



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

**DOTTORATO DI RICERCA IN
DATA SCIENCE AND COMPUTATION**

Ciclo 37

Settore Concorsuale: 03/D1 - CHIMICA E TECNOLOGIE FARMACEUTICHE,
TOSSICOLOGICHE E NUTRACEUTICO - ALIMENTARI

Settore Scientifico Disciplinare: CHIM/08 - CHIMICA FARMACEUTICA

Ensemble-Aware Computational Approaches for RNA-Targeted Drug Discovery

Presentata da: Stefano Bosio

Coordinatore Dottorato

Prof. Daniele Bonacorsi

Supervisore

Prof. Matteo Masetti

Cosupervisori

Prof. Andrea Cavalli

Dr. Mattia Bernetti

Esame finale anno 2026

*To my first mentors, my family,
Pier Luigi, Luisa, and Fabrizio*

Abstract

The expanding recognition of the central roles of RNA in cellular regulation is opening new therapeutic opportunities for diseases driven by its dysregulation. Two principal strategies are pursued in RNA based therapeutics, oligonucleotides and small molecules. While oligonucleotides have achieved clinical milestones, this thesis concentrates on small molecules, whose chemical tunability and compatibility with medicinal chemistry provide a complementary route. Despite growing interest, few marketed drugs directly target RNA, and most were identified through phenotypic screening. This thesis addresses three interconnected projects, progressing from comparative biophysical analysis to applied RNA targeting and experimental ensemble refinement.

First, we compared ligand recognition in a protein enzyme and an RNA receptor that bind the same ligand, riboflavin kinase and the flavin mononucleotide riboswitch. Molecular dynamics simulations characterized conformational space, electrostatic organization, and ligand and solvent dynamics within binding pockets. The analysis revealed comparable capacities to establish polar noncovalent interactions, alongside distinct patterns of flexibility and solvent exchange.

Second, we developed an early stage discovery framework targeting the HIV-1 trans activation response element. Mixed solvent simulations using the SHAdow Mixed solvent metAdynaMics approach generated hot spot and pocket maps that guided ensemble aware virtual screening. Conformational clustering was refined with local structural descriptors to preserve base pairing and stacking features. A hybrid pipeline integrating rigid receptor docking, a machine learning scoring function, and complementary chemical space exploration produced a shortlist of approximately two hundred candidate compounds.

Third, we refined an HIV-1 TAR conformational ensemble by combining enhanced sampling with Maximum Entropy reweighting to reproduce residual dipolar coupling data. Well Tempered Ensemble simulations expanded conformational coverage, while reweighting enforced agreement with experimental averages. The refined ensemble enabled the characterization of dominant substates and their roles in ligand recognition.

Together, these studies establish an ensemble centered framework for RNA targeted small molecule discovery.

Table of Contents

List of Acronyms and Abbreviations	VII
1 Introduction	1
1.1 Ribonucleic Acid (RNA)	1
1.1.1 Essential biological and historical aspects	2
1.1.2 Classification of RNA Functions	3
1.1.3 Chemical properties of RNA	5
1.2 RNA as a Drug and as a Druggable Target	9
1.3 Computational approaches in drug discovery	11
1.3.1 Computational approaches in RNA-targeted drug discovery	12
2 Aim of the work	15
3 Theoretical background	17
3.1 Molecular Dynamics	17
3.1.1 RNA Force Fields	19
3.1.2 Enhanced Sampling Methods	20
3.1.3 Maximum Entropy Integration of Simulations and Experiments	24
3.1.4 Mixed Solvent Approaches	27
3.2 Computational strategies for ligand screening and prioritization	30
3.2.1 Theoretical Foundations of Molecular Docking	30
3.2.2 The RNAmigos2 Virtual Screening Pipeline	31
3.3 Unsupervised Machine Learning for MD simulations	32
3.3.1 Dimensionality Reduction	32
3.3.2 Clustering	34
4 From Proteins to RNA: Comparative Analysis of Ligand Binding Determinants	37
4.1 Methods	40
4.1.1 System Setup and Simulation Protocol	40
4.2 Results and Discussions	45

4.2.1	Enhanced Conformational Flexibility in RNA without Appreciable Pocket Volume Variation	46
4.2.2	Comparable Electrostatic Contribution to Ligand Binding . .	52
4.2.3	Distinct Pattern of Hydrogen Bonds and Ligand Conformations	55
4.2.4	Faster Exchange Rate of Water in the Bound Protein Binding Site	58
4.2.5	Summary of Main Findings	61
4.2.6	Limitations of the study	64
4.3	Conclusions	65
5	Exploring RNA Druggability through SHAMAN	67
5.1	Methods	70
5.1.1	System setup and simulation protocol	70
5.1.2	Virtual screening database preparation	72
5.1.3	Structure preparation and virtual screening	73
5.1.4	Diversity-based compound selection	73
5.1.5	Scaffold and Chemical Space Network Analyses	74
5.2	Results and Discussion	76
5.2.1	Simulations and Pocket Characterization	76
5.2.2	From Ensemble Clustering to Docking Integration	79
5.2.3	Analysis of the Final Set Properties	80
5.3	Summary of Main Findings	83
5.4	Conclusions	84
6	Integrative Refinement of RNA Conformational Ensembles via the Maximum Entropy Principle	87
6.1	Methods	89
6.1.1	System setup and simulation protocol	89
6.1.2	Residual Dipolar Couplings (RDCs)	91
6.1.3	SHAPE reactivity	92
6.1.4	SiteMap	93
6.2	Results and Discussion	93
6.2.1	Maximum entropy reweighting and validation	96
6.2.2	Redistribution of conformational ensembles after reweighting	99

6.2.3	Comparison with available ligand-binding pockets	103
6.2.4	Summary of main findings	108
6.3	Conclusions	109
7	Conclusions and Outlook	111
	References	115
	List of Figures	131
	List of Tables	147
A	Theoretical Background	149
B	Similarities and Differences in Ligand Binding to Protein and RNA Tar- gets: The Case of Riboflavin	151
B.1	Calculation of sequence identity and similarity between human and bacterial RF kinase	156
C	Exploring RNA Druggability through SHAMAN	157
D	Integrative Refinement of RNA Conformational Ensembles via the Maxi- mum Entropy Principle	161
E	Data availability	167
	Acknowledgements	169

List of Acronyms and Abbreviations

ADP	Adenosine Diphosphate
AI	Artificial Intelligence
ASO	Antisense Oligonucleotide
circRNA	Circular RNA
COM	Center of Mass
Cryo-EM	Cryogenic Electron Microscopy
CV	Collective Variable
eRNA	Enhancer RNA
FMN	Flavin Mononucleotide Riboswitch
IDP	Intrinsically Disordered Protein
KD	Dissociation Constant
lncRNA	Long non-coding RNA
MaxEnt	Maximum Entropy
MD	Molecular Dynamics
MetaD	Metadynamics
ML	Machine Learning
MM	Molecular Mechanics
miRNA	microRNA
mRNA	Messenger RNA
NMR	Nuclear Magnetic Resonance
PCA	Principal Component Analysis
PDB	Protein Data Bank
piRNA	PIWI-interacting RNA
PT	Parallel Tempering
PT-WTE	Parallel Tempering in the Well-Tempered Ensemble
QT	Quality Threshold

RDC Residual Dipolar Coupling
RF Riboflavin
RK Riboflavin Kinase
RMSD Root Mean Square Deviation
RMSF Root Mean Square Fluctuation
RNA Ribonucleic Acid
rRNA Ribosomal RNA
SBDD Structure-Based Drug Discovery
scaRNA Cajal-body RNA
SHAMAN Shadow Mixed-Solvent Metadynamics
siRNA Small Interfering RNA
SMA Spinal Muscular Atrophy
SMN Survival Motor Neuron
snRNA Small Nuclear RNA
snoRNA Small Nucleolar RNA
SP Survival Probability
tRF tRNA-derived Fragments
TAR Trans-Activation Response Element
tracrRNA Trans-activating CRISPR RNA
tRNA Transfer RNA
t-SNE t-distributed Stochastic Neighbor Embedding
WTE Well-Tempered Ensemble
WTMetaD Well-Tempered Metadynamics

1 Introduction

1.1 Ribonucleic Acid (RNA)

Ribonucleic acid (RNA) is widely recognized as one of the most essential and versatile biomolecules in molecular biology, being present in all living systems.[1, 2] For decades, its role was thought to be restricted to the transmission of genetic information from DNA to proteins, as formulated in the Central Dogma of molecular biology.[3] Within this classical framework, RNA was largely regarded as a transient messenger, a simple intermediary between the genetic archive of DNA and the functional repertoire of proteins.

