgré ces résultats, l'échantillonnage et l'évaluation réussissent relativement bien dans environ la moitié des cas de fragments de protéines. Nous avons entrepris plusieurs tentatives pour affiner empiriquement les valeurs des paramètres cibles afin d'améliorer l'échantillonnage, tout en évaluant l'impact de ces ajustements sur les performances globales de l'ASF. Les résultats, malgré l'utilisation de différentes valeurs de paramètres cibles, ont révélé des performances très similaires à celles de l'ASF, soulevant des interrogations quant à la faisabilité d'optimiser le sous-ensemble de paramètres de manière isolée. Cette faisabilité peut être examinée grâce à une approche systématique, comme celle proposée dans ce chapitre.

III. HIPPO: nouveaux potentiel d'évaluation des modèles RRM-ARNsb

Nous avons présenté une nouvelle approche d'optimisation fondée sur l'histogramme et HIPPO (“HIstogram-based Pseudo-POtential”), un potentiel de notation nouvellement conçu pour les poses d'amarrage des complexes protéine-ARNsb dans la représentation à gros grains d'ATTRACT. L'originalité principale de HIPPO réside dans sa **nature composite**, rassemblant quatre potentiels de notation distinctes, chacun capable de capturer des modes de liaison protéine-ARNsb spécifiques, i.e. HIPPO repose sur l'hypothèse qu'il existe une collection d'ensembles de paramètres de notation (comme par opposition à un seul ensemble de paramètres) qui peut être utilisé pour classer efficacement les solutions d’accueil protéine-ARNs quasi-natives. Les paramètres de HIPPO sont dérivés analytiquement des fréquences de contact dans des poses d’amarrage quasi-natives et des poses incorrectes. Ces fréquences de contact, dérivées de quatre ensembles différents de poses d'amarrage, sont discrétisées par un ensemble particulier de seuils en histogrammes, conduisant à une collection de quatre ensembles d'histogrammes $\mathcal{H}$ qui forment ensemble le potentiel de notation HIPPO. Ainsi, HIPPO est un potentiel de notation composite protéine-ARNsb: généralement, les 5% supérieurs des poses selon chaque ensemble d'histogrammes sont combinés, sélectionnant 20% de toutes les poses d'amarrage au total. Pour rationaliser le processus depuis la construction de l'ensemble de données jusqu'à la génération des paramètres de notation finaux, nous avons décidé de nous concentrer exclusivement sur les RRM, car ce domaine de la protéine est particulièrement important pour l'étude des interactions protéine-ARNsb et est présent dans de nombreux (environ 65 %) des les structures protéine-ARNs disponibles.

L'application de ces potentiels, suivie de l'agrégation de leurs résultats, a abouti à un classement plus précis par rapport à l'ASF de pointe. HIPPO a particulièrement amélioré la notation du fragment le mieux ajusté au sein de chaque complexe, facilitant ainsi son utilisation comme point d'ancrage pour l'amarrage incrémentiel. De plus, nous avons introduit le concept de **BP** (“best-performing potential”), pour l'instant limité au cas test. Dans cette approche, le potentiel le plus performant parmi les quatre est identifié et utilisé individuellement pour la notation.

Ensuite, nous avons exposé les résultats de l'application de HIPPO et de BP à une référence de complexes qui n'ont pas été utilisés lors du développement de HIPPO. Dans cette comparaison, HIPPO a surpassé l'ASF dans la notation de ces complexes, et BP a nettement dépassé HIPPO et l'ASF.

De plus, nous avons utilisé le classement ASF et le classement BP pour **assembler des poses** pour un petit sous-ensemble de 14 complexes (3 fragments par complexe). Les résultats préliminaires sont très prometteurs, car BP a surpassé les performances de l'ASF pour les quatre ensembles d'hyperparamètres testés. Ce succès motive fortement le développement d'un modèle permettant de dériver BP de HIPPO pour un cas donné.

Il est à noter que le protocole utilisé pour développer HIPPO (et potentiellement le modèle permettant de passer de HIPPO à BP) est a priori applicable à d'autres types de complexes. En outre, il pourrait être utilisé pour pallier une limitation inhérente à l'approche d'amarrage basée sur les fragments “hot-spot” (voire section ‘Conclusion et Perspectives’).

