Advanced computational techniques to aid the rational design of small molecules targeting RNA

Francesco Paolo Panei

Sorbonne Université

Ecole doctorale Complexité du Vivant

Unité de Biologie Structurale Computationnelle, Institut Pasteur

Integrated Drug Discovery, Sanofi R&D

Advanced computational techniques to aid the rational design of small molecules targeting RNA

par

Francesco Paolo Panei

Thèse de doctorat de Bioinformatique et biologie des systèmes

dirigée par

Massimiliano Bonomi	*Directeur de thèse*
Paraskevi Gkeka	*Co-encadrante de thèse*

Présentée et soutenue publiquement le 30 avril 2024

Devant un jury composé de :

Alessandra Magistrato	*Rapporteuse*
Guillaume Stirnemann	*Rapporteur*
Elodie Laine	*Examinatrice et présidente du jury*
Modesto Orozco	*Examinateur*
Massimiliano Bonomi	*Directeur de thèse*
Paraskevi Gkeka	*Co-encadrante de thèse*

"Happiness is happening
The dragons have been bled
Gentleness is everywhere
Fear's just in your Head
Only in your Head
Fear is in your Head
Only in your Head
So Forget your Head
And you'll be free"
David Bowie, *Fill your heart* from Hunky Dory (1972)

Contents

Outline of the thesis

In recent years, RNA has gained significant attention as a potential therapeutic target, especially for small-molecule drugs. Despite this, the number of available RNA-targeted drugs remains limited. The experimental screening techniques used to identify most RNA binders involve iterative experimentation and do not depend on prior knowledge of binding properties. However, RNAs are highly flexible targets and both characterizing and leveraging RNA structural dynamics are needed to make RNA-targeted therapeutics even more relevant. Structure-based approaches in Computer-Aided Drug Design (CADD) employ three-dimensional information about the target to guide the rational design and optimization of potential drug candidates. Due to the limited availability of RNA-small molecule structures, relatively few computational tools have been developed so far to assist in the identification of candidate drugs. In particular, existing structure-based methods mostly rely on techniques developed for protein targets and they do not fully account for the inherent flexibility of RNA molecules. My thesis aims to address this critical gap and develop computational tools that will lay the foundation for a more comprehensive framework to capitalize on the current possibilities presented by RNA-targeted therapeutics.

In Chapter 1, I will delineate the background of my work. RNA molecules are briefly presented from several points of view: biochemical, functional, and as therapeutic targets. Then, an extended discussion will explore their complex structural dynamics, discussing the potential therapeutic opportunities and the role of computational methods to this end. Finally, I will review the state-of-the-art of structure-based rational design of RNA-targeted drugs, specifically emphasizing the consideration of RNA flexibility within existing tools.

Among the main obstacles in the development of structure-based approaches, there is the lack of a comprehensive, curated and regularly updated repository collecting all the RNA-small molecule structures. To fill this gap, the first part of my work has been dedicated to creating HARIBOSS (Harnessing RIBOnucleic acid—Small molecule Structures), an online database of RNA-small molecule structures. In Chapter 2, I will *i)* describe the technical aspects behind HARIBOSS, *ii)* show the results of the physicochemical analysis of its entries, and finally *iii)* discuss the relevance and limitations of the work for drug design purposes.

Identifying potential small molecule binding sites is a first step for rational drug design. The physicochemical analysis of HARIBOSS revealed the inaccuracies of existing tools used for the characterization of RNA binding sites, mostly working on single RNA structures. In this context, a requirement to advance structure-based drug design was the implementation of a binding site

detection tool able to fully account for the flexible nature of RNA. Therefore, the main part of my work consisted in the development of SHAMAN (SHAdow Mixed Solvent MetAdyNamics), an advanced computational method to identify small molecule binding sites in RNA conformational ensembles. In Chapter 3, I will *i)* present the technical details of the SHAMAN approach, *ii)* show the results of its benchmark against a set of representative RNA targets, and finally *iii)*discuss the relevance and the limitations of the work for drug design purposes.

The tools that I developed during my PhD help tracing guidelines to develop more effective approaches for the rational design of small molecules targeting RNA. In Chapter 4, I will summarize the key findings of my research, discuss their relevance, and outline future perspectives for the field.

List of publications

Chapters 2 and 3 are assembled from the two articles I published during my PhD [1]:

- F P Panei, R Torchet, H Ménager, P Gkeka, M Bonomi, HARIBOSS: a curated database of RNA-small molecule structures to aid rational drug design, Bioinformatics, Volume 38, Issue 17, September 2022, Pages 4185–4193, https://doi.org/10.1093/bioinformatics/btac483
- F P Panei, P Gkeka, M Bonomi, Identifying small molecules binding sites in RNA conformational ensembles with SHAMAN, biorXiv, 2024, https://doi.org/10.1101/2023.08.08.552403 (under review)

[1]The mentioned Chapters have their own bibliography, which is reported at their end. For simplicity, the main and supplementary bibliographies have been merged.

Chapter 1

Introduction

While traditionally viewed as a genetic information carrier, RNA performs a variety of crucial roles, spanning gene expression regulation, immune defense, genome maintenance, and catalysis. As a consequence, RNA molecules have become key therapeutic targets and small molecule targeting has emerged as a promising approach. However, the development of such approaches is currently hindered by the limited understanding of RNA-ligand interactions. In particular, the flexible nature of RNA molecules hampers the possibility of a comprehensive biophysical characterization by both experimental and computational techniques. At the same time, the inherent flexibility of RNA also offers unique therapeutic opportunities. Computational methods, and in particular Molecular Dynamics simulations, have the potential to describe RNA structural dynamics at atomistic details and offer a natural framework to streamline the search for RNA-targeted drugs. However, the new avenues opened by structure-based rational design are still not fully capitalizing on the opportunities introduced by RNA flexibility.

1.1 RNA molecules: essential biology

RNA is a unique nucleic acid biomolecule characterized by a greater flexibility and reactivity with respect to DNA. The variety of intra-molecular interactions that RNA can engage in the cellular environment leads RNA to fold into a vast array of conformations, despite its relatively low chemical diversity.
In this section, I will first introduce the essential biology of RNA molecules, overviewing their synthesis as well as their inter- and intra- molecular interactions. Then, I will briefly overview their heterogeneous structurome arising from RNA folding, by showing the most common secondary and tertiary structures motifs that are formed through this process. Except where explicitly indicated, the discussion is inspired by the textbook "Molecular Biology of the Cell" by B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter [1].

1.1.1 Physicochemical properties of RNA nucleotides

RNA, or ribonucleic acid, is a linear polymer macromolecule primarily composed of four building blocks, known as nucleotides. Each nucleotide is composed of a planar aromatic base, a furanose-ring sugar moiety, and a phosphate group (Fig. 1.1A). The sugar is connected to the base and the phosphate group by a glycosidic and phopshoester bond, respectively. The four different nucleotides mainly differ by the chemical composition of the aromatic base: adenine (A), guanine (G), cytosine (C), and uracil (U) (Fig. 1.1B).

The synthesis of the RNA macromolecule occurs *via* transcription, a fundamental biological process that copies the genetic information embedded in DNA onto RNA molecules. During transcription, the RNA chain is assembled by the RNA polymerase enzyme through a series of covalent bonds known as phosphodiester bonds, which link the phosphate group of one nucleotide to the hydroxyl group on the sugar of another. At the end of the process, the RNA polymer is composed of a hydrophilic and negatively charged backbone, consisting of the alternate sequence of phosphate groups and sugars, joint to the more hydrophobic nucleobases (Fig. 1.1C).
RNA is characterized by a greater flexibility with respect to DNA nucleic acids. This is mainly attributed to the distinctive hydroxyl group located at the 2' position of the ribose sugar (dashed circle in (Fig. 1.1A). First, the 2'- OH is a versatile hydrogen bond donor and acceptor and is the principal responsible for the characteristic lower chemical stability of RNA molecules. Moreover, the 2' hydroxyl group triggers RNA hydrolysis, which consists of the breaking of the RNA backbone. This occurs as the 2' oxygen atom can interact with the adjacent phosphate group, leading to cleavage of the phosphodiester bond with the 5' carbon of the next nucleotide.

1.1.2 RNA intra-molecular non-bonded interactions

Once united by phosphodiester bonds, nucleotides experience a wide range of non-bonded intra-molecular interactions all along the RNA chain. The most frequent interactions are overviewed in the following.

- **Base pairing.** Base pairing interactions involve hydrogen bonds between nucleobases and represent the strongest intra-molecular interaction among RNA. Like DNA, RNA forms canonical base pairs following Watson-Crick complementarity rules, where A pairs with U and G

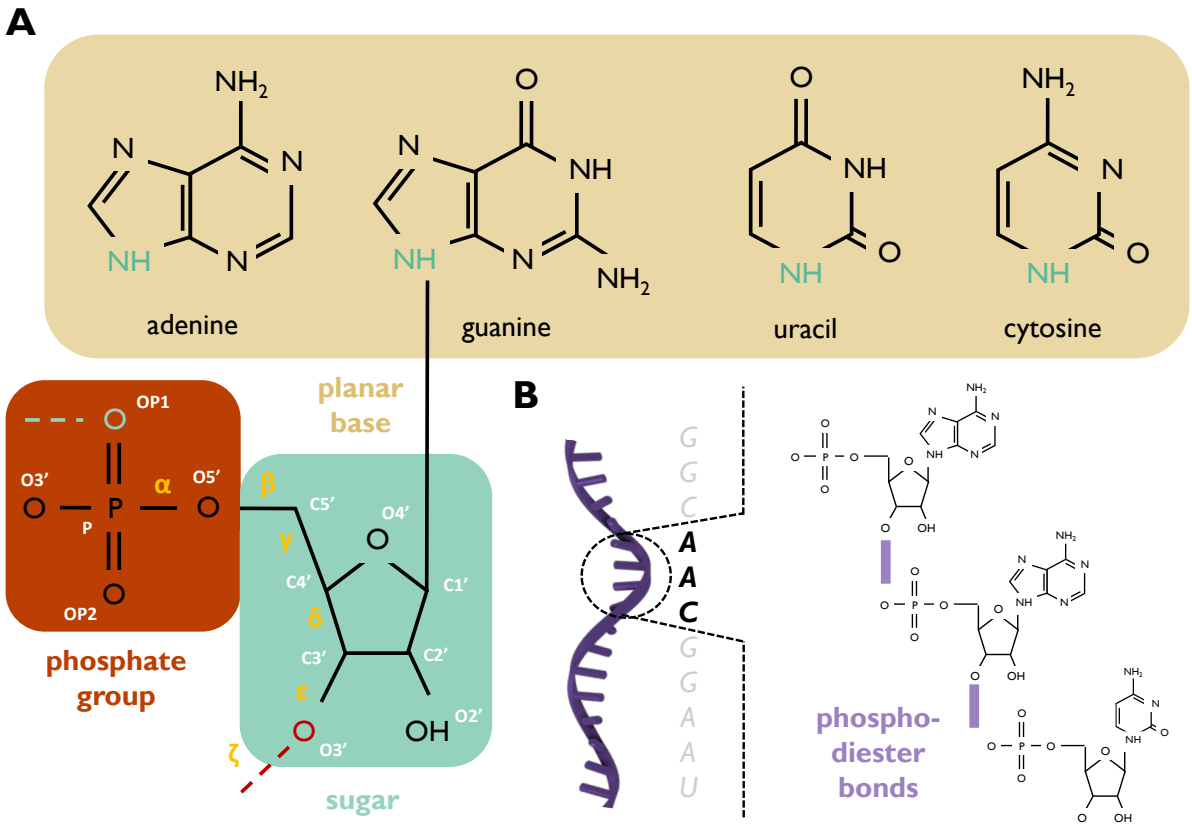


Figure 1.1: RNA building blocks. A) The generic chemical composition of an RNA nucleotide. On the bottom, the penta-carbon sugar moiety (red background), and the phosphate group (turquoise background). Covalent bonds between atoms are marked with solid lines. White text reports the standard nomenclature of RNA atoms. Yellow Greek letters indicate the dihedral angles of the RNA backbone (Sec. 1.1.3). The connection with the previous and next nucleotides is shown in color-coded dashed lines. On the top, the four different aromatic bases (khaki background): from left to right, adenine (A), guanine (G), cytosine (C), and uracil (U). The atom of the basis bonded to the sugar is highlighted in turquoise. **B)** On the left, a diagrammatic single-stranded RNA annotated with its sequence. The dashed inset focuses on a region of three nucleotides (bold text). On the right, the corresponding chemical composition of the chain composed by these three nucleotides. Phosphodiester bonds are highlighted with purple lines.

pairs with C, providing stable foundations to the RNA structure. The high favorability of Watson-Crick base pairings underscores their extreme specificity in RNA interactions. Due to its unique chemical reactivity, RNA also forms non-canonical base pairings. The stability of these pairs hinges on factors such as sterics, which require glycosidic bond rotation, protonation of the bases, and/or direct metal binding.

- $\pi - \pi$ **stacking.** Stacking interactions in RNA involve the non-covalent association between the aromatic rings of the nucleobases, which result in their stacked arrangement in space.

- **electrostatic.** The phosphate backbone of RNA is negatively charged, leading to electrostatic interactions. These include repulsion between similarly charged groups and attractions with positively charged ions or molecules, which are fundamental in a variety of biological processes.

- **van der Waals.** Van der Waals interactions are weak attractive forces that occur between

atoms in proximity. They arise from fluctuations in their electronic distribution, leading to temporary dipoles or induced dipoles in nearby atoms or molecules.

1.1.3 The heterogeneity of the RNA structurome

Despite the relatively low chemical diversity of its four building blocks, RNA adopts a wide range of conformations. This is due to the unique physicochemical properties of RNA and due to the diversity of the intra-molecular and environmental interactions that take place during RNA folding. In this section, I will briefly overview the process of RNA folding and present representative examples of the structural heterogeneity of RNA molecules.

RNA folding

RNA folding involves transforming linearly transcribed RNA molecules into specific three-dimensional shapes through intra-molecular contacts. Unlike protein folding, which relies on burying hydrophobic amino acids, RNA's secondary structure formation is governed by the hydrophobic nature of nucleobases [2]. Achieving a three-dimensional conformation requires managing electrostatic repulsion in the anionic sugar-phosphate backbone. This is primarily addressed through interactions with the surrounding environment, including water and ions. Positively charged metal ions, especially magnesium divalent ions (Mg^{2+}), accumulate near RNA molecules through electrostatic interactions with phosphate groups, facilitating and stabilizing RNA folding [3]. The polar phosphate groups also serve as primary hydration sites for surrounding water molecules [4]. The two free oxygen atoms of each phosphate group can form up to three hydrogen bonds with water molecules, crucially shielding electrostatic repulsion in the RNA backbone. Additional hydration sites include the sugar's 2' hydroxyl groups and the polar groups of nucleobases, specifically the carbonyl (C=O) and amino (NH_2) groups.

The conformations that the folded RNA molecules assume in space can be viewed from two perspectives. Before forming a three-dimensional or tertiary structure, the linear strand of RNA transcript folds into a secondary structure, which is mainly determined by base pairing interactions or by the interactions between hydrogen bond donor and acceptor groups between close nucleobases. In both cases, the rotational angles around the bonds connecting consecutive nucleotides, referred to as dihedral angles, play an important role in defining the folded structure of RNA. The most important ones, from the computational modeling perspective of this thesis, are (Fig. 1.1A):

- α**:** between the P-O5' bond, significant for the phosphate-sugar backbone structure;
- β**:** around the O5'-C5' bond, affecting the orientation of the phosphate group;
- γ**:** around the C5'-C4' bond, critical for the sugar puckering affecting the overall conformation of RNA;
- δ**:** around the C4'-C3' bond, essential for the backbone structure;
- ϵ**:** around the C3'-O3 bond, influencing the phosphodiester linkage and backbone flexibility;
- ζ**:** around the phosphate group and the O3' atom of the preceding ribose, influencing the overall conformation of the RNA chain;

In the two following sections, I will report some examples of common secondary and tertiary structures assumed by RNA molecules.

RNA secondary structure motifs

The most common and stable secondary structure in RNA forms through canonical base pairings, which give rise to a duplex domain (Fig. 1.2A). The highly favorable nature of base pairings contributes to the stability of RNA duplexes. In addition, nucleotides that are not involved in hydrogen bonds enable the formation of single-stranded secondary structures unique to RNA and referred to as loop regions (Fig. 1.2B). Common RNA loops are situated at the bottom (apical loop) or within the inner part (internal loops) of helical regions, or they may involve only one strand of the duplexes (bulge loops). In such regions, it is common that RNA sequences present multiple repeats of the same nucleotide. The 2D structure composed of a helical region capped with an apical loop is called stem-loop or hairpin (Fig. 1.2C). The latter can be found at the junction of more than one helix, resulting in a multi-way junction (Fig. 1.2D). Furthermore, non-canonical base pairing patterns contribute to the structural complexities of RNA. For instance, uridines can form base pairing on both sides of the adenines. This unique hydrogen bonding pattern is referred to as Hoogsteen, and it can result in a base triple when combined with the canonical base pair (Fig. 1.2E).

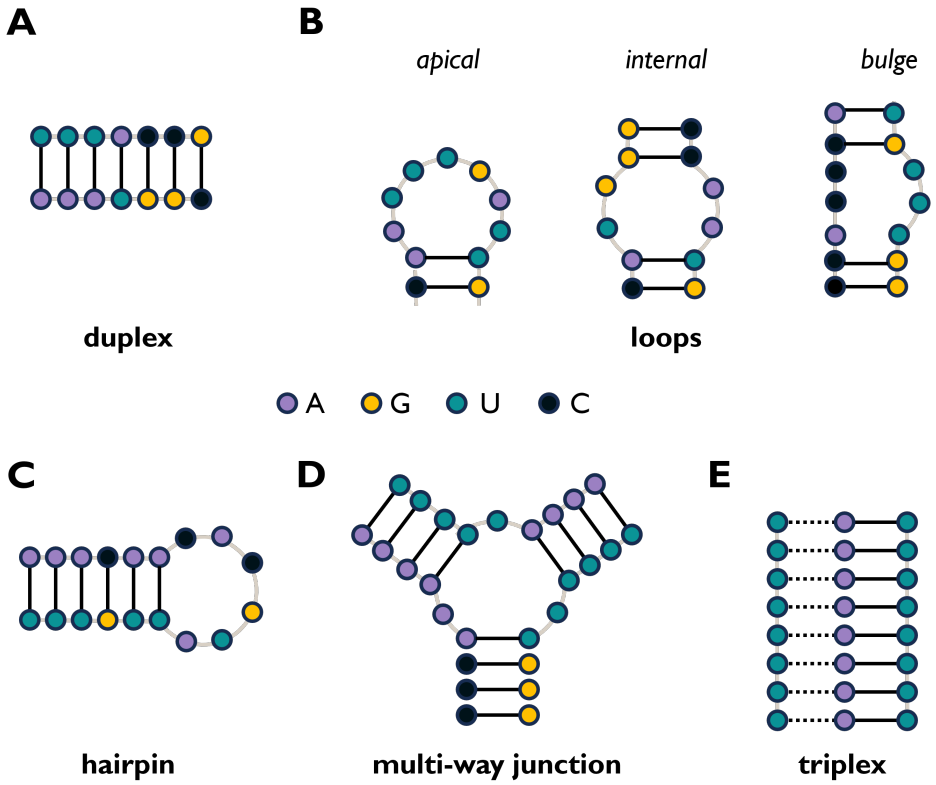


Figure 1.2: RNA secondary structure motifs. A-E) Diagrammatic examples of representative RNA secondary structure motifs: the duplex (**A**), three kinds of loop regions (**B**), the hairpin (**C**, also stem-loop), the junction between helical regions (**D**), and the triplex (**E**). Color-coded dots represent the 4 different RNA nucleotides. Abbreviations A, G, U, and C stand for adenine, uracil, and cytosine, respectively. Light grey lines indicate the phosphodiester bonds between nucleotides. Canonical and non-canonical base pairings are annotated with solid and dotted black lines, respectively.

RNA tertiary structure motifs

The tertiary structure of RNA emerges as spatial interactions occur among elements of the secondary structure. Despite potential separation in the primary or secondary structure, these areas closely interact during RNA folding. The 3D configuration of RNA duplexes generates helical domains, robustly stabilized by cooperative stacking interactions between base pairs. The overall folding architecture of RNA tertiary structures relies heavily on the coaxial stacking of adjacent helices and is influenced by the topology of RNA junctions between neighboring helices. An illustrative case exemplifying the arrangement of distinct helical domains in three-dimensional space is provided by yeast phenylalanine transfer RNA [5] (Fig. 1.3A). Alongside coaxial stacking, more complex structures may arise from tertiary interactions between independent secondary structures, as in the case of kissing loops (Fig. 1.3B) or from phosphodiester bonds between unpaired nucleotides, as in the case of pseudoknots (Fig. 1.3C).

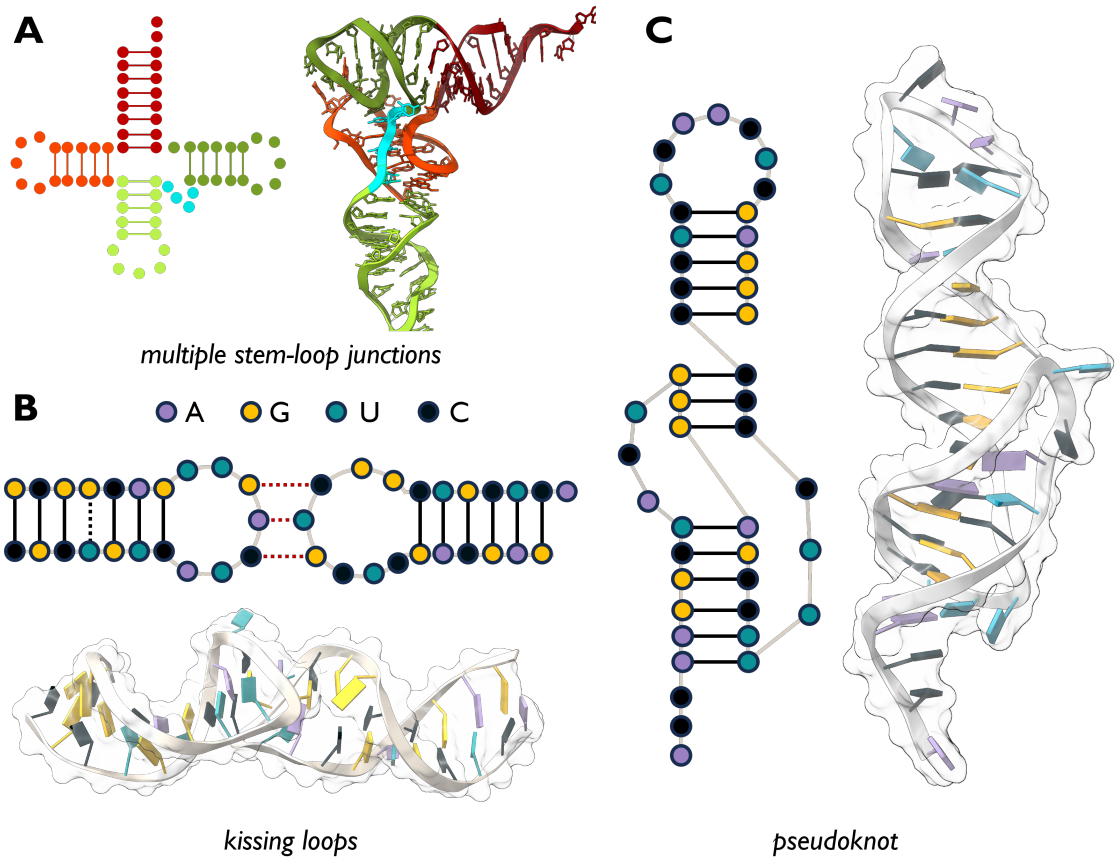


Figure 1.3: Examples of RNA tertiary structures . A) Secondary (left) and tertiary (right) structures of the yeast phenylalanine tRNA (PDB 6tna [5]). Different domains of the RNA are color-coded. **B-C)** Secondary (top in **B** and left in **C**) and tertiary (bottom in **B** and right in **C**) structures of the *Neurospora* Varkud satellite ribozyme (PDB 2mi0 [6]) and Turnip yellow mosaic virus RNA (PDB 1a60 [7]), respectively. Color-coded dots represent the 4 different RNA nucleotides. Abbreviations A, G, U, and C stand for adenine, uracil, and cytosine, respectively. Canonical and non-canonical base pairings are annotated with solid and dotted black lines, respectively.

The heterogeneous set of distinctive tertiary structures adopted by RNA provides the foundation to explore its multiple roles in cellular functions, which are presented in the next section.

1.2 RNA functions beyond coding genetic information

For a long time, RNA has been classified for its intermediary role within the framework of the so-called "Central Dogma" of molecular biology. The prevailing notion within the scientific community was that the indispensable biological functions in organisms were primarily executed by proteins, with RNA merely serving as a passive carrier of genetic information. However, over the past few decades, the understanding of RNA role within the cell has dramatically transcended these notions. Due to the increasing number of studies characterizing its wide structurome, RNA has been recognized to perform a plethora of biological functions. Most notably, RNA has been found to regulate gene expression at various stages of protein synthesis, encompassing epigenetic modifications, modulation of RNA-processing, and direct translation inhibition. Moreover, RNA plays a prominent role in the immune system of certain organisms and the maintenance of cellular homeostasis. The era referred to as the "non-coding RNA revolution" unfolded gradually, following the technological and methodological development of experimental and computational methods.
This section begins by providing a historical overview of the significant discoveries in RNA functions over the past decades. Subsequently, a classification of the RNA molecules identified to date will be presented. To elucidate the diverse roles played by RNA, I will discuss representative examples of its functions.

1.2.1 Historical outline of the "non-coding RNA revolution"

The following paragraphs present a historical outline of the key discoveries about RNA cellular functions. After introducing the role of RNA in the early years of molecular biology, I will overview the process known as the "non-coding RNA revolution" along three main phases (Fig. 1.4): *i)* the complete understanding of protein synthesis, *ii* the discovery of regulative RNAs, and *iii)* the establishment of RNA role in cellular activities. The focus here will be on the historical outline of the discoveries, while a more detailed description of the mentioned biological functions can be found in the next section 1.2.2.

Background: the Central Dogma of molecular biology The foundational concepts shaping our understanding of molecular biology, including RNA molecules, were established in the 1950s and 1960s with the formalization of the Central Dogma [8]. This framework outlines the flow of genetic information from DNA to proteins, mediated by RNA (Fig. 1.4A). In particular, the elucidation of messenger RNA (mRNA), transfer RNA (tRNA), and ribosomal RNA (rRNA) solidified our understanding of transcription and translation processes [1]: transcription converts DNA into mRNA, while translation employs mRNA as a template for protein synthesis. In this historical context, RNA was primarily perceived as a vital component for protein synthesis, which was considered central to cellular activities. During this protein-centric phase in molecular biology, non-coding transcripts, though recognized, were often dismissed as 'junk' [9]. However, starting from the late 1960s, this knowledge started to have significant changes (Fig. 1.4B).

Phase I: complete understanding of protein synthesis. In this first phase, many studies focused on the complete understanding of protein synthesis and on the characterization of the corresponding role of RNA molecules. Besides rRNA in ribosomes, which are ribonucleoprotein (RNP)

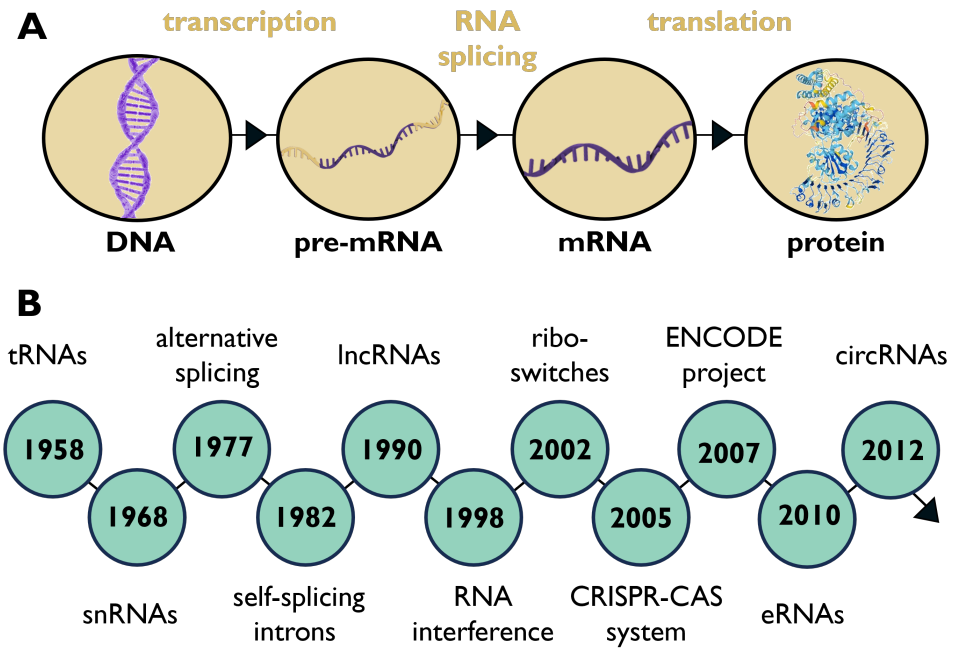


Figure 1.4: Central dogma of molecular biology and non-coding RNA revolution. A) Flowchart of the fundamental steps of protein synthesis: from left to right, the transcription from DNA to pre-mRNA, the RNA-splicing from pre-mRNA to mRNA, and the translation from mRNA to protein. **B** Timeline of the key discoveries in the context of the non-coding RNA revolution (Sec. 1.2.1)

complexes known from mid 1950s [10], the first non-coding RNA (ncRNA) to be discovered was tRNA in 1958 by Zamecnik's group. Following the discovery of novel RNA components, identified as small nuclear RNAs (snRNAs), in the late 1960s [11] groundbreaking investigations led by Sharp's [12], Roberts' [13], and Steitz's [14] groups elucidated the mechanism of RNA splicing mediated by the spliceosome enzymatic complex, composed by snRNAs and proteins. Additionally, the subsequent discovery of alternative splicing in 1977 [15] by Roberts and Sharp challenged the prevailing notion that the genetic message is definitively established during RNA synthesis, earning them the 1993 Nobel Prize in Physiology or Medicine. Concurrently, the Cech's group's discovery of self-splicing RNA molecules in the *Tetrahymena thermophila* ribosomal RNA gene showcased the ability of RNA to catalyze its own splicing process [16]. The conventional belief that catalytic functions were exclusively carried out by proteins was challenged, leading to the introduction of the term 'ribozyme' [16]. For this discovery, Cech was acknowledged with the 1989 Nobel Prize in Chemistry.

Phase II: discovery of regulative RNAs. The 1990s are characterized by the first establishments of the role of RNA beyond its involvement in protein synthesis. In particular, it became evident that RNA is an adaptive regulator of gene expression, by orchestrating cellular responses to external stimuli. In 1998, the groups led by Mello and Fire reported the targeted degradation of *C. elegans* mRNA triggered by a non-coding double-stranded RNA molecule [17]. This discovery laid the foundation of the biological pathway now referred to as "RNA interference" and earned Mello and Fire the Nobel Prize in Physiology or Medicine in 2006. In the late 1990s, investigations into bacterial systems revealed other surprising phenomena involving RNA. First, Oppenheim's group revealed the existence of non-coding RNA molecules able to modulate gene expression in response to

temperature changes [18, 19]. In 2002, the R. R. Breaker's group finally characterized riboswitches, which also are able to modulate gene expression upon the binding with a cognate metabolite.

Phase III: establishment of RNA role in cellular activities The early years of the 21st century marked the definitive recognition of the central role of RNA in cellular processes. Genomic studies conducted in the first decade by the ENCODE consortium unveiled a dynamic transcription of the majority of the animal and plant genomes into long RNAs with limited or no protein-coding potential [20, 21]. This revelation led to the categorization of long non-coding RNAs (lncRNAs), previously discovered and classified as transcriptional noise, into a distinct and functionally diverse class of RNAs [22]. In 2007, studies by the Mojica's [23] and Barrangou's [24] groups on the CRISPR mechanism highlighted the role of RNA as a guide for an adaptive immune system defending bacteria against foreign nucleic acids. By engineering the synthesis of the guide RNA in this immune system [25], Charpentier and Doudna revolutionized the field of genome editing and won the Nobel Prize in Chemistry in 2020. Over the last decade, advancements in RNA biology have revealed new non-coding RNAs. A first important example is constituted by enhancer RNAs (eRNAs), which were discovered in the early 2010s to have a crucial role in epigenetic regulation [26]. Concurrently, the discovery of circular RNAs (circRNAs) [27] introduced a novel class of single-stranded RNAs forming closed continuous loops. Despite their prevalence, the precise functional roles of these circRNAs remain largely unknown.

The presented revolution, still ongoing, fundamentally changed the previous perceptions about RNA, which is now recognized as uniquely able not only to store genetic information, like DNA, but also to catalyze chemical reactions and perform cellular functions, similar to proteins. Furthermore, recent studies suggested that life on Earth may have initially emerged through self-replicating RNA molecules, substantially shifting the paradigms of a protein-center world to an "RNA world" [28]. Looking ahead, the pace of RNA research shows no signs of slowing. Given the ongoing advancements in technology and methods of investigation, new discoveries are likely to shed further light on the landscape of RNA biology in the upcoming years.

1.2.2 The wide range of non-coding RNAs functions

Classifying the diverse functions of ncRNAs is a complex task, which goes beyond the scope of this thesis. However, it is feasible to categorize them broadly as "housekeeping" and adaptive regulatory RNAs [29]. This section provides an overview of RNA classes falling into these categories by illustrating their biological functions through pertinent examples. Except where explicitly indicated, the discussion is inspired by the textbook "Molecular Biology of the Cell" by B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter [1].

"House-keeping" RNAs

The class of "house-keeping" RNAs all the RNAs that are essential for the survival and day-to-day operations of the cell. They provide fundamental mechanisms that support protein expression, ensuring the consistent functioning of the cellular machinery. For the purposes of my research, it is interesting to point out the RNAs involved in the functions that are presented in the following

paragraphs.

Carrying out protein translation. rRNA and tRNA are essential actors in the protein translation process. rRNA forms the core scaffold of ribosomes, the cellular machinery responsible for translation, and provides a structural framework for the accurate decoding of mRNAs. Conversely, tRNA acts as a molecular intermediary, establishing a connection between mRNA coding regions ("codons") and their corresponding amino acids. This function is executed through distinct regions on different tRNAs, each specific to an amino acid, which contain sequences complementary to the codon ("anti-codons"). These anti-codons ensure the accurate incorporation of amino acids into the developing protein chain.

Catalyzing RNA splicing. RNA splicing takes place after RNA transcription. This process consists in the excision and reordering of distinct and distant coding regions of mRNA, known as exons, into the final mature mRNA that is ready for translation into a functional protein. The non-coding regions, known as introns, are either transformed into other functional forms of RNA or degraded. In the majority of organisms, the RNA splicing process is catalyzed by the RNP complex known as spliceosome, which consists of the assembly of multiple proteins and snRNAs [11]. snRNA molecules play a crucial role since they contribute to the spliceosome catalytic core, engaging base-pairing interactions with the pre-mRNA and guiding the precise excision of introns and ligation of exons. In other organisms, pre-mRNA is able to catalyze its own splicing [16]. In this case, the pre-mRNA folds into a complex secondary structure, allowing them to catalyze their splicing through a mechanism similar to the one of spliceosomes. The catalytic ability of intronic RNA is essentially dependent on the metal ions interacting with the RNA backbone chain [30]. While the presented mechanisms are producing a single mRNA isoform from a given pre-mRNA, the same gene can encode multiple mRNA variants [15]. This mechanism, known as alternative splicing, is accomplished by varying the exons that are included or excluded during the splicing event, yielding a diverse array of mRNA products.

Promoting viral life-cycle. Viruses are parasite biological entities that can not sustain life by themselves and need to infect living organism cells to survive. In many viruses, the genomic material is constituted by single strands of RNA that encode the information required for their replication in the host organism [31]. A representative example of a "house-keeping" viral RNA is constituted by Human Immunodeficiency Virus-1 Trans-Active Response (HIV-1 TAR), a critical element in the life cycle of the HIV-1 virus [32]. HIV-1 TAR forms a stem-loop structure that interacts with the viral Tat (Trans-Activator of Transcription) protein (Sec. 1.4.2). The TAR-Tat complex enhances the transcription of the viral genome, promoting efficient viral replication.

Adaptive regulatory RNAs

The class of adaptive regulatory RNAs comprehends all the RNAs that regulate various cellular processes. An important and distinctive feature of these RNAs is that they are able to modulate their activities based on external conditions and inputs. For the purposes of my research, it is interesting to point out the RNAs involved in the functions that are presented in the following paragraphs.

Modulating gene expression by small RNAs. A first important regulatory mechanism involving RNA is the RNA interference (RNAi) pathway, discovered in bacteria as part of their defense mechanisms against foreign genetic elements [33]. This process functions as a cellular regulatory mechanism, modulating gene expression by specifically silencing targeted mRNA molecules. Exogenous double-stranded RNAs of approximately 100 nucleotides are recognized by a family of enzymes known as Dicer. Dicer cleaves these long RNAs into smaller RNA molecules of about 10 nucleotides, whose sequences complement those of the target mRNA. The precise recognition and processing of precursor molecules by Dicer are highly conformation-dependent. The generated small-interfering RNAs (siRNAs) integrate into a ribonucleoprotein known as the RNA-Induced Silencing Complex (RISC). Within the RISC, proteins of the Argonaute family guide the small RNA along the target mRNA, leveraging its complementary sequence to discern and specifically bind to the corresponding mRNA targets. The recognition of the precursor interfering RNAs and the Dicer complex, and thus the subsequent expression, is regulated by a variety of processes. In the case of siRNAs, this leads to the degradation of the target mRNA by cleavage. Additionally, other RNAi pathways have been characterized, involving two other classes of interfering RNAs. The first one concerns microRNAs (miRNAs), which are processed from endogenous precursors, and generally cause translational repression rather than mRNA degradation. The second class concerns piwi-interacting RNAs (piRNAs), found primarily in the germline, and silencing transposable elements in order to keep the integrity of the genome during germ cell development and reproduction. Interestingly, circRNAs have been found to act as miRNA sponges, sequestering and inhibiting the activity of miRNAs.

Modulating gene expression by riboswitches. Another surprising phenomenon in the context of RNA regulative roles consists of RNA molecules found in bacteria that are able to up- or down-regulate the expression of certain genes. A first example of such "RNA switches" is constituted by RNA thermometers [34]: these molecules often adopt at low temperatures a hairpin structure that impedes the ribosome from accessing the start codon, effectively blocking translation initiation. As the temperature rises, the RNA structure unfolds, allowing the ribosome to bind to the mRNA and initiate translation. A second example is constituted by riboswitches [35]. These RNA molecules regulate the expression of the downstream gene undergoing a major conformational change, generally upon binding with high specificity of a cognate metabolite. Both mechanisms enable a rapid and precise cellular response to environmental changes, crucial for bacteria that need to adapt quickly to the environment of a host organism.

Protecting from foreign invaders Bacteria developed an adaptive immune system to defend against invading genetic elements like phages and plasmids [25]. This process takes place in a specific genomic *locus*, the Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) array. A segment of the invading DNA is cleaved by CRISPR-associated (Cas) proteins, processed and integrated into the CRISPR regions of the bacterial genome, which are characterized by a unique arrangement of short, repetitive DNA sequences, known as "repeats", and separated by the DNA fragments of all the past invaders, called "spacers". In this way, the bacterial genome is continuously updated, effectively storing information about past infections and representing an example of "immunologic memory". The foreign DNA elements at the CRISPR locus are transcribed and processed in a set of CRISPR RNAs (crRNAs), individually containing a single spacer sequence.

These crRNAs are loaded in specific RNP complexes together with Cas proteins (similar to the RISC complexes in RNAi) and used as guides to recognize, cleave, and finally neutralize the target invader molecule.

Long non-coding RNAs

lncRNAs are RNAs longer than 200 nucleotides and without protein-coding potential [36]. These molecules were generally considered transcription errors. However, the increasing evidence of their prevalence in living organisms has revealed an expanding repertoire of lncRNAs, showcasing their diverse functionalities, often at the boundary between "house-keeping" and regulative ones. From serving as molecular scaffolds, modulating chromatin architecture, and acting as sponge decoys for other molecules, to functioning as enhancers and cellular signals, lncRNAs represent a heterogeneous class of RNA molecules. A thorough characterization of lncRNAs is yet to be achieved and a comprehensive overview goes beyond the purposes of this section. In the following paragraphs, I will discuss three important examples of lncRNAs.

MALAT1: a director of RNA splicing. An important and representative example of lncRNAs is constituted by Metastasis-Associated Lung Adenocarcinoma Transcript 1 (MALAT1), which is ubiquitously expressed across various tissues [37]. Initially known for its involvement in metastatic lung cancer, subsequent studies have revealed that MALAT1 plays multifaceted and intricate roles in cellular physiology, particularly in the modulation of gene expression and the regulation of alternative splicing. MALAT1 is localized in nuclear speckles, which are interchromatin domains enriched in pre-mRNA splicing factors. Within these regions, MALAT1 modulates their distribution and activity, effectively orchestrating the splicing patterns of pre-mRNA targets. Able to interact with several different partners thanks to its structural flexibility, MALAT1 ensures timely and precise alterations in splicing, catering to the cell's specific needs.

XIST: grant for female survival. A second important and representative example of lncRNA is constituted by the X-inactive specific transcript (XIST) [38]. This molecule plays a pivotal role in the process of X-chromosome inactivation in female mammals. This inactivation is essential to ensure compensation between male (XY) and female (XX) genes, as it is induced by the transcriptional silencing of one of the two X chromosomes in female cells. Upon initiation of the inactivation process, XIST envelops the entirety of the X-chromosome, acting as a scaffold for diverse chromatin-modifying complexes. This orchestration results in alterations to the chromatin structure, effectively silencing gene expression across the entire chromosome. The discovery and study of XIST have greatly expanded our understanding of lncRNAs and their diverse roles in cellular processes and gene regulation.

TERRA: modulator of homeostasis. Telomeric Repeat-containing RNA (TERRA) is a long non-coding RNA found near the telomeres, the protective caps at the ends of eukaryotic chromosomes [39]. Unlike conventional RNA molecules, TERRA is unique in that it is transcribed from the telomeric DNA, and engages in specific interactions with telomeric proteins to maintain the heterochromatic state of telomeres. This is crucial for telomere maintenance: it protects DNA from

deterioration or fusion with neighboring chromosomes, contributing to the overall genomic stability along aging and correct cellular homeostasis.

1.3 The emerging therapeutic potential of targeting RNA with small molecules

Along with the discoveries of its cellular functions, RNA has emerged as a promising therapeutic target. Among different therapeutic strategies, targeting RNA with small molecules stands out for their intrinsic pharmacological properties, offering a means to target the different classes of RNAs effectively and safely. Despite this potential, the field lacks an established framework. In particular, a major obstacle is constituted by the limited biophysical characterization of RNA interactions with small molecules. In this sense, while drug design has traditionally drawn extensively from a century-long focus on protein targeting, understanding the unique physicochemical attributes of RNA binders and the structural properties of target RNAs is crucial.
This section is first dedicated to providing an overview of important examples of diseases caused by RNA dysfunctions. Then, I will finally introduce the therapeutic approaches that have been developed to target the diverse classes of RNA molecules involved in diseases. In this context, I will highlight why small molecules targeting is one of the most promising and effective strategies and discuss the general principles of this framework.

1.3.1 Pathological mechanism linked to RNAs

The range of functions carried out by RNA molecules is wide. As a consequence, the possible dysregulation of RNA functions constitutes a primary driver of serious diseases like cancer, neurodegenerative diseases, and metabolic disorders [40–42]. Moreover, RNA plays a fundamental role in both viral and bacterial infections. In the following paragraphs, I will highlight important examples of RNAs that constitute important therapeutic targets. The reported examples refer to the RNAs that have been introduced in the previous Sec. 1.2.2.

mRNAs. Pathological conditions associated with proteins often originate from their mRNA blueprints, where genetic mutations, post-transcriptional modifications, and regulatory issues encode the aberrant protein behaviors. Spinal Muscular Atrophy (SMA) is an example of a disease driven by a splicing error [43]. In this neurodegenerative disorder, the aberrant exclusion of exon 7 of the Survival Motor Neuron 2 (SMN2) gene results in the insufficient production of the SMN protein, essential for motor neuron survival.

miRNAs. miRNAs are essential in controlling protein synthesis via the RNAi pathway and sustaining cellular equilibrium. Deviations in their expression levels, whether by overexpression or underexpression, are associated with numerous diseases. For instance, miRNA-96 naturally regulates the translation of the FOXO1 protein, a key factor in cell cycle regulation and apoptosis [44]. In cancer, observed overexpression of miRNA-96 leads to the aberrant suppression of FOXO1, thus contributing to tumor progression and growth.

lncRNAs. The dysregulation of lncRNAs can favor the formation of cancers, and drive neurological as well as cardiovascular diseases [36]. For instance, MALAT1 dysfunction is particularly implicated in cancer [37]. The modulation of alternative splicing by MALAT1 is critical for cellular functions, and when disrupted, MALAT1 altered splicing and gene expression can significantly contribute to cancer progression and metastasis. The dysregulation of TERRA lncRNAs function of telomere maintenance may cause the instability of the genome and its deterioration, resulting in the primary driver for tumor-promoting mutations [45]. Moreover, the aberrant behavior of XIST lncRNA and the incorrect inactivation of the X chromosome have been recognized as the primary driver of Alzheimer's disease [46].

Bacterial and viral RNAs. Diseases may be caused by the proliferation of bacterial and viral infections. The correct functioning of the cellular agents in these pathogens constitutes their primary driver. In bacteria, an important example is constituted by riboswitches, such as the Flavin MonoNucleotide (FMN) riboswitch [47], which is able to regulate their metabolism and survival by modulating gene expression. For what concern viral pathogens, several unique RNA molecules are critical for the life cycle of the virus. An important example is constituted by the HIV-1 TAR RNA element, whose interaction with the Tat protein is essential to promote viral transcription and ultimately HIV replication [48].

RNA tandem repeats. In addition to the aforementioned RNA classes, RNA transcripts whose sequence is characterized by the presence of multiple repeated patterns have been associated with the development of diseases [49]. An important example is the r(CUG) expansion, associated with Myotonic Dystrophy Type 1 (DM1) [50]. In DM1, the expanded r(CUG) repeats in the DMPK gene result in abnormally long RNA sequences. These expanded RNAs accumulate in the nucleus, sequestering RNA-binding proteins and altering the normal splicing of various pre-mRNAs. This leads to the diverse and systemic symptoms of DM1, which include muscle wasting and myotonia.

1.3.2 RNA-targeted therapeutics on the rise

Alongside the recognition of the role of RNA molecules in several diseases, the interest in RNA-targeted therapeutics has significantly increased. The exploration of RNA as a therapeutic target holds great potential for developing effective strategies. While many approaches have historically focused on the non-coding RNA revolution, often linked to the discovery of specific classes or biological pathways of ncRNAs, there has been a recent resurgence of interest in directly targeting RNA with small molecule drugs.
This section first explores the therapeutic opportunities arising from the targeting of disease-related RNAs. Subsequently, I will outline the key criteria to assess the relevance and success of the therapeutic approaches developed so far to target RNA. This discussion serves as a concise introduction to the more in-depth exploration of small molecules targeting in Section 1.3.3.

The advantages of targeting RNA

Expanding the class of biomolecular targets to RNA molecules holds multiple promising advantages. On one side, directly targeting mRNA elements emerges as a potentially equal or more effective

strategy than targeting the corresponding expressed protein. Indeed, this approach would allow altering the expression of proteins at the transcript level, potentially addressing challenges related to proteins that are difficult to target [51]. On the other side, the shift toward RNA-modulating agents would pave the way to novel therapeutic opportunities, unavailable by the sole targeting of proteins [52]. In the context of major genomics studies conducted at the beginning of this century, it has been established that only the $\sim 1.5\%$ of the human genome encodes proteins [20, 53] (left panel, Fig. 1.5). Moreover, among this small fraction, $\sim 10 - 15\%$ is thought to be disease-related (right panel, Fig. 1.5) [54]. From a quantitative perspective, current protein-targeted drugs interact with fewer than 700 gene products, meaning that only the $\sim 0.05\%$ of the human genome has been drugged [55]. Conversely, a substantial fraction (approximately 70%) of the human genome is transcribed into non-coding RNAs (left panel, 1.5). Several classes of these non-coding RNAs represent validated therapeutic targets (Sec. 1.3.1), including around 15000 lncRNA transcripts that are still poorly characterized [53]. In summary, if it was possible to target just a fraction of the tens of thousands of mRNAs and non-coding RNAs, the extent of the druggable human genome could increase substantially [52].

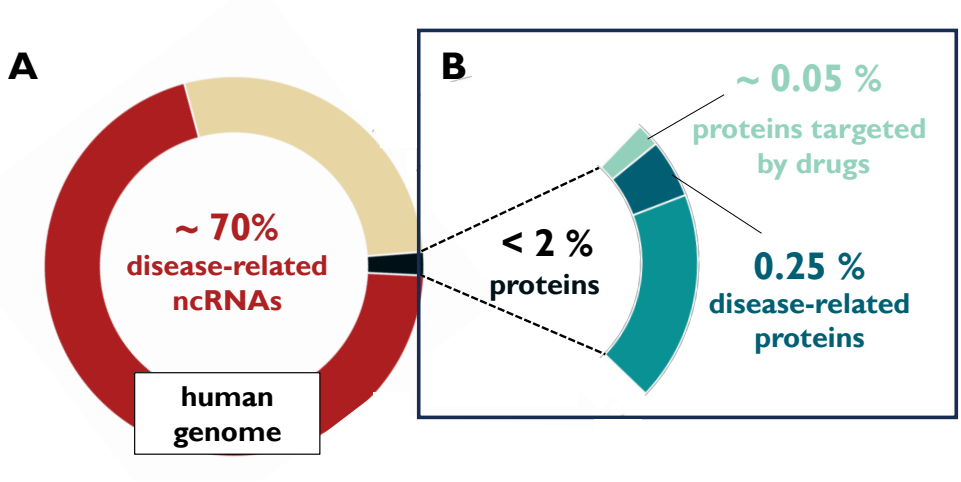


Figure 1.5: Composition of the human genome. On the left, a pie chart reporting the relative abundance of non-coding RNA transcripts that may potentially be targets for therapeutic intervention, and transcripts coding proteins (black). On the right, zoom in on the portion encoding proteins colored with different shades of blue. All reported percentages refer to the pie chart on the left.

Relevant approaches to target RNA molecules

Due to the potential of targeting RNA, a variety of therapeutic approaches have been developed in the last two decades [40–42]. To facilitate a comprehensive discussion and comparison of the different strategies, it is useful to review the pharmacological criteria that must be satisfied to be used on humans. From a broad perspective, the essential pharmacological properties for a therapeutic agent can be summarized by:

- the **efficacy** or **specificity**, that is the capability of a therapeutic agent to effectively reduce or completely suppress a specific pathological behavior, remaining stable and active under physiological conditions for the time necessary to exert their therapeutic effect;

- the **selectivity,** that is the capability of a therapeutic agent to exert a biological function only on precise targets, which may be a biomolecule or a biological pathway, reducing at the minimum possible off-targets effects;
- the **safety**, which is the capability of a therapeutic agent to not trigger adverse side effects on the human organism, like toxicity or unfavorable immune responses;
- the **bioavailability**, which is the capability of a therapeutic agent to be readily absorbed, distributed, metabolized, and excreted by the human body.

The physicochemical properties related to bioavailability concern the behavior of a therapeutic agent within the organism and are referred to as pharmacokinetic properties. The three other aspects concern the interaction between the therapeutic agent and the organism as well as its response, and they are referred to as pharmacodynamics properties. Regulatory agencies, like the European Medicine Agency (EMA) or the U.S. Food and Drug Administration (FDA), play a crucial role in ensuring that adequate clinical trial phases are rigorously completed before therapeutic agents are authorized for market distribution.
In the following paragraphs, I will introduce the different therapeutic strategies that have been developed to target RNA molecules and highlight their advantages and drawbacks with respect to the presented properties.

Antisense technologies. Currently, the majority of marketed RNA-targeted drugs are either single-stranded antisense oligonucleotides (ASOs) or double-stranded short interfering RNAs (siRNAs) [56]. ASOs are short single-strand nucleotides synthesized to bind to target mRNA by leveraging the high specificity and stability of base pairing interactions. ASOs can alter RNA splicing, stability, and translation, therefore offering a route to influence genetic pathways implicated in various diseases [57]. However, despite its potential, this technique suffers from significant limitations [58]. While ASOs are designed for high selectivity, their effectiveness can be limited by challenges in cellular uptake and distribution, primarily due to their susceptibility to enzymatic degradation. Such degradation can undermine their selectivity, leading to off-target effects and potential toxicity. In a different approach, double-stranded siRNAs function by targeting specific mRNA molecules leveraging the RNA interference mechanism, leading to their degradation and thus silencing the expression of the corresponding gene. Synthetic siRNAs to be processed in the natural RNAi pathway offer potential therapeutic applications in a wide range of diseases [59]. However, the therapeutic use of siRNAs faces notable limitations, particularly in terms of bioavailability and tissue distribution [60]: siRNAs are rapidly cleared by the kidney and exhibit limited tissue distribution, posing a significant challenge in achieving effective therapeutic concentrations in target tissues.

CRISPR-CAS-based approaches. CRISPR-Cas system uses RNA-guided enzymes to precisely alter genomic information. The engineering of this mechanism with the CRISPR-Cas9 system [25] enables in principle the targeting of almost any genomic entity. Due to its ability to introduce corrective mutations and modify genetic elements, CRISP-Cas9 constitutes a novel and suitable tool in RNA-targeted therapeutic development, especially for direct somatic cell editing in patients [61]. Despite its revolutionary potential, *in vivo* applicability of CRISPR-Cas9 therapeutics faces major challenges [62, 63]. Most importantly, the induced genome editing, which is irreversible, includes

off-target effects and thus raises fundamental safety issues. Moreover, the precise and effective delivery of the CRISPR-Cas9 system to specific cells or tissues poses challenges. Immune responses against Cas9, originating from bacterial proteins, can occur in patients, potentially affecting the treatment's efficacy.

Small molecules targeting. An alternative approach to RNA-based therapeutics is the direct targeting of RNA with small molecule drugs [52, 64–67]. These compounds can be approximately defined by the compliance to the classical druggability criteria introduced by Lipinski [68]: small molecules should not be too heavy, too polar, too hydrophilic, and too hydrophobic. As a consequence of their physicochemical properties, small molecule compounds can efficiently cross biological membranes, maintain sufficient solubility for absorption, and avoid rapid metabolic degradation or excretion. Furthermore, small molecules are ideal for selective target recognition due to their precise molecular size and structure, enabling specific interaction with biological targets that can be optimized during the process of drug discovery (Sec. 1.5.1). Most importantly, except for the CRISPR-Cas9 technique, whose *in vivo* clinical applicability is currently limited, other RNA-targeted therapeutics are tailored primarily to target mRNA. In contrast, small molecules inherently have the capacity to interact with diverse RNA targets, thereby taking full advantage of the expanded range of targetable genomic elements, including non-coding RNAs (ncRNAs). Given these considerations, the utilization of small molecules emerges as a highly promising approach for RNA targeting that I will discuss more in-depth in the next section.

1.3.3 Targeting RNA with small molecules

Small molecules that target proteins have played a crucial role in advancing medicine throughout the past century [55, 69]. The extensive utilization of biophysical assays and the widespread availability of protein-small molecule structures have contributed to a comprehensive understanding of their interactions [70–72]. However, RNA molecules differ significantly from protein targets due to their unique physicochemical characteristics (Sec. 1.1). The interactions between RNA and small molecules are not as extensively characterized, leading to ongoing debates regarding the extent to which the knowledge derived from protein targeting experiences can be applied to target RNA [52, 64, 66, 67, 73–75]. As a result, the promising framework of RNA targeting is slowed down and far from comparable results. Despite the identification of both successful and promising compounds, the scientific community is struggling to establish general principles of targeting RNA with drug-like compounds, with a collection of *ad-hoc* approaches built on the peculiar characteristics of each RNA molecule [56, 65]. In the following paragraphs, I will summarize the key general principles of targeting RNA with small molecules that have emerged so far.

The chemical space of RNA-targeted small molecules. Over the recent years, concurrent with the increasing interest in small molecules targeting RNA, there has been a gathering of information on interactions between RNA and small molecules. This has paved the way for an initial understanding of the characteristics of their molecular recognition and the chemical space spanned by RNA binders with respect to FDA-approved drugs, mostly targeting proteins [52, 70]. An initial effort has been performed by M. Disney and collaborators, who developed a small molecule library of compounds targeting r(CUG)-repeat sequences and with verified biological activity [76]. The

analysis of this library revealed that the polarity and hydrogen-bonding properties of such binders were significantly different from FDA-approved drugs and, more generally, from protein-oriented libraries. Successively, Hargrove's group implemented one of the first comprehensive repositories of RNA-ligand interactions, the R-BIND database (Sec 1.5.6), which collects RNA-targeted ligands with demonstrated biological activity [77, 78]. The analysis of R-BIND compounds confirmed that most bioactive RNA ligands differ from FDA-approved drugs in terms of relevant physicochemical properties: in addition to their pronounced polarity, these ligands were characterized by the significant presence of planar nitrogenous and aromatic rings as well as by linear regions distributed along a single axis. This common rod-like shape of RNA binders has been supported by Schneekloth's group in a systematic analysis of experimentally validated ligands [79]. Hargrove's group also carried out a systematic comparison of the interactions driving the recognition of the RNA- and protein-small molecules complexes deposited in the PDB that supported the mentioned findings: small molecules are mostly recognized by RNA through different interaction mechanisms and, in particular, intercalation into the stacking base pairs and hydrogen bonding [74]. The insights gained from the examination of the recently established ROBIN database by Schneekloth's group (Sec. 1.5.6) have further solidified this understanding [75]. In this work, the connection between the presence of nitrogenous and aromatic rings, and the topological charge has once again been emphasized in relation to the intercalation between nucleobases, which forms the basis for the recognition of most aromatic RNA binders.

The drug-likeness of RNA binders. While significant progress has been made in comprehending the interaction between RNA and ligands, the connection between their physicochemical properties and their suitability to become drugs with a favorable pharmacological profile remains unclear. In particular, the anionic nature of the RNA backbone introduces a fundamental challenge regarding the selectivity of RNA binders: it restricts the number of compatible structures and favors interactions with positively charged species [65, 80]. Moreover, a complex three-dimensionality, difficult to achieve for planar and/or linear ligands, is now recognized as likely resulting in better pharmacological profiles [80, 81]. Early drug discovery efforts mostly identified aminoglycosides antibiotics that are highly positively charged compounds with high affinity, but often bad pharmacological properties [82, 83]. Due to these difficulties, for a long time, most RNA molecules were not perceived as suitable targets for therapeutic intervention. Recently, the mentioned developments in the understanding of RNA-small molecule interactions led to the discovery of therapeutic agents that comply with the traditional rules of medicinal chemistry. Three relevant examples that are worth to mention are (Fig. 1.6):

- **linezolid** (Fig. 1.6A), the first FDA-approved antibiotic of the class of oxazolidinones [84], which inhibit bacterial protein synthesis by binding to the 50S ribosomal subunit, specifically targeting the peptidyl transferase center composed of protein and rRNA elements [85];

- **ribocil** (Fig. 1.6B), a small molecule targeting the FMN riboswitch [86]. By exploiting the same binding pathway of the FMN riboswitch natural partner, ribocil selectively and potently binds the riboswitch target and suppresses bacterial gene expression. Despite its optimal pharmacological profile, ribocil molecule is subject to antibiotic resistance and proposed modifications are still in clinical phases [87];

- **risdiplam** (Fig. 1.6C), a small molecule very recently approved by FDA in the treatment of SMA [88]. By selectively binding the SMN2 splicing site, risdiplam promotes the correct splicing and the subsequent expression of SMN2 gene, which is underexpressed in patients affected by SMA.

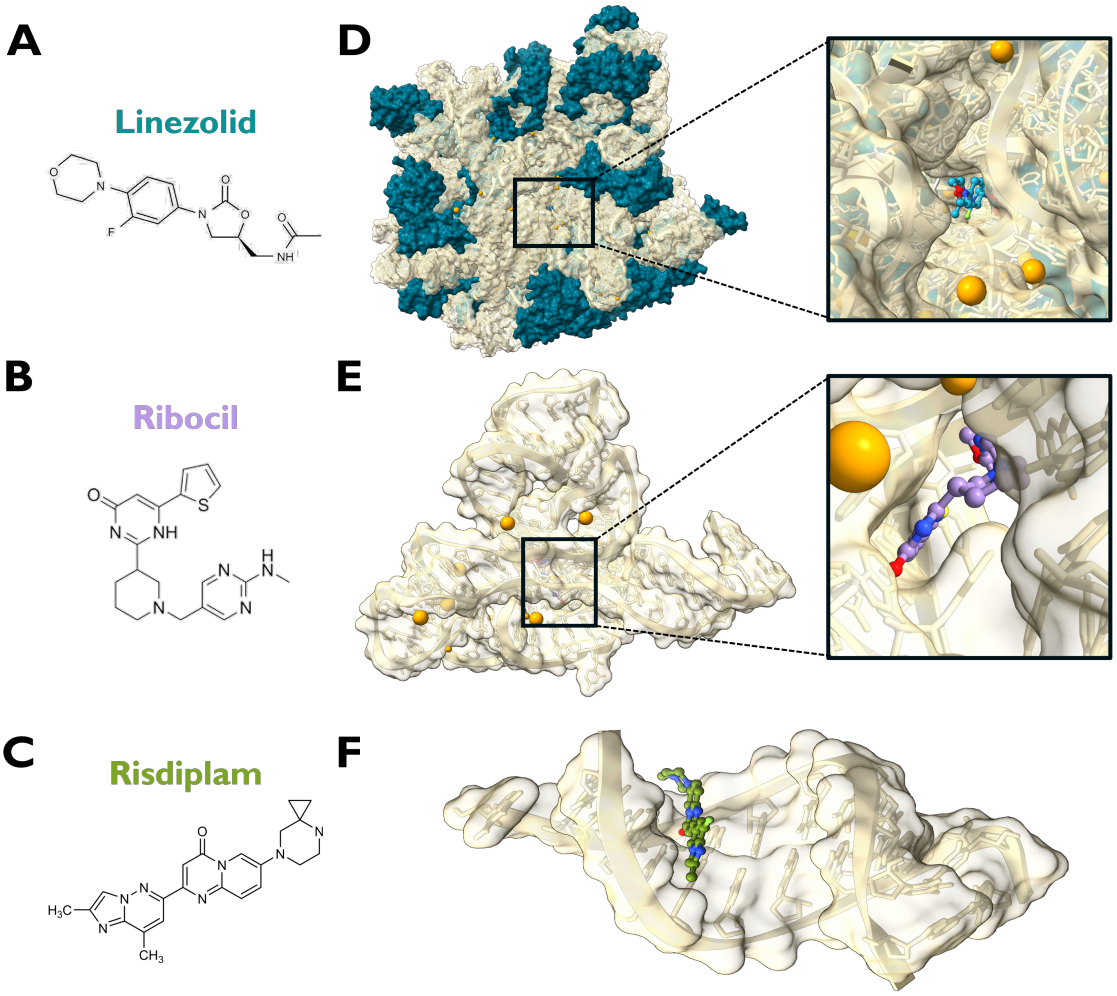


Figure 1.6: Examples of RNA-small molecule interactions. A-C) The 2D structure of three examples of small molecules targeting RNA: linezolid [84] (**A**), ribocil [86] (**B**), and risdiplam [88] (**C**). **D-F)** Molecular images of the RNA targets (khaki ribbon-surface) of the compounds in **A, B, C**: the 50S ribosomal subunit of the prokaryoti *Haloarcula Marismortui* (PDB 3cpw [89], **D**), the FMN riboswitch of the *E. coli* (PDB 5xk9 [90], **E**), and the 5'-splice site of SMN2 exon 7 (PDB 6hmo [91], **F**), respectively. The insets in **D** and **E** zoom in the binding site region. Proteins in **D** are shown as dark blue molecular surfaces. The ligands resolved in **E** and **F** do not correspond to the ones displayed in **B** and **C** due to the unavailability of resolved structures, but the binding pocket is analogous.

The presented examples showcase the feasibility of developing effective therapeutic strategies targeting RNA. However, most RNA binders that are in clinical or pre-clinical phases present deficiencies from the point of view of conventional medicinal chemistry [52]. In this regard, it is important to mention that the nature of drug-like compounds is highly debated in the literature and that the classic definition of drug-likeness by Lipinski may not comprehensively include all the potential therapeutic agents [92–94]. In particular, it is possible that RNA-binding small molecules will have unique properties eventually falling outside Lipinski's rule of 5 and to hypothesize that they may act as therapeutics agents [65, 95]. A comprehensive discussion on this point goes beyond

the scope of this Introduction, but it is worth mentioning some examples that are already reported in the literature. In the mentioned R-BIND database, ligands composed of multiple RNA-binding cores connected by linker regions were reported as bioactive RNA binders, even though heavier than traditional small molecules [77, 78]. Another important class of binders is constituted by Ribonuclease-Targeting Chimeras (RIBOTAC), large compounds able to specifically degrade target RNAs [96]. Finally, drugs composed of metal atoms, which showed great potential in the selective binding of DNA nucleic acids despite possible toxic effects, are gaining momentum in RNA-targeted applications [97].

The binding pockets of RNA targets. The physicochemical nature of RNA binders reflects the characteristics of the RNA receptor molecule. In this sense, a second major element to facilitate the design of compound targeting RNA is understanding the structural properties of RNA binding pockets, highlighting the similarities and the differences with respect to protein pockets. In general, the current evidence shows that RNA needs to fold into conformations with enough structural complexity to form buried pockets that can engage specific and high-affinity interactions with small-molecule compounds [52, 98]. This is indeed the case of linezolid and ribocil, whose binding pockets are deeply buried into the target RNA molecule, resembling a typical hydrophobic protein pocket (Fig. 1.6DE). However, both cases present unique features that may not broadly apply to the targeting of RNA [52]: linezolid binds rRNA, which is the most abundant RNA in cells and therefore requires achieving a modest binding affinity, while ribocil binds in the same pocket of the natural metabolite of the FMN riboswitch, which is predisposed to small molecule binding. Given the highly electronegative and limited buried surface of most RNAs, many potential targets may form shallow cavities or present a lower structural complexity [65, 80]. In this regard, a recent computational analysis of the RNA-ligand structures deposited in the PDB database by Schneekloth and collaborators suggested that RNA binding pockets are much less hydrophobic than protein ones [79]. The most important example of an exposed cavity can be found in the binding site of risdiplam, which is located in the splice site of SMN2 pre-mRNA within a superficial cleft formed between the helical domain (Fig. 1.6F). As observed in the latter work and confirmed by the R-BIND analysis, the nature of such cavities may accommodate rod-like ligands commonly found to bind RNA. [77]. However, mostly due to the limited availability of RNA-small molecule structures, their structural pattern of recognition remains unexplored

To summarize, while recent years experienced an increase in the relevance of RNA-targeted drug discovery, the development of compounds that are able to selectively and specifically bind RNA targets remains challenging. In view of understanding and characterizing the unique properties of RNA molecules and of the interactions they engage with small molecules, one key element is still often neglected. Indeed, unlike most protein targets, a significant portion of potential RNA targets, such as mRNAs and viral RNAs, is highly flexible and does not assume a single static structure in the cellular environment. A comprehensive evaluation of the possible interactions that one compound can engage with a given RNA target can not overlook the possibility that these interactions depend on its structural dynamics. In the direction of developing effective therapeutic strategies to target RNAs, the next section will explore the principles of RNA structural dynamics and the methods employed in its characterization.

1.4 The dynamic and elusive nature of RNA targets

RNA molecules are dynamic entities that adopt a variety of interconverting conformations in solution. Their intrinsic flexibility constitutes at the same time a key element in performing cellular functions and a significant challenge for their biophysical characterization. Theoretically, RNA flexibility is best described using the thermodynamic model of the free-energy landscape, which accounts for the hierarchical nature of motions and interactions influencing RNA structural dynamics. Practically, determining accurate RNA conformational ensembles is difficult, despite the tremendous advancements of experimental techniques to study its structural dynamics. Computational approaches open prominent avenues to overcome some of the challenges of RNA structure determination by experimental techniques. In particular, Molecular Dynamics (MD) simulations enable an atomistic description of RNA behavior in solutions and constitute one of the most comprehensive computational methods to capture the elusive dynamics of RNA molecules.
This section is first dedicated to discuss the importance of RNA flexibility for performing its cellular functions. In this context, I will introduce the principles of RNA structural dynamics from a thermodynamic perspective. Then, I will provide an overview of the experimental and computational methods that have been developed to determine RNA structure. Finally, I will discuss the state-of-the-art of MD simulations in describing the structural dynamics of RNA molecules.

1.4.1 The role of structural dynamics in the cellular functions of RNAs

In an aqueous environment, RNA does not maintain a single static structure, but rather dynamically samples a vast array of conformations [99]. The functionality of most RNAs is directly determined by this "conformational propensity" [100]. By referring to the functions that have been elucidated in Section 1.2, I will here discuss some important examples of how RNA functions depend on structural changes taking place at the secondary, tertiary, and quaternary structural levels (Fig. 1.7).

A first example is constituted by the mentioned riboswitches, which modulate gene expression in bacteria [35]. Often upon ligand binding, these molecules are able to transition from a secondary structure to a significantly different one (Fig. 1.7A). The ribosomes are able to distinguish the two conformations and to modulate consequently the bacterial gene expression. In some viral systems, similar behaviors have been characterized. The HIV-1 RNA genome can alter its secondary structure and impact its dimerization, which is necessary for the subsequent translation of viral proteins [101] (Fig. 1.7B). Furthermore, the catalytic role of RNA is largely dependent on RNA structural changes (Fig. 1.7C). In the representative case of self-splicing ribozymes, the effective catalysis of the splicing reactions often involves cycling through various tertiary structures [102]. Conformational changes are also crucial to form RNA quaternary assemblies. RNA-binding proteins often bind to regions of RNA that are single-stranded, as in the case of factors involved in alternative splicing [103]. Such molecular recognition only takes place after the "unwinding" of RNA secondary structure (Fig. 1.7D). The interaction with proteins may also drive the conformational rearrangement of RNA molecules to inhibit their activity. Another important example is given by the LIN28A protein that induces the formation of a conformation of the let-7 pre-microRNA that is not recognized by the Dicer enzyme [104] (Fig. 1.7E).

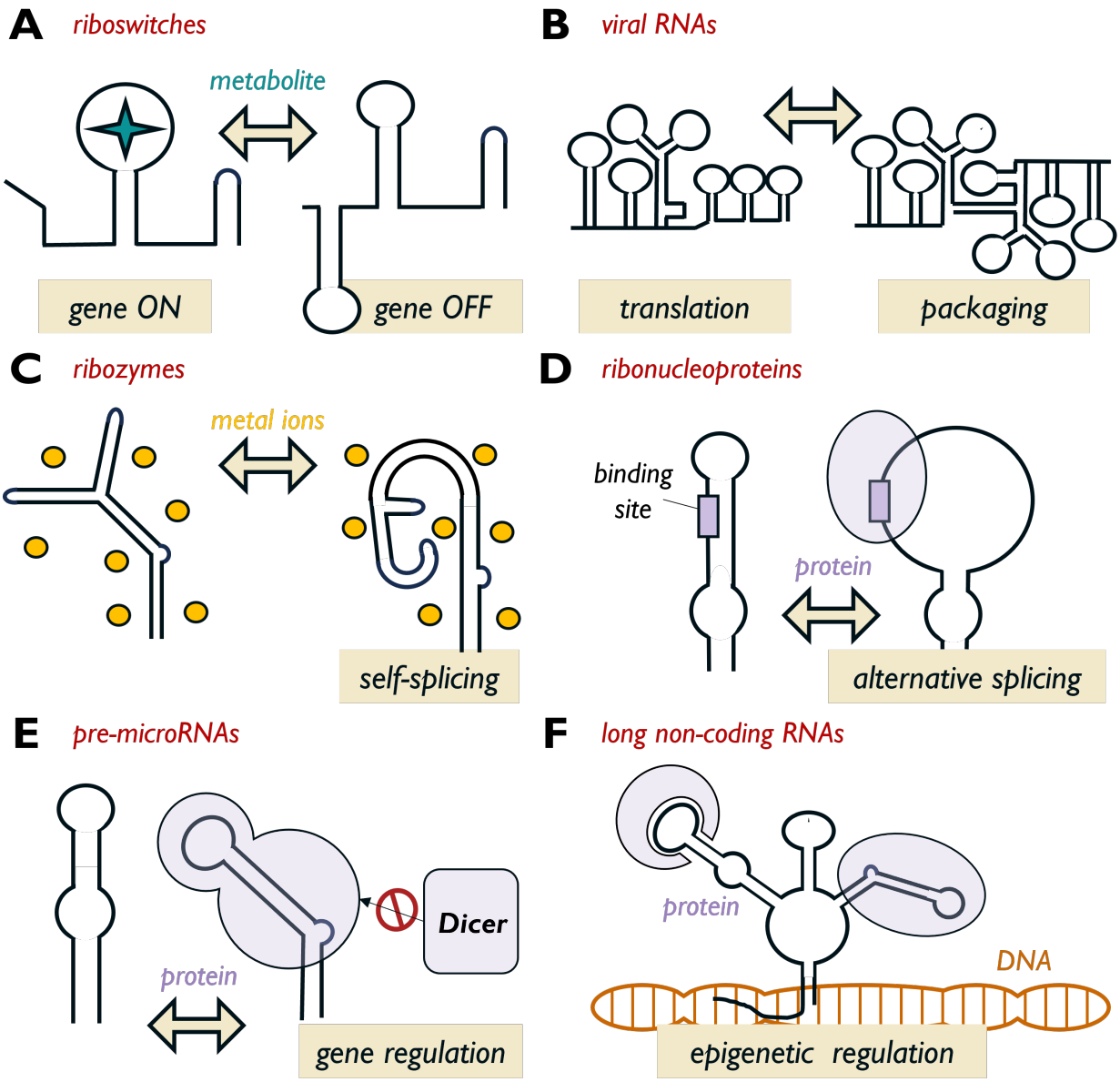


Figure 1.7: Conformational changes in RNA functions. Diagrammatic representations of the conformational changes occurring in several RNA molecules and their link with biological functions. **A)** Gene regulation by a generic riboswitch, upon binding with its cognate metabolite (turquoise star). **B)** Translation or packaging by HIV-1 RNA genome. **C)** Self-splicing of the ribozymes upon interaction with the cationic environment. **D)** Formation of spliceosome complex upon unwinding of double-stranded regions. **E)** Gene regulation in the RNA interference pathway upon binding of protein partners that impede the recognition with the Dicer enzyme. **F)** Epigenetic regulation by HOTAIR lncRNA upon binding with protein partners [105]. The figure has been adapted from Ganser *et al.* [99].

A final example is constituted by lncRNAs. Despite an accurate biophysical characterization of these molecules is lacking, they likely undergo conformational changes when acting as scaffolds for assembling proteins, DNA, and RNA molecules, as in the case of HOTAIR lncRNA [105] (Fig. 1.7F).

The highlighted examples of RNA conformational changes underscore the critical interplay between RNA dynamics and its functional roles: a remarkable diversity of cellular mechanisms is facilitated by the conformational propensity of RNA molecules [100]. To gain a deeper insight into the intricate

nature of RNA structural dynamics, in the next section, I will elucidate some fundamental principles governing their conformational dynamics.

1.4.2 Principles of RNA structural ensembles

From a thermodynamic standpoint, RNA in solution exhibits a dynamic behavior represented by a statistical ensemble of multiple conformations that interconvert over various timescales, spanning from picoseconds to hours [106, 107]. The understanding of RNA structural dynamics necessitates the application of the theoretical concept of free-energy landscapes, which is first elucidated in this section. This framework will enable the establishment of an energetic hierarchy among RNA motions, whose characterization constitutes the core of the discussion.

The RNA conformational space: a thermodynamic description

The formalism of free-energy landscape, first developed to describe complex systems such as glasses and, later, proteins [108], provides a powerful framework for describing RNA dynamic ensembles [109]. The set of all possible conformations can be represented by a continuous free-energy landscape, punctuated by local minima, or basins, that correspond to highly populated states. The population of a given basin is determined by its stability relative to other ones, and the interconversion rate between them depends on the corresponding energetic barriers as well as on the temperature of the system. A transition from one state to another is often a rare event, contingent on the height of the associated energetic barrier and the likelihood of spontaneous occurrence through thermal fluctuations. A representative example to mention is the HIV-1 TAR RNA (Fig. 1.8A, Sec. 1.2.2), whose structural dynamics have been characterized by several studies in recent years [99, 110, 111].

The free-energy landscape of HIV-1 TAR (Fig. 1.8B) has been proposed to be dominated by a single native secondary structure, consisting of two helical regions surrounding a bulge region, along with an apical loop. The 3D orientation of the two helices can shift from a closely stacked and rigid configuration ($\sim 40\%$) to a more bent and flexible arrangement ($\sim 40\%$) on the picosecond-to-microsecond timescale (Fig. 1.8B). Also observed are non-native secondary structures with populations of $\sim 10\%$, $\sim 0.1\%$, and $\sim 0.01\%$ and transition rates on the microsecond-to-millisecond timescale. These states exhibit variations in base pairing patterns within and around the bulge and apical loops. The viral activity of HIV-1 is mainly determined by its molecular recognition with Tat protein (Sec. 1.2.2). This interaction stabilizes the coaxial conformation of HIV-1 TAR, characterized by non-canonical triple base-pairs involving the bulge. From a thermodynamics perspective, the binding with Tat cellular modifier results in the redistribution of the HIV-1 TAR populations along its free-energy landscape (Fig. 1.8C). This redistribution of populations incurs an energetic cost that needs to be compensated [112]. In this example of the binding between two biomolecules, this cost is balanced by the additional formation of favorable intermolecular contacts [99].