Over the past decades, however, this view has expanded considerably. Advances in molecular biology and biochemistry have revealed that RNA molecules are not passive carriers but active regulators and catalysts. The classical RNA species, messenger RNA (mRNA), transfer RNA (tRNA), and ribosomal RNA (rRNA), remain indispensable for protein synthesis, yet an ever-growing number of non-coding RNAs (ncRNAs) have been shown to control gene expression, modulate chromatin states, and mediate cellular stress responses. [1] The discovery of catalytic RNAs, or ribozymes, together with the pervasive presence of regulatory RNAs in virtually all domains of life, has firmly positioned RNA at the core of cellular regulation.[2]

Beyond its biological relevance, RNA has also emerged as a central element in biotechnology and medicine. The success of RNA-based vaccines has demonstrated the feasibility of exploiting RNA molecules as therapeutic agents, while antisense oligonucleotides, RNA interference, and RNA-targeted small molecules exemplify the growing pharmacological importance of this field.[4, 5] This dual nature of RNA, as both a biological regulator and a pharmacological tool, has reshaped the landscape of modern drug discovery and inspired intense research efforts across academia and industry.

This introductory chapter, therefore, aims to provide an overview of the essential biological aspects of RNA, tracing its transition from a supporting role within the Central Dogma to a central actor in the regulation of cellular processes. It will also

emphasize the increasing importance of RNA in pharmaceutical sciences, where it is now investigated both as a therapeutic agent and as a molecular target with vast potential applications.

1.1.1 Essential biological and historical aspects

To understand the multifaceted roles of RNA, it is essential to begin with its fundamental biological features. Chief among these is RNA biosynthesis, which takes place through transcription, a central process by which genetic information is transferred from DNA to RNA molecules. During transcription, RNA polymerase catalyzes the stepwise assembly of the RNA strand by forming covalent bonds between nucleotides.

Within the framework of molecular biology, this property was long interpreted in light of the Central Dogma,[3] originally formulated in the 1960s. According to this model, RNA was considered a passive intermediary in the linear flow of genetic information from DNA to functional proteins (Fig. 1.1).

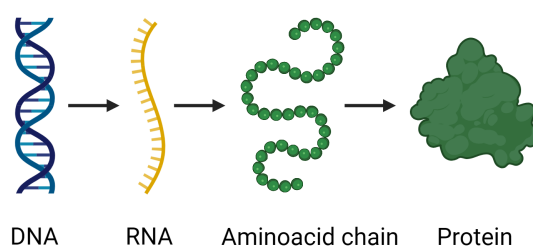


Figure 1.1. Schematic representation of the Central Dogma, illustrating the flow of genetic information from DNA to RNA to proteins.

This perspective was reinforced by the identification of key RNA species such as messenger RNA (mRNA), transfer RNA (tRNA), and ribosomal RNA (rRNA), which collectively support the protein synthesis machinery. For several decades, transcripts lacking coding potential were largely disregarded and often labeled as "junk RNA".[6]

Only in the closing years of the twentieth century did a paradigm shift occur, demonstrating that RNA molecules can carry out a wide spectrum of regulatory functions beyond translation. A decisive breakthrough came in 1998, when Fire and Mello showed that double-stranded RNA could trigger the degradation of specific mRNA sequences in *C. elegans*. [7] This process, later termed RNA interference (RNAi),

provided conclusive proof that RNA could modulate gene expression at the post-transcriptional level.

Building on these foundations, subsequent studies revealed additional classes of non-coding RNAs, including microRNAs[8] and long non-coding RNAs,[9] which regulate gene expression through diverse mechanisms such as translational repression, mRNA destabilization, or chromatin remodeling.[9] Moreover, RNA-mediated regulation proved to be responsive to environmental cues, including temperature fluctuations, highlighting the ability of RNA molecules to act as dynamic sensors and effectors within complex gene regulatory networks.[10]

In recent years, increasing attention has been directed toward the "RNA World" hypothesis,[11] a theoretical framework proposing that early life on Earth may have originated from self-replicating RNA molecules capable of both storing genetic information and catalyzing chemical reactions. In this view, RNA would have preceded the evolution of DNA and proteins, fulfilling dual roles as both genotype and phenotype in the earliest biological systems. This concept is supported by the discovery of catalytic RNAs, or ribozymes, as well as by the central role of RNA in contemporary biological machineries such as the ribosome, which preserves functional elements of ancient RNA-based catalysis.[12]

1.1.2 Classification of RNA Functions

RNA molecules can be classified according to several criteria, including coding potential, size, and functional role.[13–15] The fundamental distinction lies between coding and non-coding RNAs, which fulfill divergent roles and therefore exhibit marked heterogeneity in both structure and function.[16]

1. Coding RNAs Messenger RNAs (mRNAs) act as intermediaries that transfer genetic information from DNA to ribosomes, where they direct protein synthesis.

2. Housekeeping non-coding RNAs These RNAs carry out constitutive and indispensable cellular functions. Key examples include:

- Transfer RNAs (tRNAs), adaptor molecules that deliver specific amino acids to the ribosome during translation.

- Ribosomal RNAs (rRNAs), central structural and catalytic components of ribosomes, essential for protein assembly.
- Small nuclear RNAs (snRNAs), which participate in pre-mRNA splicing.
- Small nucleolar RNAs (snoRNAs) and Cajal-body RNAs (scaRNAs), which guide chemical modifications of other RNAs, such as rRNAs and snRNAs.

3. Regulatory non-coding RNAs This category comprises RNAs that modulate gene expression, epigenetic states, and related regulatory processes. They are commonly divided by length:

- **Small non-coding RNAs** (typically less than 200 nucleotides), including:
 - i) microRNAs (miRNAs), which regulate mRNA stability and translation; [16]
 - ii) small interfering RNAs (siRNAs), derived from endogenous or exogenous sources to induce post-transcriptional silencing; [7]
 - iii) PIWI-interacting RNAs (piRNAs), crucial for transposon silencing and germline genome integrity;
 - iv) tRNA-derived fragments (tRFs), increasingly recognized as modulators during cellular stress.
- **Long non-coding RNAs (lncRNAs)**, a heterogeneous group of transcripts longer than 200 nucleotides, involved in epigenetic regulation, transcriptional control, and post-transcriptional processes. [16, 17]

4. Other non-coding RNAs Several additional classes have also been identified:

- Enhancer RNAs (eRNAs), transcribed from enhancer elements and thought to facilitate chromatin architecture and transcriptional activation.[18, 19]
- Circular RNAs (circRNAs), characterized by covalently closed structures, implicated in gene regulation and increasingly studied as disease biomarkers.[20]
- Trans-activating CRISPR RNAs (tracrRNA) in prokaryotes, which guide CRISPR-associated (Cas) proteins to recognize and degrade invading nucleic acids as part of adaptive immunity.[21]

1.1.3 Chemical properties of RNA

RNA is composed of four fundamental monomeric units, the nucleotides. Each nucleotide contains three main components: a planar aromatic base, a ribose sugar in its furanose form, and a phosphate group (Figure 1.2). The chemical identity of the nitrogenous base determines the nature of the nucleotide, which may be adenine (A), guanine (G), cytosine (C), or uracil (U).

Structurally, RNA is a linear polynucleotide chain in which nucleotides are covalently linked by 3'–5' phosphodiester bonds. A defining feature of RNA, when compared to DNA, is the presence of a hydroxyl group at the 2' position of the ribose ring. This additional functional group acts both as a hydrogen-bond donor and acceptor, thereby broadening the spectrum of possible intra and intermolecular interactions. The 2' hydroxyl group also confers high chemical reactivity: it can perform intramolecular nucleophilic attack on the adjacent phosphate group, leading to cleavage of the phosphodiester bond between the 3' end of one nucleotide and the 5' carbon of the next. As a result, RNA is intrinsically prone to spontaneous hydrolysis under physiological conditions.[22]

Compared to DNA, RNA exhibits much greater structural flexibility. This conformational plasticity arises largely from the 2' hydroxyl group, which perturbs the regularity of the sugar-phosphate backbone and promotes the formation of diverse secondary and tertiary structural motifs.

A central aspect in the study of RNA dynamics lies in the network of non-covalent interactions along the chain. In addition to classical interactions also present in proteins, such as electrostatic forces, especially relevant due to the highly negatively charged phosphate backbone, and van der Waals interactions, RNA displays distinctive base–base interactions.[23] These fall into two main categories: base pairing and base stacking.