IV. Pipeline d’amarrage RRM-ARNsb

RRM-RNA dock

Nous avons présenté une approche basée sur les données compatible avec l'amarrage basé sur des fragments. Il utilise ce que l'on appelle des modèles d'ancrage, qui représentent les positions moyennes de l'ancre (nucléotide d'empilement), pour piloter l'amarrage. Ces modèles ont été dérivés du regroupement de structures RRM-ARN présentant des interactions d'empilement conservées. Ma contribution réside dans le développement du pipeline d'amarrage conçu pour faciliter l'amarrage basé sur les données de fragments contenant deux ancres sur RRM. Ce pipeline possède une interface de ligne de commande simple. Il utilise AlphaFold DB pour obtenir un modèle du RRM, InteR3Mdb pour identifier les acides aminés d'ancrage, les modèles d'ancrage pour identifier un emplacement approximatif des ancres par rapport aux acides aminés d'ancrage et le moteur d'amarrage ATTRACT pour effectuer l'amarrage avec des contraintes suite à un processus préalablement établi protocole piloté par ancre (Fig 6).

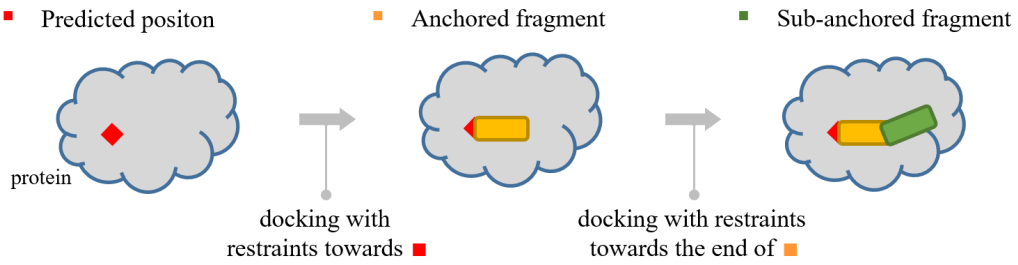


Fig 6: Image schématique de l'amarrage ancré du fragment ancré (en jaune) et d'un fragment adjacent (en vert). La position prévue de l'ancre est affichée en rouge.

Bien que présentant des fonctionnalités restreintes, ce pipeline simplifie significativement le processus d'amarrage moléculaire basé sur les données en éliminant la nécessité d'une préparation manuelle de la structure du récepteur et de la création manuelle de restraint fichiers. Par conséquent, elle rend l'amarrage moléculaire basé sur les données plus accessible, particulièrement pour les chercheurs moins expérimentés en biologie structurale. Comme anticipé, l'efficacité de ce système d'amarrage surpasse celle obtenue par des méthodes d'amarrage *ab initio*.

Restrictions expérimentales supplémentaires

Nous avons présenté un ensemble de données expérimentales non structurelles, collectées manuellement à partir de la littérature. Cet ensemble de données pourrait être utilisé comme source de contraintes d'amarrage

supplémentaires pour, à terme, étendre le RRM-RNA dock à un pipeline général d'amarrage protéine-ARNsb.

V. Conclusion et Perspectives

L'objectif principal de cette thèse était d'améliorer l'amarrage protéine-ARNsb en abordant le problème du scoring. Notre approche de ce problème impliquait le développement d'une nouvelle fonction de notation, HIPPO, spécifiquement adaptée aux interactions protéine-ARNs. De plus, nous avons développé un pipeline d'accueil convivial « RRM-RNA dock », adapté aux complexes RRM-ARNsb. Dans cette section, nous résumons les deux contributions et soulignons les principaux résultats, ainsi qu'abordons brièvement certains petits projets entrepris dans le cadre de cette recherche doctorale.

HIPPO

La base de HIPPO était une optimisation infructueuse par recuit simulé de Monte Carlo de l'ensemble des paramètres d'amarrage d'origine d'ASF, suivie de la conception de l'approche d'optimisation basée sur l'histogramme. Le premier projet nous a conduit à l'hypothèse selon laquelle un ensemble de paramètres singuliers est insuffisant pour évaluer avec précision les fragments protéine-ARNsb. Le deuxième projet était un préalable à la mise en œuvre du protocole permettant de dériver HIPPO. Contrairement aux fonctions de notation existantes, HIPPO est composé de 4 potentiels de notation distincts, capables de prendre en compte différents modes de liaison.