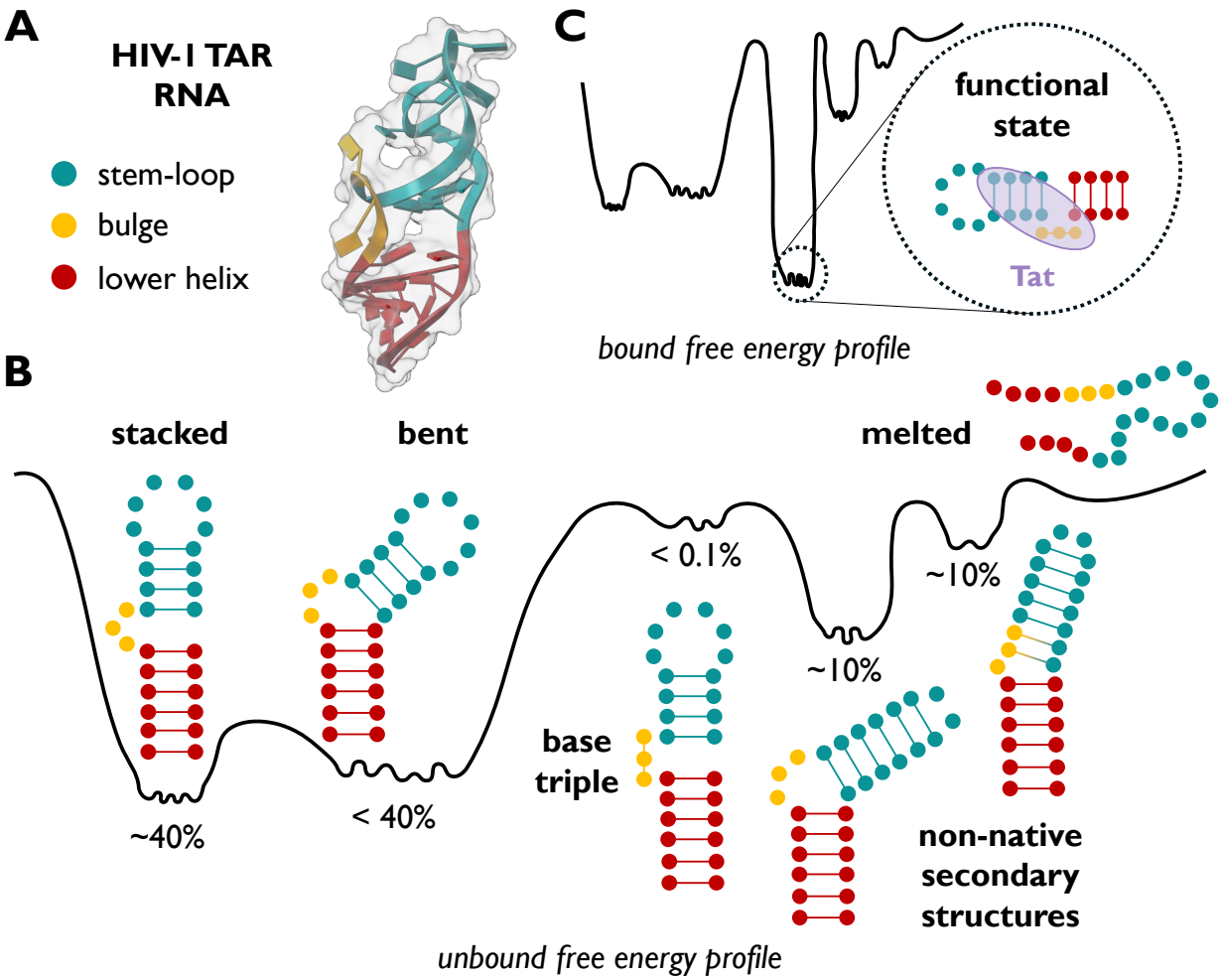


Figure 1.8: The conformational landscape of HIV-1 TAR RNA. A) Cartoon-surface molecular image of the HIV-1 TAR RNA (from PDB 1uts [113]). The different domains of the molecule are color-coded. **B)** The conformational landscape of the unbound HIV-1 TAR RNA as characterized by Ganser *et al.* [99]. The relative energetic stabilities are represented by the depth of the free-energy basins, annotated with their corresponding populations. **C)** The free-energy profile of the HIV-1 TAR RNA bound to the Tat viral protein. In the circular inset, a diagrammatic representation of the HIV-1 TAR-Tat complex, which corresponds to the viral functional state.

Due to its inherent flexibility, a thermodynamic ensemble representation of RNA proves essential. Such a representation not only reflects the true nature of the molecule but also aids in describing and accurately predicting key steps in processes that involve conformational changes, such as the binding with other biomolecules. However, characterizing the rugged free energy landscape of RNA molecules with its stable as well as metastable states poses a major challenge. To this end, understanding the structural and energetic variations across different timescales of RNA motions is crucial. In the upcoming section, I will discuss these variations, which result in a hierarchical organization of RNA dynamics.

Hierarchical timescales of RNA motion

The local minima of RNA free energy landscape are hierarchically organized in three different tiers corresponding to higher or lower barriers (Fig. 1.9), resulting in shorter or longer conversion rates,

respectively [114–116]. In the following paragraphs, I will overview the different motions that RNA experiences in these tiers, providing examples of their functional implications.

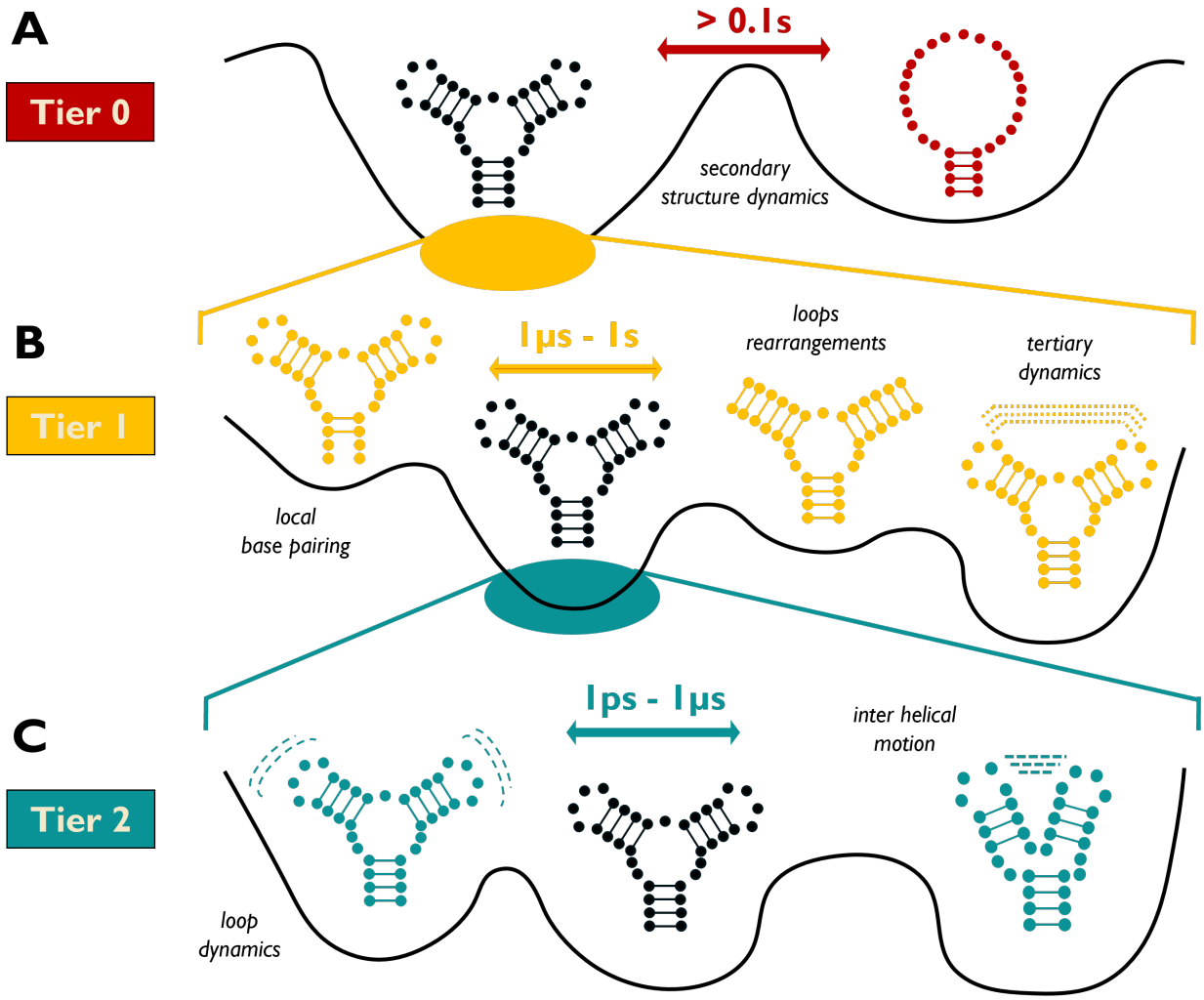


Figure 1.9: The tiers of RNA structural dynamics. A-C) The free-energy profile of a generic RNA molecule (black diagram) across the three different tiers of its structural dynamics (Sec. 1.4.2). Colored regions of the free-energy profile are zoomed in up to the lower tier. The colored text above the arrows denotes the characteristic timescale of transition between the conformational states of each tier (color-coded). Italic text reports the nature of the motion associated with the conformational transition. Tertiary interactions in **B** and jittering motions in **C** are represented by colored dotted and dashed lines, respectively.

Tier 0: Secondary structure dynamics. The slowest conformational dynamics happening in RNA molecules involve changes in its secondary structure, which span timescales of the order of 0.1 s or even slower for large molecules (Fig. 1.9A) [117]. This is due to the essential role that the strongest RNA intra-molecular interactions (base pairings and $\pi-\pi$) play in its secondary structure folding. Interestingly, since different base-pairing combinations and stacking interactions may have comparable energetic profiles, RNA molecules can exist with similar probability in distinct secondary structures. From a functional perspective, this may lead to kinetic trapping in a nonfunctional state [118]. In such cases, external energy contributions can trigger the transition toward a specific state. In the example of riboswitches, often characterized by two distinct secondary structures with similar energetic contributions, it is the binding with a metabolite that results in the regulation of gene

expression in bacteria systems [35].

Tier 1: base-pairing and tertiary dynamics. A given secondary structure may experience faster motions on the timescales ranging from microseconds to milliseconds that do not globally affect the secondary structure dynamics (Fig. 1.9B). It is possible to classify:

- **localized base-pairs dynamics**, involving changes in localized base-pairing interactions. First, base pairs may transiently melt into an open, energetically disfavored state due to the strength of stacking interactions and the base pair location within the structure, [119]. Second, base-pairs partners may undergo rearrangements in and around non-canonical structures, such as apical and internal loops [120]. Such phenomena rearrange temporarily the residues exposed to solvent and are crucial in the recognition of protein binding partners, like in the case of HIV-1 TAR with Tat protein [120]. Finally, two nucleic bases can pair in multiple configurations, varying by factors like rotatable bonds angle and protonation state [121]. This can change the 3D helical structure, thus impacting molecular recognition, like in the case of ion-binding to group I introns [122].

- **Tertiary-structure dynamics**, involving long-range tertiary contacts between distal RNA loops. The structural elements participating in tertiary interactions can undergo localized base-pairs dynamics, which affect the stability and structure of the RNA on timescales ranging from microseconds to seconds. Such rearrangements have been shown to modulate RNA catalytic cycles [123], substrate exchange [102], and ligand binding [124] in systems like ribozymes and riboswitches. Tertiary dynamics can also toggle a molecule between active and inactive conformations, as in the unique case of *Murine leukemia virus* (MLV) mRNA translation [125]. In ribosomes, precise tertiary-structure dynamics stabilize correct mRNA-tRNA pairs during decoding, contributing to high specificity in tRNA selection [126].

Tier 2: Jittering dynamics. Within the free-energy basin of a given global structure, and while undergoing tertiary interactions, RNAs experience a wide range of faster motions (Fig. 1.9C). Such dynamics mainly involve the jittering of structural motifs and span timescales from picoseconds to microseconds. The following main classes of dynamic motions can be distinguished:

- **Inter-helical dynamics**, involving large collective motions of RNA helical domains influenced by the cooperative stacking interactions of the flexible functions that interconnect them. This inter-helical motion is crucial for the overall RNA architecture, especially for the relative positioning of groups that participate in long-range tertiary interactions, catalytic activity, and protein binding [127]. An important example is constituted by the discussed HIV-1 TAR, which transitions between a bent conformation, stabilized by stacking interactions between the bulge and the lower helix, and a higher populated coaxially stacked conformation (Fig. 1.8).

- **Loop dynamics**. RNA secondary structure primarily comprises helical domains connected and capped by loops, which serve as flexible sites for interaction with proteins, RNAs, ligands, and small molecules. These loop regions are highly dynamic, with conformational changes occurring over a range of timescales, from picoseconds to microseconds, facilitating

various intermolecular interactions. An example of loop dynamics can be found in CUUG tetraloops, which are crucial elements involved in the molecular recognition of bacterial rRNA with proteins [128].

Despite the presented subdivision, it is important to remark that the energetic contribution between different tiers is interconnected [106]. For example, only a single set of tertiary interactions (Tier 1) may be possible for a given secondary structure (Tier 0). Alternatively, the possible loop conformations (Tier 2) can influence the entropic cost associated with the formation of tertiary interactions (Tier 1).

Modularity of structural motifs

Despite their diverse sequences, RNA molecules frequently adopt a limited set of secondary and tertiary structural motifs, resulting in similar structural configurations across various RNA types [114]. Tetraloops, hairpin structures that often cap RNA helices, exemplify this phenomenon: their structure remains consistent across different RNAs independent of the RNA sequence and the structural context outside the motif itself [128]. From a thermodynamics perspective, this indicates that the probability of forming a given motif conformation, and thus the corresponding ensemble, is dictated by its internal properties. In this so-called "RNA reconstitution model", the free energy landscape of constituent motifs are then added together to reconstitute the ensemble of an RNA assembly [129, 130]. Following this principle of ensemble modularity, the challenging description and characterization of large RNA-protein assembly, as well as of long non-coding RNAs may be facilitated. For this reason, tetraloops constitute a very important model system for both experimental and computational techniques aimed at describing RNA conformational ensembles [131] (Sec. 1.4.4).

The understanding of the hierarchical and interconnected nature of RNA motion and, at the same time, of the possibilities introduced by their modularity across different structural motifs, is fundamental. Such knowledge may indeed ease the design and realization of both experimental and computational studies aimed at characterizing its dynamic nature. A more comprehensive biophysical assessment of the different classes of RNAs, as well as the development of therapeutic strategies targeting them, is dependent on the quality and accuracy of such studies. In the perspective of the concepts discussed so far, the next sections are dedicated to overview the most important approaches in the characterization of RNA structural dynamics.

1.4.3 Experimental and computational approaches to determine RNA structure

Unraveling the structure of biomolecules is fundamental for comprehending their biological functions. Yet, the inherent dynamic nature of RNA presents a formidable challenge. This section provides an overview of state-of-the-art methods employed to investigate the structural dynamics of RNA molecules. The discussion commences with an exploration of experimental techniques employed for RNA structure determination. Then, I will overview the computational tools designed for RNA structure prediction. Special attention is given to assessing their efficacy in characterizing the rugged conformational landscape of RNA molecules.

Overview of the experimental methods for RNA structure determination

X-ray crystallography. X-ray crystallography has been the gold standard for high-resolution structure determination [132]. In this technique, immobilized crystals of the target molecule are exposed to X-rays. The produced diffraction pattern is transformed into an electron-density map that serves as a basis to model the 3D structure of the molecule. Despite the high resolution, which is the most important advantage of this technique, X-ray structures present different drawbacks. First, they capture only one conformation of the crystallized molecule, which may be largely insufficient in the context of RNA. Moreover, due to crystal packing effects, the resolved conformation may be not representative of biologically relevant conditions.

Cryo-Electron Microscopy (cryo-EM). Cryo-EM is a recently developed technique that revolutionized structural biology [133]. In cryo-EM, the sample is rapidly frozen at cryogenic temperatures, forming a vitreous ice where water molecules do not have time to form a crystal. Then, the scattering of electron beams is analyzed by electron microscopy producing 2D images that are computationally reconstructed into a 3D map. Cryo-EM provides information on a quasi-native state of the studied molecule, including snapshots of different conformational states prior to flash freezing. However, it does not provide kinetic information regarding their interconversion. This technique is suitable for large biomolecules, like ribosomes [134], but can not resolve molecules lower than 50 kDa thus limiting the application to many ncRNAs.

Nuclear Magnetic Resonance (NMR) spectroscopy. NMR spectroscopy makes use of magnetic fields to analyze a molecule of interest by observing the absorption and emission of energy in the radio-frequency range. Generally, this technique allows studying the molecule under a wide range of solution conditions and with very high spatio-temporal resolution for both local and global dynamics, spanning 12 orders of magnitude in time [99]. For these reasons, NMR is one of the most powerful methods in the study of RNAs, resolving biologically relevant conformations at low populations and at varying salt concentrations, even if only for short molecules of ~ 50 nucleotides [120, 135, 136]. A multitude of different measurements can be made in the context of NMR. Chemical shifts, residual dipolar and scalar couplings, and relaxation dispersion methods provide a comprehensive toolkit for probing RNA structural dynamics, offering insights into chemical nature, molecular orientation, torsional angles, conformational transitions, and structural ensembles. A detailed description of the corresponding experiments can be found in Ref. [99].

Chemical Probing. Specific chemicals are used to modify the chemistry of nucleotides in a structure-specific manner, depending on base-pairing, like in the case of dimethylsulfate reagents, or flexibility, like in the case of SHAPE (Selective 2'-Hydroxyl Acylation analyzed by Primer Extension) reagents [137]. The sites of chemical modification are then mapped using reverse transcription and sequencing methods like the recently developed Mutate and Map [138]. The timescales probed by this technique go up to minutes and therefore it is used to study RNA at secondary structure rearrangements level. The footprint of the molecule, whose reactivity is measured at the nucleotide level, is used to model single conformations and as a structural constraint for computational predictions [139–141] (Sec. 1.4.3).

Small-Angle X-ray Scattering (SAXS). SAXS provides structural information at low resolution and is particularly useful for studying large RNA complexes in solution [142]. By analyzing the scattering pattern of X-rays passed through the sample, one can deduce the overall shape and dimensions of the molecule.

Single-Molecule Förster Resonance Energy Transfer (smFRET). smFRET involves labeling RNA with donor and acceptor fluorophores and measuring the energy transfer between them. These energy measurements is then used to extract valuable structural information about the distances and dynamic interactions within the labeled regions of the RNA molecule. smFRET is particularly useful for studying RNA conformational changes, since it allows generating distributions of experimental observables over a structural ensemble [143], and it provides data at the single-molecule level.

Overview of the computational methods for RNA structure prediction.

When an experimental structure is not available, computational methods can be used for structure prediction [144]. Computational tools for structure predictions are able to generate the secondary or tertiary structure of a given biomolecule starting from its sequence. For what concerns proteins, the recent release of AlphaFold2 [145], which is able to predict the 3D structure of proteins with accuracy comparable to experimental techniques, the field of structure prediction is quite advanced. The development of AlphaFold2 pushed the relevance of computational structure prediction beyond the boundaries of specific applications to embrace a wider context of research, ranging from basic biological research to biotechnology. In this context, a variety of computational tools for RNA structure prediction have been developed [146, 147] and in this section I will provide an overview of their state-of-the-art.

Due to the complexity of the tertiary dynamics of RNA molecules, early efforts in RNA structure prediction focused on secondary structure. The knowledge of the secondary structure is indeed useful in the description of several biological processes. Secondary structure prediction tools can be divided into the following categories:

- **energy-based methods**, based on free-energy calculations. A representative example of these methods is given by RNAStructure by D. H. Matthews's group [139]. The energy of RNA secondary structures using the nearest-neighbor thermodynamic model, which sums the free energy contributions of base pair interactions and structural elements identified in the RNA sequence. The top-scored structure is reported together with a set of suboptimal structures.
- **evolutionary-based methods**, based on the evolutionary covariation of homologous sequences. A representative example of these methods is R-Scape by Eddy's group [148]. Information about base-pairs coevolution is extracted from a multiple sequences alignment and used to model a single RNA structure.
- **machine learning-based methods**, leveraging the structural information gathered from extensive training datasets. A representative example of these methods is Ufold by Xie's

group [149]. An image-like representation of the input RNA sequence is evaluated using a convolutional neural network with an encoder-decoder architecture, and a single structure is modeled on the basis of the predicted base-pairing interactions.

More recent advancements in the field made possible the implementation of computational tools for 3D structure prediction. Tertiary structure prediction tools may require both sequence and secondary structure as inputs and can be classified into:

- ***ab initio* folding methods**, based on template-free energy calculations, often carried out during molecular simulations. A representative example of these methods is SimRNA by Bujnicki's group [150]. A Monte Carlo sampling of the conformational space identifies thermodynamically relevant conformations that correspond to potential alternative structures.
- **fragment-assembly methods**, based on assembling structural motifs from a template library. A representative example of these methods is FARFAR2 [151] developed by Das' group, and included in the Rosetta suite. In this software, sets of fragments of three nucleotides extracted from similar RNA structures are assembled by a high-resolution scoring function to predict an ensemble of structures.
- **comparative modeling methods**, based on template structures. A representative example of these methods is VfoldLA by Chen's group [152]. VfoldLA first classifies single-stranded loops into four types and then assembles 3D structures of RNA molecules based on loop-helix connections using loop/junction templates.
- **machine learning methods**, based on artificial intelligence architectures trained on datasets of existing RNA-ligand structures. A representative example of these methods is given by ARES, by the Das' and Dror's groups [153]. ARES re-scores the tertiary structures generated by FARFAR2 based on a deep-learning-based scoring function. Even if trained on only 18 known RNA structures, ARES remarkably outperformed other state-of-the-art structure prediction tools.

Main limitations and possible solutions

In the scenario of the "non-coding RNA revolution" (Sec.1.2.2), an explosion of both experimental and computational research has provided crucial insights into the dynamics of RNA molecules. On the experimental side, advanced NMR techniques provided crucial insights into the role of RNA conformational propensity in cellular activity [100] and into the recognition between miRNAs and mRNA in the RNAi context [154]. Cryo-EM enabled the determination of atomic-resolution structures of large RNAs [155], such as the spliceosomes [156]. Chemical probing enlightened the mechanism of recognition between RNA and proteins [157]. However, especially in the case of complex systems, experimental methods have many limitations. Besides being prone to random and systematic errors, current experimental techniques do not have enough time resolution to distinguish between the relevant conformations that RNA populates [158, 159]. The resulting ensemble-averaged measurements need to be deconvoluted computationally in order to extract reliable information on the conformational heterogeneity of the system [160].
On the other side, computational structure prediction provides an alternative approach to solve the

problem of RNA structure determination [146]. In principle, these tools are able to predict relevant structures of a given RNA molecule with atomistic details. In practice, they are limited by two main factors: *i)* inaccuracies in the theoretical model used to score/generate the structures, and *ii)* the limited sampling of the RNA conformational space (Sec. 1.4.4). The advent of machine-learning methods offers a promising avenue for the future of RNA structure prediction. However, the current development of machine-learning models for RNA is dramatically hindered by the lack of extensive training sets of RNA molecules and their accuracy is not comparable with models used in protein structure prediction [161, 162]. The use of FARFAR2 for the sampling of RNA conformational space, coupled with the ARES deep-learning scoring function [153], has proven to be one of the most accurate approaches currently available. To further improve the accuracy of the generated models, experimental data can be integrated into the structure prediction calculations [163, 164]. In a very recent work, Al-Hashimi and coworkers proposed a novel method that, by integrating FARFAR2 predictions with NMR experimental data, allows for the simultaneous determination of relevant populated 3D structures, their relative abundance, and kinetic rates of interconversion [165].

The previously discussed experimental and computational techniques have been noted for their challenges in accurately modeling the intrinsic flexibility of RNA molecules. Computational structure prediction, with few exceptions, has traditionally prioritized the determination of a single structure rather than a structural ensemble. Notably, the RNA-puzzles competition [166], a significant global event in RNA structural assessment akin to CASP for proteins [167], focuses on predicting individual static structures. In contrast, Molecular Dynamics (MD) holds the potential to simulate the time evolution of a given RNA molecule at an atomistic level. In the next section, I will discuss the role of MD simulations in the description of the structural dynamics of RNA molecules.

1.4.4 RNA structural dynamics as captured by Molecular Dynamics

MD simulations are a physical approach for studying atomic and molecular interactions based on Newtonian physics [168]. By integrating Newton's laws of motion, successive configurations of the system are generated, yielding trajectories that track particle positions and velocities over time. These trajectories allow the calculation of various properties such as free energy and kinetics measures, offering insights beyond the sensibility of experimental techniques. Consequently, MD simulations serve as a natural framework to explore the structural dynamics of RNA molecules [169–171].
In this section, I will first introduce the fundamental principles of atomistic MD simulations. Then, I will overview the state-of-the-art in the context of RNA systems, and I will discuss the main limitations of MD together with the corresponding solutions adopted by the scientific community. Except when explicitly indicated, the discussion is based on the textbook "Understanding Molecular Simulations" by D. Frenkel and B. Smit [168].

Sampling with Molecular Dynamics

MD is a computational simulation technique that provides a time-dependent characterization of many-body molecular systems. Given N interacting components, the core of MD is the prediction of their trajectory, namely positions and velocities, by numerically integrating the equations of

motion. Depending on the aim of the study, a system can be indeed modeled at different levels of granularity:

- **atomistic MD**, where the system is represented explicitly as a set of interacting atoms. Integrating classical Newton equations of motions, the accessible timescales are currently of the order of microseconds;
- **coarse-grained MD**, where groups of atoms are represented by single interaction sites ('beads') to reduce the complexity of the system. Integrating Newton's equation of motion, accessible timescales range from microseconds to milliseconds.
- **quantum-mechanics MD**, where a more accurate description of the translational, rotational, or vibrational degree of freedom is obtained by modeling the quantum effects of the electronic environment. Integrating Schrödinger's equation of motion, the accessible timescales are of the order of picoseconds to nanoseconds.
- **hybrid MD**, where the above techniques are combined and different parts of the system may be modeled with different approaches. Timescales and equations of motion depend on the choices of the representation.

In the remainder, I will focus on atomistic explicit-solvent Molecular Dynamics simulations, but most of the concepts can in principle be generalized to other classes of molecular simulations.

MD basic formalism and assumptions MD simulations are equivalent, to some extent, to single-molecule experiments with atomistic resolution: the position $\mathbf{r}_i(t)$ and momentum $\mathbf{p}_i(t)$ of each atom i at time t are given by integrating Newton's law:

$$\frac{d\mathbf{p}_i}{dt} = -\frac{\partial U_i(\mathbf{r}(t))}{\partial \mathbf{r}_i} \quad i = 1, ..., N \tag{1.1}$$

where $U(\mathbf{r}(t))$ is the force field analytical expression depending on the interactions of the system in the configuration defined by the positions $\mathbf{r} = \mathbf{r}_1, ..., \mathbf{r}_N$ and the momenta $\mathbf{p} = \mathbf{p}_1, ..., \mathbf{p}_N$ at time t. The knowledge of the initial conditions of positions and momenta, and of the functional form of the potential energy are necessary and sufficient conditions to solve Eq. (1.1) and determine the evolution of the system. A generic variable A depending on positions $\mathbf{r}_i$ and momenta $\mathbf{p}_i$ can be therefore measured as the time average $\overline{A}$ over the MD simulation:

$$\overline{A} = \frac{1}{T_0} \int_{t=0}^{T_0} A(\mathbf{r}(t), \mathbf{p}(t))\, dt \tag{1.2}$$

Under the assumption of the ergodic hypothesis[172], the time-averaged value $\overline{A}$ over a sufficiently long simulation time T_0 corresponds to its statistical average $\langle A \rangle$ in the phase space of positions $\mathbf{r}_i$ and momenta $\mathbf{p}_i$

$$\lim_{T_0 \to \inf} \overline{A} \equiv \langle A \rangle = \int A(\mathbf{r}; \mathbf{p})\, \rho(\mathbf{r}; \mathbf{p})\, d\mathbf{r}\, d\mathbf{p} \tag{1.3}$$

where $\rho(\mathbf{r}_1, ..., \mathbf{r}_N; \mathbf{p}_1, ..., \mathbf{p}_N)$ is the probability distribution in the phase-space. The compliance to the ergodic hypothesis is at the basis of the accuracy of MD predictions: it ensures that a sufficiently long trajectory can provide a reliable representation of the system's macroscopic properties from its microscopic states.

Realistic cellular conditions: NPT and NVT ensembles Biomolecular simulations aim to realistically reproduce the cellular environment. The direct integration of Newton's equations (Eq. (1.1)) leads to a simulation in the microcanonical ensemble (NVE). The NVE ensemble describes a system with conserved total energy and, therefore, cannot provide an accurate description of the behavior of biomolecules in cells, which constantly exchange energy with the environment. To address this issue, the simulated system is coupled with an external thermostat and barostat to keep temperature and pressure constant, respectively. To this end, stochastic energy terms are added to the equation of motion (Eq. (1.1)) to model the thermal and/or pressure baths.
Depending on the aims of the study, MD simulations are normally conducted in:

- *isothermal-isobaric ensemble* (NPT), where N atoms interact at constant temperature T and pressure P. In this case, the probability distribution in Eq. (1.3) is given by

$$\rho(\mathbf{r};\mathbf{p}) = \frac{e^{-\beta(H(\mathbf{r};\mathbf{p})+PV)}}{\Xi(N,P,T)} \tag{1.4}$$

  where $\beta = \frac{1}{k_B T}$, $k_B = 2.479\ kJ/mol$ is the Boltzmann constant, H the Hamiltonian describing the system, V is the volume and $\Xi(N,P,T) = \int e^{-\beta(H(\mathbf{r},\mathbf{p})+PV)}\ d\mathbf{r}d\mathbf{p}$ is the NPT partition function.

- *canonical ensemble* (NVT), where N atoms interact at constant temperature T and volume V. In this case, the probability distribution in Eq. (1.3) is given by

$$\rho(\mathbf{r};\mathbf{p}) = \frac{e^{-\beta H(\mathbf{r};\mathbf{p})}}{Z(N,V,T)} \tag{1.5}$$

where $Z(N,V,T) = \int e^{-\beta H(\mathbf{r},\mathbf{p})}\ d\mathbf{r}d\mathbf{p}$ is the NVT canonical partition function.

Force fields An atomistic force field consists of an analytical expression of the interatomic potential energy and therefore implies the introduction of several parameters. The determination of these parameters is performed *ab initio*, by quantum mechanics calculations, or empirically, by fitting experimental data [173], or both. Force fields can be further classified based on their approach to modeling atomic polar characteristics. In non-polarizable force fields, each atom is assigned a fixed partial charge that does not change during the simulation. In contrast, polarizable force fields allow partial charges on atoms to vary in response to their local environment, mimicking real molecular polarization effects. Both classes of force fields are generally expressed as a series of pairwise additive terms modeling bonded and non-bonded interactions. In addition, polarizable force fields include a polarization term that models the dynamic response of the electronic distribution, often by allowing the charge of an atom to fluctuate in response to the local electric field. This more complex formulation of polarizable force fields results in a higher accuracy, but a lower computational efficiency. For this reason, non-polarizable force fields are more extensively used in biological applications, even if the fine-tuning of force parameters can not yield the same level of accuracy than polarizable ones [169].
The class of empirical non-polarizable force-field has generally a simple functional form. A representative example is constituted by the form used in the force field of AMBER simulation package,

developed in 1995 by Cornell *et al.* [174], where the different energetic contributions are summed together:

$$U = U_{\text{stretch}} + U_{\text{angle}} + U_{\text{dihedral}} + U_{\text{nonbond}} \tag{1.6}$$

Indicating with $r(t)$, $\theta(t)$, and $\psi(t)$ bonds length, bonds angle, and dihedrals angle at time t, the bonded interactions are represented by a set of two harmonic springs:

$$U_{\text{bond}} = \sum_{\text{bonds}} k_r[r(t) - r_0]^2 \tag{1.7}$$

and

$$U_{\text{angle}} = \sum_{\text{angles}} k_\theta[\theta - \theta_0]^2 \tag{1.8}$$

with force constants for bond stretch k_r, and angle bending k_θ and equilibrium values r_0, and θ_0, respectively. In addition, the dihedrals are represented by a sum of periodic functions:

$$U_{\text{dihedral}} = \sum_{\text{dihedrals}} \frac{U_n}{2}[1 + \cos(n\phi(t) - \gamma)] \tag{1.9}$$

with amplitude U_n, periodicity n and phase γ.
Non-bonded interactions are modeled by the sum of Lennard-Jones and Coulomb contributions:

$$U_{\text{nonbond}} = \sum_{i,j}^{N} \epsilon_{ij} \left[\left(\frac{R^0_{ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{R^0_{ij}}{r_{ij}} \right)^{6} \right] + \sum_{i,j}^{N} q_i q_j \left(\frac{1}{4\pi\epsilon_0} \right) \frac{1}{r_{ij}} \tag{1.10}$$

where r_{ij} is the distance between atoms i and j, ϵ_{ij} is the potential well depth, R^0_{ij} is the Van der Waals distance, q_i the charge of atom i, and ϵ_0 the permittivity of free space.

The quality of the force field used in an MD simulation is the critical element for the accuracy of the predictions. In this sense, it is fundamental to review the state-of-the-art of RNA atomistic force fields.

State-of-the-art of RNA atomistic force fields

Compared to proteins, the development of RNA force fields has progressed at a slower pace [175]. However, in parallel with the growing interest in RNA molecules, the increase in computational power, and the need to address limitations observed in long simulations of RNA molecules, several developments have been implemented [176]. In the following sections, I will introduce the three most widely used families of RNA force fields [177–179]: AMBER, CHARMM, and DESRES. Then, given their importance in RNA structural dynamics (Sec. 1.1.3), I will overview the state-of-the-art of water and ions force fields. The dihedral angles mentioned in this section are visualized in Fig. 1.1.

AMBER force fields. The most popular RNA force field belongs to the Assisted Model Building with Energy Refinement (AMBER) family [177]. The first version supporting nucleic acids simulation was presented by Cornell *et al.* [174], and it is generally referred to as AMBERff94. The parameters of AMBERff94, including both proteins and nucleic acids, were modified in 1999 [180]

and 2000 [181], resulting in the major branch of AMBER force fields called AMBERff99. Concerning nucleic acids, these modifications involved the refinement of the sugar puckering and χ dihedral parameters. A major correction, known as *bsc*0 for RNA and DNA was introduced by Orozco's group in 2007 [182]. In this correction, α and γ dihedral angles of the nucleic acid backbone were modified to reduce unrealistic helical twists in RNA stems. In 2011, another milestone was set by Zgarbova *et al.* [183] with the reparameterization of the χ dihedral to prevent the formation of entirely untwisted and ladder-like structures known as χ_{OL3} refinement.
It is important to underline that alternative modifications have been proposed for the AMBER force field, but the *bsc*0 and χ_{OL3} ones stand out for their critical stabilization of nucleic acid simulations and lower presence of undesirable side effects [177]. In 2010, Yildrim *et al.* [184] proposed a χ reparameterization to correct the *syn/anti* balance on the basis of NMR data, achieving performance similar to χ_{OL3}. However, the subsequent extension AMBER99TOR [185], introducing a reparameterization of all backbone torsion angles and thus having a completely different set with respect to χ_{OL3}, performs suboptimally for helical regions of RNA [186]. Further tuning of the force field involved the modifications of the non-bonded terms, as for the balance between stacking, H-bonding, and solvation in Chen and Garcia work [187], or the modified phosphate oxygen parameters in the AMBERFFLJbb force field proposed by Bergonzo and Cheatham [188]. More recent attempts to improve the AMBER force field for RNA are the dihedral reparameterization conducted in the Mathews' group [189]. In more recent years, alternative schemes were also proposed, including the introduction of an additional term to better describe hydrogen-bond interactions [190] and a grid-based energy correction map (CMAP) term [191].

DESRES force fields. For both proteins and nucleic acids, a substantial contribution to the development of atomistic force fields comes from the D.E. Shaw RESearch (DESRES) group. DESRES developed one of the most powerful computational facilities, the supercomputer Anton, which allows simulating the behavior of biomolecules on timescales inaccessible to most of the scientific community [192]. The work of DESRES has been focused first on the improvement of the AMBER force fields. The current state-of-the-art version of the AMBER force field was implemented in 2010, under the name of Amberff99SB-ILDN, which significantly improved torsion parameters [193]. All further developments of the AMBER force field, discussed in the previous paragraph, did not modify the electrostatic parameters, as this would have required a complete reparameterization of the torsional terms in the FF [176]. Remarkably, DESRES introduced in 2018 a new force-field for nucleic acids by altering AMBER nucleobase charges, along with the recalibration of Lennard-Jones and several torsion parameters [194]. The DESRES force field improved nucleobase stacking, base pairing, and key torsional conformers and demonstrated accuracy comparable with the state-of-the-art protein force fields. Very recently, DESRES introduced the DES-Amber force field, built by transferring non-bonded parameters for RNA from the DESRES force field and adjusting the torsion ones to obtain electrostatics prediction more compatible with the Amberff99SB-ILDN force field [195, 196].

CHARMM force fields. The Chemistry at HARvard Macromolecular Mechanics (CHARMM) is another major family of potentials developed for molecular simulations, originally developed by Karplus's group in the 1980s [197]. The first parametrization of nucleic acid potential for CHARMM

molecular simulations was performed in 1995, with the development of CHARMM22 [198]. An improved version of the force field was CHARMM27, which introduced a reparameterization of the dihedrals contributions and addressed base-pair opening issues [199]. However, the stability of nucleic acid duplex was reported as underestimated by different groups, which observed fraying phenomena, unnatural partial unfolding, or separation of the nucleotides at the ends of RNA strands [200]. The latest CHARMM release for nucleic acids is CHARMM36 [201], which only partially addressed the fraying issue [169]. The CHARMM community is leading the efforts in the development of polarizable force fields for nucleic acids [202, 203], even if the majority of biological applications are performed by the less computationally expensive non-polarizable ones [177]. It is important to remark that important studies in the context of RNA-targeted drug discovery used a CHARMM force field, as in the case of the dynamic ensemble of HIV-1 TAR determined by Al-Hashimi and co-workers [110] (Sec. 1.4.4), and in the SILCS-RNA approach developed by Mackerell and co-workers [204] to identify small-molecules binding site in RNA molecules (Sec. 1.5.4). However, the application of CHARMM for RNA studies has been tested less thoroughly compared to the AMBER family [177].

Water and ions. The cellular environment is aqueous and its representation significantly influences the accuracy of simulated molecular interactions. For explicit-solvent MD simulations, the modeling of water molecules is crucial to obtain accurate and reliable predictions, especially in the case of flexible systems such as RNAs [205]. Moreover, the optimization of the water model is a possible way to correct the unbalance of RNA stacking and hydrogen bonding interactions mentioned above. In this sense, the work by Bergonzo *et al.* identified the OPC water model [206] as a better alternative with respect to other models, as proven by the better agreement with the conformer populations derived from NMR experiments for the repeated tetraloop r(GACC) [188]. Analogously, DESRES observed significant improvements in their force field predictions for highly flexible systems when accounting for water dispersion interactions with the TIP4P-D water model [207].
Most importantly, the anionic RNA molecules directly interact with cation salts in solutions, which play a fundamental role in RNA folding (Sec.1.1.3, dynamics, catalysis, and, in general, function [4, 208]. Therefore, their inclusion and accurate modeling is also crucial to obtain an accurate representation of cellular conditions and reliable predictions of RNA dynamics [209]. To this end, one of the most important parametrizations of monovalent ions in the framework of AMBER force fields is the one made by T. E. Cheatham [210]. In cells, RNA molecules mostly interact with divalent cations, in particular Mg^{2+} [211]. Due to their diverse roles in RNA biology, obtaining accurate force field parameters for Mg^{2+} has been proven to be particularly challenging [212]. Indeed, the electronic configuration of Mg^{2+} is not negligible in its non-bonded interactions with RNA and the introduction of *ad hoc* parameters determined from the quantum-mechanics calculations is generally adopted [213]. However, these applications are currently limited by the computational cost required to simulate relevant biological timescales [212] and therefore state-of-the-art approaches for the parametrization of Mg^{2+} ions rely on standard MD simulations, like in the case of Villa's parameters [214] adopted in the work of this thesis.

Main limitations of MD simulations

MD simulations constitute a powerful tool to simulate and study biomolecules. In theory, MD can accurately predict the evolution of a given system for an arbitrary simulation time. In practice, two main issues make the former statement untrue and they are discussed in the following paragraphs.

Inaccuracy of force fields. The first issue is the inherent inaccuracy of the force field potential, an empirical model designed to mimic the real interatomic forces acting on the simulated molecular system. Despite the titanic efforts of AMBER, CHARMM, and DESRES communities, the accuracy of RNA force fields is still limited. One of the most concerning issues is the unbalance between $\pi-\pi$ stacking [215] and hydrogen-bond interactions [216] and the improper hydration of RNA functional groups [188]. These fundamental difficulties lead to overestimating the populations of non-native RNA conformations, especially concerning imbalances between folded and unfolded states. The development efforts of the last decade were triggered by the necessity of overcoming such problems and the field is now in an exciting, but turbulent phase in which the choice of the force field is not trivial [176]. The applicability of a particular version of a force field can fluctuate not only among simulated biomolecules but also within distinct regions of a singular simulated molecule. In this sense, a reparameterization of the force field may improve the accuracy for specific systems but also fail for others.

Sampling the conformational space. The second issue is the exhaustive sampling of the conformational space of the simulated system. The constraints imposed by current computer hardware introduce a feasibility upper threshold for the simulation time and the compliance to the ergodic hypothesis (Eq. (1.3)) can not be ensured in most cases. From a quantitative point of view, the sampled timescales in MD simulations typically do not go beyond the order of microseconds [177]. Relevant biological processes involving RNA conformational changes may take place in significantly longer timescales (Sec. 1.4). For example, ligand-binding and catalytic reactions take place on timescales ranging from microseconds to hours [107]. Moreover, simulations are performed starting from single geometries and the explored ensemble is generally highly dependent on the initial state [217]. Due to substantial energetic barriers between distinct states, accurately sampling all available states in the free-energy landscape requires exploring multiple rare events. This task, however, surpasses the capabilities of standard MD simulations [218]. To this end, it is common to make use of advanced statistical mechanics techniques that respectively improve and enhance the sampling of the conformational space [107, 219, 220]. The two following sections will overview the main principles of these advanced techniques and present their state-of-the-art in the context of RNA molecules.

Improving force fields by integrating experimental data

A promising strategy to effectively improve MD prediction quality is combining them with experimental information to obtain more accurate structural ensembles. This strategy has been originally developed for proteins [221] and more recently has been applied to RNA [222]. In terms of general principles, integrative methods can be classified into two general categories:

- **system-dependent methods**, in which ensemble-averaged experimental data are used either

on-the-fly during molecular simulations or *a posteriori* to refine the population. Results from this approach are not transferable to other systems, even if they may be used to predict new experiments on the same system. Existing techniques are often based on Maximum Entropy principle [223] and Bayesian inference [224].

- **system-independent methods**, in which experimental data are exploited to improve force field parameters [190]. In this context, machine-learning approaches constitute a promising strategy and enable the prediction of the impact of parameter modifications. Results obtained with this methodology are in theory transferable to other systems.

Such approaches allowed the accurate determination of the structural ensembles of several RNAs, including oligonucleotides [225] as well as the more challenging tetraloops [131, 226]. A comprehensive review of the general principles of integrative approaches and of the current advancements in the RNA field goes beyond the scope of this thesis and the reader is addressed to the seminal review papers of Refs. [158, 222], respectively.

Enhanced sampling of the RNA conformational space

The key concept of all enhanced sampling techniques is to alter the dynamics of a system to facilitate the crossing of free-energy barriers and ultimately to accelerate the occurrence of rare events. From the point of view of general principles, the majority of these approaches can be categorized into [219, 227]:

- approaches that modify the probability distribution of selected degrees of freedom of the system, named "collective variables" (CVs), whose transitions are slow and difficult to sample;
- approaches that alter the Hamiltonian parameters, such as temperature, to facilitate barrier crossing.

In the two following paragraphs, I will briefly overview the main principles, strengths, and limitations of the two approaches as well as their state-of-the-art in the context of RNA molecules. The reader is addressed to the referenced literature for a more comprehensive overview of enhanced sampling techniques.

CV-based enhanced sampling techniques. A CV is a function of the configuration of the system that characterizes its collective behavior, providing insights into specific dynamic processes or transitions. Different methods have been developed to introduce in the simulation a "bias" potential acting on a selected set of CVs and thereby modifying the corresponding free-energy landscape. Examples of these methods include umbrella sampling [228] and metadynamics [229]. In order for these methods to be effective, the choice of a good set of collective variables is crucial. First, CVs should properly describe the slow motions of the system, which need to be accelerated in order to have an exhaustive sampling. Second, they should be as few as possible to describe the system in a reduced dimensionality framework, pinpointing the essential properties responsible for the complex phenomenon under investigation. In a first approximation, CVs to study biological systems may be chosen by physicochemical intuition. For instance, specific interatomic distances are suited to sample chemical reactions. This kind of CVs allowed the characterization of RNA-ligand

and RNA-ions interaction for a variety of RNA motifs [230, 231], as well as RNA-induced catalysis, for bacterial [232] and viral RNA systems [233]. In these examples, the *a priori* knowledge of the properties of the system guided the choice of CVs. However, when knowledge is limited, the task of selecting CVs represents a main challenge for these enhanced sampling techniques. Moreover, the more ambitious goal of exhaustively sampling the conformational space of an RNA molecule requires the choice of more complex CVs and a much greater computational cost. Existing approaches implemented CVs consisting of the combination of tertiary contacts [234] or RNA-specific metrics [235, 236]. In recent years, artificial intelligence (AI) has taken a prominent role in the selection of collective variables [237] and its applications are gaining momentum in the context of RNA and, in particular, RNA-small molecules interactions [238].

Techniques that modify the Hamiltonian. A second class of approaches includes those based on modifications of selected parameters of the potential energy function as well as the temperature of the system. An important example of this category is Replica Exchange Molecular Dynamics (REMD), based on performing different replica simulations at different values of an "exchange" variable [219], traditionally the temperature (T-REMD [239]). This exchange mechanism allows escaping local minima by passing to a higher temperature condition. While requiring less *a priori* information on the system, the replica exchange approach is much more demanding from a computational cost perspective than enhancing the sampling of a set of few CVs [107]. Despite the cost, such methods are widely used by the scientific community to explore RNA structural dynamics. Several efforts were conducted to sample the folding of different RNA systems, ranging from small tetraloops [240] to larger riboswitches [241]. A combination of tempering approaches with CV-based ones has also been proposed to explore the conformational landscape of tetranucleotides [242].

1.5 Strategies to identify small-molecules targeting RNA: a computational perspective

In constituting a major challenge for the biophysical characterization of RNA molecules, their flexibility also impedes the comprehensive understanding of molecular recognition mechanisms with binding partners. In this context, the therapeutic relevance of targeting RNA with small molecules is still modest. The clinically most advanced compounds so far were first identified by costly and time-consuming screening experiments with a restricted knowledge *a priori* of the binding with the RNA target. An alternative and less expensive approach is constituted by the rational design of small molecules targeting RNA, leveraging the knowledge of the physicochemical properties that drive the interactions between targets and ligands. In this perspective, the challenges introduced by RNA flexibility also present therapeutic opportunities to design RNA-targeted drugs. The discussed computational methods to investigate RNA structural dynamics, and in particular MD simulations, offer a natural approach to effectively address these opportunities and they may be crucial in the identification of RNA-targeted therapeutic agents. However, the existing computational tools present several limitations in fully accounting for the inherent flexibility of RNA molecules.

In this section, I will first provide an overview of the drug discovery pipeline, introducing the role of computational methods during the process. Then, I will discuss the challenges and opportunities

introduced by the inherent flexibility of RNA as drug target. In this view, the main successes and limitations of the experimental approaches employed in RNA drug discovery will be summarized to introduce the need of advanced computational techniques. Among these, I will finally discuss the state-of-the-art in addressing the problems of *i)* identifying a binding site on a given target and *ii)* find a suitable partner for the binding.

1.5.1 The drug discovery pipeline and Computer-Aided Drug Design

A drug discovery campaign is a multifaceted journey aimed at identifying potential therapeutic agents that can interact effectively with a biological target [243]. Depending on the success of the drug discovery campaign, the candidate molecules are then subjected to the subsequent stages of drug development, including clinical tests, further optimization, and, finally, marketing. In the context of this thesis, I will focus on the drug discovery pipeline and I will emphasize the opportunities introduced by computational methods to streamline the process by aiding the rational design of drugs.

Traditional drug discovery pipeline

The drug discovery pipeline (Fig. 1.10A) can be summarized into a series of essential steps [243]), which I am going to overview in the following paragraphs.

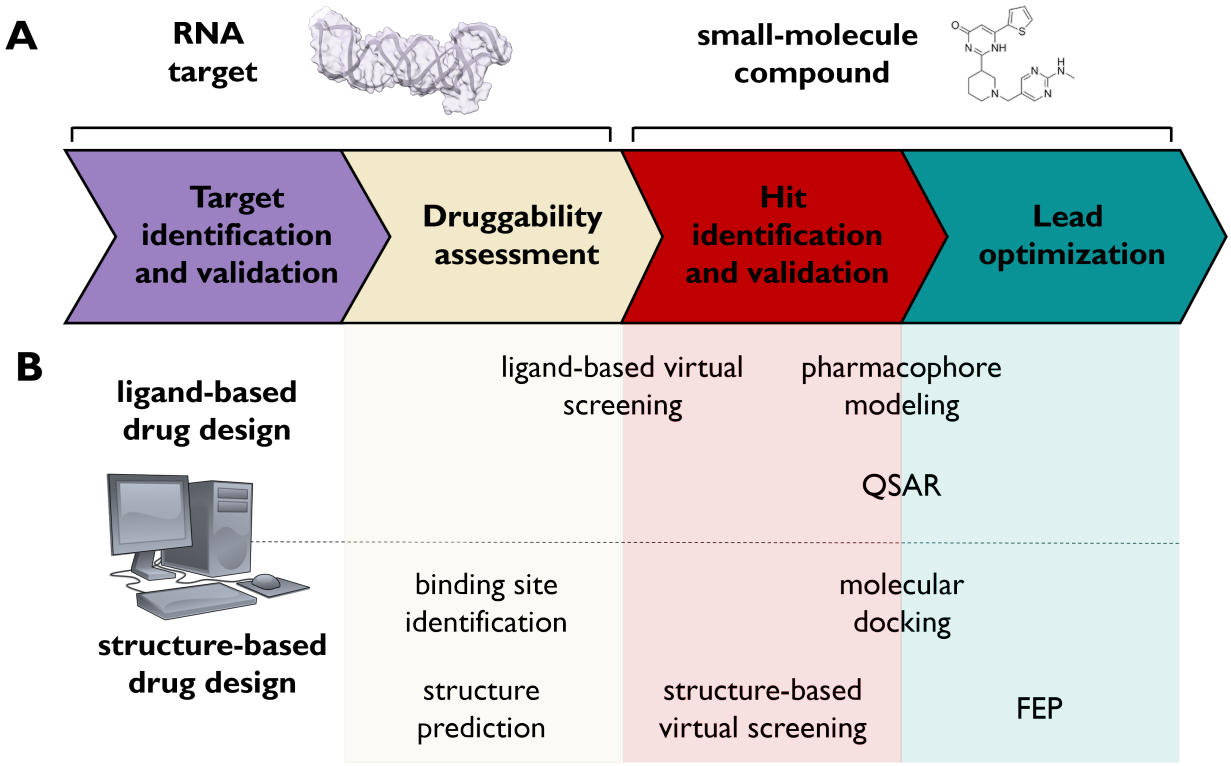


Figure 1.10: The role of computational methods in the drug discovery pipeline. A) Flowchart of the main steps of a drug discovery campaign (Sec. 1.5.1). On the top of the flowchart, RNA and ligand representations are shown to underline the steps that focus on the target and on the drug, respectively. **B)** Computational operations discussed in the text for ligand-based and structure-based drug design, respectively above and below the dotted line.

Target Identification and Validation. The operation of target identification marks the initial phase of drug discovery, where a biological molecule is recognized as a potential therapeutic target due to its involvement in a specific disease pathway. Once a given target has been identified, it is crucial to validate its biological relevance. This process involves confirming *i)* the target's accessibility to small molecules within the cellular environment and *ii)* assessing whether modulating its activity or function produces the desired impact on the disease phenotype. This foundational step ensures that the selected target is amenable to therapeutic intervention and that it can be the object of a drug discovery campaign.

Target Druggability Assessment. After identification, the target undergoes a comprehensive evaluation to determine its "druggability," assessing its potential to interact with and respond to small-molecule drugs. This evaluation occurs from both a physicochemical and functional perspective. The physicochemical analysis is carried out by structural determination techniques and biological assays (Sec. 1.5.3) that provide insights exploitable in further compound development. On the functional side, cellular or biochemical assays are employed to characterize the impact of ligand binding on the target's function. The knowledge acquired in this step is crucial for the subsequent steps of the pipeline.

Identification of hit small molecules. Once the druggability of the target is assessed, the subsequent phase involves the identification of compounds suitable for the interaction. Generally, screening experiments of the target against large libraries of compounds are employed in this step (Sec. 1.5.3). Often, given the low rate of success of such experiments, a sufficient criterion to retain a certain molecule is a high binding affinity with the target. At this stage, these compounds are termed "hits": they may not exhibit therapeutic effects, or they may possess poor pharmacological properties. These issues will eventually be addressed in successive stages.

Hit Validation. Due to their high-throughput nature, screening experiments are often conducted in simplified and controlled laboratory conditions. In the phase of hit validation, the biological activity of identified hits is rigorously confirmed within a biologically relevant context that mimics the physiological condition of the human body. Evaluation encompasses a thorough assessment of hit affinity and specificity, coupled with an examination of their functional impact through cellular or biochemical assays. A crucial goal of hit validation is the elimination of false positives and the assessment of the therapeutic relevance of the effect of the hits on the target. A successful validation advances the hits to the next stage of optimization and further development.

Small-Molecule Design and Lead Optimization. The validated hits are subjected to iterative cycles of optimization to improve their pharmacological and pharmacokinetic properties (Sec. 1.3.2). This process, often known as lead optimization, employs a combination of medicinal chemistry, computational modeling, and empirical testing to refine these molecules into viable drug candidates to be proposed for clinical phases.

Computer-Aided Drug Design

Each step of the drug discovery pipeline integrates multidisciplinary expertise in biology, chemistry, biophysics, and pharmacology. Despite the tremendous theoretical and technological progress in all these fields, the overall process remains slow and expensive [244]. In particular, employing the screening of huge libraries of compounds to find a drug candidate by trial and error, early-discovery and pre-clinical efforts are estimated to account for more than the 43% of total costs [245]. At the same time, the majority of time- and cost- saving opportunities lie in these early stages, through a primary characterization of the molecular mechanism of recognition adopted by a given target and its potential ligands. In this perspective, *in silico* rational drug design, also referred to as Computer-Aided Drug Discovery (CADD), is an alternative approach that focuses on the design of compounds specifically tailored to interact with a given biological target in well-characterized binding pockets [246]. Currently, both academia and the pharmaceutical industry are investing in computational drug discovery, particularly in the emerging deep-learning and artificial intelligence applications [247, 248].

CADD approaches are generally split into two categories [249]: Ligand-Based Drug Design (LBDD), and Structure-Based Drug Design (SBDD). In the following paragraph, I will briefly overview the most common operations of both categories and their role in the drug discovery pipeline.

Ligand-based drug design. LBDD is centered on the known ligand interactions with the target. Generally, this approach relies on the basic assumption that compounds with similar structural or physicochemical properties exhibit similar activities [250]. The information given by known active ligands against a macromolecular target is used to infer the relevant physicochemical properties responsible for the activity and to drive the design of new compounds [249, 251]. In the LBDD framework, the operations that can be performed to support the drug discovery pipeline are listed in the following.

- **Quantitative Structure-Activity Relationship.** Quantitative Structure-Activity Relationship (QSAR) is a computational approach aimed at establishing the relationship between the chemical structures of a compound series and a specific chemical or biological activity. The process begins by identifying a set of chemical entities or lead molecules that demonstrate the sought biological activity. A quantitative correlation is then established between the physicochemical properties of these active molecules and their biological effectiveness. QSAR models are employed in the hit identification stage, as filtering criteria in ligand-based virtual screenings, and in the lead optimization stage, to refine the active compounds.

- **Pharmacophore modeling.** A pharmacophore is an abstract and geometry-based description of the molecular features that are necessary for molecular recognition of a ligand by a specific biological macromolecule. Differently from QSAR, this approach is more versatile since it focuses on the 3D spatial configuration of these properties in order to obtain a certain activity, more than finding a quantitative relationship. Pharmacophore models are employed in the hit identification stage, as templates in ligand-based virtual screenings, and in the lead optimization stage, to refine the active compounds.

- **Ligand-based virtual screening.** Ligand-based virtual screening (LBVS) is used to identify potential hits on a given target. In this approach, large libraries of compounds are screened to identify the ones that have a better match with a given QSAR or pharmacophore model and may be potential lead candidates.

Structure-based drug design. SBDD is based on the structural information about the target and, eventually, of the interacting ligand. The knowledge of binding sites within the three-dimensional structure of the macromolecular target guides the design and assessment of ligands, based on the potential interactions with the given binding site [244]. In this context, the computational operations that support the drug discovery pipeline are listed in the following (Fig. 1.10).

- **Target structure prediction.** When experimentally determined structures are not available, the prediction of a 3D structure of the target is the first essential step in SBDD. To this end, a variety of computational methods have been developed and the subsequent steps of the pipeline rely on the accuracy of such predictions. It is important to remark that structure prediction tasks may require the correct modeling of the flexibility of the target in order to be effective for drug design purposes.

- **Binding site detection.** Once the structure of the molecular target is determined, the subsequent step is to identify potential hotspots for small molecule binding, namely regions suited for a stable and specific interaction. Computational methodologies designed for binding site detection can delineate favorable interaction sites on a specific target, offering atomic-level structural insights. Such approaches provide valuable physicochemical information that might be prohibitively expensive or challenging to obtain through experiments, especially in the context of non-native conformations.

- **Molecular docking.** If the structures of both the target molecule and potential ligand are available, molecular docking is a computational technique used to predict their complex structure and the corresponding binding affinity. This potent tool aids in the identification of potential hits through its application in virtual screening campaigns. Additionally, it plays a crucial role in lead optimization by furnishing a thorough physicochemical analysis of the interactions between a candidate compound and the target.

- **Structure-based Virtual Screening.** Structure-based Virtual Screening (SBVS) is a computational technique used to identify potential binders by the systematic application of molecular docking to large libraries of compounds. This step represents the computational counterpart of experimental screenings (Sec. 1.5.3) and can be used in the context of hits identification to aid the identification of candidate molecules for the subsequent *in vitro* biological assays conducted during lead optimization.

- **Free-energy perturbations.** Free-energy perturbation (FEP) is a computational technique used to estimate the free energy change associated with small modifications to a molecular system, such as ligand binding to a target. By simulating how slight alterations in a ligand structure affect its binding affinity, FEP helps in predicting the impact of these changes on the biological activity. The strength of FEP lies in its ability to provide detailed insights into the

thermodynamics of ligand-receptor interactions, guiding the optimization of lead compounds with higher precision.

Hybrid approaches. When data about both the structure of ligand-target complexes and similarity relationships to active compounds are available, the integration of both ligand-based and structure-based approaches may lead to a more comprehensive framework suitable to enhance the success of computational drug discovery applications [252, 253]. It is worth it to mention these two approaches:

- **Combined virtual screening.** Ligand-based virtual screening is performed to identify potential hit compounds, peculiarly focusing on their desired biological activity and without detailed structural information about the binding itself. Oppositely, structure-based virtual screening campaigns identify potential hits compounds focusing on the predicted affinity of the binding, disregarding direct considerations about the desired biological activity, which can be embedded in the screened library or addressed in the subsequent lead optimization stage. In this perspective, the combination of the two methodologies, for example including pharmacophore restraints in SBVS, may enhance the probability of successfully identifying active and selective small-molecule compounds.

- **Sub-structure similarity search.** A common strategy for identifying molecules likely to possess a desired affinity or biological activity is to screen existing repositories collecting interaction information of entirely or partially similar targets and/or compounds. Such databases, especially when available through dedicated webservers, are of fundamental importance for the development of rational design approaches. While similarity search has been traditionally applied from the ligand perspective seeking a desired activity in similar ligands in LBVS [254], it is in principle applicable to the target, by looking for known targets similar in its primary, secondary, or tertiary structures [255].

1.5.2 The importance of modeling RNA flexibility in drug discovery

The pace of advancement of RNA-targeted drug discovery lags significantly behind the protein-targeted counterpart. A critical obstacle is constituted by the inherent flexibility of RNA targets, which is hindering the development of approaches aimed at the rational design of small molecules targeting RNAs [52, 64–67]. However, besides constituting a major challenge, the flexible nature of RNA targets holds promising potential to develop novel and effective therapeutics strategies [64, 256].
In this section, I will first recall the basic principles of biomolecules recognition. Then, the discussion will focus on the insights that can be inferred from the characterization of the thermodynamics and kinetics properties of RNA-small molecule binding. Finally, I will overview how the rugged conformational space of RNAs can be leveraged to develop therapeutic strategies.

Describing the mechanism of biomolecular recognition

Highly specific and tightly regulated interactions between macromolecules are at the basis of all biological processes in living organisms. In particular, the molecular recognition between a macro-

molecule, like a protein or an RNA molecule, and a smaller ligand leads to the formation a complex system characterized by specific binding properties. First, the interaction between a receptor macromolecule and a ligand is not a static encounter. Furthermore, conformational changes in all interacting partners influence both the thermodynamic and kinetic properties of the binding process [257]. In the following paragraphs, I will discuss the main theoretical models of molecular recognition and its thermodynamic and kinetic character.

Models of biomolecules recognition. Historically, two models have been proposed to describe the process of molecular recognition: the "lock-and-key" [258] and the "induced fit" [259] models. The "lock-and-key" model considers the binding entities as rigid structures characterized by a pre-existing shape complementarity. On the other hand, the "induced fit" model posits that the receptor can adjust its shape to fit the substrate upon binding, acknowledging the flexibility of the recognition process. However, more recent experimental evidence has led to a deeper understanding of recognition processes, giving rise to the conformational selection model [257]. This model builds on the thermodynamic understanding that the receptor, in its unbound state, undergoes a dynamic equilibrium of various conformations, including those found in the bound state. The interaction with a ligand selects the most favorable conformation for its binding. The common current opinion of structural biologists is that the initial recognition is based on the conformational selection model and the target and ligand molecules undergo smaller local rearrangements induced by the binding [257].

Thermodynamic and kinetic properties of binding. The binding affinity of a molecular complex is generally measured by the equilibrium constant of dissociation K_d, which indicates the propensity to reversibly separate in its individual components. Given a receptor R and a ligand L in an aqueous environment, their K_d can be expressed as:

$$K_d = \frac{[R][L]}{[RL]} \tag{1.11}$$

that is the ratio between the product of the receptor and ligand equilibrium concentrations, respectively $[R]$ and $[L]$, and the equilibrium concentration of their complex structure $[RL]$. However, K_d is determined by the underlying thermodynamics of the system. Assuming the reversibility of the binding, the dissociation constant K_d is determined by the difference in Gibbs free energy ΔG between the bound and unbound states. By assuming a standard state concentration of $C^\circ = 1\ mol/liter$, the relation between the two quantities reads:

$$\Delta G = k_B T\ \ln(\frac{K_d}{C^\circ}) \tag{1.12}$$

where T is the temperature of the system, and k_B is the Boltzmann constant. Since the Gibbs free energy ΔG can be expressed in terms the enthalpy ΔH and entropy ΔS as $\Delta G = \Delta H - T\Delta S$, the dissociation constant can finally be expressed as:

$$K_d = \frac{e^{\frac{\Delta H}{k_B T}}}{e^{\frac{\Delta S}{k_B}}} \tag{1.13}$$

which clearly shows the role of the enthalpic and entropic contribution to the binding. The enthalpic contribution is mainly determined by non-covalent interactions between receptor and ligand, such as hydrogen bonds, electrostatic interactions, and Van der Waals forces. The entropic contribution is instead related to changes in the degrees of freedom of the system upon binding. The intrinsic unfavorable entropic contribution due to the structural constraints forming in the bound state of a ligand-receptor complex is often balanced by the formation of additional intermolecular contacts, which cause a favorable enthalpic contribution and increase the overall affinity.
The insights into the differences between the unbound and bound thermodynamic states do not by themselves reveal the process of molecular recognition. A further characterization is obtained with the kinetic properties of the binding, describing how quickly and efficiently interactions occur and disengage. The binding affinity can be expressed as:

$$K_d = \frac{K_{off}}{K_{on}} \tag{1.14}$$

where k_{on} is the rate at which the ligand binds to the receptor per unit concentration, while k_{off} is the rate at which the complex dissociates.

Insights from thermodynamics and kinetics of RNA - small molecule binding

The discussed properties of molecular recognition, initially acquired by studying protein receptors, were shown to apply accurately even in the case of RNA molecules [109, 127]. In particular, the conformational selection model has been experimentally validated for the RNA-ligand recognition, demonstrating how conformations may individually engage specific and different interactions with the binding partner [109, 127, 260, 261]. In this context, a keen understanding of the thermodynamics and kinetics character of the mechanisms driving molecular recognition may be crucial in the discovery and development of compounds targeting RNA. First, from a thermodynamics perspective, the enthalpic and entropic contributions to the binding free energy (Eq. (1.13)) can be distinctly informative of the interaction mechanism and drive the design of novel compounds. For instance, the binding of coralyne to double-stranded poly(A) RNA has been reported to be predominantly enthalpy-driven, thus indicating strong, specific interactions characteristic of flat aromatic compounds [262]. Conversely, the selectivity of compound B-12 for RNA octaloops is largely attributed to an entropic gain, which is likely due to the release of cations upon binding [263]. Furthermore, from a kinetic perspective, determining the binding rates (Eq. (1.14)) may be crucial to identify competitive mechanisms introduced by RNA binders [264]. A relevant example is provided by the 2H-4 small molecule targeting the $r(CUG)_{10}$ trinucleotide RNA repeat: this compound binds the target with a faster k_{on} with respect to the alternative splicing regulator muscleblind-like 1 protein (MBNL1) involved in myotonic dystrophy type 1. The insights gained from the thermodynamics and kinetics of RNA-small molecule interactions provide a valuable foundation for the overall success rate and efficiency of the drug discovery process. The characterization of the different contributions to the binding process may inform decisions related to target selection, screening strategies, and hit identification, offering a comprehensive approach to the rational design of small molecules targeting RNA.

Leveraging the exploration of the RNA conformational landscape

RNA structural dynamics can be described by a wide conformational landscape dotted with many hierarchical free-energy minima, each one populated with a certain probability (Sec. 1.4.2). In the following paragraphs, I will progressively show how the accurate exploration of the conformational space of a target RNA can be seen as a unique therapeutic opportunity.

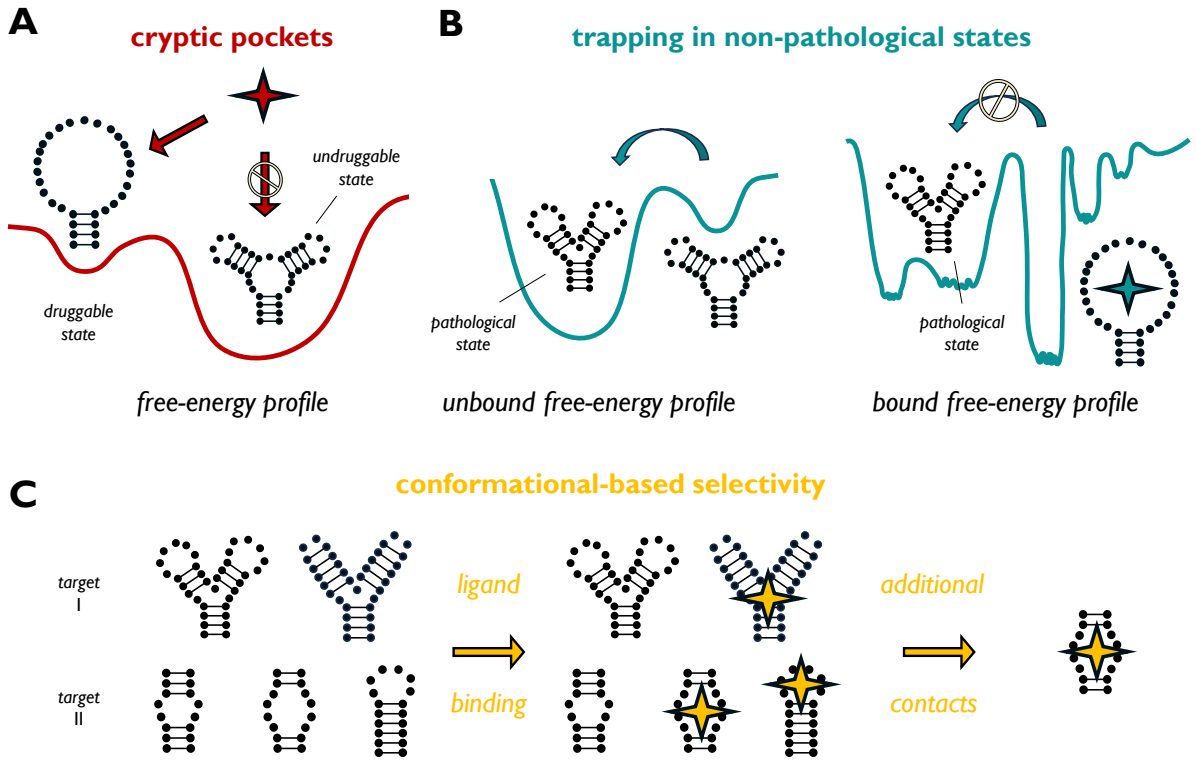


Figure 1.11: The therapeutic opportunities introduced by the flexibility of RNA molecules. A) The two-state free-energy profile of a generic RNA molecule. The metastable state (left basin) is characterized by the presence of a druggable pocket that is not found in the native structure of the molecule (right basin). **B)** The diagram reporting the free-energy profile of a generic RNA molecule in unbound (left panel) and bound (right panel) conformations. The ligand is represented by a cyan star. **C)** The molecular recognition between a generic aminoglycoside (red star) and two target RNAs structure: *i)* the target RNA exists in a dynamic equilibrium of conformations (left panel), *ii)* the presence of the ligand increases the population of only a subset of the conformations populated by the RNAs (center panel), and *iii)* the formation of additional contacts finally increase the affinity of the binding only with one target (right panel).

Cryptic pockets. In the context of high-affinity binding with small molecules, pockets that are well-suited for such interactions may only form in conformational states with marginal populations. Characterizing these states poses inherent challenges (Sec. 1.4.4), yet their presence expands targeting opportunities: a small molecule ligand could only bind to a meta-stable conformation within the ensemble, emphasizing the importance of exploring these less populated states (Fig. 1.11A). An interesting example is given by the work of Varani and collaborators, which characterized and stabilized an unstable conformation of HIV-1 TAR RNA and discovered a small-molecule binding site close to the HIV-1 TAR active site [265]. Subsequently, the authors identified hit compounds binding in this region with a surprising level of selectivity and provided a promising foundation for

the development of inhibitors capable of inhibiting HIV-1 TAR viral activity [48]. Importantly, the identified druggable binding site was not forming in the most stable and accessible conformation of HIV-1 TAR. This result exemplifies how the characterization of meta-stable conformations may be decisive in identifying such cryptic yet druggable pockets.

Modulating RNA thermodynamics. In the thermodynamic framework, the binding of RNA with a ligand is a phenomenon that alters the probabilities of certain conformations and their transition rates, eventually resulting in the stabilization of the bound conformation. The meta-stable conformation where a cryptic pocket is found may become predominant upon binding. This is particularly evident in the case of riboswitches, which assume, upon binding with their cognate metabolite, a conformation that causes the enhancement or suppression of gene expression [35]. An emerging therapeutic strategy is thus to characterize transient and non-functional conformations of pathological RNAs and trap them in such conformations by increasing their thermodynamic stability upon small-molecule binding [109, 266, 267] (Fig. 1.11B). In an important example, Al-Hashimi and co-workers studied the effects of the stabilization of an excited conformation of HIV-1 TAR RNA, which has a poor binding affinity with Tat protein partner responsible for viral replication [120]. Performing virtual screening on an NMR-informed structural ensemble of HIV-1 TAR (Sec. 1.5.5), they subsequently identified a compound that stabilizes the aforementioned conformation with moderate affinity [111]. In a recent follow-up work, the groups led by Al-Hashimi and Hargrove identified the compound DMA-169. This molecule demonstrated the capability to impede the viral activity of HIV-1 TAR by sequestering nucleotide residues essential for recognition with the Tat protein.

Conformational-based selectivity of RNA binders. The mentioned therapeutic strategy is effective because a compound may engage in specific interactions only in the meta-stable conformation. Such phenomenon may be extremely useful toward the development of RNA-targeted therapeutic agents. A representative example is given by the class of aminoglycosides antibiotics, widely used in therapeutics despite their promiscuity in RNA target binding [83]. Indeed, these highly polar molecules achieve selectivity with a specific RNA conformation through the formation of extra intermolecular contacts that enhance binding affinity after the initial recognition [80, 268]. Unbound RNA molecules populate a dynamic ensemble of conformations (Sec. 1.4.2, left panel in Fig. 1.11C). According to the "conformational selection" thermodynamic model (Sec. 1.5.2), the presence of the ligand first selects a subset of the conformations populated by the target RNA. This phenomenon depends on the formation of enough favorable molecular interactions. Due to the high positive charge of aminoglycosides, such favorable interactions may be engaged with different RNA targets in different conformations (center panel in Fig. 1.11C). However, subsequent structural adjustments refine the binding site structure and the formation of additional contacts enhances the selectivity of the binding to a single target (right panel in Fig. 1.11C). The discussed mechanism of recognition has been documented, for example, in the case of the binding between the aminoglycoside tobramycin with Asp tRNA, which constitutes a competitive inhibitor mechanism to the initiation of bacterial translation [269]. Additionally, if a compound lacks sufficient selectivity for a specific RNA target, it can be intentionally improved by conjugating it with moieties that are able to establish specific interactions with the target [270]. Such powerful strategy has been employed,

for example, in the modification of neomycin aminoglycoside to bind the HIV-1 TAR RNA [271].

In view of the concepts discussed in this section, it is possible to propose a paradigm shift from viewing RNA dynamics only as an obstacle to harnessing this property towards the identification of therapeutic agents able to engage selective recognition with the target. The subsequent section aims to elucidate the constraints of experimental approaches in addressing these opportunities and introduces the potential contribution of computational methods in achieving this goal.

1.5.3 Successes and challenges of the experimental techniques employed in the search of RNA drugs

Many experimental techniques are employed in drug discovery. A subset of the methodologies used in RNA drug discovery have been initially developed for protein targets over the past 50 years. On the other side, along with the progress in technology and the increasing importance of RNA as a therapeutic target, RNA-specific methodologies have been developed. Most of the RNA-targeted compounds identified so far rely on these experimental techniques. While such approaches are fundamental in the identification and validation of new therapeutic agents, they often overlook the dynamic character of RNA-small molecule recognition. Oppositely, CADD tools hold the potential to fully address the inherent flexibility of RNA molecules and may be crucial toward the development of effective therapeutic strategies.

In this section, I will first overview the state-of-the-art experimental approaches used in RNA drug discovery. Following the subdivision of the drug discovery pipeline presented in Sec. 1.5.1, this overview is separated into methods to identify small molecules targeting RNA and methods to validate hit compounds. Finally, I will discuss the main limitations of such approaches and introduce the importance of structure-based rational design in RNA drug discovery.

Experimental methods to identify small molecules targeting RNA

The search for RNA-targeted compounds generally relies on large-scale screening experiments aimed at discriminating potential binders among large libraries. The most widely used techniques are high throughput, fragment-based, DNA-encoded library, and phenotypic screenings. In the following paragraphs, I will briefly overview these technologies and underline their most relevant success in the identification of RNA-targeted compounds.

High throughput screenings. In drug discovery, High-Throughput Screening (HTS) rapidly tests the biological activity of a large number of sample compounds on a given target to identify potential lead candidates [272]. Such powerful procedures are extremely costly and require rigorous assay design and advanced statistical analysis of results. Furthermore, their effectiveness depends on the diversity and specificity of the screened chemical libraries. The main experimental techniques that have successfully been used in the HTS of compounds targeting RNA are presented in the following itemized list.

- **Affinity-based mass spectrometry.** Affinity selection mass spectrometry (AS-MS) detects target-ligand interactions by separating bound complexes from unbound ligands using

size-exclusion chromatography, followed by mass spectrometry identification [273]. The main drawback of this method is that it is not suitable for highly flexible molecules. An extension of this method, known as automated ligand identification system (ALIS), is recently gaining importance in the context of RNA targeting [274]. ALIS screenings led to the discovery of ribocil, a synthetic ligand targeting the FMN riboswitch (Sec. 1.3.3).

- **Micro-arrays.** Small Molecule Microarray (SMM) screening immobilizes small molecules on surfaces like glass slides and assesses interactions with a biological target using fluorescence-based detection [275]. Its main drawback is that the immobilized ligand may not represent its true behavior in solution. The Disney lab developed an extension of this technique, called the two-dimensional combinatorial screen (2DCS) [276]. In 2DCS, the secondary structure of a given RNA target is predicted from its sequence and each loop structural motif is compared with an internal library of known RNA motif–small molecule binding partners. 2DCS is integrated into the InfoRNA platform [255] (Sec. 1.5.6) and led, among the others, to the identification of pre-miRNA96 binders that inhibit its biogenesis and induce apoptosis of cancer cells [277].

- **Fluorescence-based assays.** Fluorescence-based assays, such as FRET-based assays and Fluorescent Indicator Displacement (FID), measure changes in fluorescence upon small molecule binding to RNA [278]. A more sensitive extension of this method combines time-resolved fluorimetry with FRET assays to reduce fluorescence background signal and has been successfully used to identify inhibitors for myotonic dystrophy type 1 (DM1) [276]. Recent studies showed that FID may be useful in identifying inhibitors of enterovirus 71 (EV71) viral translation and replication [279].

Fragment-based screenings. HTS methods were the primary hit discovery approach until the 2000s, screening millions of $\sim$ 500 Da compounds seeking nanomolar affinities. However, HTS faced the challenges introduced by the anionic nature of RNA targets, often yielding few hits or false positives [280]. Together with the increased focus on RNA as a pharmaceutical target, fragment-based drug design (FBDD) methods emerged as an alternative approach [281]. By screening smaller libraries of low-molecular-weight compounds (approximately 200 Da), FBDD reduces the costs while covering a similar or broader chemical space than HTS [282]. FBDD is particularly advantageous for RNA targeting since it can focus on neutral ligand scaffolds, offering higher specificity with respect to the RNA binders screened in HTS, composed of multiple charged groups [283]. However, FBDD often uncovers fragments characterized by low binding affinities, necessitating rigorous optimization to transform them into potent and effective lead compounds in the drug discovery process. FBDD has successfully identified novel ligands for the *E. coli* thiamine pyrophosphate (TPP) *thiM* riboswitch [284] and small molecules binding to TERRA lncRNA [285].