Base pairing involves hydrogen bonding between nucleobases. RNA most often adopts canonical Watson–Crick pairs, with adenine (A) binding uracil (U) and guanine (G) binding cytosine (C). However, the structural properties of RNA, particularly the 2' hydroxyl group, allow for numerous non-canonical geometries. These include

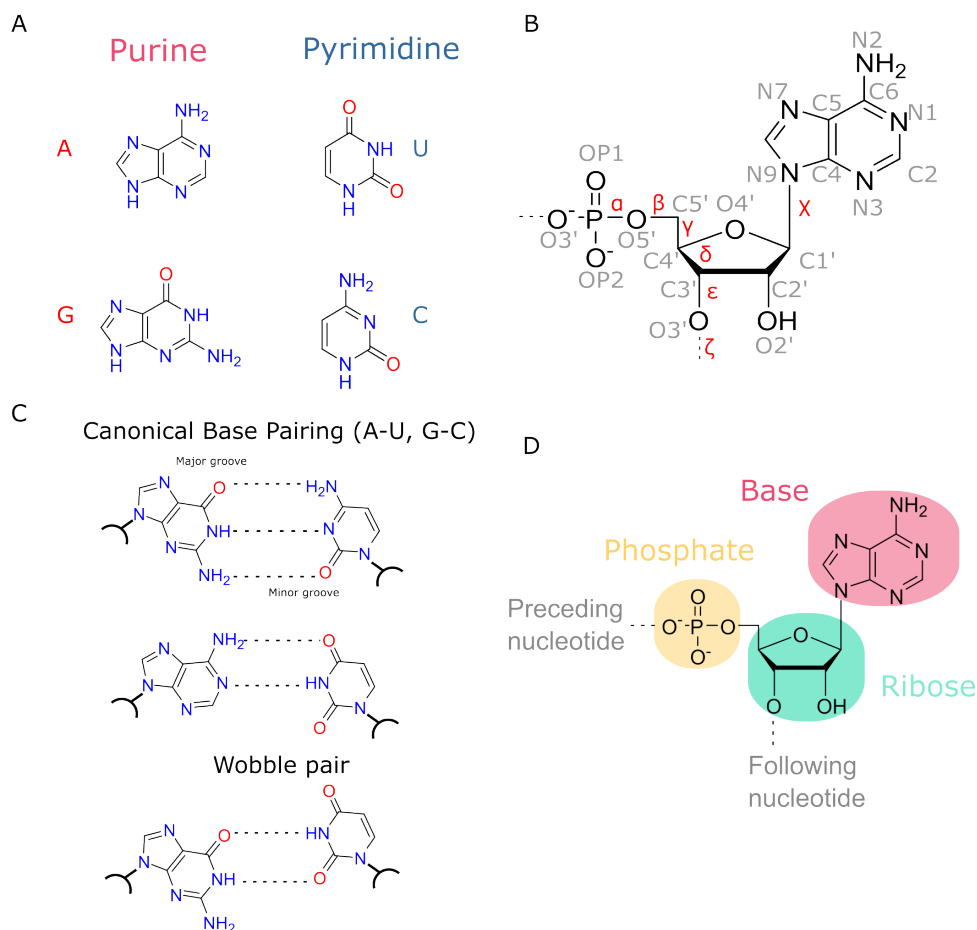


Figure 1.2. Structural features of RNA nucleotides and base pairing. **(A)** Chemical structures of the four canonical RNA nucleobases: purines (adenine, A; guanine, G) and pyrimidines (cytosine, C; uracil, U). **(B)** Structure of an RNA nucleotide, highlighting the ribose sugar, phosphate group, and nitrogenous base (adenine shown as an example). The principal torsional angles of the RNA backbone (α , β , γ , δ , ϵ , ζ , and χ) are indicated. **(C)** Base-pairing schemes: canonical Watson–Crick pairs (A–U and G–C) and the non-canonical wobble G–U pair. Hydrogen bonds and the orientation of major and minor grooves are shown. **(D)** Schematic representation of a nucleotide within the RNA backbone, showing the phosphate group (yellow), ribose sugar (green), and nitrogenous base (red), and their connectivity to adjacent nucleotides.

wobble pairs such as G–U, as well as many other arrangements that deviate from standard Watson–Crick rules. A classification of such interactions, based on glycosidic bond orientation and the edges of the bases involved, is summarized in Figure 1.2C.

Base stacking refers instead to attractive interactions between the aromatic rings of adjacent nucleobases. Governed by the principles of π – π stacking, these interactions provide substantial stabilization to both secondary and tertiary RNA structures. Stacking is highly sequence- and context-dependent, with its strength influenced by solvent conditions, ion concentrations, and backbone conformation.

1.1.3.1 RNA structural features

Secondary structure motifs The primary structure of RNA, typically a single linear strand of covalently linked ribonucleotides, is conceptually simple. By contrast, its secondary and tertiary organizations display remarkable complexity.[24, 25] Secondary structure arises mainly from local base pairing and base stacking interactions, which drive the formation of energetically stable double-helical domains. These regions usually adopt an A-form helix stabilized by canonical Watson–Crick base pairs, particularly G–C and A–U, and serve as the fundamental building blocks of RNA folding.[26]

Not all nucleotides, however, participate in canonical hydrogen bonding. Unpaired nucleotides generate characteristic single-stranded motifs, which are hallmarks of RNA architecture. Common examples include apical loops (located at helix termini), internal loops (interrupting helical continuity), and bulge loops (unpaired residues protruding from one strand of a duplex).[27] The combination of a helical segment capped by an apical loop forms a stem-loop, or hairpin structure, one of the most ubiquitous secondary motifs. When multiple helices converge at a single site, more complex architectures known as multi-helix junctions emerge (Figure 1.3).[28]

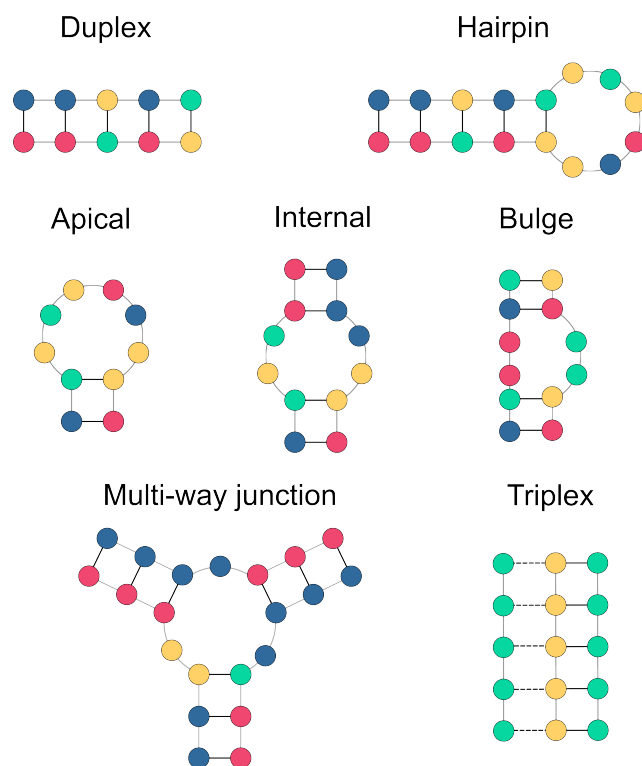


Figure 1.3. Schematic representations of common RNA secondary structure motifs, including duplexes, internal loops, bulges, hairpins, multiloops, multi-way junctions, and RNA triplexes. Nucleotides are color-coded as follows: adenine (A) in yellow, uracil (U) in green, guanine (G) in blue, and cytosine (C) in red. Light grey lines indicate phosphodiester backbone connections. Black solid lines denote canonical base pairs (Watson-Crick), while dotted lines indicate non-canonical interactions.

Beyond canonical base pairing, RNA can establish a broad spectrum of non-canonical interactions that substantially enrich its structural and functional repertoire. These include alternative hydrogen-bonding schemes, such as Hoogsteen and reverse Hoogsteen pairing, which can support base triplets and higher-order motifs. For instance, U·A·U base triples (i.e., a uridine forming a Hoogsteen interaction with the adenine of a pre-existing A–U Watson–Crick pair) are widespread in RNA triple helices and help stabilize tertiary folds.[29] Such contacts are central to the stabilization of RNA tertiary structure and to mediating long-range intramolecular communication.

Tertiary structure motifs Tertiary structure refers to the three-dimensional organization of the RNA chain and emerges from long-range spatial interactions among secondary elements. While certain motifs may be distant in the linear sequence, or even appear disconnected in two-dimensional representations, their folding within the cellular environment can bring them into close proximity, enabling the establishment of stable and functionally relevant contacts.[30]

Canonical duplexes often stack cooperatively, giving rise to extended helical domains. Their stability and orientation are governed largely by coaxial stacking, where adjacent helices align along a common axis, maximizing base overlap and enhancing thermodynamic stability. Junction regions, where several helices meet, are key determinants of global architecture: they dictate relative helical orientation and provide structural flexibility.[30, 31]

More elaborate topologies result from tertiary contacts between independently folded secondary motifs. Notable examples include kissing-loop interactions, where two hairpin loops form complementary base pairs, and pseudoknots, in which a single-stranded region folds back to base pair with a previously distant sequence, producing a topologically interlaced structure.[32] These motifs enhance the compactness, rigidity, and functional diversity of folded RNA molecules and frequently serve as scaffolds for ligand binding or protein recognition.