L'évaluation expérimentale de HIPPO a été réalisée en notant un ensemble de complexes protéine-ARNsb. HIPPO a surpassé l'ASF, l'état de l'art en matière d'amarrage basé sur des fragments de protéines et d'ARNsb dans une représentation à gros grains. De plus, ces résultats ont prouvé la généralisabilité de HIPPO, car il était dérivé exclusivement de complexes RRM-ARNsb. De plus, l'utilisation BP pour chaque cas de notation a donné un meilleur classement par rapport à ASF et HIPPO. Les résultats préliminaires de l'assemblage de 3 fragments de plusieurs complexes suggèrent que BP serait une fonction de notation appropriée pour l'amarrage incrémentiel.

RRM-RNA dock

Les interactions d'empilement entre les acides aminés dans des positions conservées et les nucléotides non appariés peuvent servir d'ancres et piloter l'amarrage protéine-ARNsb. Cette approche d'ancrage préexistante nécessite des informations sur les positions possibles du nucléotide d'empilement (ancre) par rapport à l'acide aminé d'empilement, c'est-à-dire des motifs ancrés. Un ensemble de modèles d'ancrage a été généré par Hrishikesh Dhonge via le regroupement des structures 3D déterminées expérimentalement de complexes RRM-ARN. En unissant ces modèles d'ancrage à la méthodologie d'ancrage-docking, nous avons créé un pipeline basé sur ATTRACT pour l'amarrage de fragments RRM-ARNsb. Ce pipeline obtient un modèle de RRM auprès d'AlphaFoldDB et exécute l'amarrage ATTRACT pour un fragment avec deux nucléotides empilés, avec des restrictions de distance maximale vers chaque position d'ancrage possible.

Comme prévu, ce pipeline offre un meilleur échantillonnage par rapport à l'amarrage ab initio. Son avantage notable réside dans son accessibilité aux non-experts en biologie structurale computationnelle. Les utilisateurs sont dispensés des tâches de préparation de la structure 3D du récepteur ou d'identification des positions des acides aminés, de construction de contraintes, etc.

Perspectives

Une approche incrémentale vs. Les potentiels doubles pour la 'hot-spot' et 'cold-spot'

Une difficulté majeure liée à l'amarrage basé sur des fragments concerne la distinction entre les liaisons aux points chauds ("hot-spot", HS) et aux points froid ("cold-spot", CS). Ce problème peut être adressé en utilisant HIPPO en conjonction avec une stratégie d'amarrage incrémentale, dans laquelle un unique fragment lié à un HS est amarré avec une grande précision, et le reste de la chaîne est modélisé fragment par fragment à partir des poses du premier fragment. HIPPO augmente considérablement la proportion de poses quasi-natives parmi les poses les mieux classées pour le fragment le mieux amarré d'un complexe, facilitant ainsi ce type de modélisation incrémentale. Dans ce contexte, le principal obstacle réside dans l'identification du fragment lié au HS avant l'amarrage (au moins avant l'assemblage). Lorsque l'identification s'avère impossible, une approche consiste à traiter chaque fragment comme étant lié à un HS, à itérer sur tous les fragments et à regrouper les résultats les mieux classés de chaque itération (Fig. 7).

En alternative, il est possible d'explorer le développement de potentiels de score doubles, capables d'évaluer précisément les poses d'amarrage des fragments liés aux HS ou aux CS de manière distincte, c'est-à-dire en utilisant une approche basée sur des fragments classique. Pour obtenir de tels potentiels doubles, les cas de données d'entraînement (fragments de référence pour dériver les potentiels de score) doivent être étiquetés comme étant liés à des HS ou des CS avant l'entraînement. Il existe deux approches pour cela :

- Approche 1. Définissons le fragment HS comme celui pour lequel le HIPPO/BP actuel est couronné de succès;
- Approche 2. Définissons le fragment HS comme étant à proximité immédiate des acides aminés HS. Les acides aminés HS pourraient être identifiés en utilisant une approche spécialisée (des tests initiaux sur les protéines-ssRNA peuvent être nécessaires).