DNA-encoded library screenings. A DNA-encoded library (DEL) is a collection of small molecules, each tagged with a unique DNA segment that encodes the structure of the attached compound [286]. In DEL screening, compounds are exposed to a target, and high-affinity binders are identified by sequencing the attached DNA tags [287]. This technique offers a rapid, cost-effective alternative to HTS, allowing the screening of vast compound libraries. However, DEL can

yield false positives due to DNA tags non-specifically binding to RNAs. Recently, DEL was combined with 2DCS to screen around 70000 small molecules against about 4000 RNAs, leading to the discovery of pre-miRNA27a, a compound with potential therapeutic effects targets in breast cancer [288].

Phenotypic screenings. Phenotypic drug discovery (PDD), unlike previously mentioned methods that are "target-centric", is "target-agnostic" and screens biological pathways (like alternative pre-mRNA splicing or bacterial growth) to identify molecules that induce a desired, potentially therapeutic, phenotype[289]. In a typical experiment, a compound library is screened with high-throughput methods on big biological samples, such as cells or tissues. The phenotypic effects and the involved targets are assessed using a variety of assays, including high-content imaging systems to capture the intricate details of cellular morphology. Phenotypic screenings allow for targeting diseases even when the knowledge about the biological target is limited, identifying multi-target compounds, and have generally a low false positive rates [290]. However, phenotypic screenings incur higher costs, both economically and in terms of time, due to the iterative application of HTS methods. Importantly, PDD has led to significant advances in RNA-targeted therapeutics. In 2014, PTC Therapeutics and Hoffmann-LaRoche used PDD to identify SMN2 splicing modifiers in a mouse model of spinal muscular atrophy (SMA) [291], successively leading to the discovery of risdiplam [88]. PTC-Roche and Novartis identified potential SMA therapeutics, such as branapalm, which is now in clinical trials [292]. In 2015, Merck's PDD against *E. coli* led to the discovery of ribocil, an FMN riboswitch-binding molecule suppressing bacterial activity [86].

Experimental methods to validate candidate compounds

During hit validation, a variety of methods are employed to confirm and characterize the binding of small molecules to their RNA targets. The resolution of RNA-small molecules complexes by either X-ray crystallography, NMR spectroscopy, or cryo-EM (Sec. 1.5.3), constitutes an important validation and allows characterizing the molecular interactions at high resolution. Moreover, considering the dynamic nature of RNA-ligand recognition (Sec. 1.5.2), there is a growing reliance on experimental techniques capable of providing thermodynamic and kinetic insights into the binding process [56, 256]. In the following paragraphs, I will provide a brief overview of the most important experimental techniques used to validate the small molecules that have been identified to target RNA.

Isothermal Titration Calorimetry. Isothermal titration calorimetry (ITC) is a technique that measures the heat change associated with molecular interactions, providing direct insights into the thermodynamics of binding between small molecules and RNA targets [293]. This technique is valuable for understanding binding affinities and enthalpy changes and it offers a quantitative approach to characterize interaction strengths (Eq. (1.12)). While ITC is a reliable method, it necessitates large amounts of samples and requires that the small-molecule compound is highly soluble.

Microscale Thermophoresis. Microscale thermophoresis (MST) is a technique to quantify the thermodynamics of the binding by observing the movement of a labeled target RNA molecule in response to temperature changes and at varying concentrations of a ligand [294]. MST is a versatile

and emerging technique in the context of RNA-small molecule interactions since it does not require target immobilization. However, it is less sensitive than ITC and the sample preparation may alter small molecules binding properties.

Surface Plasmon Resonance. Surface plasmon resonance (SPR) is an optical technique for measuring the binding of molecules to a surface, commonly used for real-time, label-free analysis of the interaction between small molecules and RNA [295]. SPR provides kinetic data, including association and dissociation rates (Eq. (1.14)), thus enabling a detailed understanding of the binding process. SPR is sensitive and efficient with minimal sample needs, but it requires stable target immobilization and significant expertise for successful execution.

Biolayer Interferometry. Another optical technique for studying biomolecular interactions is Biolayer interferometry (BLI), which measures changes in the thickness of a biological layer on a biosensor as molecules bind to or dissociate from the surface [296]. This method is particularly useful for analyzing the kinetics of small molecule-RNA interactions. Compared to SPR, the increasing popularity of BLI stems from its cost-effectiveness, user-friendliness, higher throughput, and reduced technical complexity.

Structure-based rational design of RNA drugs

The experimental techniques introduced in the previous section are crucial in the research of drugs targeting RNA and they serve as primary methods for the initial identification of candidate compounds, their validation, and subsequent optimization. However, the majority of the presented approaches present several limitations. Except for specific techniques able to capture the thermodynamic and kinetic properties of the binding, the biophysical characterization arising from such experiments often consists solely of the evaluation of binding affinities [297]. In particular, most of the screening experiments are conducted with no insights into the properties of the binding and proceed by multiple trial-and-error cycles. For instance, the binding site of risdiplam, which was discovered through phenotypic experiments [88], was identified and characterized years after its discovery [91]. The exact mechanism of the molecular recognition of risdiplam with SMN2 pre-mRNA is still not fully understood [298].
Computer-Aided Drug Design (CADD) methods hold the potential for a detailed characterization of the behavior of the target RNA in solution and of its binding with small molecules. Indeed, CADD enables distinguishing and studying the different conformations explored by RNA targets where small molecules can bind with atomistic resolution. In particular, tools based on MD simulations are the equivalent of single-molecule experiments and can follow the evolution of a given system enabling a full exploration of the dynamic properties of molecular recognition (Sec. 1.4.4). For these reasons, CADD approaches inherently address the challenges and capitalize on the opportunities presented by RNA flexibility (Sec. 1.5.2). The prediction accessible with computational methods may inform the experimental techniques and facilitate the rational design of small molecules targeting RNA [299, 300].
Within the context of this thesis, it is interesting to focus on the computational techniques that aid RNA drug discovery by answering the following two fundamental questions [301]:

- In the cellular environment, does the RNA target explore conformations with binding pockets accessible to small molecules?
- Is it possible to identify or design a compound that binds the target RNA with high affinity, selectivity and specificity?

In the following section, I will give a detailed overview of the state-of-the-art computational tools that address these two questions.

1.5.4 *In silico* identification of RNA-small molecules binding sites

The first question introduced at the end of Sec. 1.5.3 and concerning the identification of druggable conformations of the target RNA can be addressed by binding site detection tools (Fig. 1.10). This operation consists in the accurate location of favorable regions for interaction with small molecules on the structure of a given RNA target, previously determined by experimental or computational techniques. Simultaneously with the growing interest in RNA as a therapeutic target in recent decades, a range of computational tools initially designed for proteins has been tailored for RNA molecules or developed specifically to identify binding sites on RNA. By focusing on their ability to capture the inherent flexibility of RNA, the available tools can be broadly classified into the categories of single structure-based, dynamics-based, and network and machine learning-based methods [301, 302].
In this section, I will delve into the operational characteristics of each category of binding site detection tools, providing insights through commentary on representative examples. The discussion will address the general advantages and limitations of the categories. To conclude the section, I will discuss and compare the state-of-the-art for RNA molecules, aiming to establish guidelines for selecting the most suitable software. The reader can find a comprehensive list of available software for binding site detection on RNA molecules in Tab. 1.1.

Single structure-based methods

These methods make use of algorithms that are implemented to detect binding sites on a single static structure of the target biomolecule. Historically, the operation of binding site detection was first implemented by the GRID software developed by P. Goodford [303]. Since then, many computational tools have been built starting from the GRID algorithm. From a technical viewpoint, the latter is composed of the following essential steps.

1. **Subdivision in a spatial grid.** The 3D space that embeds the target molecule is subdivided by the definition of a spatial grid.
2. **Identification of target surface.** The grid voxels that are not buried enough within the target structure are filtered out from subsequent calculations.
3. **Identification of cavities.** Each surface voxel undergoes systematic evaluation to determine the presence of favorable characteristics for ligand binding. Generally, small probes are rolled along the surface to study the target shape complementarity by means of geometrical criteria, and the likeness of the binding by means of energetic calculations;

4. **Clustering of interacting regions.** Neighbor voxels that were classified as potentially interacting are clustered together and define the identified binding site. The nature of the probes that have been clustered in a given binding pocket is leveraged to infer its physico-chemical characteristics, in terms of properties such as volume, buriedness, hydrophobicity, and hydrogen bonding donor/acceptor character.

	Name	Year	Reference	Availability
Single structure based	PocketFinder	2005	[304]	Commercial
	SiteMap	2009	[305]	Commercial
	POCASA	2010	[306]	Webserver
	3V	2010	[307]	Webserver
	AutoSite	2016	[308]	Download
	mkgridXf	2019	[309]	Free for academic use
Network or knowledge based	Rsite	2015	[310]	Download
	Rsite2	2016	[311]	Download
	Rbinds	2020	[312]	Webserver
	RNAsite	2021	[313]	Webserver
	BiteNet	2021	[314]	Webserver
	RLBind	2023	[315]	Webserver
Dynamics based	SILCS-RNA	2022	[204]	Free for academic use

Table 1.1: State-of-the-art of binding site detection tools for RNA. A comprehensive list of available computational tools for binding site detection on RNA molecules. The listed tools are grouped by the corresponding category (Sec. 1.5.4), which is reported on the left. Tools specifically developed for RNA have a light khaki background. The "Download" and "Webserver" flags in the Availability are hyperlinks addressing the corresponding website.

Two widely used computational tools that belong to this category and are also available for RNA molecules are SiteMap [305] and PocketFinder [304], implemented in the commercial packages of Schrödinger Maestro Suite and ICM Molsoft, respectively. In SiteMap, the surface voxels are identified by evaluating the fraction of the surrounding space that is occupied by the receptor. Then, the van der Waals interaction energy is calculated *via* a Lennard-Jones potential applied to selected probe atoms. The binding site identification is carried out similarly in PocketFinder with the additional filtering of clustered regions with an estimated volume smaller than a given threshold. Single structure-based computational tools are known for their remarkable speed. However, they also exhibit two crucial limitations. First, they operate on a single static conformation of the target molecule, neglecting the crucial dynamic nature of RNA interactions. Secondly, the impact of aqueous solvation is often overlooked, while it is indispensable for the realistic modeling of RNA binding events. Due to the intrinsic constraints of single structure-based methods, no tool within this category has been explicitly tailored for RNA molecules (Tab. 1.1).

Dynamics-based methods

To overcome the limitations of single structure-based methods, a newer class of tools leverages molecular dynamics (MD) simulations directly. One approach involves predicting binding sites *a posteriori* from pre-existing MD trajectories, exemplified by the well-established PyMol plug-in CAVER [316]. Alternatively, recent developments focus on real-time exploration of pocket evolution during MD simulations, like MDPocket by Barril's group [317]. From a technical viewpoint, the mentioned simulations-based methods are equivalent to single structure-based ones. Indeed, the binding site identification is carried out with similar algorithms. At the same time, they possess the unique ability to capture the flexibility of the target by considering multiple conformations generated throughout an MD trajectory. Moreover, the effects of aqueous solvation can be explicitly accounted for in MD simulations, enhancing the accuracy of the predictions. However, the higher accuracy of simulations-based methods comes at the expense of a significantly higher computational time.
A further development of dynamics-based methods is achieved by the more recent mixed-solvent techniques [302, 318–320]. In these methods, small fragments are directly introduced in the simulated system to explore the surface of the target molecule and characterize favorable interaction hotspots by analyzing their thermodynamic behavior. While being time-consuming, such techniques hold significant promise in simultaneously addressing the two key questions enunciated at the end of Sec. 1.5.3. First, they identify small molecule binding sites on a given RNA target, allowing for its full flexibility and accounting for the potential role of ions and solvent effect in ligand recognition. Second, they can be regarded as the computational analog of fragment-based screenings (Sec.1.5.3), potentially giving insights on preferred ligand interactions that may inform the design of potential RNA binders [321, 322]. A representative example of mixed-solvent techniques is given by the SILCS-RNA approach developed by Mackerrel's group [204]. In SILCS-RNA, Gran Canonical Monte Carlo - MD (GCMC-MD) simulations are used to sample the conformations of the RNA in the presence of different small compounds, each one representing a particular flavor of interaction. The most favorable binding hotspots are defined in terms of the occupancy frequency of the small compounds on given subregions of the target RNA surface. The final output of SILCS-RNA consists of a set of binding affinity maps ("FragMaps") for each probe compound used in the simulations.

Network and machine learning-based methods

A more recent class of tools for RNA binding site detection models the RNA–ligand interaction as a network of contacting atoms. Many tools have been developed in this framework, varying based on the specific network representation chosen for their depiction. Two pioneering examples in the realm of RNA drug discovery include RSite [310] and RSite2 [311]. These tools utilize inter-nucleotide Euclidean distance networks, derived from 3D or 2D structures respectively, to predict functional sites for RNA–ligand binding by identifying maximally closely clustered nucleotides. However, Rsite and Rsite2 inter-nucleotide networks lack the ability to distinguish various connection types between nucleotides, resulting in frequent false positive predictions [301]. To overcome this issue, the newer RBind transforms RNA structures into graphs, representing nucleotides as nodes and non-covalent contacts as edges and outperforms both RSite and RSite2 [312, 323].
On the other hand, the newest class of binding site detection methods has been developed with

the recent improvements in machine-learning models. These approaches rely on the training from existing structural data. A first representative example is RNASite, which is based on a random-forest model trained on a set of 60 RNA-ligand structures [313]. By extracting various features for each nucleotide, including geometric, topological, and evolutionary ones, RNASite predicts whether a given nucleotide belongs to the functional sites. The quality of the prediction of RNASite was benchmarked against RSite and RSite2, providing significantly better results. However, the state-of-the-art of machine-learning binding site detection is constituted by deep-learning models [324]. A representative tool is BiteNet$_N$, which outperformed all the mentioned tools of this category [314]. One of the key ingredients of BiteNet$_N$ is its training set, which includes both RNA and DNA-ligand complexes and considers NMR models and X-ray co-crystals as independent entries, for a total of $\sim$ 2000 structures. Given an RNA target, BiteNet$_N$ builds a voxel-based representation associating each portion of the space to eight different atomic densities of a particular type. These voxelized representations are then fed to a 3D convolutional neural network (cNN) that scores segments in nucleic acid structures concerning the binding sites. As output, BiteNet$_N$ provides the coordinates of binding site interface centers, the probability scores for each center, and scores for each nucleotide in a binding site.

In general, network- and machine learning- based methods have the great advantage of being fast. Moreover, they intrinsically do not suffer from the inaccuracies of methods based on physical assumptions. In particular, the newer deep learning models have the potential to discern intricate patterns in the input data and make predictions without relying on predefined rules specific to the network structure. However, these tools strongly depend on the availability of training sets of structural data and their accuracy is highly dependent on the choice of the representation model.

State-of-the-art of binding site detection for RNA

Currently, there is no comprehensive evaluation of the performances of computational tools for RNA binding site detection across different classes of methods. Existing benchmark studies, as mentioned, concern only the network and machine-learning based tools and were conducted on test sets with relatively small sizes. This absence seems to be largely due to the heterogeneity of the algorithms employed and, as a consequence, to the arbitrary output that software may have to match a given definition of the binding site. However, it is possible to trace some guidelines that may help in tailoring the choice of a suitable computational tool for a given application.

From a broad perspective, single structure-based methods constitute a unique choice to assess in a reasonable time the physico-chemistry of RNA binding pockets. In this sense, PocketFinder [304] has been used by Schneekloth's group in one of the unique systematic assessments of the physicochemical properties of resolved RNA structures [79] (Sec. 1.3.3). While this study provided valuable insights into the structural characterization of RNA-small molecule pockets, it is important to underline two main issues. First, the intrinsic overlooking of the dynamic nature of molecular recognition is not an optimal choice for RNA molecules. Second, all the available tools were initially developed for protein targets and rely on threshold parameters that may not be optimized for RNA molecules. Therefore, while single structure-based methods may be useful for big and stable RNA molecules, they seem inappropriate for highly flexible RNA targets if their conformational space has not been explored previously.

Simulations-based methods start from overcoming the challenges of single structure-based tools However, the development of simulations-based methods for RNA remains limited. Being a unique example, SILCS-RNA [204] represents a milestone in the context of RNA CADD. However, this tool presents relevant limitations. The first and most important one consists of the limited accounting for the flexibility of the RNA target molecule. Indeed, to facilitate the thermodynamics calculations for the binding affinity, the RNA is restrained to its initial configuration during the entire MD simulation. While SILCS-RNA is able to capture the dynamic nature of the local recognition between the target RNA and ligand as well as the solvation effects, it is unable to sample globally different conformations. As observed before, binding pockets may be hidden in metastable conformations and the exploration of the conformational landscape of the target RNA is crucial for drug-discovery purposes (Sec. 1.5.2). Furthermore, oppositely to common CADD tools which quantitatively provide a ranked list of the identified pockets, the final output of SILCS-RNA consists of a set of affinity maps defined on the surface of the whole RNA. As highlighted by the authors themselves, subsequent molecular docking or virtual screening applications require a visual inspection of the results to identify the most probable interacting sites.

Finally, the development of machine learning, and in particular deep learning, methods represents one of the most important alternatives for binding site detection on RNA molecules. However, the limited available structural knowledge of RNA binding sites is dramatically hampering the development of the field. Most importantly, it is currently unclear whether the future higher number of resolved structures in the training sets of machine learning models would be enough to account for the flexibility of RNA molecules [325–327]. At least until the number of RNA-ligand structures became comparable to the one of proteins, the unique category of tools that fully address the flexible nature of RNA is the category of dynamics-based tools. In this perspective, the future implementation of machine-learning algorithms into MD simulations represents a promising approach to comprehensively and effectively support the discovery of small molecules targeting RNA [324, 328].

1.5.5 *In silico* identification small molecules binding RNA

The second question raised at the end of section 1.5.3 concerned the identification and/or design of potential RNA binders for a given RNA target. In theory, this may be carried out by multiple kinds of structure-based computational tools. The first are computational tools for RNA structure prediction. However, the computational prediction of quaternary complex structures is still in its nascent stages for proteins [329] and has not been developed for RNA molecules. From a broader perspective, MD simulations have the capability to predict the atomistic structure of a given RNA target in a complex with a ligand. However, ligand binding is a complex process occurring at extended timescales, necessitating the application of enhanced sampling techniques. The computational cost associated with such techniques renders MD simulations impractical for predicting RNA-ligand complex structures in the context of drug design. In practice, the most expeditious and widely utilized method for identifying RNA-small molecule binding modes is virtual screening, which relies on the systematic application of molecular docking.
From a technical viewpoint, molecular docking requires as inputs the independent structures of the

	Name	Year	Reference	Availability
stochastic optimization	ICM	1994	[331]	Commercial
	GOLD	1997	[332]	Commercial
	AutoDock	1998	[333]	Download
	RiboDock	2004	[334]	Download
	Glide	2004	[335]	Commercial
	PLANTS	2007	[336]	Download
	FITTED	2007	[337]	Commercial
	AutoDock VINA	2010	[338]	Download
	rDock	2014	[339]	Download
Incremental construction	Surflex-Dock	2003	[340]	Commercial
	DOCK 6	2009	[341]	Download
Molecuar dynamics	MORDOR	2008	[342]	Freefor academic use
Multiconformer docking	RLDOCK	2021	[343]	Download
	NLDock	2023	[344]	Download

Table 1.2: State-of-the-art of molecular docking engines for RNA. A comprehensive list of available computational tools for molecular docking of binding site detection on RNA molecules. The listed tools are grouped by the algorithm used for the sampling of the ligand conformational space (Molecular docking step I: Sec. 1.5.5), which is reported on the left. Tools specifically developed for RNA have a light khaki background. The "Download" flag in the Availability are hyperlinks addressing the corresponding website.

target RNA and a ligand. Then, the docking operation is composed of two independent steps that I will first discuss in this section. For each of the two steps, I will overview the category of tools that are currently available for the molecular docking of RNA. At the end of the section, I will present the results of existing benchmark analyses and have a final discussion in order to trace some guidelines in the choice of computational tools for molecular docking.

Molecular docking step I: sampling RNA-ligand conformational space

The primary objective of molecular docking is to predict the optimal binding position and orientation of receptor and ligand molecules. This task presents several challenges, encompassing concerns related to biological reliability and computational expenses [330]. A main obstacle is constituted by the dynamic nature of molecular recognition and the flexibility of both the RNA receptor and ligand molecules (Sec. 1.5.2). In the following two paragraphs, I will discuss the details of the docking algorithms that have been implemented to account for the flexibility of both receptor and target molecules. Then, I will discuss the strategies employed by docking algorithms to account for the effects of solvation, which are fundamental for the accurate description of the RNA-small molecule interactions. The reader can find a comprehensive list of available software for molecular docking on RNA molecules in Tab. 1.2.

Incorporating the ligand flexibility. A first approach to account for the dynamic nature of molecular recognition within docking algorithms is to incorporate the flexibility of the ligand [301, 345]. Different algorithms have been implemented to this end and they can be classified in the following listed categories.

- **stochastic optimization.** This method involves random or probabilistic algorithms to sample the ligand conformational space, like genetic algorithms, simulated annealing, or Monte Carlo. A representative example of this method is given by rDock [339] developed by the Barill and Morley groups. RDock generates ligand poses using a multi-stage process that includes Genetic Algorithm search for initial pose sampling, followed by Monte Carlo sampling and Simplex minimization [346] to refine the poses into low-energy configurations.

- **incremental construction.** This method is based on the local building of the ligand in the binding site by gradually adding fragments and considering the energy contribution of each of them. A representative example is DOCK 6 [341]. DOCK 6 utilizes an "anchor-and-grow" incremental construction method where the largest rigid portion of the ligand is first oriented in the active site of the target and then flexible parts are added and optimized.

- **multiconformer docking**. This method first generates and evaluates a range of possible conformations of the ligand ("rotamers") to be then accounted for during docking calculations. A representative example of this method is the RLDOCK developed by Chen's group [343]. This tool identifies potential anchor sites on the RNA target based on ligand-provided geometric criteria. Subsequently, it generates a pool of diverse ligand conformers by systematically varying parameters like rotatable bond angles. The generated conformers are clustered according to their root-mean-square deviation (RMSD) and finally docked onto the RNA target.

While multiconformer docking offers a fast choice to consider the ligand flexibility before docking, its performance relies on the quality of the generated conformer ensemble. In contrast, stochastic and incremental sampling methods can treat ligand flexibility during docking. A main drawback of such on-the-fly methods is that small errors in the early steps may be amplified during later calculations. Moreover, stochastic approaches suffer from the issues of the exhaustive sampling of the conformational space already discussed in Sec. 1.4.4.

Incorporating the receptor flexibility. Historically, docking algorithms considered fixed receptor-ligand geometries due to computational limitations. However, recent progress in computational methods has facilitated a shift from the "rigid" docking approach to one that considers molecular flexibility. This transition is particularly crucial for providing accurate descriptions of molecular recognition, especially in the context of RNA molecules. Existing methods fall into the categories itemized in the following [301, 330, 345, 347]:

- **soft docking.** This approach accommodates subtle conformational adjustments by permitting slight steric overlaps between the ligand and the receptor. Subsequently, the regions experiencing clashes undergo energy minimization, potentially resulting in the refinement of both ligand and receptor coordinates to enhance their mutual fit. This technique is suited for binding processes that do not alter significantly the overall structure of the binding site. A representative example of this category is given by Glide [335], which is a software distributed in the commercial Maestro Schrödinger Suite and initially developed for proteins. Glide performs torsional energy minimization and Monte Carlo pose refinement, scaling down the van der Waals radii of specific receptor and/or ligand atoms. This process creates additional space in the binding pocket and therefore allows for more accurate fitting of ligands.

- **fully-flexible docking.** An alternative to soft docking is constituted by explicitly taking into account the flexibility of the receptor during docking by means of MD simulations. An important example of this category is constituted by MORDOR [342]. This method introduces in an MD simulation a driving force that moves the ligand. Starting from a random position around the receptor, the ligand explores its surface by an additional root-mean-square-deviation type of force (Path Exploration With Distance Constraints method [348]). In this way, the ligand is constrained to explore the conformational space following a low-energy pathway. This technique is computationally more demanding, but can provide more accurate prediction of the binding event for very flexible molecules.

- **ensemble docking.** This strategy consists of the iterative application of the docking operation on a previously generated conformational ensemble of the target molecule. An important application of this approach in RNA drug discovery has been realized by Al-Hashimi and co-workers, who performed a virtual screening campaign on an MD-generated conformational ensemble of HIV-1 TAR [111]. Given the high flexibility of HIV-1 TAR molecule, the ensemble approach was in this case crucial to identify a compound able to stabilize a conformation of HIV-1 TAR poorly recognized by the partner protein responsible for viral replication [111, 349].

The selection of an optimal docking strategy to accommodate the flexibility of the RNA receptor hinges on the specific objectives of the study. The majority of available docking software for RNA allows the choice between rigid and soft approaches. Virtual screening campaigns, designed to broadly discriminate among extensive compound libraries, often favor the efficiency of the rigid option. Conversely, during the lead optimization phase of drug discovery, soft molecular docking is preferred for its accuracy. However, the conformational variability introduced by soft docking may not adequately capture substantial conformational changes in the target RNA, which is accounted as populating almost a single conformation. Due to the intrinsic inaccuracy of this approach, full-flexible docking is a great alternative to adequately describe the recognition process, naturally accounting for the induced fit effects of molecular recognition and the kinetics of the binding. Systematic applications of this approach in virtual screening campaigns currently rely on prior filtering of compound libraries carried out with rigid docking. In the presented context, ensemble docking may provide a valuable compromise, as it can easily be carried out in parallel on the multiple conformations of the target molecule. The correct accounting of the target flexibility in ensemble docking heavily depends on the accuracy of the previously generated ensemble. Despite the high potential of this approach for RNA drug discovery, the number of applications of ensemble docking is currently limited.

Accounting solvent effects. In conclusion, the accurate prediction of binding poses for RNA and ligands necessitates accounting for solvent-mediated interactions. The negatively charged sugar-phosphate backbone of RNA induces the accumulation of water molecules and metal ions, mediating interactions with ligands (Sec. 1.1.3). To address this problem, some docking algorithms have been modified to explicitly account for the solvation effect. For instance, AutoDock has been equipped with a new potential function that simulates dynamically bound water molecules to the RNA [350]. However, due to the higher computational cost, most molecular docking software do not explicitly

account for the presence of water and ions. Two main strategies are commonly employed to address this issue. First, molecular simulations with explicit water and/or ions are conducted to refine RNA structures [350, 351]. The positions of important ions can then be retained for subsequent ligand docking. However, this approach faces challenges due to the high sensitivity of ligand-RNA interactions to the positions and orientations of water and ions, which may be inaccurately sampled during simulations. Alternatively, it is possible to predict the binding of water molecules and ions to the RNA prior to docking This approach relies on independent tools, such as the Tightly Binding Ion (TBI) model [352] and the Monte Carlo TBI (MCTBI) model [353], which sample discrete ion distributions. Additionally, 3D-RISM [354] predicts the distribution of both solvent and ions around a macromolecule, and the more recent SPLASH'EM (Solvation Potential Laid around Statistical Hydration on Entire Macromolecules) [355] is a model designed to predict bridging water molecules in nucleic acid–ligand complexes.

Molecular docking step II: scoring the sampled conformations

The second step of molecular docking consists in the assessment of the relative probability of the generated poses by means of a scoring function (SF). The score associated with the generated poses of a given complex structure is aimed to correlate with binding affinity experimental data. Being the two steps of docking independent, many studies focused on the implementation of SFs that perform an additional scoring *a posteriori* of the docking poses found by another tool. Broadly, scoring approaches can be categorized into the categories discussed in the next paragraphs [301, 356]. The reader can find a comprehensive list of available SFs for the molecular docking of RNA molecules in Tab. 1.3.

Physics-based methods. The development of atomistic RNA force fields for MD simulations enabled their application in the context of drug discovery and, in particular, molecular docking. Various docking software rely on such force-field potentials to assess the binding affinity between the target and the ligand depending on their physicochemical interaction potential. While these scoring methods offer accurate insights into the molecular mechanisms of interaction, their computational cost makes them less suitable for large-scale virtual screening. To reduce the computational cost of this approach, RNA atomistic force fields for RNA are often coupled with implicit solvent models. Such models offer an optimal balance between speed and accuracy. A representative tool of this category is DOCK 6 [341, 357], which combines the implicit Generalized Born model augmented with the hydrophobic solvent accessible surface area [358] in combination with the AMBER force field. An alternative, computationally less intensive approach involves energy calculation using a simplified form of the potential, which is composed of a weighted sum of different components of interactions such as van der Waals, electrostatic, and hydrogen bonds. In these empirical approaches, the weighted coefficients of the energetic terms are generally fitted by optimizing the success rate of computational predictions for a training set. Several tools originally developed for proteins and adapted to RNA molecules make use of such scoring functions: the fully empirical scoring function of AutoDock VINA [338], the GoldScore in GOLD [332], and the GlideScore in Glide [335]. Among the more recent tools specifically developed for RNA, a representative example is given by RLDOCK scoring function [343]. The main limitations of these methods consist in the inaccuracy introduced by neglecting the correlations of the different energetic contributions and in the low transferability

of the weight coefficients between different biological systems. Their success highly depends on the quality of the curated training set, which is generally limited for RNA-small molecule interactions due to the limited availability of their complex structures.

	Name	Score type	Year	Reference	Availability
Physics based	GoldScore	Empirical terms	1997	[332]	Commercial
	Surflex	Empirical terms	2003	[340]	Commercial
	GlideScore	Empirical terms	2004	[335]	Commercial
	PLANTS	Empirical terms	2007	[336]	Download
	AutoDock 4	Empirical terms	2007	[359]	Download
	MORDOR	Force fields	2008	[342]	Free for academic use
	DOCK 6	Force fields	2009	[341]	Download
	AutoDock VINA	Empirical terms	2010	[338]	Download
	IMDLScore2	Empirical terms	2012	[360]	-
	rDock	Empirical terms	2014	[339]	Download
	RLDOCK	Empirical terms	2020	[343]	Download
Knowledge based	Kscore	Statistical potentials	2008	[361]	-
	DrugScoreRNA	Statistical potentials	2011	[362]	Free for academic use
	LigandRNa	Statistical potentials	2013	[363]	Deprecated
	SPA-LN	Iterative statistical potentials	2017	[364]	-
	Tbind	Iterative statistical potentials	2018	[365]	-
	ITScore-NL	Iterative statistical potentials	2020	[366]	-
Machine learning based	RFScore-VS	Gradient boosting trees	2017	[367]	Download
	RNAPosers	Random forest	2020	[368]	Download
	AnnapuRNA	k NearestNeighbour	2021	[369]	Download

Table 1.3: State-of-the-art of scoring functions for RNA-ligand molecular docking. A comprehensive list of available SF of RNA-ligand conformations. The listed tools are grouped by their category (Molecular docking step II: Sec. 1.5.5), which is reported on the left. If the scoring function is described in the same publication of the corresponding docking engine, the Tab reports its standard scoring function. Tools specifically developed for RNA have a light khaki background. The "Download" flag in the Availability are hyperlinks addressing the corresponding website. Dash in availability indicates that no information on availability is reported in the referenced publication. Italic text indicates standalone scoring functions.

Knowledge-based methods. A statistical potential approach for evaluating the binding affinity of a target-ligand complex is based on the statistical analysis of known complexes, under the empirical assumption that frequently observed interactions are energetically favorable. Such methods differ from each other by the functional forms of potential energy terms. Generally, pairwise interaction terms are derived by the occurrence frequency of atom pairs in a database using the inverse Boltzmann relation [370]. Built on a previous version for protein receptors, the first example of knowledge-based SF for RNA is by DrugScoreRNA [371]. In addition to the distance-dependent pairwise potential, more complex interactions can be inferred by the relative orientation between different atom pairs. In this direction, LigandRNA SF added a three-body anisotropic poten-

tial term, demonstrating higher accuracy [363]. However, neglecting the many-body correlations between different interaction terms constitutes an important limitation of these approaches. To address this challenge, a typical approach involves iterative refinement of the energy function, until the simulated probability distribution for various atom pairs matches the observed distribution from experimental data. A representative example of this class is ITScore-NL [366], which is the SF implemented in the NLDOCK docking software [344]. Thanks to its iterative nature, ITScore-NL employs statistical potentials that combine atomic pair interactions, nucleobase-ligand stacking, and electrostatics, achieving greater accuracy than other knowledge-based scoring functions. However, despite these improvements, such data-driven approaches suffer from limited structure data of RNAs and RNA-ligand complexes.

Machine learning-based methods The recent advent of machine learning methods revolutionized the field of docking scoring functions in the realm of RNA-small molecules interactions [372]. Oppositely to knowledge-based approaches, such methods have the advantage of relying on multiple trainable parameters that may better leverage the available structural knowledge. A representative example of these methods is RNAPosers [368]. This tool makes use of a set of pose classifiers that can estimate the "nativeness" of a ligand for a given RNA and ligand structure, by means of a random forest method. For this category of tools, the choice of the input features is crucial to distinguish noise from the large structural knowledge embedded in the available data and to avoid an excessively high dimension of the parameters space. In this sense, an optimal engineered feature should maximally simplify the input information while capturing the critical factors that govern RNA-ligand docking outcomes. AnnapuRNA is a recent scoring function that proposed a coarse-grained model for feature engineering. In general, machine-learning methods are prone to overfit the available experimental data. This limitation is more important in the field of RNA drug discovery due to the limited structural knowledge of RNA-ligand complexes.

Performance comparison across RNA-small molecules docking software

Given the multitude of heterogeneous docking software available, determining the preferred choice for RNA-small molecule complexes is challenging. One crucial consideration is the software's ability to accurately predict the correct complex structure among the top-scored ones. A variety of different benchmark studies have been conducted during the last decade in order to assess the quality of available docking software, mainly in coincidence of the release of one specific tool. In this context, it is important to also remark that such benchmark studies generally involve only open-source software. A general trend emerging from these studies revealed that computational tools specifically developed for RNA were shown to outperform the tools originally developed for proteins [301]. NLDOCK, the most recent engine for RNA-ligand docking has been released with its own SF ITSCore-NL and it has been compared to rDock, AutoDock, and DOCK6 using four different test sets, consistently demonstrating superior prediction quality [366]. A second tool that consistently gave good predictions is rDOCK docking engine, coupled with both its own SF and several standalone ones (SPA-LN, RNAPosers, and ITSCore-NL) [301]. In coincidence of its release, AnnapuRNA outperformed other tested SFs, showing that the implemented coarse-grained model successfully captured the core of RNA-ligand interaction data despite the approximation [369].

Besides the success trend arising from the mentioned benchmark analyses, it is important to highlight that all the mentioned studies were conducted on relatively small test sets, with a maximum size of 77 RNA-ligand structures. Their results may therefore not be comprehensively informative of the quality of the available docking software. Very recently, Jiang *et al.* addressed this limitation by conducting a benchmark analysis against 800 RNA and DNA-ligand complexes, the most extensive nucleic-acid test set so far [373]. The benchmark was conducted using rDock, RLDOCK, and several other docking software originally developed for proteins (Surflex-Dock, DOCK 6, AutoDock, AutoDock Vina, and PLANTS). Interestingly, the authors varied the conformation of the input ligand between its experimental pose, a rotated pose, and a randomly generated pose. From their results, they revealed that RLDOCK predictions are not robust upon variation of the input conformation of the ligand. The only RNA-specific tool able to make robust and reliable predictions is rDock [339], suggesting the stochastic algorithm for the sampling of the ligand conformation may capture more reliably than the others the intricacy of nucleic acid-ligand recognition. Surprisingly, PLANTS obtained comparable and in some cases more promising results. NLDOCK was not included in the described analysis due to the unavailability of its code at the moment of its realization, but Jiang *et al.* conducted an additional analysis on four smaller test sets and showed that PLANTS outperforms also NLDOCK.

When evaluating the predictive capabilities of existing software for binding affinity, the field demonstrates diminished promise in contrast to pose identification. In the mentioned study by Jiang *et al.*, which assessed binding affinity predictions across various tools using a dataset comprising 89 RNA and DNA-ligand complexes, overall correlations with experimental data were generally low [373]. Upon variation of the input conformer, the tool that gave averagely the best predictions was rDock with its own SF. However, the tool that had the highest correlation with experimental data was PLANT. In both cases, the SF falls into the category of physics-based ones. Interestingly, such methods showed more reliability than machine learning models trained only on RNA in the prediction of RNA- and DNA- ligand binding affinities. However, before these recent results, the best prediction on a set of 77 RNA-ligand complexes was obtained by the re-scoring of a pool of conformations generated by AutoDock docking engine with the ITScoreNL scoring function [366]. This result is representative of a general trend among the other available benchmark studies conducted against RNA-ligand structures, indicating comparable or superior predictions of machine learning-based SFs in predicting the binding affinity of RNA-small molecules complexes [301].

Local and blind docking

To conclude this section, it is fundamental to mention the difference between local and blind docking. Indeed, molecular docking may be informed by a previously detected binding site ("local" docking) or it may be performed on a target without prior identification of binding hotspots ("blind" docking) [301]. To this end, all computational tools rely on an algorithm of binding site detection. Commercial molecular docking tools generally rely on algorithm for binding site identification which is also available as standalone software, like in the case of SiteMap [305] for Glide [335] and PocketFinder [304] for ICM [331]. Other tools implemented an inner binding site detection method, which generally belongs to the category of single structure-based methods, like in the case of rDock [339] and NLDock [344]. In almost all cases, the binding site detection by molecular docking tools

is carried out on a static depiction of the receptor RNA. As mentioned before, this framework is not the most suitable to address the inherently flexible RNA molecules.
Despite its importance for RNA-targeted drug discovery, the mentioned benchmark literature did not consider the capacity of blind docking, restricting the comparison between different software to their ability to dock a known ligand onto its native partner. Very few studies compared the capabilities of different software in performing blind docking. In coincidence with the release of NLDOCK [344]. From their results, they showed NLDOCK greater ability to correctly identify the unknown binding site with respect to rDock [339], AutoDock [333] and DOCK6 [341]. However, the success rate was always inferior to the 32% for the top-scored prediction and never superior to the 50 % considering all predictions.

1.5.6 Chemical libraries of RNA binders and databases of RNA-small molecule structures

As a compendium to the discussed techniques, efforts to identify RNA-targeting compounds are enhanced by the knowledge of RNA already targeted by small molecules. A fundamental contribution in this sense would be the building of specific chemical libraries of known RNA binders to be used in virtual screening campaigns. Given the importance that RNA targeting acquired in drug discovery, many important pharmaceutical industries started to build their own internal libraries [374]. However, as discussed in Sec. 1.3.3, the chemical space of the RNA binders is highly debated within the scientific community, and this scarce knowledge is hindering the development of RNA-specific libraries. Within this framework, machine-learning models are showing promising potential to leverage experimentally-derived chemical libraries to identify compounds with greater specificity and selectivity for RNA [75].
In a broader context, the creation and maintenance of databases that gather information on RNA-small molecule interactions have become fundamentally important. These repositories, frequently curated through manual or semi-manual processes and enhanced with web interfaces, seek to offer a thorough compilation of compounds with documented interactions with RNA. They include experimental details, binding affinities, and structural data for known RNA-ligand complexes. Moreover, the set of RNA binding partners annotated in a given database constitutes an effective chemical library that can be screened virtually or experimentally in the search for new therapeutic agents. The nature of such databases highly depends on the nature of the data annotated together with the entries and the features that have been implemented, such as web browsing of the database, and screening for similarity in sequence and/or structure. In the following, I will describe three representative examples of RNA-small molecule databases. The reader can find in Tab. 1.4 a comprehensive list of available RNA-small molecules databases.

R-BIND. The R-BIND database curated by Hargrove's group classifies RNA-binding ligands based on their bioactivity in cell culture or animal models, differently from previous efforts that assessed the properties of RNA binders *in vitro* [78, 379]. R-BIND repository reports mainly the physicochemical and pharmacological properties of the ligands, with relatively little structural information on the RNA target. From a general perspective, R-BIND is an example of how such databases are a fundamental tool in the understanding of RNA-ligand interactions, providing insights into the chemical space of RNA binders (1.3.3). More specifically, R-BIND is a widely used

Name	Year	RNA data	ligand data	Origin	Ref	Availability
SMMRNA	2014	-	· physico-chemistry · 2D structure	Literature	[375]	Deprecated
NALDB	2016	· sequence	· physico-chemistry · 2D structure	Literature	[376]	Deprecated
InfoRNA	2018	· 2D structure motif	· physico-chemistry · 2D structure	Literature and experiments	[255]	Free for academic use
R-BIND	2020	· sequence · 2D structure	· physico-chemistry · 2D structure	Literature	[78]	Webserver
RNALigands	2021	· sequence · 2D structure	-	InfoRNA, R-Bind, PDB	[377]	Webserver
RPocket	2021	· sequence · pocket topology	· physico-chemistry	PDB	[378]	Webserver
ROBIN	2022	-	· physico-chemistry	Experiments	[75]	Download

Table 1.4: Available database of RNA-small molecules binding data. A comprehensive list of available databases of RNA-small molecule interactions. The listed tools are grouped by their category (Molecular docking step II: Sec. 1.5.5), which is reported on the left. If the scoring function is described in the same publication of the corresponding docking engine, the Tab reports its standard scoring function. Tools specifically developed for RNA have a light khaki background. The "Webserver" and "Download" flags in the Availability are hyperlinks addressing the corresponding website.

tool in drug design since it allows for the screening of similar targets and/or ligands starting from a query of input ligand/fragment or, alternatively, RNA secondary structure. Very recently, R-BIND ligands were further filtered and classified into more than 800 compounds, giving birth to the largest academic library of RNA-targeted compounds, the Duke RNA-Targeted Library (DRTL) [**WicksProbingMolecules**].

InfoRNA. Beyond being an RNA-small molecules database, InfoRNA is an innovative computational platform developed by Disney's group to identify small molecules targeting RNA from sequence [255]. InfoRNA functions by mining 2D motifs of target RNAs inferred from their sequence and comparing these motifs to a comprehensive database of known RNA motifs−small molecule interactions, which are derived from a systematic integration of scientific literature, advanced structural prediction and bioinformatic methods, such as 2DCS (Sec. 1.4.3). InfoRNA applications led to the identification of Targaprimir-96, a bioactive small molecule targeting the Drosha processing site of the oncogenic pre-miRNA96 and inhibiting its biogenesis [277].

ROBIN. A more recent example is the Repository Of BInders to Nucleic acids (ROBIN) database, developed by Schneekloth's group [75]. This repository compiles nucleic acid binders identified through microarray screenings (see Section 1.5.3) and provides results from physicochemical analyses comparing these binders with drug-like and, more broadly, protein binders, leveraging machine learning models. Specifically designed to enhance the understanding of the chemical space of RNA binders, ROBIN aims to delineate the boundary between RNA- and protein-binding small molecules. This distinction facilitates the design of chemical libraries and individual ligands targeting RNA structures.

While databases play a crucial role in drug discovery, there is a noticeable absence of comprehensive repositories specifically dedicated to RNA-small molecule complexes. Existing databases (Tab.

1.4) tend to focus solely on binders or to incorporate only information related to the secondary structure of RNA targets. Very rarely, databases collect information on binding affinity, which can instead give fundamental insights to assess the selectivity and specificity of compounds with a given target. The absence of a curated repository providing 3D structural data for RNA-small molecule binding pockets is a significant gap in the field. Future advances in SBDD could greatly benefit from the establishment of such repositories. These repositories would serve as a starting point for both benchmarking existing computational tools and gaining a deeper understanding of the physicochemical properties of RNA binding pockets and the structural patterns governing RNA-small molecule recognition.

Bibliography

[1] Bruce Alberts et al. *Molecular Biology of the Cell - NCBI Bookshelf*. New York, NY: Garland Science, 2002.

[2] Jennifer A. Doudna and Elizabeth A. Doherty. "Emerging themes in RNA folding". In: *Folding and Design* 2.5 (Oct. 1997), R65–R70.

[3] Vinod K. Misra and David E. Draper. "The linkage between magnesium binding and RNA folding". In: *Journal of Molecular Biology* 317.4 (Apr. 2002), pp. 507–521.

[4] Benjamin Philipp Fingerhut. "The mutual interactions of RNA, counterions and water – quantifying the electrostatics at the phosphate–water interface". In: *Chemical Communications (Cambridge, England)* 57.96 (Dec. 2021), p. 12880.

[5] Joel L. Sussman et al. "Crystal structure of yeast phenylalanine transfer RNA: I. Crystallographic refinement". In: *Journal of Molecular Biology* 123.4 (Aug. 1978), pp. 607–630.

[6] Patricia Bouchard and Pascale Legault. "Structural insights into substrate recognition by the neurospora varkud satellite ribozyme: Importance of u-turns at the kissing-loop junction". In: *Biochemistry* 53.1 (Jan. 2014), pp. 258–269.

[7] Michaël H. Kolk et al. "NMR structure of a classical pseudoknot: Interplay of single- and double-stranded RNA". In: *Science* 280.5362 (Apr. 1998), pp. 434–438.

[8] Matthew Cobb. "60 years ago, Francis Crick changed the logic of biology". In: *PLoS Biology* 15.9 (Sept. 2017).

[9] Vivien Marx. "How noncoding RNAs began to leave the junkyard". In: *Nature Methods* 19.10 (Oct. 2022), pp. 1167–1170.

[10] G. E. PALADE. "A SMALL PARTICULATE COMPONENT OF THE CYTOPLASM". In: *The Journal of Biophysical and Biochemical Cytology* 1.1 (Jan. 1955), pp. 59–68.

[11] Robert A. Weinberg and Sheldon Penman. "Small molecular weight monodisperse nuclear RNA". In: *Journal of Molecular Biology* 38.3 (Dec. 1968), pp. 289–304.

[12] Susan M. Berget, Claire Moore, and Phillip A. Sharp. "Spliced segments at the 5' terminus of adenovirus 2 late mRNA". In: *Proceedings of the National Academy of Sciences* 74.8 (Aug. 1977), pp. 3171–3175.

[13] Louise T. Chow, Thomas R. Broker, and James B. Lewis. "Complex splicing patterns of RNAs from the early regions of adenovirus-2". In: *Journal of Molecular Biology* 134.2 (Oct. 1979), pp. 265–303.

[14] Michael R. Lerner et al. “Are snRNPs involved in splicing?” In: *Nature* 283.5743 (Jan. 1980), pp. 220–224.

[15] Louise T. Chow et al. “An amazing sequence arrangement at the 5’ ends of adenovirus 2 messenger RNA”. In: *Cell* 12.1 (Sept. 1977), pp. 1–8.

[16] Kelly Kruger et al. “Self-splicing RNA: Autoexcision and autocyclization of the ribosomal RNA intervening sequence of tetrahymena”. In: *Cell* 31.1 (Nov. 1982), pp. 147–157.

[17] Andrew Fire et al. “Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans”. In: *Nature* 391.6669 (Feb. 1998), pp. 806–811.

[18] Shoshy Altuvia et al. “Alternative mRNA structures of the cIII gene of bacteriophage λ determine the rate of its translation initiation”. In: *Journal of Molecular Biology* 210.2 (Nov. 1989), pp. 265–280.

[19] G. Storz. “An RNA thermometer”. In: *Genes & Development* 13.6 (Mar. 1999), pp. 633–636.

[20] Ian Dunham et al. “An integrated encyclopedia of DNA elements in the human genome”. In: *Nature 2012 489:7414* 489.7414 (Sept. 2012), pp. 57–74.

[21] Sarah Djebali et al. “Landscape of transcription in human cells”. In: *Nature 2012 489:7414* 489.7414 (Sept. 2012), pp. 101–108.

[22] John S. Mattick. “A Kuhnian revolution in molecular biology: Most genes in complex organisms express regulatory RNAs”. In: *BioEssays* 45.9 (Sept. 2023).

[23] Francisco J.M. Mojica et al. “Intervening Sequences of Regularly Spaced Prokaryotic Repeats Derive from Foreign Genetic Elements”. In: *Journal of Molecular Evolution* 60.2 (Feb. 2005), pp. 174–182.

[24] Rodolphe Barrangou et al. “CRISPR Provides Acquired Resistance Against Viruses in Prokaryotes”. In: *Science* 315.5819 (Mar. 2007), pp. 1709–1712.

[25] Martin Jinek et al. “A Programmable Dual-RNA–Guided DNA Endonuclease in Adaptive Bacterial Immunity”. In: *Science* 337.6096 (Aug. 2012), pp. 816–821.

[26] Tae Kyung Kim et al. “Widespread transcription at neuronal activity-regulated enhancers”. In: *Nature 2010 465:7295* 465.7295 (Apr. 2010), pp. 182–187.

[27] Julia Salzman et al. “Circular RNAs Are the Predominant Transcript Isoform from Hundreds of Human Genes in Diverse Cell Types”. In: *PLoS ONE* 7.2 (Feb. 2012).

[28] Harold S. Bernhardt. “The RNA world hypothesis: the worst theory of the early evolution of life (except for all the others)a”. In: *Biology Direct* 7.1 (July 2012), pp. 1–10.

[29] Peijing Zhang et al. “Non-Coding RNAs and their Integrated Networks”. In: *Journal of integrative bioinformatics* 16.3 (July 2019).

[30] Martha J. Fedor. “The role of metal ions in RNA catalysis”. In: *Current Opinion in Structural Biology* 12.3 (June 2002), pp. 289–295.

[31] Bryan R. Cullen. “Viral RNAs: Lessons from the Enemy”. In: *Cell* 136.4 (Feb. 2009), pp. 592–597.

[32] D Harrich, C Ulich, and R B Gaynor. "A critical role for the TAR element in promoting efficient human immunodeficiency virus type 1 reverse transcription". In: *Journal of Virology* 70.6 (June 1996), pp. 4017–4027.

[33] Neema Agrawal et al. "RNA Interference: Biology, Mechanism, and Applications". In: *Microbiology and Molecular Biology Reviews* 67.4 (Dec. 2003), p. 657.

[34] Jens Kortmann and Franz Narberhaus. "Bacterial RNA thermometers: molecular zippers and switches". In: *Nature Reviews Microbiology* 10.4 (Apr. 2012), pp. 255–265.

[35] Alexander Serganov and Evgeny Nudler. "A Decade of Riboswitches". In: *Cell* 152.1-2 (Jan. 2013), pp. 17–24.

[36] John S. Mattick et al. "Long non-coding RNAs: definitions, functions, challenges and recommendations". In: *Nature Reviews Molecular Cell Biology* 24.6 (June 2023), pp. 430–447.

[37] Gayatri Arun, Disha Aggarwal, and David L. Spector. "MALAT1 Long Non-Coding RNA: Functional Implications". In: *Non-coding RNA* 6.2 (June 2020).

[38] Rodrigo Aguilar et al. "Targeting Xist with compounds that disrupt RNA structure and X inactivation". In: *Nature* 604.7904 (Apr. 2022), pp. 160–166.

[39] Juan José Montero et al. "Telomeric RNAs are essential to maintain telomeres". In: *Nature Communications 2016 7:1* 7.1 (Aug. 2016), pp. 1–13.

[40] Melanie Winkle et al. "Noncoding RNA therapeutics — challenges and potential solutions". In: *Nature Reviews Drug Discovery* 20.8 (Aug. 2021), pp. 629–651.

[41] Feng Wang, Travis Zuroske, and Jonathan K. Watts. "RNA therapeutics on the rise". In: *Nature Reviews Drug Discovery* 19.7 (July 2020), pp. 441–442.

[42] James C. Kaczmarek, Piotr S. Kowalski, and Daniel G. Anderson. "Advances in the delivery of RNA therapeutics: from concept to clinical reality". In: *Genome Medicine* 9.1 (Dec. 2017), p. 60.

[43] Jianhua Zhou, Xuexiu Zheng, and Haihong Shen. "Targeting RNA-splicing for SMA treatment". In: *Molecules and Cells* 33.3 (Mar. 2012), pp. 223–228.

[44] Feng Gao and Wenhui Wang. "MicroRNA-96 promotes the proliferation of colorectal cancer cells and targets tumor protein p53 inducible nuclear protein 1, forkhead box protein O1 (FOXO1) and FOXO3a". In: *Molecular medicine reports* 11.2 (Feb. 2015), pp. 1200–1206.

[45] Diego Oliva-Rico and Luis A. Herrera. "Regulated expression of the lncRNA TERRA and its impact on telomere biology". In: *Mechanisms of ageing and development* 167 (Oct. 2017), pp. 16–23.

[46] Kaushik Chanda and Debashis Mukhopadhyay. "LncRNA Xist, X-chromosome Instability and Alzheimer's Disease". In: *Current Alzheimer research* 17.6 (Aug. 2020), pp. 499–507.

[47] Haley M. Wilt et al. "FMN riboswitch aptamer symmetry facilitates conformational switching through mutually exclusive coaxial stacking configurations". In: *Journal of Structural Biology: X* 4 (2020), p. 100035.

[48] Sai Shashank Chavali, Rachel Bonn-Breach, and Joseph E. Wedekind. "Face-time with TAR: Portraits of an HIV-1 RNA with diverse modes of effector recognition relevant for drug discovery". In: *Journal of Biological Chemistry* 294.24 (June 2019), pp. 9326–9341.

[49] Kushal J. Rohilla and Keith T. Gagnon. "RNA biology of disease-associated microsatellite repeat expansions". In: *Acta neuropathologica communications* 5.1 (Aug. 2017), p. 63.

[50] J. David Brook et al. "Molecular basis of myotonic dystrophy: Expansion of a trinucleotide (CTG) repeat at the 3' end of a transcript encoding a protein kinase family member". In: *Cell* 68.4 (Feb. 1992), pp. 799–808.

[51] Noreen F. Rizvi et al. "Discovery of Selective RNA-Binding Small Molecules by Affinity-Selection Mass Spectrometry". In: *ACS Chemical Biology* 13.3 (Mar. 2018), pp. 820–831.

[52] Katherine Deigan Warner, Christine E. Hajdin, and Kevin M. Weeks. "Principles for targeting RNA with drug-like small molecules". In: *Nature Reviews Drug Discovery* 17.8 (2018), pp. 547–558.

[53] Jennifer Harrow et al. "GENCODE: The reference human genome annotation for The ENCODE Project". In: *Genome Research* 22.9 (Sept. 2012), pp. 1760–1774.

[54] Scott J. Dixon and Brent R. Stockwell. "Identifying druggable disease-modifying gene products". In: *Current Opinion in Chemical Biology* 13.5-6 (Dec. 2009), pp. 549–555.

[55] Rita Santos et al. "A comprehensive map of molecular drug targets". In: *Nature Reviews Drug Discovery 2016 16:1* 16.1 (Dec. 2016), pp. 19–34.

[56] Ella Czarina Morishita. "Discovery of RNA-targeted small molecules through the merging of experimental and computational technologies". In: *Expert Opinion on Drug Discovery* 18.2 (Feb. 2023), pp. 207–226.

[57] Davide Di Fusco et al. "Antisense oligonucleotide: Basic concepts and therapeutic application in inflammatory bowel disease". In: *Frontiers in Pharmacology* 10.MAR (Mar. 2019), p. 440751.

[58] C. Frank Bennett. "Therapeutic Antisense Oligonucleotides Are Coming of Age". In: *Annual Review of Medicine* 70.1 (Jan. 2019), pp. 307–321.

[59] Xiuhui Chen et al. "RNA interference-based therapy and its delivery systems". In: *Cancer and Metastasis Reviews* 37.1 (Mar. 2018).

[60] Katrin Tiemann and John J. Rossi. "RNAi-based therapeutics–current status, challenges and prospects". In: *EMBO Molecular Medicine* 1.3 (June 2009), p. 142.

[61] Christof Fellmann et al. "Cornerstones of CRISPR–Cas in drug discovery and therapy". In: *Nature Reviews Drug Discovery 2016 16:2* 16.2 (Dec. 2016), pp. 89–100.

[62] D.C. Luther et al. "Delivery approaches for CRISPR/Cas9 therapeutics in vivo : advances and challenges". In: *Expert Opinion on Drug Delivery* 15.9 (Sept. 2018), pp. 905–913.

[63] Zhihan Zhao et al. "Prime editing: advances and therapeutic applications". In: *Trends in Biotechnology* 41.8 (Aug. 2023), pp. 1000–1012.

[64] James P. Falese, Anita Donlic, and Amanda E. Hargrove. "Targeting RNA with small molecules: from fundamental principles towards the clinic". In: *Chemical Society Reviews* 50.4 (2021), pp. 2224–2243.

[65] Jessica L. Childs-Disney et al. “Targeting RNA structures with small molecules”. In: *Nature Reviews Drug Discovery* (Aug. 2022).

[66] Colleen M. Connelly, Michelle H. Moon, and John S. Schneekloth. “The Emerging Role of RNA as a Therapeutic Target for Small Molecules”. In: *Cell Chemical Biology* 23.9 (Sept. 2016), pp. 1077–1090.

[67] A. Di Giorgio and M. Duca. “Synthetic small-molecule RNA ligands: future prospects as therapeutic agents”. In: *MedChemComm* 10.8 (2019), pp. 1242–1255.

[68] Christopher A Lipinski et al. “Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings 1PII of original article: S0169-409X(96)00423-1. The article was originally published in Advanced Drug Delivery Reviews 23 (1997) 3–25. 1”. In: *Advanced Drug Delivery Reviews* 46.1-3 (Mar. 2001), pp. 3–26.

[69] Hartmut Beck et al. “Small molecules and their impact in drug discovery: A perspective on the occasion of the 125th anniversary of the Bayer Chemical Research Laboratory”. In: *Drug Discovery Today* 27.6 (June 2022), pp. 1560–1574.

[70] Andrew L. Hopkins and Colin R. Groom. “The druggable genome”. In: *Nature Reviews Drug Discovery* 1.9 (Sept. 2002), pp. 727–730.

[71] Brett Lomenick, Richard W. Olsen, and Jing Huang. “Identification of Direct Protein Targets of Small Molecules”. In: *ACS Chemical Biology* 6.1 (Jan. 2011), p. 34.

[72] Lijuan Wang et al. “Analytical methods for obtaining binding parameters of drug–protein interactions: A review”. In: *Analytica Chimica Acta* 1219 (Aug. 2022), p. 340012.

[73] Fareed Aboul-Ela. “Strategies for the design of RNA-binding small molecules”. In: *Future Medicinal Chemistry* 2.1 (Jan. 2010), pp. 93–119.

[74] G. Padroni et al. “Systematic analysis of the interactions driving small molecule-RNA recognition”. In: *RSC Medicinal Chemistry* 11.7 (2020), pp. 802–813.

[75] Kamyar Yazdani et al. “Machine Learning Informs RNA-Binding Chemical Space**”. In: *Angewandte Chemie* 135.11 (Mar. 2023), e202211358.

[76] Suzanne G. Rzuczek, Mark R. Southern, and Matthew D. Disney. “Studying a Drug-like, RNA-Focused Small Molecule Library Identifies Compounds That Inhibit RNA Toxicity in Myotonic Dystrophy”. In: *ACS Chemical Biology* 10.12 (Dec. 2015), pp. 2706–2715.

[77] Brittany S. Morgan et al. “Discovery of Key Physicochemical, Structural, and Spatial Properties of RNA-Targeted Bioactive Ligands”. In: *Angewandte Chemie - International Edition* 56.43 (2017), pp. 13498–13502.

[78] Anita Donlic et al. “R-BIND 2.0: An Updated Database of Bioactive RNA-Targeting Small Molecules and Associated RNA Secondary Structures”. In: *ACS Chemical Biology* 17.6 (June 2022), pp. 1556–1566.

[79] William M. Hewitt, David R. Calabrese, and John S. Schneekloth. “Evidence for ligandable sites in structured RNA throughout the Protein Data Bank”. In: *Bioorganic and Medicinal Chemistry* 27.11 (2019), pp. 2253–2260.

[80] Sandra Kovachka et al. “Small molecule approaches to targeting RNA”. In: *Nature Reviews Chemistry 2024 8:2* 8.2 (Jan. 2024), pp. 120–135.

[81] Frank Lovering, Jack Bikker, and Christine Humblet. “Escape from flatland: increasing saturation as an approach to improving clinical success”. In: *Journal of medicinal chemistry* 52.21 (Nov. 2009), pp. 6752–6756.

[82] Lirui Guan and Matthew D. Disney. “Recent advances in developing small molecules targeting RNA”. In: *ACS Chemical Biology* 7.1 (Jan. 2012), pp. 73–86.

[83] Kevin M. Krause et al. “Aminoglycosides: An Overview”. In: *Cold Spring Harbor Perspectives in Medicine* 6.6 (June 2016).

[84] Robert C. Moellering. “Linezolid: The first oxazolidinone antimicrobial”. In: *Annals of Internal Medicine* 138.2 (Jan. 2003), pp. 135–142.

[85] Daniel J. Diekema and Ronald N. Jones. “Oxazolidinone antibiotics”. In: *Lancet (London, England)* 358.9297 (Dec. 2001), pp. 1975–1982.

[86] John A. Howe et al. “Selective small-molecule inhibition of an RNA structural element”. In: *Nature* 526.7575 (Oct. 2015), pp. 672–677.

[87] Stephen E. Motika et al. “A Gram-Negative Antibiotic Active Through Inhibition of an Essential Riboswitch”. In: *Journal of the American Chemical Society* 142.24 (June 2020), p. 10856.

[88] Hasane Ratni et al. “Discovery of Risdiplam, a Selective Survival of Motor Neuron-2 (SMN2) Gene Splicing Modifier for the Treatment of Spinal Muscular Atrophy (SMA)”. In: *Journal of Medicinal Chemistry* 61.15 (Aug. 2018), pp. 6501–6517.

[89] Joseph A. Ippolito et al. “Crystal structure of the oxazolidinone antibiotic linezolid bound to the 50S ribosomal subunit”. In: *Journal of Medicinal Chemistry* 51.12 (June 2008), pp. 3353–3356.

[90] John A. Howe et al. “Atomic resolution mechanistic studies of ribocil: A highly selective unnatural ligand mimic of the E. coli FMN riboswitch”. In: *RNA Biology* 13.10 (Oct. 2016), pp. 946–954.

[91] Sébastien Campagne et al. “Structural basis of a small molecule targeting RNA for a specific splicing correction”. In: *Nature Chemical Biology 2019 15:12* 15.12 (Oct. 2019), pp. 1191–1198.

[92] Ming Qiang Zhang and Barrie Wilkinson. “Drug discovery beyond the ‘rule-of-five’”. In: *Current Opinion in Biotechnology* 18.6 (Dec. 2007), pp. 478–488.

[93] Chris J. Radoux et al. “The druggable genome: Twenty years later”. In: *Frontiers in Bioinformatics* 2 (Sept. 2022), p. 958378.

[94] Michael D. Shultz. “Two Decades under the Influence of the Rule of Five and the Changing Properties of Approved Oral Drugs”. In: *Journal of Medicinal Chemistry* 62.4 (Feb. 2019), pp. 1701–1714.

[95] Timothy E H Allen et al. “Physicochemical Principles Driving Small Molecule Binding to RNA”. In: *bioRxiv* (Feb. 2024), p. 2024.01.31.578268.

[96] Sourav K. Dey and Samie R. Jaffrey. "RIBOTACs: Small Molecules Target RNA for Degradation". In: *Cell Chemical Biology* 26.8 (Aug. 2019), pp. 1047–1049.

[97] Farukh Arjmand et al. "Recent advances in metallodrug-like molecules targeting non-coding RNAs in cancer chemotherapy". In: *Coordination Chemistry Reviews* 387 (May 2019), pp. 47–59.

[98] James M. Carothers et al. "Informational Complexity and Functional Activity of RNA Structures". In: *Journal of the American Chemical Society* 126.16 (Apr. 2004), pp. 5130–5137.

[99] Laura R. Ganser et al. "The roles of structural dynamics in the cellular functions of RNAs". In: *Nature Reviews Molecular Cell Biology* 20.8 (2019), pp. 474–489.

[100] Megan L. Ken et al. "RNA conformational propensities determine cellular activity". In: *Nature* 617.7962 (May 2023), pp. 835–841.

[101] Sarah C. Keane et al. "NMR detection of intermolecular interaction sites in the dimeric 5'-leader of the HIV-1 genome". In: *Proceedings of the National Academy of Sciences of the United States of America* 113.46 (Nov. 2016), pp. 13033–13038.

[102] Ailong Ke et al. "A conformational switch controls hepatitis delta virus ribozyme catalysis". In: *Nature* 429.6988 (May 2004), pp. 201–205.

[103] J. Matthew Taliaferro et al. "RNA Sequence Context Effects Measured In Vitro Predict In Vivo Protein Binding and Regulation". In: *Molecular Cell* 64.2 (Oct. 2016), pp. 294–306.

[104] Yunsun Nam et al. "Molecular Basis for Interaction of let-7 MicroRNAs with Lin28". In: *Cell* 147.5 (Nov. 2011), pp. 1080–1091.

[105] Srinivas Somarowthu et al. "HOTAIR forms an intricate and modular secondary structure". In: *Molecular cell* 58.2 (Apr. 2015), p. 353.

[106] Anthony M. Mustoe, Charles L. Brooks, and Hashim M. Al-Hashimi. "Hierarchy of RNA Functional Dynamics". In: *https://doi.org/10.1146/annurev-biochem-060713-035524* 83 (June 2014), pp. 441–466.

[107] Vojtěch Mlýnský and Giovanni Bussi. "Exploring RNA structure and dynamics through enhanced sampling simulations". In: *Current Opinion in Structural Biology* 49 (Apr. 2018), pp. 63–71.

[108] Hans Frauenfelder, Stephen G. Sligar, and Peter G. Wolynes. "The energy landscapes and motions of proteins". In: *Science (New York, N.Y.)* 254.5038 (1991), pp. 1598–1603.

[109] Hashim M Al-Hashimi and Nils G Walter. "RNA dynamics: it is about time". In: *Current Opinion in Structural Biology* 18.3 (June 2008), pp. 321–329.

[110] Aaron T. Frank et al. "Constructing RNA dynamical ensembles by combining MD and motionally decoupled NMR RDCs: New insights into RNA dynamics and adaptive ligand recognition". In: *Nucleic Acids Research* 37.11 (2009), pp. 3670–3679.

[111] Andrew C. Stelzer et al. "Discovery of selective bioactive small molecules by targeting an RNA dynamic ensemble". In: *Nature Chemical Biology* 7.8 (2011), pp. 553–559.

[112] Yang Qi et al. "Continuous Interdomain Orientation Distributions Reveal Components of Binding Thermodynamics". In: *Journal of Molecular Biology* 430.18 (Sept. 2018), pp. 3412–3426.

[113] Alastair I.H. Murchie et al. “Structure-based Drug Design Targeting an Inactive RNA Conformation: Exploiting the Flexibility of HIV-1 TAR RNA”. In: *Journal of Molecular Biology* 336.3 (Feb. 2004), pp. 625–638.

[114] Ignacio Tinoco and Carlos Bustamante. “How RNA folds”. In: *Journal of Molecular Biology* 293.2 (Oct. 1999), pp. 271–281.

[115] Philippe Brion and Eric Westhof. “HIERARCHY AND DYNAMICS OF RNA FOLDING”. In: *https://doi.org/10.1146/annurev.biophys.26.1.113* 26 (Nov. 2003), pp. 113–137.

[116] Anthony M. Mustoe et al. “New insights into the fundamental role of topological constraints as a determinant of two-way junction conformation”. In: *Nucleic Acids Research* 40.2 (Jan. 2012), pp. 892–904.

[117] Boris Fürtig et al. “Conformational dynamics of bistable RNAs studied by time-resolved NMR spectroscopy”. In: *Journal of the American Chemical Society* 129.51 (Dec. 2007), pp. 16222–16229.

[118] Daniel Herschlag. “RNA Chaperones and the RNA Folding Problem”. In: *Journal of Biological Chemistry* 270.36 (Sept. 1995), pp. 20871–20874.

[119] Congju Chen et al. “Structural energetics and base-pair opening dynamics in sarcin-ricin domain RNA”. In: *Biochemistry* 45.45 (Nov. 2006), pp. 13606–13613.

[120] Elizabeth A. Dethoff et al. “Visualizing transient low-populated structures of RNA”. In: *Nature 2012 491:7426* 491.7426 (Oct. 2012), pp. 724–728.

[121] Neocles B. Leontis, Jesse Stombaugh, and Eric Westhof. “The non-Watson–Crick base pairs and their associated isostericity matrices”. In: *Nucleic Acids Research* 30.16 (Aug. 2002), pp. 3497–3531.

[122] Brent M. Znosko et al. “Structural features and thermodynamics of the J4/5 loop from the Candida albicans and Candida dubliniensis group I introns”. In: *Biochemistry* 43.50 (Dec. 2004), pp. 15822–15837.

[123] Xiaowei Zhuang et al. “Correlating structural dynamics and function in single ribozyme molecules”. In: *Science* 296.5572 (May 2002), pp. 1473–1476.

[124] Qi Zhang et al. “Comparison of solution and crystal structures of PreQ 1 riboswitch reveals calcium-induced changes in conformation and dynamics”. In: *Journal of the American Chemical Society* 133.14 (Apr. 2011), pp. 5190–5193.

[125] Brian Houck-Loomis et al. “An equilibrium-dependent retroviral mRNA switch regulates translational recoding”. In: *Nature 2011 480:7378* 480.7378 (Nov. 2011), pp. 561–564.

[126] Rebecca M Voorhees and V Ramakrishnan. “BI82CH08-Ramakrishnan ARI 1 May 2013 16:8 Structural Basis of the Translational Elongation Cycle *”. In: ().

[127] José Almeida Cruz and Eric Westhof. “The Dynamic Landscapes of RNA Architecture”. In: *Cell* 136.4 (Feb. 2009), pp. 604–609.

[128] Vladimir Baumruk et al. “Comparison between CUUG and UUCG tetraloops: thermodynamic stability and structural features analyzed by UV absorption and vibrational spectroscopy”. In: *Nucleic Acids Research* 29.19 (Oct. 2001), pp. 4089–4096.

[129] Sarah Knight Denny et al. “High-Throughput Investigation of Diverse Junction Elements in RNA Tertiary Folding”. In: *Cell* 174.2 (July 2018), pp. 377–390.

[130] Namita Bisaria et al. “Kinetic and thermodynamic framework for P4-P6 RNA reveals tertiary motif modularity and modulation of the folding preferred pathway”. In: *Proceedings of the National Academy of Sciences* 113.34 (Aug. 2016).

[131] Andreas Oxenfarth et al. “Integrated NMR/Molecular Dynamics Determination of the Ensemble Conformation of a Thermodynamically Stable CUUG RNA Tetraloop”. In: *Journal of the American Chemical Society* 145.30 (Aug. 2023), pp. 16557–16572.

[132] M. S. Smyth and J. H.J. Martin. “x Ray crystallography”. In: *Molecular Pathology* 53.1 (2000), p. 8.

[133] Dmitry Lyumkis. “Challenges and opportunities in cryo-EM single-particle analysis”. In: *The Journal of Biological Chemistry* 294.13 (Mar. 2019), p. 5181.

[134] Alan Brown and Sichen Shao. “Ribosomes and cryo-EM: a duet”. In: *Current Opinion in Structural Biology* 52 (Oct. 2018), pp. 1–7.

[135] Honglue Shi et al. “Rapid and accurate determination of atomistic RNA dynamic ensemble models using NMR and structure prediction”. In: *Nature Communications 2020 11:1* 11.1 (Nov. 2020), pp. 1–14.

[136] Maja Marušič, Maria Toplishek, and Janez Plavec. “NMR of RNA - Structure and interactions”. In: *Current Opinion in Structural Biology* 79 (Apr. 2023), p. 102532.

[137] Kevin M. Weeks and David M. Mauger. “Exploring RNA Structural Codes with SHAPE Chemistry”. In: *Accounts of chemical research* 44.12 (Dec. 2011), p. 1280.

[138] Nathan A. Siegfried et al. “RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP)”. In: *Nature Methods 2014 11:9* 11.9 (July 2014), pp. 959–965.

[139] Aleksandar Spasic et al. “Modeling RNA secondary structure folding ensembles using SHAPE mapping data”. In: *Nucleic Acids Research* 46.1 (Jan. 2018), pp. 314–323.

[140] Pablo Cordero and Rhiju Das. “Rich RNA Structure Landscapes Revealed by Mutate-and-Map Analysis”. In: *PLOS Computational Biology* 11.11 (Nov. 2015), e1004473.

[141] Clarence Yu Cheng et al. “Consistent global structures of complex RNA states through multidimensional chemical mapping”. In: *eLife* 4.JUNE2015 (June 2015).

[142] Yujie Chen and Lois Pollack. “SAXS studies of RNA: structures, dynamics, and interactions with partners”. In: *Wiley Interdisciplinary Reviews: RNA* 7.4 (July 2016), pp. 512–526.

[143] Benjamin Schuler. “Single-molecule FRET of protein structure and dynamics - a primer.” In: *Journal of nanobiotechnology* 11 Suppl 1.1 (Dec. 2013), pp. 1–17.

[144] Bing Li et al. “Advances in RNA 3D Structure Modeling Using Experimental Data”. In: *Frontiers in Genetics* 11 (Oct. 2020), p. 574485.

[145] John Jumper et al. “Highly accurate protein structure prediction with AlphaFold”. In: *Nature 2021 596:7873* 596.7873 (July 2021), pp. 583–589.

[146] Jinsong Zhang et al. “Advances and opportunities in RNA structure experimental determination and computational modeling”. In: *Nature Methods 2022 19:10* 19.10 (Oct. 2022), pp. 1193–1207.

[147] Jie Deng et al. “RNA structure determination: From 2D to 3D”. In: *Fundamental Research* 3.5 (Sept. 2023), pp. 727–737.

[148] Elena Rivas, Jody Clements, and Sean R. Eddy. “A statistical test for conserved RNA structure shows lack of evidence for structure in lncRNAs”. In: *Nature Methods 2016 14:1* 14.1 (Nov. 2016), pp. 45–48.

[149] Laiyi Fu et al. “UFold: fast and accurate RNA secondary structure prediction with deep learning”. In: *Nucleic Acids Research* 50.3 (Feb. 2022), e14–e14.

[150] Michal J. Boniecki et al. “SimRNA: a coarse-grained method for RNA folding simulations and 3D structure prediction”. In: *Nucleic Acids Research* 44.7 (Apr. 2016), e63–e63.

[151] Andrew Martin Watkins, Ramya Rangan, and Rhiju Das. “FARFAR2: Improved De Novo Rosetta Prediction of Complex Global RNA Folds”. In: *Structure* 28.8 (Aug. 2020), pp. 963–976.

[152] Xiaojun Xu, Chenhan Zhao, and Shi Jie Chen. “VfoldLA: A web server for loop assembly-based prediction of putative 3D RNA structures”. In: *Journal of Structural Biology* 207.3 (Sept. 2019), pp. 235–240.

[153] Raphael J.L. Townshend et al. “Geometric deep learning of RNA structure”. In: *Science* 373.6558 (Aug. 2021), pp. 1047–1051.

[154] Lorenzo Baronti et al. “Base-pair conformational switch modulates miR-34a targeting of Sirt1 mRNA”. In: *Nature 2020 583:7814* 583.7814 (May 2020), pp. 139–144.

[155] Kalli Kappel et al. “Accelerated cryo-EM-guided determination of three-dimensional RNA-only structures”. In: *Nature Methods 2020 17:7* 17.7 (July 2020), pp. 699–707.

[156] Max E. Wilkinson et al. “Structural basis for conformational equilibrium of the catalytic spliceosome”. In: *Molecular Cell* 81.7 (Apr. 2021), pp. 1439–1452.

[157] Miles Kubota, Catherine Tran, and Robert C. Spitale. “Progress and challenges for chemical probing of RNA structure inside living cells”. In: *Nature Chemical Biology 2015 11:12* 11.12 (Nov. 2015), pp. 933–941.

[158] Massimiliano Bonomi et al. “Principles of protein structural ensemble determination”. In: *Current Opinion in Structural Biology* 42 (2017), pp. 106–116.

[159] Alisha N. Jones and Michael Sattler. “Challenges and perspectives for structural biology of lncRNAs—the example of the Xist lncRNA A-repeats”. In: *Journal of Molecular Cell Biology* 11.10 (Oct. 2019), pp. 845–859.

[160] Sharon Aviran and Danny Incarnato. “Computational Approaches for RNA Structure Ensemble Deconvolution from Structure Probing Data”. In: *Journal of Molecular Biology* 434.18 (Sept. 2022), p. 167635.

[161] Haopeng Yu, Yiman Qi, and Yiliang Ding. “Deep Learning in RNA Structure Studies”. In: *Frontiers in Molecular Biosciences* 9 (May 2022), p. 869601.

[162] Kengo Sato and Michiaki Hamada. “Recent trends in RNA informatics: a review of machine learning and deep learning for RNA secondary structure prediction and RNA drug discovery”. In: *Briefings in Bioinformatics* 24.4 (July 2023), pp. 1–13.

[163] Yang Wu et al. “Improved prediction of RNA secondary structure by integrating the free energy model with restraints derived from experimental probing data”. In: *Nucleic Acids Research* 43.15 (Sept. 2015), pp. 7247–7259.

[164] Nikolay V. Dokholyan. “Experimentally-driven protein structure modeling”. In: *Journal of Proteomics* 220 (May 2020), p. 103777.

[165] Rohit Roy et al. “Kinetic Resolution of the Atomic 3D Structures Formed by Ground and Excited Conformational States in an RNA Dynamic Ensemble”. In: *Journal of the American Chemical Society* 145.42 (Oct. 2023), pp. 22964–22978.

[166] José Almeida Cruz et al. “RNA-Puzzles: A CASP-like evaluation of RNA three-dimensional structure prediction”. In: *RNA* 18.4 (Apr. 2012), p. 610.

[167] John Moult et al. “A large-scale experiment to assess protein structure prediction methods”. In: *Proteins: Structure, Function, and Bioinformatics* 23.3 (Nov. 1995), pp. ii–iv.

[168] Daan Frenkel and Berend Smit. *Understanding Molecular Simulation: From Algorithms to Applications, Third Edition*. Elsevier, Jan. 2023, pp. 1–728.

[169] Jiří Šponer et al. “RNA Structural Dynamics As Captured by Molecular Simulations: A Comprehensive Overview”. In: *Chemical Reviews* 118.8 (Apr. 2018), pp. 4177–4338.

[170] Giulia Palermo et al. “Understanding the mechanistic basis of non-coding RNA through molecular dynamics simulations”. In: *Journal of Structural Biology* 206.3 (June 2019), pp. 267–279.