1.2 RNA as a Drug and as a Druggable Target

As previously discussed, RNA has progressively emerged as both a central regulator of cellular physiology and a versatile target for therapeutic strategies.[33, 34] Beyond its canonical role in transmitting genetic information, the discovery of non-coding RNAs, riboswitches, and post-transcriptional modifications has revealed an unprecedented spectrum of structural and functional diversity. These insights have driven the development of RNA-based therapies and, more recently, the direct targeting of RNA with small molecules. Current strategies can be broadly divided into two complementary approaches: RNA as a pharmaceutical agent, including oligonucleotide drugs such as antisense oligonucleotides, siRNAs, miRNAs, and CRISPR-based technologies, and RNA as a therapeutic target for small molecules.[4, 5, 33, 35,

36] The following sections briefly review these two paradigms.

RNA as a Pharmaceutical Agent Oligonucleotide therapeutics exploit Watson–Crick complementarity to modulate gene expression at the RNA level with high specificity. Within this class, antisense oligonucleotides (ASOs) hybridize to complementary mRNA sequences to alter splicing or elicit RNase H–mediated degradation.[37] FDA-approved examples include Nusinersen, which promotes exon 7 inclusion in the survival motor neuron 2 (SMN2) transcript to treat spinal muscular atrophy, [38, 39] and Eteplirsen, an exon-skipping agent used in Duchenne muscular dystrophy.[40, 41] In parallel, siRNAs have reached the clinic: landmark approvals include Patisiran, the first FDA-approved siRNA that silences transthyretin (TTR) mRNA in hereditary transthyretin amyloidosis, [42, 43] followed by Givosiran for acute hepatic porphyria[44] and Lumasiran for primary hyperoxaluria type 1,[45] underscoring the therapeutic breadth of RNAi. Therapeutic modulation of miRNAs is also under active investigation, with synthetic oligonucleotides designed either to mimic tumor-suppressive miRNAs or to inhibit oncogenic ones.[46]

A distinct modality is represented by mRNA-based vaccines. Unlike ASOs and siRNAs, which modulate endogenous transcripts, these formulations deliver exogenous mRNA encoding antigenic proteins to elicit protective immune responses. The clinical success of the Pfizer–BioNTech BNT162b2[47] and Moderna mRNA-1273 [48] vaccines against SARS-CoV-2 established the feasibility and scalability of mRNA therapeutics, and their adaptability to rapidly emerging pathogens. Beyond infectious diseases, mRNA vaccines are being explored in oncology to encode tumor-associated antigens and stimulate anti-tumor immunity.[49]

Small Molecules Targeting RNA In parallel with oligonucleotide-based approaches, growing efforts have focused on developing small molecules that selectively bind RNA structures.[4, 34] Unlike oligonucleotide drugs, small molecules offer favorable pharmacokinetic properties, including oral bioavailability and tissue penetration, while retaining the ability to modulate RNA function. The feasibility of RNA as a small-molecule target is supported by numerous clinical and preclinical studies, as well as advances in structural biology and computational drug design. [33, 50] Classic examples include antibiotics such as aminoglycosides, which bind

bacterial rRNA to inhibit protein synthesis, and linezolid, which targets the ribosomal peptidyl transferase center.[51] Riboswitches, naturally occurring regulatory RNA elements, have also been exploited: synthetic ligands for the preQ₁ riboswitch, for example, have shown that rational design can achieve potent and selective modulation of RNA tertiary structures.[52] Similarly, group II introns, which function as ribozymes, have been inhibited by compounds such as intronistat B, with potential antifungal applications.[53]

A particularly notable achievement in human therapeutics is the development of small-molecule modulators of pre-mRNA splicing. Risdiplam target the SMN2 pre-mRNA, stabilizing the interaction between the U1 snRNP and the transcript to promote exon 7 inclusion.[54–57] Risdiplam, in particular, represents a landmark success: it is the first orally bioavailable RNA-targeting small molecule approved by the FDA for the treatment of spinal muscular atrophy (SMA). SMA is a severe neuromuscular disorder caused by insufficient levels of SMN protein due to deletions or mutations in the SMN1 gene. Patients rely on the paralogous SMN2 gene, which predominantly produces transcripts lacking exon 7 and therefore encodes unstable proteins. Risdiplam modifies SMN2 splicing, thereby increasing the production of full-length SMN protein. [54]

1.3 Computational approaches in drug discovery

Over the past few decades, computational methods have become essential tools in modern drug discovery, complementing and accelerating experimental pipelines. Structure-based approaches, such as molecular docking, molecular dynamics (MD) simulations, and free energy calculations, are now routinely employed to explore biomolecular recognition and guide hit-to-lead optimization.[58–60] Docking enables the rapid screening of large compound libraries against target structures, providing putative binding poses and rough estimates of affinity. MD simulations extend this framework by explicitly accounting for the dynamic nature of biomolecules under realistic conditions, thereby elucidating mechanistic features and conformational adaptability.[61] Free energy methods, including alchemical and endpoint calculations, further allow quantitative assessment of ligand binding thermodynamics and are increasingly used to prioritize compounds with higher accuracy.[62]

In recent years, the field has also been reshaped by the integration of artificial intelligence (AI) and machine learning (ML), which are revolutionizing both ligand and structure-based approaches. Data-driven models are now applied to predict binding affinities, generate de novo molecules, and optimize compound libraries, thereby complementing physics-based simulations and enhancing their predictive power. [63] The integration of these strategies has consolidated computational drug discovery as a core element in pharmaceutical research, capable of reducing costs, shortening timelines, and providing molecular-level insights unattainable by experimental techniques alone.

1.3.1 Computational approaches in RNA-targeted drug discovery

Although computational methods have long reached the status of established tools in protein-based drug discovery,[58–60] their systematic application to RNA targets is still at an early stage.[33, 34, 64] Structure-based design and virtual screening pipelines are now routine in campaigns directed against protein targets, where decades of crystallographic data and refined force fields provide reliable structural and energetic frameworks. By contrast, RNA drug discovery poses unique challenges, including its pronounced conformational plasticity, the scarcity of high-resolution experimental structures, and the limited accuracy of force fields specifically optimized for nucleic acids.[23, 34] Moreover, the chemical space of RNA binders remains only partially explored, with emerging evidence suggesting distinctive physicochemical requirements that deviate from protein-oriented rules.[65, 66]

Despite these limitations, recent progress highlights the growing feasibility of computational approaches in RNA drug discovery. Ensemble-based docking strategies, initially developed for proteins, are being adapted to capture RNA dynamics and identify transiently formed binding pockets.[67–69] Similarly, MD simulations, complemented by enhanced sampling and experimental data integration, are increasingly used to generate conformational ensembles consistent with solution observations.[70] In parallel, RNA-specific docking protocols and scoring functions have emerged, outperforming protein-derived tools in reproducing experimentally validated RNA–ligand complexes. [69, 71–74]

AI and ML have also introduced a new tools in RNA-targeted drug discovery. data-

driven scoring functions like RNAmigos2 extend these advances to the prediction of RNA–small molecule interactions, using graph-based deep learning to capture pocket-specific recognition patterns.[75] Alongside scoring functions such as RNAmigos2, another notable contribution is BiteNet, a 3D convolutional neural network designed to detect ligand-binding sites in nucleic acid structures.[76] These advancements are rapidly consolidating into a robust framework that, when integrated with experimental pipelines, will expand the therapeutic space beyond proteins.