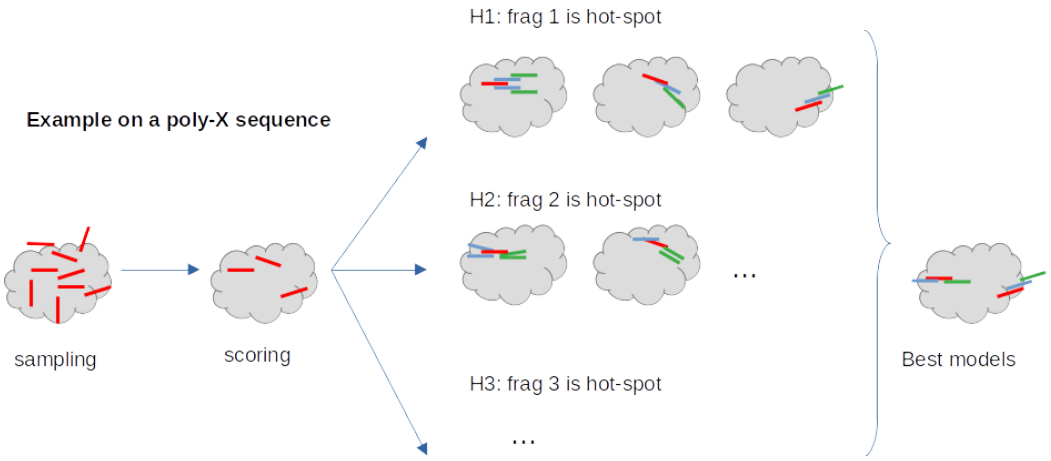


Fig 7: Illustration simplifiée de l'amarrage itératif utilisant chaque fragment comme un HS.

Sur la base de ces deux ensembles d'entraînement, un double potentiel de score, HIPPO-HS et HIPPO-CS, peut être dérivé. Il est possible que moins de quatre ensembles de paramètres de score soient nécessaires pour chaque ensemble d'entraînement, ce qui pourrait simplifier l'application des potentiels résultants. Si, à ce stade, il n'existe aucun modèle capable d'identifier le type de fragment (lié à un HS ou à un CS) avant l'amarrage, chaque fragment subirait deux tours de notation - une fois avec HIPPO-HS et une fois avec HIPPO-CS. Plusieurs assemblages (k assemblages), où k représente le nombre total de fragments dans le complexe, seraient effectués.

Le développement d'un potentiel de score double aboutirait à huit ensembles de paramètres distincts (quatre pour HIPPO-HS et quatre pour HIPPO-CS). Bien que cela puisse sembler lointain, il pourrait exister une voie pour passer des concepts HIPPO-HS et HIPPO-CS au concept de BP (“best-performing potential”). Cette transition pourrait être réalisée en formant un classificateur pour identifier les paramètres BP appropriés pour chaque fragment, basé sur les données d'entrée d'amarrage ou directement à partir de l'ensemble des poses d'amarrage.

Caractérisation des modes de liaison Protéine–ANRsb en utilisant BP

HIPPO est capable de prendre en compte les différents modes de liaison protéine-ARNsb, et en particulier RRM-ARNsb, en utilisant 4 potentiels de score distincts. Le concept de BP consiste à identifier, pour un complexe et un fragment donnés, le $\mathcal{H}$ unique qui surpasse les trois autres $\mathcal{H}$ dans le classement des poses quasi-natives dans les 5% supérieurs. Cette approche peut permettre la caractérisation des différents modes de liaison protéine-fragment - ensembles d'interactions protéine-ARNsb distinctes ayant une signification biologique - en distinguant essentiellement un mode des autres, sur la base du BP.

Le BP peut être facilement identifié pour chaque cas de test. Par conséquent, tous les cas (l'ensemble du benchmark) peuvent être catégorisés en cinq classes, avec quatre classes correspondant à chaque $\mathcal{H}$, et la classe restante représentant les cas atypiques, où aucun des $\mathcal{H}$ n'est couronné de succès (le critère de succès reste à déterminer). Idéalement, il y aura des ensembles distincts de distances entre perles pour chaque classe, ou des caractéristiques de niveau supérieur (par exemple, distances et angles). L'identification initiale de ces caractéristiques pourrait être réalisée par un examen manuel, suivi, par exemple, de l'application des approches d'exploration de motifs.