[171] Sweta Vangaveti, Srivathsan V. Ranganathan, and Alan A. Chen. “Advances in RNA molecular dynamics: a simulator’s guide to RNA force fields”. In: *Wiley Interdisciplinary Reviews: RNA* 8.2 (Mar. 2017), e1396.

[172] Jan von Plato. “Boltzmann’s ergodic hypothesis”. In: *Archive for History of Exact Sciences* 42.1 (Mar. 1991), pp. 71–89.

[173] Alexander D. Mackerell. “Empirical force fields for biological macromolecules: Overview and issues”. In: *Journal of Computational Chemistry* 25.13 (Oct. 2004), pp. 1584–1604.

[174] Christopher I. Bayly et al. “A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules”. In: *Journal of the American Chemical Society* 117.19 (1995), pp. 5179–5197.

[175] Petra Kührová et al. “Computer Folding of RNA Tetraloops: Identification of Key Force Field Deficiencies”. In: *Journal of Chemical Theory and Computation* 12.9 (Sept. 2016), pp. 4534–4548.

[176] Pablo D. Dans et al. “Modeling, Simulations, and Bioinformatics at the Service of RNA Structure”. In: *Chem* 5.1 (Jan. 2019), pp. 51–73.

[177] Jiri Sponer et al. “RNA structural dynamics as captured by molecular simulations: A comprehensive overview”. In: *Chemical Reviews* 118.8 (2018), pp. 4177–4338.

[178] Louis G. Smith et al. “Physics-based all-atom modeling of RNA energetics and structure”. In: *Wiley Interdisciplinary Reviews: RNA* 8.5 (Sept. 2017), e1422.

[179] Alexa M Salsbury and Justin A Lemkul. “Recent developments in empirical atomistic force fields for nucleic acids and applications to studies of folding and dynamics”. In: *Current Opinion in Structural Biology* 67 (Apr. 2021), pp. 9–17.

[180] Thomas E. Cheatham, Piotr Cieplak, and Peter A. Kollman. “A modified version of the Cornell et al. force field with improved sugar pucker phases and helical repeat”. In: *Journal of biomolecular structure & dynamics* 16.4 (1999), pp. 845–862.

[181] Junmei Wang, Piotr Cieplak, and Peter A Kollman. “How Well Does a Restrained Electrostatic Potential (RESP) Model Perform in Calculating Conformational Energies of Organic and Biological Molecules? Keywords: additive force field; nonadditive force field; restrained electrostatic potential (RESP); torsional angle parameterization”. In: *Journal of Computational Chemistry* 21.12 (2000), pp. 1049–1074.

[182] Alberto Pérez et al. “Refinement of the AMBER Force Field for Nucleic Acids: Improving the Description of α/γ Conformers”. In: *Biophysical Journal* 92.11 (June 2007), pp. 3817–3829.

[183] Marie Zgarbová et al. “Refinement of the Cornell et al. Nucleic Acids Force Field Based on Reference Quantum Chemical Calculations of Glycosidic Torsion Profiles”. In: *Journal of Chemical Theory and Computation* 7.9 (Sept. 2011), pp. 2886–2902.

[184] Ilyas Yildirim et al. “Reparameterization of RNA χ torsion parameters for the AMBER force field and comparison to NMR spectra for cytidine and uridine”. In: *Journal of Chemical Theory and Computation* 6.5 (May 2010), pp. 1520–1531.

[185] Ilyas Yildirim et al. “Revision of AMBER torsional parameters for RNA improves free energy predictions for tetramer duplexes with GC and iGiC base pairs”. In: *Journal of Chemical Theory and Computation* 8.1 (Jan. 2012), pp. 172–181.

[186] Jiří Šponer et al. “The DNA and RNA sugar–phosphate backbone emerges as the key player. An overview of quantum-chemical, structural biology and simulation studies”. In: *Physical Chemistry Chemical Physics* 14.44 (Oct. 2012), pp. 15257–15277.

[187] Alan A. Chen and Angel E. García. “High-resolution reversible folding of hyperstable RNA tetraloops using molecular dynamics simulations”. In: *Proceedings of the National Academy of Sciences of the United States of America* 110.42 (Oct. 2013), pp. 16820–16825.

[188] Christina Bergonzo and Thomas E. Cheatham. “Improved Force Field Parameters Lead to a Better Description of RNA Structure”. In: *Journal of Chemical Theory and Computation* 11.9 (Sept. 2015), pp. 3969–3972.

[189] Asaminew H. Aytenfisu et al. “Revised RNA Dihedral Parameters for the Amber Force Field Improve RNA Molecular Dynamics”. In: *Journal of Chemical Theory and Computation* 13.2 (Feb. 2017), pp. 900–915.

[190] Thorben Fröhlking et al. “Automatic Learning of Hydrogen-Bond Fixes in the AMBER RNA Force Field”. In: *Journal of chemical theory and computation* 18.7 (July 2022), pp. 4490–4502.

[191] Jun Chen et al. “RNA-Specific Force Field Optimization with CMAP and Reweighting”. In: *Journal of Chemical Information and Modeling* 62.2 (Jan. 2022), pp. 372–385.

[192] David E. Shaw et al. "Anton, a special-purpose machine for molecular dynamics simulation". In: *Communications of the ACM* 51.7 (July 2008), pp. 91–97.

[193] Kresten Lindorff-Larsen et al. "Improved side-chain torsion potentials for the Amber ff99SB protein force field". In: *Proteins: Structure, Function, and Bioinformatics* 78.8 (June 2010), pp. 1950–1958.

[194] Dazhi Tan et al. "RNA force field with accuracy comparable to state-of-the-art protein force fields". In: *Proceedings of the National Academy of Sciences of the United States of America* 115.7 (Feb. 2018), E1346–E1355.

[195] Stefano Piana et al. "Development of a Force Field for the Simulation of Single-Chain Proteins and Protein-Protein Complexes". In: *Journal of Chemical Theory and Computation* 16.4 (Apr. 2020), pp. 2494–2507.

[196] Maxwell R. Tucker et al. "Development of Force Field Parameters for the Simulation of Single- and Double-Stranded DNA Molecules and DNA-Protein Complexes". In: *Journal of Physical Chemistry B* 126.24 (June 2022), pp. 4442–4457.

[197] Bernard R. Brooks et al. "CHARMM: A program for macromolecular energy, minimization, and dynamics calculations". In: *Journal of Computational Chemistry* 4.2 (June 1983), pp. 187–217.

[198] Alexander D. MacKerell et al. "An All-Atom Empirical Energy Function for the Simulation of Nucleic Acids". In: *Journal of the American Chemical Society* 117.48 (1995), pp. 11946–11975.

[199] Alexander D MacKerell, Jr Nilesh Banavali, and Nicolas Foloppe. "Development and Current Status of the CHARMM Force Field for Nucleic Acids". In: (2001).

[200] Ignacio Faustino, Alberto Pérez, and Modesto Orozco. "Toward a Consensus View of Duplex RNA Flexibility". In: *Biophysical Journal* 99.6 (Sept. 2010), pp. 1876–1885.

[201] Elizabeth J. Denning et al. "Impact of 2'-hydroxyl sampling on the conformational properties of RNA: Update of the CHARMM all-atom additive force field for RNA". In: *Journal of Computational Chemistry* 32.9 (July 2011), pp. 1929–1943.

[202] Justin A. Lemkul and Alexander D. MacKerell. "Polarizable Force Field for DNA Based on the Classical Drude Oscillator: II. Microsecond Molecular Dynamics Simulations of Duplex DNA". In: *Journal of chemical theory and computation* 13.5 (May 2017), p. 2072.

[203] Justin A. Lemkul and Alexander D. MacKerell. "Polarizable Force Field for RNA Based on the Classical Drude Oscillator". In: *Journal of computational chemistry* 39.32 (Dec. 2018), p. 2624.

[204] Abhishek A. Kognole, Anthony Hazel, and Alexander D. MacKerell. "SILCS-RNA: Toward a Structure-Based Drug Design Approach for Targeting RNAs with Small Molecules". In: *Journal of Chemical Theory and Computation* 18.9 (Sept. 2022), pp. 5672–5691.

[205] Omkar Singh, Pushyaraga P. Venugopal, and Debashree Chakraborty. "Effect of Water Models on The Stability of RNA: Role of Counter-Ions". In: *Chemical Physics Impact* 7 (Dec. 2023), p. 100313.

[206] Saeed Izadi, Ramu Anandakrishnan, and Alexey V. Onufriev. "Building Water Models: A Different Approach". In: *The Journal of Physical Chemistry Letters* 5.21 (Nov. 2014), pp. 3863–3871.

[207] Stefano Piana et al. "Water dispersion interactions strongly influence simulated structural properties of disordered protein states". In: *Journal of Physical Chemistry B* 119.16 (Apr. 2015), pp. 5113–5123.

[208] Sarah A. Woodson. "Metal ions and RNA folding: a highly charged topic with a dynamic future". In: *Current opinion in chemical biology* 9.2 (2005), pp. 104–109.

[209] Petra Kührová et al. "Sensitivity of the RNA Structure to Ion Conditions as Probed by Molecular Dynamics Simulations of Common Canonical RNA Duplexes". In: *Journal of Chemical Information and Modeling* 63.7 (Apr. 2023), pp. 2133–2146.

[210] In Suk Joung and Thomas E. Cheatham. "Determination of Alkali and Halide Monovalent Ion Parameters for Use in Explicitly Solvated Biomolecular Simulations". In: *The Journal of Physical Chemistry B* 112.30 (July 2008), pp. 9020–9041.

[211] Jessica C. Bowman et al. "Cations in charge: magnesium ions in RNA folding and catalysis". In: *Current opinion in structural biology* 22.3 (June 2012), pp. 262–272.

[212] Lorenzo Casalino and Alessandra Magistrato. "Structural, dynamical and catalytic interplay between Mg2+ ions and RNA. Vices and virtues of atomistic simulations". In: *Inorganica Chimica Acta* 452 (Oct. 2016), pp. 73–81.

[213] Maria Carola Colombo et al. "Hybrid QM/MM Car-Parrinello Simulations of Catalytic and Enzymatic Reactions". In: *CHIMIA* 56.1-2 (Jan. 2002), p. 13.

[214] Olof Allnér, Lennart Nilsson, and Alessandra Villa. "Magnesium Ion–Water Coordination and Exchange in Biomolecular Simulations". In: *Journal of Chemical Theory and Computation* 8.4 (Apr. 2012), pp. 1493–1502.

[215] Michael V. Schrodt, Casey T. Andrews, and Adrian H. Elcock. "Large-Scale Analysis of 48 DNA and 48 RNA Tetranucleotides Studied by 1 μs Explicit-Solvent Molecular Dynamics Simulations". In: *Journal of Chemical Theory and Computation* 11.12 (Nov. 2015), pp. 5906–5917.

[216] Christina Bergonzo et al. "Highly sampled tetranucleotide and tetraloop motifs enable evaluation of common RNA force fields". In: *RNA* 21.9 (Sept. 2015), pp. 1578–1590.

[217] Jiří Šponer et al. "Molecular dynamics simulations of nucleic acids. from tetranucleotides to the ribosome". In: *Journal of Physical Chemistry Letters* 5.10 (May 2014), pp. 1771–1782.

[218] Albert C. Pan et al. "Demonstrating an Order-of-Magnitude Sampling Enhancement in Molecular Dynamics Simulations of Complex Protein Systems". In: *Journal of Chemical Theory and Computation* 12.3 (Mar. 2016), pp. 1360–1367.

[219] Rafael C. Bernardi, Marcelo C.R. Melo, and Klaus Schulten. "Enhanced sampling techniques in molecular dynamics simulations of biological systems". In: *Biochimica et Biophysica Acta (BBA) - General Subjects* 1850.5 (May 2015), pp. 872–877.

[220] Alessandro Barducci, Massimiliano Bonomi, and Michele Parrinello. "Metadynamics". In: *WIREs Computational Molecular Science* 1.5 (Sept. 2011), pp. 826–843.

[221] Andrew B. Ward, Andrej Sali, and Ian A. Wilson. "Integrative structural biology". In: *Science* 339.6122 (Feb. 2013), pp. 913–915.

[222] Mattia Bernetti and Giovanni Bussi. "Integrating experimental data with molecular simulations to investigate RNA structural dynamics". In: *Current Opinion in Structural Biology* 78 (Feb. 2023), p. 102503.

[223] E. T. Jaynes. "Information Theory and Statistical Mechanics". In: *Physical Review* 106.4 (May 1957), p. 620.

[224] Massimiliano Bonomi et al. "Metainference: A Bayesian inference method for heterogeneous systems". In: *Science Advances* 2.1 (Jan. 2016).

[225] Sandro Bottaro et al. "Conformational ensembles of RNA oligonucleotides from integrating NMR and molecular simulations". In: *Science Advances* 4.5 (May 2018), eaar8521.

[226] Christina Bergonzo, Alexander Grishaev, and Sandro Bottaro. "Conformational heterogeneity of UCAAUC RNA oligonucleotide from molecular dynamics simulations, SAXS, and NMR experiments". In: *RNA* 28.7 (July 2022), pp. 937–946.

[227] Omar Valsson, Pratyush Tiwary, and Michele Parrinello. "Enhancing Important Fluctuations: Rare Events and Metadynamics from a Conceptual Viewpoint". In: *Annual Review of Physical Chemistry* 67.1 (May 2016), pp. 159–184.

[228] G. M. Torrie and J. P. Valleau. "Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling". In: *Journal of Computational Physics* 23.2 (Feb. 1977), pp. 187–199.

[229] Alessandro Laio and Michele Parrinello. "Escaping free-energy minima". In: *PNAS* 99.20 (2002), pp. 12562–12566.

[230] Joseph A. Liberman et al. "Structural analysis of a class III preQ1 riboswitch reveals an aptamer distant from a ribosome-binding site regulated by fast dynamics". In: *Proceedings of the National Academy of Sciences of the United States of America* 112.27 (July 2015), E3485–E3494.

[231] Anna Bochicchio et al. "Molecular View of Ligands Specificity for CAG Repeats in Anti-Huntington Therapy". In: *Journal of Chemical Theory and Computation* 11.10 (Oct. 2015), pp. 4911–4922.

[232] Lorenzo Casalino et al. "Who Activates the Nucleophile in Ribozyme Catalysis? An Answer from the Splicing Mechanism of Group II Introns". In: *Journal of the American Chemical Society* 138.33 (Aug. 2016), pp. 10374–10377.

[233] Pallavi Thaplyal et al. "Inverse thio effects in the hepatitis delta virus ribozyme reveal that the reaction pathway is controlled by metal ion charge density". In: *Biochemistry* 54.12 (Mar. 2015), pp. 2160–2175.

[234] Jigneshkumar Dahyabhai Prajapati, José N. Onuchic, and Karissa Y. Sanbonmatsu. "Exploring the Energy Landscape of Riboswitches Using Collective Variables Based on Tertiary Contacts". In: *Journal of Molecular Biology* 434.18 (Sept. 2022), p. 167788.

[235] Sandro Bottaro, Francesco Di Palma, and Giovanni Bussi. "The role of nucleobase interactions in RNA structure and dynamics". In: *Nucleic Acids Research* 42.21 (Dec. 2014), pp. 13306–13314.

[236] Sandro Bottaro et al. "Free Energy Landscape of GAGA and UUCG RNA Tetraloops". In: *Journal of Physical Chemistry Letters* 7.20 (Oct. 2016), pp. 4032–4038.

[237] Paraskevi Gkeka et al. "Machine Learning Force Fields and Coarse-Grained Variables in Molecular Dynamics: Application to Materials and Biological Systems". In: *Journal of Chemical Theory and Computation* 16.8 (Aug. 2020), pp. 4757–4775.

[238] Yihang Wang et al. "Interrogating RNA-Small Molecule Interactions with Structure Probing and Artificial Intelligence-Augmented Molecular Simulations". In: *ACS Central Science* 8.6 (June 2022), pp. 741–748.

[239] Yuji Sugita and Yuko Okamoto. "Replica-exchange molecular dynamics method for protein folding". In: *Chemical Physics Letters* 314.1-2 (Nov. 1999), pp. 141–151.

[240] Vojtěch Mlýnský et al. "Toward Convergence in Folding Simulations of RNA Tetraloops: Comparison of Enhanced Sampling Techniques and Effects of Force Field Modifications". In: *Journal of Chemical Theory and Computation* 18.4 (Apr. 2022), pp. 2642–2656.

[241] Xu Xue, Wang Yongjun, and Li Zhihong. "Folding of SAM-II riboswitch explored by replica-exchange molecular dynamics simulation". In: *Journal of Theoretical Biology* 365 (Jan. 2015), pp. 265–269.

[242] Alejandro Gil-Ley and Giovanni Bussi. "Enhanced conformational sampling using replica exchange with collective-variable tempering". In: *Journal of Chemical Theory and Computation* 11.3 (Mar. 2015), pp. 1077–1085.

[243] JP Hughes et al. "Principles of early drug discovery". In: *British Journal of Pharmacology* 162.6 (Mar. 2011), pp. 1239–1249.

[244] Maria Batool, Bilal Ahmad, and Sangdun Choi. "A Structure-Based Drug Discovery Paradigm". In: *International Journal of Molecular Sciences 2019, Vol. 20, Page 2783* 20.11 (June 2019), p. 2783.

[245] Anastasiia V. Sadybekov and Vsevolod Katritch. "Computational approaches streamlining drug discovery". In: *Nature* 616.7958 (Apr. 2023), pp. 673–685.

[246] Victor T. Sabe et al. "Current trends in computer aided drug design and a highlight of drugs discovered via computational techniques: A review". In: *European Journal of Medicinal Chemistry* 224 (Nov. 2021), p. 113705.

[247] Tanaji Talele, Santosh Khedkar, and Alan Rigby. "Successful Applications of Computer Aided Drug Discovery: Moving Drugs from Concept to the Clinic". In: *Current Topics in Medicinal Chemistry* 10.1 (Jan. 2010), pp. 127–141.

[248] Mohit Pandey et al. "The transformational role of GPU computing and deep learning in drug discovery". In: *Nature Machine Intelligence 2022 4:3* 4.3 (Mar. 2022), pp. 211–221.

[249] Stephani Joy Y. Macalino et al. "Role of computer-aided drug design in modern drug discovery". In: *Archives of Pharmacal Research* 38.9 (Sept. 2015), pp. 1686–1701.

[250] Mark A. Johnson, Gerald M. Maggiora, and Calif.) American Chemical Society. Meeting (196th : 1988 : Los Angeles. "Concepts and applications of molecular similarity". In: (1990), p. 393.

[251] Chayan Acharya et al. "Recent Advances in Ligand-Based Drug Design: Relevance and Utility of the Conformationally Sampled Pharmacophore Approach". In: *Current computer-aided drug design* 7.1 (Mar. 2011), p. 10.

[252] Olivier Sperandio, Maria Miteva, and Bruno Villoutreix. "Combining Ligand- and Structure-Based Methods in Drug Design Projects". In: *Current Computer Aided-Drug Design* 4.3 (Sept. 2008), pp. 250–258.

[253] Malgorzata N. Drwal and Renate Griffith. "Combination of ligand- and structure-based methods in virtual screening". In: *Drug Discovery Today: Technologies* 10.3 (Sept. 2013), e395–e401.

[254] Jürgen Bajorath. "Integration of virtual and high-throughput screening". In: *Nature Reviews Drug Discovery 2002 1:11* 1.11 (Nov. 2002), pp. 882–894.

[255] Matthew D. Disney et al. "Inforna 2.0: A Platform for the Sequence-Based Design of Small Molecules Targeting Structured RNAs". In: *ACS Chemical Biology* 11.6 (June 2016), pp. 1720–1728.

[256] Aline Umuhire Juru, Neeraj N. Patwardhan, and Amanda E. Hargrove. "Understanding the Contributions of Conformational Changes, Thermodynamics, and Kinetics of RNA–Small Molecule Interactions". In: *ACS Chemical Biology* 14.5 (May 2019), pp. 824–838.

[257] David D. Boehr, Ruth Nussinov, and Peter E. Wright. "The role of dynamic conformational ensembles in biomolecular recognition". In: *Nature Chemical Biology 2009 5:11* 5.11 (Oct. 2009), pp. 789–796.

[258] F. Cramer. "Biochemical correctness: Emil Fischer's lock and key hypothesis, a hundred years after — an essay". In: *Pharmaceutica Acta Helvetiae* 69.4 (Mar. 1995), pp. 193–203.

[259] D. E. Koshland. "Application of a Theory of Enzyme Specificity to Protein Synthesis". In: *Proceedings of the National Academy of Sciences* 44.2 (Feb. 1958), pp. 98–104.

[260] Lise Pascale et al. "Thermodynamic studies of a series of homologous HIV-1 TAR RNA ligands reveal that loose binders are stronger Tat competitors than tight ones". In: *Nucleic Acids Research* 41.11 (June 2013), pp. 5851–5863.

[261] Qi Zhang et al. "Visualizing spatially correlated dynamics that directs RNA conformational transitions". In: *Nature* 450.7173 (Dec. 2007), pp. 1263–1267.

[262] Prabal Giri and Gopinatha Suresh Kumar. "Binding of protoberberine alkaloidcoralyne with double stranded poly(A): a biophysical study". In: *Molecular BioSystems* 4.4 (Mar. 2008), pp. 341–348.

[263] Jason R. Thomas, Xianjun Liu, and Paul J. Hergenrother. "Biochemical and thermodynamic characterization of compounds that bind to RNA hairpin loops: Toward an understanding of selectivity". In: *Biochemistry* 45.36 (Sept. 2006), pp. 10928–10938.

[264] Peter J. Tonge. "Drug–Target Kinetics in Drug Discovery". In: *ACS Chemical Neuroscience* 9.1 (Jan. 2018), pp. 29–39.

[265] Amy Davidson et al. “A Small-Molecule Probe Induces a Conformation in HIV TAR RNA Capable of Binding Drug-Like Fragments”. In: *Journal of Molecular Biology* 410.5 (July 2011), pp. 984–996.

[266] Gregory M. Lee and Charles S. Craik. “Trapping moving targets with small molecules”. In: *Science (New York, N.Y.)* 324.5924 (Apr. 2009), pp. 213–215.

[267] Laura R. Ganser et al. “Probing RNA Conformational Equilibria within the Functional Cellular Context”. In: *Cell Reports* 30.8 (Feb. 2020), pp. 2472–2480.

[268] Thomas Hermann. “Rational ligand design for RNA: the role of static structure and conformational flexibility in target recognition”. In: *Biochimie* 84.9 (Sept. 2002), pp. 869–875.

[269] Frank Walter et al. “Binding of tobramycin leads to conformational changes in yeast tRNAAsp and inhibition of aminoacylation”. In: *EMBO Journal* 21.4 (Feb. 2002), pp. 760–768.

[270] Klara Aradi, Audrey Di Giorgio, and Maria Duca. “Aminoglycoside Conjugation for RNA Targeting: Antimicrobials and Beyond”. In: *Chemistry (Weinheim an der Bergstrasse, Germany)* 26.54 (Sept. 2020), pp. 12273–12309.

[271] Kenneth F. Blount and Yitzhak Tor. “A tale of two targets: differential RNA selectivity of nucleobase-aminoglycoside conjugates”. In: *Chembiochem : a European journal of chemical biology* 7.10 (Oct. 2006), pp. 1612–1621.

[272] Ricardo MacArron et al. “Impact of high-throughput screening in biomedical research”. In: *Nature Reviews Drug Discovery 2011 10:3* 10.3 (Mar. 2011), pp. 188–195.

[273] Takashi Motoyaji. “Revolution of Small Molecule Drug Discovery by Affinity Selection-Mass Spectrometry Technology”. In: *Chemical and Pharmaceutical Bulletin* 68.3 (Mar. 2020), pp. 191–193.

[274] Noreen F. Rizvi and Elliott B. Nickbarg. “RNA-ALIS: Methodology for screening soluble RNAs as small molecule targets using ALIS affinity-selection mass spectrometry”. In: *Methods* 167 (Sept. 2019), pp. 28–38.

[275] Arturo J. Vegas, Jason H. Fuller, and Angela N. Koehler. “Small-molecule microarrays as tools in ligand discovery”. In: *Chemical Society Reviews* 37.7 (June 2008), pp. 1385–1394.

[276] Raman Parkesh et al. “Design of a bioactive small molecule that targets the myotonic dystrophy type 1 RNA via an RNA motif-ligand database and chemical similarity searching”. In: *Journal of the American Chemical Society* 134.10 (Mar. 2012), pp. 4731–4742.

[277] Sai Pradeep Velagapudi, Steven M Gallo, and Matthew D Disney. “Sequence-based design of bioactive small molecules that target precursor microRNAs”. In: *Nature Chemical Biology* 10.4 (Apr. 2014), pp. 291–297.

[278] Kayleigh R. McGovern-Gooch and Nathan J. Baird. “Fluorescence-based investigations of RNA-small molecule interactions”. In: *Methods* 167 (Sept. 2019), pp. 54–65.

[279] Jesse Davila-Calderon et al. “IRES-targeting small molecule inhibits enterovirus 71 replication via allosteric stabilization of a ternary complex”. In: *Nature Communications 2020 11:1* 11.1 (Sept. 2020), pp. 1–13.

[280] Daniel A. Erlanson et al. “Twenty years on: the impact of fragments on drug discovery”. In: *Nature Reviews Drug Discovery 2016 15:9* 15.9 (July 2016), pp. 605–619.

[281] Kasper P. Lundquist et al. “Fragment-Based Drug Discovery for RNA Targets”. In: *ChemMedChem* 16.17 (Sept. 2021), pp. 2588–2603.

[282] Brett A. Tounge and Michael H. Parker. “Designing a Diverse High-Quality Library for Crystallography-Based FBDD Screening”. In: *Methods in Enzymology* 493 (Jan. 2011), pp. 3–20.

[283] Roba Moumné et al. “Fragment-based design of small RNA binders: Promising developments and contribution of NMR”. In: *Biochimie* 94.7 (July 2012), pp. 1607–1619.

[284] Liuhong Chen et al. “A fragment-based approach to identifying ligands for riboswitches”. In: *ACS Chemical Biology* 5.4 (Apr. 2010), pp. 355–358.

[285] Miguel Garavís et al. “Discovery of selective ligands for telomeric RNA G-quadruplexes (TERRA) through 19F-NMR based fragment screening”. In: *ACS Chemical Biology* 9.7 (July 2014), pp. 1559–1566.

[286] Matthew A. Clark et al. “Design, synthesis and selection of DNA-encoded small-molecule libraries”. In: *Nature Chemical Biology 2009 5:9* 5.9 (Aug. 2009), pp. 647–654.

[287] Willy Decurtins et al. “Automated screening for small organic ligands using DNA-encoded chemical libraries”. In: *Nature Protocols 2016 11:4* 11.4 (Mar. 2016), pp. 764–780.

[288] Raphael I. Benhamou et al. “DNA-encoded library versus RNA-encoded library selection enables design of an oncogenic noncoding RNA inhibitor”. In: *Proceedings of the National Academy of Sciences of the United States of America* 119.6 (Feb. 2022), e2114971119.

[289] Fabien Vincent et al. “Phenotypic drug discovery: recent successes, lessons learned and new directions”. In: *Nature Reviews Drug Discovery 2022 21:12* 21.12 (May 2022), pp. 899–914.

[290] Alleyn T. Plowright and Lauren Drowley. “Phenotypic Screening”. In: *Annual Reports in Medicinal Chemistry* 50 (2017), pp. 263–299.

[291] Nikolai A. Naryshkin et al. “SMN2 splicing modifiers improve motor function and longevity in mice with spinal muscular atrophy”. In: *Science* 345.6197 (Aug. 2014), pp. 688–693.

[292] James Palacino et al. “SMN2 splice modulators enhance U1–pre-mRNA association and rescue SMA mice”. In: *Nature Chemical Biology 2015 11:7* 11.7 (June 2015), pp. 511–517.

[293] Gopinatha Suresh Kumar and Anirban Basu. “The use of calorimetry in the biophysical characterization of small molecule alkaloids binding to RNA structures”. In: *Biochimica et Biophysica Acta (BBA) - General Subjects* 1860.5 (May 2016), pp. 930–944.

[294] Michelle H. Moon et al. “Measuring RNA-Ligand Interactions with Microscale Thermophoresis”. In: *Biochemistry* 57.31 (Aug. 2018), pp. 4638–4643.

[295] Tam Vo et al. “Biosensor-surface plasmon resonance: A strategy to help establish a new generation RNA-specific small molecules”. In: *Methods* 167 (Sept. 2019), pp. 15–27.

[296] Jessica L. Childs-Disney et al. “Structure of the myotonic dystrophy type 2 RNA and designed small molecules that reduce toxicity”. In: *ACS Chemical Biology* 9.2 (Feb. 2014), pp. 538–550.

[297] Aline Umuhire Juru and Amanda E. Hargrove. "Frameworks for targeting RNA with small molecules". In: *Journal of Biological Chemistry* 296 (2021), p. 100191.

[298] Zhichao Tang et al. "Recognition of single-stranded nucleic acids by small-molecule splicing modulators". In: *Nucleic Acids Research* 49.14 (2021), pp. 7870–7883.

[299] Mattia Bernetti et al. "Computational drug discovery under RNA times". In: *QRB Discovery* 3 (Nov. 2022), e22.

[300] Jacopo Manigrasso, Marco Marcia, and Marco De Vivo. "Computer-aided design of RNA-targeted small molecules: A growing need in drug discovery". In: *Chem* 7.11 (Nov. 2021), pp. 2965–2988.

[301] Yuanzhe Zhou, Yangwei Jiang, and Shi-Jie Chen. "RNA–ligand molecular docking: Advances and challenges". In: *WIREs Computational Molecular Science* 12.3 (May 2022).

[302] Daniel Alvarez-Garcia and Xavier Barril. "Molecular Simulations with Solvent Competition Quantify Water Displaceability and Provide Accurate Interaction Maps of Protein Binding Sites". In: *Journal of Medicinal Chemistry* 57.20 (Oct. 2014), pp. 8530–8539.

[303] P. J. Goodford. "A computational procedure for determining energetically favorable binding sites on biologically important macromolecules". In: *Journal of Medicinal Chemistry* 28.7 (July 1985), pp. 849–857.

[304] Jianghong An, Maxim Totrov, and Ruben Abagyan. "Pocketome via comprehensive identification and classification of ligand binding envelopes". In: *Molecular & cellular proteomics : MCP* 4.6 (June 2005), pp. 752–761.

[305] Thomas A. Halgren. "Identifying and Characterizing Binding Sites and Assessing Druggability". In: *Journal of Chemical Information and Modeling* 49.2 (Feb. 2009), pp. 377–389.

[306] Jian Yu et al. "Roll: a new algorithm for the detection of protein pockets and cavities with a rolling probe sphere". In: *Bioinformatics* 26.1 (Jan. 2010), pp. 46–52.

[307] N. R. Voss and M. Gerstein. "3V: cavity, channel and cleft volume calculator and extractor". In: *Nucleic Acids Research* 38.Web Server (July 2010), W555–W562.

[308] Pradeep Anand Ravindranath and Michel F. Sanner. "AutoSite: an automated approach for pseudo-ligands prediction—from ligand-binding sites identification to predicting key ligand atoms". In: *Bioinformatics* 32.20 (Oct. 2016), pp. 3142–3149.

[309] Damien Monet et al. "mkgridXf : Consistent Identification of Plausible Binding Sites Despite the Elusive Nature of Cavities and Grooves in Protein Dynamics". In: *Journal of Chemical Information and Modeling* 59.8 (Aug. 2019), pp. 3506–3518.

[310] Pan Zeng et al. "Rsite: a computational method to identify the functional sites of noncoding RNAs". In: *Scientific Reports* 5.1 (Mar. 2015), p. 9179.

[311] Pan Zeng and Qinghua Cui. "Rsite2: an efficient computational method to predict the functional sites of noncoding RNAs". In: *Scientific Reports* 6.1 (Jan. 2016), p. 19016.

[312] Huiwen Wang and Yunjie Zhao. "RBinds: A user-friendly server for RNA binding site prediction". In: *Computational and Structural Biotechnology Journal* 18 (Jan. 2020), pp. 3762–3765.

[313] Hong Su, Zhenling Peng, and Jianyi Yang. "Recognition of small molecule–RNA binding sites using RNA sequence and structure". In: *Bioinformatics* 37.1 (Apr. 2021), pp. 36–42.

[314] Igor Kozlovskii and Petr Popov. "Structure-based deep learning for binding site detection in nucleic acid macromolecules". In: *NAR Genomics and Bioinformatics* 3.4 (Oct. 2021).

[315] Kaili Wang et al. "RLBind: a deep learning method to predict RNA–ligand binding sites". In: *Briefings in Bioinformatics* 24.1 (Jan. 2023), bbac486.

[316] Eva Chovancova et al. "CAVER 3.0: A Tool for the Analysis of Transport Pathways in Dynamic Protein Structures". In: *PLOS Computational Biology* 8.10 (2012), e1002708.

[317] Peter Schmidtke et al. "MDpocket: open-source cavity detection and characterization on molecular dynamics trajectories". In: *Bioinformatics* 27.23 (Dec. 2011), pp. 3276–3285.

[318] Phani Ghanakota and Heather A. Carlson. "Driving Structure-Based Drug Discovery through Cosolvent Molecular Dynamics". In: *Journal of Medicinal Chemistry* 59.23 (Dec. 2016), pp. 10383–10399.

[319] Gerard Martinez-Rosell et al. "PlayMolecule CrypticScout: Predicting Protein Cryptic Sites Using Mixed-Solvent Molecular Simulations". In: *Journal of Chemical Information and Modeling* 60.4 (Apr. 2020), pp. 2314–2324.

[320] Christina E. Faller et al. "Site Identification by Ligand Competitive Saturation (SILCS) Simulations for Fragment-Based Drug Design". In: 2015, pp. 75–87.

[321] Lucas Defelipe et al. "Solvents to Fragments to Drugs: MD Applications in Drug Design". In: *Molecules* 23.12 (Dec. 2018), p. 3269.

[322] Maicol Bissaro, Mattia Sturlese, and Stefano Moro. "Exploring the RNA-Recognition Mechanism Using Supervised Molecular Dynamics (SuMD) Simulations: Toward a Rational Design for Ribonucleic-Targeting Molecules?" In: *Frontiers in Chemistry* 8 (Feb. 2020), p. 515945.

[323] Kaili Wang et al. "RBind: computational network method to predict RNA binding sites". In: *Bioinformatics* 34.18 (Sept. 2018), pp. 3131–3136.

[324] Greta Bagnolini, TinTin B. Luu, and Amanda E. Hargrove. "Recognizing the power of machine learning and other computational methods to accelerate progress in small molecule targeting of RNA". In: *RNA* 29.4 (Apr. 2023), pp. 473–488.

[325] Suresh Dara et al. "Machine Learning in Drug Discovery: A Review". In: *Artificial Intelligence Review 2021 55:3* 55.3 (Aug. 2021), pp. 1947–1999.

[326] Lucas S.P. Rudden, Mahdi Hijazi, and Patrick Barth. "Deep learning approaches for conformational flexibility and switching properties in protein design". In: *Frontiers in Molecular Biosciences* 9 (Aug. 2022), p. 928534.

[327] Tadeo Saldanõ et al. "Impact of protein conformational diversity on AlphaFold predictions". In: *Bioinformatics* 38.10 (May 2022), pp. 2742–2748.

[328] Jun Zhang et al. "Artificial Intelligence Enhanced Molecular Simulations". In: *Journal of Chemical Theory and Computation* 19.14 (July 2023), pp. 4338–4350.

[329] Zhuoran Qiao et al. "State-specific protein–ligand complex structure prediction with a multiscale deep generative model". In: *Nature Machine Intelligence 2024* (Feb. 2024), pp. 1–14.

[330] Anna Maria Ferrari et al. "Soft docking and multiple receptor conformations in virtual screening". In: *Journal of Medicinal Chemistry* 47.21 (Oct. 2004), pp. 5076–5084.

[331] Ruben Abagyan, Maxim Totrov, and Dmitry Kuznetsov. "ICM—A new method for protein modeling and design: Applications to docking and structure prediction from the distorted native conformation". In: *Journal of Computational Chemistry* 15.5 (May 1994), pp. 488–506.

[332] Gareth Jones et al. "Development and validation of a genetic algorithm for flexible docking 1 1Edited by F. E. Cohen". In: *Journal of Molecular Biology* 267.3 (Apr. 1997), pp. 727–748.

[333] Garrett M Morris et al. "Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function". In: *Journal of computational chemistry* 19.14 (1998), pp. 1639–1662.

[334] S. David Morley and Mohammad Afshar. "Validation of an empirical RNA-ligand scoring function for fast flexible docking using RiboDock®". In: *Journal of Computer-Aided Molecular Design* 18.3 (2004), pp. 189–208.

[335] Richard A. Friesner et al. "Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy". In: *Journal of Medicinal Chemistry* 47.7 (Mar. 2004), pp. 1739–1749.

[336] Oliver Korb, Thomas Stützle, and Thomas E. Exner. "An ant colony optimization approach to flexible protein–ligand docking". In: *Swarm Intelligence* 1.2 (Nov. 2007), pp. 115–134.

[337] Christopher R. Corbeil, Pablo Englebienne, and Nicolas Moitessier. "Docking Ligands into Flexible and Solvated Macromolecules. 1. Development and Validation of FITTED 1.0". In: *Journal of Chemical Information and Modeling* 47.2 (Mar. 2007), pp. 435–449.

[338] Oleg Trott and Arthur J. Olson. "AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading". In: *Journal of Computational Chemistry* 31.2 (Jan. 2010), pp. 455–461.

[339] Sergio Ruiz-Carmona et al. "rDock: A Fast, Versatile and Open Source Program for Docking Ligands to Proteins and Nucleic Acids". In: *PLoS Computational Biology* 10.4 (Apr. 2014), e1003571.

[340] Ajay N. Jain. "Surflex: Fully Automatic Flexible Molecular Docking Using a Molecular Similarity-Based Search Engine". In: *Journal of Medicinal Chemistry* 46.4 (Feb. 2003), pp. 499–511.

[341] P. Therese Lang et al. "DOCK 6: combining techniques to model RNA-small molecule complexes". In: *RNA (New York, N.Y.)* 15.6 (June 2009), pp. 1219–1230.

[342] Christophe Guilbert and Thomas L. James. "Docking to RNA via Root-Mean-Square-Deviation-Driven Energy Minimization with Flexible Ligands and Flexible Targets". In: *Journal of Chemical Information and Modeling* 48.6 (June 2008), pp. 1257–1268.

[343] Yangwei Jiang and Shi-Jie Chen. "RLDOCK method for predicting RNA-small molecule binding modes". In: *Methods* 197 (Jan. 2022), pp. 97–105.

[344] Yuyu Feng et al. "NLDock: a Fast Nucleic Acid–Ligand Docking Algorithm for Modeling RNA/DNA–Ligand Complexes". In: *Journal of Chemical Information and Modeling* 61.9 (Sept. 2021), pp. 4771–4782.

[345] Bohdan Waszkowycz, David E. Clark, and Emanuela Gancia. "Outstanding challenges in protein–ligand docking and structure-based virtual screening". In: *Wiley Interdisciplinary Reviews: Computational Molecular Science* 1.2 (Mar. 2011), pp. 229–259.

[346] J. A. Nelder and R. Mead. "A Simplex Method for Function Minimization". In: *The Computer Journal* 7.4 (Jan. 1965), pp. 308–313.

[347] Xuan-Yu Meng et al. "Molecular Docking: A powerful approach for structure-based drug discovery". In: *Current computer-aided drug design* 7.2 (June 2011), p. 146.

[348] Christophe Guilbert, David Perahia, and Liliane Mouawad. "A method to explore transition paths in macromolecules. Applications to hemoglobin and phosphoglycerate kinase". In: *Computer Physics Communications* 91.1-3 (Sept. 1995), pp. 263–273.

[349] Laura R. Ganser et al. "Demonstration that Small Molecules can Bind and Stabilize Low-abundance Short-lived RNA Excited Conformational States". In: *Journal of Molecular Biology* 432.4 (Feb. 2020), pp. 1297–1304.

[350] Nicolas Moitessier, Eric Westhof, and Stephen Hanessian. "Docking of aminoglycosides to hydrated and flexible RNA". In: *Journal of Medicinal Chemistry* 49.3 (Feb. 2006), pp. 1023–1033.

[351] So Jung Park, Yang Gyun Kim, and Hyun Ju Park. "Identification of rna pseudoknot-binding ligand that inhibits the - 1 ribosomal frameshifting of SARS-coronavirus by structure-based virtual screening". In: *Journal of the American Chemical Society* 133.26 (July 2011), pp. 10094–10100.

[352] Yuhong Zhu, Zhaojian He, and Shi Jie Chen. "TBI server: A web server for predicting ion effects in RNA folding". In: *PLoS ONE* 10.3 (Mar. 2015).

[353] Li Zhen Sun, Jing Xiang Zhang, and Shi Jie Chen. "MCTBI: a web server for predicting metal ion effects in RNA structures". In: *RNA (New York, N.Y.)* 23.8 (Aug. 2017), pp. 1155–1165.

[354] George M. Giambaşu et al. "Ion counting from explicit-solvent simulations and 3D-RISM". In: *Biophysical journal* 106.4 (Feb. 2014), pp. 883–894.

[355] Wanlei Wei et al. "Predicting Positions of Bridging Water Molecules in Nucleic Acid-Ligand Complexes". In: *Journal of Chemical Information and Modeling* 59.6 (June 2019), pp. 2941–2951.

[356] Rocco Meli, Garrett M. Morris, and Philip C. Biggin. "Scoring Functions for Protein-Ligand Binding Affinity Prediction Using Structure-based Deep Learning: A Review". In: *Frontiers in Bioinformatics* 2 (June 2022), p. 885983.

[357] William J. Allen et al. "DOCK 6: Impact of New Features and Current Docking Performance". In: *Journal of computational chemistry* 36.15 (June 2015), p. 1132.

[358] Di Qiu et al. "The GB/SA Continuum Model for Solvation. A Fast Analytical Method for the Calculation of Approximate Born Radii". In: *The Journal of Physical Chemistry A* 101.16 (Apr. 1997), pp. 3005–3014.

[359] Ruth Huey et al. "A semiempirical free energy force field with charge-based desolvation". In: *Journal of Computational Chemistry* 28.6 (Apr. 2007), pp. 1145–1152.

[360] Lu Chen, George A. Calin, and Shuxing Zhang. “Novel insights of structure-based modeling for RNA-targeted drug discovery”. In: *Journal of Chemical Information and Modeling* 52.10 (Oct. 2012), pp. 2741–2753.

[361] Dennis M. Krüger et al. “Target Flexibility in RNA−Ligand Docking Modeled by Elastic Potential Grids”. In: *ACS Medicinal Chemistry Letters* 2.7 (July 2011), pp. 489–493.

[362] Xiaoyu Zhao et al. “An improved PMF scoring function for universally predicting the interactions of a ligand with protein, DNA, and RNA”. In: *Journal of Chemical Information and Modeling* 48.7 (2008), pp. 1438–1447.

[363] Anna Philips et al. “LigandRNA: computational predictor of RNA–ligand interactions”. In: *RNA* 19.12 (Dec. 2013), pp. 1605–1616.

[364] Zhiqiang Yan and Jin Wang. “SPA-LN: a scoring function of ligand–nucleic acid interactions via optimizing both specificity and affinity”. In: *Nucleic Acids Research* 45.12 (July 2017), e110–e110.

[365] Zixuan Cang and Guo-Wei Wei. “Integration of element specific persistent homology and machine learning for protein-ligand binding affinity prediction”. In: *International Journal for Numerical Methods in Biomedical Engineering* 34.2 (Feb. 2018).

[366] Yuyu Feng and Sheng You Huang. “ITScore-NL: An Iterative Knowledge-Based Scoring Function for Nucleic Acid-Ligand Interactions”. In: *Journal of Chemical Information and Modeling* 60.12 (Dec. 2020), pp. 6698–6708.

[367] Maciej Wójcikowski, Pedro J. Ballester, and Pawel Siedlecki. “Performance of machine-learning scoring functions in structure-based virtual screening”. In: *Scientific Reports* 7.1 (Apr. 2017), p. 46710.

[368] Sahil Chhabra, Jingru Xie, and Aaron T. Frank. “RNAPosers: Machine Learning Classifiers for Ribonucleic Acid–Ligand Poses”. In: *The Journal of Physical Chemistry B* 124.22 (June 2020), pp. 4436–4445.

[369] Filip Stefaniak and Janusz M. Bujnicki. “AnnapuRNA: A scoring function for predicting RNA-small molecule binding poses”. In: *PLOS Computational Biology* 17.2 (Feb. 2021), e1008309.

[370] Paul D. Thomas and Ken A. Dill. “An iterative method for extracting energy-like quantities from protein structures.” In: *Proceedings of the National Academy of Sciences* 93.21 (Oct. 1996), pp. 11628–11633.

[371] Patrick Pfeffer and Holger Gohlke. “DrugScore ^{RNA} Knowledge-Based Scoring Function To Predict RNA−Ligand Interactions”. In: *Journal of Chemical Information and Modeling* 47.5 (Sept. 2007), pp. 1868–1876.

[372] Eduardo Habib Bechelane Maia et al. “Structure-Based Virtual Screening: From Classical to Artificial Intelligence”. In: *Frontiers in Chemistry* 8 (Apr. 2020).

[373] Dejun Jiang et al. “How Good Are Current Docking Programs at Nucleic Acid-Ligand Docking? A Comprehensive Evaluation”. In: *Journal of Chemical Theory and Computation* 19.16 (Aug. 2023), pp. 5633–5647.

[374] Ken Garber. “Drugging RNA”. In: *Nature Biotechnology* 41.6 (June 2023), pp. 745–749.

[375] Ankita Mehta et al. “SMMRNA: a database of small molecule modulators of RNA”. In: *Nucleic Acids Research* 42.D1 (Jan. 2014), pp. D132–D141.

[376] Subodh Kumar Mishra and Amit Kumar. “NALDB: nucleic acid ligand database for small molecules targeting nucleic acid”. In: *Database* 2016 (Feb. 2016), baw002.

[377] Saisai Sun, Jianyi Yang, and Zhaolei Zhang. “RNALigands: a database and web server for RNA–ligand interactions”. In: *RNA* 28.2 (Feb. 2022), pp. 115–122.

[378] Ting Zhou et al. “RPocket: an intuitive database of RNA pocket topology information with RNA-ligand data resources”. In: *BMC Bioinformatics* 22.1 (Dec. 2021), p. 428.

[379] Brittany S. Morgan et al. “R-BIND: An Interactive Database for Exploring and Developing RNA-Targeted Chemical Probes”. In: *ACS Chemical Biology* 14.12 (Dec. 2019), pp. 2691–2700.

Chapter 2

HARIBOSS: a curated database of RNA-small molecules structures to aid rational drug design

RNA molecules are implicated in numerous fundamental biological processes and many human pathologies, such as cancer, neurodegenerative disorders, muscular diseases, and bacterial infections. Modulating the mode of action of disease-implicated RNA molecules can lead to the discovery of new therapeutical agents and even address pathologies linked to 'undruggable' protein targets. This modulation can be achieved by direct targeting of RNA with small molecules. As of today, only a few RNA-targeting small molecules are used clinically. One of the main obstacles that has hampered the development of a rational drug design protocol to target RNA with small molecules is the lack of a comprehensive understanding of the molecular mechanisms at the basis of RNA-small molecule recognition. Here, we present HARIBOSS, a curated collection of RNA-small molecule structures determined by X-ray crystallography, nuclear magnetic Resonance spectroscopy and cryo-electron microscopy. HARIBOSS facilitates the exploration of drug-like compounds known to bind RNA, the analysis of ligands and pockets properties, and ultimately the development of *in silico* strategies to identify RNA-targeting small molecules. HARIBOSS can be explored via a web interface available at https://hariboss.pasteur.cloud.

2.1 Introduction

During the past two decades, RNA molecules have been shown to perform a variety of vital biological functions besides being a passive carrier of genetic information from DNA to protein. An explosion of research in the field of RNA biology has provided information about RNA diversity with several new definitions of RNA types as well as structural and functional information. For example, it is now well established that RNA is implicated in the regulation of gene expression at the levels of transcription, RNA processing and translation, in the regulation of epigenetic modifications and in the protection of the nucleus from foreign nucleic acids [1–4]. In conjunction with these discoveries, modulating RNA function is emerging as a promising therapeutic approach against pathologies such as cancer, viral infections, cardiovascular and muscular diseases, and neurodegenerative disorders [5, 6]. Traditionally, modulation of coding and non-coding RNA has been achieved using oligonucleotides such as small interfering RNA (siRNA), antisense, aptamers, and other RNA moieties or direct RNA-editing by CRISPR-Cas9 [7–10]. While oligonucleotides have been successful in binding to and modulating RNA, their drug bioavailability and membrane penetration have been quite challenging. Moreover, part of these molecules carry a large negative charge and, therefore, are susceptible to degradation by RNAses [11–13]. Small molecules able to selectively bind to RNA provide a more attractive alternative from a bioavailability and delivery perspective [14–17].

Direct targeting of coding and non-coding RNA with small molecules has the potential to be revolutionary. Only 1.5% of the human genome encodes proteins and only a small fraction (12%) of this percentage is related to diseases and targeted by existing drugs (3%). Strategies like targeting non-coding RNAs, which represent instead the majority of the human genome, or the mRNA of undruggable proteins will therefore allow to significantly expand the space of potential targets [18]. Several small molecules that interfere with RNA functions have already been identified [19–23], suggesting that therapeutics based on small molecules targeting RNA may be possible. However, only a few compounds have been approved so far by the United States Food and Drug Administration (FDA), namely linezolid, ribocil and risdiplam. Linezolid (Zyvox), initially discovered in the mid 1990s and approved for commercial use in 2000, is a broad-spectrum antibacterial agent that binds to the large RNA subunit of the ribosome and interferes with the positioning of the tRNA [24]. Ribocil, discovered by Merck in 2015, is a drug that selectively binds the flavin mononucleotide (FMN) riboswitch (RNA-mediated regulator of gene expression in bacteria) and silences gene expression, making it effective in the treatment of bacterial infections [25]. Interestingly, ribocil binds in the same pocket as the natural flavin mononucleotide ligand. Risdiplam, discovered by Roche-Genentech in 2018, is a drug that modulates the splicing of the Survival Motor Neuron 2 (SMN2) mRNA and mitigates the pathological SMN2 protein states related to Spinal Muscular Atrophy [26]. All these three compounds as well as most of those under pre-clinical or clinical evaluation have been discovered using loss/gain-of-function studies, phenotypic screening, or animal models.

Computer-aided drug design (CADD) has the potential to guide the rational development of small molecules targeting RNA [27, 28]. To date, this strategy is hampered by our limited understanding of RNA structural and dynamic properties as well as of the mechanisms of RNA-small molecule (RNA-SM) recognition [14, 18]. Previous efforts to characterize the physico-chemical prop-

erties of RNA binders and their intersection with the chemical space of drugs targeting proteins provided insights into their molecular interaction with RNA [16, 29–31]. Databases that collect all the known compounds binding RNA can be exploited for ligand- [32, 33] and 2D structure-based [34–36] virtual screening. DrugPred_RNA is, to the best of our knowledge, the only tool that has been trained and tuned using 3D structure data in order to characterize RNA binding sites [37]. In addition, most of current CADD pipelines have been developed for protein targets and might not be directly applicable to RNA. A first step in the development of structure-based approaches therefore requires an extensive assessment of the existing tools for pocket detection and ligand docking and possibly the development of new tools that exploit all the available structural information on RNA-SM complexes. While there have been previous efforts to collect such data [38–41], there is currently no comprehensive, curated, and regularly updated repository available to the scientific community.

Here, we present "Harnessing RIBOnucleic acid – Small molecule Structures" (HARIBOSS), a curated online database of RNA-SM structures. Each entry in HARIBOSS corresponds to a structure deposited in the RCSB Protein Data Bank (PDB) database [42] containing at least one chemical compound matching a list of basic drug-like criteria and interacting with at least one RNA chain. Ligands are annotated by their physico-chemical properties and by the number and composition of interacting RNA molecules. RNA pockets occupied by ligands are characterized in terms of geometric properties, such as volume, hydrophobicity and hydrophilicity, and overall propensity to bind small molecules and drug-like compounds. HARIBOSS is freely accessible via a web interface (`http://hariboss.pasteur.cloud`) and will facilitate understanding the nature of the interactions that drive RNA-SM recognition and benchmarking existing tools for in silico drug design with RNA targets.

2.2 Materials and Methods

2.2.1 Construction of the database

The HARIBOSS database was built in 3 steps (Fig. 2.1A), which were implemented in a series of python scripts. The operations described below are performed on a monthly basis to update HARIBOSS with the new structures deposited in the PDB.

I - Initial fetching from the PDB

The first step consisted of querying the PDB and collecting all the structures that contained at least one RNA molecule and one ligand. This operation was performed using a solr-based search API developed by the European Molecular Biology Laboratory of the European Bioinformatics Institut (EMBL-EBI) [43]. At this stage, we collected structures in which small molecules interact with RNA, DNA, proteins or a combination of these biological entities.

II - Identifying structures with drug-like compounds bound to RNA

Among the initially fetched structures, we selected those that contained drug-like compounds non-covalently bound only to RNA. To accomplish this, we adopted the following procedure. For each

entry, the correspondent mmCIF file was processed by MDAnalysis v. 2.0.0-beta [44] and OpenMM v 7.5.1 [45] to classify its constituents into RNA, DNA, protein, ions, and water molecules. Modified residues were assigned to RNA, DNA, or protein molecules based on the information about covalent bonds retrieved from the mmCIF with Biopython v. 1.79 [46]. All the molecules not included in the categories defined above were classified as ligands. We considered a ligand to be a drug-like compound if it satisfied the following criteria [38]:

- mass within 160 and 1000 Da, as reported in the mmCIF file [46];
- presence of at least one C atom;
- presence of only C, H, N, O, Br, Cl, F, P, Si, B, S, Se atoms;

For each compound fulfilling these criteria, we defined as interacting all the molecules within 6 Å from the ligand atoms. RNA chains with less than 10 atoms in the radius of 6 Å around the ligand were not considered as interacting (Table 2.1). For the first release of HARIBOSS, we retained in the database only the structures in which ligands interact exclusively with RNA chains. At this stage, we annotated each entry with the following information: total mass of the system, molecular composition (RNA, RNA/DNA, RNA/protein, RNA/DNA/protein), experimental method used to determine the structure (experiment type, resolution and deposition year), properties of bound ligands (PDB name, residue number and chain id, molecular mass, identity of the interacting RNA chains). The RNA-SM complexes obtained at the end of this filtering step constitute the *redundant* version of the HARIBOSS database.

III - Clustering of RNA/ligand complexes

We defined a clustering procedure to select representative RNA/ligand complexes and build a *non-redundant* version of HARIBOSS. Since the database contained ligands interacting with more than one RNA chain, we adopted the following procedure:

1. We created a FASTA file with all the sequences of the individual RNA chains, each annotated with its interacting ligand, which we processed with CD-HIT-EST [47] to cluster them at 90% sequence identity.
2. In case of ligands interacting with multiple RNA chains, we defined two structures to belong to the same cluster if the individual chains from the two entries were clustered together at step 1.
3. To obtain a variety of different RNA/ligand interactions, RNA chains belonging to the same cluster but bound to different ligands were classified in separate subclusters.
4. These subclusters were further classified based on the structural similarity of the pocket atoms by performing a hierarchical clustering using the RMSD of the aligned pocket atoms of the nucleic backbone as metrics and a cutoff of 1.5 Å.
5. For each cluster, we selected as representative the entry closest to the cluster center with the highest experimental resolution, which we consider an appropriate choice for the use of HARIBOSS in computational structural studies.

2.2.2 Analyzing the database

The HARIBOSS entries were analyzed based on general information about the structure, physico-chemical properties of the ligands and of the RNA pockets.

Exploration of the chemical space of RNA binders. To present an overview of the chemical space of the RNA binders included in our database, the TMAP visualization method was used [48]. In TMAP, the molecules are represented by their fingerprints and indexed in an LSH forest structure. Based on the distances calculated during this step, an undirected weighted c-approximate k-nearest neighbor graph (c-k-NNG) is used to construct a minimum spanning tree. This tree is then projected onto the Euclidean plane. For the creation of the spanning tree, the MHFP6 fingerprints and a point size of 20 were used. For more information about the method, we refer the readers to the original publication [48]. TMAP was chosen over simple clustering or other chemical space visualization tools for the informative nature of its tree-based layout, which enables to locally as well as globally locate specific clusters and their relative position compared to other clusters. Moreover, such an approach will be able to accommodate in the future the expanding chemical space of RNA binders.

Analysis of the ligand properties. QikProp (Schrodinger Suite 2021.v3) was employed for the calculation of the following ligand properties: solvent-accessible surface area (SASA) and its hydrophobic and hydrophilic content (FOSA and FISA respectively), predicted IC50 value for blockage of HERG K+ channels, Caco-2 cell permeability and brain/blood partition coefficient. Furthermore, the volume of each ligand was calculated using the calc_volume.py tool included in Schrödinger Suite 2021.v3.

Cavity volume analysis with mgridXf. To identify the cavities occupied by ligands in the structures of our HARIBOSS database and measure their volume, we followed a 3-step procedure.

1. **Cavity identification.** We started by using mkgridXf [49] to identify potential cavities in each structure. Ligands, ions, and water molecules were first removed from the corresponding mmCIF file with MDAnalysis [44]. The system was then processed by mkgridXf with inner and outer radii spheres equal to 1 Å and 8 Å, respectively. The cavities found by mkgridXf were extremely large, often extending throughout the entire RNA structure. We therefore decided to further classify them into sub-cavities.

2. **Sub-cavity classification.** We segmented each cavity found by mkgridXf using watershed, an image processing algorithm that can be used to detect contours and separate adjacent objects in a 3D volume [50]. The minimum distance between the centers of two sub-cavities was set at 6 Å based on the dimension of the smallest ligands in our database.

3. **Identification of sub-cavities occupied by ligands and volume calculation.** We used Delaunay triangulation [51] to identify all the sub-cavities occupied by each ligand. To avoid including sub-cavities populated by a few ligand atoms, we considered a sub-cavity to be occupied only when at least 20% of the ligand atoms were found inside its volume. Finally, for each ligand, we merged all the occupied sub-cavities and calculated the total volume. The

use of different cutoffs for the sub-cavity occupation in the range 10%-30% did not significantly alter the final volume distribution (Fig. 2.5).

Pocket analysis with Schrodinger suite. We analyzed each pocket in the HARIBOSS database using the tools available in Schrödinger Maestro Suite v. 2021-3. This analysis is articulated in 4 steps:

1. We extracted the region within 12 Å of the ligand in order to optimize the computational cost of the analysis, as our database contained large systems up to 10^4 kDa;

2. We used Maestro [52] to prepare each substructure by adding missing hydrogens, optimizing their assignment at pH of 7.4 with PROPKA [53], and minimizing the resulting model using steepest descent in combination with OPLS3 [54];

3. We used SiteMap [55, 56] to first define pockets using the coordinates of the bound ligand and a sitebox equal to 6 Å. In cases in which multiple pockets were identified for the same ligand, we used Delaunay triangulation to select the pockets occupied by at least 20% of the ligand atoms (Table 2.2).

4. For the pockets occupied by ligands, we analyzed: volume, number of site points, hydrophobicity, hydrophilicity, enclosure (or buriedness), donor/acceptor character, SiteScore, and Dscore. Except for the volume, these are unitless quantities whose ranges were calibrated on a benchmark set of protein-ligand complexes. In particular, SiteScore and Dscore quantify the propensity of a pocket to bind ligands and drug-like molecules, respectively. Both scoring functions are defined in terms of pocket size (n), enclosure (e), and hydrophilicity (p) as:

$$\text{SiteScore} = a_1 * \sqrt{n} + a_2 * e + a_3 * p \tag{2.1}$$

$$\text{Dscore} = a_1 n + a_2 e + a_3 p \tag{2.2}$$

with different coefficients for SiteScore ($a_1 = 0.0733$, $a_2 = 0.6688$, $a_3 = -0.20$) and Dscore ($a_1 = 0.094$, $a_2 = 0.60$, $a_3 = 0.324$).

2.2.3 HARIBOSS website

The HARIBOSS database is available online through a web application (`https://hariboss.pasteur.cloud`). The online version of HARIBOSS is enriched with additional cross-links and properties, either fetched from RCSB PDB or computed using the RDKit library. The web application allows to query, visualize and download the data using either a compound-centric or a complex-centric perspective.

The compound-centric perspective (`https://hariboss.pasteur.cloud/compounds/`) allows to access the list of compounds identified as RNA binders in HARIBOSS. Multiple options allow to filter these compounds based on their properties (e.g. molecular weight) or on the properties of the

PDB complexes where they have been identified (e.g. experimental resolution). It is possible to visualize the compounds either as thumbnails (the default representation), as a list or as a table. The details of each of these compounds include different identifiers (SMILES, IUPAC, InChi, InChiKey), as well as links to external databases, compliance with some of the drug-likeness criteria, and the list of RNA-SM complexes in which each compound has been identified. Multiple elements of this section of the web application were heavily inspired by the iPPI-DB database [57]. The graphical representation of the compounds uses the SmilesDrawer component [58].

The complex-centric perspective (`https://hariboss.pasteur.cloud/complexes/`) provides an access to the list of RNA-SM complexes. It also includes filtering options based on the properties of the complexes, as well as of the associated ligands. For each complex, a detail page provides a graphical representation of the complex, using the NGL library [59], and its main properties. This page also lists the identified pockets, color-coded according to their SiteScore ligandability score. Selecting the pockets allows to highlight them in the graphical representation of the complex. Both the compound- and complex-centric perspectives provide similar download features to retrieve either the entire query represented on the screen or a set of items that can be selected using the corresponding checkboxes.

2.3 Results

HARIBOSS is constructed in 3 steps (Fig. 2.1A), Materials and Methods): initial fetching of RNA structures from the PDB, filtering of the database to identify systems with at least one small molecule bound to RNA (redundant HARIBOSS), and clustering based on RNA sequence identity and pocket structural similarity (non-redundant HARIBOSS). As of May 2022, the redundant version of HARIBOSS contains 716 PDB structures of RNA-SM complexes, for a total of 1226 pockets occupied by 267 unique ligands. After clustering, the non-redundant version of HARIBOSS contains 484 PDB structures, with a total of 676 pockets occupied by 267 unique ligands. In the following paragraphs, we present an overview of the general properties of the structures included in the HARIBOSS database, the physico-chemical properties of the ligands bound to RNA, and of the pockets and cavities.

2.3.1 General properties of the HARIBOSS structures

The molecular composition of the structures changed at different stages of the construction of the database. Initially, HARIBOSS contains systems composed of RNA/protein complexes (59.3%), RNA molecules only (23.1%), RNA/protein/DNA (13.4%), and RNA/DNA complexes (4.2%) (Fig. 2.1A, cyan). All the RNA/protein/DNA complexes along with several RNA/protein complexes were filtered out because molecules other than RNA were involved in ligand binding and therefore were beyond the scope of the present study (Fig. 2.1A, red). In the non-redundant version of HARIBOSS, the majority of structures (62.0%) contained only RNA molecules, while the remaining part consisted of either RNA/protein (37.8%) or RNA/DNA (0.2%) complexes. (Fig. 2.1A, yellow). The systems initially included in our database significantly vary in size, with a total mass ranging from 0.5 to 104 kDa regardless of the presence of a small molecule (Fig. 2.1B, cyan). Interestingly, all

the structures above 102 kDa are ribosomal RNA/protein complexes. Furthermore, the filtering of the initial database reduced the number of structures with mass below 5 kDa from 23% to 9% (Fig. 2.1B, red). This significant reduction supports the idea that RNA must have sufficient structural complexity to bind small molecules [14, 18, 40].

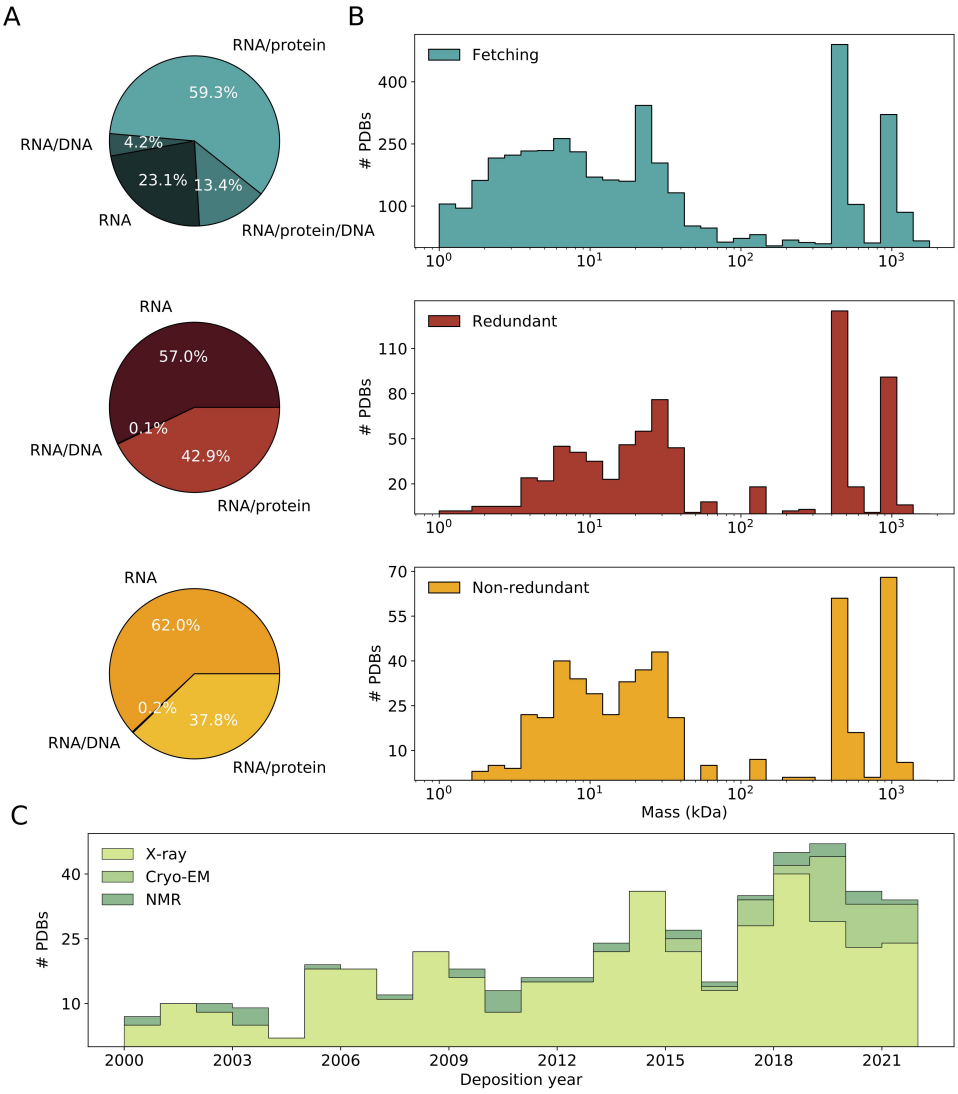


Figure 2.1: General properties of the RNA-SM structures included in the HARIBOSS database. A) Composition of HARIBOSS in terms of RNA, RNA-protein, RNA-DNA and RNA-DNA-protein structures at the different steps of the database construction: initial fetching from the PDB database (cyan), filtering based on ligand composition and nature of the interacting partners (redundant HARIBOSS, red), clustering based on sequence and pocket structure similarity (non-redundant HARIBOSS, yellow). B Distribution of the total mass of the system at the three different steps of HARIBOSS construction. **C)** Number of PDBs in the non-redundant HARIBOSS resolved by X-ray crystallography (X-ray, light green), cryo-electron microscopy (cryo-EM, green) and Nuclear Magnetic Resonance spectroscopy (NMR, dark green) as a function of the deposition year.

The majority of structures in the non-redundant version of HARIBOSS ($\sim$ 83%) were determined by X-ray crystallography with a typical resolution of 2.8 Å. However, over the past 5 years the number of cryo-electron microscopy (cryo-EM) structures, in particular of large ribosomal RNA/protein complexes, has steadily been increasing (Fig. 2.1C) and currently represents 11% of our database

with a typical resolution of 3.0 Å. Finally, structures determined by Nuclear Magnetic Resonance (NMR) spectroscopy constitute the remaining 6% of the database. Cryo-EM and NMR structures are particularly interesting from the point of view of rational drug design as they provide dynamic information about the RNA molecules that can be exploited in virtual screenings [60]. Overall, the number of RNA-SM structures deposited in the PDB every year is constantly increasing as a reflection of the growing interest of the community in studying RNA-SM interactions (Fig. 2.1C).

2.3.2 Properties of the HARIBOSS ligands

To explore the chemical space of RNA binders, we first created a dataset of unique ligands, based on their PDB identifier. As of May 2022, the total number of unique ligands in the HARIBOSS database is 267. Among these ligands, there are 11 that appear in more than 10 structures of the non-redundant HARIBOSS (Table 2.3). To provide an overview of the chemical matter present in our dataset, we used a minimum spanning tree representation (Materials and Methods) (Fig. 2.2A). Diverse ligand scaffolds are present, with some of them belonging to known classes of therapeutic agents. There is a significant number of structures of known antibiotics, like Linezolid and other oxazolidinones, Tiamulin, Eravacycline, and other tetracyclines (Fig. 2.6). It is not surprising that GTP and nucleoside analogs as well as long polar molecules, like PEG and spermidine-derived polyamines, also appeared as common RNA binders. We have also identified compounds that are not likely to be specific ligands, for example buffer components and crystallization enhancers. We did not exclude such compounds as they are potentially useful for future fragment-based ligand design efforts. The high polarity of RNA binders is depicted in their TPSA distribution, with half of the molecules being beyond the Veber drug-likeness threshold (140 ^{2}) (Fig. 2.2B). This high polarity is in line with the overall good solubility of the ligands in our HARIBOSS dataset (Fig. 2.2C). The compounds in our database significantly vary in size with a molecular mass spanning a range from 162 Da to 972 Da. The vast majority of these compounds (82%) have a mass lower than 600 Da (Fig. 2.2D). Approximately 45% of the RNA binders had a number of donors plus acceptors above the drug-likeness threshold (Fig. 2.2F and Fig. 2.7). This is again consistent with the high polar nature of these small molecules as well as the RNA targets. Moreover, RNA structurally forms "warm-like" long cavities and therefore long and flexible small molecules are expected to be among its binders. This is confirmed by the distribution of the total number of rotatable bonds of the HARIBOSS small molecules, which in 12.4% of the cases exceed the Lipinski's threshold of 10 rotatable bonds (Fig. 2.2E).

All the above-mentioned properties are reported also on the HARIBOSS website (`https://hariboss.pasteur.cloud`). To complement the analysis of the small-molecule chemical space, QikProp (Schrödinger Suite 2021.v3) was used to calculate the following properties: solvent-accessible surface area (SASA) and its hydrophobic and hydrophilic content (FOSA and FISA respectively), predicted IC50 value for blockage of hERG K^+ channels (QPlogHERG), Caco-2 cell permeability (QPPCaco) and brain/blood partition coefficient (QPlogBB). Based on this analysis, SASA was found, as expected, to be proportional to the MW with the exception of very large molecules ($MW > 600\ Da$) in which the ligand conformation may significantly alter SASA (Fig. 2.8). On the contrary, weaker correlations were found between MW and both FOSA (Fig. 2.9) and FISA (Fig. 2.10), and between FISA and FOSA (Fig. 2.11). Interestingly, 25.2% of our ligand database is above the FISA drug-like

threshold, while only 3% shows high FOSA, in agreement with TPSA (Fig. 2.2B). Among the set of unique RNA binders 55% do not show potential hERG liabilities (Fig. 2.12). Approximately half of the ligands (52%) are predicted to have poor Caco-2 cell permeability (Fig. 2.13) while more 70% of the molecules have a good predicted brain/blood partition coefficient (Fig. 2.14).

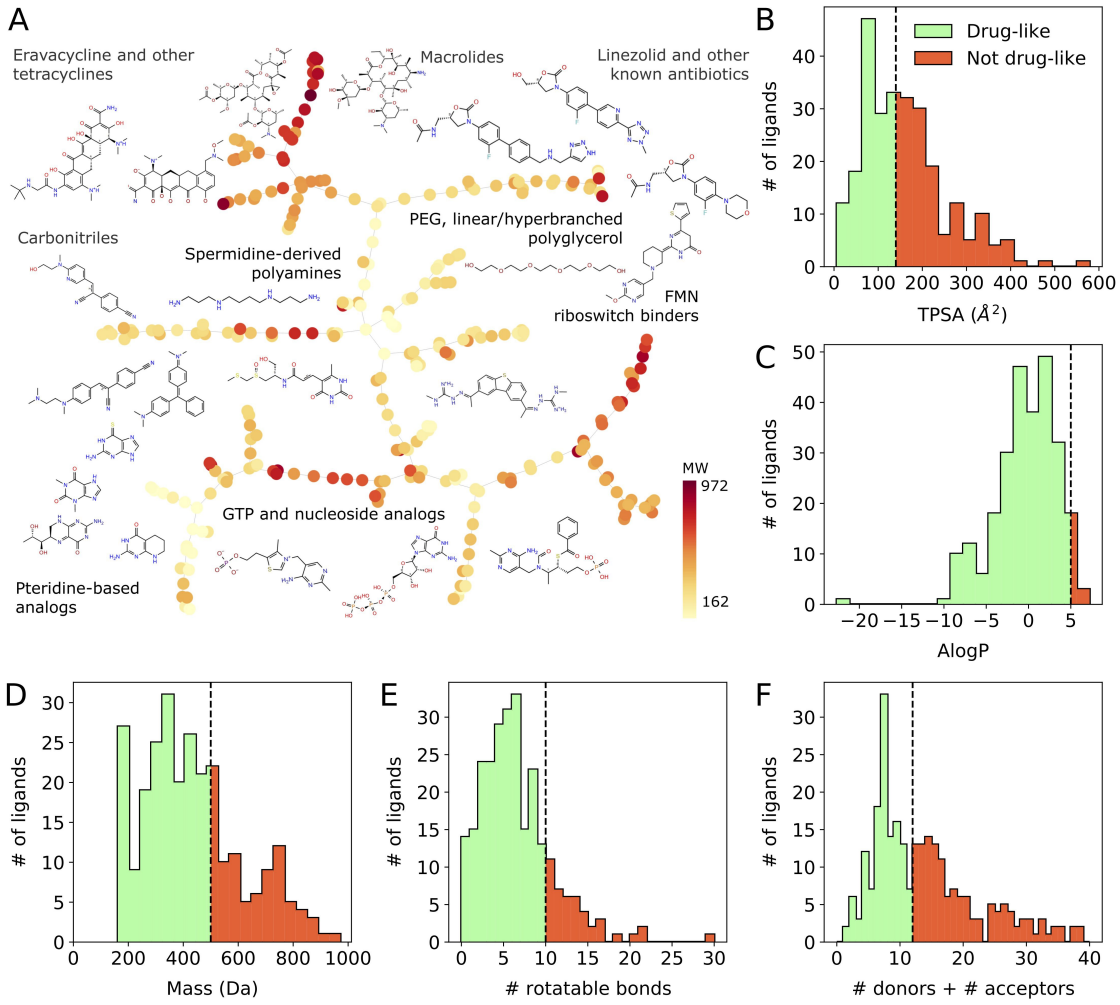


Figure 2.2: Pharmaco-chemical properties of the ligands in the HARIBOSS entries. A) 2-D representation of the chemical space of the 267 unique RNA binders obtained by an undirected weighted c-approximate k-nearest neighbor graph (Materials and Methods). Each node represents a ligand and is colored by its molecular weight (MW). **B - F)** Distributions of pharmaco-chemical properties of the ligands: the topological polar surface area (TPSA, B), the octanol-water partition coefficient (AlogP, C), the molecular mass (D), the number of rotatable bonds (E), and the sum of hydrogen bond donors and acceptors (F). For each panel, the dashed line indicates the threshold that defines drug-like compounds based on Lipinski's rule of 5 [61] or Veber rule [62]. Green/red portions of the histogram represent the regions satisfying/violating these criteria.

2.3.3 Properties of the RNA pockets and cavities

In the non-redundant version of HARIBOSS, the majority of pockets occupied by ligands and identified by SiteMap and mkgridXf (Materials and Methods) are formed by a single RNA chain (67%), while the remaining are at the interface of two (29%) or more (4%) RNA chains (Fig. 2.3A, yellow bars). Overall, the majority of pockets (67%) were considered to be potential ligand binding sites (ligandable) according to the SiteScore scoring function (SiteScore $\geq$ 0.8, Fig. 2.3B). However,

according to the Dscore score, only 35% of the pockets were classified as druggable (Dscore ≥ 0.98), 24% as difficult targets ($0.83 \leq$ Dscore < 0.98), and the remaining 41% as undruggable (Fig. 2.3C).