2 Aim of the work

The aim of this thesis is to contribute to RNA targeted small molecule discovery through computational means by assembling and validating a computational framework that progresses from comparative case studies to an application on a pharmaceutically relevant RNA. Chapter 3, Theoretical Background, provides the operative backbone of the work, with algorithms and methodologies used across projects. Building on this foundation, the projects increase in complexity. In the first project, described in Chapter 4, From Proteins to RNA: Comparative Analysis of Ligand Binding Determinants, we aim to investigate how a single ligand engages two distinct molecular contexts, a protein target and an RNA target. We select riboflavin as the common ligand and examine riboflavin kinase and the flavin mononucleotide riboswitch as representative systems that bind the same chemical entity. This design controls for intrinsic ligand properties and isolates the contribution of the binding environment. Specifically, the chapter addresses three questions. First, how do the conformational ensembles of the two targets differ in regions relevant for recognition, and how do these differences modulate the accessible binding modes. Second, what is the quantitative role of electrostatics in shaping interaction patterns and effective complementarity within the binding site. Third, how do ligand dynamics and water dynamics within the pockets contribute to stability and selectivity, including possible signatures of solvent rearrangement that accompany binding. In Chapter 5, Exploring RNA Druggability through SHAMAN, we adopt a drug discovery oriented perspective and target a challenging yet pharmaceutically relevant RNA, the HIV-1 trans activation response (TAR) element. We develop an exploration and prioritization workflow based on the recently introduced SHadow Mixed solvent metAdyNamics, SHAMAN, that maps binding hotspots within conformational ensembles, and ranks them according to druggability oriented criteria. The workflow then couples these maps to downstream virtual screening by selecting representative conformations, evaluating pocket quality with complementary predictors, seeking consensus with the mixed solvent maps, and guiding docking toward the most promising states. This chapter demonstrates that ensemble-aware pocket mapping is essential for flexible RNA targets, sharpens the focus of structure based screening, and delivers a final set of commercially available compounds that are selected for experimental testing to further validate the approach.

Finally, in Chapter 6, Integrative Refinement of RNA Conformational Ensembles via the Maximum Entropy Principle, we address a central challenge for HIV-1 TAR, namely the reconstruction of an experimentally validated conformational ensemble that is suitable for drug discovery campaigns. We begin from extensive enhanced sampling and refine the resulting distribution with a Maximum Entropy reweighting scheme to enforce agreement with independent observables while preserving the ensemble nature of the target. The objective is to obtain an experimentally compliant ensemble that can be analyzed and described. In parallel, by comparing ensemble derived pockets of unbound TAR with ligand bound structures available in the Protein Data Bank, we aim to assess the relative contribution of conformational selection and induced fit, clarifying how pre-existing conformations and ligand driven rearrangements cooperate to produce recognition competent states.

Together, these chapters aim to establish a coherent and reproducible path from methods to application that supports rational design on flexible RNA systems.

3 Theoretical background

3.1 Molecular Dynamics

Molecular Dynamics (MD) simulations are a powerful computational technique for investigating the structural and energetic evolution of molecular systems at atomic resolution over time. Over the past decades, their application in the pharmaceutical field has become increasingly consolidated.[61, 77, 78] Within this framework, atoms are treated as classical particles whose motions are governed by Newton's laws, and interatomic interactions are described by empirical potential energy functions, known as force fields. The system is modeled according to the principles of molecular mechanics (MM) and the Born-Oppenheimer approximation, which assumes that electronic motions can be decoupled from nuclear dynamics.[79]

Atoms are connected by virtual springs that mimic chemical bonds, allowing oscillations around equilibrium geometries. The force field encompasses both bonded (bonds, angles, dihedrals) and non-bonded (van der Waals and electrostatic) interactions:

$$\begin{aligned} V(r^N) = & \sum_{b \in \text{bonds}} k_b (l_b - l_b^0)^2 \\ & + \sum_{b \in \text{angles}} k_a (\theta_b - \theta_b^0)^2 \\ & + \sum_{b \in \text{torsions}} \frac{1}{2} V_b [1 + \cos(n_b \phi_b - \gamma_b)] \\ & + \sum_{i=1}^{N-1} \sum_{j=i+1}^N f_{ij} \left[\epsilon_{ij} \left(\left(\frac{r_{ij}^0}{r_{ij}} \right)^{12} - 2 \left(\frac{r_{ij}^0}{r_{ij}} \right)^6 \right) + \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \right] \end{aligned} \quad (3.1)$$

The parameters defining these interactions are typically derived from quantum mechanical calculations, experimental measurements, or a combination of both. Once the potential energy surface is established, the force on each atom is computed as the negative gradient of the potential:

$$F_i = -\frac{dV}{dr_i} = m_i \frac{d^2 r_i}{dt^2} \quad (3.2)$$

This equation yields the acceleration of each particle, from which velocities and positions are iteratively updated at each timestep according to the second law of motion:

$$a_i = \frac{d^2 r_i}{dt^2} \quad (3.3)$$

Initial atomic coordinates were obtained from experimentally determined structures (X-ray crystallography, NMR spectroscopy, cryo-EM) or, when unavailable, predicted with state-of-the-art tools such as AlphaFold [80, 81] and Boltz.[82, 83] Initial velocities were drawn from a Maxwell–Boltzmann distribution at the target temperature. The dynamics were propagated by numerically integrating the classical equations of motion with a fixed time step Δt , thereby generating time-ordered ensemble of configurations (i.e., trajectories) of molecular conformations that encode the dynamical behavior of the system.

One of the principal advantages of employing MD simulations for the characterization of molecular properties lies in their theoretical foundation on the ergodic hypothesis. According to this principle, provided that the simulation time is sufficiently long to enable exhaustive sampling of the configurational space, the time average of a given observable along the trajectory can be considered equivalent to its ensemble average. Therefore, the properties computed from the simulation are, in principle, comparable to those measured experimentally under experimental conditions.[84] MD simulations are widely used to elucidate complex biophysical phenomena, ranging from protein folding, allostery,[85] ligand-binding, kinetics[86] Intrinsically Disordered Proteins (IDPs) conformational dynamics,[87, 88] and binding-affinity estimation.[64]

When designing and conducting MD simulations, two fundamental aspects must be carefully considered: (i) the reliability of the employed force field and (ii) the associated computational cost. The force field, which defines the potential energy function governing interatomic interactions, critically influences the accuracy of the simulation

outcomes. As such, the more accurately a force field captures the underlying physical chemistry of the system, the more reliable will be the simulated observables. In the context of protein modeling, years of extensive refinement and validation have led to the establishment of highly accurate and standardized force fields. Conversely, the parameterization of RNA-specific force fields remains an area of active development.[23] Recent years have seen notable improvements in this domain, as discussed in Paragraph 3.1.1 .

A further challenge regard the convergence of MD simulations. Conformational transitions of biological macromolecules often occur on the millisecond timescale, whereas conventional equilibrium MD simulations are typically limited to the microsecond regime, despite recent advances in computational hardware and algorithms. As a consequence, routine MD may fail to adequately explore the full conformational landscape, thus violating the conditions required for the ergodic hypothesis to hold. In particular, the sampling of infrequent but biologically relevant conformational states may be severely limited. To address this limitation, several enhanced sampling strategies have been developed. These methods introduce an external biasing potential to accelerate the exploration of the conformational space while maintaining a reasonable computational cost. These methods are deeper discussed in the Paragraph 3.1.2.

3.1.1 RNA Force Fields

Molecular mechanics force fields provide an empirical description of interatomic interactions and constitute the foundation of classical molecular dynamics simulations. In the case of RNA, considerable efforts have been devoted to improving the accuracy of all-atom force fields, although the development of a fully general-purpose model remains incomplete.

In the RNA field, the most widely used family is AMBER (Assisted Model Building with Energy Refinement), whose Cornell et al. parametrization (AMBER ff94) has been progressively refined over the years (See Table A.1).[89–91] Notable improvements include the reparameterization of the α/γ backbone torsions in the parmbsc0 correction, which reduced unrealistic helical twists in RNA stems [92], and the reparameterization of the glycosidic χ torsion in the AMBER χ OL3 model, which prevented the formation of ladder-like structures by improving the syn/anti balance of

nucleobases. [93–95] Further more recent tuning of AMBER has focused on phosphate oxygen parameters,[96] hydrogen-bonding imbalances [97, 98], and the use of CMAP correction maps.[99] More recently, the NBfix_{0BPh} correction introduced a pair specific weakening of short range Lennard–Jones repulsion between intranucleotide base C–H donors and the O5' phosphate oxygen, which alleviates an artificial 0BPh steric clash, improves the syn/anti balance of nucleobases, and stabilizes native A form like conformations in short RNAs without detectable side effects on larger structured motifs.[100] Additionally, a targeted nonbonded adjustment that weakens the short range Lennard Jones repulsion for specific $\text{CH} \cdots \text{O}$ contacts, was recently introduced to stabilize native RNA stem loop structures without altering the remainder of the force field. [101] Another important contribution was introduced by the D. E. Shaw Research group, which developed the DESRES modification based on AMBER ff14.[102] This parametrization recalibrated nucleobase charges, torsional profiles, and Lennard–Jones parameters, improving the balance of base pairing and stacking interactions.[103]

Beyond AMBER, the CHARMM (Chemistry at Harvard Macromolecular Mechanics) community has established a parallel set of RNA force fields. Early versions, such as CHARMM22[104] and CHARMM27,[105] refined backbone dihedrals but suffered from excessive terminal base fraying. CHARMM36[106, 107] partially mitigated this issue, while more recent efforts introduced a polarizable force field based on the Drude oscillator, providing a more accurate description of charge-transfer phenomena that are essential for RNA–ion interactions and tertiary folding.[108]