Application étendue de HIPPO

- La principale orientation de travail concernant HIPPO est la formation d'un classificateur pour permettre l'utilisation du BP au-delà du cas de l'ensemble d'entraînement. Un tel modèle pourrait être obtenu en se basant sur la séquence du fragment et la séquence et/ou structure de la protéine, et/ou sur les poses d'amarrage. Activer le BP améliorerait considérablement l'échantillonnage, comme montré dans le §4.4;
- Comme mentionné précédemment, HIPPO pourrait être utilisé pour l'échantillonnage en utilisant le moteur d'amarrage ATTRACT et la procédure de minimisation de Monte-Carlo. C'est l'une des perspectives les plus prometteuses, car cela pourrait atténuer le problème d'échantillonnage;
- Lors de la dérivation de HIPPO, le seuil de LRMSD pour les poses quasi-natives a été assoupli de 3Å à 5Å afin d'obtenir plus de cas de données avec un nombre plus élevé de quasi-natives. Il serait intéressant d'explorer si un nombre inférieur de quasi-natives est suffisant pour la dérivation d'un HIPPO efficace;
- Il serait intéressant de tester la précision de HIPPO sur les chaînes complètes de ARNsb ;
- Enfin, le protocole peut être appliqué à d'autres types de complexes, tels que protéine-ARNsb au-delà du domaine RRM, protéine-ADNsb et protéine-peptides.

Ce chapitre présente les autres perspectives, telles que: 1/ un pipeline pour l'amarrage itératif, 2/ l'expansion du benchmark protéine-ARNsb, 3/ une validation croisée rigoureuse pour HIPPO et 4/ une enquête sur la flexibilité.

Fragment-based modelling of protein-RNA complexes for protein design

Abstract

Protein-RNA complexes play crucial roles in cell regulation. Predicting their 3D structure has applications in protein design and drug development. The ITN project RNAct aimed to combine experimental and computational methods to design new "RNA recognition motifs" (RRM) - protein domains interacting with single-stranded RNA (ssRNA) - for applications in synthetic biology and bioanalysis.

Modelling protein-ssRNA complexes (*docking*) is an arduous task due to the flexibility of ssRNA, which lacks a proper structure in its free form. Traditional docking methods sample the relative positions (*poses*) of 2 molecular structures and score them to select the correct (near-native) ones. It is not directly applicable here due to the absence of free ssRNA structures, nor is deep learning due to the too low number of known structures for training. Fragment-based docking (FBD), the state-of-the-art approach for ssRNA, docks all possible conformations of RNA fragments onto a protein and assembles their best-scored poses combinatorially. ssRNA'TTRACT, our FBD method, uses the well-known ATTRACT docking software, with its coarse-grained representation that replaces atom groups by one bead. Yet the RNA-protein parameters of ATTRACT scoring function (ASF) are not ssRNA-specific and require optimisation. Additionally, RRM-specific features can be learned and used to guide the docking.

With my colleague H. Dhondge, we have developed a data-driven FBD pipeline for RRM-ssRNA complexes, as an updated version of an existing strategy. RRMs have two aromatic amino acids (aa) in conserved positions, each stacking with a nucleotide of the bound ssRNA. H. Dhondge collected all known RRM-ssRNA structures with such stacking and clustered them to obtain a set of prototypes for the 3D coordinates of such interactions in RRM. I then set up a docking pipeline with as input the RRM and RNA sequences and the identification of the stacked nucleotides. The pipeline retrieves the RRM structure from AlphaFoldDB, identifies possible 3D positions of the stacked nucleotides and runs ssRNA'TTRACT with maximal distance restraints toward each position.

In parallel, we addressed the weakness of ASF for ssRNA by deriving HIPPO (HIstogram-based Pseudo-POtential), a new scoring potential for ATTRACT poses of ssRNA on RRM, based on the frequency of bead-bead distances in near-native versus wrong poses. It combines 4 distinct parameter sets (four $\mathcal{H}$) into a consensus scoring, to better account for the diverse RRM-ssRNA binding modes. Tested in a leave-one-out approach, HIPPO reaches a 3-fold enrichment of near-natives in 20% top-scored poses for ½ of the ssRNA fragments, versus ¼ with ASF. It even reaches a 4-fold enrichment for ⅓ of the fragments, versus 7% of the fragments with ASF. Surprisingly, HIPPO performed better than ASF also on a benchmark of non-RRM proteins, while trained only on RRMs.

Most FBD approaches encounter inherent scoring issues, probably due to some fragments binding more specifically/strongly than others. To address this point, we examined the best-scored fragment per complex and found that HIPPO consistently selects more near-natives than ASF for this fragment. This inspired an incremental docking approach: the top-ranked poses of one fragment are used as a starting point to build a full RNA chain incrementally. This strategy eliminates the need for known conserved contacts, which have been required so far to obtain accurate models, making it generalizable to non-RRM proteins.