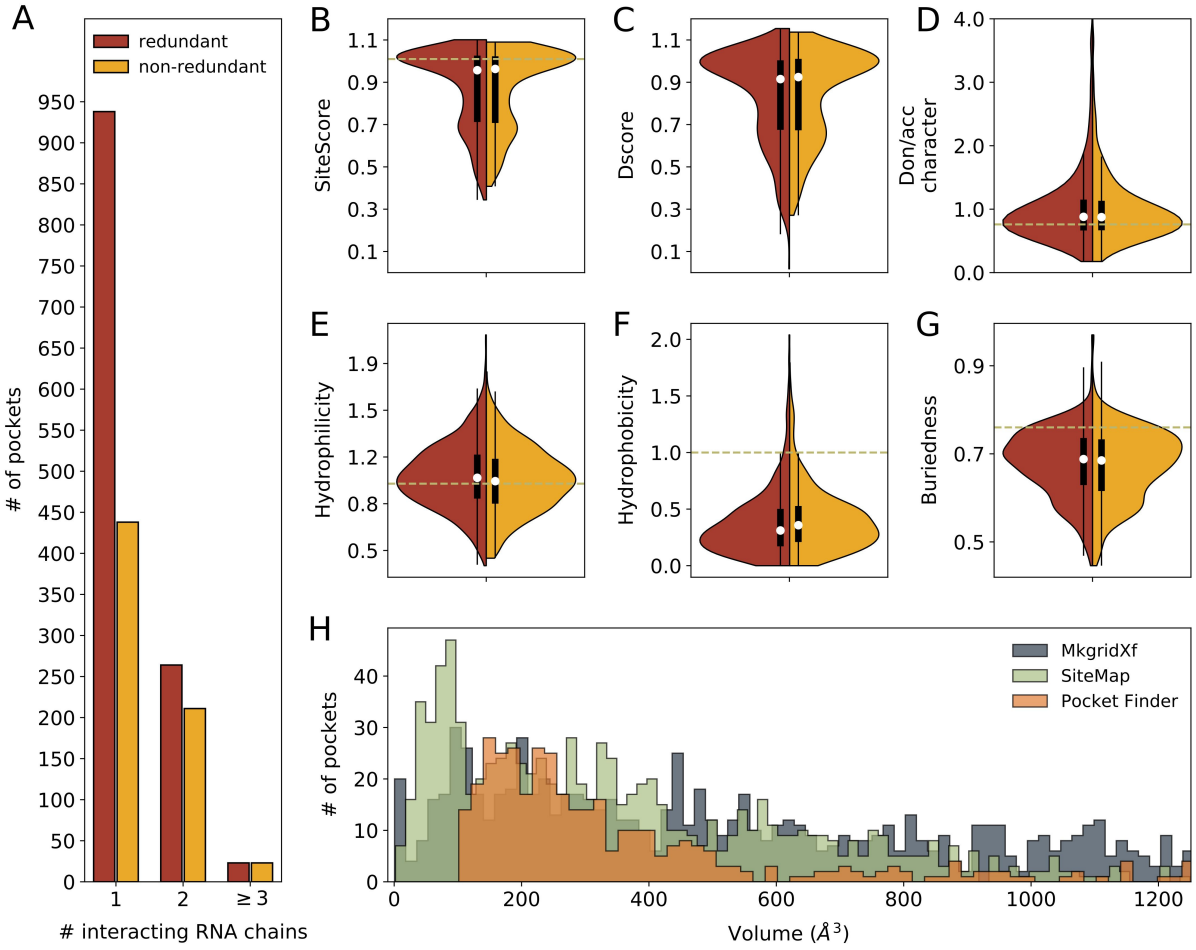


Figure 2.3: Pharmaco-chemical properties of RNA-SM pockets. A) Number of RNA-SM pockets as a function of the number of interacting RNA chains in the redundant (red) and non-redundant (yellow) HARIBOSS database. **B – G)** Violin plots representing the distributions of properties of RNA-SM pockets calculated by SiteMap: ligandability (SiteScore, B) and druggability scores (Dscore, C), hydrogen bond donor/acceptor character (D), hydrophilicity (E), hydrophobicity (F), and buriedness (G). The average values of these properties calculated on a benchmark of protein-SM complexes is represented by dashed lines [55, 56]. **H)** Volume distribution of RNA-SM pockets calculated with SiteMap (light green) and mkgridXf [49] (dark green) on the redundant HARIBOSS database, and with PocketFinder (orange) in Hewitt et al. [40]. The x-axis is limited at 1250 Å^3.

SiteScore and Dscore depend on several physico-chemical properties of the pocket and have been optimized for protein molecules. Among these properties, RNA pockets presented a typical hydrogen-bond donor/acceptor character (Fig. 2.3D) and hydrophilicity (Fig. 2.3E) similar to proteins. However, RNA pockets appeared to be less hydrophobic (Fig. 2.3F) and more exposed to solvent (Fig. 2.3G). This different character is a consequence of the more polar nature of RNA molecules. A minority of pockets, corresponding to 33% of cases, was considered non ligandable (SiteScore $<$ 0.8, Fig. 2.3B). In many cases, low ligandability and druggability can be interpreted in terms of the physico-chemical properties of the pocket. For example, hydrophilic pockets particularly exposed to solvent were generally considered not druggable (Fig. 2.15), as in the case of the structure of a

RNA primer–template bound to ligand 5GP (PDB 5dhb, Fig. 2.4A). In contrast, the hydrophobicity of RNA-SM pockets appears not to be strongly correlated with ligandability (Fig. 2.16). Highly hydrophobic pockets, like the site in which ligand EKM binds the Mango-II Fluorescent RNA Aptamer structure (PDB 6c64, Fig. 2.4B), and less hydrophobic pockets, such as the binding site of ligand UG4 in Fusibacterium ulcerans ZTP riboswitch (PDB 6wzs, Fig. 2.4C), were found to be equally ligandable. However, druggable pockets have a tendency to be more hydrophobic (Fig. 2.17).

In agreement with previous studies [40], RNA-SM pockets span a broad range of volumes, independently from the software and dataset used for pocket identification and volume calculation (Fig. 2.3H). The size of the most druggable and ligandable pocket in HARIBOSS, corresponding to the binding site of 747 to the Corn RNA aptamer (PDB 5bjp and 5bjo, Fig. 2.4D), is equal to ~ 200 Å^3. However, cavities larger than ~ 300 Å^3 were also classified as ligandable (Fig. 2.20). Given the size of the small molecules in HARIBOSS (Fig. 2.2A), these large cavities cannot be fully occupied by a ligand. Indeed, although for small pockets the ligand volume is often bigger than the pocket volume, suggesting that the ligand is exposed to water, for larger pockets the ligand volume is consistently smaller compared to the pocket volume (Fig. 2.19). Further investigations are needed to understand whether this finding is due to the fact that RNA molecules form large cavities that are only partly occupied by ligands or by artifacts of the software used for detecting cavities and calculate their volume. Finally, all cavities with volume smaller than 100 Å^3 were considered neither druggable (Fig. 2.18) nor ligandable (Fig. 2.20).

2.4 Conclusions

Here, we presented HARIBOSS, a curated database of structures of RNA-small molecule complexes, built to aid the development of computational drug design pipelines. For each HARIBOSS entry, we provided general structural information and we analyzed the physico-chemical properties of the ligands bound to RNA, and of the respective pockets. For the majority of structures in our database, the experimentally-resolved pockets were confirmed as ligandable. Only one third of all pockets was ranked as good for drug design purposes, the remainder part being a difficult target or undruggable. Our analysis indicates that low druggability is due to the fact that RNA pockets are less hydrophobic and more exposed to solvent than protein pockets. In line with these findings, RNA binders in the HARIBOSS database were mostly highly polar and water-soluble ligands. Known classes of antibiotics and endogenous polar ligands are among the most frequent RNA binders in PDB. Cell permeability is, as expected, a major issue for a significant part of these molecules. Future studies will be aimed at identifying HARIBOSS compounds that are RNA-selective and characterizing the physico-chemical interactions that determine their selectivity.

As of today, HARIBOSS contains only static RNA-SM structures. This is already an important step that will aid the discovery of RNA binders using 3D structure-based, rational drug discovery approaches. However, this is only the first step in the identification of RNA inhibitors and modulators. Demonstrating that binding translates to changes in dynamics and function is the most challenging part. The next natural step is to include in HARIBOSS structural ensembles representing the highly-dynamic nature of RNA. Characterizing the dynamic properties of RNA molecules

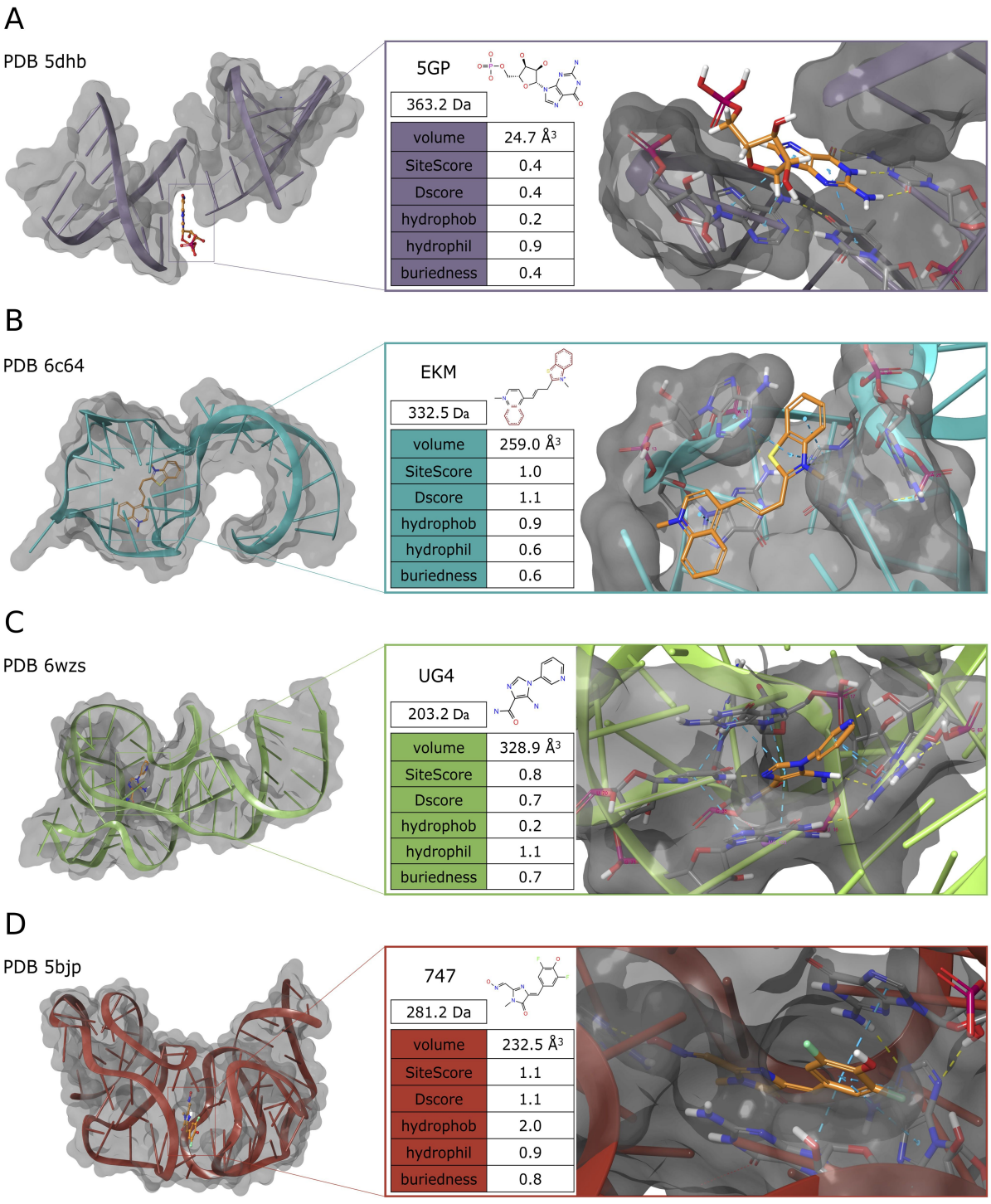


Figure 2.4: Examples of RNA-SM structures included in HARIBOSS. A) An RNA primer–template bound to ligand 5GP (PDB 5dhb). **B)** The Mango-II Fluorescent RNA aptamer bound to ligand EKM (PDB 6c64). **C)** The Fusibacterium ulcerans ZTP riboswitch bound to UG4 (PDB 6wzs). D) The Corn RNA aptamer bound to ligand 747 (PDB 5bj). Left column: PDB code and structure of the RNA-SM complex. Right column: name, molecular weight, 2D representation of the small molecule, table with ligand properties calculated by SiteMap, and close view of the pocket occupied by the small molecule. The following types of RNA-SM interactions are highlighted by dashed lines: hydrogen bonds (yellow), salt bridges (purple), and π-π stacking (cyan).

is particularly important to identify selective and specific compounds using structure-based approaches that target individual members of RNA conformational ensembles [60]. Furthermore, in the future we will identify apo RNA structures deposited in the PDB and structural ensembles from MD simulations to determine whether ligand binding to RNA molecules can be better described as an induced-fit or conformational selection process.

HARIBOSS can be accessed via a web interface available at `https://hariboss.pasteur.cloud` and explored using a compound or pocket perspective. Our database will facilitate: i) assessing the accuracy of existing protein-oriented drug design computational tools, identifying areas of improvement, and optimizing them for RNA molecules; ii) investigating the nature of RNA-SM interactions; iii) defining the chemical space of RNA binders and their potential to be used as drugs; iv) identifying new potential RNA targets based on pocket druggability or starting from a specific compound. In conclusion, our comprehensive, curated, and regularly updated database of RNA-SM structures is a stepping stone for the scientific community to develop novel in silico approaches to discover compounds for direct RNA targeting.

2.5 Supplementary Information

2.5.1 Supplementary figures

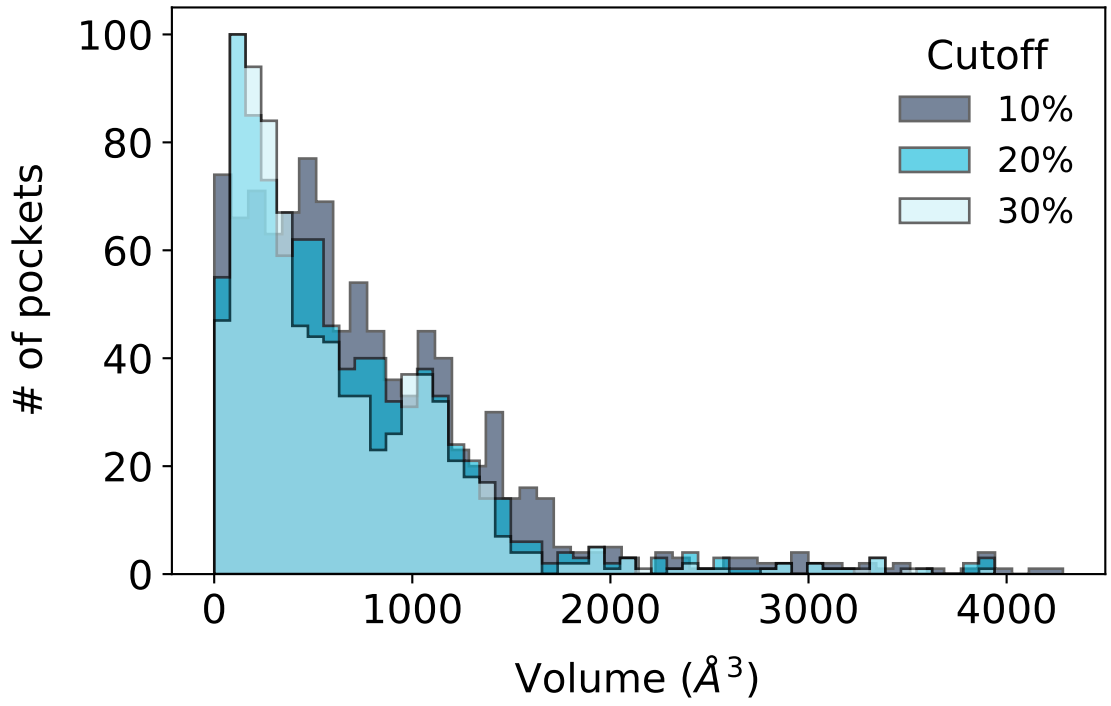


Figure 2.5: Cavity analysis with mkgridXf. Volume distributions of the cavities found by mkgridXf on the redundant HARIBOSS database using a cutoff for the sub-cavity occupation equal to 10% (dark blue), 20% (cyan), and 30% (light blue).

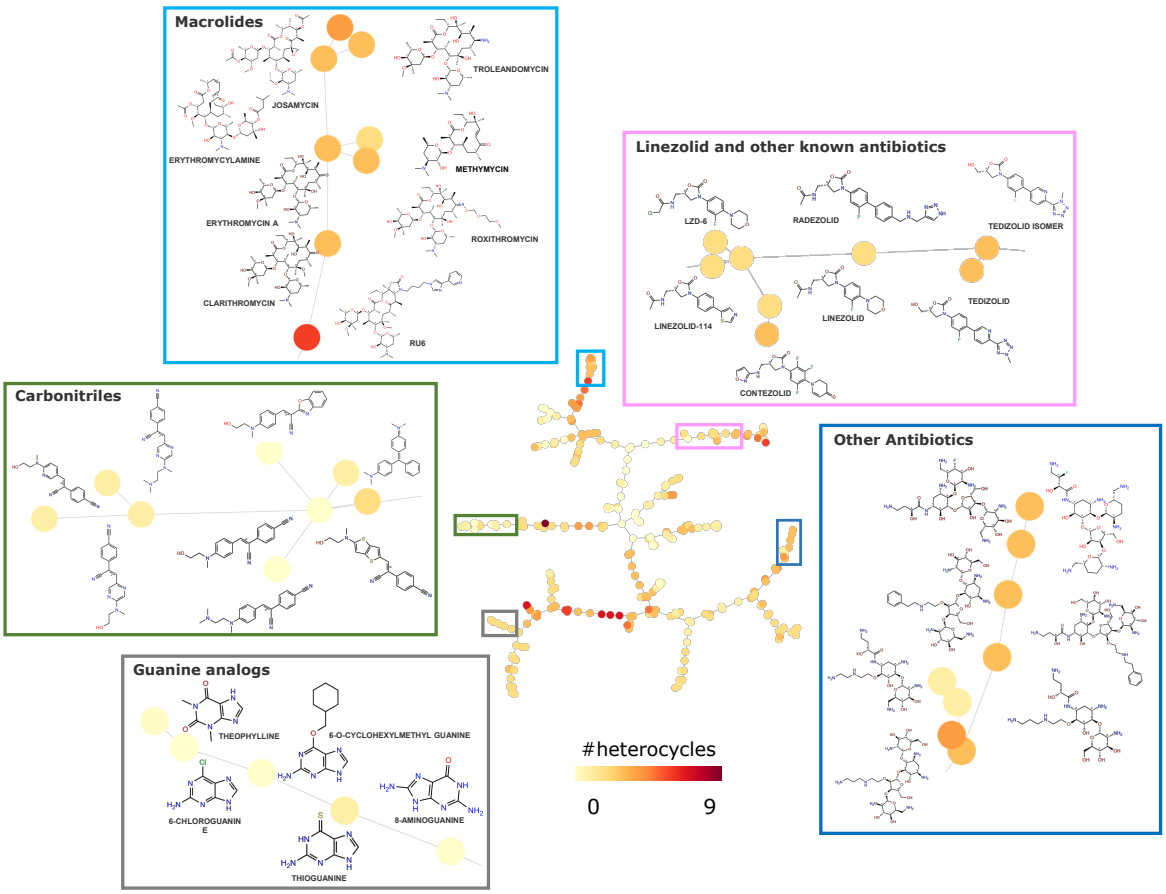


Figure 2.6: Minimum spanning tree representation of the HARIBOSS small molecule database. Five characteristic families of RNA binders belonging to the highlighted branches are shown in detail. Each node is colored according to the number of heterocyclic groups present in the specific molecule.

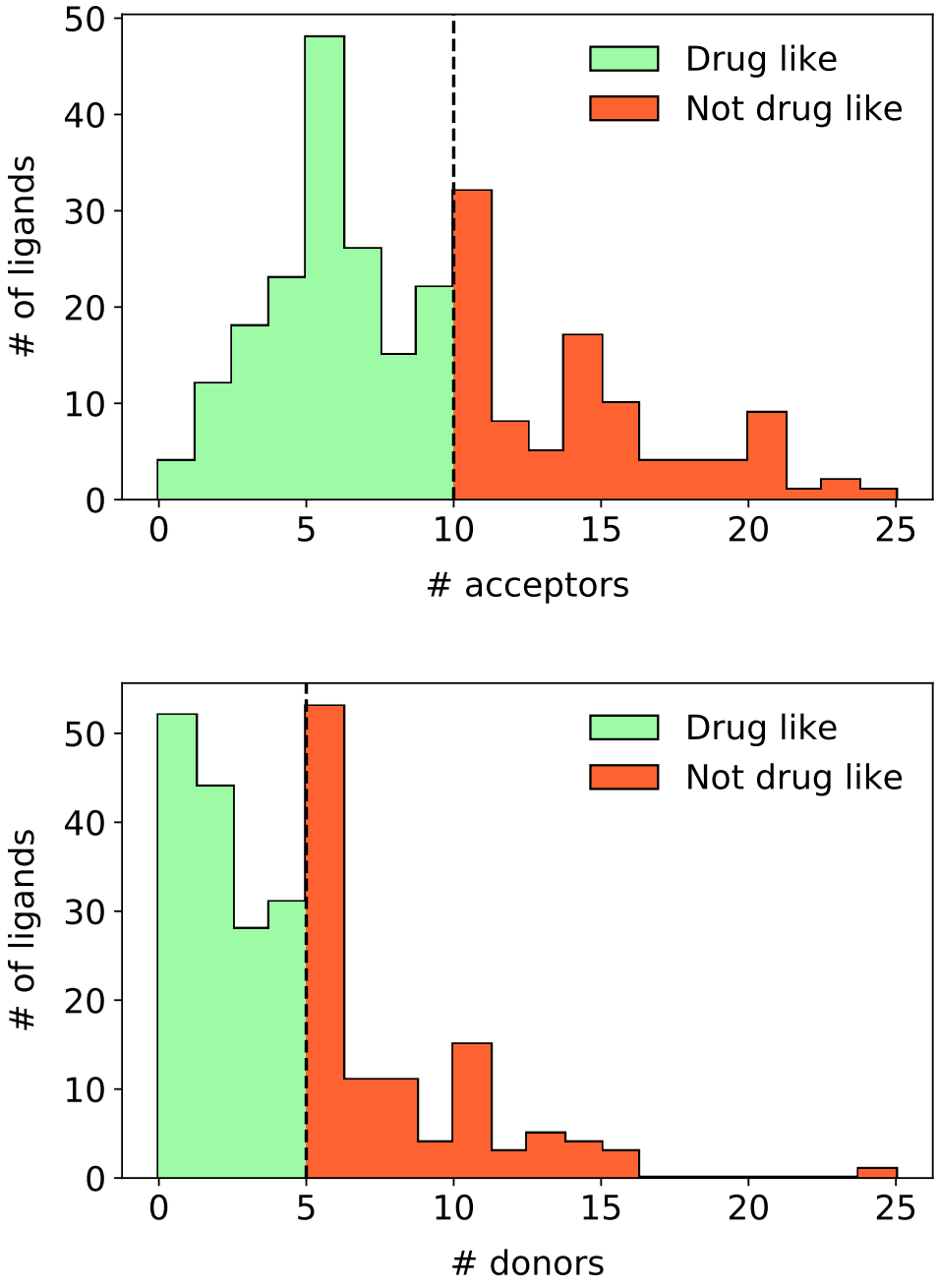


Figure 2.7: Distribution of the number of hydrogen bond donors (top) and hydrogen bond acceptors (bottom) of the HARIBOSS ligands. The dashed line indicates the threshold that defines drug-like compounds based on Lipinski's rule of 5 [61] or Veber rule [62]. Green/red portions of the histogram represent the regions satisfying/violating these criteria.

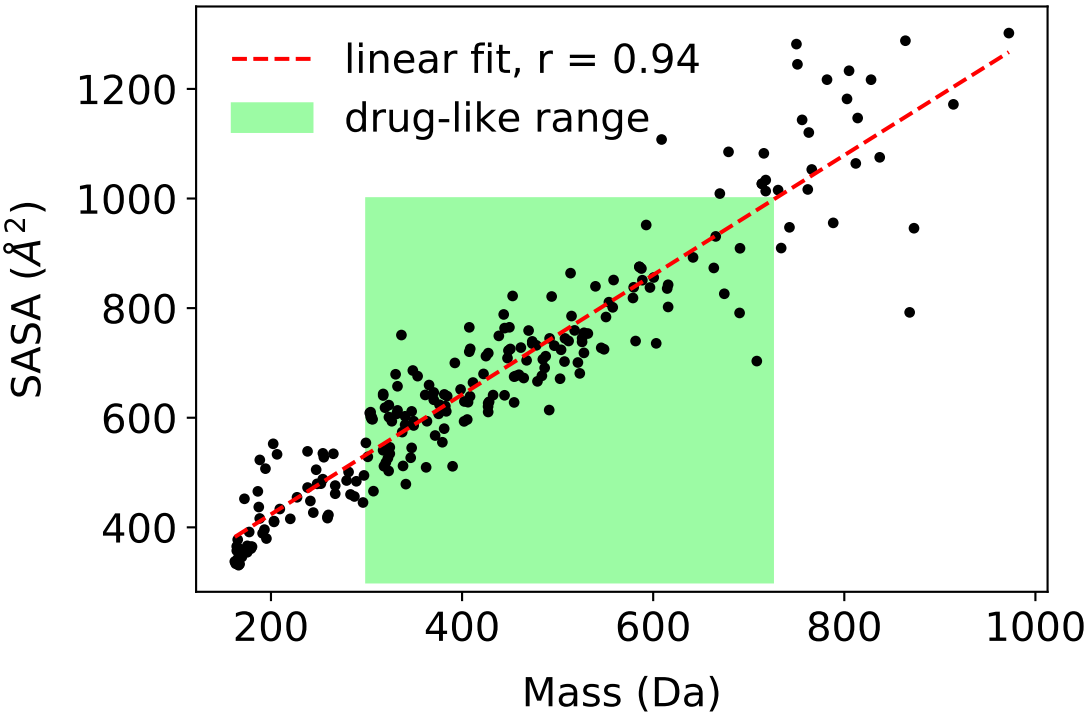


Figure 2.8: Scatter plot of ligand mass vs Solvent Accessible Surface Area (SASA). The properties were calculated on the set of unique ligands using QikProp. The green rectangle indicates the range of values corresponding to drug-like molecules as defined in [56], the red dotted line the linear fit.

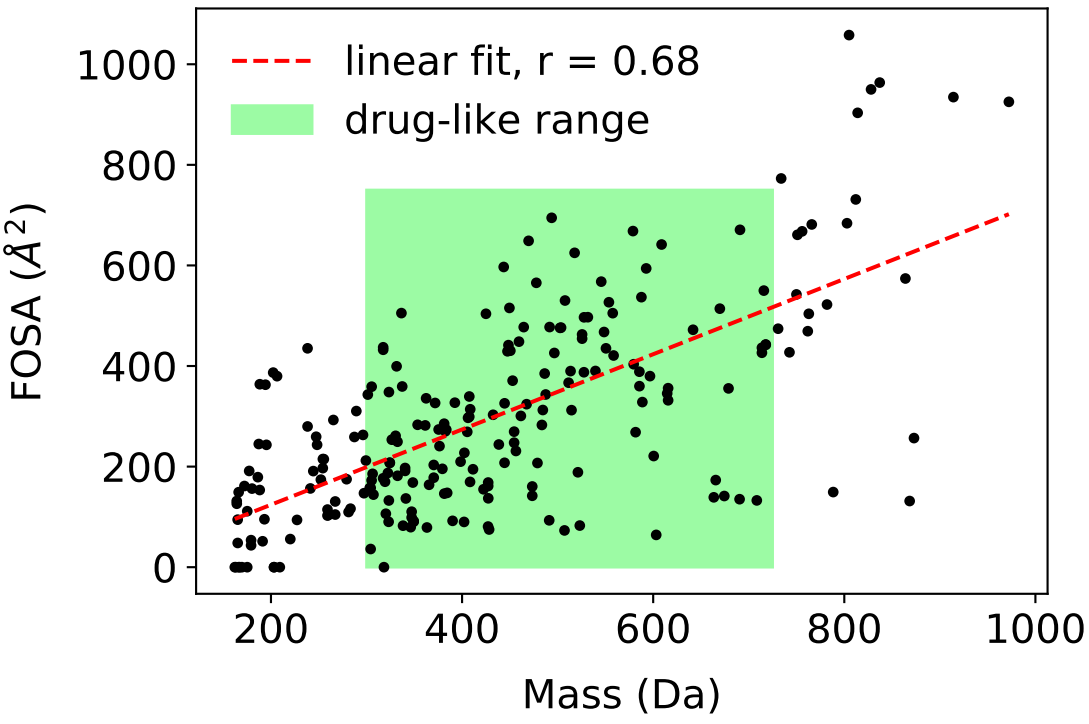


Figure 2.9: Scatter plot of ligand mass vs Hydrophobic solvent accessible surface area (FOSA). The properties were calculated on the set of unique ligands using QikProp. The green rectangle indicates the range of values corresponding to drug-like molecules as defined in [56], the red dotted line the linear fit.

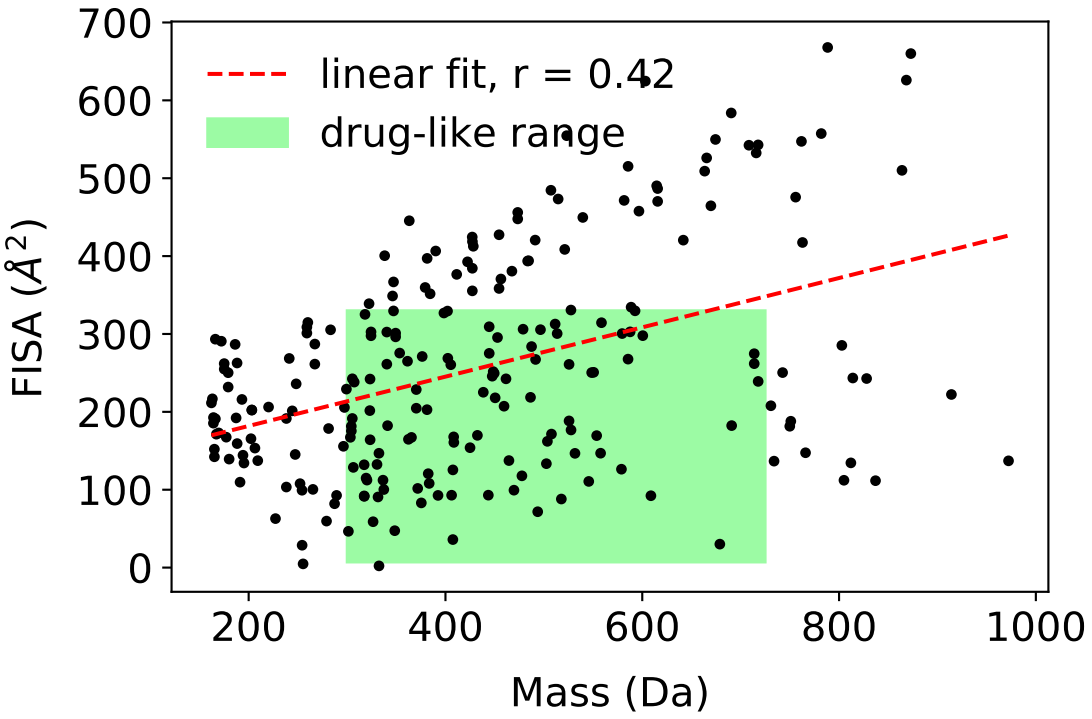


Figure 2.10: Scatter plot of ligand mass vs Hydrophilic solvent accessible surface area (FISA). The properties were calculated on the set of unique ligands using QikProp. The green rectangle indicates the range of values corresponding to drug-like molecules as defined in [56], the red dotted line the linear fit.

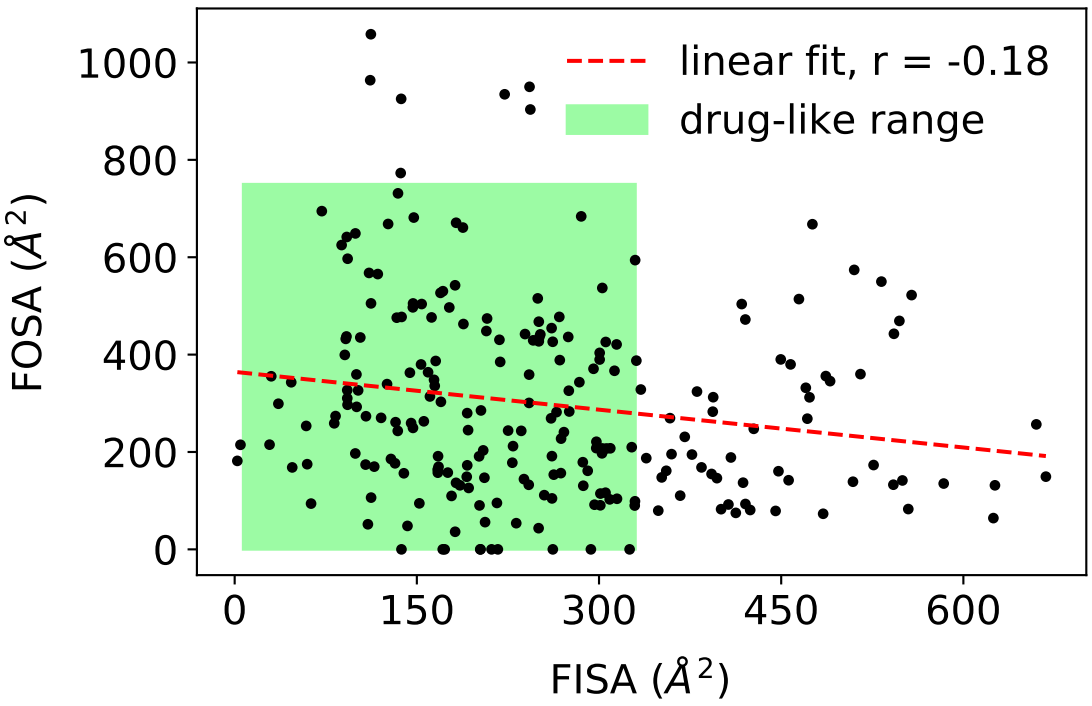


Figure 2.11: Scatter plot of ligand FISA vs FOSA. The properties were calculated on the set of unique ligands using QikProp. The green rectangle indicates the range of values corresponding to drug-like molecules as defined in [56], the red dotted line the linear fit.

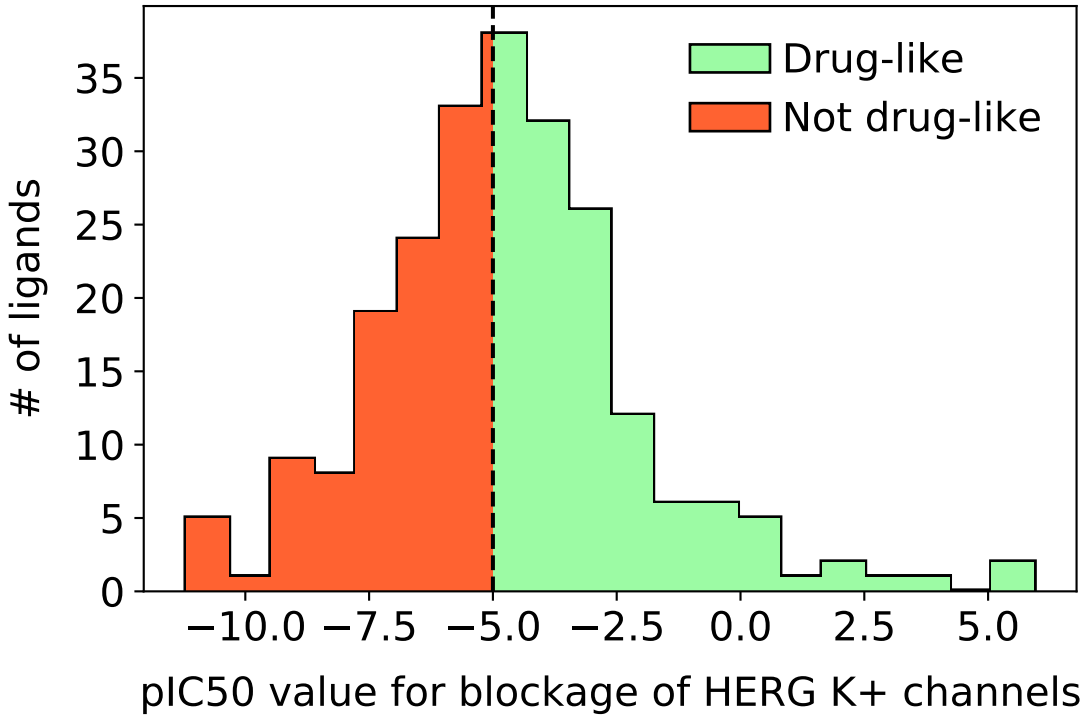


Figure 2.12: Distribution of the predicted IC50 value for blockage of HERG K+ channels (QPlogHERG) of the HARIBOSS ligands. The property was calculated on the set of unique ligands using QikProp. Green/red portions of the histogram represent the regions satisfying/violating the drug-likeness criterion for QPlogHERG as defined in [56].

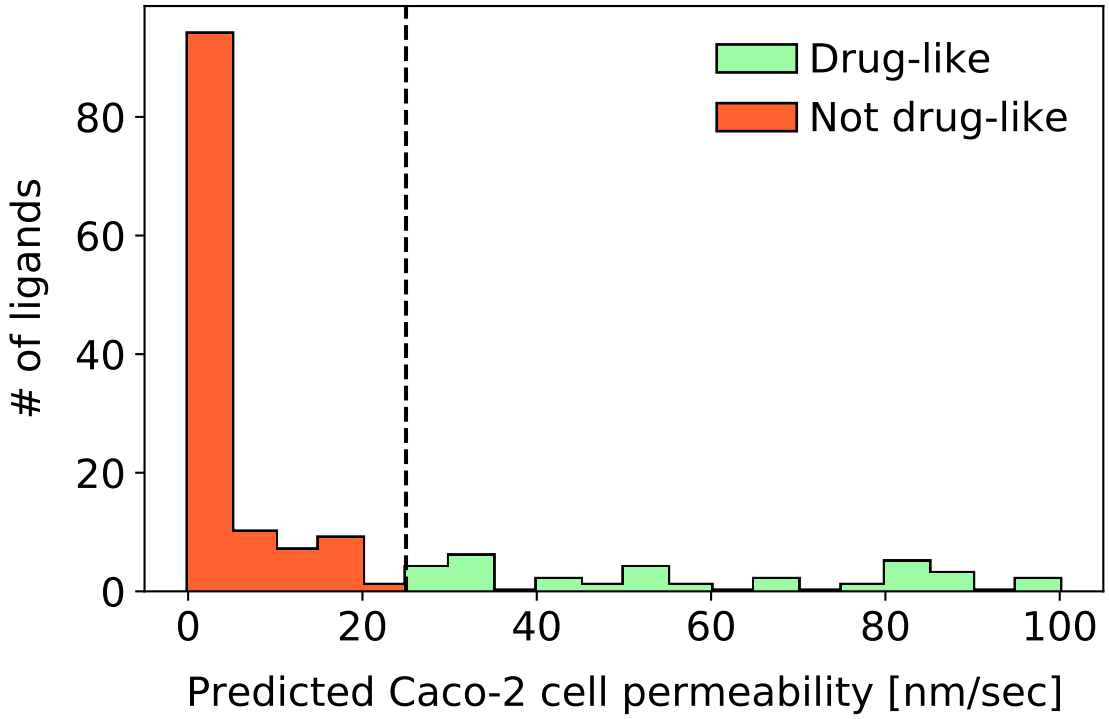


Figure 2.13: Distribution of the predicted Caco-2 cell permeability (QPPCaco) of the HARIBOSS ligands. The property was calculated on the set of unique ligands using QikProp. Green/red portions of the histogram represent the regions satisfying/violating the drug-likeness criterion for QPPCaco as defined in [56].

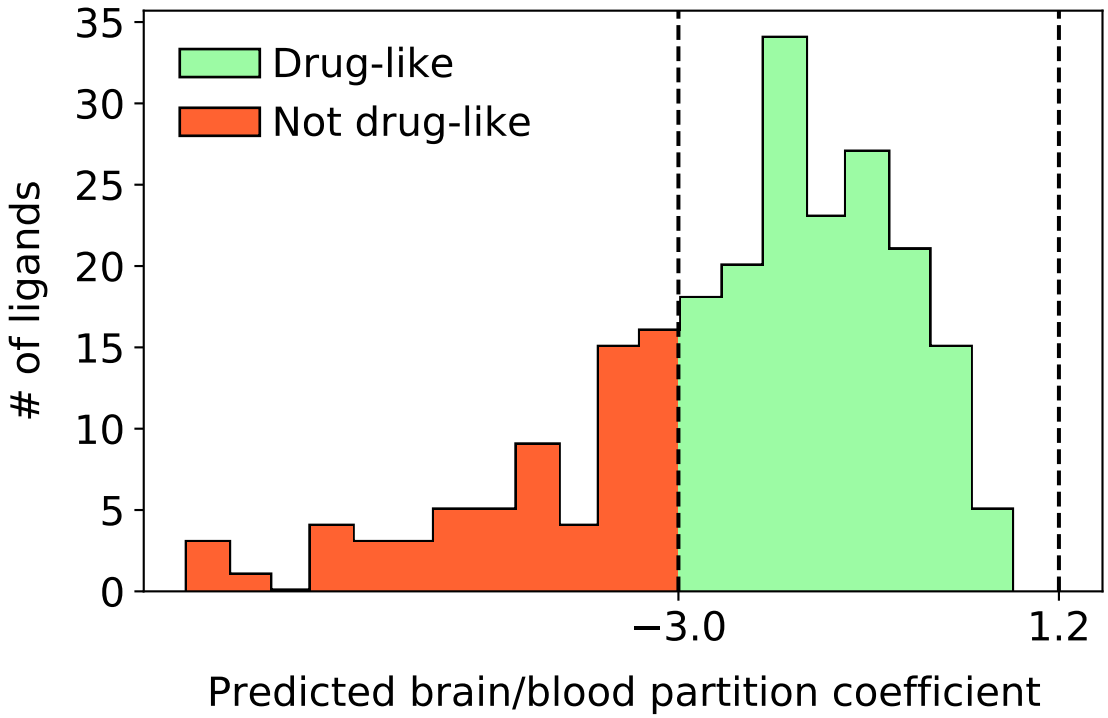


Figure 2.14: Distribution of the predicted brain/blood partition coefficient (QPlogBB) of the HARIBOSS ligands. The property was calculated on the set of unique ligands using QikProp. Green/red portions of the histogram represent the regions satisfying/violating the drug-likeness criterion for QPlogBB as defined in [56].

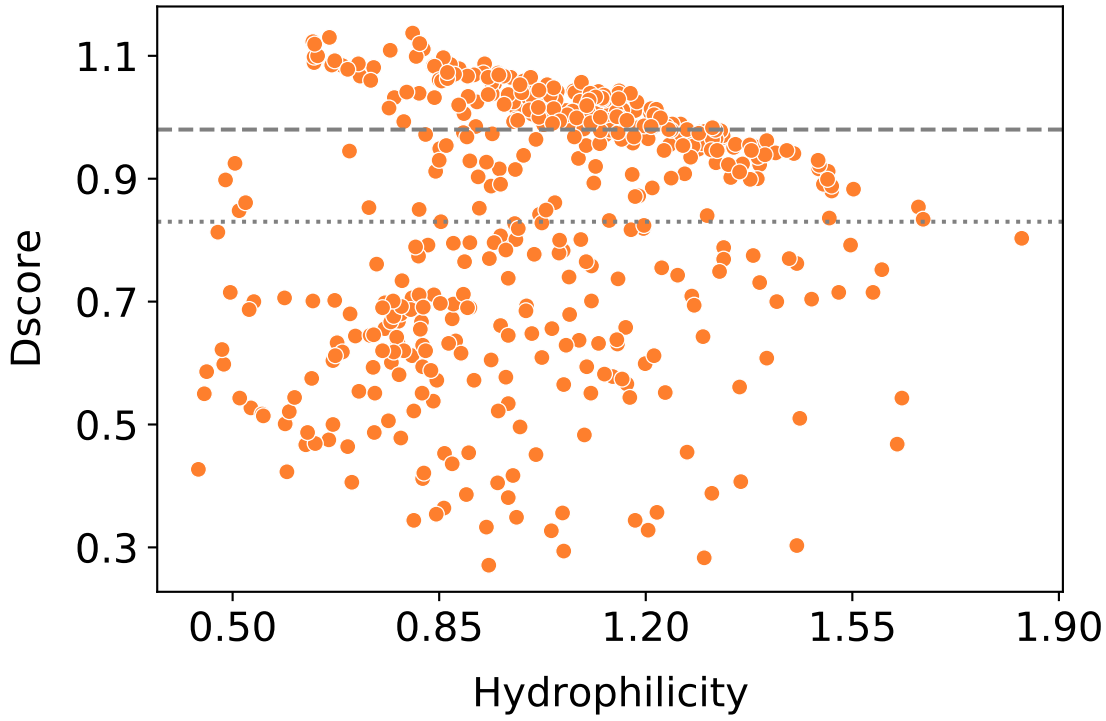


Figure 2.15: Scatter plot of pocket hydrophilicity vs druggability score (Dscore). The properties were calculated on the non-redundant HARIBOSS database using SiteMap. The dashed and dotted lines represent the thresholds for druggable and difficult-target pockets, respectively.

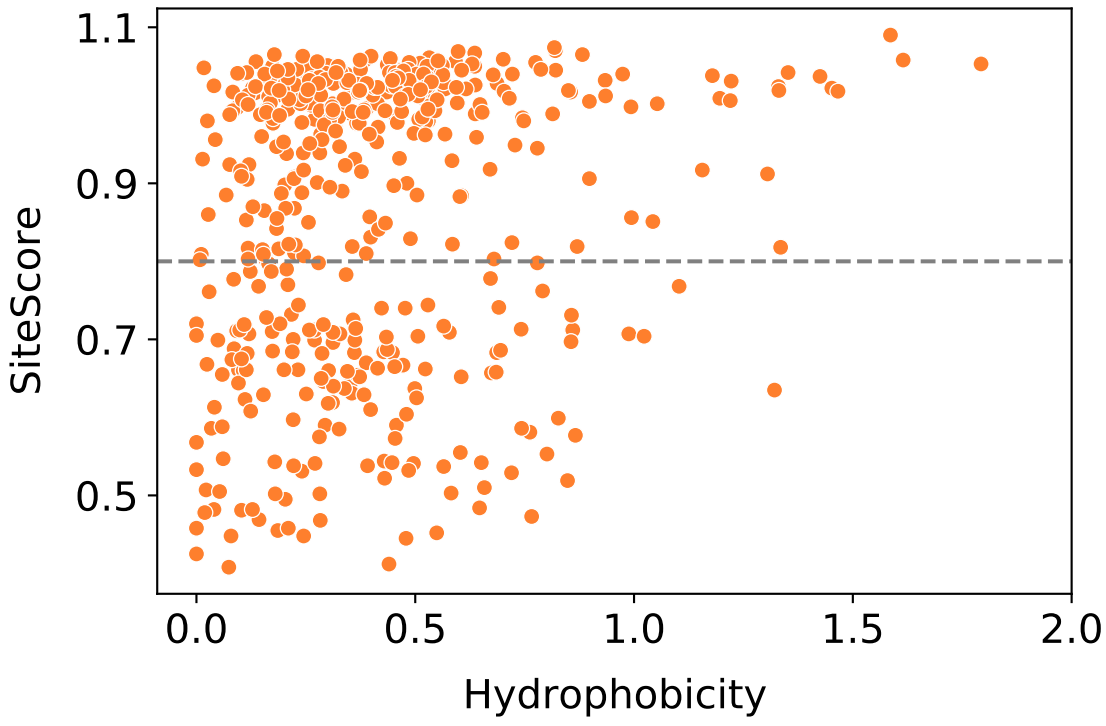


Figure 2.16: Scatter plot of pocket hydrophobicity vs ligandability score (SiteScore). The properties were calculated on the non-redundant HARIBOSS database using SiteMap. The dashed line represents the threshold for ligandable pocket.

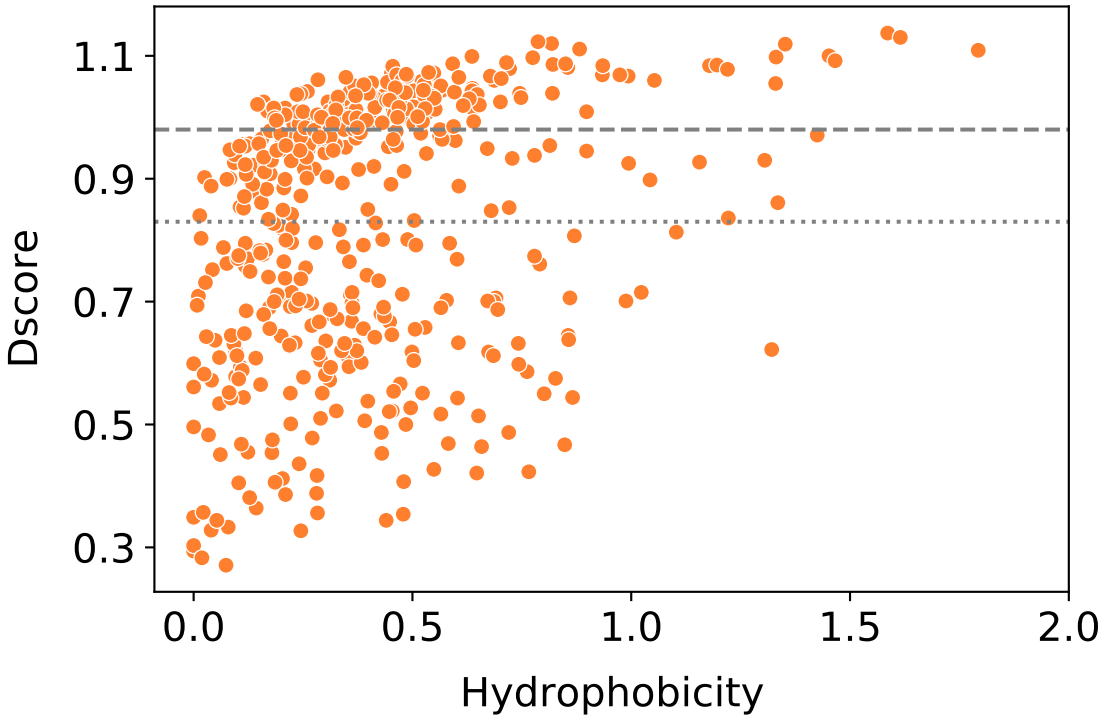


Figure 2.17: Scatter plot of pocket hydrophobicity vs druggability score (Dscore). The properties were calculated on the non-redundant HARIBOSS database using SiteMap. The dashed and dotted lines represent the thresholds for druggable and difficult-target pockets, respectively.

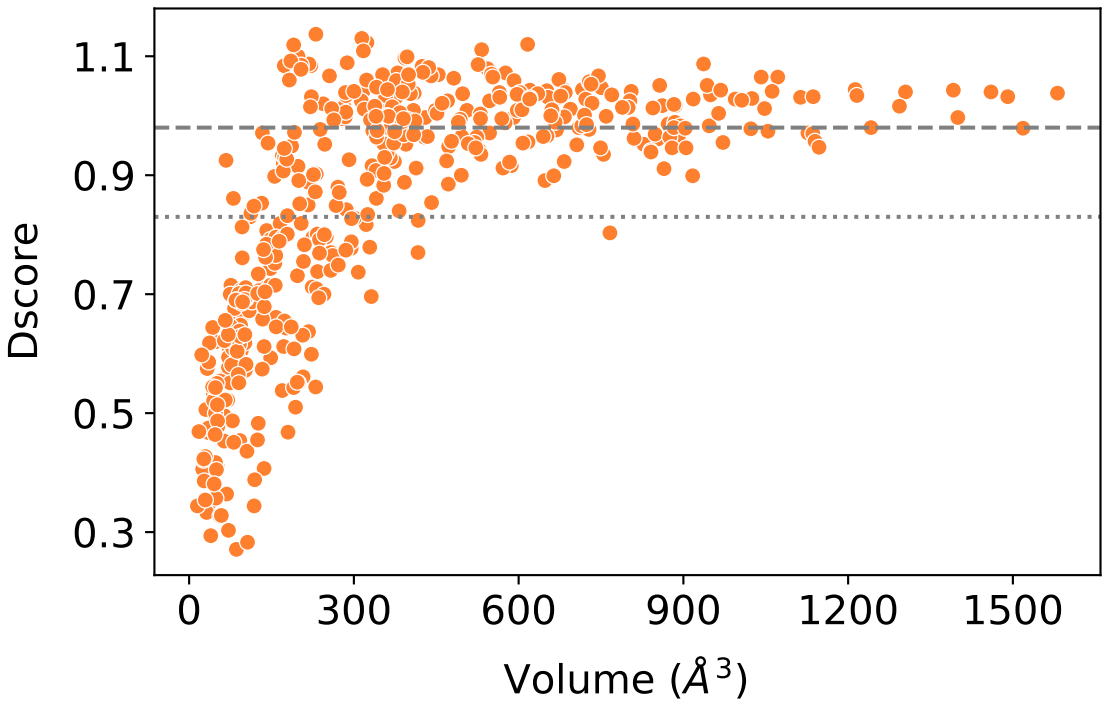


Figure 2.18: Scatter plot of pocket volume vs druggability score (Dscore). The properties were calculated on the non-redundant HARIBOSS database using SiteMap. The dashed and dotted lines represent the thresholds for druggable and difficult-target pockets, respectively.

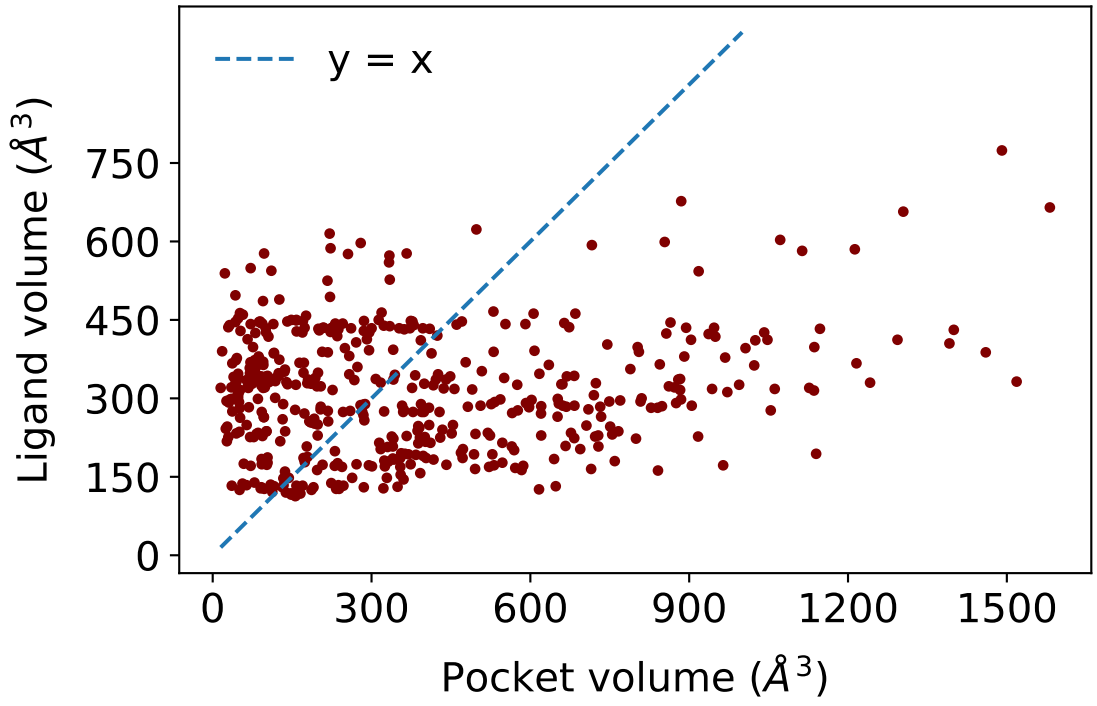


Figure 2.19: Scatter plot of pocket volume vs ligand volume. The properties were calculated on the non-redundant HARIBOSS database using SiteMap and the volume calculation script from Schrodinger Suite.

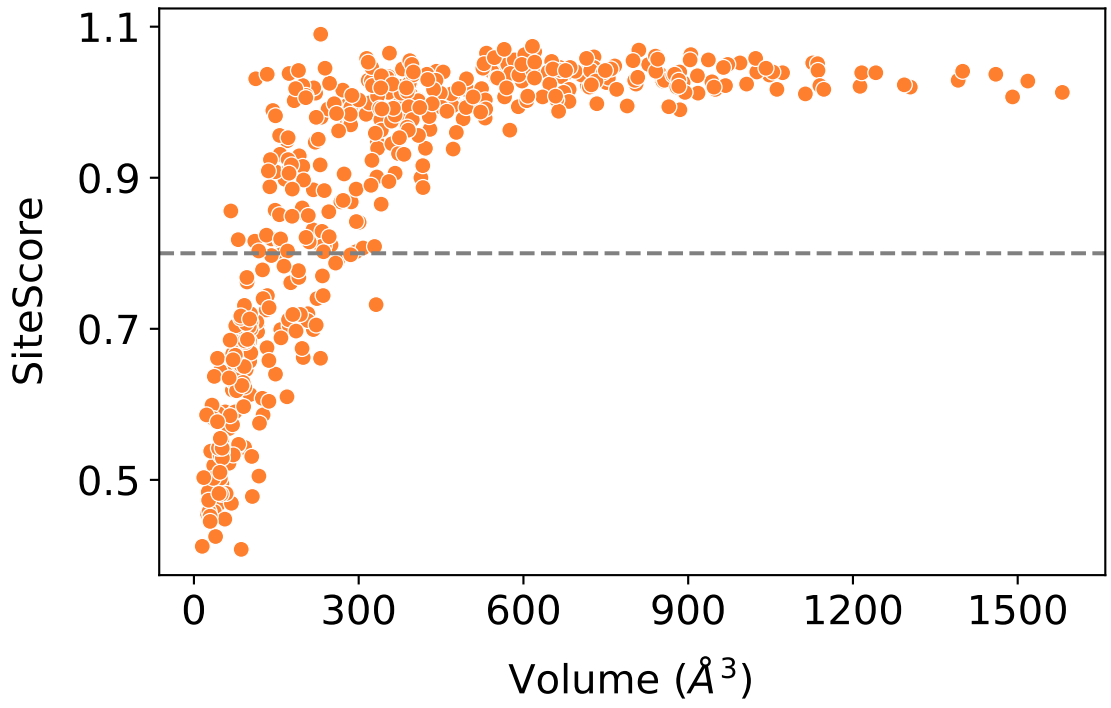


Figure 2.20: Scatter plot of pocket volume vs ligandability score (SiteScore). The properties were calculated on the non-redundant HARIBOSS database using SiteMap. The dashed line represents the threshold for ligandable pocket.

2.5.2 Supplementary tables

Cutoff	Redundant HARIBOSS # of interacting RNA chains					Non-redundant HARIBOSS # of interacting RNA chains				
	Total	**1**	**2**	**3**	**4**	**Total**	**1**	**2**	**3**	**4**
0	1158	794	323	38	3	610	380	211	17	2
5	1158	863	258	34	3	579	393	170	14	2
10	1145	873	248	23	1	573	405	159	9	0
20	1151	915	235	1	0	564	420	144	0	0

Table 2.1: Number of pockets in the redundant and non-redundant HARIBOSS databases, as a function of the number of RNA interacting chains and the minimum number of atoms (cutoff) for a chain to be considered as interacting. This analysis was performed on the HARIBOSS database updated in February 2022.

Pocket analysis stage	**# cases**
Input	1226
Preparation output	1180
Evaluation	1017

Pocket composition and occupancy	**# cases**
1 pocket	809
2 subpockets, with only 1 populated	116
2 subpockets, none populated	21
2 subpockets, both populated	7
More than 2 subpockets	12
No pockets	52

Table 2.2: Statistics of the pocket analysis by SiteMap.

Non-redundant HARIBOSS		
Ligand PDB ID	**Occurrence**	**Name**
PAR	50	Paromomycin
SPM	30	Spermine
NMY	24	Neomycin
LLL	20	Gentamicin C1A
GP3	19	Diguanosine-5'-Triphosphate
SAM	17	S-Adenosylmethionine (SAMe)
8UZ	16	TC007
GET	16	Geneticin (G418)
GTP	15	Guanosine-5'-triphosphate
AM2	13	Apramycin
NEG	10	Negamycin

Table 2.3: Occurrence of the 15 most frequent ligands in non-redudant HARIBOSS.

Bibliography

[1] Jennifer Cable et al. “Noncoding RNAs: biology and applications—a Keystone Symposia report”. In: *Annals of the New York Academy of Sciences* 1506.1 (Dec. 2021), pp. 118–141.

[2] Thomas R. Cech and Joan A. Steitz. “The noncoding RNA revolution - Trashing old rules to forge new ones”. In: *Cell* 157.1 (2014), pp. 77–94.

[3] Run-Wen Yao, Yang Wang, and Ling-Ling Chen. “Cellular functions of long noncoding RNAs”. In: *Nature Cell Biology* 21.5 (May 2019), pp. 542–551.

[4] Stefanie A. Mortimer, Mary Anne Kidwell, and Jennifer A. Doudna. “Insights into RNA structure and function from genome-wide studies”. In: *Nature Reviews Genetics* 15.7 (2014), pp. 469–479.

[5] Melanie Winkle et al. “Noncoding RNA therapeutics — challenges and potential solutions”. In: *Nature Reviews Drug Discovery* 20.8 (Aug. 2021), pp. 629–651.

[6] Ai-Ming Yu, Young Hee Choi, and Mei-Juan Tu. “RNA Drugs and RNA Targets for Small Molecules: Principles, Progress, and Challenges”. In: *Pharmacological Reviews* 72.4 (Oct. 2020), pp. 862–898.

[7] Feng Wang, Travis Zuroske, and Jonathan K. Watts. “RNA therapeutics on the rise”. In: *Nature Reviews Drug Discovery* 19.7 (July 2020), pp. 441–442.

[8] James C. Kaczmarek, Piotr S. Kowalski, and Daniel G. Anderson. “Advances in the delivery of RNA therapeutics: from concept to clinical reality”. In: *Genome Medicine* 9.1 (Dec. 2017), p. 60.

[9] Ryszard Kole, Adrian R. Krainer, and Sidney Altman. “RNA therapeutics: beyond RNA interference and antisense oligonucleotides”. In: *Nature Reviews Drug Discovery* 11.2 (Feb. 2012), pp. 125–140.

[10] Niels Dammes and Dan Peer. “Paving the Road for RNA Therapeutics”. In: *Trends in Pharmacological Sciences* 41.10 (Oct. 2020), pp. 755–775.

[11] Thomas C. Roberts, Robert Langer, and Matthew J. A. Wood. “Advances in oligonucleotide drug delivery”. In: *Nature Reviews Drug Discovery* 19.10 (Oct. 2020), pp. 673–694.

[12] Tulsi Ram Damase et al. “The Limitless Future of RNA Therapeutics”. In: *Frontiers in Bioengineering and Biotechnology* 9 (Mar. 2021).

[13] Noreen F. Rizvi and Graham F. Smith. “RNA as a small molecule druggable target”. In: *Bioorganic & Medicinal Chemistry Letters* 27.23 (Dec. 2017), pp. 5083–5088.

[14] James P. Falese, Anita Donlic, and Amanda E. Hargrove. “Targeting RNA with small molecules: from fundamental principles towards the clinic”. In: *Chemical Society Reviews* 50.4 (2021), pp. 2224–2243.

[15] Hafeez S. Haniff et al. “Target-Directed Approaches for Screening Small Molecules against RNA Targets”. In: *SLAS Discovery* 25.8 (Sept. 2020), pp. 869–894.

[16] Amanda E. Hargrove. “Small molecule-RNA targeting: Starting with the fundamentals”. In: *Chemical Communications* 56.94 (2020), pp. 14744–14756.

[17] Matthew D. Disney. “Targeting RNA with Small Molecules To Capture Opportunities at the Intersection of Chemistry, Biology, and Medicine”. In: *Journal of the American Chemical Society* 141.17 (May 2019), pp. 6776–6790.

[18] Katherine Deigan Warner, Christine E. Hajdin, and Kevin M. Weeks. “Principles for targeting RNA with drug-like small molecules”. In: *Nature Reviews Drug Discovery* 17.8 (2018), pp. 547–558.

[19] Anita Donlic et al. “Regulation of MALAT1 triple helix stability and in vitro degradation by diphenylfurans”. In: *Nucleic Acids Research* 48.14 (Aug. 2020), pp. 7653–7664.

[20] Suzanne G Rzuczek et al. “Precise small-molecule recognition of a toxic CUG RNA repeat expansion”. In: *Nature Chemical Biology* 13.2 (Feb. 2017), pp. 188–193.

[21] Anthony Bugaut et al. “Small molecule-mediated inhibition of translation by targeting a native RNA G-quadruplex”. In: *Organic & Biomolecular Chemistry* 8.12 (2010), p. 2771.

[22] Andrew C. Stelzer et al. “Discovery of selective bioactive small molecules by targeting an RNA dynamic ensemble”. In: *Nature Chemical Biology* 7.8 (2011), pp. 553–559.

[23] Hafeez S. Haniff et al. “Target-Directed Approaches for Screening Small Molecules against RNA Targets”. In: *SLAS Discovery* 25.8 (2020), pp. 869–894.

[24] Seyed MohammadReza Hashemian, Tayebeh Farhadi, and Mojdeh Ganjparvar. “Linezolid: a review of its properties, function, and use in critical care”. In: *Drug Design, Development and Therap* 12 (June 2018), pp. 1759–1767.

[25] John A. Howe et al. “Selective small-molecule inhibition of an RNA structural element”. In: *Nature* 526.7575 (Oct. 2015), pp. 672–677.

[26] Hasane Ratni et al. “Discovery of Risdiplam, a Selective Survival of Motor Neuron-2 (SMN2) Gene Splicing Modifier for the Treatment of Spinal Muscular Atrophy (SMA)”. In: *Journal of Medicinal Chemistry* 61.15 (Aug. 2018), pp. 6501–6517.

[27] Jacopo Manigrasso, Marco Marcia, and Marco De Vivo. “Computer-aided design of RNA-targeted small molecules: A growing need in drug discovery”. In: *Chem* 7.11 (Nov. 2021), pp. 2965–2988.

[28] Yanqiu Shao and Qiangfeng Cliff Zhang. “Targeting RNA structures in diseases with small molecules”. In: *Essays in Biochemistry* 64.6 (2020), pp. 955–966.

[29] G. Padroni et al. “Systematic analysis of the interactions driving small molecule-RNA recognition”. In: *RSC Medicinal Chemistry* 11.7 (2020), pp. 802–813.

[30] Noreen F. Rizvi et al. "Targeting RNA with Small Molecules: Identification of Selective, RNA-Binding Small Molecules Occupying Drug-Like Chemical Space". In: *SLAS DISCOVERY: Advancing the Science of Drug Discovery* 25.4 (Apr. 2020), pp. 384–396.

[31] Brittany S. Morgan et al. "Discovery of Key Physicochemical, Structural, and Spatial Properties of RNA-Targeted Bioactive Ligands". In: *Angewandte Chemie - International Edition* 56.43 (2017), pp. 13498–13502.

[32] Ankita Mehta et al. "SMMRNA: a database of small molecule modulators of RNA". In: *Nucleic Acids Research* 42.D1 (Jan. 2014), pp. D132–D141.

[33] Subodh Kumar Mishra and Amit Kumar. "NALDB: nucleic acid ligand database for small molecules targeting nucleic acid". In: *Database* 2016 (Feb. 2016), baw002.

[34] Matthew D. Disney et al. "Inforna 2.0: A Platform for the Sequence-Based Design of Small Molecules Targeting Structured RNAs". In: *ACS Chemical Biology* 11.6 (June 2016), pp. 1720–1728.

[35] Brittany S. Morgan, Jordan E. Forte, and Amanda E. Hargrove. "Survey and summary insights into the development of chemical probes for RNA". In: *Nucleic Acids Research* 46.16 (2018), pp. 8025–8037.

[36] Saisai Sun, Jianyi Yang, and Zhaolei Zhang. "RNALigands: a database and web server for RNA–ligand interactions". In: *RNA* 28.2 (Feb. 2022), pp. 115–122.

[37] Illimar Hugo Rekand and Ruth Brenk. "DrugPred_RNA—A Tool for Structure-Based Druggability Predictions for RNA Binding Sites". In: *Journal of Chemical Information and Modeling* 61.8 (Aug. 2021), pp. 4068–4081.

[38] Filip Stefaniak and Janusz M. Bujnicki. "AnnapuRNA: A scoring function for predicting RNA-small molecule binding poses". In: *PLOS Computational Biology* 17.2 (Feb. 2021), e1008309.

[39] Zhihai Liu et al. "PDB-wide collection of binding data: current status of the PDBbind database". In: *Bioinformatics* 31.3 (Feb. 2015), pp. 405–412.

[40] William M. Hewitt, David R. Calabrese, and John S. Schneekloth. "Evidence for ligandable sites in structured RNA throughout the Protein Data Bank". In: *Bioorganic and Medicinal Chemistry* 27.11 (2019), pp. 2253–2260.

[41] Ting Zhou et al. "RPocket: an intuitive database of RNA pocket topology information with RNA-ligand data resources". In: *BMC Bioinformatics* 22.1 (Dec. 2021), p. 428.

[42] H. M. Berman. "The Protein Data Bank". In: *Nucleic Acids Research* 28.1 (Jan. 2000), pp. 235–242.

[43] Saqib Mir et al. "PDBe: towards reusable data delivery infrastructure at protein data bank in Europe". In: *Nucleic Acids Research* 46.D1 (Jan. 2018), pp. D486–D492.

[44] Naveen Michaud-Agrawal et al. "MDAnalysis: A toolkit for the analysis of molecular dynamics simulations". In: *Journal of Computational Chemistry* 32.10 (July 2011), pp. 2319–2327.

[45] Peter Eastman et al. "OpenMM 7: Rapid development of high performance algorithms for molecular dynamics". In: *PLOS Computational Biology* 13.7 (July 2017), e1005659.

[46] P. J. A. Cock et al. “Biopython: freely available Python tools for computational molecular biology and bioinformatics”. In: *Bioinformatics* 25.11 (June 2009), pp. 1422–1423.

[47] Limin Fu et al. “CD-HIT: accelerated for clustering the next-generation sequencing data”. In: *Bioinformatics* 28.23 (Dec. 2012), pp. 3150–3152.

[48] Daniel Probst and Jean-Louis Reymond. “Visualization of very large high-dimensional data sets as minimum spanning trees”. In: *Journal of Cheminformatics* 12.1 (Dec. 2020), p. 12.

[49] Damien Monet et al. “mkgridXf : Consistent Identification of Plausible Binding Sites Despite the Elusive Nature of Cavities and Grooves in Protein Dynamics”. In: *Journal of Chemical Information and Modeling* 59.8 (Aug. 2019), pp. 3506–3518.

[50] Stéfan van der Walt et al. “scikit-image: image processing in Python”. In: *PeerJ* 2 (June 2014), e453.

[51] Pauli Virtanen et al. “SciPy 1.0: fundamental algorithms for scientific computing in Python”. In: *Nature Methods* 17.3 (Mar. 2020), pp. 261–272.

[52] G. Madhavi Sastry et al. “Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments”. In: *Journal of Computer-Aided Molecular Design* 27.3 (Mar. 2013), pp. 221–234.

[53] Mats H. M. Olsson et al. “PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical p K _a Predictions”. In: *Journal of Chemical Theory and Computation* 7.2 (Feb. 2011), pp. 525–537.

[54] Edward Harder et al. “OPLS3: A Force Field Providing Broad Coverage of Drug-like Small Molecules and Proteins”. In: *Journal of Chemical Theory and Computation* 12.1 (Jan. 2016), pp. 281–296.

[55] Tom Halgren. “New Method for Fast and Accurate Binding-site Identification and Analysis”. In: *Chemical Biology & Drug Design* 69.2 (Feb. 2007), pp. 146–148.

[56] Thomas A. Halgren. “Identifying and Characterizing Binding Sites and Assessing Druggability”. In: *Journal of Chemical Information and Modeling* 49.2 (Feb. 2009), pp. 377–389.

[57] Rachel Torchet et al. “The iPPI-DB initiative: a community-centered database of protein–protein interaction modulators”. In: *Bioinformatics* 37.1 (Apr. 2021), pp. 89–96.

[58] Daniel Probst and Jean-Louis Reymond. “SmilesDrawer: Parsing and Drawing SMILES-Encoded Molecular Structures Using Client-Side JavaScript”. In: *Journal of Chemical Information and Modeling* 58.1 (Jan. 2018), pp. 1–7.

[59] Alexander S Rose et al. “NGL viewer: web-based molecular graphics for large complexes”. In: *Bioinformatics* 34.21 (Nov. 2018), pp. 3755–3758.

[60] Laura R. Ganser et al. “The roles of structural dynamics in the cellular functions of RNAs”. In: *Nature Reviews Molecular Cell Biology* 20.8 (2019), pp. 474–489.

[61] Christopher A Lipinski et al. “Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings 1PII of original article: S0169-409X(96)00423-1. The article was originally published in Advanced Drug Delivery Reviews 23 (1997) 3–25. 1”. In: *Advanced Drug Delivery Reviews* 46.1-3 (Mar. 2001), pp. 3–26.

[62] Daniel F. Veber et al. “Molecular Properties That Influence the Oral Bioavailability of Drug Candidates”. In: *Journal of Medicinal Chemistry* 45.12 (June 2002), pp. 2615–2623.

Chapter 3

Identifying small molecules binding sites in RNA conformational ensembles with SHAMAN

The rational targeting of RNA with small molecules is hampered by our still limited understanding of RNA structural and dynamic properties. Most *in silico* tools for binding site identification rely on static structures and therefore cannot face the challenges posed by the dynamic nature of RNA molecules. Here we present SHAMAN, a computational technique to identify potential small-molecule binding sites in RNA structural ensembles. SHAMAN enables exploring the conformational landscape of RNA with atomistic molecular dynamics and at the same time identifying RNA pockets in an efficient way with the aid of probes and enhanced sampling techniques. In our benchmark composed of large, structured riboswitches as well as small, flexible viral RNAs, SHAMAN successfully identified all the experimentally resolved pockets and ranked them among the most favorite probe hotspots. Overall, SHAMAN sets a solid foundation for future drug design efforts targeting RNA with small molecules, effectively addressing the long-standing challenges in the field[1].

[1]The reader can find a step-by-step-tutorial of SHAMAN in Appendix A

3.1 Introduction

RNA molecules, initially thought to be only carriers of genetic information from gene to proteins, are now known to perform a variety of biological functions, such as regulating the process of protein synthesis and defending against the entry of foreign nucleic acids into cells [1–4]. Alongside these findings, modulation of RNA functions is becoming a promising therapeutic approach for treating diseases such as cancer, viral infections, cardiovascular and muscular disorders, and neurodegenerative conditions [5–7]. Besides classical approaches, such as the design of antisense oligonucleotides interfering with mRNAs or directly editing RNA with CRISPR-Cas9, targeting RNA with small molecules is emerging as a promising strategy [8–11] in terms of number of potential targets, bioavailability, and delivery [11–15]. Although in recent years the research in this field has surged [16, 17], the number of FDA-approved drugs is still limited and the compounds currently available on the market were identified exclusively by costly and time-consuming experimental screenings [18–20]

Computer-aided drug design (CADD) provides several essential tools to assist various stages of drug discovery, from druggability assessment to virtual screening for hit identification, binding affinity calculations, and generative methods for lead optimization. While these tools are well established for proteins, their application to RNA molecules is still in its infancy. The available biochemical and structural data is gradually elucidating the chemical properties of RNA binders [21] and the structural properties of RNA binding sites [22]. This knowledge has been stimulating the development of ligand-[23, 24] and 2D structure-[25–27] based virtual screening approaches, 3D binding-site detection tools [28–32], docking software [33–36] and scoring functions [37–40] specific for RNA molecules. However, our understanding of the structural and dynamic properties of RNA molecules and their interaction with small molecules still remains limited, thus ultimately hindering the rational design of novel and effective compounds [41].

In the cellular context, function-specific biological signals trigger complex multi-step RNA conformational changes that in turn guide a variety of RNA functions, such as ligand sensing and signaling, catalysis, or co-transcriptional folding [42, 43]. These conformational changes and the underlying dynamics are influenced both by the inherent flexibility of RNA molecules, i.e. many large-scale motional modes spanning a variety of timescales, and other cellular co-factors [44]. Despite the significant efforts to characterize RNA dynamics using both experimental [45], *in silico* [46], and integrative approaches [47], most available tools for CADD, and in particular for the identification of small molecules binding sites, still rely on a static description of RNA structure [28–32]. The only exception is SILCS-RNA31 where potential binding sites are identified by exploring the conformation of the target RNA with small cosolvent probes, similar to mixed-solvent approaches already extensively used for proteins [48]. While SILCS-RNA can describe small structural rearrangements induced by the probes, it is not designed to capture large RNA conformational changes and, therefore, it is not able to detect binding sites present in metastable states that are marginally populated yet crucial for therapeutic applications [41–43, 49].

Here, we present SHAdow Mixed solvent metAdyNamics (SHAMAN), a computational technique for binding site identification in dynamic RNA structural ensembles. Thanks to its unique parallel

architecture, SHAMAN allows at the same time to: *i)* explore the conformational landscape of RNA with atomistic explicit-solvent molecular dynamics (MD) simulations driven by state-of-the-art forcefields and *ii)* identify potential small-molecules binding sites in an efficient way with the aid of probes and the metadynamics [50] enhanced-sampling technique. SHAMAN was benchmarked on a set of biologically relevant target systems, including large, structured riboswitches as well as smaller highly dynamic RNAs involved in viral proliferation. Our method successfully identified all the experimentally resolved pockets present in our benchmark set and was able to rank them among the most favorite probe hotspots. Our work constitutes an advanced computational pipeline for binding site identification in dynamic RNA structural ensembles, thus providing crucial information for structure-based rational design of novel compounds targeting RNA.

3.2 Results

This section is organized as follows. First, we provide a general overview of SHAMAN and illustrate its accuracy in identifying experimentally resolved binding sites in a set of biologically relevant RNA targets. Second, we focus on the probes used in our SHAMAN simulations and investigate their relation to physico-chemical features of both the RNA pockets and the small molecules bound to them in known experimental structures. We then compare SHAMAN with state-of-the-art tools for binding site prediction in RNA. Finally, we present two case studies, the FNM riboswitch and the HIV-1 TAR, to i) demonstrate how SHAMAN can be used to study well-structured as well as more flexible RNAs; ii) highlight the main strengths of our technique in modeling both local and global flexibility of the target. A complete analysis of the systems in our benchmark set is reported in Supplementary Information (Sec. 3.5.1 and Figs. 3.8-3.12).

3.2.1 Overview on the SHAMAN approach

SHAMAN is a computational technique that uses small fragments or probes and all-atomistic explicit-solvent MD simulations to identify potential small-molecule binding sites in RNA structural ensembles (Fig. 3.1A). SHAMAN is based on a unique architecture in which multiple replicas of the system are simulated in parallel (Fig. 3.1B). A mother simulation, containing only RNA and possibly structural ions, explores the conformational landscape of the target and communicates the positions of the RNA atoms to the replicas. Each replica contains a different probe that explores the RNA conformation provided by the mother simulation using the metadynamics enhanced-sampling approach [50]. Soft positional restraints applied to the RNA backbone atoms of the replica allow for local induce-fit effects caused by the probes, while following or "shadowing" the conformational changes of the mother RNA simulation. This parallel architecture enables an efficient exploration of the same RNA conformation by different probes and to build, for each representative cluster of RNA conformations, a set of potential small-molecule binding sites or SHAMAPs (Fig. 3.1C). Each SHAMAP corresponds to a region of space occupied with high probability by at least one probe and is ranked by the binding free energy ΔG of the probe(s) to a specific RNA conformation (Fig. 3.1D). A more detailed description of the SHAMAN method can be found in Sec. 3.4.1.