3.1.2 Enhanced Sampling Methods

Despite the advancements in MD engines and computational resources, the exhaustive exploration of biomolecular free energy landscapes remains a major challenge. This limitation is primarily due to the presence of high free energy barriers that hinder transitions between metastable states, which are often crucial for understanding the functional behavior of biological macromolecules. To overcome this issue, a wide range of enhanced sampling methods has been developed. These methods aim to facilitate barrier crossing and improve the statistical characterization of the underlying structural ensemble by either (i) introducing a bias along a set of predefined slow

degrees of freedom, commonly referred to as collective variables (CVs),[109–111] or (ii) modifying the thermodynamic ensemble through strategies such as temperature scaling[112] or solute tempering.[113]

Among CV-based approaches, umbrella sampling[109] and metadynamics[110] as well as their developments[111] have gained substantial relevance due to their ability to dynamically adapt the biasing potential and reconstruct the underlying free energy surface in a computationally efficient manner. Metadynamics-based methods are often formulated within the theoretical framework of extended ensembles, where the simulation samples from a modified distribution that can be reweighted to recover equilibrium thermodynamic observables.[114]

In the following paragraphs, the theoretical foundations of metadynamics, well-tempered metadynamics, parallel tempering, and the well-tempered ensemble will be presented and discussed.

3.1.2.1 Metadynamics

Metadynamics (MetaD) is one of the most widely used enhanced sampling strategies based on collective variables (CVs).[115] It overcomes the intrinsic timescale limitations of standard molecular dynamics by promoting transitions between metastable states. The method introduces a history-dependent bias potential $V(S, t)$ as a function of a set of CVs $S(x) = (S_1, \dots, S_d)$, with x denoting the microscopic coordinates. This bias discourages revisiting already explored regions of the CV space, thereby promoting rare events and improving sampling of the free-energy landscape.

To drive the system out of local minima, the bias is constructed as a time-dependent sum of Gaussian functions deposited at regular intervals along the trajectory in CV space

$$V(S, t) = \sum_{k\tau < t} W(k\tau) \exp\left(-\sum_{i=1}^d \frac{[S_i(x) - S_i(x(k\tau))]^2}{2\sigma_i^2}\right), \quad (3.4)$$

where $W(k\tau)$ is the Gaussian height and σ_i the width along the i -th CV; the sum runs over past deposition times t' . At convergence, the free-energy surface $F(S)$ can be estimated from the accumulated bias as

$$F(S) = -\lim_{t \rightarrow \infty} V(S, t) + C, \quad (3.5)$$

with C an arbitrary additive constant. Accurate estimation requires that the bias has (quasi-)converged and that the chosen CVs capture the relevant slow modes; otherwise, orthogonal slow dynamics remain unsampled and the reconstructed free energy may be biased.

3.1.2.2 Well-Tempered Metadynamics

To control the unbounded growth and oscillations of the bias in standard metadynamics, the well-tempered formulation (WTMetaD) introduces a bias factor $\gamma = (T + \Delta T)/T \geq 1$.^[111]

As in standard MetaD, the history-dependent bias is written as a sum of Gaussians deposited at regular intervals τ (Eq. 3.4). However, in well-tempered metadynamics, the Gaussian height is attenuated according to the bias accumulated locally in CV space:

$$W(k\tau) = W_0 \exp \left[- \frac{V(S(x(k\tau)), k\tau)}{k_B \Delta T} \right]. \quad (3.6)$$

Under mild conditions, the long-time bias converges to a scaled version of the underlying free energy:

$$V(S, t \rightarrow \infty) = - \frac{\Delta T}{T + \Delta T} F(S) + C = - \left(1 - \frac{1}{\gamma} \right) F(S) + C. \quad (3.7)$$

As in all CV-based methods, keeping the dimensionality d as low as possible, while retaining the essential slow modes, facilitates repeated transitions and robust free-energy estimation.

3.1.2.3 Parallel Tempering

Parallel tempering (PT), also known as replica exchange molecular dynamics (REMD), is a widely used non-CV-based enhanced sampling technique.^[112] Unlike methods relying on collective variables, which require the user to identify suitable slow degrees of freedom, PT acts globally on the entire system. The basic idea is to simulate in parallel N replicas of the system at different temperatures $\{T_1, T_2, \dots, T_N\}$. At fixed intervals, exchanges of configurations between adjacent replicas are attempted and accepted according to a Metropolis criterion.^[112]

Given two replicas i and j at temperatures T_i and T_j , with configurations x_i and x_j of potential energies $U(x_i)$ and $U(x_j)$, respectively, the acceptance probability reads

$$P(i \rightarrow j) = \min[1, \exp(\Delta w_{ij})], \quad (3.8)$$

with

$$\Delta w_{ij} = (\beta_i - \beta_j) (U(x_j) - U(x_i)), \quad (3.9)$$

where $\beta = 1/(k_B T)$ is the inverse temperature.

At high-temperature replicas, energy barriers are more easily overcome, thus promoting configurational transitions that would be extremely rare at the lowest temperature. The replica at the lowest temperature, T_1 , corresponds to the physical system of interest and samples the canonical distribution, while benefiting from exchanges with higher temperature replicas that broaden its sampling.

The efficiency of PT depends on the overlap between the potential energy distributions of adjacent replicas. This is quantified by the acceptance ratio, i.e. the fraction of successful exchanges over attempted ones. An optimal setup would ensure nearly constant acceptance ratios across the temperature ladder, leading to efficient “round trips,” defined as the time required for a replica to diffuse from the lowest to the highest temperature and back. Since energy fluctuations grow with temperature, the temperature spacing between replicas can be progressively increased along the ladder. A geometric progression is often employed:

$$T_i = T_{\min} \left(\frac{T_{\max}}{T_{\min}} \right)^{\frac{i-1}{N-1}}, \quad i = 1, \dots, N, \quad (3.10)$$

where $T_{\min}$ and $T_{\max}$ denote the lowest and highest temperatures, respectively.

Despite its robustness, PT can be computationally demanding because covering a broad temperature range with sufficient overlap often requires a large number of replicas. This practical limitation can be alleviated by combining PT with the well-tempered ensemble, as discussed in the next paragraph.

3.1.2.4 Well-Tempered Ensemble

The well-tempered ensemble (WTE) is obtained by applying WTMetaD to the potential energy, i.e. by choosing the CV as $S(x) = U(x)$. Using Eq. (3.7) with $S = U$, the long-time bias reads

$$V(U, t \rightarrow \infty) = -\left(1 - \frac{1}{\gamma}\right) F(U) + C, \quad (3.11)$$

where the free energy in energy space is

$$F(U) = U - \beta^{-1} \ln N(U) + C', \quad (3.12)$$

with $N(U) = \int dR \delta(U - U(R))$ denoting the configurational density of states. By integrating over configurations, the stationary energy distribution sampled in WTE is a tempered version of the canonical one

$$P_\gamma(U) \propto \left[N(U) e^{-\beta U} \right]^{1/\gamma}, \quad (3.13)$$

which corresponds to the partition function

$$Z_\gamma = \int dU \left[N(U) e^{-\beta U} \right]^{1/\gamma}. \quad (3.14)$$

As a consequence, energy fluctuations are enhanced in a γ -controlled manner, improving barrier crossing while preserving unbiased thermodynamics via standard reweighting.

3.1.3 Maximum Entropy Integration of Simulations and Experiments

Integrating molecular simulations with experimental data is a powerful way to improve the predictive accuracy of computational models, especially for flexible and heterogeneous systems such as IDPs and RNA. Among available strategies, the Maximum Entropy (ME or MaxEnt) formalism provides a rigorous framework that minimally perturbs an MD-derived prior ensemble while enforcing agreement with experimental observables. This is particularly relevant for flexible biological targets, where preserving the ensemble nature is essential both for understanding their behavior and for effective ligand design.

In general, the entropy of a probability distribution $P(\mathbf{q})$ relative to a given prior $P_0(\mathbf{q})$ is defined as

$$S[P||P_0] = - \int d\mathbf{q} P(\mathbf{q}) \ln \frac{P(\mathbf{q})}{P_0(\mathbf{q})}. \quad (3.15)$$

where $\mathbf{q}$ is any set of continuous variables describing the system.