Future research aims to identify the best-performing $\mathcal{H}$ for each fragment, potentially using (deep) machine learning. Our workflow to derive scoring parameters is in principle applicable to any protein/ligand type and we plan to expand it to other RNA-binding protein domains, as well as ssDNA and long peptides.

Modélisation par fragments de complexes protéine-ARN pour le design de protéines

Résumé

Les complexes protéine-ARN jouent un rôle crucial dans la régulation cellulaire. La prédiction de leur structure 3D a des applications dans la conception de protéines et de médicaments. Le projet ITN RNAct visait à combiner des méthodes expérimentales et informatiques pour concevoir de nouveaux "motifs de reconnaissance de l'ARN" (RRM) - domaines protéiques interagissant avec l'ARN simple brin (ARNsb) - pour la biologie synthétique et la bioanalyse.

La modélisation des complexes protéine-ARNsb (amarrage) est ardue car l'ARNsb n'a pas de structure propre dans sa forme libre. L'amarrage traditionnelle échantillonne les positions relatives (poses) de 2 structures moléculaires et les note pour sélectionner les plus probables. Il n'est pas directement applicable ici en raison de l'absence de structures libres d'ARNsb, pas plus que l'apprentissage profond en raison du nombre trop faible de structures connues. L'amarrage par fragments, état de l'art pour l'ARNsb, amarre toutes les conformations possibles de fragments d'ARN sur une protéine et assemble les poses les mieux notées de manière combinatoire. Notre méthode ssRNA'TTRACT utilise le logiciel d'amarrage ATTRACT et sa représentation gros grain qui remplace des groupes d'atomes par une bille. Cependant, les paramètres ARN-protéine de sa fonction de notation (ASF) ne sont pas spécifiques à l'ARNsb et peuvent être optimisés. De plus, des caractéristiques spécifiques aux RRM peuvent être apprises et guider l'amarrage.

Nous avons développé un pipeline d'amarrage RRM-ssRNA basé sur les données, pour actualiser une stratégie existante. Les RRM ont 2 acides aminés aromatiques de position conservée, chacun liant par empilement un nucléotide de l'ARN. Mon collègue H. Dhondge a regroupées les structures RRM-ARNsb connues sur critère géométrique et obtenu un ensemble de prototypes de coordonnées 3D de tels empilements dans les RRM. J'ai créé un pipeline qui prend en entrée une séquence de RRM et d'ARN et l'identification des nucléotides empilés, récupère la structure du RRM dans AlphaFoldDB, identifie les positions 3D possibles des nucléotides empilés et exécute ssRNA'TTRACT avec des contraintes de distance maximales vers chaque position.

En parallèle, nous avons dérivé HIPPO (HIstogram-based Pseudo-POtential), un potentiel de notation pour les poses gros-grain RRM-ARNsb basé sur la fréquence des distances bille-bille dans les poses quasi-natives versus erronées. HIPPO combine 4 ensembles de paramètres (quatre $\mathcal{H}$) en une note consensus, afin de prendre en compte les divers modes de liaison RRM-ARNsb. Testé dans une approche "leave-one-out", il atteint un enrichissement d'un facteur 3 en quasi-natives dans les 20% de poses mieux notées pour ½ des cas contre ¼ avec ASF, et 'un facteur 4 pour ⅓ des cas contre 7% avec ASF. Surprenamment, HIPPO obtient aussi de meilleurs résultats qu'ASF sur un ensemble test de protéines sans RRM, bien que entraîné sur des RRM.

Les approches par fragment rencontrent un problème intrinsèque de notation car certains fragments se lient plus spécifiquement/fortement que d'autres. Or nous avons constaté que, pour le fragment le mieux noté par complexe, HIPPO sélectionne systématiquement plus de quasi-natifs qu'ASF. Cela nous a inspiré une approche d'amarrage incrémentale: chacune des poses bien notées d'un fragment sont utilisées comme graine pour construire une chaîne d'ARN complète de manière incrémentale. Cette stratégie élimine le besoin de contacts conservés connus, jusqu'alors nécessaires pour obtenir des modèles précis, ce qui la rend généralisable aux protéines sans RRM.

Nos recherches futures visent à identifier le $\mathcal{H}$ le plus performant pour chaque fragment, potentiellement par apprentissage automatique (profond). Notre approche pour dériver des paramètres de notation est en principe applicable à tout type de protéine/ligand et nous prévoyons de l'étendre à d'autres domaines de protéines liant l'ARN, ainsi qu'à l'ADNsb et aux peptides longs.