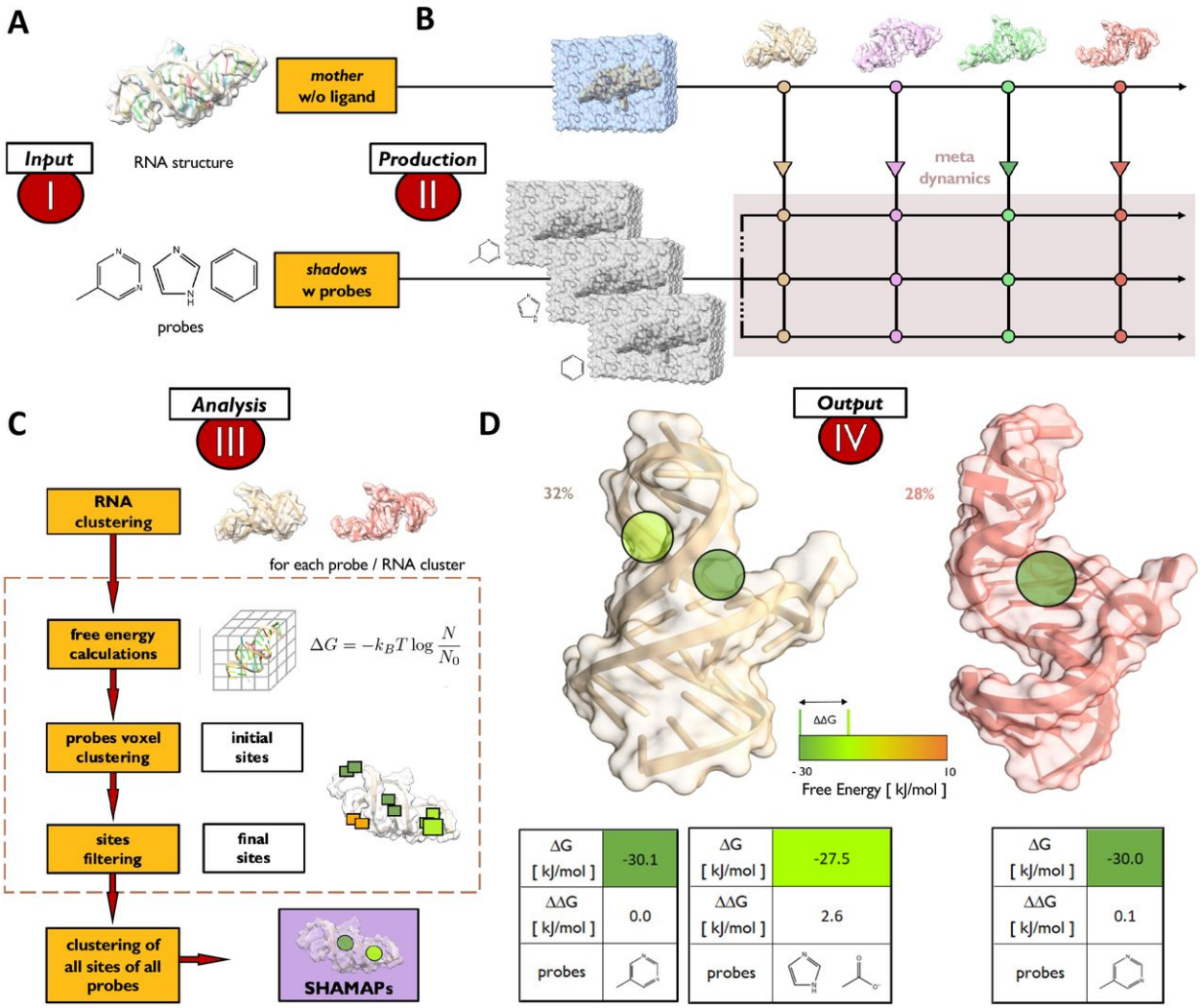


Figure 3.1: Overview of the SHAMAN approach. A) *Input Stage:* Selection of the RNA target structure, as well as the selection of the probes to initialize the *mother* and replica systems, each one with a different probe. **B)** *Production Stage:* The unbiased/unrestrained MD simulation of the *mother* system communicates the positions of the RNA backbone atoms to the replicas, which are restrained to follow the mother like shadows. The probe exploration of the RNA conformational space is accelerated by metadynamics. **C)** *Analysis Stage:* From top to bottom: i) The sampled RNA ensemble is clustered into a set of representative conformations; ii) For each cluster and probe, a free-energy map is calculated from the probe occupancy during the course of the trajectory simulation; iii) Voxels in the free-energy maps are clustered together into an initial set of interacting sites; iv) For each interacting site, free energy and buriedness scores are evaluated, and sites too exposed to solvent are discarded; v) For each RNA cluster, all interacting sites obtained from all probes are clustered together into SHAMAPs. **D)** *Output Stage:* Two RNA representative clusters with populations equal to 32% (light brown cartoon, left panel) and 28% (red/pink cartoon, lower right panel) with the corresponding SHAMAPs (represented by green circles). For each SHAMAP, we provide the binding free energy to RNA (ΔG) and the difference with respect to the lowest free energy (top-scored) SHAMAP ($\Delta\Delta G$) along with a list of probes that explored the corresponding regions.

3.2.2 Benchmark of the SHAMAN accuracy

The accuracy of SHAMAN in identifying experimentally resolved binding sites was evaluated on 7 biologically relevant systems, including riboswitches (Fig. 3.2A) and viral RNAs (Fig. 3.2B). For each system, SHAMAN simulations were initialized from both holo conformations after the removal of the ligand (holo-like) and, when available, apo conformations, resulting in a total of 12 runs (Tab. 3.1 and 3.2). The validation set was composed of 14 unique binding pockets obtained from 69 experimental structures of riboswitches (Tab. 3.3) and viral RNAs in complex with different ligands (Tab. 3.4). For each simulation, the accuracy was defined in terms of the distance between our SHAMAPs and the ligand position in the reference experimental structures (Eq. (3.10) and Fig. 3.2C).

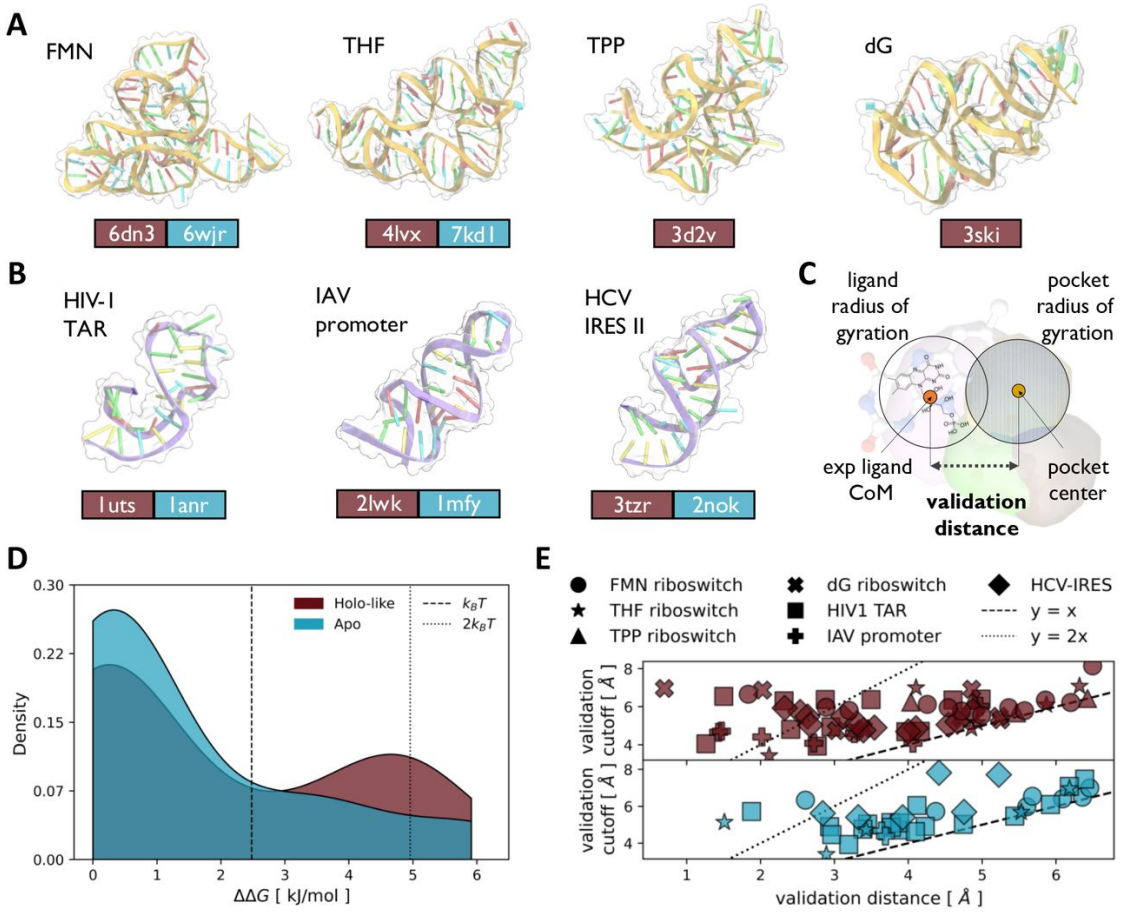


Figure 3.2: Assessment of SHAMAN accuracy. A) A cartoon-surface representation of the four riboswitches present in our benchmark set (Tab. 3.1), with the corresponding name in the upper left of each panel. In the lower part, the PDB id of the starting structure used for in our SHAMAN simulations is reported in a red brown and blue cyan box for the holo-like and apo case (when available), respectively. The cartoon representations correspond to the holo-like input structures. **B)** As in panels A), for the three viral systems RNAs of our benchmark set (Tab. 3.1). **C)** Definition of the validation distance (Eq. (3.10)) as the distance between the free-energy weighted center of predicted an interacting sites and the center of mass of the experimental ligand. **D)** $\Delta\Delta G$ distribution of the probes that correctly identified known experimental pockets for holo-like (brown) and apo simulations (cyan). **E)** Scatter plots of the validation distance (x axis) and cutoff defined by Eq. (3.10) (y axis) for holo-like (brown, upper panel) and apo (cyan, lower panel) simulations. The dashed line indicates validation distances equal to the validation cutoff, while the dotted line corresponds to half the validation cutoff. Each system is identified by a different marker shape, as defined in the legend.

SHAMAN was able to identify all the experimentally resolved pockets present in all the systems of our benchmark set, both when initializing the simulations from holo-like and apo conformations (Tab. 3.5 and 3.6). Most importantly, the experimental binding sites were ranked among the most probable SHAMAPs in each corresponding run. To quantify the rank, we defined the difference in binding free energy $\Delta\Delta G$ between each SHAMAP and the one with the lowest free energy (Eq. 3.9). When starting from the apo conformation of the RNA target molecule, the $\Delta\Delta G$ of the SHAMAPs overlapping with the ligands was in 80% of cases below k_BT and in 100% of cases below $2k_BT$ (Fig. 3.2D). When starting from holo-like conformations, these percentages dropped to 64% and 84% (Fig. 3.2D). Ranking the experimental binding pockets among the SHAMAPs with the lowest free energy (top scored) is fundamental in the context of CADD, and in particular in virtual screening applications (Sec. 3.3).

The geometrical proximity of our SHAMAPs to the experimental binding sites present in our benchmark set was noteworthy. The average distance between the centers of the interacting sites overlapping with a ligand and its position in the experimental structure was equal to 3.8 Å and 4.4 Å in the holo-like (Fig. 3.2E, upper panel) and apo (Fig. 3.2E, lower panel) cases, respectively. Both values are relatively small when compared to the distance threshold used in our validation criterion (Eq. 3.10), which was defined as the sum of the radii of gyration of the SHAMAP (on average 1.6 Å, Fig. 3.2E) and the ligand (on average 3.7 Å, Fig. 3.2F). As expected, this proximity to the experimental binding sites was remarkably greater in the simulations initiated from holo-like conformations in which the binding sites were already present. As a matter of fact, 22% of the successful interacting sites identified in the holo-like simulations were close to the experimental pocket by half of our distance threshold, while this holds only for 1% of the apo simulations.

3.2.3 Analysis of the probes

Two sets of probes were used in the SHAMAN benchmark described in the previous section. The first set of 8 probes (Tab. 3.7) was previously used in the development of SILCS-RNA [31] and was mostly composed of aliphatic compounds selected to represent specific types of interaction with the RNA target. This set includes: acetate (ACEY), benzene (BENX), dimethyl-ether (DMEE), formamide (FORM), imidazole (IMIA), methyl-ammonium (MAMY), methanol (MEOH), and propane (PRPX). A second set of 5 probes (Tab. 3.8) was generated in this work using a fragmentation protocol (Sec. 3.4.2) applied to the ligands present in *i)* the HARIBOSS [22] database of RNA-ligand resolved structures; and *ii)* the R-BIND [26] database of bioactive small molecules targeting RNA. This second set includes mostly aromatic compounds: benzene (BENX), dihydro-pyrido-pyrimidinone-Imidazo-pyridine (BENF), benzothiophene (BETH), methyl-pyrimidine (MEPY), and the cyclic non-aromatic piperazine (PIRZ).

We first explored the relation between the probes that successfully identified experimental binding sites and some of the structural characteristics features of the RNA pockets. Aromatic probes showed a preference for exploring cavities buried deep inside the RNA structure (Fig. 3.3A, dark green bars), with an estimated average buriedness of 0.75 ± 0.06, which is relatively high compared to known RNA-small molecule pockets (Fig. 3.3B). On the other hand, non-aromatic or aliphatic probes displayed two distinct patterns. FORM, MEOH, and MAMY selectively explored shallow

pockets with an average buriedness of 0.59 ± 0.04 (Fig. 3.3A, olive green bars), while DMEE, PRPX, and ACEY promiscuously explored pockets with varying solvent exposure and an average buriedness of 0.70 ± 0.08 (Fig. 3.3, olive green bars). PIRZ exhibited an intermediate behavior, with an average buriedness of 0.65 ± 0.06 (Fig. 3.3A, brown bar). As a consequence, aromatic probes were particularly successful (66% of cases) in identifying riboswitch binding sites, which in our validation set typically resided in buried cavities (Fig. 3.3C). For example, the location of the representative riboswitch binder GNG (PDB 3ski bound to 2'-Deoxyguanosine riboswitch) was exclusively identified by aromatic probes (Fig. 3.3D). On the other hand, aliphatic probes identified pockets with high likelihood (70%) in viral RNAs (Fig. 3.3E), whose inherent flexibility resulted in shallow cavities exposed to solvent. An example is the binding site of SS0, a typical viral RNA binder (PDB 3tzr), which was primarily identified by aliphatic probes (Fig. 3.3F).

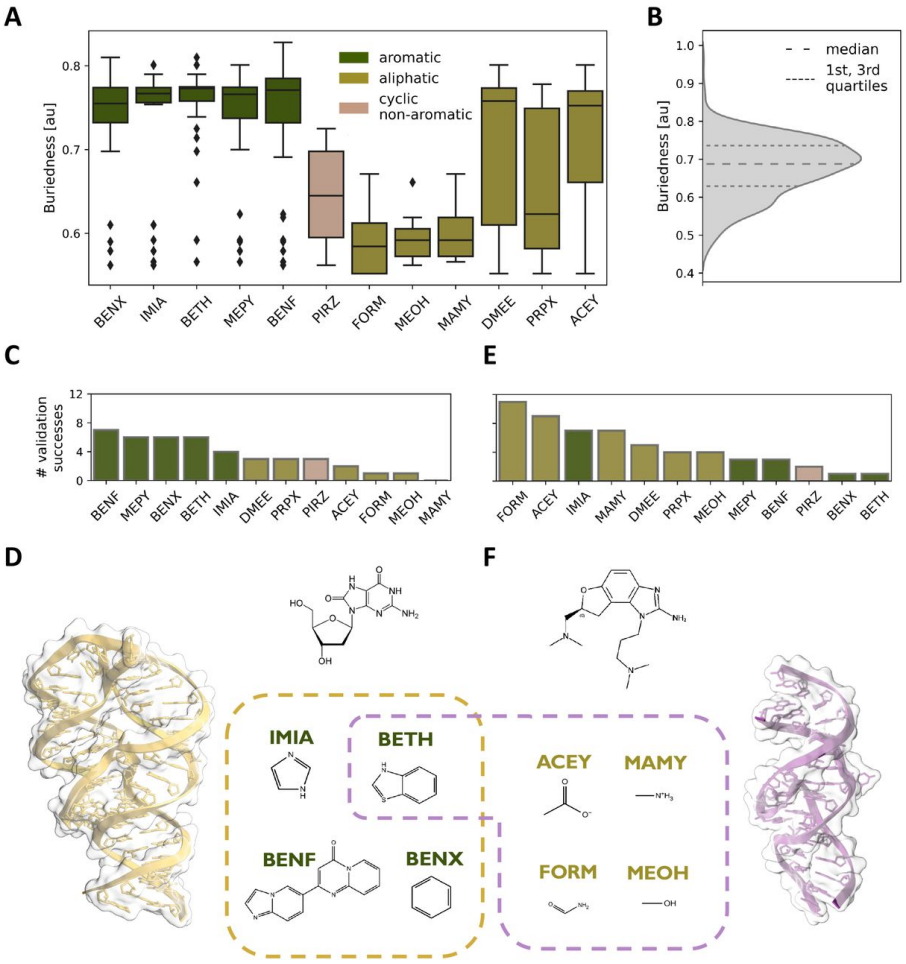


Figure 3.3: Analysis of the SHAMAN probes. A) Violin plots representing the buriedness of the experimental pockets (y-axis) successfully identified by a given SHAMAN probe (x-axis). Buriedness values were extracted from the HARIBOSS database [22] (Tab. 3.3 and 3.4). Outliers are shown as black diamonds. **B)** Buriedness distribution for the RNA pockets occupied by ligands in all the structures deposited in HARIBOSS. **C)** Total number of times that a probe explored an experimental binding site in the riboswitches of our validation set. **D)** Cartoon representation of the 2'-deoxyguanosine (dG) riboswitch (PDB 3ski) with 2D structure of the GNG binder. In the dashed box, the 2D structures of the probes that identified the GNG binding site. **E)** As in panel C, for the viral RNAs of our validation set. **F)** Cartoon representation of the RNA from the Hepatitis C Virus (PDB 3tzr) with 2D structure of the SS0 binder. In the dashed box, the 2D structures of the probes that identified the SS0 binding site.

Although the main goal of SHAMAN is pocket identification, motivated by its perspective use in virtual screening and ligand optimization (Sec. 3.3), we also investigated a possible link between the similarity of a given probe to a ligand and its ability to identify the corresponding experimental pocket. We started by comparing standard physico-chemical properties of the entire ligand or the corresponding Murcko scaffold (Sec. 3.4.2). Our analysis did not reveal a strong correlation between ligands and probes (Tab. 3.9). We then calculated the Tanimoto similarity using different fingerprints (Sec. 3.4.2). Our analysis suggested that we cannot predict whether a probe would be successful based on its similarity with a ligand (Fig. 3.14). However, based on a statistical classification point of view (Sec. 3.4), we can conclude that probes that did not resemble the ligand were highly unlikely to successfully identify the corresponding binding site, with a negative predictive value (NPV) equal to 0.82 (Sec. 3.4.2, Eq. (3.11) and Tab. 3.10).

3.2.4 Comparison with other tools

We compared SHAMAN with three state-of-the-art computational tools for small-molecule binding site prediction on RNA molecules: SiteMap [51], BiteNet [52], and RBinds[53, 54]. For all the systems in our benchmark set, we tested the ability of these tools to correctly predict the RNA nucleotides interacting with small molecules in experimentally determined structures (Matherials and Methods). First, we determined the quality of the predictions obtained from holo-like conformations using only the corresponding experimental holo structure as ground truth (Tab. 3.1, red column). SHAMAN and BiteNet outperformed SiteMap and RBinds (Fig. 3.4A) in terms of Matthews Correlation Coefficient (MCC score), a comprehensive measure of predictive quality for binary classifiers (Sec. 3.4.4). The low MCC scores of SiteMap and RBinds were mostly due to their low accuracy and precision. While the quality of the predictions obtained with SHAMAN and BiteNet was comparable, the precision of our approach was more variable across our benchmark set, with a tendency to overestimate the number of interacting nucleotides. Given that SHAMAN accounts for the flexibility of the RNA target, we hypothesized that this was the result of the prediction of alternative binding pockets not present in the single holo structure used as ground truth. To verify this hypothesis, we assessed the quality of predictions by considering as ground truth for each system the set of interacting nucleotides in all the experimental binding sites of our validation set (Tab. 3.3 and 3.4), Sec. 3.4.4). With this definition, SHAMAN precision and overall MCC score improved (Fig. 3.4B), in support of our hypothesis. Finally, to simulate a common drug discovery scenario in which only the structure of the apo state is available, we tested the quality of the predictions obtained from apo conformations (Tab. 3.1, cyan column). In this case, the quality of SHAMAN predictions was superior to BiteNet (Fig. 3.4C) as our approach was able to identify with high accuracy and precision the correct set of interacting nucleotides in all the reference experimental structures. These results clearly indicate that prediction tools that do not account for the flexibility of the RNA target are not able to predict binding sites formed upon local or global structural rearrangements.

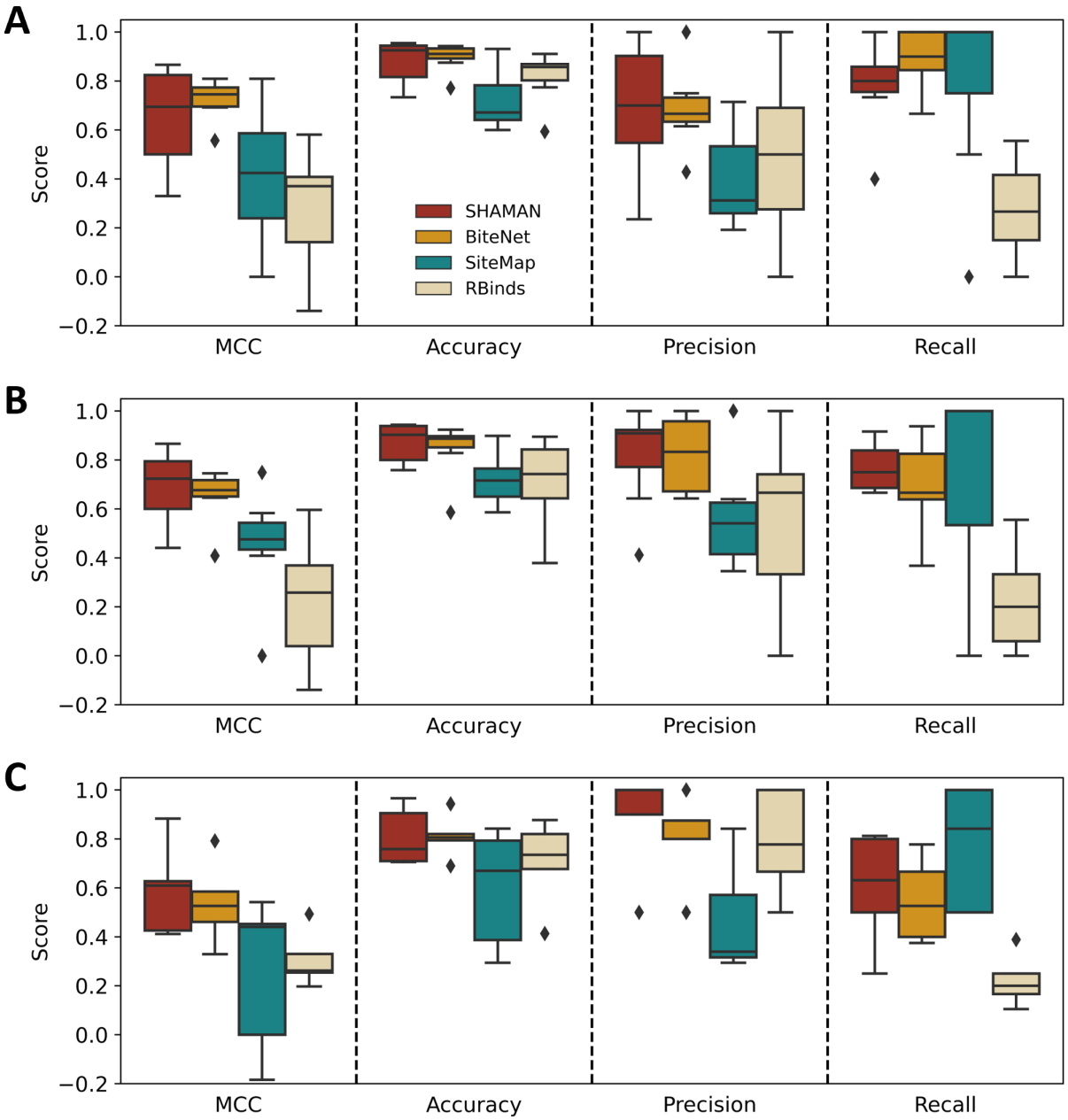


Figure 3.4: Comparison with other tools. From left to right, boxplots reporting the predictive quality of different binding site prediction tools evaluated by four statistical metrics for binary classifiers (Sec. 3.4.4). **A)** Binding site prediction on the holo-like systems (Tab. 3.1, red column) validated against the single corresponding experimental structure. **B-C)** Binding site prediction on holo-like (**B**) and apo (**C**) systems (Tab. 3.1, red and cyan columns) against all the validation structures (Tab. 3.3 and 3.4, Sec. 3.4.4). Each box represents the interquartile range between the first and third quartiles, with the median indicated by a horizontal black line. Outliers are marked as black diamonds.

3.2.5 The case of FMN riboswitch

The Flavin MonoNucleotide (FMN) riboswitch is an RNA molecule found in bacteria that regulates FMN gene expression via binding the FMN metabolite [55]. Being the target of ribocil [18], one of the few FDA-approved compounds targeting RNA, the FMN riboswitch constitutes a natural test case for our drug-design purposes. As of today, 19 X-ray structures of the FMN riboswitch are deposited in the PDB database, 3 in apo and 16 in holo conformations. The 9 unique small molecules resolved in the holo structures fall into three main families: the cognate FMN family, the synthetic ribocil family, and the tetracyclic DKM binder (Fig. 3.15). The ligands belonging to the FMN and ribocil families share a U-shaped conformation and occupy the same binding site, buried into the RNA structure within the junctional region of the six stems between the A-48 and A-85 bases (Fig. 3.5A). The DKM tetracyclic ligand exhibits instead a distinct binding mode [56] as it induces a flip in A-48 and stacks face-to-face between A-48 and G-62, resembling the apo form (Fig. 3.5B). We therefore challenged our SHAMAN approach to capture the local rearrangements of the FMN riboswitch and to identify both types of binding poses starting from a single static structure.

We tested SHAMAN starting from both holo-like (PDB 6dn3 [57]) and apo (PDB 6wjr [55]) structures (Fig. 3.5C-D). One major RNA cluster, including the initial conformations, was populated for 99% and 84% of the holo-like and apo trajectories. This limited conformational variability observed in our simulations is consistent with the structural variety resolved experimentally (Tab. 3.11), supporting the accuracy of the force field used in our SHAMAN simulations. In this predominant RNA structural cluster, our method successfully located the experimental binding sites (Fig. 3.5C-D) with very high accuracy, in the best case with a discrepancy of only 1.5 Å and 1.7 Å in the holo-like and apo simulations, respectively (Tab. 3.5). Moreover, the experimental pocket was ranked in both cases among the most probable SHAMAPs (Fig. 3.5D), with a $\Delta\Delta G$ (Eq. (3.9)) of 0.04 kJ/mol and 0.08 kJ/mol, respectively (Tab. 3.5). These results are even more remarkable if we consider the buried character of the FMN riboswitch pocket, which made it difficult for the probes to access it and sample it accurately. As discussed above (Fig. 3.3), most of the probes that successfully identified this buried pocket were aromatic, both in the holo-like (83%) and apo (75%) cases (Fig. 3.5E).

Notably, the two distinct binding modes of FMN and DKM ligands were identified with comparable accuracy in both runs starting from holo-like and apo conformations. Each of these starting conformations was representative of one single binding mode: in the holo-like structure, the A-48 basis faces A-85, while in the apo case it is flipped onto A-49. SHAMAN enabled the identification of both binding modes, including the one not present in the starting conformation, something not possible with algorithms based on static structures. This is highlighted by superimposing the SHAMAPs found in the holo-like and apo simulations to the corresponding starting structure (Fig. 3.5C-D, insets). The detection of both binding modes was made possible by simulating different probes in parallel and allowing for induce-fit effects in the RNA conformation sampled by the mother simulation (Sec. 3.3). In the holo-like case, the BENX and IMIA probes captured the tail of the FMN binder (left panel, Fig. 3.5F, black and green surfaces, respectively), while BENF and MEPY overlapped with the tetracyclic part of DKM (right panel Fig. 3.5F, orange and celeste surfaces, respectively). In the apo case, MEPY interacting site overlapped with both ligands, but the tetracyclic part of DKM was captured only by IMIA (Fig. 3.5G).

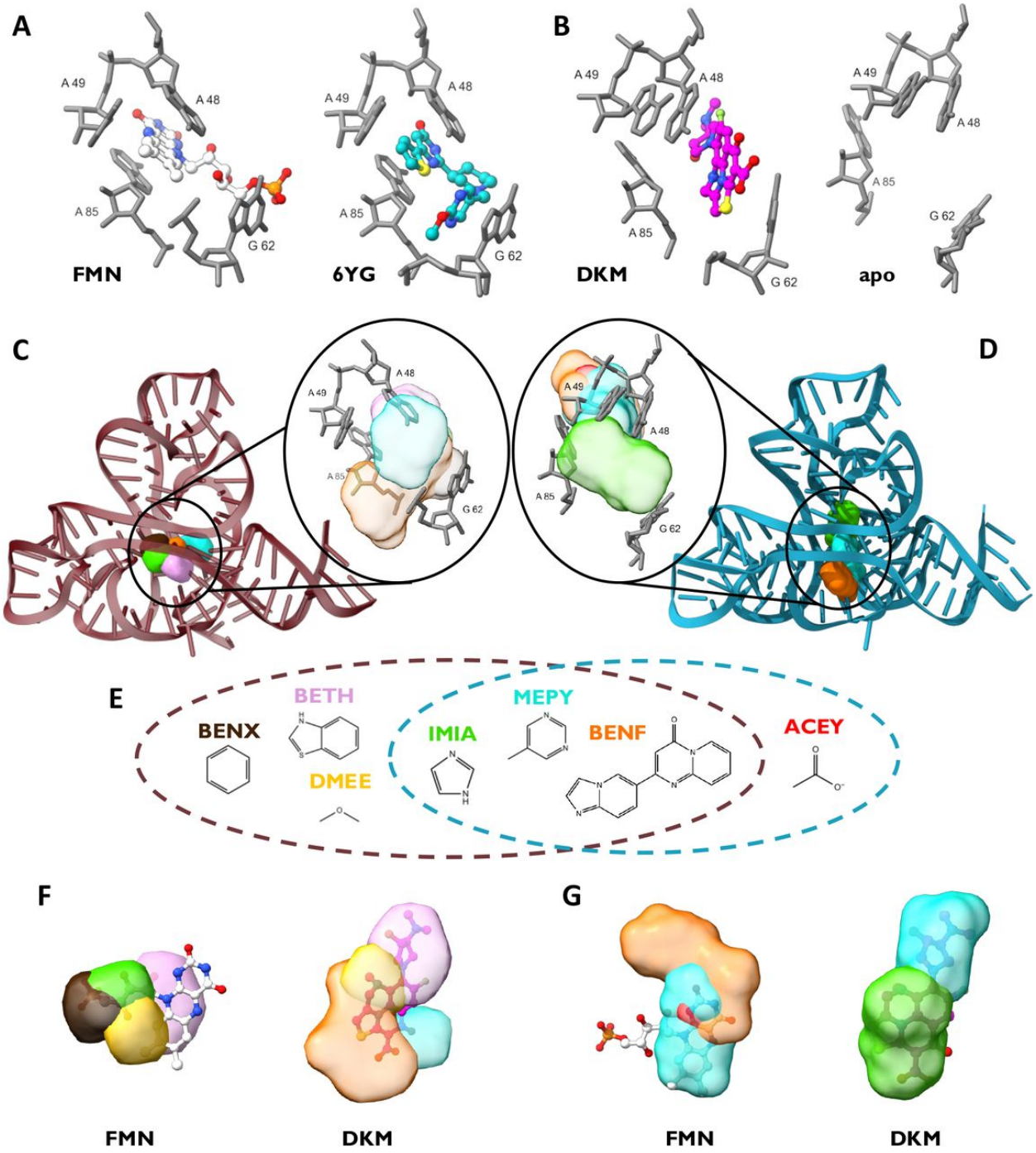


Figure 3.5: The case of the FMN riboswitch. A) Key RNA binding site residues for the FMN ligand (PDB 2yie) and ribocil (PDB 5kx9) families. **B)** Key RNA binding site residues for the DKM ligand (PDB 6bfb) and in the apo conformation (PDB 6wjr). **C-D)** Cartoon representation of holo-like (C) and apo (D) starting structures used in the SHAMAN simulations of the FMN riboswitch. In the insets, the key binding site residues are overlayed with the probe densities (colors as in Tab. 3.7 and 3.8. **E)** 2D structures of the probes that successfully identified the experimental binding sites in the FMN riboswitch. The brown and cyan dashed circles indicate the successful probes in the holo-like and apo simulations, respectively. **F-G)** For the holo-like (F) and apo (G) simulations, the SHAMAPs with best overlap with FMN (left) and DKM (right) ligands, representing the two different binding modes of the FNM riboswitch.

3.2.6 The case of HIV-1 TAR element

The HIV-1 Trans-activation response element (HIV-1 TAR) is a highly flexible, non-coding RNA molecule responsible for regulating HIV-1 gene expression through binding with Tat protein [58, 59]. Understanding its conformational dynamics is crucial for drug development but remains challenging due to the major structural changes occurring upon binding diverse partners [60, 61]. This conformational plasticity of HIV-1 TAR is reflected in the more than 20 resolved structures, primarily determined by NMR, alone or bound to different ligands in water-exposed cavities. Our validation set was composed of 5 holo structures bound to different small molecules with different binding modes (Fig. 3.16) in the groove between the bulge UCU and the apical loop CUGGGA (residues 23-25 and 30-35, Fig. 3.6A). This is a crucial region that also encodes the Tat protein binding site [62]. One of these structures (PDB 2l8h) indicates the presence of a transient and functionally relevant pocket formed upon binding to the MV2003 small molecule [60]. Given its complex dynamics, HIV-1 TAR constitutes an important benchmark of the capabilities of SHAMAN to detect binding sites appearing upon global conformational changes of the target molecule.

We tested SHAMAN starting from two structures of HIV-1 TAR, one in holo-like (PDB 1uts [63]) and one in apo (PDB 1anr [64]) conformation. Both simulations recapitulated the expected flexibility of the target RNA molecule and identified multiple significantly populated structural clusters (Fig. 3.6B-C). A significant portion of the SHAMAPs was in the major groove of HIV-1 TAR (Fig. 3.6B-C) with a relatively high probability ($\Delta\Delta G$ within $2k_BT$). Among these, SHAMAN identified all the 5 experimental binding sites, even though the overall similarity of the RNA to the deposited structures was never below 3 Å backbone RMSD (Fig. 3.17). The most accurate overlaps with the experimental ligands were obtained with SHAMAPs detected in conformations *b* and *e* in the holo case (Fig. 3.6D) and conformations *a*, *c*, and *d* (Fig. 3.6E) in the apo case, mostly with aliphatic probes (Fig. 3.6F). The geometric accuracy in identifying the binding sites was inferior compared to the FMN riboswitch, with an average distance between binding and interacting sites of 4.0 Å and 4.1 Å for the holo-like and apo cases, respectively (Tab. 3.6). However, we consider this distance still acceptable given the high flexibility of the molecule and the shallow nature of the experimental binding sites.

Notably, SHAMAN was able to identify the cryptic binding pocket proposed by Davidson et al. in 2011 [60] (orange residues in Fig. 3B of their paper). In our simulations, this site was detected in conformation e (orange residues in Fig. 3.6C) by the ACEY and MAMY probes (color code red and pink densities, respectively). While in the work of Davidson et al. the cryptic pocket appeared in the presence of the MV2003 small molecule bound to HIV-1-TAR, here its detection was made possible by the ability of SHAMAN to describe large conformational changes of small RNAs and account for induce-fit effects of the probes (Sec. 3.3).

3.3 Discussion

Here we presented SHAMAN, a computational technique for small-molecule (SM) binding site identification in RNA structural ensembles based on all-atom MD simulations accelerated by metadynamics. We benchmarked the accuracy of our approach using a set of known RNA-SM structures,

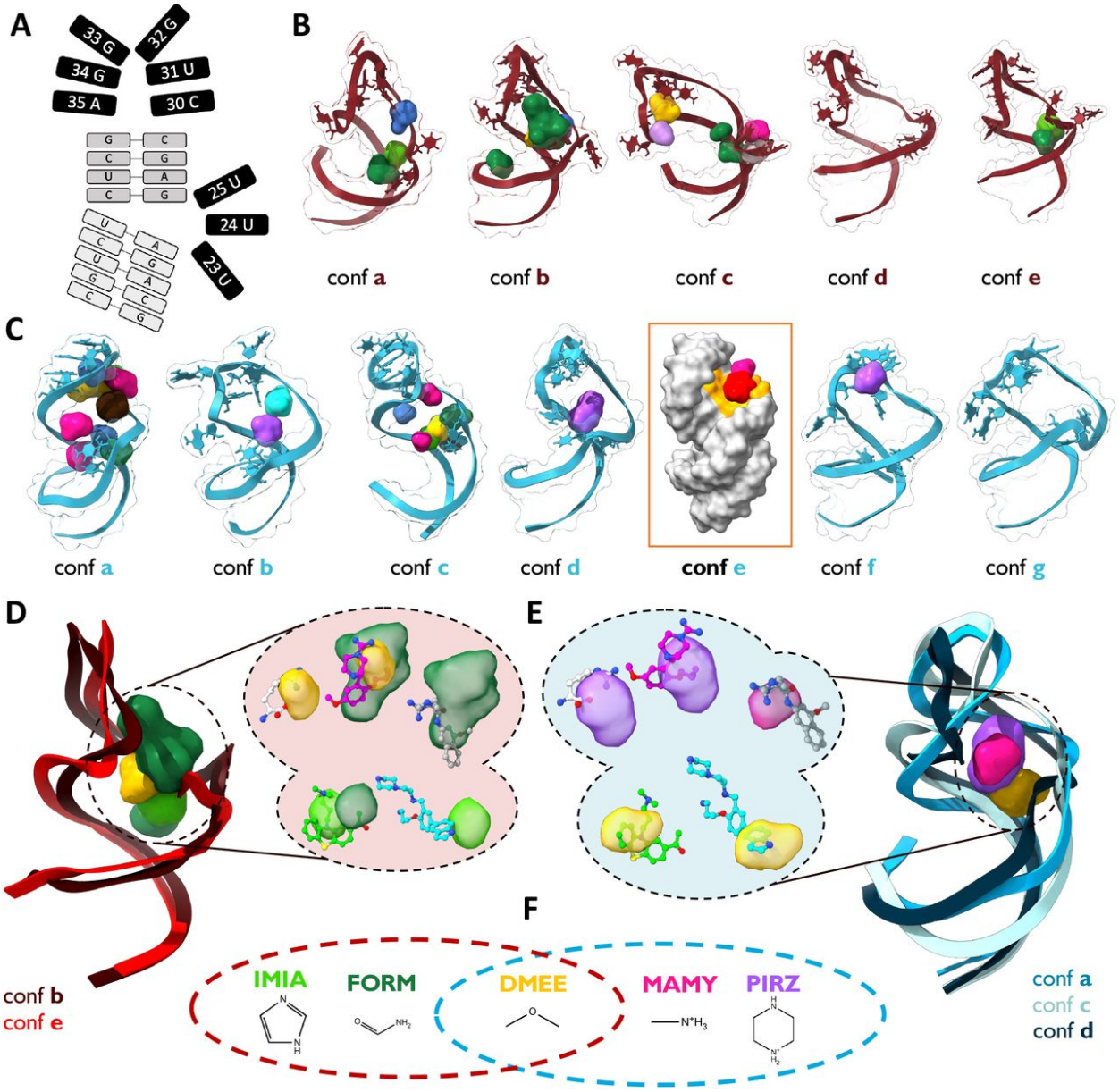


Figure 3.6: The case of the HIV-1 TAR. A) 2D structure of the HIV-1 TAR. The two stem regions are indicated in light grey; the bulge (residues 23-25) and the apical loop (residues 30-35) in black. **B-C)** Representative RNA clusters determined by the SHAMAN simulations initiated from the holo-like (B) and apo (C) conformations. SHAMAPs are visualized as solid surfaces with the color code defined in Tab. 3.7 and 3.8. The RNA state labeled as "conf e" in panel C is represented as a grey surface to highlight the orange region explored by ACEY (red density) and MAMY (rose density). This area corresponds to the cryptic binding site identified by Davidson *et al.* [60]. **D-E)** Representative RNA conformations and SHAMAPs with best overlap with the experimental binding sites found in the simulations initiated from the holo-like (D) and apo (E) conformations. In the insets, SHAMAPs that best identified the 5 ligands present in our validation set (Tab. 3.4): clockwise from top left, ARG in PDB 1arj, PMZ in PDB 1lvj, P13 in PDB 1uts, P12 in PDB 1uui, MV2003 in PDB 2l8h. **F)** 2D structures of the probes that successfully identified the experimental binding sites. The brown and cyan dashed circles indicate the successful probes in the holo-like and apo simulations, respectively.

which included large, stable riboswitches and smaller, highly flexible viral RNAs. SHAMAN was able to identify all the binding pockets observed in the experimental structures and rank them among the most favorable probe interacting hotspots, both when starting from holo-like and apo conformations of the target. The interacting sites found by the SHAMAN simulations initiated from holo-like conformations were closer to the experimental pockets than those found in the apo cases. However, in the latter case the SHAMAPs corresponding to experimental binding sites were still very accurate and ranked as the top scored interacting sites for the majority of systems. Furthermore, our predictions were more accurate in the case of rigid riboswitches, with the regions explored by the probes perfectly matching the experimental binding sites. The accuracy was still very satisfying also for viral RNA molecules considering their high flexibility.

SHAMAN emerges as one of the most advanced physics-based approaches for binding site identification in RNA structural ensembles. A major limitation of existing CADD tools in this framework is the inadequate treatment of RNA flexibility. In these regards, SILCS-RNA [31] represents the state-of-the-art computational techniques by modeling the flexibility of the target RNA using a mixed-solvent MD approach. However, the method proposed by the MacKerell group presents two important limitations. First, it makes use of positional restraints on the RNA backbone atoms and therefore is not designed to detect cavities formed upon major conformational changes. Second, SILCS-RNA was tested only by starting the MD simulations from holo structures after the removal of the bound ligand, therefore restraining the RNA target in a conformation in which the binding site is already formed. On the contrary, SHAMAN has been designed to enable the identification of pockets in dynamic RNA conformational ensembles characterized by both local and global conformational changes. The FMN riboswitch case study highlights how the target RNA molecules simulated in the replica systems have enough freedom to undergo local rearrangements induced by the probes and ultimately to capture the two distinct binding modes observed in the experimental structures. Furthermore, the challenging case study of HIV-1 TAR demonstrates that cryptic pockets formed upon global conformational rearrangements [60] can also be successfully identified by SHAMAN.

Despite the potentialities discussed above, the current implementation of SHAMAN presents two important limitations. First, the unbiased MD simulation of the RNA target in the *mother* replica will hardly ever provide a comprehensive exploration of the conformational space at low computational cost. However, this might not be a severe limitation if the scope is to determine potential druggable sites in the proximity of the metastable holo-like and apo RNA conformations resolved experimentally. To achieve a more global conformational exploration, in the future we will accelerate sampling of the RNA target in the mother replica by using enhanced-sampling techniques distributed with the PLUMED library, where SHAMAN is also implemented. Another limitation of our approach resides in the accuracy of the RNA force fields used in our MD simulations. Despite tremendous progress [65], the accuracy of molecular mechanics force fields for nucleic acids is still as high as for proteins. One way to effectively improve the underlying force field is to integrate experimental data into MD simulations. A large variety of integrative approaches, often based on Maximum Entropy and Bayesian principles [66] have been developed in the past 10 years to use ensemble-averaged experimental data, such as many NMR observables, to model accurate struc-

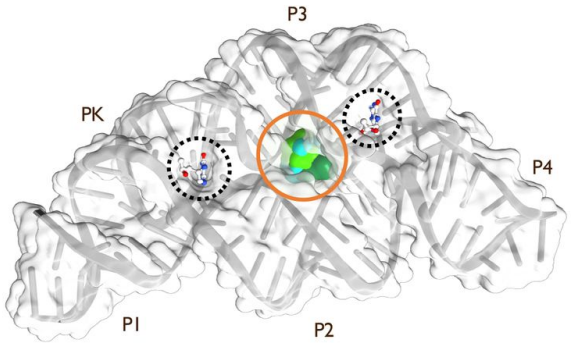
tural ensembles of dynamic proteins. These approaches have been more recently applied to the determination of RNA structural ensembles [49, 67] and will be used in the future to improve the accuracy of the RNA ensembles determined by SHAMAN. However, it should be noted that in the current implementation of SHAMAN the probe (pseudo) binding free energy is calculated without accounting for the population of the RNA structural cluster in which the binding site is found. Therefore, improving the cluster populations by means of integrative approaches will not have a significant impact on the accuracy of SHAMAN, provided that the sampling of the conformational landscape of RNA molecules is exhaustive in the first place.

In the future we foresee multiple different applications of SHAMAN in the context of CADD, in particular in combination with virtual screening applications and fragment-based drug design. Here our approach was used only to identify binding sites occupied by ligands in experimentally resolved structures. In this process, we also detected potential alternative binding sites that were in many cases ranked among the top-scored SHAMAPs. For example, in the case of the THF riboswitch, we identified a top-scored SHAMAP at the center of the RNA molecule between helix P2 and P3 (Fig. 3.7). In this region, to our knowledge, no binders have been experimentally determined yet. In the future, we will attempt at experimentally validating this pocket and eventually targeting it in a virtual screening campaign. Even more exciting is the application of SHAMAN to novel targets for which a small molecule has not been found yet. In these regards, the fact that top-scored SHAMAPs often corresponded to known binding sites will allow us to restrict virtual screening campaigns to a few localized regions.

Despite the fact that we did not find a strong correlation between successful probes and ligands, we believe that SHAMAN can provide some guidance to tailor the choice of small molecules for virtual screening or to optimize known ligands. For example, in the case of riboswitches characterized by buried cavities and viral RNA with shallower and more exposed cavities, the results of our analysis suggested the use of molecules rich in aromatic or aliphatic moieties, respectively. In addition, areas close to the location of known ligands identified by certain probes as strong interacting hotspots could provide insights about how to modify the ligand to improve its affinity or even clues about ligand binding pathways (Fig. 3.18).

One of the growing concerns with rational drug discovery approaches for RNA targeting is selectivity. Although in the present study we apply SHAMAN to RNA molecules with low sequence identity, one could consider employing our protocol to examine the uniqueness of a binding site in one target against a set of undesirable targets close in sequence (antitargets). In the case where a binding site is located in the same area across all examined RNA molecules, but it has different physico-chemical and structural properties, a cross-docking approach, i.e. docking to multiple RNAs and selecting molecules with predicted affinity for the desired target significantly higher compared to the others, can be used to identify potentially selective compounds.

In conclusion, our method provides a novel and promising foundation for future drug design efforts targeting RNA. The accuracy, reliability, and versatility of SHAMAN in identifying small-molecule binding sites across diverse RNA systems with various degrees of flexibility highlight its potential



ΔG [kJ/mol]	-31.5
ΔΔG [kJ/mol]	0.9
probes	FORM

ΔG [kJ/mol]	-32.1		
ΔΔG [kJ/mol]	< 0.1		
probes	MEPY	FORM	IMIA

ΔG [kJ/mol]	-27.3
ΔΔG [kJ/mol]	4.5
probes	BENX

Figure 3.7: Identification of an alternative pocket in the THF riboswitch. In the upper panel, cartoon representation and molecular surface of the center of the most populated RNA cluster found in the SHAMAN simulation initiated from a holo-like conformation (PDB 4lvx). The THF riboswitch presents two binding pockets (dashed circles), one in a three-way junction (HB4 ligand bound between helical domains P2, P3 and P4, right side) and the other in a pseudoknot (HB4 ligand bound in PK region, left side). The experimental ligands in PDB 4lvx are superimposed by aligning the coordinates to the RNA cluster center. Our protocol detected a low free-energy SHAMAP in the middle of the THF riboswitch between helix P2 and P3 (surfaces surrounded by orange circle, colored as defined in Tab. 3.7 and 3.8). In the lower panel, the light grey and light orange tables report the details of the SHAMAPs that identified the two experimental and the alternative binding sites, respectively.

value in the field. By integrating SHAMAN in virtual screening pipelines, we aim in the future at creating an advanced platform for the rational *in silico* design of RNA-targeting molecules, effectively addressing the longstanding challenges in the field.

3.4 Materials and Methods

3.4.1 Details of the SHAMAN algorithm

SHAMAN consists of four main stages, each one composed of a set of operations described in detail in the following sections. At the beginning of each stage, we provide a brief non-technical overview to facilitate the reading.

I - Input stage

The initial input of SHAMAN consists of the 3D structures of the target RNA and of a set of N probes. Starting from this information, we generate a reference *mother* system, including the RNA and possibly structural ions, and N *replicas*, each one with the addition of a different probe.

Setup of the mother Simulation. The 3D structures of all the systems (Tab. 3.1) were obtained from the PDB database [68]. In the case of RNA structures determined by NMR experiments, the first model was selected. In the case of holo structures, the ligand was removed. Furthermore, to correctly model the RNA with our forcefield, the following elements were also eliminated, if

present: crystal waters, PO3 group in the 3' terminal, modified residues at both terminals, and ions not modeled by our forcefield (SO4 in PDB 3tzr, 3ski, and 7kd1). The resulting model was then prepared by adding hydrogen atoms using UCSF Chimera [69] at pH=7.4 and processed by the OpenMM library [70] v. 7.7.0 to generate an initial GROMACS configuration and topology files. The forcefield used for RNA was AMBER99SB-ILDN* [71] with the BSC0 correction on torsional angles [72] and the χ_{OL3} correction on anti-g shifts [73]. Ions were modeled using the Joung and Cheatam parameters [74] with the Villa et al. correction for magnesium [75]. Water molecules were modeled with the OPC force field [76]. Forcefield parameters were obtained from `https://github.com/srnas/ff`.

Setup of the replica simulations. The 3D structures of the probe were generated as described in the section "Details of the Probes." One replica of the mother system was generated for each probe. A single probe conformer was generated using the RDkit python library v. 2022.3 and inserted in a random position and orientation, with a minimum and maximum distance of its center of mass from the RNA atoms equal to 0.2 nm and 1.0 nm, respectively. The force field and topology of the probe were created with OpenFF Sage 2.0 [77].

General details of the MD simulations. Both mother and replica systems were solvated in a triclinic box with dimensions chosen in such a way that each edge of the box was 1.0 nm away from the closest RNA atom. K^+ and Cl^- were added to ensure charge neutrality in the system at a salt concentration of $0.15M$. In all simulations, the equations of motion were integrated by a leap-frog algorithm with a timestep of 2 fs. The smooth particle mesh Ewald method [78] was used to calculate electrostatic interactions with a cutoff of 0.9 nm. Van der Waals interactions were gradually switched off at 0.8 nm and cut off at 0.9 nm. All simulations were performed with GROMACS [79] v. 2021.5 equipped with a development version of PLUMED [80] v. 2.8 (GitHub master branch).

II - Production stage

After independently equilibrating mother and replica systems, the SHAMAN simulation proceeds in parallel. The RNA in the mother simulation is freely evolving and the positions of the RNA backbone atoms are communicated to the replica systems. A restraint is added to the positions of the backbone RNA atoms in the replica systems to make sure that they follow like shadows the conformation sampled by the mother. To accelerate the exploration of the RNA surface, the sampling of the probe in the replica systems is enhanced by metadynamics.

Equilibration procedure. All systems were independently equilibrated before the production stage. This procedure consisted of i) energy minimization with steepest descent; ii) a 10 ns-long equilibration in the NPT ensemble using the Berendsen barostat [81] at 1 atm; iii) a 10 ns-long equilibration in the NVT ensemble using the Bussi-Donadio-Parrinello thermostat [82] at 300K. During the last two steps, harmonic restraints with harmonic constant equal to $400kJ/mol/nm^2$ were applied to the positions of the RNA backbone as well as probe atoms.

SHAMAN simulations. The systems were simulated in parallel for 1 µs each. The following settings were implemented using PLUMED. First, the position of the atoms of the RNA backbone in the mother system were communicated at each MD step to all the replicas with a stride equal to 0.2 ps and the corresponding atoms were restrained to have a maximum RMSD of 0.2 nm from the mother configuration using an upper harmonic wall with intensity equal to 10000 $kJ/mol/nm^2$. Second, to accelerate the probe exploration of the RNA surface, we used metadynamics [50]. As collective variables $\mathbf{S}(\mathbf{R})$, we used the xyz coordinates of the center of mass of the probe, defined after aligning the atoms of the RNA backbone to the initial reference conformation using the FIT_TO_TEMPLATE action in PLUMED. The well-tempered variant of metadynamics [83] was used with bias factor equal to 10. Gaussians with initial height of $1.2kJ/mol$ and width of 0.1 nm were deposited every 1 ps. Finally, we restrained the position of the center of mass of the probe in order to be at most 1.0 nm away from the closest RNA atoms using an upper harmonic wall with intensity equal to 10000 $kJ/mol/nm^2$.

III - Analysis stage

For each representative cluster of RNA conformations explored by SHAMAN, we *i)* identified the regions with high probe occupancy; *ii)* defined a set of potential interacting sites for each probe; *iii)* clustered together the sites found by all probes to create the final SHAMAPs.

Metadynamics reweighting. We removed the effect of the metadynamics bias potential on the probe trajectories by calculating for each frame the unbiasing weight w_t as [84]:

$$w_t \propto \exp\left(\frac{V_G(\mathbf{S}(\mathbf{R_t}), \bar{t})}{k_B T}\right) \tag{3.1}$$

where $V_G(\mathbf{S}(\mathbf{R_t}), \bar{t})$ is the well-tempered metadynamics potential accumulated at the end of the simulation $\bar{t}$ and evaluated on the conformation R_t. All these operations were performed independently for each simulation using the *driver* utility of PLUMED and independently for each simulation.

RNA clustering. We first concatenated all the trajectories of the mother and replica simulations, after removal of the probes, and fixed the discontinuities due to the periodic boundary conditions. Then, we clustered all the RNA conformations using the Gromos algorithm [85] implemented in GROMACS. The clustering employed the RMSD calculated on the RNA backbone atoms with a cutoff of 0.3 nm as the metric. To reduce memory requirements, the clustering was initially performed on a subset of frames (1 every 10), and the excluded frames were subsequently assigned to the closest cluster using a Python script based on the MDAnalysis library [86] (version 2.2.0). For each state, the cluster center was considered as the representative structure. The cluster populations were calculated independently for the mother and each replica simulation, and clusters with populations less than 10% were discarded in the subsequent analysis.

Calculation of probe free energy maps. The following analysis was performed independently for each replica and probe system as well as for each RNA cluster. We first extracted from each trajectory the frames corresponding to the selected cluster and aligned all the conformations to the RNA backbone atoms of the cluster center. We then defined a grid in 3D space with a voxel size of

0.1 nm and computed the corresponding probe binding free energy δG_{ijk} for each voxel (ijk) using the formula:

$$\delta G_{ijk} = -k_B T \log\left(\frac{N_{ijk}}{N_0}\right) \tag{3.2}$$

where $k_B T = 2.494339\,\mathrm{kJ/mol}$ and N_{ijk} is the sum over all probe atoms of the normalized metadynamics unbiasing weights (Eq. (3.1)) of the frames in which the probe atom explored the voxel ijk. N_0 represents the probe occupancy in the bulk solvent and is defined as:

$$N_0 = \frac{n_{\mathrm{probe}} \cdot V_{\mathrm{voxel}}}{V_{\mathrm{MD}}} \tag{3.3}$$

where n_{probe} is the number of probe atoms, V_{voxel} is the volume of the voxels, and V_{MD} is the simulation box's volume. δG_{ijk} quantifies the propensity of finding a probe atom within voxel ijk rather than in the bulk solvent: voxels with a low value of δG_{ijk} represent therefore potential strong binding sites to the RNA molecule. We estimated the associated error σ_G by calculating the standard deviation of δG_{ijk} calculated in the first and second half of the trajectory (Fig. 3.19).

Voxels selection, clustering into interacting sites, and filtering. To exclude weak affinity regions, and independently for each probe, we first selected all the voxels within 10 kJ/mol from the minimum value of δG_{ijk} across all voxels. The selected voxels were then clustered into *interacting sites* using the DBSCAN algorithm implemented in the scikit python library [87] v. 1.8.1, with a maximum distance between points equal to 0.2 nm and a minimum number of samples equal to 5. For each interacting site, we calculated the associated binding free energy ΔG_l as:

$$\Delta G_l = -k_B T \log\left(\sum_{ijk} p_{ijk}\right) \tag{3.4}$$

where $p_{ijk} = \exp\left(-\frac{\delta G_{ijk}}{k_B T}\right)$, and the sum is over all the voxels belonging to the site. For each interacting site, we also defined its center $\mathbf{g}_l$ as the free-energy-weighted average position of the voxel centers $\mathbf{r}_{ijk}$:

$$\mathbf{g}_l = \frac{\sum_{ijk} p_{ijk}\mathbf{r}_{ijk}}{\sum_{ijk} p_{ijk}} \tag{3.5}$$

and a free-energy-weighted radius of gyration R_l as:

$$R_l = \sqrt{\frac{\sum_{ijk} p_{ijk} \cdot d(\mathbf{r}_{ijk}, \mathbf{g}_l)^2}{\sum_{ijk} p_{ijk}}} \tag{3.6}$$

where d is the Euclidean distance. Finally, we calculated the buriedness score x^l_{bur} of an interacting site to quantify its exposure to solvent. For each voxel ijk, we first defined the RNA density N^{RNA}_{ijk} as the sum of the metadynamics unbiasing weights (Eq. 1) of the frames in which an RNA atom explored the voxel ijk. We then defined x^l_{bur} as:

$$x^l_{\mathrm{bur}} = \frac{100}{N_l} \sum_{ijk} N^{\mathrm{RNA}}_{ijk} \tag{3.7}$$

where the sum runs over all the voxels N_l voxels at the surface of the interacting site. Interacting sites with low buriedness score correspond to regions surrounded by few RNA atoms, *i.e.*, exposed to solvent. All the sites with a buriedness score lower than 0.15 were filtered out.

Calculation of the final SHAMAPs. For each representative cluster of RNA conformations, we defined a set of SHAMAPs by clustering together all the interacting sites found by all probes. To perform this operation, we used the DBSCAN algorithm applied to the centers of the interacting sites $\mathbf{g}_l$, with a maximum distance between points given by $2 * \left(\bar{R}_l + \sigma_R\right)$, where $\bar{R}_l$ is the average radius of gyration across all sites, and σ_R is their standard deviation, and a minimum number of samples equal to 1. For each SHAMAP, we defined the binding free energy ΔG_S as the minimum free energy over all the interacting sites that clustered into this SHAMAP:

$$\Delta G_S = \min_{l \in S} \left(\Delta G_l\right) \tag{3.8}$$

And $\Delta\Delta G_S$ is the difference between the binding free energy of a SHAMAP and the minimum value across all SHAMAPs (*top scored*):

$$\Delta\Delta G_S = \Delta G_S - \min_{S} \left(\Delta G_S\right) \tag{3.9}$$

IV - Output stage

The SHAMAPs obtained at the end of the previous analysis stage constituted the final set of hotspots associated with a given conformational state of the RNA target. The SHAMAPs are reported in a table and ordered by ΔG_S. Along with this information, each SHAMAP is annotated with the properties of its constituent interacting sites: a list of probes that explored the site, their corresponding ΔG_l, the population of the RNA cluster in which the site has been visited, the coordinates of the centers $\mathbf{g}_l$, and the radius of gyration R_l.

3.4.2 Details of the SHAMAN benchmark

Details of the target RNAs. For our SHAMAN simulations, we selected 7 RNA systems, whose structures in complex with at least one ligand were deposited in the PDB databank [68] (Tab. 3.1). To initiate our SHAMAN simulations, we selected 1 holo structure per system and, when available, an apo structure of the same RNA molecule. In total we performed 12 SHAMAN simulations. A summary of all simulations performed along with details about the systems are reported in Tab. 3.2.

Details of the PDB structures used for validation. To benchmark the accuracy of our approach, we first retrieved for each system all the holo structures deposited in the PDB with different ligands and binding poses. We then visually inspected each structure and identified 14 structures with unique binding poses and pockets. All the structures used for validation along with details about the RNA, the ligand, and the experimental method and resolution are reported in Tab. 3.3 and Tab. 3.4.

Details of the probes. The set of probes used in our protocol is composed of two subsets. First, we included 8 probes already used in the SILCS-RNA study [31], namely acetate (ACEY), benzene (BENX), dimethyl-ether (DMEE), formamide (FORM), imidazole (IMIA), methyl-ammonium (MAMY), methanol (MEOH), and propane (PRPX) (Tab. 3.7). These fragments had been selected in the original study as a representative set of functional groups. Second, we developed the following approach to identify fragments with higher probability to bind to RNA molecules. Two databases were used, namely HARIBOSS [22] comprising 265 experimentally validated RNA binders (`https://hariboss.pasteur.cloud`) and RBIND [26] that includes 159 RNA bioactive molecules (`https://rbind.chem.duke.edu`). In an effort to identify chemical groups that exist in both libraries, we prepared the Murcko scaffolds from the molecules derived from both databases and compared the corresponding sets. 6 Murcko scaffolds appear in both HARIBOSS and RBIND molecules (Tab. 3.8). From these, 5 representative scaffolds were selected for the SHAMAN simulations, namely benzene (BENX), dihydro-pyrido-pyrimidinoneimidazo-pyridine (BENF), benzothiophene (BETH), methyl-pyrimidine (MEPY), and piperazine (PIRZ). The preparation and comparison of the HARIBOSS and RBIND libraries was done using a KNIME 4.6 protocol that includes the following steps: i) molecule preparation using Epik [88] at pH 7.4, ii) conversion to canonical SMILES using RDkit v. 2022.3, iii) Murcko scaffold derivation using the RDkit Murcko Scaffolds KNIME node, iv) set comparison using the 'Compare Ligand Sets' node provided by Schrödinger v. 2022.3, and finally v) a fragmentation of the common scaffolds using the RECAP fragmentation method [89] (implemented as the 'Fragments from Molecules' node provided by Schrödinger). All probes used in the SHAMAN simulations have been prepared using the LigPrep module of Schrödinger Suite [90] at pH 7.4. BETH was intentionally modeled in a protonated state, as it appears in the origin molecules from RBind and HARIBOSS.

Details of the validation procedure. To benchmark the accuracy of our approach in identifying binding sites occupied by a ligand in known experimental structures, we used the following procedure:

i. **Multiple sequence alignment:** for each simulated system, we aligned the sequence of our target RNA with the sequences of all the validation PDBs using CLUSTALW [91] v. 2.0.

ii. **Structural alignment of validation PDBs to SHAMAN cluster centers:** for each validation PDB, we defined the binding site as the set of nucleotides with at least one atom within 0.6 nm of a ligand atom. The nucleic backbone atoms of the validation PDB belonging to this region were then structurally aligned to the corresponding nucleotides in each RNA cluster center, based on the sequence alignment defined above.

iii. **Definition of success for a probe interacting site:** for each validation PDB, we defined an *experimental sphere* centered on the center of mass of the heavy atoms of the ligand $\mathbf{g}_{\text{exp}}$ and with a radius given by its radius of gyration R_{exp}. For each probe interacting site, we defined a *validation sphere* centered on the free-energy weighted center of the interacting site $\mathbf{g}_l$ (Eq. (3.5)) and with a radius given by its free-energy weighted radius of gyration R_l (Eq. (3.6)). We then considered a probe interacting site as successful if the *validation sphere* was overlapping with the *experimental sphere*:

$$d(\mathbf{g}_l, \mathbf{g}_{\text{exp}}) \leq R_l + R_{\text{exp}} \tag{3.10}$$

In case of a match with multiple validation structures, we retained only the one corresponding to the interacting site with a lower free-energy gap $\Delta\Delta G$ (Eq. (3.9)) from the top-scored SHAMAP.

iv. **Definition of success for a SHAMAP:** a SHAMAP was considered successful in identifying a known ligand binding site if at least one of the probe interacting sites that compose the SHAMAP was successful according to the criterion defined above.

3.4.3 Probes-ligands comparison

For probes and ligands in the SHAMAN simulations initiated from holo structures, we first calculated the following set of descriptors with RDKit v. 2022.3: molecular weight, number of aromatic rings, number of H-bond donors/acceptors, number of H-bond acceptors, topological polar surface area (TPSA), and number of heterocycles. The correlation between probes and ligands descriptors for probes and ligands was computed with scipy v. 1.8.1 using the Pearson correlation coefficient. The analysis was performed using either the entire ligand or its Murcko scaffold. We also quantified the similarity between ligands and successful probes using different types of fingerprints (FPs) as implemented in RDKit. In particular, we used Morgan (radius = 2, 2048 bits), RDKit (2048 bits), and MACCS FPs. Using these FPs and the Tanimoto distance, we calculated the similarity between successful probes and reference ligands, considered either as entire ligands or by using their corresponding Murcko scaffold.

To further investigate a possible correlation between ligands and successful probes, we formulated the following hypothesis: the ability of a probe to identify a pocket binding site is related to its similarity to the corresponding ligand. We then compared each of the 13 probes (Tab. 3.7 and 3.8) with all the 8 ligands resolved in the experimental pockets (Tab. 3.1) and considered a probe to be similar (dissimilar) to a ligand if the Tanimoto distance calculated with MACCS FP was greater (lower) than 0.4 (0.2). Based on the SHAMAN results in our benchmark, we built a confusion matrix of the four possible outcomes (Tab. 3.10) and defined the SHAMAN negative predictive value (NPV) as the ratio between true negatives (TN) and the total number of negatives (TN+FN):

$$NPV = \frac{TN}{TN + FN} \tag{3.11}$$

3.4.4 Comparison with other tools

We selected three state-of-the-art tools for RNA binding site detection: SiteMap [51], BiteNet [52], and RBinds [54]. We evaluated the ability of these tools to predict the RNA nucleotides that belong to an experimentally detected binding site in the 7 systems of our benchmark set, including holo-like and apo structures, for a total of 12 conformations (Tab. 3.1).

Definition of the ground truth. For each system, the reference set of binding site nucleotides was defined as follows:

1. We performed a multiple sequence alignment of all the systems in our validation set (Tab. 3.3 and 3.4) using CLUSTALW [91] v. 2.0;
2. We discarded all the nucleotides that were not resolved in all the validating structures;
3. In each validating structure, we defined as interacting with the small molecule all the nucleotides with at least one atom within 4 Å of an atom of the ligand;
4. To compare the predictions against all the validating structures (Fig. 3.4), we defined as interacting nucleotides the union of all the interacting nucleotides across all the validating structures.

Prediction of interacting nucleotides. For each software, the input was the same PDB file that was used as the starting structure for our SHAMAN simulations (Details of the SHAMAN algorithm, II. Production stage). The set of predicted interacting nucleotides was defined as follows:

- **SHAMAN:** Each interacting site predicted by SHAMAN is stored in a file as the set of coordinates of the centers of the grid voxels (Details of the SHAMAN algorithm, III. Analysis stage). We defined as interacting all the nucleotides found in the RNA cluster center with at least one atom closer than 4 from the coordinates of all the interacting sites belonging to the SHAMAPs that identified the experimental pockets considered for validation (Tab. 3.5 and 3.6).
- **SiteMap:** For each structure, a local installation of SiteMap (v. 2023-4) was run from the command line with the options: *-keepvolpts* and *-modbalance yes*. The output was a PDB-like file containing the coordinates of the predicted binding sites. Among the predicted binding sites, we visually selected the one that best overlapped with the position of the experimentally resolved ligand. Finally, we defined as interacting all the nucleotides with at least one atom within 4 of the pseudo-atoms defined in the output PDB file.
- **BiteNet:** For each structure, BiteNet was executed using a standalone version of the software. The input parameter "input probability score threshold" was set at its default value of 0.1, and the "RNA-small molecule binding site" option was selected. The binary classification of interacting/non-interacting nucleotides was defined in the output file "*predictions.csv*".
- **RBinds:** For each structure, RBinds was executed via the webserver available at `http://zhaoserver.com.cn/RBinds/RBinds.html`. The list of predicted interacting nucleotides was defined in the "sites" card in the output file "*RNAcentrality.json*".

Comparison metrics. The quality of the prediction of interacting nucleotides was defined based on the following metrics for binary classifiers:

- The Matthew Correlation Coefficient (MCC), which is a global measure of prediction quality recognized for its comprehensiveness and reliability compared to other standard metrics [92]. The MCC score accounts for the quality in all the four classes of the confusion matrix:

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}} \tag{3.12}$$

- The accuracy, which is the fraction of correct (positive and negative) predictions:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{3.13}$$

- The precision, which is the fraction of relevant instances among the retrieved instances:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3.14}$$

- The recall (or sensitivity), which is the fraction of relevant instances that were retrieved:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3.15}$$

3.4.5 Software and data availability

SHAMAN simulations can be run with the development version (GitHub master branch) of PLUMED (`https://github.com/plumed/plumed.github.io`). The GROMACS topology files and PLUMED input files used in our benchmark are available on PLUMED-NEST (`www.plumed-nest.org`), the public repository of the PLUMED consortium [93] as plumID:23.031. Scripts to facilitate the preparation of the input files and the analysis of the results, as well as a complete tutorial (Appendix A), will be released soon under a license "free for academics, not for commercial use".

3.5 Supplementary Information

3.5.1 Supplementary analysis

TPP riboswitch

The thiamine pyrophosphate (TPP) riboswitch is a highly conserved riboswitch found in archaea, bacteria, and eukaryotes, which directly modulates gene expression through a variety of mechanisms [94]. Upon binding to its cognate partner, the TPP, in the core region of two multi-way junctions between four helices (Fig. 3.8A), the TPP riboswitch assumes a stable three-dimensional structure [95]. Fragment-based screening experiments revealed the binding of several fragments in the same region in which TPP binds, but with an unexpected conformational change of residue G72, a key element in the recognition of the pyrophosphate [96] (Fig. 3.8B). These results indicate that alternative conformations of the TPP riboswitches may be targeted in drug design efforts, making this molecule an interesting case study for our pipeline.

We tested SHAMAN starting from a holo-like conformation of the TPP riboswitch (PDB 3d2v [97]). During this simulation, the TPP riboswitch populated a single structural cluster. The top-scored SHAMAN identified in this state (Tab. 3.3) corresponds to the region of the TPP binding site (left panel, Fig. 3.8C). The geometric accuracy in identifying the binding site resolved in the experimental structure is also impressive: the best match was obtained with ligand HPA (PDB 4nyd [96]), with a validation distance (Eq. (3.10)) of 0.64 Å, corresponding to the best overlap of our benchmark (Tab. 3.5). Remarkably, our probe methyl-pyrimidine (MEPY) is perfectly overlapping with the aromatic rings of the fragment used in the aforementioned experimental screening (right panel, Fig. 3.8C). Importantly, a second SHAMAN with very good scoring ($\Delta\Delta G < 0.1\ kJ/mol$), identified mostly by formamide (FORM) and acetate (ACEY) probes (Fig. 3.8D), corresponds to the region that interacts with the tail of the TPP ligand in the cognate bound state (Fig. 3.8B). This case study is important since, while our SHAMAN simulation captured the conformation of the TPP-bound state (red sticks in Fig. 3.8D), it was also able to identify the alternative binding modes by allowing for the rearrangement of the involved residues.

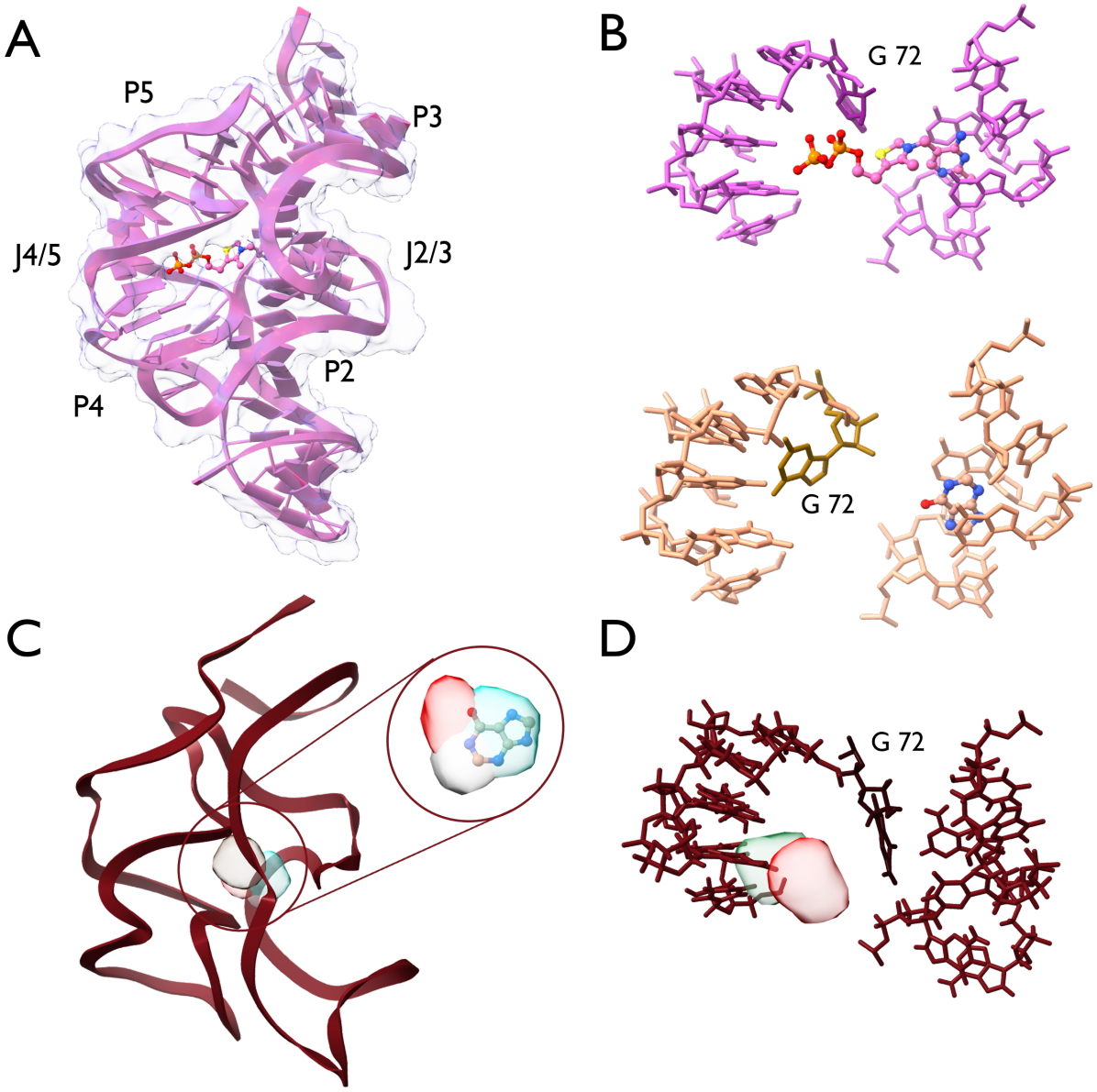


Figure 3.8: The case of the TPP riboswitch. A) Cartoon-surface representation of the TPP riboswitch in the bound conformation, as in PDB 2hoj [98]. P2-P5 indicate the helical regions, while J2/3 and J4/5 the multi-way junctions. **B)** In the upper panel, the TPP binding mode, as in PDB 2hoj [98]. In the lower panel, the binding mode of HPA, one of the fragments screened in the work related to PDB 4nyd [96]. **C)** On the left, a cartoon representation of the main conformation explored during SHAMAN simulations started from the holo-like state of the TPP riboswitch. The SHAMAPs identifying the experimental binding site are visualized as solid surfaces with the color code defined in Tab. 3.7 and 3.8. In the inset, the superposition of the mentioned SHAMAPs with ligand HPA, represented in sticks with CPK standard colors. Hydrogen atoms are visualized only when resolved in the experimental structure. **D)** The position of the interacting residues defined in B in the main conformation of TPP riboswitch explored during SHAMAN simulations. In panel B and D, the key residue G72 is highlighted by a darker color. The RNA nucleotides and ligands are represented using licorice and CPK styles, respectively.

THF riboswitch

The tetrahydrofolate (THF) riboswitch, primarily found in bacteria, regulates the expression of genes involved in the synthesis and transport of the THF vitamin, which is essential for bacterial metabolism [99]. The THF riboswitch in its functional state forms two characteristic binding hotspots (Fig. 3.9A). Interestingly, the two pockets are both key factors of the riboswitch regulatory function and their formation and stability are interconnected [100]. In the unbound conformation of the THF riboswitch, the absence of ligands causes the unwinding of the pseudoknot PK and the misalignment of P1 and P3 helices [101] (Fig. 3.9B). The intrinsic structural dynamics of riboswitches and the dual-ligand binding capability of the THF riboswitch makes it an interesting case study for our pipeline.

We tested SHAMAN starting from two structures of the THF riboswitch, one in holo-like (PDB 4lvx [100]) and one in apo (PDB 7kd1 [101]) conformation. In both cases SHAMAN was able to identify both binding sites among the most probable binding pockets (left panels of Fig. 3.9CD, Tab. 3.3). Interestingly, in both apo and holo-like cases the two pockets are identified within the same RNA conformation. The geometric accuracy is very high for all the identified pockets, with an average validation distance (Eq. (3.10)) of 1.92 Å for both holo-like and apo cases (Tab. 3.5). The latter case is particularly remarkable since the binding helical regions in the starting apo structure are not coaxially aligned. By focusing on the P1 and P3 helices, their relative conformation explored during SHAMAN simulations (Fig. 3.9E, blue ribbons) resembles that present in the bound state (Fig. 3.9E, golden ribbons) significantly more than in the starting structure (Fig. 3.9E, cyan ribbons). This result exemplifies the main strength of our approach, which can capture binding pockets formed upon major conformational rearrangements.