Given a set of experimental constraints s_i^{exp} , the optimal posterior distribution $P_{\text{ME}}(\mathbf{q})$ is obtained by maximizing the Shannon entropy relative to the prior $P_0(\mathbf{q})$, while reproducing the ensemble averages of the M measured observables and preserving normalization. This constrained optimization problem can be written as

$$\begin{cases} P_{\text{ME}}(\mathbf{q}) = \arg \max_{P(\mathbf{q})} S[P||P_0] \\ \int d\mathbf{q} s_i(\mathbf{q}) P(\mathbf{q}) = \langle s_i(\mathbf{q}) \rangle = s_i^{\text{exp}}, \quad i = 1, \dots, M \\ \int d\mathbf{q} P(\mathbf{q}) = 1 \end{cases} \quad (3.16)$$

where $s_i(\mathbf{q})$ is the ensemble average of M computed over $P(\mathbf{q})$. The problem can be solved by introducing Lagrange multipliers, searching for the stationary point of the associated Lagrangian

$$\mathcal{L} = S[P||P_0] - \sum_{i=1}^M \lambda_i \left(\int d\mathbf{q} s_i(\mathbf{q}) P(\mathbf{q}) - s_i^{\text{exp}} \right) - \mu \left(\int d\mathbf{q} P(\mathbf{q}) - 1 \right), \quad (3.17)$$

where λ_i and μ are Lagrange multipliers. The functional derivative of $\mathcal{L}$ with respect to $P(\mathbf{q})$ is

$$\frac{\delta \mathcal{L}}{\delta P(\mathbf{q})} = - \ln \frac{P(\mathbf{q})}{P_0(\mathbf{q})} - 1 - \sum_{i=1}^M \lambda_i s_i(\mathbf{q}) - \mu. \quad (3.18)$$

Setting the functional derivative to zero, the posterior belongs to the exponential family; up to normalization it reads:

$$P_{\text{ME}}(\mathbf{q}) \propto e^{- \sum_{i=1}^M \lambda_i s_i(\mathbf{q})} P_0(\mathbf{q}) \quad (3.19)$$

In practice, it can be incorporated into MD simulation in different ways. It is possible to enforce the ensemble on the fly or to re-weight a posteriori the ensemble. In this thesis, the second option will be used. For a molecular dynamics trajectory, consider an ensemble of N_s structures. We assign a new weight to each configuration x_t as

$$w_t \propto e^{-\sum_{i=1}^M \lambda_i^* s_i(x_t)}, \quad (3.20)$$

and we normalize the weights so that $\sum_{t=1}^{N_s} w_t = 1$.

Here $s_i(x_t)$ is the value of the i th observable computed for structure x_t , and λ_i^* denotes the optimal Lagrange multipliers that minimize the following quantity:

$$\Gamma(\boldsymbol{\lambda}) = \ln \left(\frac{1}{N_s} \sum_{t=1}^{N_s} e^{-\sum_{i=1}^M \lambda_i s_i(x_t)} \right) + \sum_{i=1}^M \lambda_i s_i^{\text{exp}} + \Gamma_{\text{err}}(\boldsymbol{\lambda}). \quad (3.21)$$

The last term models experimental error, acts as a regularization that mitigates overfitting, and is defined as

$$\Gamma_{\text{err}}(\boldsymbol{\lambda}) = \frac{1}{2} \sum_{i=1}^M \lambda_i^2 \sigma_i^2, \quad (3.22)$$

where σ_i specifies the uncertainty associated with observable s_i .

This framework has been applied in recent years to a broad spectrum of experimental data, spanning CryoElectron Microscopy (Cryo-EM) density maps[116] to Small-Angle X-ray Scattering.[88, 117, 118] In all these cases a forward model is required to compute observables for each molecular configuration along the MD trajectory. Such models are often empirical and of limited accuracy, which introduces systematic errors in the calculated observables. This limitation, present for proteins, is even more pronounced for RNA systems

After reweighting, only a subset of prior configurations contributes meaningfully to the posterior ensemble. This concentration can be quantified by the Kish effective sample size

$$K = \frac{(\sum_{t=1}^{N_s} w_t)^2}{\sum_{t=1}^{N_s} w_t^2}. \quad (3.23)$$

With normalized weights, $\sum_{t=1}^{N_s} w_t = 1$, this reduces to $K = 1 / \sum_{t=1}^{N_s} w_t^2$. Uniform weights give $K = N_s$ while a distribution dominated by a single configuration gives $K \approx 1$.

3.1.4 Mixed Solvent Approaches

In structure-based drug discovery (SBDD), identifying ligand-binding hotspots is critical for molecular recognition and for prioritizing and optimizing bioactive compounds. Since its introduction in 2009,[119] mixed-solvent MD has become a versatile tool for early discovery, enabling binding-site detection and ligandability assessment.[120] Its fully dynamic nature also facilitates the discovery of cryptic pockets and allosteric sites, often missed by conventional approaches.[85, 121, 122] The resulting interaction maps can guide lead optimization or feed downstream pipelines (e.g., pharmacophore modeling and docking),[123–125] while explicitly characterizing solvent-mediated interactions and the role of water in binding.

The roots of mixed-solvent mapping trace back to Goodford’s GRID (1985), which projected probe atoms onto a grid to locate favorable interaction regions (Figure 3.1A) and laid the basis for Molecular Interaction Field (MIF) analysis [126, 127].

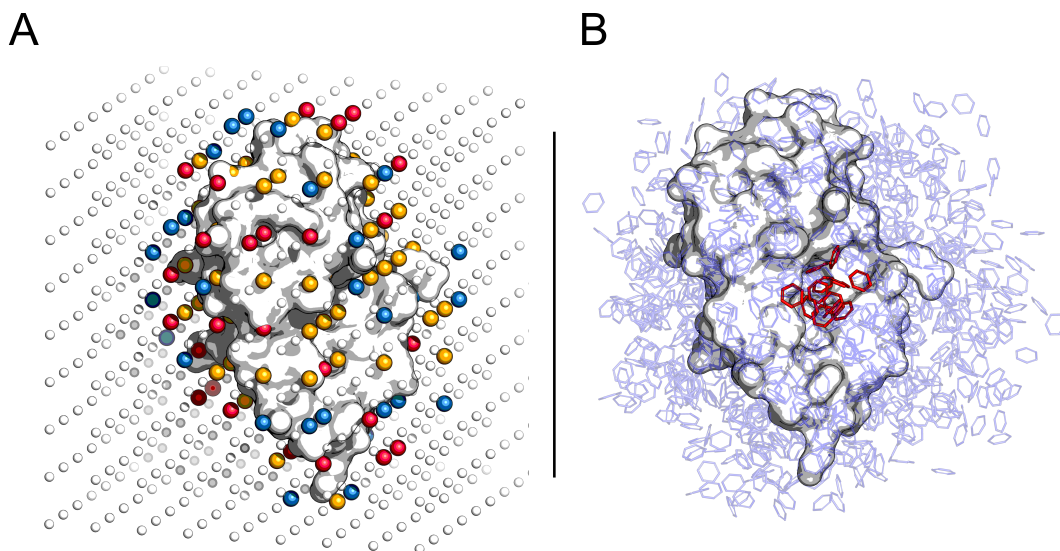


Figure 3.1. (A) Classical GRID approach: a cubic lattice surrounds the protein surface and nodes within 6 Å are colored according to probe type (DRY: hydrophobic, orange; N1: H-bond donor, blue; O: H-bond acceptor, red). (B) Multiple-copy simultaneous search (MCSS) with benzene probes placed on the protein surface.

GRID-like methods proved successful [128–130] but are limited by their static treatment of flexibility and solvation, two determinants of druggability. [131] To address these issues, Karplus and Miranker introduced MCSS (1991), placing multiple non-interacting probes on protein surfaces (Figure 3.1B). [132] Variants combining MCSS with Locally Enhanced Sampling extended the idea to MD, [133, 134] opening the way to mixed-solvent (cosolvent) MD, where small organic probes in explicit solvent dynamically explore biomolecular surfaces to reveal hotspots (Figure 3.2).