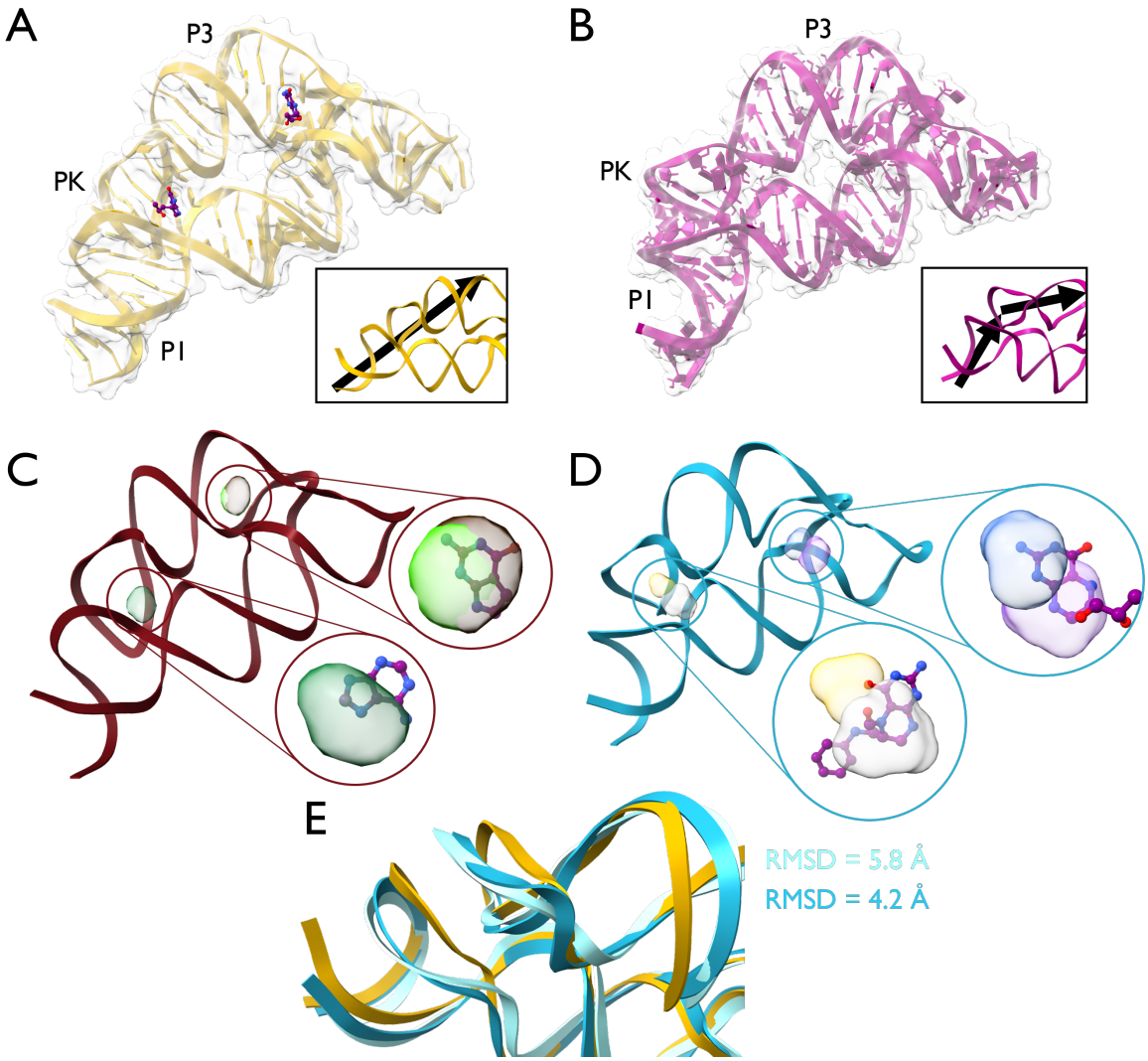


Figure 3.9: The case of the THF riboswitch. A-B) Cartoon-surface representation of the THF riboswitch in its bound (**A**, PDB 4lvx [100]) and unbound (**B**, PDB 7kd1 [101]) states. P1 and P3 denote the helical regions while PK denotes the pseudoknot in the molecule. The experimental ligands of the THF riboswitch are visualized in CPK style. In the inset, the relative orientation of the pseudoknot (PK) and the adjacent helical domains is highlighted by arrows. **C-D)** In left panels, cartoon representation of the main conformation explored during our SHAMAN simulations started from the holo-like (**C**) and apo (**D**) states. The SHAMAPs identifying experimental binding sites in our validation set (Tab. 3.3) are visualized as solid surfaces with the color code defined in Tab. 3.7 and 3.8. In the insets, the SHAMAPs are overlapped with the experimental ligands: **C)** from left to right, ADE (PDB 4lw0 [100]) and 7DG (PDB 4lvw [100]); **D)** from left to right, FFO (PDB 3sd3 [99]) and H4B (PDB 4lvx [100]). Hydrogen atoms are visualized only when resolved in the experimental structure. **E)** Superposition of the holo (yellow ribbon), initial (cyan ribbons), and most populated (blue ribbon) structures found in the apo simulation, with a focus on the pseudoknot region.

dG riboswitch

The deoxyguanosine (dG) riboswitch is an RNA molecule found in bacteria and is involved in the regulation of metabolism by modulating their gene expression [102]. The binding site of the dG riboswitch is deeply buried in the core of a three-way junction between P1, P2 and P3 helical domains (Fig. 3.10A). Within this hydrophobic region, the dG riboswitch is able to recognize and bind the deoxyguanosine nucleoside via stable and specific base-pairing interactions [102] (Fig. 3.10B). Several studies have revealed that the dG riboswitch is also able to undergo conformational changes and to form more solvent-exposed pockets[103]. In such pockets, other ligands may bind and, due to the absence of sugar moieties, disrupt the cognate hydrogen bonding patterns (Fig. 3.10C). In our validation set (Sec. 3.4.2, Tab. 3.3), these two different binding modes were considered as distinct pockets. This ability of recognizing both cognate and non-cognate ligands through conformational rearrangements makes the dG riboswitch an interesting target for therapeutic approaches and an important case study for our method.

We tested SHAMAN starting from a holo-like conformation of the dG riboswitch after removal of its cognate ligand (PDB 3ski [104]). Interestingly, this riboswitch showed a relatively high flexibility, populating 5 distinct conformations in the timescale of our simulation. The two different experimental pockets were identified in two different RNA clusters, in both cases accurately characterizing the buried region of interaction (Fig. 3.10D). The pocket of the cognate ligand was identified by a top-scored SHAMAP ($\Delta\Delta G < 0.1$ kJ/mol, Tab. 3.5), which is 1.6 Å away from the position of the experimental ligand and constituted solely by the BENF probe (Fig. 3.10E). Such result is remarkable since BENF presents similar chemical characteristics with respect to the cognate binding partners (Fig. 3.3). The alternative binding mode was identified by a SHAMAP with lower score ($\Delta\Delta G = 3.3$ kJ/mol), but with good geometric accuracy (3.3 Å, Fig. 3.10F). While our starting structure was derived from our cognate ligand bound conformation, SHAMAN simulations explored a conformation (red sticks in Fig. 3.10G) locally more similar to the alternative binding mode (cyan sticks in Fig. 3.10G). This important result demonstrates the power of our pipeline in identifying binding pockets after conformational changes of the RNA target.

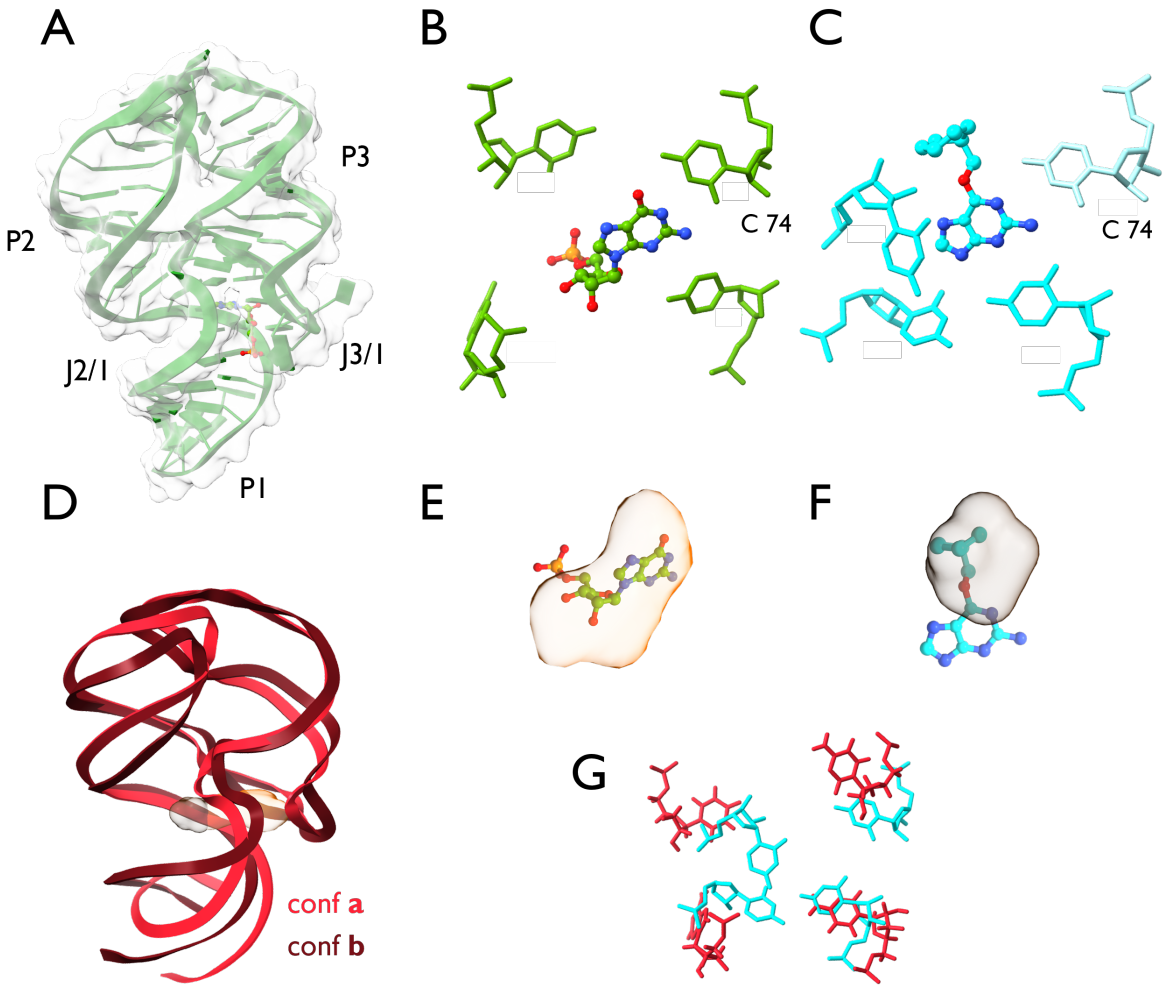


Figure 3.10: The case of the dG riboswitch. A) Cartoon-surface representation of the dG riboswitch in the bound conformation, as in PDB 3slw [104]. P1-3 indicate the helical regions, while J2/1 and J3/1 the multiway junctions. The 5GP ligand is visualized in CPK style. **BC)** The position of the key residues of interaction in the cognate (**B**, PDB 3slq [104] with ligand 5GP) and alternative (**C**, PDB 6uc9 [103] with ligand CMG10) bound states of the dG riboswitch. **D.** Cartoon representation of the two main conformations (conf a and b) explored by SHAMAN started from the holo-like state of the dG riboswitch. The SHAMAPs identifying the experimental binding site are visualized as solid surfaces with the color code defined in Tab. 3.7 and 3.8. **EF)** Superposition of the SHAMAPs with the 5GP (**E**) and CMG (**F**) ligands in the corresponding experimental structures. Hydrogen atoms are visualized only when resolved in the experimental structure. **G)** Superposition of the key interacting residues in conf a (red sticks) with the alternative bound structure of the dG riboswitch, after alignment of the binding regions (Sec. 3.4.2).

HCV IRES IIa RNA

The Hepatitis C Virus (HCV) Internal Ribosome Entry Site (IRES) IIa RNA is a crucial element of the HCV viral genome, as it facilitates the initiation of HCV proteins synthesis. Its functioning mechanism is peculiar: while the activity of most viral RNAs depends on the presence of translation initiation factors, HCV IRES IIa directly interacts with the host ribosomal machinery [105]. The recruitment of the ribosomal subunit to the HCV RNA is driven by its ordered folding in a L-shaped bent conformation, stabilized by divalent metal ions [106] (Fig. 3.11A). Stabilizing alternative conformations of the HCV RNA might therefore alter its capability to recognize the host ribosome and disrupt the viral replication, making this molecule an interesting therapeutics target. Several small molecules have been identified to bind HCV RNA [107] and to induce conformational changes that decrease the affinity to the ribosome (Fig. 3.11B). Among the available structures of HCV RNA bound to ligands, we identified two distinct binding sites located in the same region within the central groove of the RNA helix, but with different binding modes (Fig. 3.11CD, Tab. 3.4). The characteristic structural dynamics of the HCV RNA makes this molecule an important case study for our approach.

We tested SHAMAN starting from two conformations of the HCV IRES IIa RNA, one holo-like (PDB 3tzr [106]) and one apo (PDB 2nok [106]). In both simulations, the RNA molecule showed a higher stability compared to other viral systems studied, such as the HIV-1 TAR. This is suggested by the exploration of only two and one relevant structural clusters in the holo-like and apo simulations, respectively (left panels in Fig. 3.11EF). This stability may be attributed to the presence, in both apo and holo starting structures of HCV IRES IIa, of divalent metal ions simulated in the mother replica of SHAMAN runs (Sec. 3.4.1). In the holo-like case, SHAMAN was able to identify both experimental binding sites (right panel, Fig. 3.11E) as the most probable ones ($\Delta\Delta G = 0$ and $\Delta\Delta G = 0.2$ kJ/mol, respectively) and with high geometric accuracy (2.67 and 2.31 Å, respectively). In the latter case, despite the molecule populated most of the time a bent conformation close to the apo state, SHAMAN was able to detect both experimental binding sites (right panel, Fig. 3.11F) among the most probable ones ($\Delta\Delta G = 2.9$ and $\Delta\Delta G < 0.1$ kJ/mol, respectively), and with a slightly lower accuracy (2.84 and 3.20 Å, respectively). This case study demonstrates the ability of our approach to correctly identify the interacting hotspots independently of the starting structure of the target.

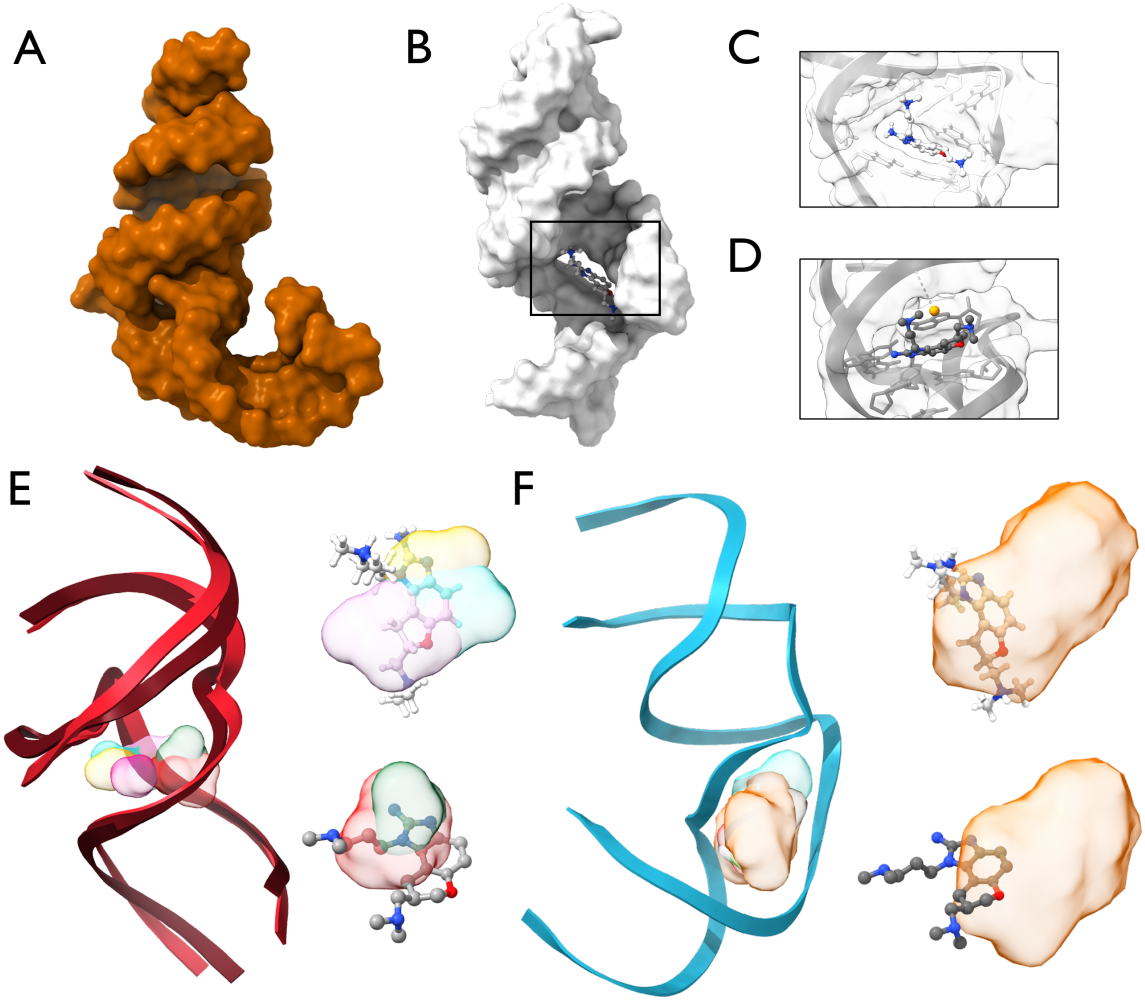


Figure 3.11: The case of the HCV IIa IRES RNA. AB) Surface representation of the HCV RNA in its apo (**A**, PDB 2nok [106]) and holo (**B**, PDB 2ku0 [107]) conformations. The ISI ligand of the IAV promoter is visualized in CPK style. **CD)** Focus on the binding modes of HCV RNA bound to the ISI (PDB 2ku0 [107], **C**) and SS0 (PDB 3tzr [108], **D**) ligands. Resolved magnesium ions are represented by orange spheres. **EF)** In left panels, cartoon representation of the main conformations explored during our SHAMAN simulations started from the holo-like (**E**) and apo (**F**) states of the HCV IRES IIa RNA. The SHAMAPs identifying experimental binding sites in our validation set (Tab. 3.4) are visualized as solid surfaces with the color code defined in Tab. 3.7 and 3.8. In the insets, the SHAMAPs are overlapped with the experimental ligands: **C)** from top to bottom, SS0 (PDB 3tzr [106]) and ISI (PDB 2ku0 [107]); **D)** from top to bottom SS0 (PDB 3tzr [108] and ISH (PDB 2ktz [107]). Hydrogen atoms are visualized only when are resolved in the experimental structure.

IAV promoter RNA

The Influenza A Virus (IAV) promoter, positioned in the 5' untranslated region (5' UTR) of the viral genome, is a crucial player in the virus replication cycle [109]. The IAV promoter exhibits a distinctive partial duplex structure, characterized by a panhandle-like shape of the two bent helical domains (Fig. 3.12A). This conformation is selectively recognized by the viral polymerase, which then initiates viral transcription. In its sole available bound structure, the IAV promoter has been resolved in complex with the 0EC small molecule [110]. The latter stabilizes an alternative conformation of IAV promoter, characterized by the widening of the major RNA groove and the coaxial alignment of the angle between the helical regions (Fig. 3.12B). Such conformation has a lower affinity to bind the IAV polymerase enzyme and causes the inhibition of the viral activity. The challenging nature of IAV promoter and the relation between its dynamics and functions make it an important case study for our method.

We tested SHAMAN starting from two structures of the IAV promoter, one in holo-like (PDB 2lwk [110]) and one in apo (PDB 1mfy [111]) conformation. Our simulations indicated that this RNA molecule is highly flexible and populated three and seven different structural clusters in the holo-like (Fig. 3.12C) and apo case (Fig. 3.12D), respectively. The region that exhibited the highest flexibility is the lower domain of the IAV promoter: the bending of the angle between helical regions varied significantly among all the explored conformations (Fig. 3.12D). In the holo-like simulation, the experimental pocket was identified in the most populated cluster among the top-scored SHAMAPs ($\Delta\Delta G = 0.2$ kJ/mol, left panel of Fig. 3.12E). Oppositely, during the apo simulation, the experimental pocket was identified in a scarcely populated cluster and with lower score ($\Delta\Delta G = 5.9$ kJ/mol, left panel Fig. 3.12F). The geometric accuracy was remarkable in both cases: the experimental pocket was identified with a validation distance (Eq. (3.10)) of 1.43 and 3.64 Å, respectively (top right panel, Fig. 3.12EF). The lower accuracy of the apo simulation results can be explained by the structural dynamics of the helical domains. While in the holo simulation the two helical regions of the IAV promoter form a coaxial alignment characteristic of the bound state (lower right panel, Fig. 3.12E), in the apo simulation the interhelical angle is bent and more similar to the unbound state of IAV promoter (lower right panel, Fig. 3.12F). Overall, SHAMAN results for IAV promoter highlight its ability to correctly identify binding hotspots in unfavorable conditions where the target molecule is stacked in a conformation that resembles the unbound state.

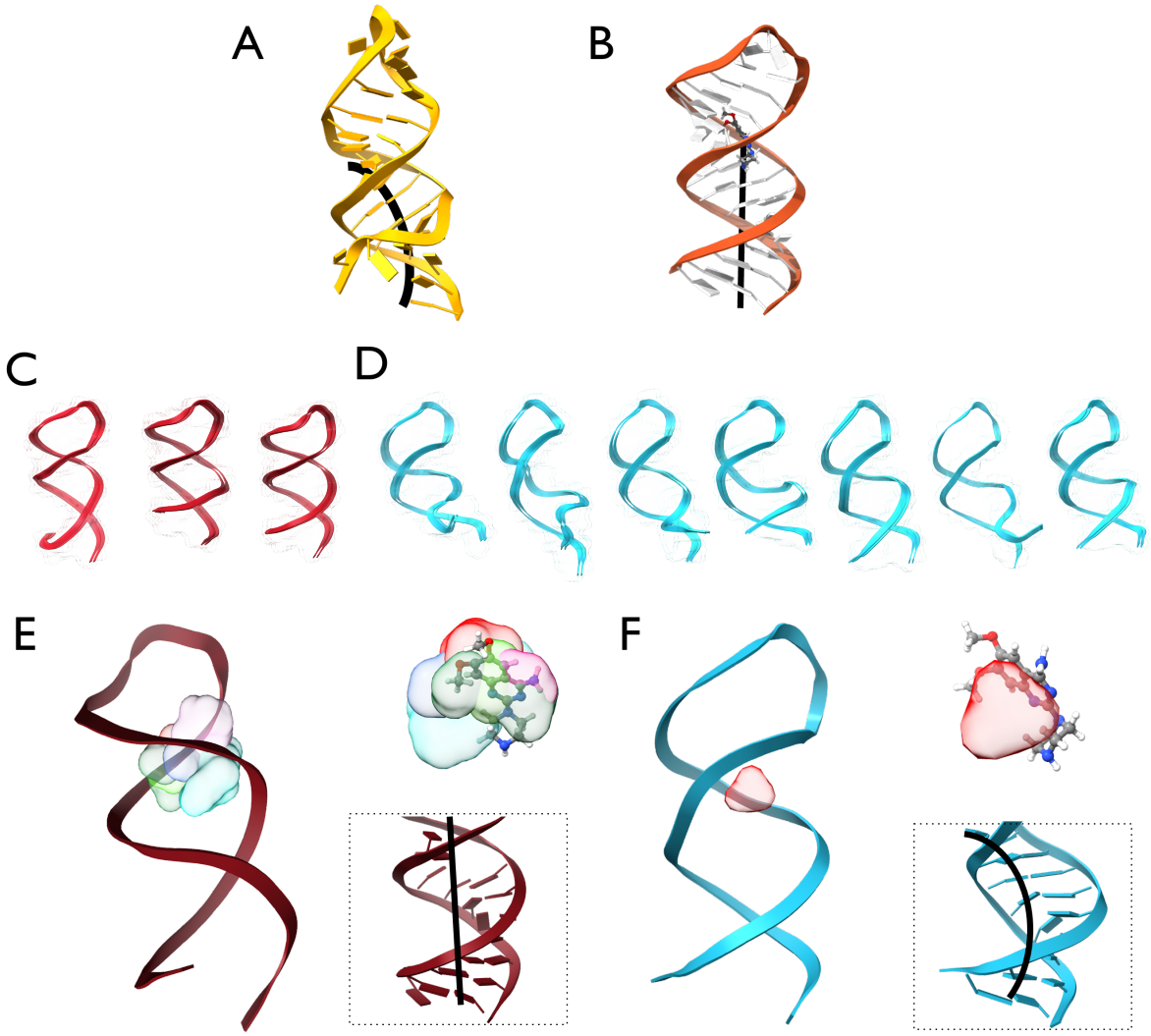


Figure 3.12: The case of the IAV promoter. AB) Cartoon representation of the HCV RNA in its apo (**A**, PDB 1mfy [111]) and bound (**B**, PDB 2lwk [110]) states. Black arrows highlight the alignment of the helical domains. The 0EC ligand of the IAV promoter is reported with CPK style. **CD)** The conformations explored during the holo-like (**C**) and apo (**D**) SHAMAN simulations of the IAV promoter. **EF)** In the left panels, a cartoon representation of the conformation in which the experimental binding site was identified in the holo-like (**E**) and apo (**F**) simulations. The SHAMAPs are visualized as solid surfaces with the color code defined in Tab. 3.7 and 3.8. In the top right panel, the corresponding SHAMAPs are superimposed to the 0EC ligand. Hydrogen atoms are visualized only when resolved in the experimental structure. In the low right panel, the bending of the interhelical regions of the IAV promoter is highlighted by black arrows.

3.5.2 Supplementary figures

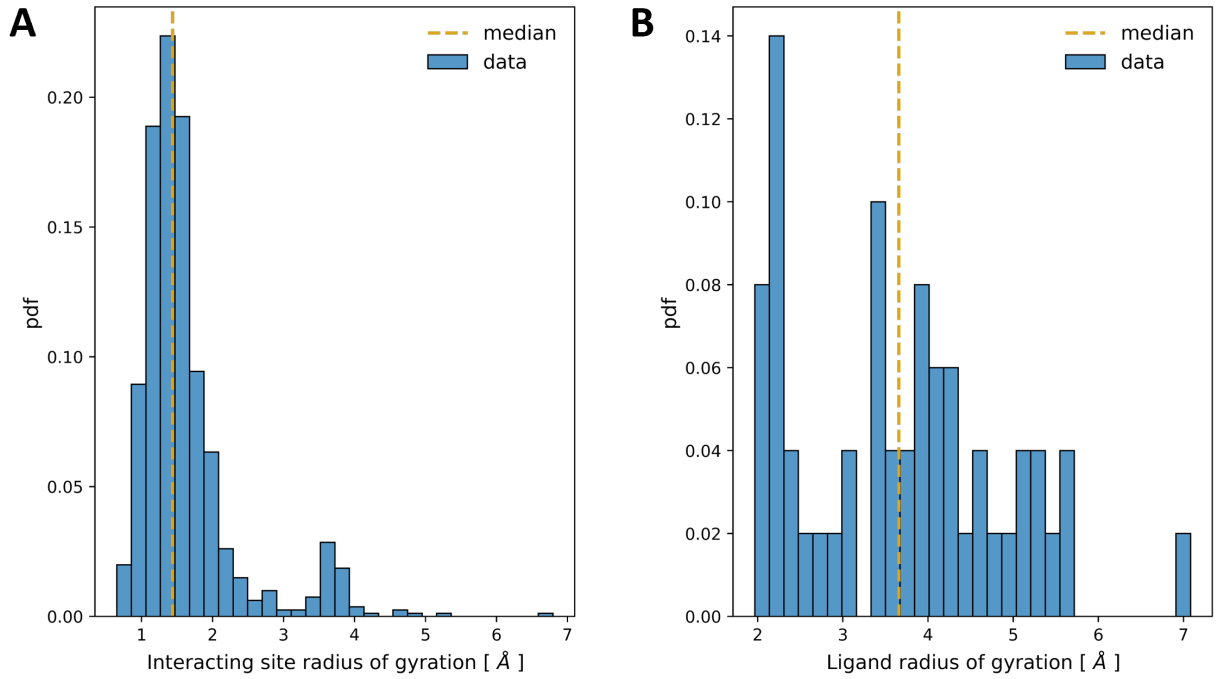


Figure 3.13: Radius of gyration of SHAMAN interacting sites and ligands. A) Normalized distribution of the radius of gyration of all the interacting sites detected by SHAMAN in all the systems of our benchmark set (Tab. 3.1). **B)** Normalized distribution of the radius of gyration of all the unique ligands present in the structures of our validation set (Tab. 3.3 and 3.4). In both plots, the median of the distribution is reported as a yellow dashed line.

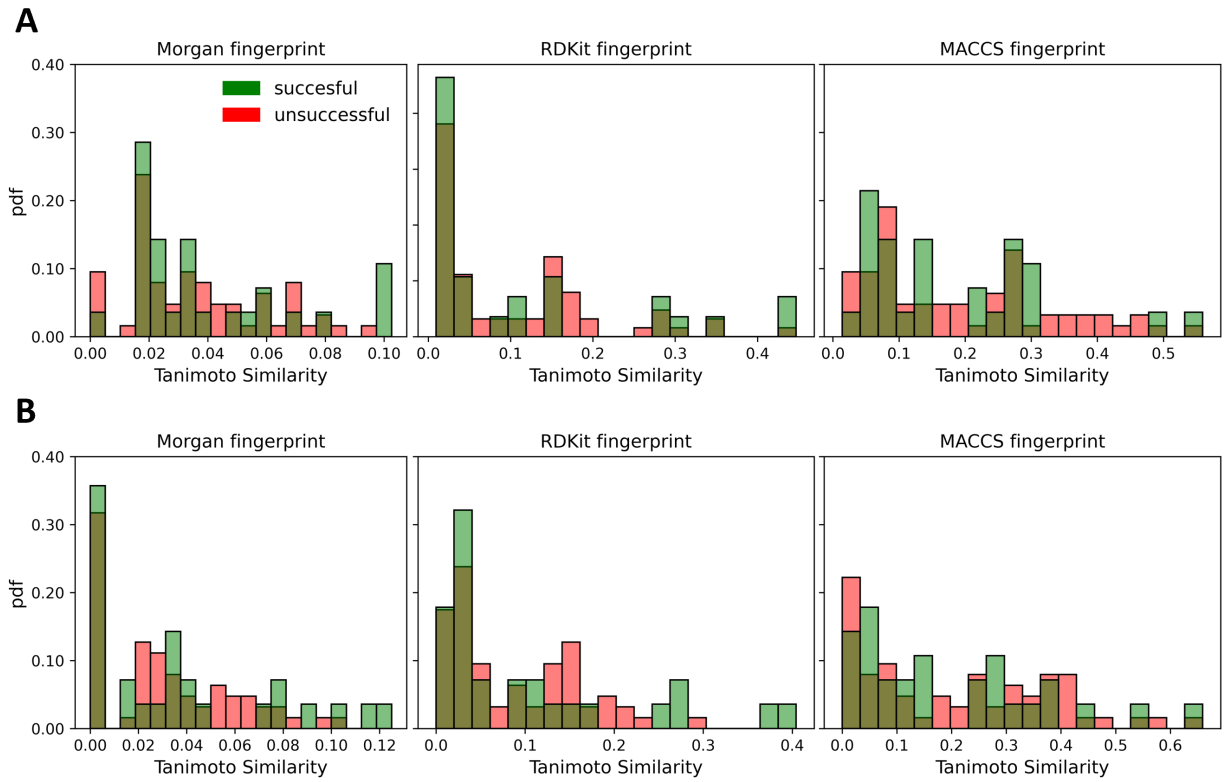


Figure 3.14: Analysis of the similarity between ligands and probes. A) Distributions of the Tanimoto similarity between the ligands present in the experimental structures of our benchmark set (Tab. 3.1) and the probes. Successful probes that identified the ligand are colored in green, and unsuccessful probes in red. The analysis is limited to the SHAMAN simulations initiated from holo-like structures. From left to right, the analysis is performed with the 3 different fingerprints: Morgan, RDKit, MACCS. **B)** As in panel A, with similarities calculated using the Murcko scaffold instead of the entire ligand.

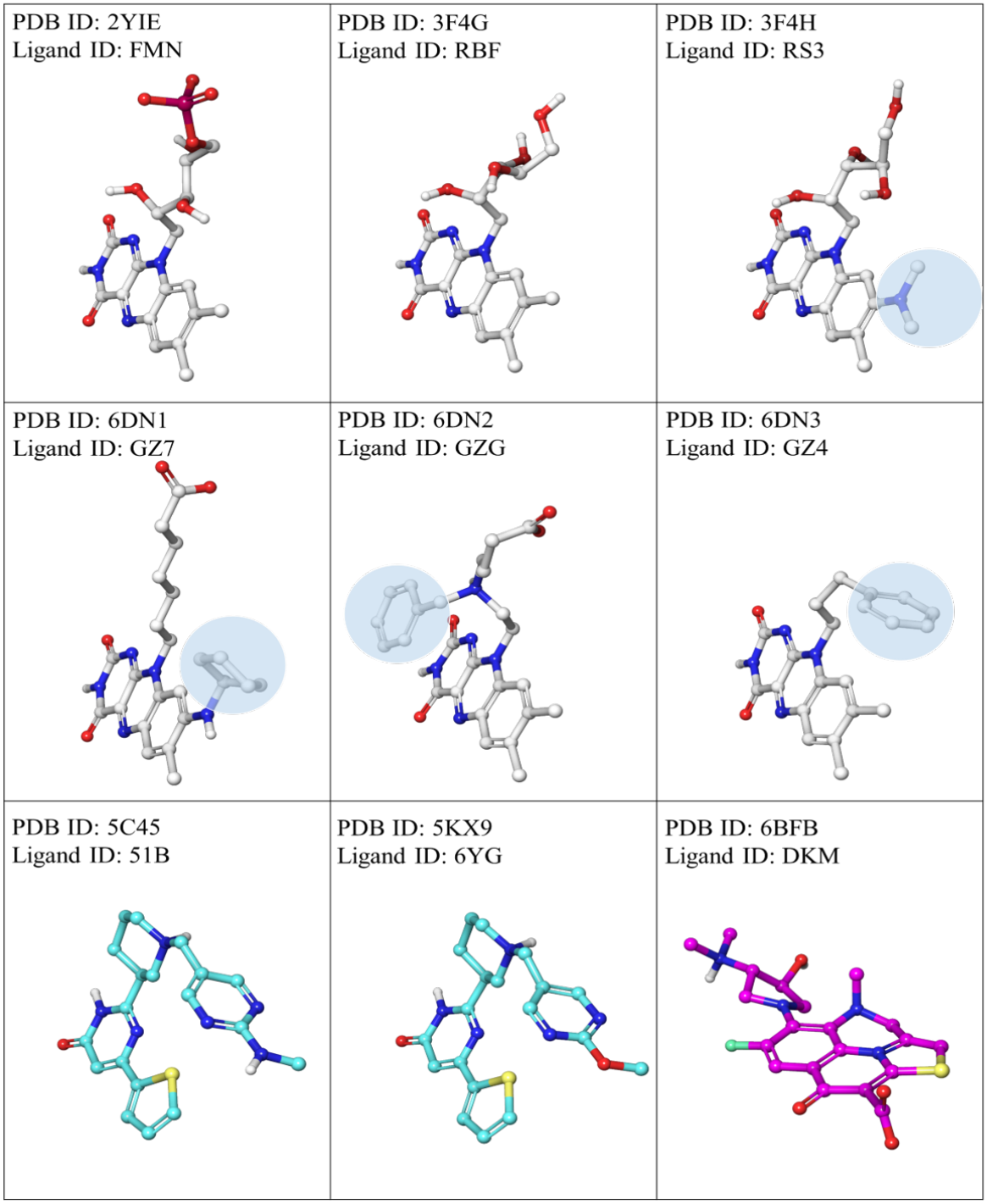


Figure 3.15: Conformers of the FMN riboswitch binders and chemical families. The 3D conformers of the 9 unique ligands of the validation set of the FMN riboswitch are shown in each panel (Tab. 3.3). Following previous studies [56], we subdivided the FMN binders into 3 chemical families: the FMN, ribocil, and DKM families. The carbon atoms of the 3 families are reported in light grey, cyan, and violet, respectively. The remaining atoms are reported with standard CPK colors. Significantly different chemical decorations among the members of the FMN family are highlighted by cyan circles. All ligands are aligned to facilitate the comparison between their 3D conformations.

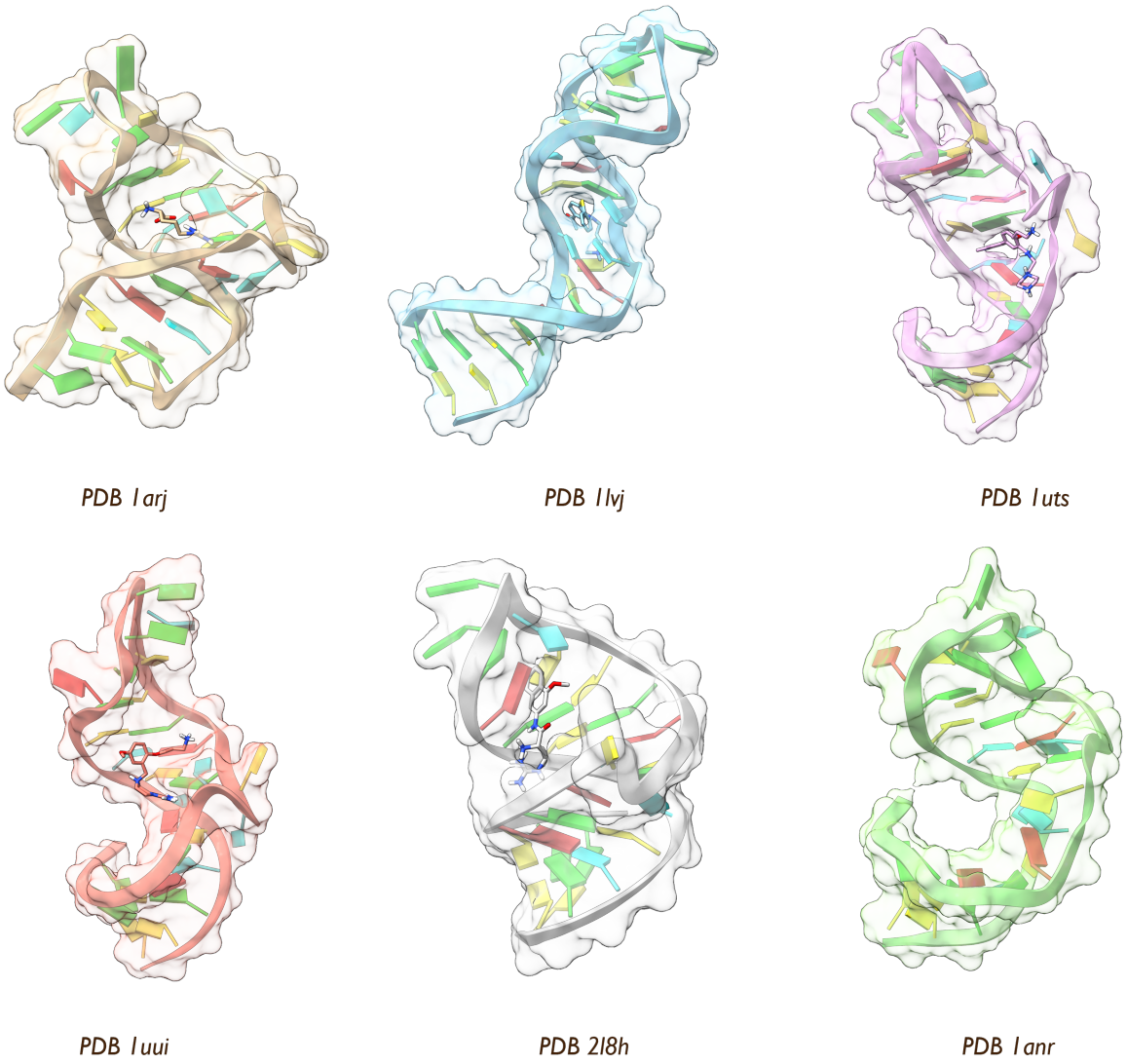


Figure 3.16: Structural diversity of the HIV-1 TAR RNA. Surface-ribbon representation of the structures of the 6 different conformations of the HIV-1 TAR RNA analyzed in this work along with the corresponding PDB id. From left to right and from top to bottom, the first 5 panels report the holo structures used in our validation set (Tab. 3.4) with the corresponding ligand highlighted in ball-and-stick representation. In the last panel, we illustrate an apo conformation of HIV-1 TAR.

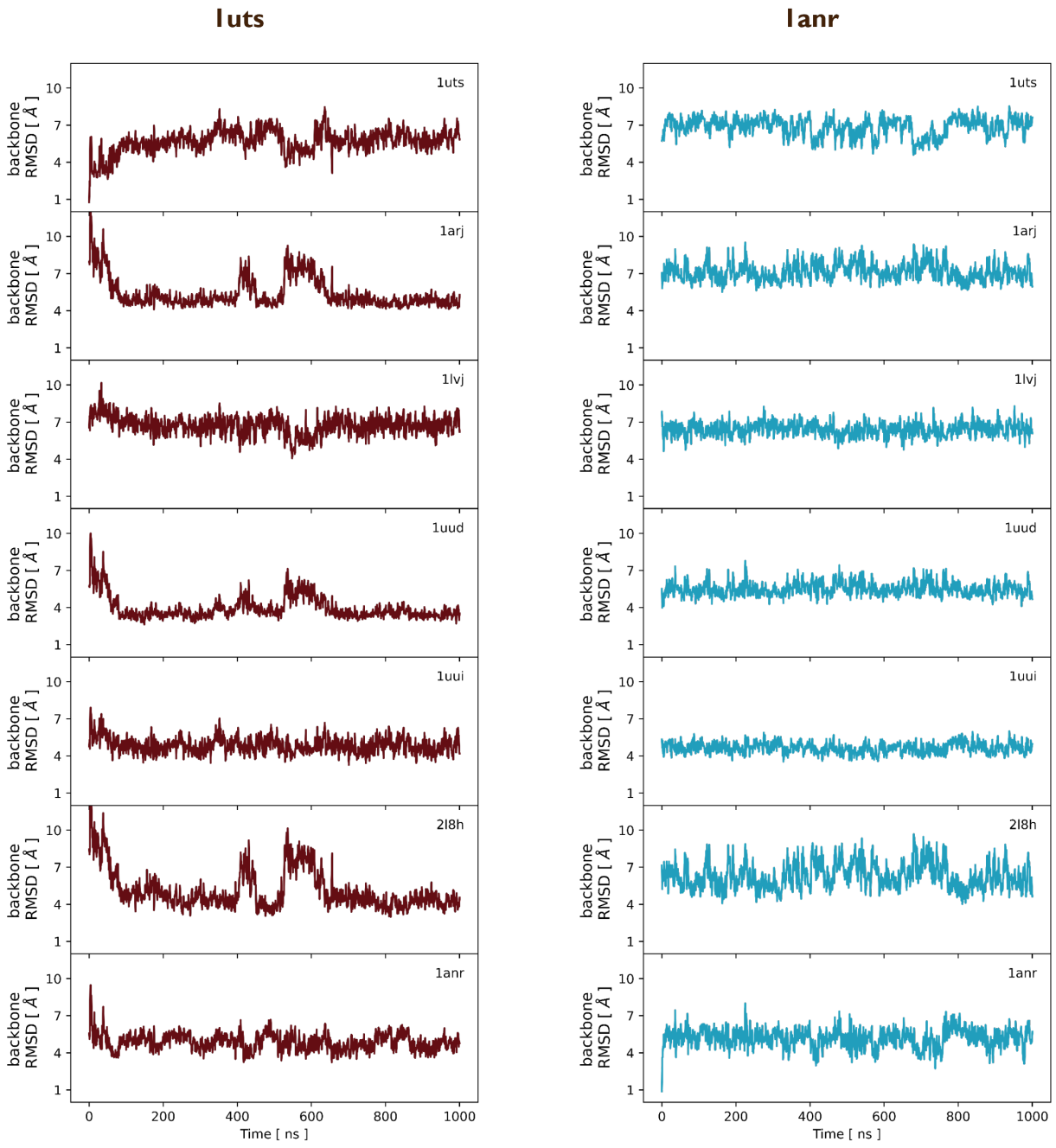


Figure 3.17: Structural variety of the HIV-1 TAR RNA ensembles explored by SHAMAN. RMSD of the backbone atoms of the HIV-1 TAR RNA in the SHAMAN mother simulation with respect to various experimental structures present in our validation set (Tab. 3.4), as a function of simulation time. Each panel corresponds to a different experimental structure, whose PDB id is indicated in the upper right corner. Simulations initiated from holo-like (PDB 1uts) and apo (PDB 1anr) conformations are highlighted in brown (left) and cyan (right), respectively.

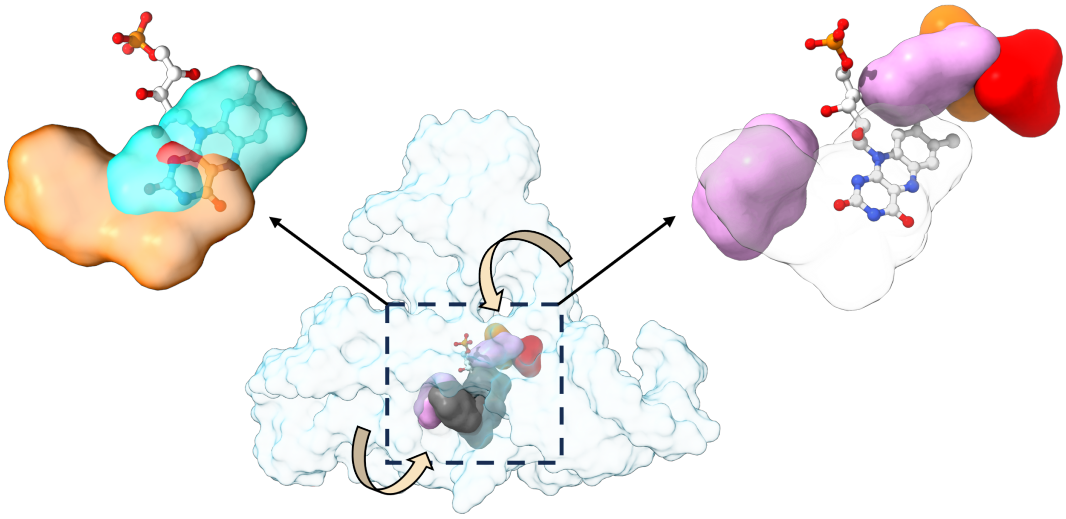


Figure 3.18: Potential applications of SHAMAN in CADD. In the center, the molecular surface of the most populated conformation of the FMN riboswitch obtained in the SHAMAN simulation initiated from an apo structure (PDB 6wjr). The dashed box indicated the FMN resolved binding site. The FMN ligand (PDB 2yie) as well as the SHAMAP that identified this binding site are superimposed by aligning the coordinates to the RNA cluster center. Different colors indicate different probes (color code as in Tab. 3.7 and 3.8). In the left panel, probe interacting sites of which the SHAMAP is composed of and their overlap with the ligand. In the right panel, interacting sites adjacent to the ligand binding site, with arrows suggesting two possible pathways to access the binding site.

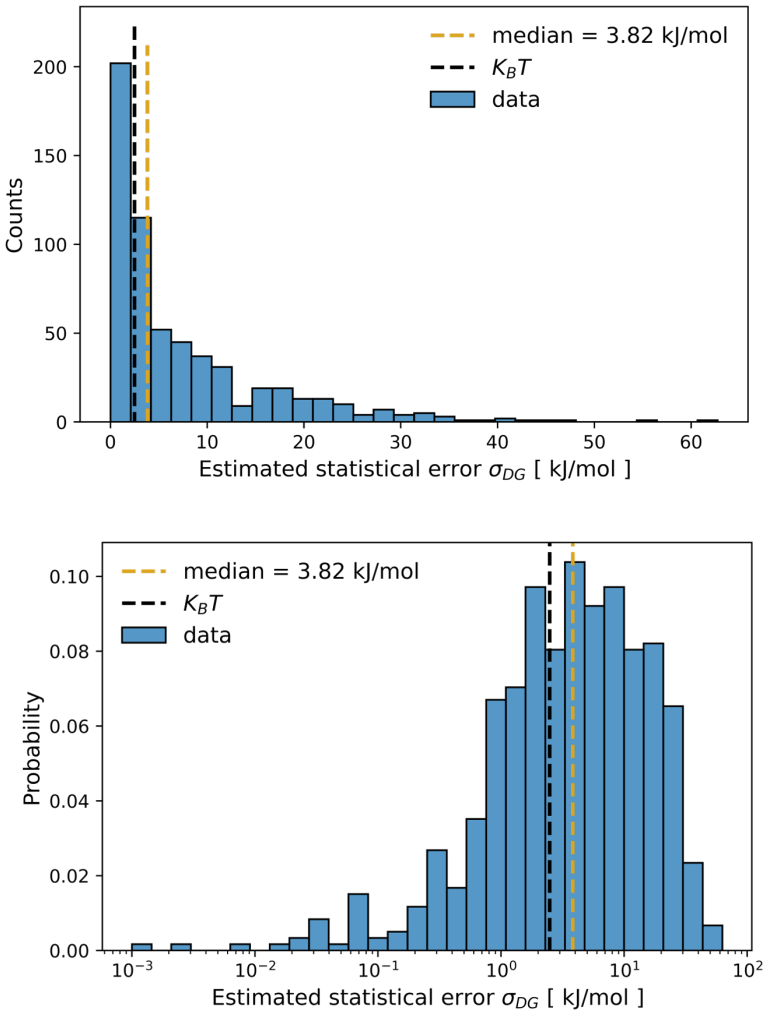


Figure 3.19: Statistical error in the free energy of the probes interacting sites. The normalized distribution of the calculated statistical error $\sigma^l_{\Delta G}$ on the free energy of each probe interacting site l (Sec. 3.4.1) calculated across all probes and systems. The median value of this distribution (3.82 kJ/mol), is reported as a yellow dashed line.

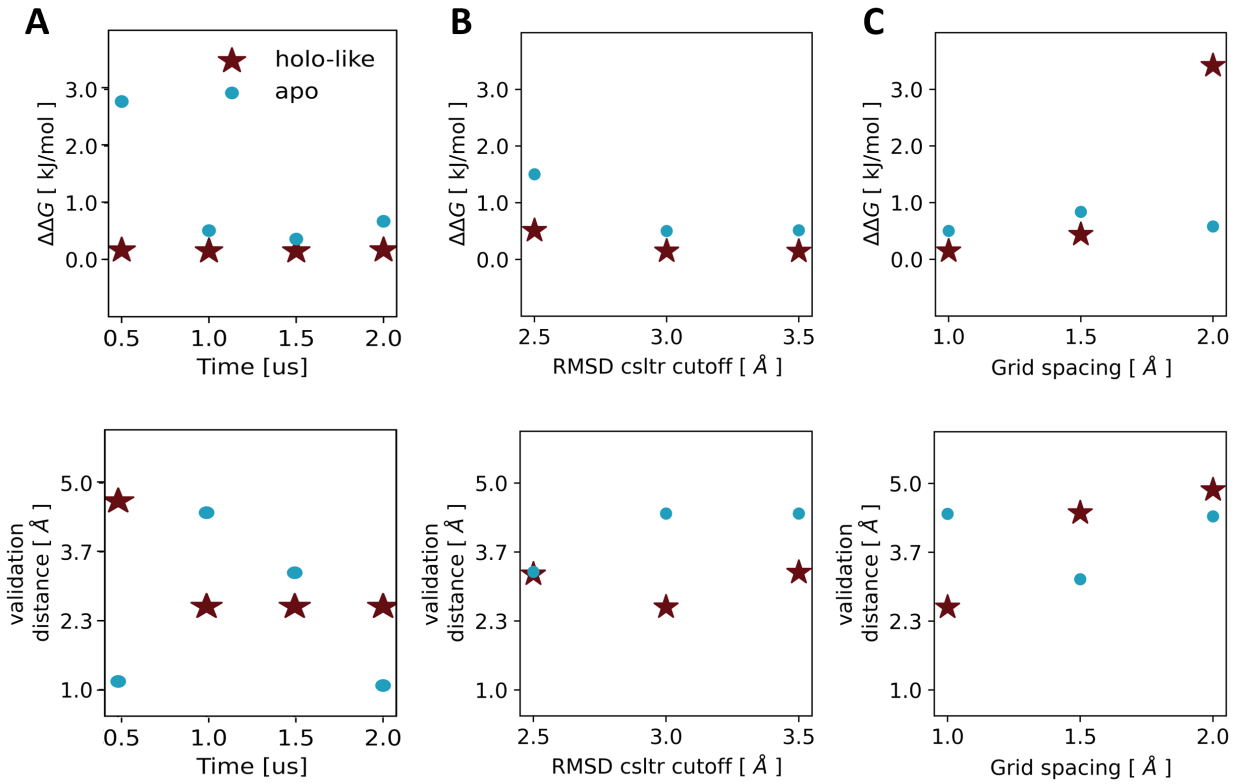


Figure 3.20: Effect of the choice of input parameters on SHAMAN accuracy. A-C) Scatter plot of $\Delta\Delta G$ (top panels) and validation distance (Eq. (3.10), bottom panels) in simulations initiated from holo-like (red) and apo (blue) conformations upon variation of: **A)** simulation time, **B)** clustering cutoff, and **C)** grid spacing in the probe binding free energy calculations. When one input parameter is varied, the other two are kept at their reference value used in the SHAMAN simulations ($T = 1\mu s$, RMSD cutoff = 3.0 Å, grid spacing = 1.0 Å). The data reported in this figure refer to SHAMAN simulations performed on the HIV-1 TAR PDB structure 1uts.

3.5.3 Supplementary tables

system	holo	ligand id	apo	global RMSD [Å]	binding site RMSD [Å]
FMN riboswitch	6dn3	GZ4	6wjr	1.82	1.09
THF riboswitch	4lvx	H4B, H4B	7kd1	3.74	0.36, 2.00
TPP riboswitch	3d2v	PYI	N/A	-	-
dG riboswitch	3ski	GNG	N/A	-	-
HIV-1 TAR	1uts	P13	1anr	5.59	3.82
HCV-IRES-IIa	3tzr	SS0	2nok	7.43	3.97
IAV promoter	2lwk	0EC	1mfy	4.40	0.20

Table 3.1: SHAMAN benchmark set. The first column lists the systems chosen for our benchmark, with the riboswitches and the viral RNAs highlighted in gold and violet, respectively. For each system, the columns with brown and cyan headers report the PDB id of the experimentally resolved holo and (when available) apo structures, respectively, along with the ligand id in the holo structure. The last two columns report the RMSD between holo and apo structures calculated on the common RNA backbone atoms of the whole molecules and of the binding site region only. The THF riboswitch has two copies of the H4B ligand and therefore two values are reported for the binding site RMSD.

starting PDB	# nt	# water molecules	# structural ions	# counter ions [K+/CL-]	total # atoms	simulation time [µs * N]
6dn3	109	17210	9	143/50	72552	1 * 19
6wjr	111	15194	2	155/45	64554	1 * 19
4lvx	89	12093	0	125/37	51405	1 * 17
7kd1	89	14005	2	133/41	59075	1 * 17
3d2v	77	11284	1	97/33	47757	1 * 17
3ski	67	8346	4	95/25	35663	1 * 17
1uts	29	5423	0	44/16	22682	1 * 17
1anr	29	5158	0	43/15	21620	1 * 17
3tzr	36	6028	9	46/18	25346	1 * 17
2nok	44	7988	7	51/23	33440	1 * 17
2lwk	32	4862	0	45/14	20530	1 * 17
1mfy	31	4515	0	43/13	19106	1 * 17

Table 3.2: Details of the SHAMAN simulations. For each of the 12 SHAMAN runs, we report the following details of the mother system simulations. From left to right: the PDB id of the starting structure (with the riboswitches and the viral RNAs indicated in gold and violet, respectively), the number of RNA nucleotides (nt), the number of water molecules, the number of structural ions present in the deposited PDB, the number of extra ions added to neutralize the simulation box at 0.15 M, the total number of atoms, the total simulation time (1 µs for mother and each replica system).

FMN riboswitch	PDB ids	# nt	ligand PDB id	buriedness [au]	deposition year	experimental technique	resolution [Å]	ref	HARIBOSS entry
pocket 1	2yie	156	FMN	0.78	2011	X-ray	2.94	DOI	2yie
	3f2q	128	FMN	0.77	2008	X-ray	2.95	DOI	3f2q
	3f2t	130	FMN	0.77	2008	X-ray	3.00	DOI	3f2t
	3f2w	134	FMN	0.77	2008	X-ray	3.45	DOI	3f2w
	3f2x	129	FMN	0.77	2008	X-ray	3.11	DOI	3f2x
	3f2y	127	FMN	0.78	2008	X-ray	3.20	DOI	3f2y
	3f4e	127	FMN	0.77	2008	X-ray	3.05	DOI	3f4e
	3f4g	128	RBF	0.79	2008	X-ray	3.01	DOI	3f4g
	3f4h	125	RS3	0.80	2008	X-ray	3.00	DOI	3f4h
	3f30	128	FMN	0.79	2008	X-ray	3.15	DOI	3f30
	5c45	119	51B	0.76	2015	X-ray	2.93	DOI	5c45
	5kx9	121	6YG	0.75	2016	X-ray	2.90	DOI	5kx9
	6bfb	124	DKM	0.77	2017	X-ray	2.82	DOI	6bfb
	6dn1	147	GZ7	0.76	2018	X-ray	3.03	DOI	6dn1
	6dn2	126	GZG	0.78	2018	X-ray	2.88	DOI	6dn2
	6dn3	124	GZ4	0.76	2018	X-ray	2.80	DOI	6dn3
THF riboswitch	PDB ids	# nt	ligand PDB id	buriedness [au]	deposition year	experimental technique	resolution [Å]	ref	HARIBOSS entry
pocket 1, pocket 2	3sd3	90	FFO	0.62	2011	X-ray	1.95	DOI	3sd3
	4lvv	**90**	**FFO**	**0.59, 0.57**	**2013**	**X-ray**	**2.10**	**DOI**	**4lvv**
	4lvw	**90**	**7DG**	**-**	**2013**	**X-ray**	**1.77**	**DOI**	**-**
	4lvx	**90**	**H4B**	**0.56, 0.60**	**2013**	**X-ray**	**1.90**	**DOI**	**4lvx**
	4lvy	**90**	**LYA**	**0.58, 0.64**	**2013**	**X-ray**	**2.00**	**DOI**	**4lvy**
	4lvz	90	6AP	-	2013	X-ray	1.77	DOI	-
	4lw0	90	ADE	-	2014	X-ray	1.89	DOI	-
TPP riboswitch	PDB ids	# nt	ligand PDB id	buriedness [au]	deposition year	experimental technique	resolution [Å]	ref	HARIBOSS entry
pocket 1	2gdi	79	TPP	0.76	2006	X-ray	2.05	DOI	2gdi
	2hoj	79	TPP	0.73	2006	X-ray	2.50	DOI	2hoj
	2hok	79	TPP	-	2006	X-ray	3.20	DOI	-
	2hol	80	TPP	0.73	2006	X-ray	2.90	DOI	2hol
	2hom	81	TPS	0.75	2006	X-ray	2.89	DOI	2hom
	2hoo	84	BFT	0.75	2006	X-ray	3.00	DOI	2hoo
	2hop	77	218	0.70	2006	X-ray	3.30	DOI	2hop
	3d2g	78	TPP	0.74	2008	X-ray	2.25	DOI	3d2g
	3d2v	78	PYI	0.78	2008	X-ray	2.00	DOI	3d2v
	3d2x	78	D2X	0.75	2008	X-ray	2.50	DOI	3d2x
	4nya	79	2QB	0.77	2013	X-ray	2.65	DOI	4nya
	4nyb	80	2QC	0.75	2013	X-ray	3.10	DOI	4nyb
	4nyc	80	SVN	-	2014	X-ray	2.90	DOI	-
	4nyd	80	HPA	-	2013	X-ray	3.15	DOI	-
	4nyg	79	VIB	0.72	2013	X-ray	3.05	DOI	4nyg

dG riboswitch	PDB ids	# nt	ligand PDB id	buriedness [au]	deposition year	experimental technique	resolution [Å]	ref	HARIBOSS entry
pocket 1	3ski	68	GNG	0.81	2011	X-ray	2.30	DOI	3ski
	3skl	65	GNG	0.74	2011	X-ray	2.90	DOI	3skl
	3skr	65	GNG	0.78	2011	X-ray	3.10	DOI	3skr
	3skt	65	GNG	0.77	2011	X-ray	3.10	DOI	3skt
	3skw	65	GNG	0.75	2011	X-ray	2.95	DOI	3skw
	3skz	67	GMP	0.71	2011	X-ray	2.61	DOI	3skz
	3slm	67	DGP	0.73	2011	X-ray	2.70	DOI	3slm
	3slq	67	5GP	0.70	2011	X-ray	2.50	DOI	3slq
pocket 2	3fo4	64	6GU	0.77	2008	X-ray	1.90	DOI	3fo4
	3fo6	68	6GO	0.79	2008	X-ray	1.90	DOI	3fo6
	3g4m	68	2BP	-	2009	X-ray	2.40	DOI	-
	3gao	68	XAN	-	2010	X-ray	1.90	DOI	-
	3ger	68	6GU	0.81	2009	X-ray	1.70	DOI	3ger
	3ges	68	6GO	0.79	2009	X-ray	2.15	DOI	3ges
	3gog	66	6GU	0.76	2009	X-ray	2.10	DOI	3gog
	3got	68	A2F	-	2010	X-ray	1.95	DOI	-
	3rkf	68	DX4	0.83	2011	X-ray	2.50	DOI	3rkf
	6ubu	68	GUN	-	2019	X-ray	1.60	DOI	-
	6uc7	68	Q44	0.69	2019	X-ray	1.80	DOI	6uc7
	6uc8	68	ANG	0.81	2019	X-ray	1.90	DOI	6uc8
	6uc9	68	CMG	0.73	2019	X-ray	1.94	DOI	6uc9

Table 3.3: SHAMAN riboswitch validation set. For each riboswitch in our benchmark set (gold cells), we report the details of the holo structures used for validation. Structures are grouped based on pocket similarity (Sec. 3.4.2). For each structure, we report from left to right: link to PDB entry, number of RNA nucleotides (nt), link to ligand PDB id, pocket buriedness (as calculated in HARIBOSS [22]), deposition year, experimental technique, resolution, link to reference publication, and link to HARIBOSS entry, if present. The THF riboswitch is sometimes resolved with two bound ligands (bold entries) therefore some properties are reported as a comma-separated list.

HIV-1 TAR	PDB ids	# nt	ligand PDB id	buriedness [au]	deposition year	experimental technique	resolution [Å]	ref	HARIBOSS entry
pocket 1	1uts	30	P13	0.55	2003	NMR	-	DOI	1uts
pocket 2	1arj	30	ARG	0.67	1995	NMR	-	DOI	1arj
pocket 3	1lvj	32	PMZ	0.61	2002	NMR	-	DOI	1lvj
pocket 4	1uud	30	P14	0.62	2003	NMR	-	DOI	1uud
	1uui	30	P12	-	2003	NMR	-	DOI	1uui
pocket 5	2l8h	30	MV2003		2011	NMR	-	DOI	-
HCV-IRES-IIa	PDB ids	# nt	ligand PDB id	buriedness [au]	deposition year	experimental technique	resolution [Å]	ref	HARIBOSS entry
pocket 1	3tzr	36	SS0	0.66	2011	X-ray	2.21	DOI	3tzr
pocket 2	2ku0	39	ISI	0.57	2010	NMR	-	DOI	2ku0
	2ktz	39	ISH	0.59	2010	NMR	-	DOI	2ktz
IAV promoter	PDB ids	# nt	ligand PDB id	buriedness [au]	deposition year	experimental technique	resolution [Å]	ref	HARIBOSS entry
pocket 1	2lwk	33	0EC	-	2012	NMR	-	DOI	2lwk

Table 3.4: SHAMAN viral RNAs validation set. Columns defined as in Tab. 3.3.

FMN riboswitch	ΔΔG [kJ/mol]	rank	best match	distance [Å]	ΔΔG [kJ/mol]	rank	best match	distance [Å]
pocket 1	<0.1	2	1-1 04-DMEE 5kx9 6YG	1.49	<0.1	2	0-1 05-MEPY 6dn3 GZ4	1.73
THF riboswitch	ΔΔG [kJ/mol]	rank	best match	distance [Å]	ΔΔG [kJ/mol]	rank	best match	distance [Å]
pocket 1	<0.1	2	1-1 04-DMEE 5kx9 6YG	2.37	<0.1	2	0-1 05-MEPY 6dn3 GZ4	1.51
pocket 2	<0.2	3	1-1 04-DMEE 5kx9 6YG	1.48	<0.2	3	0-1 05-MEPY 6dn3 GZ5	2.31
TPP riboswitch	ΔΔG [kJ/mol]	rank	best match	distance [Å]	ΔΔG [kJ/mol]	rank	best match	distance [Å]
pocket 1	0.0	1	0-1 05-MEPY 4nyd HPA	0.65	-	-	-	
dG riboswitch	ΔΔG [kJ/mol]	rank	best match	distance [Å]	ΔΔG [kJ/mol]	rank	best match	distance [Å]
pocket 1	<0.1	4	0-2 06-BENF 3slq 5GP	1.57	-	-	-	
pocket 2	3.3	41	2-3 01-BENX 5kx9 6YG	3.27	-	-	-	

Table 3.5: Details of the SHAMAPS corresponding to the experimental binding sites in the riboswitch benchmark set. For each system and unique pocket, from left to right: ΔΔG (Eq. \eqref{}) of the SHAMAP with best overlap with the ligand in the experimental structures (Sec. 3.4.1), rank, details of the interacting site with the best match (interacting site id with index of the RNA cluster, probe name, PDB id of the matching experimental structure, ligand name), and validation distance (Eq. 10). Columns with brown and cyan headers report the results of the SHAMAN simulations initiated from holo-like and (when available) apo structures, respectively.

HIV-1 TAR RNA	ΔΔG [kJ/mol]	rank	best match	distance [Å]	ΔΔG [kJ/mol]	rank	best match	distance [Å]
pocket 1	5.0	47	8-2 10-FORM 1arj ARG	4.6	1.0	23	2-4 09-PIRZ 1arj ARG	3.89
pocket 2	0.2	11	0-8 08-IMIA 1lvj PMZ	2.4	0.2	8	0-3 04-DMEE 1lvj PMZ	2.94
pocket 3	1.9	27	2-8 16-FORM 1uts P13	2.32	0.2	8	0-3 04-DMEE 1uts P13	6.25
pocket 4	5.0	47	8-2 10-FORM 1uui P12	5.38	1.0	23	2-4 09-PIRZ 1uui P12	3.84
pocket 5	5.0	47	8-2 10-FORM 2l8h MV2003	5.45	0.1	5	0-1 12-MAMY 2l8h MV2003	3.76
HCV-IRES IIa RNA	ΔΔG [kJ/mol]	rank	best match	distance [Å]	ΔΔG [kJ/mol]	rank	best match	distance [Å]
pocket 1	0.2	3	0-3 13-ACEY 3tzr SS0	2.67	2.9	14	1-1 14-MAMY-ions 3tzr SS0	2.84
pocket 2	0.0	1	0-1 02-BETH 2ku0 ISI	2.31	<0.1	5	0-1 08-IMIA 2ktz ISH	3.17
IAV RNA promoter	ΔΔG [kJ/mol]	rank	best match	distance [Å]	ΔΔG [kJ/mol]	rank	best match	distance [Å]
pocket 1	0.2	3	0-1 15-IMIA 2lwk 0EC	1.43	5.9	47	1-3 13-ACEY 2lwk 0EC	3.64

Table 3.6: Details of the SHAMAPS corresponding to the experimental binding sites in the viral RNA benchmark set. Columns as in Tab. 3.5

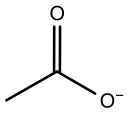
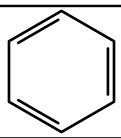
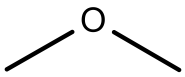
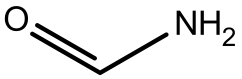
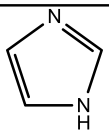
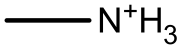
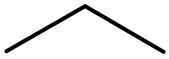
Probe	2D structure	Full name	Color
ACEY		Acetate	
BENX		Benzene	
DMEE		Dimethyl ether	
FORM		Formamide	
IMIA		Imidazole	
MAMY		Methyl-ammonium	
MEOH		Methanol	
PRPX		Propane	

Table 3.7: First set of SHAMAN probes. These probes were used in the development of SILCS-RNA2. From left to right: probe id, 2D structure, probe name, and color code.

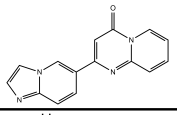
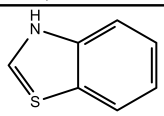
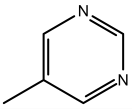
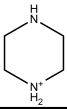
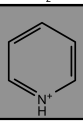
Probe	2D structure	Full name	Color
BENX		Benzene	
BENF		Dihydro-pyrido-pyrimidinone-imidazo-pyridine	
BETH		Benzothiophene	
MEPY		Methyil-pyrimidine	
PIRZ		Piperazine	
PYRD		Pyrimidine	

Table 3.8: Second set of SHAMAN probes. Probes determined in this work from the fragmentation of known RNA binders (Sec. 3.4.2). Except for PYRD (dark grey), this constituted the second set of probes used in our SHAMAN simulations. From left to right: probe id, 2D structure, probe name, and color code.

whole ligands	Pearson correlation	p-value	Murcko scaffolds	Pearson correlation	p-value
molecular weight	-0.08	0.70	molecular weight	0.16	0.40
# aromatic rings	0.20	0.31	# aromatic rings	0.19	0.32
# H-bond donors	-0.21	0.28	# H-bond donors	0.14	0.45
# H-bond acceptors	-0.16	0.24	# H-bond acceptors	0.05	0.82
topological polar surface area	-0.24	0.21	topological polar surface area	0.76	0.70
# heterocycles	0.34	0.08	# heterocycles	0.36	0.08

Table 3.9: Correlation between physico-chemical properties of ligands and successful probes. For each property: name of the property, Pearson correlation coefficient (P_{CC}) and the corresponding p-value calculated between ligands and probes that successfully identified the corresponding binding site. Only SHAMAN simulations initiated from holo-like structures were considered for this analysis. Values were calculated considering either the entire ligand (left) or its Murcko scaffold (right).

	successful probes	**unsuccessful probes**
similar probes	TP = 4	FN = 9
dissimilar probes	FP = 18	TN = 42

Total population	73
TPR	0.31
TNR	0.70
PPV	0.08
NPV	0.82

Table 3.10: Analysis of the relation between probe-ligand similarity and being a successful probe. (Left) Confusion matrix to test the hypothesis that probes similar to ligands are successful, with number of cases with similar/dissimilar probes and successful/unsuccessful probes. Only SHAMAN simulations initiated from holo-like structures were considered for this analysis. (Right) True positive rate (TPR), true negative rate (TNR), positive predictive value (PPV), and negative predictive value (NPV).

	2yie	3f2q	3f2t	3f2w	3f2x	3f2y	3f4e	3f4g	3f4h	3f30	5c45	5kx9	6bfb	6dn1	6dn2	6dn3
2yie		0.55	0.58	0.67	0.57	0.57	0.41	0.61	0.76	0.60	0.75	0.75	0.74	0.31	0.65	0.81
3f2q			0.48	0.53	0.35	0.32	0.36	0.56	0.66	0.54	0.66	0.70	0.60	0.69	0.55	0.70
3f2t				0.47	0.44	0.50	0.49	0.61	0.64	0.43	0.69	0.67	0.74	0.57	0.56	0.65
3f2w					0.50	0.50	0.50	0.69	0.75	0.55	0.73	0.70	0.76	0.66	0.65	0.68
3f2x						0.34	0.41	0.55	0.64	0.57	0.68	0.59	0.69	0.58	0.50	0.64
3f2y							0.44	0.58	0.63	0.55	0.73	0.64	0.66	0.63	0.57	0.77
3f4e								0.52	0.63	0.46	0.71	0.61	0.73	0.50	0.51	0.69
3f4g									0.31	0.62	0.69	0.51	0.56	0.67	0.58	0.77
3f4h										0.65	0.70	0.58	0.51	0.72	0.66	0.85
3f30											0.79	0.73	0.70	0.60	0.59	0.73
5c45												0.39	0.69	0.78	0.69	0.66
5kx9													0.55	0.79	0.63	0.56
6bfb														0.68	0.60	0.74
6dn1															0.62	0.74
6dn2																0.59
6dn3																

Table 3.11: Similarity between the experimentally determined structures of the FMN riboswitch. RMSD between all pairs of FMN riboswitch structures present in our validation set (Tab. 3.3). The RMSD was calculated on all the matching pairs of backbone atoms of the two RNA molecules. All values are in Angstrom.

Bibliography

[1] Thomas R. Cech and Joan A. Steitz. "The noncoding RNA revolution - Trashing old rules to forge new ones". In: *Cell* 157.1 (2014), pp. 77–94.

[2] Jennifer Cable et al. "Noncoding RNAs: biology and applications—a Keystone Symposia report". In: *Annals of the New York Academy of Sciences* 1506.1 (Dec. 2021), pp. 118–141.

[3] Stefanie A. Mortimer, Mary Anne Kidwell, and Jennifer A. Doudna. "Insights into RNA structure and function from genome-wide studies". In: *Nature Reviews Genetics* 15.7 (2014), pp. 469–479.

[4] Run-Wen Yao, Yang Wang, and Ling-Ling Chen. "Cellular functions of long noncoding RNAs". In: *Nature Cell Biology* 21.5 (May 2019), pp. 542–551.

[5] Feng Wang, Travis Zuroske, and Jonathan K. Watts. "RNA therapeutics on the rise". In: *Nature Reviews Drug Discovery* 19.7 (July 2020), pp. 441–442.

[6] Tulsi Ram Damase et al. "The Limitless Future of RNA Therapeutics". In: *Frontiers in Bioengineering and Biotechnology* 9 (Mar. 2021).

[7] François Halloy et al. "Innovative developments and emerging technologies in RNA therapeutics". In: *RNA Biology* 19.1 (Dec. 2022), pp. 313–332.

[8] Noreen F. Rizvi and Graham F. Smith. "RNA as a small molecule druggable target". In: *Bioorganic & Medicinal Chemistry Letters* 27.23 (Dec. 2017), pp. 5083–5088.

[9] James P. Falese, Anita Donlic, and Amanda E. Hargrove. "Targeting RNA with small molecules: from fundamental principles towards the clinic". In: *Chemical Society Reviews* 50.4 (2021), pp. 2224–2243.

[10] Matthew D. Disney. "Targeting RNA with Small Molecules To Capture Opportunities at the Intersection of Chemistry, Biology, and Medicine". In: *Journal of the American Chemical Society* 141.17 (May 2019), pp. 6776–6790.

[11] Katherine Deigan Warner, Christine E. Hajdin, and Kevin M. Weeks. "Principles for targeting RNA with drug-like small molecules". In: *Nature Reviews Drug Discovery* 17.8 (2018), pp. 547–558.

[12] Ryszard Kole, Adrian R. Krainer, and Sidney Altman. "RNA therapeutics: beyond RNA interference and antisense oligonucleotides". In: *Nature Reviews Drug Discovery* 11.2 (Feb. 2012), pp. 125–140.

[13] James C. Kaczmarek, Piotr S. Kowalski, and Daniel G. Anderson. "Advances in the delivery of RNA therapeutics: from concept to clinical reality". In: *Genome Medicine* 9.1 (Dec. 2017), p. 60.

[14] Melanie Winkle et al. “Noncoding RNA therapeutics — challenges and potential solutions”. In: *Nature Reviews Drug Discovery* 20.8 (Aug. 2021), pp. 629–651.

[15] D.C. Luther et al. “Delivery approaches for CRISPR/Cas9 therapeutics in vivo : advances and challenges”. In: *Expert Opinion on Drug Delivery* 15.9 (Sept. 2018), pp. 905–913.

[16] Brittany S. Morgan, Jordan E. Forte, and Amanda E. Hargrove. “Survey and summary insights into the development of chemical probes for RNA”. In: *Nucleic Acids Research* 46.16 (2018), pp. 8025–8037.

[17] Samantha M. Meyer et al. “Small molecule recognition of disease-relevant RNA structures”. In: *Chemical Society Reviews* 49.19 (2020), pp. 7167–7199.

[18] John A. Howe et al. “Selective small-molecule inhibition of an RNA structural element”. In: *Nature* 526.7575 (Oct. 2015), pp. 672–677.

[19] Hasane Ratni et al. “Discovery of Risdiplam, a Selective Survival of Motor Neuron-2 (SMN2) Gene Splicing Modifier for the Treatment of Spinal Muscular Atrophy (SMA)”. In: *Journal of Medicinal Chemistry* 61.15 (Aug. 2018), pp. 6501–6517.

[20] Seyed MohammadReza Hashemian, Tayebeh Farhadi, and Mojdeh Ganjparvar. “Linezolid: a review of its properties, function, and use in critical care”. In: *Drug Design, Development and Therap* 12 (June 2018), pp. 1759–1767.

[21] Kamyar Yazdani et al. “Machine Learning Informs RNA-Binding Chemical Space**”. In: *Angewandte Chemie* 135.11 (Mar. 2023), e202211358.

[22] F P Panei et al. “HARIBOSS: a curated database of RNA-small molecules structures to aid rational drug design”. In: *Bioinformatics* 38.17 (Sept. 2022), pp. 4185–4193.

[23] Ankita Mehta et al. “SMMRNA: a database of small molecule modulators of RNA”. In: *Nucleic Acids Research* 42.D1 (Jan. 2014), pp. D132–D141.

[24] Subodh Kumar Mishra and Amit Kumar. “NALDB: nucleic acid ligand database for small molecules targeting nucleic acid”. In: *Database* 2016 (Feb. 2016), baw002.

[25] Saisai Sun, Jianyi Yang, and Zhaolei Zhang. “RNALigands: a database and web server for RNA–ligand interactions”. In: *RNA* 28.2 (Feb. 2022), pp. 115–122.

[26] Anita Donlic et al. “R-BIND 2.0: An Updated Database of Bioactive RNA-Targeting Small Molecules and Associated RNA Secondary Structures”. In: *ACS Chemical Biology* 17.6 (June 2022), pp. 1556–1566.

[27] Matthew D. Disney et al. “Inforna 2.0: A Platform for the Sequence-Based Design of Small Molecules Targeting Structured RNAs”. In: *ACS Chemical Biology* 11.6 (June 2016), pp. 1720–1728.

[28] Illimar Hugo Rekand and Ruth Brenk. “DrugPred_RNA—A Tool for Structure-Based Druggability Predictions for RNA Binding Sites”. In: *Journal of Chemical Information and Modeling* 61.8 (Aug. 2021), pp. 4068–4081.

[29] Pan Zeng and Qinghua Cui. “Rsite2: an efficient computational method to predict the functional sites of noncoding RNAs”. In: *Scientific Reports* 6.1 (Jan. 2016), p. 19016.

[30] Kaili Wang et al. “RLBind: a deep learning method to predict RNA–ligand binding sites”. In: *Briefings in Bioinformatics* 24.1 (Jan. 2023), bbac486.

[31] Abhishek A. Kognole, Anthony Hazel, and Alexander D. MacKerell. “SILCS-RNA: Toward a Structure-Based Drug Design Approach for Targeting RNAs with Small Molecules”. In: *Journal of Chemical Theory and Computation* 18.9 (Sept. 2022), pp. 5672–5691.

[32] Hong Su, Zhenling Peng, and Jianyi Yang. “Recognition of small molecule–RNA binding sites using RNA sequence and structure”. In: *Bioinformatics* 37.1 (Apr. 2021), pp. 36–42.

[33] Sergio Ruiz-Carmona et al. “rDock: A Fast, Versatile and Open Source Program for Docking Ligands to Proteins and Nucleic Acids”. In: *PLoS Computational Biology* 10.4 (Apr. 2014), e1003571.

[34] Yuyu Feng et al. “NLDock: a Fast Nucleic Acid–Ligand Docking Algorithm for Modeling RNA/DNA–Ligand Complexes”. In: *Journal of Chemical Information and Modeling* 61.9 (Sept. 2021), pp. 4771–4782.

[35] Yangwei Jiang and Shi-Jie Chen. “RLDOCK method for predicting RNA-small molecule binding modes”. In: *Methods* 197 (Jan. 2022), pp. 97–105.

[36] Christophe Guilbert and Thomas L. James. “Docking to RNA via Root-Mean-Square-Deviation-Driven Energy Minimization with Flexible Ligands and Flexible Targets”. In: *Journal of Chemical Information and Modeling* 48.6 (June 2008), pp. 1257–1268.

[37] Filip Stefaniak and Janusz M. Bujnicki. “AnnapuRNA: A scoring function for predicting RNA-small molecule binding poses”. In: *PLOS Computational Biology* 17.2 (Feb. 2021), e1008309.

[38] Sahil Chhabra, Jingru Xie, and Aaron T. Frank. “RNAPosers: Machine Learning Classifiers for Ribonucleic Acid–Ligand Poses”. In: *The Journal of Physical Chemistry B* 124.22 (June 2020), pp. 4436–4445.

[39] Patrick Pfeffer and Holger Gohlke. “DrugScore ^{RNA} Knowledge-Based Scoring Function To Predict RNA−Ligand Interactions”. In: *Journal of Chemical Information and Modeling* 47.5 (Sept. 2007), pp. 1868–1876.

[40] Anna Philips et al. “LigandRNA: computational predictor of RNA–ligand interactions”. In: *RNA* 19.12 (Dec. 2013), pp. 1605–1616.

[41] Jacopo Manigrasso, Marco Marcia, and Marco De Vivo. “Computer-aided design of RNA-targeted small molecules: A growing need in drug discovery”. In: *Chem* 7.11 (Nov. 2021), pp. 2965–2988.

[42] Laura R. Ganser et al. “The roles of structural dynamics in the cellular functions of RNAs”. In: *Nature Reviews Molecular Cell Biology* 20.8 (2019), pp. 474–489.

[43] Megan L. Ken et al. “RNA conformational propensities determine cellular activity”. In: *Nature* 617.7962 (May 2023), pp. 835–841.

[44] Hashim M Al-Hashimi and Nils G Walter. “RNA dynamics: it is about time”. In: *Current Opinion in Structural Biology* 18.3 (June 2008), pp. 321–329.

[45] Komal Soni et al. “Structural basis for specific RNA recognition by the alternative splicing factor RBM5”. In: *Nature Communications* 14.1 (July 2023), p. 4233.

[46] Jiří Šponer et al. “RNA Structural Dynamics As Captured by Molecular Simulations: A Comprehensive Overview”. In: *Chemical Reviews* 118.8 (Apr. 2018), pp. 4177–4338.

[47] Mattia Bernetti and Giovanni Bussi. “Integrating experimental data with molecular simulations to investigate RNA structural dynamics”. In: *Current Opinion in Structural Biology* 78 (Feb. 2023), p. 102503.

[48] Lucas Defelipe et al. “Solvents to Fragments to Drugs: MD Applications in Drug Design”. In: *Molecules* 23.12 (Dec. 2018), p. 3269.

[49] Loïc Salmon et al. “A general method for constructing atomic-resolution RNA ensembles using NMR residual dipolar couplings: The basis for interhelical motions revealed”. In: *Journal of the American Chemical Society* 135.14 (2013), pp. 5457–5466.

[50] Alessandro Laio and Michele Parrinello. “Escaping free-energy minima”. In: *PNAS* 99.20 (2002), pp. 12562–12566.

[51] Thomas A. Halgren. “Identifying and Characterizing Binding Sites and Assessing Druggability”. In: *Journal of Chemical Information and Modeling* 49.2 (Feb. 2009), pp. 377–389.

[52] Igor Kozlovskii and Petr Popov. “Structure-based deep learning for binding site detection in nucleic acid macromolecules”. In: *NAR Genomics and Bioinformatics* 3.4 (Oct. 2021).

[53] Kaili Wang et al. “RBind: computational network method to predict RNA binding sites”. In: *Bioinformatics* 34.18 (Sept. 2018), pp. 3131–3136.

[54] Huiwen Wang and Yunjie Zhao. “RBinds: A user-friendly server for RNA binding site prediction”. In: *Computational and Structural Biotechnology Journal* 18 (Jan. 2020), pp. 3762–3765.

[55] Haley M. Wilt et al. “FMN riboswitch aptamer symmetry facilitates conformational switching through mutually exclusive coaxial stacking configurations”. In: *Journal of Structural Biology: X* 4 (2020), p. 100035.

[56] Noreen F. Rizvi et al. “Discovery of Selective RNA-Binding Small Molecules by Affinity-Selection Mass Spectrometry”. In: *ACS Chemical Biology* 13.3 (Mar. 2018), pp. 820–831.

[57] Quentin Vicens et al. “Structure–Activity Relationship of Flavin Analogues That Target the Flavin Mononucleotide Riboswitch”. In: *ACS Chemical Biology* 13.10 (Oct. 2018), pp. 2908–2919.

[58] D Harrich, C Ulich, and R B Gaynor. “A critical role for the TAR element in promoting efficient human immunodeficiency virus type 1 reverse transcription”. In: *Journal of Virology* 70.6 (June 1996), pp. 4017–4027.

[59] Sai Shashank Chavali, Rachel Bonn-Breach, and Joseph E. Wedekind. “Face-time with TAR: Portraits of an HIV-1 RNA with diverse modes of effector recognition relevant for drug discovery”. In: *Journal of Biological Chemistry* 294.24 (June 2019), pp. 9326–9341.

[60] Amy Davidson et al. “A Small-Molecule Probe Induces a Conformation in HIV TAR RNA Capable of Binding Drug-Like Fragments”. In: *Journal of Molecular Biology* 410.5 (July 2011), pp. 984–996.

[61] Catherine Musselman, Hashim M. Al-Hashimi, and Ioan Andricioaei. "iRED Analysis of TAR RNA Reveals Motional Coupling, Long-Range Correlations, and a Dynamical Hinge". In: *Biophysical Journal* 93.2 (July 2007), pp. 411–422.

[62] Konrad Krawczyk et al. "Tertiary Element Interaction in HIV-1 TAR". In: *Journal of Chemical Information and Modeling* 56.9 (Sept. 2016), pp. 1746–1754.

[63] Alastair I.H. Murchie et al. "Structure-based Drug Design Targeting an Inactive RNA Conformation: Exploiting the Flexibility of HIV-1 TAR RNA". In: *Journal of Molecular Biology* 336.3 (Feb. 2004), pp. 625–638.

[64] F Aboul-ela. "Structure of HIV-1 TAR RNA in the absence of ligands reveals a novel conformation of the trinucleotide bulge". In: *Nucleic Acids Research* 24.20 (Oct. 1996), pp. 3974–3981.

[65] Alexa M Salsbury and Justin A Lemkul. "Recent developments in empirical atomistic force fields for nucleic acids and applications to studies of folding and dynamics". In: *Current Opinion in Structural Biology* 67 (Apr. 2021), pp. 9–17.

[66] Massimiliano Bonomi et al. "Principles of protein structural ensemble determination". In: *Current Opinion in Structural Biology* 42 (2017), pp. 106–116.

[67] Sandro Bottaro et al. "Conformational ensembles of RNA oligonucleotides from integrating NMR and molecular simulations". In: *Science Advances* 4.5 (May 2018), eaar8521.

[68] H. M. Berman. "The Protein Data Bank". In: *Nucleic Acids Research* 28.1 (Jan. 2000), pp. 235–242.

[69] Eric F. Pettersen et al. "UCSF Chimera?A visualization system for exploratory research and analysis". In: *Journal of Computational Chemistry* 25.13 (Oct. 2004), pp. 1605–1612.

[70] Peter Eastman et al. "OpenMM 7: Rapid development of high performance algorithms for molecular dynamics". In: *PLOS Computational Biology* 13.7 (July 2017), e1005659.

[71] Kresten Lindorff-Larsen et al. "Improved side-chain torsion potentials for the Amber ff99SB protein force field". In: *Proteins: Structure, Function, and Bioinformatics* 78.8 (June 2010), pp. 1950–1958.

[72] Alberto Pérez et al. "Refinement of the AMBER Force Field for Nucleic Acids: Improving the Description of α/γ Conformers". In: *Biophysical Journal* 92.11 (June 2007), pp. 3817–3829.

[73] Marie Zgarbová et al. "Refinement of the Cornell et al. Nucleic Acids Force Field Based on Reference Quantum Chemical Calculations of Glycosidic Torsion Profiles". In: *Journal of Chemical Theory and Computation* 7.9 (Sept. 2011), pp. 2886–2902.

[74] In Suk Joung and Thomas E. Cheatham. "Determination of Alkali and Halide Monovalent Ion Parameters for Use in Explicitly Solvated Biomolecular Simulations". In: *The Journal of Physical Chemistry B* 112.30 (July 2008), pp. 9020–9041.

[75] Olof Allnér, Lennart Nilsson, and Alessandra Villa. "Magnesium Ion–Water Coordination and Exchange in Biomolecular Simulations". In: *Journal of Chemical Theory and Computation* 8.4 (Apr. 2012), pp. 1493–1502.

[76] Saeed Izadi, Ramu Anandakrishnan, and Alexey V. Onufriev. "Building Water Models: A Different Approach". In: *The Journal of Physical Chemistry Letters* 5.21 (Nov. 2014), pp. 3863–3871.

[77] Simon Boothroyd et al. "Development and Benchmarking of Open Force Field 2.0.0: The Sage Small Molecule Force Field". In: *Journal of Chemical Theory and Computation* 19.11 (June 2023), pp. 3251–3275.

[78] Ulrich Essmann et al. "A smooth particle mesh Ewald method". In: *The Journal of Chemical Physics* 103.19 (Nov. 1995), pp. 8577–8593.

[79] Mark James Abraham et al. "GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers". In: *SoftwareX* 1-2 (Sept. 2015), pp. 19–25.

[80] Gareth A. Tribello et al. "PLUMED 2: New feathers for an old bird". In: *Computer Physics Communications* 185.2 (Feb. 2014), pp. 604–613.

[81] H. J. C. Berendsen et al. "Molecular dynamics with coupling to an external bath". In: *The Journal of Chemical Physics* 81.8 (Oct. 1984), pp. 3684–3690.

[82] Giovanni Bussi, Davide Donadio, and Michele Parrinello. "Canonical sampling through velocity rescaling". In: *The Journal of Chemical Physics* 126.1 (Jan. 2007), p. 014101.

[83] Alessandro Barducci, Giovanni Bussi, and Michele Parrinello. "Well-Tempered Metadynamics: A Smoothly Converging and Tunable Free-Energy Method". In: *Physical Review Letters* 100.2 (Jan. 2008), p. 020603.

[84] Davide Branduardi, Giovanni Bussi, and Michele Parrinello. "Metadynamics with Adaptive Gaussians". In: *Journal of Chemical Theory and Computation* 8.7 (July 2012), pp. 2247–2254.

[85] Xavier Daura et al. "Peptide Folding: When Simulation Meets Experiment". In: *Angewandte Chemie International Edition* 38.1-2 (Jan. 1999), pp. 236–240.

[86] Richard Gowers et al. "MDAnalysis: A Python Package for the Rapid Analysis of Molecular Dynamics Simulations". In: 2016, pp. 98–105.

[87] Fabian Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[88] Ryne C. Johnston et al. "Epik: pKa and Protonation State Prediction through Machine Learning". In: *Journal of Chemical Theory and Computation* 19.8 (Apr. 2023), pp. 2380–2388.

[89] Xiao Qing Lewell et al. "RECAPRetrosynthetic Combinatorial Analysis Procedure: A Powerful New Technique for Identifying Privileged Molecular Fragments with Useful Applications in Combinatorial Chemistry". In: *Journal of Chemical Information and Computer Sciences* 38.3 (May 1998), pp. 511–522.

[90] Schrödinger. "Schrödinger Release 2023-1: LigPrep". In: *LCC, New York, NY* (2023).

[91] M.A. Larkin et al. "Clustal W and Clustal X version 2.0". In: *Bioinformatics* 23.21 (Nov. 2007), pp. 2947–2948.

[92] Davide Chicco and Giuseppe Jurman. "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation". In: *BMC Genomics* 21.1 (Jan. 2020), pp. 1–13.

[93] The PLUMED consortium. "Promoting transparency and reproducibility in enhanced molecular simulations". In: *Nature Methods* 16.8 (Aug. 2019), pp. 670–673.

[94] Atiqah Subki et al. "Identification and characterisation of thiamine pyrophosphate (TPP) riboswitch in Elaeis guineensis". In: *PLOS ONE* 15.7 (July 2020), e0235431.

[95] Deborah Antunes et al. "Unraveling RNA dynamical behavior of TPP riboswitches: a comparison between Escherichia coli and Arabidopsis thaliana". In: *Scientific Reports 2019 9:1* 9.1 (Mar. 2019), pp. 1–13.

[96] Katherine Deigan Warner et al. "Validating Fragment-Based Drug Discovery for Biological RNAs: Lead Fragments Bind and Remodel the TPP Riboswitch Specifically". In: *Chemistry & Biology* 21.5 (May 2014), pp. 591–595.

[97] Stéphane Thore, Christian Frick, and Nenad Ban. "Structural basis of thiamine pyrophosphate analogues binding to the eukaryotic riboswitch". In: *Journal of the American Chemical Society* 130.26 (July 2008), pp. 8116–8117.

[98] Thomas E. Edwards and Adrian R. Ferré-D'Amaré. "Crystal Structures of the Thi-Box Riboswitch Bound to Thiamine Pyrophosphate Analogs Reveal Adaptive RNA-Small Molecule Recognition". In: *Structure* 14.9 (Sept. 2006), pp. 1459–1468.

[99] Jeremiah J. Trausch et al. "The Structure of a Tetrahydrofolate-Sensing Riboswitch Reveals Two Ligand Binding Sites in a Single Aptamer". In: *Structure* 19.10 (Oct. 2011), pp. 1413–1423.

[100] Jeremiah J. Trausch and Robert T. Batey. "A Disconnect between High-Affinity Binding and Efficient Regulation by Antifolates and Purines in the Tetrahydrofolate Riboswitch". In: *Chemistry & Biology* 21.2 (Feb. 2014), pp. 205–216.

[101] Haley M. Wilt et al. "Tying the knot in the tetrahydrofolate (THF) riboswitch: A molecular basis for gene regulation". In: *Journal of Structural Biology* 213.1 (Mar. 2021), p. 107703.

[102] Michal M. Matyjasik and Robert T. Batey. "Structural basis for 2'-deoxyguanosine recognition by the 2'-dG-II class of riboswitches". In: *Nucleic Acids Research* 47.20 (Nov. 2019), pp. 10931–10941.

[103] Michal M. Matyjasik, Simone D. Hall, and Robert T. Batey. "High Affinity Binding of N2-Modified Guanine Derivatives Significantly Disrupts the Ligand Binding Pocket of the Guanine Riboswitch". In: *Molecules 2020, Vol. 25, Page 2295* 25.10 (May 2020), p. 2295.

[104] Olga Pikovskaya et al. "Structural principles of nucleoside selectivity in a 2'-deoxyguanosine riboswitch". In: *Nature Chemical Biology 2011 7:10* 7.10 (Aug. 2011), pp. 748–755.

[105] Peter J. Lukavsky. "Structure and function of HCV IRES domains". In: *Virus Research* 139.2 (Feb. 2009), pp. 166–171.

[106] Sergey M Dibrov et al. "Functional Architecture of HCV IRES Domain II Stabilized by Divalent Metal Ions in the Crystal and in Solution". In: *Angewandte Chemie International Edition* 46.1-2 (Jan. 2007), pp. 226–229.

[107] Ryan B. Paulsen et al. “Inhibitor-induced structural change in the HCV IRES domain IIa RNA”. In: *Proceedings of the National Academy of Sciences of the United States of America* 107.16 (Apr. 2010), pp. 7263–7268.

[108] Sergey M. Dibrov et al. “Structure of a hepatitis C virus RNA domain in complex with a translation inhibitor reveals a binding mode reminiscent of riboswitches”. In: *Proceedings of the National Academy of Sciences of the United States of America* 109.14 (Apr. 2012), pp. 5223–5228.

[109] Erin Noble et al. “Biophysical Analysis of Influenza A Virus RNA Promoter at Physiological Temperatures”. In: *Journal of Biological Chemistry* 286.26 (July 2011), pp. 22965–22970.

[110] Mi Kyung Lee et al. “A novel small-molecule binds to the influenza A virus RNA promoter and inhibits viral replication”. In: *Chemical Communications* 50.3 (Dec. 2013), pp. 368–370.

[111] Mi Kyung Lee et al. “A single-nucleotide natural variation (U4 to C4) in an influenza A virus promoter exhibits a large structural change: implications for differential viral RNA synthesis by RNA-dependent RNA polymerase”. In: *Nucleic Acids Research* 31.4 (Feb. 2003), pp. 1216–1223.

Chapter 4

Conclusions and perspectives

This thesis addresses a critical lack in the state-of-the-art approaches for structure-based rational design of small molecules targeting RNA. While the elusive structural dynamics of RNA molecules constitutes an impediment to their targeting, computational methods enable leveraging this property toward the development of unique therapeutic strategies (Sec. 1.5.2). Nonetheless, in the landscape of available computational tools, no well-established approach was comprehensively addressing the challenges stemming from the flexibility of RNA molecules. Overcoming this limitation would enhance the relevance of RNA-targeted therapeutics, which is currently limited to a small number of available medicines. In this direction, my PhD project was designed to develop an advanced computational framework to enhance the role of computational methods in aiding the identification of small molecules targeting RNA.

A starting point for any structure-based approach is the utilization of the available structural information. The analysis of known examples of RNA-small molecule interactions may be informative toward the rational design of therapeutic agents. In this sense, the curation of comprehensive repositories collecting all the available structures of RNA and ligands would be of primary importance. Mostly due to the challenges faced by experimental and computational methods in RNA structure determination (Sec. 1.4), the number of available RNA structures is very limited. Most importantly, curated repositories collecting the structural properties of RNA-ligand complexes were missing at the beginning of the project. To fill this lack, the first part of my PhD was dedicated to the creation of HARIBOSS, a curated database of RNA-small molecule structures retrieved from the PDB. HARIBOSS is available *via* a dedicated web interface and is regularly updated with all the structures resolved by X-ray, NMR, and cryo-EM, in which ligands with drug-like properties interact with RNA molecules. Each HARIBOSS entry is annotated with the physicochemical properties of ligands and RNA pockets. HARIBOSS will facilitate RNA drug discovery for two principal reasons:

1. it constitutes a repository of RNA-small molecules that can be used to benchmark computational tools as well as a training set for machine-learning models in the context of structure prediction, binding site detection, and molecular docking;

2. the analysis of HARIBOSS entries will contribute to the understanding of the molecular mechanism of interaction underlying RNA-small molecule recognition.

Along with the release of HARIBOSS, a thorough analysis of the available RNA-small molecule structures was conducted. The findings revealed that the majority of RNA binding pockets are not suited for interaction with drug-like molecules (Fig. 2.3). This inadequacy was primarily attributed to the lower hydrophobicity and increased exposure to solvent exhibited by RNA cavities compared to the binding sites typical of proteins. Accordingly, RNA binders deposited in the PDB are on average highly polar and possess sub-optimal pharmacological properties, especially in terms of polarity and cell permeability (Fig. 2.2). These outcomes align with the picture introduced in Chapter 1, portraying RNA as a flexible and elusive target. To what extent RNA binding pockets can accommodate effective therapeutic agents remains to be understood. The recent FDA approval of risdiplam holds significant promise for the future, demonstrating the feasibility of developing drug-like compounds that selectively and effectively bind pockets lacking the buried character of protein pockets. However, the current number of deposited RNA-small molecule structures is insufficient for a comprehensive assessment of the druggability of RNA molecules and of the drug-likeness of RNA binders. Additionally, the PDB database, and HARIBOSS by consequence, do not encompass a significant portion of data from pharmaceutical industries, which are currently at the forefront of RNA-targeted drug discovery. Finally, the drug-likeness itself has not a unique definition and the upcoming discoveries elucidating the mechanisms of recognition between RNA and small molecules may lead to a shift of what is currently intended for a drug (Sec. 1.3.3). In less than two years since its release, the number of structures in HARIBOSS has increased by 17.5%, and it is reasonable to anticipate that future years will provide further insights into the physicochemical nature of the compounds able to target RNA.

From a broader perspective, HARIBOSS analysis underscored the fact that evaluating the druggability of RNA targets using criteria established for proteins may yield an incomplete portrait. First, HARIBOSS database is currently composed only of static structures. This depiction of RNA molecules as static entities intrinsically overlooks their dynamic nature and may not fully capture their interaction mechanism. Furthermore, the assessment of the properties of RNA pockets relied on computational tools designed to identify and evaluate binding sites on such static structures. The reliability of these tools to accurately capture the physicochemical characteristics of the binding pockets of the flexible RNA targets remains uncertain. A notable example is found in the evaluation of physical properties, such as the pocket volume (Fig. 2.3), which consistently yielded estimates excessively large compared to the average size of small molecules across various software platforms. Such inaccurate characterization of binding sites may be useless toward the development of an effective approach to identify small molecule binding sites on RNA. One of the main improvements of the future versions of HARIBOSS will therefore be the inclusion of structural ensembles representing the highly dynamic nature of RNA. However, the accurate modeling of RNA flexibility in the detection of unknown RNA binding sites is beyond the reach of the mentioned tools. While MD simulations offer a natural framework to comprehensively describe RNA structural dynamics (Sec. 1.4.4), no existing tool for binding site identification was explicitly designed to leverage the mentioned therapeutic opportunities introduced by the exploration of RNA conformational landscapes. To overcome this critical limitation, the main work of my PhD consisted in the development of SHAMAN, a computational technique to identify potential small-molecule binding sites in RNA structural ensembles. SHAMAN uniqueness relies on its ability to perform at the same time:

- the exploration of the conformational landscape of the target RNA with atomistic Molecular Dynamics (MD) simulations;

- the identification of potential binding pockets by means of small probe compounds that explore the RNA surface with the aid of enhanced sampling techniques.

SHAMAN constitutes an important achievement in the field of RNA-targeted drug discovery. Indeed, it is able to provide accurate predictions of experimentally-verified binding sites in both cases of large and stable riboswitches molecules and more flexible viral RNAs (Fig. 3.2). More importantly, SHAMAN predictive accuracy has been demonstrated in realistic drug discovery scenarios, where the target molecule structure is available only in apo conformations and with no prior information on potential binding sites. In this regard, from the analysis of riboswitches, it is indeed evident how SHAMAN can accurately predict alternative binding modes after local structural rearrangements. Notably, the analysis of viral RNAs highlighted how accounting for RNA flexibility is crucial to identify binding pockets formed upon large structural rearrangements. Overall, the architecture of SHAMAN enables the detection of interacting hotspots that are difficult or invisible to other approaches based on static structures (Fig. 3.4). The reliability of predictions obtained with SHAMAN holds promise for the step that follows binding site identification, that is the identification of potential RNA binders. Indeed, future users of SHAMAN will be able to restrict virtual screening campaigns to the few interacting regions that were identified as most probable in the conformations explored by the target RNA. Moreover, the probability densities of the different probes that identified a given binding site (SHAMAPs, Sec. 3.4.1) may be used by docking software to obtain a more accurate pose identification. Future developments of the method will explore the integration of SHAMAN predictions with molecular docking engines, in order to identify potential binders for a given RNA target and to predict their complex structure. The integration of advanced computational approaches optimized for specific tasks establishes a framework to effectively address the challenges of drug discovery as a whole. This holds promise for enhancing the relevance of RNA-targeted therapeutics in the near future.
The ultimate goal of a drug discovery campaign is the development of compounds that are able to bind RNA with high selectivity and specificity. While the characterization of the latter binding property is currently exclusive of experimental techniques, computational methods have the potential to aid the discovery of selective RNA binders. The tools that I developed do not explicitly address this challenge and the difficult endeavor currently relies on the subsequent step of virtual screening: preferential binding patterns of a given compound in a pocket identified by SHAMAN can be highlighted by the screening against different targets. However, it is important to highlight that the mentioned ultimate goal can not be effectively achieved without first identifying pockets with the potential to bind small molecules. SHAMAN was designed to carry out this operation by exploring the conformational landscape of RNA molecules, addressing a first urgent need for drug discovery. Furthermore, it is possible to leverage the information currently given by SHAMAN toward the design of selective compounds. Indeed, the exploration of the conformational space of a target RNA may implicitly give information for the selective binding of small molecules. As discussed in Sec. 1.5.2, a potential binder is susceptible to engage more favorable intermolecular contacts with a specific conformation of the RNA target, thereby enhancing the selectivity of the binding with respect to other conformations.

In order to explicitly address the issues related to selectivity, it is possible to foresee the main challenges to be addressed in future implementations of the computational tools that I developed during my PhD. First, the available information about binding affinity, selectivity, and specificity of HARIBOSS ligands needs to be included in the database. Such information will be fundamental in the development and benchmark of advanced computational tools aimed at the prediction of the binding properties. Moreover, it will allow the identification of chemical fragments related to specific binding properties and, ultimately, the definition of multiple sets of probes to be used in more tailored applications of the SHAMAN protocol. Indeed, a current main limitation of SHAMAN is that the set of probes used to identify binding sites was extracted from the HARIBOSS database, without insights into the selectivity for specific RNAs. Moreover, as discussed before, the exploration of the conformational space may be informative of selective mechanisms of molecular binding. Since the current implementation of SHAMAN relies on unbiased MD to explore the conformational space of the target RNA, future developments will need to introduce enhanced sampling techniques to perform more exhaustive simulations.

An ultimate aspect concerning the future is whether MD simulations will retain their role in drug discovery. Currently, this technique constitutes one of the most advanced computational approaches to characterize the recognition phenomenon between two biomolecules. The strength of SHAMAN itself relies on the state-of-the-art of RNA atomistic force fields for MD simulations (Sec. 1.4.4). Nevertheless, the computational expenses required for the comprehensive development of simulation-based methods can be impractical for drug design or beyond the reach of most of the scientific community. Oppositely, methods based on machine learning (ML) models, and especially deep learning, constitute a faster and powerful alternative approach whose influential role will be inevitable in future applications of Computer Aided Drug Design (CADD). First, the accuracy of RNA structure prediction is anticipated to approach the achievements seen in protein structure prediction. Subsequently, it is plausible that ML will soon demonstrate its capability to predict the complex structure of a macromolecule receptor with a specified ligand, constituting a complete alternative to both binding site detection and molecular docking. However, ML methods have important limitations. A first one concerns their strong dependence on the availability of the training data, which are currently very limited for RNA-ligand structures. As already highlighted, the role of HARIBOSS in collecting and optimally organizing the available structural information about RNA-small molecule interactions will be precious. Independently of the forthcoming increase in resolved RNA-small molecule structures, a second challenge arises in the effectiveness of ML methods in accounting for the conformational flexibility of RNA targets in their recognition with small molecules. This challenge primarily arises due to the intricate demands on feature engineering and model implementation in order to capture the thermodynamics and kinetic properties of the binding. To this end, MD is likely to maintain a guiding role, at least in the near future. Upcoming studies and applications will be crucial in elucidating the respective contributions of ML and MD. An important possibility is the synergistic combination of the relative main strengths, already implemented for example in the force field parametrization and in the choice of collective variables for enhanced MD simulations. Despite these uncertainties, the predominant outlook is that the field of RNA-targeted drug discovery will soon witness significant advancements through the application of advanced structure-based computational methods.

Appendix A

Running the SHAMAN pipeline on HCV IRES RNA

The SHAMAN pipeline presented in Chap. 3 will be soon available in a dedicated GitHub repository. The repository is organized in the following directories:

- `PYTHON`, containing the necessary python scripts for pre-processing, post-processing and analysis of SHAMAN simulations;
- `BASH`, containing a series of bash scripts that run the entire SHAMAN pipeline in an automated way;
- `forcefield`, containing all the necessary force field files that I used in SHAMAN simulations;
- `mdps`: the .mdp GROMACS instruction files for the different stages of SHAMAN pipeline;
- `tutorial`: a step-by-step example of a typical SHAMAN run.

In this Appendix, I report a step-by-step tutorial to guide future users of SHAMAN, constituting the text of `README.md` file contained in the `tutorial` directory. All the reported paths refer to the `tutorial` directory, which is replaced by a generic `DATA` directory in the text.

Step-by-step tutorial to run SHAMAN

Hello, world!

We report here a step-by-step tutorial to run the SHAMAN pipeline (Fig. 3.1), gathering all the information that you need to run our protocol with detailed remarks and warnings on the crucial steps. The system examined in this tutorial is the Hepatitis C Virus Internal Ribosome Entry Site (HCV IRES), as resolved in the 3tzr PDB entry. HCV IRES is essential for the synthesis of the HCV proteins, promoting its proliferation in cells, and therefore constitutes an interesting target for drug design (Sec. 3.5.1 and Fig. 3.11). For this tutorial, we selected a set of 4 probes, 2 aromatic (benzene and benzotiophene, which we will call BENX and BETH respectively) and 2 aliphatic (formamide and methyl ammonium, which we will call FORM and MAMY).

0 - Preliminaries

SHAMAN input consists of the 3D structure of the RNA target in PDB format and the SDF file of the selected probes. Both require an external pre-processing, detailed below.

WARNING: Since its creation, SHAMAN strongly relies on a pre-defined directory architecture (Fig. A.1, which that must be carefully enforced. Any modification may cause potential errors, so please be aware!

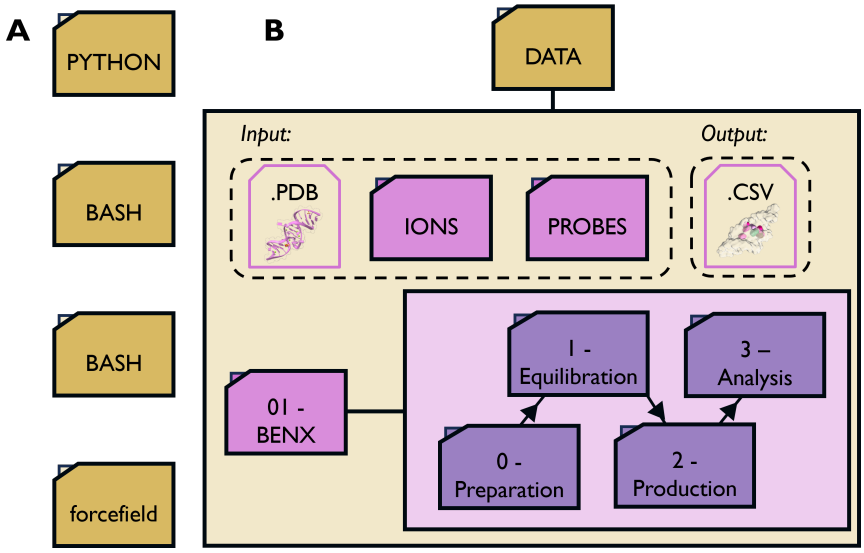


Figure A.1: The directories of SHAMAN pipeline. A) The directories of the SHAMAN GitHub repository. **B)** The content of the `DATA` directory. The input inset includes the PDB file of the target RNA (here PDB 3tzr), the directory `IONS` and the directory `PROBES`. The output inset includes the .CSV file with the until the prediction of the binding sites defined by the SHAMAPs. On the bottom, the content is progressively added in one of the probe directories (01-BENX): from left to right, the directory of the preliminary stage (0-Preparation), the directory of the equilibration stage (1-Equilibration), the directory of the production stage (3-Analysis). Elements that belong to the same directory are shown with the same color.

0.1 - Preprocessing the RNA

For the RNA molecule, the following steps need to be performed:

- Removal of any ligand, if present;
- Removal of all crystallized water molecules;
- Removal of the $PO3$ groups (P, $O1$, $O2$ atoms) eventually present at the beginning of the PDB file;
- Removal of any ion that is not modeled by the force field and listed in the `../forcefield/amber_na.ff/ions.itp` file in the `forcefield` directory;
- Addition of eventually missing hydrogen atoms at a given pH;
- Creation of a PDB file of the RNA target in which there are no ions.

REMARK The removed ions will be treated as our probes using the openFF force field and a separated SDF file with their coordinates in the PDB file is needed (header of `../BASH/1-Preprocessing.sh` for more details). This SDF file can be downloaded directly from the PDB database.

In this tutorial, we start from `./3tzr.pdb`, the Xray structure of the HCV-II IRES RNA deposited in PDB database bound to the ligand SS0, and resolved with 3 $SO4$ ions and 6 Mg ions. We first generate the `./3tzr_apo.pdb` file, where, as anticipated, we remove the $SS0$ ligand, the $SO4$ ions and the crystallized water molecules and then we add the H atoms at pH=7.4 with Chimera. The instance coordinates of the $SO4$ ions are downloaded from the 3tzr PDB entry and moved into the `./IONS` directory. Finally, we create a `./3tzr_apo_noions.pdb` file where we remove all ions.

0.2 - Preprocessing the probes

The user should choose the set of probes and supply an SDF file of each probe, prepared at a given pH. All these files must be placed in the `./PROBES` directory, each one numbered uniquely starting from 01. Here, we have prepared at pH=7.4 the `./PROBES/01-BENX.sdf`, `./PROBES/02-BETH.sdf`, `./PROBES/03-FORM.sdf`, and texttt04-MAMY.sdf probes. In addition, the `./PROBES/01-BENX.sdf` probe comes from a fragmentation protocol applied to known RNA binders, as explained in the SHAMAN paper.

1 - Input stage

As a first step, we generate the topology files for the subsequent production. This is carried out by the script `../BASH/1-Preprocessing.sh`, to be run from the `DATA` directory independently for each replica. A detailed description of how the script works can be found in its header. The input arguments of the script are, in order: the name of the PDB file of the system, the number name of the replica and the number of ions in the `IONS` directory.

REMARK The third argument, in the current set-up of SHAMAN that inserts ions only in the mother simulation, is relevant only for the 00-APO replica.

Here, we first generate the mother system 00-APO in which there will be only the RNA molecule with all the structural ions. Therefore, we will indicate the presence of the 3 $SO4$ ions initially present in 3tzr.

```
source ../BASH/1-Preprocessing.sh 3tzr 00-APO 3
```

Second, for each probe we run the same command, for example 02-BETH

```
source ../BASH/1-Preprocessing.sh 3tzr 02-BETH
```

At the end of this stage, you will find a subdirectory `0-Preparation` generated for all the replicas (Fig. A.2A). The file necessary for the next step will be:

- **system.top**: GROMACS topology file;
- **system_box_water_ions.gro**: GROMACS structure file of the system;
- **XTC_system.ndx**: GROMACS index file with the atoms whose coordinates will be included in the compressed output trajectory (the RNA plus the probe and ions, if present);
- **probe.sdf**: a processed SDF file with the probe conformer.

WARNING You may want to check that the topologies are correct. For example, in the mother `system.top` file we can see that the system contains the two RNA chains and the MG and $SO4$ ions, together with the water molecules (SOL) and the counterions KCL (Fig. A.2A). On the other hand, the system.top for the 02-BETH system, contains the RNA, the BETH probe and the solution without the structural ions (MG and $S04$, Fig. A.2B).

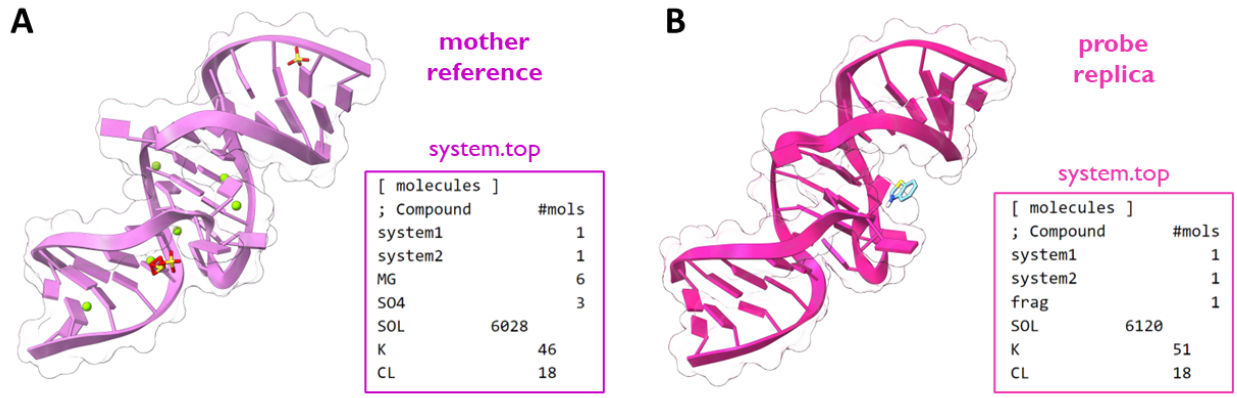


Figure A.2: Difference between mother and shadows topologies. **A)** Representation of the main elements in the topology of the mother system: the RNA chains, and the structural ions. In the lower right, the box reports a screenshot of the `system.top` file. **B)** Same as panel A, for the 02-BETH replica system.

2 - Production stage

2.1 - Pre-Production

The systems need to be prepared for the Production stage: independently for each replica we need to equilibrate the system and generate the PLUMED input file, necessary to perform the mother-replica shadowing. These operations are performed by the `../BASH/2.1-Pre-Production.sh` script, to be run from the `DATA` directory independently for each replica. A detailed description of how the script works can be found in its header. The input arguments are, in order: the numbered name of the replica, and the number of steps for NPT and NVT equilibrations.

Here, we execute a 10 ns-long NPT and NVT equilibration (5000000 steps of 2 fs each) for the 02-BETH replica:

```
source ../BASH/2.1-Pre-Production.sh 02-BETH 5000000
```

The same operation needs to be performed for all the replicas (00-APO included!) and you might want to adapt the run depending on the hardware architecture. At the end, you will find a `1-Equilibration` directory (Fig. A.1B) with all the I/O files of the equilibration procedure and a `2-Production` directory with the following files necessary for the next step:

- **shaman.tpr**, GROMACS binary file with the compiled instructions for the production;
- **conf_emin_PLUMED.pdb**, a PDB file needed by PLUMED;
- **plumed.dat**, PLUMED input file for production.

WARNING You may want to check that the PLUMED input files are correct, especially for what concerns the differences between mother and probe systems (Fig. A.3A). The `plumed.dat` file for the mother system, beside defining the RNA backbone atoms, is setting the system as the reference for shadowing (Fig. A.3B). For the replica systems, as in the case of 02-BETH, the `plumed.dat` file is reporting the following additional instructions: *i)* align on-the-fly the RNA backbone atoms to a reference (solid box in Fig. A.3C); *ii)* applying metadynamics to the center of mass of the probe (dashed box in Fig. A.3C); *iii)* apply a restraint to keep the probe close to RNA (dotted box in Fig. A.3C); *iv)* apply a restraint to the RNA backbone atoms to follow the evolution of the mother simulation (dotted box in Fig. A.3C).

2.2 - Production

SHAMAN production is launched using the `../BASH/2.2-Production.sh` script, to be run from the `DATA` directory. The production stage is run in parallel for all systems (*–multidir* option in `gmx mdrun`). The argument of the script is the number of steps for the production.
For the sake of this exercise, we perform a short production run of 200 ns (100000000 steps):

```
source ../BASH/2.2-Production.sh 100000000
```

At the end of the production, you will find in the `2-Production` subdirectory (Fig. A.1B) of each replica (referred as the i-th) the following files necessary for the next steps:

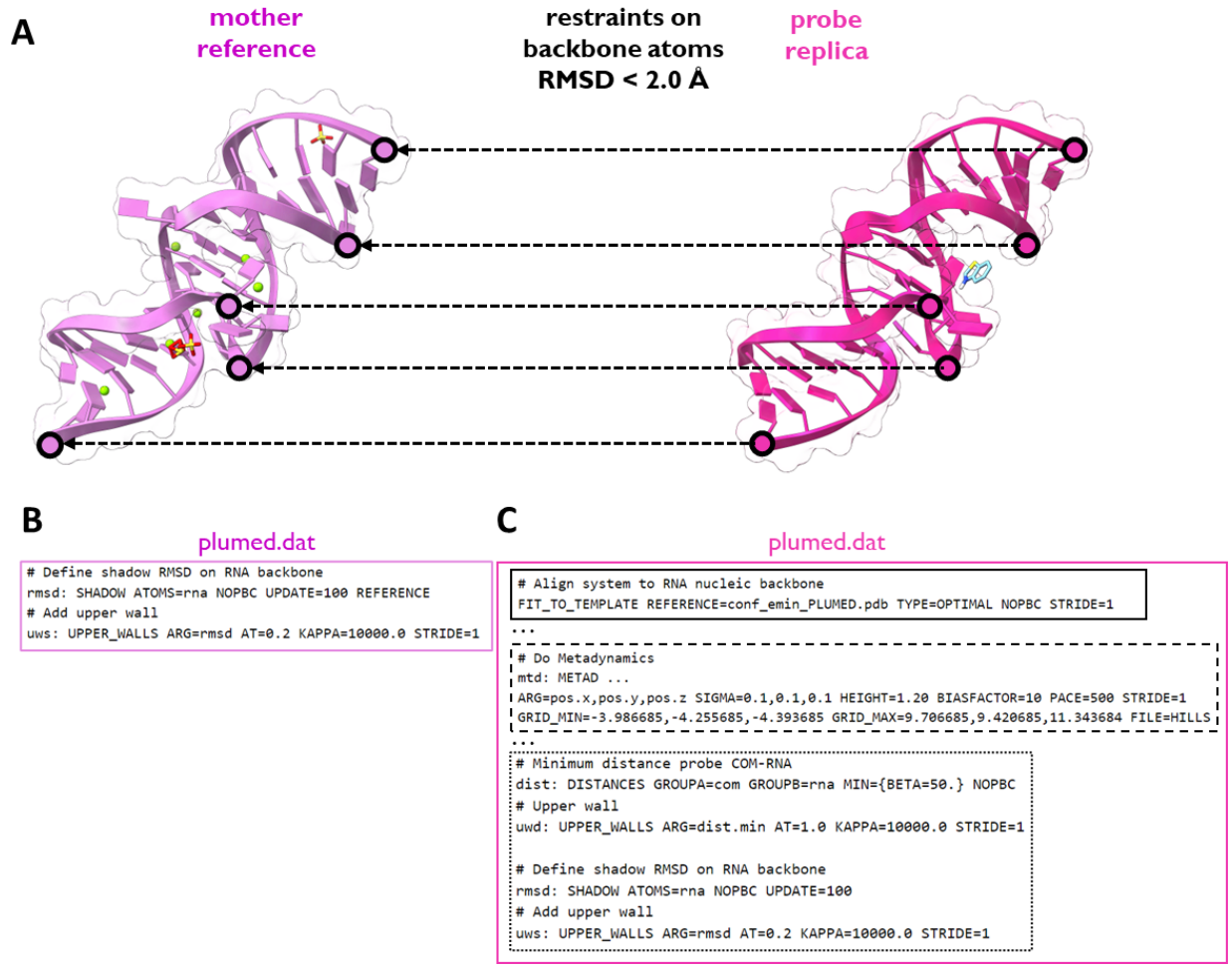


Figure A.3: Difference between PLUMED input files of mother and replica systems. A) A diagram representing the restraint on the RNA backbone atoms that is set between the mother reference system and all the probe replica systems. **B)** PLUMED input file of the mother system, setting it as the reference for shadowing. **C)** PLUMED input file for 02-BETH, showing the on-the-fly alignment of RNA to a reference conformation (solid box), the introduction of the metadynamics bias (dashed box), and the applications of the restraints to the probes and RNA atoms (dotted box).

- **shaman.xtc**, the compressed trajectory file with the coordinates of the RNA, the probe, and the eventual ions,
- **COLVAR.i**, the PLUMED output file with details on probe position;
- **HILLS.i**, the PLUMED output file with details on metadynamic bias.

2.3 - Post-Production

We now need to fix the periodic boundary conditions and align the system to a reference structure for each simulation, and to reweight the frames of the probe replica simulations that have been accelerated by the metadynamics biasing potential. This is realized by the `../BASH/2.3-Post-Production.sh` script, to be run independently for each replica from the `DATA` directory.
Here we run it for the 02-BETH replica,

```
source ../BASH/2.3-Post-Production.sh 02-BETH
```

The operation needs to be performed for each replica (00-APO included!). At the end, you will find in the `2-Production` subdirectory of each replica:

- **traj-pbc-PLUMED.xtc**, a compressed trajectory file with RNA, probe, and eventual ions with PBC-fixed coordinates aligned to the initial structure;
- **weights.dat**, a text file indicating, for each frame (row), the corresponding metadynamics weight.

3 - Analysis stage

3.1 - RNA clustering

We now identify the conformations that the target RNA explored during the simulation. This analysis is performed on a concatenated trajectory, which takes into account the RNA conformations sampled by all replicas. The clustering is performed using the `../BASH/3.1-Clustering.sh` script, to be run from the `DATA` directory upon indicating the total number of replica. Here, we run it for 5 replica (4 probes + apo) and with RMSD cutoff of 0.3 nm:

```
source ../BASH/3.1-Clustering.sh 5 0.3
```

At the end, you will find the `clstr_analysis` directory has been created, with the following files:

- **traj-cat-conv.xtc**: the concatenation of the `traj-pbc-PLUMED.xtc` trajectories for all the systems;
- **cluster_cat_conv.log**: the GROMACS output file (from `gmx cluster` routine) with details of the cluster members and populations;
- **i_clstr_center-all.pdb**: a PDB file with the structure of each cluster center.

REMARK The PDB files will be necessary for the visualization of the pockets (Fig. A.4), since their coordinates are used as reference for the density maps. The trajectories are not necessary for the following SHAMAN analysis, but they may be interesting for the study of the conformational space explored by the RNA.

In the `2-Production` subdirectory of each replica of the system, you will find:

- **cluster_log.dat**: a text file with the list of RNA clusters explored with the respective population and cluster center (frame index with respect to `traj-cat-conv.xtc`);
- **cluster_traj.dat**: a text file with the RNA cluster index of each frame of the replica trajectories (from the `traj-pbc-PLUMED.xtc` file).

REMARK In most cases, due to the architecture of SHAMAN, the RNA conformations explored by the different probe replicas will be the same, possibly with different populations. This is due to the fact that during the simulation replicas can deviate from the configuration of the mother by at most 0.2 nm. As you can see in the `cluster_log.dat` file for 02-BETH, the global population of conformation 3 is the 1.5%, but, in presence of the probe and relatively only to its sub-trajectory, the effective population reached 22%.

Free Energy Grid analysis

Now, on the most populated conformations identified in 3.1 and independently for each probe, we will define a grid in the 3D space and compute, for each voxel, its occupancy and free energy (FE). This quantifies the probability that a probe is found in a certain region rather than in the bulk solvent. Clustering of the voxels will be performed to identify probable binding hotspots explored by each probe.
The FE analysis is performed by the `../BASH/3.2-FEG-analysis.sh` script, to be run independently for each replica from the `DATA` directory, giving as input also the population threshold for an RNA cluster to be considered in the analysis.

Here, we run the script for the 02-BETH replica, taking into account only the RNA clusters with populations greater than 10%:

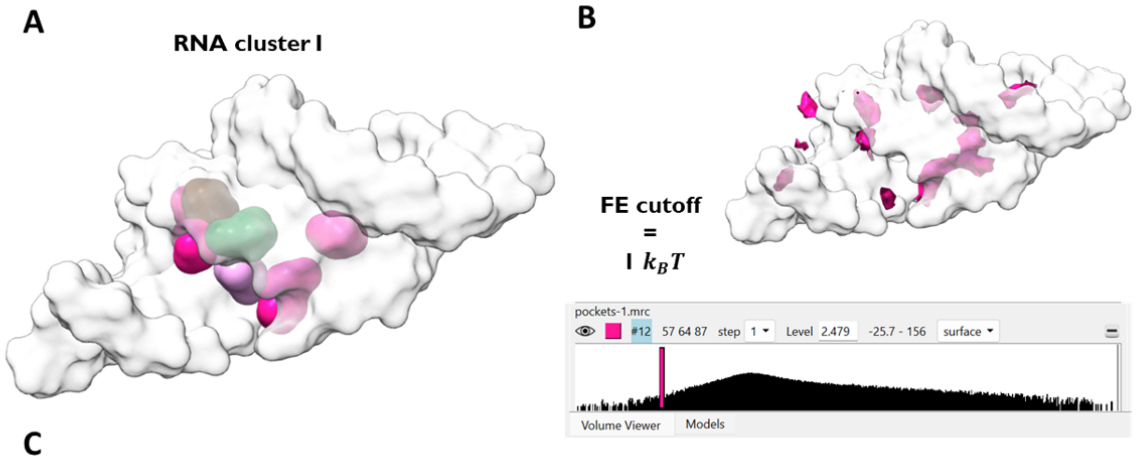
```
source ../BASH/3.2-FEG-analysis.sh 02-BETH 0.1
```

REMARK In order to take into account the effect induced by the presence of the probes, the population threshold refers to the sub-trajectory of each replica.

At the end, the `3-Analysis` directory (Fig. A.1B) has been created for each replica of the system, with the following output files:

- **traj-cluster-i.xtc:** the sub-trajectory corresponding to the i-th RNA cluster,
- **pockets-i.dat:** a text file with the list of the sites explored by the probe, for each RNA cluster, with the corresponding free energy (plus error), and the estimates of the pocket volume and buriedness,
- **pockets_df-i.json:** the same information but stored in a pandas DataFrame,
- **pockets-i.mrc:** an MRC file with the free-energy grid of the probe for the i-th conformation,
- **pockets_coordinates:** a directory with the coordinates of the identified sites (nomenclature: increasing numbers as pocket identifiers, together with ID of the RNA cluster, i.e., 0-1 means pocket 0 in cluster 1, 1-1 means pocket 1 in cluster 1, etc ...). The files here are:
  - **.npy:** numpy binary files with the xyz coordinates of the voxels belonging to this interacting site,
  - **.xyz:** a dummy coordinate file with the xyz coordinates of the voxels,
  - **_weights.npy:** numpy binary file with the free-energy grid.

In our example, we can see that, on the most populated RNA cluster 1 (Fig. A.4A), the probes 01-BENX (dark brown surface), 02-BETH (light violet surface), and 03-FORM (dark green surface) explored 1 potential interacting site, while 04-MAMY explored three (pink surfaces).



C

	best_DG	DDG	clstr_id	clstr_pop	delta	probe	pocket_ID	rel_pop	DG	sigma	buriedness	volume	g	gyr_p
0	-30.452783	0.000000	1	0.965	7.31306	[02-BETH]	[0]	[0.764]	[-30.4527834216]	[None]	[2.214151136]	[21.0]	[[27.665876733594786, 24.29681738296652, 31.90...	[2.0018638101]
1	-30.301989	0.150795	2	0.017	7.31306	[03-FORM]	[0]	[0.118]	[-30.3019885595]	[None]	[0.9040161181]	[5.0]	[[23.448200463692828, 22.20666495967819, 35.50...	[1.2759057289]
2	-30.289039	0.163745	1	0.965	7.31306	[01-BENX, 04-MAMY]	[0, 2]	[0.822, 0.676]	[-30.2890388933, -22.2795408129]	[23.2079869974, 18.3885379859]	[4.0267282165, 2.8860569418]	[28.0, 5.0]	[[24.164161606011631, 30.854111659962765, 29.19...	[1.4963238539, 0.8245714192]
3	-30.244509	0.208274	3	0.015	7.31306	[02-BETH]	[0]	[0.218]	[-30.2445092597]	[0.98473231411000001]	[1.2458801111]	[38.0]	[[37.137622257683887, 30.322396428432857, 29.05...	[2.0025075327]
4	-30.151516	0.301267	1	0.965	7.31306	[03-FORM]	[0]	[0.768]	[-30.1515159491]	[12.3841691001]	[2.6151871771]	[12.0]	[[23.756701654992181, 23.453331544431665, 35.670...	[1.4356272256]
5	-29.802734	0.650049	2	0.017	7.31306	[04-MAMY]	[1]	[0.183]	[-29.8027339841]	[None]	[1.7565061307]	[5.0]	[[29.882115549745407, 25.532956358690306, 42.4...	[0.779722495]
6	-29.107830	1.344954	1	0.965	7.31306	[04-MAMY]	[0]	[0.676]	[-29.1078297229]	[1.8338810173]	[2.3891098425]	[9.0]	[[29.606229057046377, 25.761756680630175, 42.6...	[0.9599541893]
7	-28.470948	1.981836	2	0.017	7.31306	[01-BENX]	[1]	[0.113]	[-28.4709476034]	[None]	[1.0318083277]	[8.0]	[[16.293345465131367, 32.95075800947035, 31.21...	[1.4904760195]
8	-28.135681	2.317103	2	0.017	7.31306	[01-BENX]	[0]	[0.113]	[-28.1356806339]	[1.563994582]	[4.4515102976]	[12.0]	[[23.146087724454 15, 31.387418543559786, 29.69...	[1.6113398003000001]
9	-26.662062	3.790721	1	0.965	7.31306	[04-MAMY]	[3]	[0.676]	[-26.6620624485]	[14.269662431]	[4.0015473436]	[9.0]	[[22.514904375052414, 23.187574118184457, 30.1...	[1.2445034128]
10	-25.934194	4.518589	2	0.017	7.31306	[04-MAMY]	[0]	[0.183]	[-25.9341939556]	[None]	[1.9772890397]	[7.0]	[[28.04656045584604, 27.315491019480167, 26.19...	[1.1932080250000001]
11	-24.290346	6.162437	1	0.965	7.31306	[04-MAMY]	[1]	[0.676]	[-24.2903459836]	[2.2419508436]	[2.675871688]	[13.0]	[[32.190751692627906, 21.53719979484257, 35.11...	[1.5846867953000001]

Figure A.4: Visualization of interacting sites and SHAMAPs. **A)**The most populated RNA cluster together with a surface representation of the interacting sites explored by the 4 probes. Probe colors are defined in our SHAMAN manuscript. **B)** Visualization of the MAMY mrc file on cluster 1. In the lower panel, the VolumeViewer tab of ChimeraX, showing how users can slide across all free energy values and visualize the corresponding explored regions. **C** Visualization of `shaman.csv` and its content (see text for a detailed description).

REMARK The files in the `pockets_coordinates` directory come from clustering the probe free-energy map reported in the .mrc file (cfr MAMY results in purple between Fig. A.4A and Fig. A.4B). The user can visualize these files with ChimeraX (Fig. A.4B) and explore the density at the desired free-energy value. Setting the threshold at +10 kJ/mol from the global minimum shows the regions that will be clustered in the final sites. By looking at higher or lower free-energy thresholds (1 K_BT in the example of Fig. A.4B), the user can inspect all the regions explored by MAMY up to the given FE level.

Advice for ChimeraX users To visualize the files, you need to open the "Volume Viewer" utility (Tools > Volume Data). The slide acts on the free energy distribution. We report to our users a current bug of ChimeraX with .mrc files (v. 1.5 2022-11-24). To correctly visualize the file, you will need to *i)* open the "Surface Zone" utility (Tools > Volume Data), *ii)* select your system, *iii)* apply a reasonable "Radius" (ex. 10).

4 - Output stage

4.1 - Clustering of results of all probes

To obtain the final output of our pipeline, the SHAMAPs, we will cluster all the interacting sites identified by all on a given RNA cluster. Such operation constitutes one of the main and unique characteristics of our approach, made possible by the parallel exploration of a similar RNA conformation simultaneously by all the probes.
The final clustering is performed by the `../BASH/4-SHAMAPs.sh` script, which we run from the `DATA` directory:

```
source ../BASH/4-SHAMAPs.sh
```

At the end, you will find in the 4-Output directory, the following files:

- **shaman.csv**, a CSV file (Fig. A.1B) with the final list of SHAMAPs obtained on each RNA cluster;
- **DF.csv**, a CSV file with the pandas DataFrame containing all the sites identified by all the probes. This file is the concatenation of the `pockets_df-i.json` files generated in the stage 3.2.

An example of the final table `shaman.csv` is reported in Fig. A.4C. You can see that, in the third best free-energy SHAMAP (index = 2), two sites explored by BENX and MAMY were clustered together. The SHAMAPs are ordered by the lowest free-energy value associated to the probe sites composing the SHAMAP ('best_DG' in Fig. A.4C) and annotated with:

- the difference in free energy with respect to the top-scored SHAMAP ('DDG');
- the index of the corresponding RNA cluster ('clstr_id');
- and its population ('clstr_pop');
- the typical error on free energy ('delta').

In addition, each SHAMAP entry is supplied, in the format of a list, with the properties of the constituent sites:

- the probes that explored them ('probe');
- the ID of the sites ('pocket_ID', as numbered in the output of 3.2);
- the population of the RNA cluster relative to the probe subtrajectory ('rel_pop');
- their corresponding estimated free energy ('DG');
- the associated error ('sigma');
- their exposure score ('bur');
- the volume ('volume');

- the coordinates of the free-energy weighted geometrical centers ('g');
- their gyration radii ('gyr_p').

The visualization of a given SHAMAP is realized by opening all the corresponding pocket coordinates files, as reported in the `shaman.csv` file (Fig. A.4).

List of Figures

List of Tables

Acknowledgements

My PhD experience has been, well before science, what allowed me to go through one of the most intense and important growths of my life. Looking at these two badge photos, now three years old, of the first days I entered the Institut Pasteur in Rue du Docteur Roux and Sanofi in Chilly Mazarin gives me a strong, clear sensation: a lot of things have changed. My world is richer in heterogeneity, shining, and full of happiness. It is in this perspective that I wanted to have the following final words on my PhD. Now that it is over, I feel able and also wishful to thank all the people who contributed to this growth. Every little thing I experienced with each one of you helped shape the person I would like to be and the one that I would not. Therefore, I want to say a word for each one of you. Even if no one will read it entirely, I also decided to write the entire section in this universal language. I am speaking to myself who will read these few pages in some years: this is you, at the age of 28, after your PhD and you want everyone to be able to share with you this final moment.

The first people I would like to thank, since they made the aforementioned growth possible, are my supervisors, Max and Evi. I leave with an extended scientific baggage of at least one order of magnitude compared to when I arrived. I am truly grateful for the opportunity I had and for all I learned during these years. I truly appreciated the dedication that Max put into revising this manuscript, sometimes against external conditions. Most of all, I am glad for the possibility to meet and exchange with the scientific community all around the world: being ready to discuss, defend, and finally, share our ideas with so many different scientists is for sure the most exciting part of research. I am sorry we never managed to harmonize our human souls and I hope what we went through will be useful in the future for all of us.
My PhD has been divided between two labs. I will start with Institut Pasteur, which has effectively been my home for these three years. Thanks to Michael for initially welcoming me into his unit, Arnaud for his propensity to share his knowledge and his funny way of compensating for my carelessness and avoiding food waste, Guillaume for his unconditional availability and his elegant and humble way of transferring his vast expertise, Nathalie for her kindness in sharing with me almost anything I asked and for not being too mad in front of the Monday dishes to clean, and Tru for his professional support in any task that should be carried out. I would also like to thank Nadia for the sweetness she always communicated to me even if we never had the opportunity to talk much. A special thanks goes to Marcel and Guillaume, my tutors during these years, who supported me and gave me advice in many difficult moments. Thanks also to Germano for all the time we spent talking about life and music and for the lucky and unforgettable occasion in which we had the opportunity to meet. Above all, these years would have been too hard without Maxi and Sam. Thank you for sharing with me so much humanity, science, fun, in one word, life. The lunches and discussions on

François Jacob's roof with Maxi will for sure be among the things I will miss the most. Among the others, I can't avoid mentioning the silly questions of Sam in the late (or early?) hours of the night, whether they were in Paris, or in some strange Californian motel. I am sure such questions will continue to be asked and I look forward to being there to listen to them and laugh together. Then, I am not sure how I can express my gratitude to Carla: you were always there to help me in a smarter way than I would have helped you and I am happy we managed to create a wonderful relationship that I hope to keep in our future. Thanks also to Vincent, I am sorry that our period in Pasteur did not overlap the whole time, but we both know that we share many ideas about life and I hope our story is not finished yet. After these permanent ones, some people participated partially in my Pasteurian experience. First, Ania: I need to thank you for your exceptional strength and determination, and I always smile when I think about the few moments we shared together. Last but not least, Elisa: thank you for sharing with me your enthusiasm in science as well as in music, I am very happy that our paths have crossed. A final Pasteurian acknowledgment goes to a new colleague and, most importantly, a friend that I unfortunately met only around the very end of my PhD, Stefano. Sharing those few moments with you (I am sure you understand which ones) has meant a lot to me from a human point of view and they were also fundamental during the final rush. However, I know the little time spent together was more than enough to understand that our friendship can last for a very long time.
Now it's time to thank my Sanofi team. I spent less time there, I went through the post-covid absenteeism and a relocation, but I am grateful to have had the opportunity to face this world that was completely different from my previous background. Above all, thanks to Jean-Philippe for welcoming me into his unit. I truly loved the exchanges, however ephemeral, we had about science, food, and life. Especially, I would like to thank you for your admirable quality to always keep a positive attitude, consciously in the middle of serious and facetious, and to share this attitude with whoever you're talking with without emphasizing that you were the boss. Thanks to Yann for his multiple advices and for the interest he always had in my research, thanks to Kwame for his enthusiasm and the friendship he always showed to me, thanks to Claire for her availability to share her knowledge. Thanks to Pierre, Charlotte, and Floriane for the time spent together sharing our exceptional experience of being students and employees.
I never lost contact with my former lab and I want to underline how important has been your support all along these intense years. First of all, thanks to Giancarlo for the openness of his mind and for always being available and interested in my situation. Thanks to Edo who is my first tight supervisor and who taught me more than anyone else how to give answers to the scientific questions arising from his vibrant and passionate curiosity. Thanks to Lorenzo for his recent involvement in my old project and for his availability in spending his time with me and supporting me.

It is time to spend some lines highlighting the most important contributions that so many people made during these exceptional years.
I cannot express with simple words how much I thank my beloved parents Patrizia and Roberto. If I try, I suddenly cry: everything for me started with your love and will continue on the path traced by the inestimable human, emotional, and cultural heritage that you were able to transmit to me. And thanks also to my brother Riccardo who is the only one who can be deeply aware of the extent of this heritage and who added several other layers to it with his own creative contribution. I really

look forward to building our future together. Despite its necessity for my personal development, being far from all of you and from the grandmas has probably been the hardest obstacle during these years. I am sorry and I apologize because this physical distance, which in a significant part has been imposed, is at the basis of an emotional one and it caused a difficulty in playing my role in our family. Having the time to dedicate to all of you will be a cornerstone of my future.
Indeed, what is the route of life without cornerstones? I am lucky to have different ones, who shared everything with me and to whom I am connected forever. Thanks to Giulia who, as she knows, has a special place in my heart: thanks for your constant support, for the fun that we always have, for all the moments we shared together in so many different contexts. Thanks to Francescone who is always the one I am calling whenever I have a problem, and he's always there and always has something deep and meaningful to suggest and to discuss. Thanks to Jack also for being there anytime I come back with the same love and the same spirit. Coming back to Rome without you will not be the same. Thanks to the deeper and deeper relationship I have with Daniele across the years and for his mastering of all the arts and sharing this knowledge with me. Thanks to Francesca who also is always there: even if this is the longest period that we spent apart, your place is unmodified. A special thanks is for Pietro, who made me feel having another house in the beautiful countryside of Florence: your creativeness pushes me always exploring new ideas and that constantly reminds me that we can build a greater future than the one that is depicted from the current geopolitical situation. Thanks to Svezia, Roscio, Pira, Elena, and all that the PAMF includes for making me always part of the group even if I was not there anymore. Thanks to Rosalba for keeping her role of smartest person I know, and to Marghe for keeping her role of greatest person I know. I am so grateful to all of you that, each time we meet, we are able (I am able?) to pretend that our lives are not happening in the same place. Hopefully, it will continue in this way.
As a conclusion of this roman thoughts, I need to express my sincere gratitude to Elena, who pushed me towards this Parisian adventure and who also pushed everything in the most difficult but best direction, even though she shared with me such a huge part of my life.

While I left several families in Rome, I was able to build several new ones here in Paris and nothing I did has been independent from the awareness of having such a powerful network of people. I lost count of the number of families I now have but I can say that I strongly enjoy thinking about all of you as a Kandinsky painting: chaotic, colorful, connected, beautiful.
The biggest thanks goes to the backbone of this family, Chiare, who has been my special partner in building the happiness I mentioned at the beginning. While I almost lost trains, flights, and life occasions due to my chaotic nature, you never lost a single opportunity to show me your love and support, making me catch what I was about to loose. The sunshine you once wrote me to hold needs a power supply: well, it has been nourished by the constant energy that you, with your sweet and stubborn self-giving, dedicated to me and, more generally, you channel in everything you do.
In the future, we will maybe tell about that incredible first night when I met another pillar of my Parisian life, Matteo, who gave me a kebab from the window against a molecular image of a protein. Thank you Matteo for saving me from hunger that night and so many days and nights after, thank you for your unconditional friendship, thank you for the intense and always smart discussions we had, thank you for being a person that I was missing in my life and that I will miss in the future that we will spend apart. You know, and your pizza will always have a place in the dream of a

perfect future together.
My life would have been much less motivating and much emptier without encountering three other people, whom we somehow managed to unify as the four fundamental forces and to make the Cosmos Electrique born. Pacome, you represent what I was looking for when I arrived in Paris. Thank you for the multiple afternoons we spent together that are among the most precious times I passed in the last three years. Still today, when I tell the story of our unusual encounter at the Belvedere, I can't believe that it really happened and that we are still here making music, pushing our limits, and producing unforgettable memories. In the same spirit, Caty, you also represent something I was looking for in Paris: a sister. Thank you for how much you loved me and took care of me during this period, especially the last one, which happily saw us living under the same roof. You two demonstrated to me the spirit that we can find in this city: against any cultural difference, we can be welcome and build our lives together. Our picture, however, is not complete without our beau-gosse, our star and master, Murilo. Thank you for investing so much time in deepening our relationship, thank you for sharing with happiness your knowledge, that goes so beyond Brazilian music, thank you for deciding not to give up when faced with the possibility to do so. And also, thank you for making us laugh even with your not funny jokes.
And what to say about the luck I had in meeting, on my first Parisian Saturday, Ylenia and eating the aubergines she cooked? Thanks to those eggplants because they opened multiple doors that contained the ingredients to build the majority of my actual life: an invaluable friend and comrade for limitless adventures, an immense Italian culture that I discovered only in Paris, an indispensable network of wonderful people. Each member of i Carbonari has a special place in my heart and needs to be mentioned. Thank you Mina for your meaningful eyes and your spiritual way of living which deeply inspires me and communicates probably more than what you already communicate with words. Thank you Dario for your true friendship, for your beautiful generosity (aka 'e sarsicce che se magnamo) and for the immense musical knowledge that you offer to me. Thank you Monica for your pop classicism, your unique nature and thoughts, and for the enthusiasm that you put into everything you do. Thank you Gianluca for the powerful passion that you put while playing and for being always the one on the side of pushing it further, as that night when we were only two in the huge square of Cardile. Thank you Serena for your involvement in singing, for your stubborn dedication, and your frankness; our discussions are always a hint to be better. Thank you Manola for your peaceful attitude, your beautiful polyphonies, and for the concerts when the tarantola bites you and injects some madness. Thank you Laura for your spontaneity, your sweetness, and for each single word of each single time that I can listen to your voice, daughter of the Vesuvio. Thank you Byron for your enthusiasm in playing with us, the carefulness you have in anything you hear and say, and the awareness of the greatness of Nature that your being exudes.
Many others constitute the vibrating and shining nodes of my Parisian network. The recent times are characterized by a deeper and deeper relationship with my homonymous Francesco, who has always centered on theater more than music: our encounter and collaboration may or may not be the starting point of something big, but it does not really matter. The happiness that I feel when I hear your scream above all the audible crowd sounds assures me that our souls are aside. Thanks to Chiara, who is now a constant sister for all the inconstant occasions we shared, that musical night on the Seine above all. Thanks to Cristina for being at my side and against me at the same time so many times. You two initiated me into a huge and different culture that I was barely aware of, the

Brazilian one, and that is now so present in my life. Thanks to your curiosity I also have the opportunity now to know such an exceptional guy as Max and to share with him so many adventures. Thanks to Margherita for the unexpressed but manifested love we share for each other, thanks to Aurora for her sweetness and her readiness to live life with the enthusiasm I practice, thanks to Maura for the countless nights spent together celebrating our simple will to celebrate, our passion to enjoy life, and thanks also to Giggio, who, with his unique way of living, was able to come here for one week and finally stayed nearly two years. Paris has been the confirmation of certitudes like the love I feel for Cama and for his original way of living, as well as the opening to new worlds like the one of the so-called "Trentini", including Lollo, whose peaceful and determined attitude is a truly inspiring quality, and Francois, who is born French but deserves a place in this category by osmosis. After some seasons of this Parisian episode, new people called "Cinesi" (even though everyone is 100% from Italy) entered my life and excited it. Thanks to Martina for representing this group of people and for her sweetness, I always think of all the times that you told me how much you're happy we met. Thanks to Niccolò, a new brother that I was only waiting to meet, for his innovative and inspiring ideas, thanks to Mattia for his uniqueness and for the high quality of the conversations with him, thanks to Erika for the beauty of the contrast between her quiet exterior and the original fire she has inside.
For sure, among Italian people abroad, we can form a closed tortoise, with a network efficiency that I would have never predicted without experiencing it. Although I will always be convinced of the greater cultural heritage and better food of Italy, and all kidding aside, these years also revealed to me to what extent we are all made from the same matter. I met so many people that welcomed me into their lives and that are now active parts of mine. The first ones I want to mention, who opened all the doors of music and entertainment in this city are the inhabitants of the so-called Maison du Bonheur: merci Antonn for all that you organized and for the exceptional quality that characterizes whatever you organize, merci Thierry for your sincere curiosity and your love, merci Florence for your kindness and welcoming spirit, merci Eva for being so styled when you play the bass, merci Alice for your sweetness. As a unique flower growing only at certain latitudes, merci to Tinj and to the light that he glows, you are able to make the beauty shine in everything you see and to tell this modern tale to the lucky ones you meet in your street. Merci Barbara for your constant example of strength, artistic sensibility, and fun, days are always better when you're there.

Now, it's time to stop acknowledging and start facing the new life that is going to begin. In this perspective, quoting a person I love, I would like to express the gratitude I have for myself for finally completing this long journey and for being able to build everything I built.

Paris and other nearby places, July 2024

Résumé

Les molécules d'ARN sont devenues des cibles thérapeutiques majeures, et le ciblage par petites molécules se révèle particulièrement prometteur. Cependant, malgré leur potentiel, le domaine est encore en développement, avec un nombre limité de médicaments spécifiquement conçus pour l'ARN. La flexibilité intrinsèque de l'ARN, bien qu'elle constitue un obstacle, introduit des opportunités thérapeutiques que les outils computationnels actuels ne parviennent pas pleinement à exploiter malgré leur prédisposition. Le projet de cette thèse est de construire un cadre computationnel plus complet pour la conception rationnelle de composés ciblant l'ARN. La première étape pour toute approche structure-based est l'analyse des connaissances structurales disponibles. Cependant, il manquait une base de données complète, organisée et régulièrement mise à jour pour la communauté scientifique. Pour combler cette lacune, j'ai créé HARIBOSS, une base de données de toutes les structures expérimentalement déterminées des complexes ARN-petites molécules extraites de la base de données PDB. Chaque entrée de HARIBOSS, accessible *via* une interface web dédiée, est annotée avec les propriétés physico-chimiques des ligands et des poches d'ARN. Cette base de données constamment mise à jour facilitera l'exploration des composés drug-like liées à l'ARN, l'analyse des propriétés des ligands et des poches, et en fin de compte, le développement de stratégies *in silico* pour identifier des petites molécules ciblant l'ARN. Lors de sa sortie, il a été possible de souligner que la majorité des poches de liaison à l'ARN ne conviennent pas aux interactions avec des molécules drug-like. Cela est dû à une hydrophobicité moindre et une exposition au solvant accrue par rapport aux sites de liaison des protéines. Cependant, cela résulte d'une représentation statique de l'ARN, qui peut ne pas capturer pleinement les mécanismes d'interaction avec de petites molécules. Il était nécessaire d'introduire des techniques computationnelles avancées pour une prise en compte efficace de la flexibilité de l'ARN. Dans cette direction, j'ai mis en œuvre SHAMAN, une technique computationnelle pour identifier les sites de liaison potentiels des petites molécules dans les ensembles structuraux d'ARN. SHAMAN permet d'explorer le paysage conformationnel de l'ARN cible par des simulations de dynamique moléculaire atomistique. Dans le même temps, il identifie efficacement les poches d'ARN en utilisant de petits fragments dont l'exploration de la surface de l'ARN est accélérée par des techniques d'enhanced sampling. Dans un ensemble de données comprenant divers riboswitches structurés ainsi que de petits ARN viraux flexibles, SHAMAN a précisément localisé des poches résolues expérimentalement, les classant les régions d'intéraction préférées. Notamment, la précision de SHAMAN est supérieure à celle d'autres outils travaillant sur des structures statiques d'ARN dans un scénario réaliste de découverte de médicaments où seules les structures apo de la cible sont disponibles. Cela confirme que SHAMAN est une plateforme robuste pour les futures initiatives de conception de médicaments ciblant l'ARN avec de petites molécules, en particulier compte tenu de sa pertinence potentielle dans les campagnes de criblage virtuel. Dans l'ensemble, ma recherche contribue à améliorer notre compréhension et notre utilisation de l'ARN en tant que cible pour les médicaments à petites molécules, ouvrant la voie à des stratégies thérapeutiques plus efficaces dans ce domaine en évolution.

Mots clés: petites molécules ciblant l'ARN, conception de médicaments assistée par ordinateur, approches basées sur la structure, dynamique moléculaire, flexibilité de l'ARN

Abstract

RNA molecules have recently gained huge relevance as therapeutic targets. The direct targeting of RNA with small molecule drugs emerges for its wide applicability to different classes of RNAs. Despite this potential, the field is still in its infancy and the number of available RNA-targeted drugs remains limited. A major challenge is constituted by the highly flexible and elusive nature of the RNA targets. Nonetheless, RNA flexibility also presents unique opportunities that could be leveraged to enhance the efficacy and selectivity of newly designed therapeutic agents. To this end, computer-aided drug design techniques emerge as a natural and comprehensive approach. However, existing tools do not fully account for the flexibility of the RNA. The project of this PhD work aims to build a computational framework toward the rational design of compounds targeting RNA. The first essential step for any structure-based approach is the analysis of the available structural knowledge. However, a comprehensive, curated, and regularly updated repository for the scientific community was lacking. To fill this gap, I curated the creation of HARIBOSS ("Harnessing RIBOnucleic acid – Small molecule Structures"), a database of all the experimentally determined structures of RNA-small molecule complexes retrieved from the PDB database. HARIBOSS is available *via* a dedicated web interface and is regularly updated with all the structures resolved by X-ray, NMR, and cryo-EM, in which ligands with drug-like properties interact with RNA molecules. Each HARIBOSS entry is annotated with physico-chemical properties of ligands and RNA pockets. HARIBOSS repository, constantly updated, will facilitate the exploration of drug-like compounds known to bind RNA, the analysis of ligands and pockets properties and, ultimately, the development of in silico strategies to identify RNA-targeting small molecules. In coincidence of its release, it was possible to highlight that the majority of RNA binding pockets are unsuitable for interactions with drug-like molecules, attributed to the lower hydrophobicity and increased solvent exposure compared to protein binding sites. However, this emerges from a static depiction of RNA, which may not fully capture their interaction mechanisms with small molecules. In a broader perspective, it was necessary to introduce more advanced computational techniques for an effective accounting of RNA flexibility in the characterization of potential binding sites. In this direction, I implemented SHAMAN, a computational technique to identify potential small-molecule binding sites in RNA structural ensembles. SHAMAN enables the exploration of the target RNA conformational landscape through atomistic molecular dynamics. Simultaneously, it efficiently identifies RNA pockets using small probe compounds whose exploration of the RNA surface is accelerated by enhanced sampling techniques. In a benchmark encompassing diverse large, structured riboswitches as well as small, flexible viral RNAs, SHAMAN accurately located experimentally resolved pockets, ranking them as preferred probe hotspots. Notably, SHAMAN accuracy was superior to other tools working on static RNA structures in the realistic drug discovery scenario where only apo structures of the target are available. This establishes SHAMAN as a robust platform for future drug design endeavors targeting RNA with small molecules, especially considering its potential applicability in virtual screening campaigns. Overall, my research contributed to enhancing our understanding and utilization of RNA as a target for small molecule drugs, paving the way for more effective drug design strategies in this evolving field.

Keywords: RNA-targeted small molecules, computer-aided drug design, structure-based approaches, Molecular Dynamics, RNA flexibility

:wq