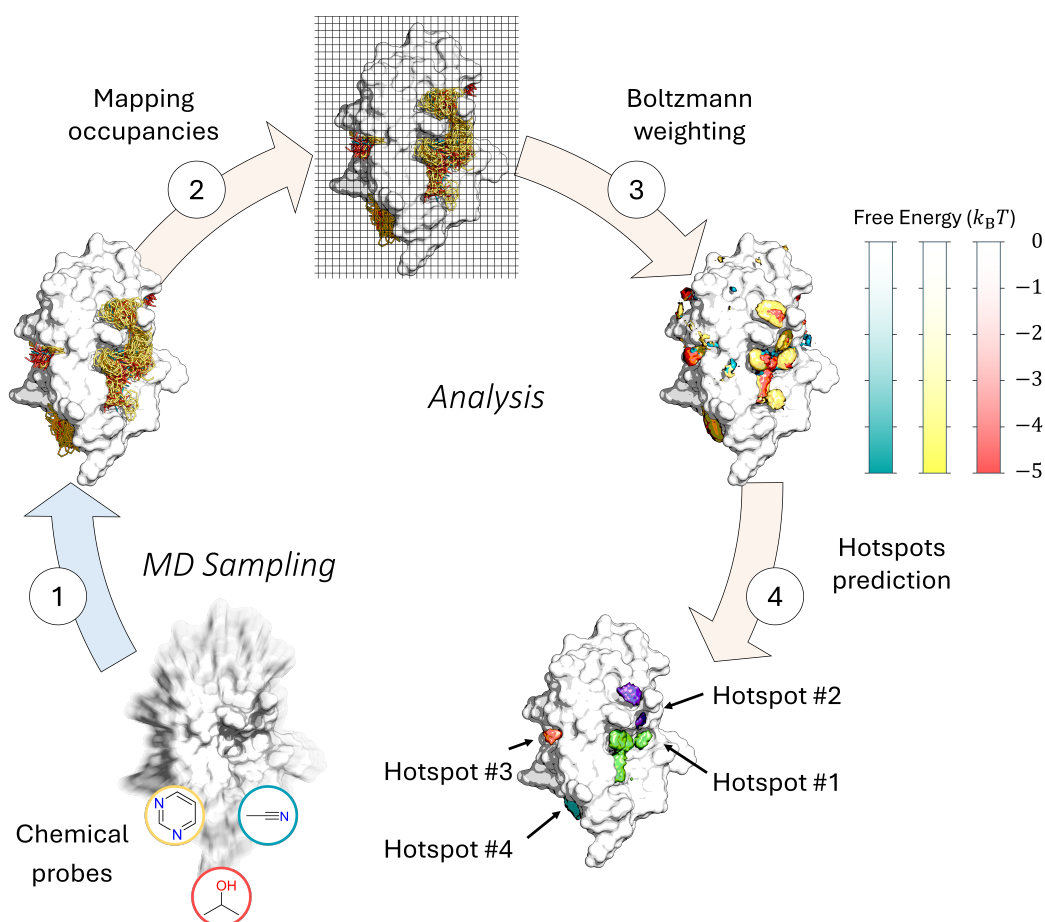


Figure 3.2. Mixed-solvent MD simulations: small organic cosolvents are added to explicit solvent and monitored during MD to map binding hotspots and cryptic pockets on biomolecules.

Despite their advantages, mixed-solvent MD simulations face potential difficulties, such as probe aggregation, phase separation, or even target destabilization, which may compromise the reliability of the analysis.[121] In the following paragraph, the

foundations of the Shadow Mixed-Solvent Metadynamics (SHAMAN) will be discussed.

3.1.4.1 SHadow Mixed-solvent metAdyNamics (SHAMAN)

Shadow Mixed-Solvent Metadynamics (SHAMAN) was designed to identify small-molecule binding sites in dynamic RNA ensembles.[135] It integrates atomistic MD with WTMetaD, employing a central “mother” simulation of RNA in explicit solvent, complemented by multiple “shadow” replicas containing probes. The probe set is user-defined; several probes have been extensively characterized by the developers. [135, 136] The RNA backbone in the shadows is restrained to the mother trajectory, ensuring consistent probe exploration of the RNA surface and accelerating convergence.

To enhance probe RNA sampling, WTMetaD[110, 111] is applied, biasing probe center of mass coordinates. At the end of the simulation, the unbiasing weights are computed as

$$w_t \propto \exp\left(\frac{V_G(S(\mathbf{R}_t), \bar{t})}{k_B T}\right), \quad (3.24)$$

where $V_G(S(\mathbf{R}_t), \bar{t})$ is the well tempered metadynamics potential accumulated at the end of the simulation $\bar{t}$ and evaluated on the conformation $\mathbf{R}_t$.

After simulation, RNA conformations are clustered to obtain representative states. Probe occupancies are converted into voxel-wise binding free energies (ΔG_i) through the Boltzmann inversion:

$$\Delta G_i^{bind} = -k_B T \ln \left(\frac{N_i}{N_0} \right), \quad (3.25)$$

where k_B is the Boltzmann constant, T the absolute temperature, N_i denotes the probe occupancy in voxel i , defined as the sum over all probe atoms of the normalized metadynamics unbiasing weights (Eq. 3.24) over all simulation frames in which a given atom is found within voxel i , and N_0 the expected counts in the absence of interactions:

$$N_0 = n_{\text{probe}} \frac{V_{\text{voxel}}}{V_{\text{MD}}} \quad (3.26)$$

with V_{MD} the total simulation volume, V_{voxel} the volume of the voxels, and n_{probe} the number of probe atoms.

Voxels with highly negative ΔG_i values indicate regions of high probe affinity, which are subsequently clustered into candidate binding sites (SHAMAPs) that aggregate signal across all probe types. These SHAMAPs are ranked by their relative binding free energy ($\Delta\Delta G_s$) and filtered based on solvent exposure to retain only biologically relevant pockets.

The SHAMAN pipeline (setup, trajectory generation, clustering, free-energy mapping, and SHAMAP analysis) represents the first open-source framework for assessing RNA druggability in dynamic ensembles. The shadow-replica mechanism has been implemented in PLUMED,[137–139] freely available at <https://www.plumed.org/>.

3.2 Computational strategies for ligand screening and prioritization

3.2.1 Theoretical Foundations of Molecular Docking

Molecular docking is a computational technique designed to predict the most probable binding mode of a small molecule when interacting with a macromolecular target, such as a protein or RNA. The principal aim of docking is to provide a reliable structural model of the ligand–target complex and to estimate the associated binding affinity.[140] This process relies on the physical and chemical complementarity between the ligand and the binding site, and involves two essential components: a search algorithm and a scoring function.

From a theoretical standpoint, docking can be conceptualized as an optimization problem in a high-dimensional conformational space. The translational, rotational, and internal degrees of freedom of the ligand are systematically or stochastically explored to identify energetically favorable binding poses. Search algorithms used in molecular docking can be broadly classified into two categories: systematic and stochastic.[141, 142] Systematic approaches exhaustively cover the search space using deterministic rules but are computationally demanding. Stochastic methods, including genetic algorithms, simulated annealing, and Monte Carlo sampling, introduce randomness to enhance efficiency while maintaining sufficient coverage of the

conformational landscape. More recently, hybrid approaches incorporating machine learning have been developed to accelerate docking and improve accuracy.[143–145]

The second crucial component of docking is the scoring function, which estimates the binding affinity of each candidate pose. Scoring functions are typically grouped into three broad classes: force-field-based, empirical, and knowledge-based. Force-field-based functions derive from molecular mechanics and compute non-bonded interactions such as van der Waals forces and electrostatics, often including approximate treatments of solvation and entropy. Empirical scoring functions combine weighted terms fitted to experimental binding data, including hydrogen bonding, hydrophobic contacts, and desolvation energies. Knowledge-based approaches rely on statistical potentials extracted from large datasets of experimentally determined ligand–target complexes, leveraging recurrent structural motifs observed in crystallographic data.

A significant limitation of most docking protocols lies in their simplified treatment of molecular flexibility. Typically, the receptor is modeled as rigid or semi-flexible, simplifying the calculation but compromising biological realism. Several advanced methods attempt to address this constraint. Ensemble docking, for instance, employs multiple receptor conformations obtained from molecular dynamics simulations or experimental structures to account for conformational heterogeneity. Induced fit docking (IFD) methods, such as those implemented in Glide,[146] iteratively refine the protein–ligand complex by allowing limited receptor flexibility during the docking process.

3.2.2 The RNAmigos2 Virtual Screening Pipeline

Beyond docking, numerous machine-learning methods have been developed to predict ligand binding poses or affinities. Among these, RNAmigos2 stands out for its performance on RNA targets. RNAmigos2 is a machine learning framework designed to rank candidate ligands against RNA binding pockets by combining graph representations of RNA structure with learned molecular embeddings. [75, 147] Each RNA structure is represented as a 2.5D graph (a 2D base-pairing graph augmented with local three-dimensional features) that captures canonical and non-canonical interactions according to the Leontis–Westhof classification,[148] and is invariant to global rigid-body transformations. A graph neural network encoder, trained with self-supervision

on unlabeled RNA substructures, converts this graph into a continuous vector representation.

Ligands are represented as molecular graphs and embedded using a variational autoencoder pretrained on ChEMBL compounds. Two decoding heads are employed: a compatibility head trained to distinguish native ligands from property-matched decoys, and an affinity head trained on a synthetic dataset of pose-independent docking scores. During inference, the two scores are integrated to produce a final ranking that can be applied to libraries containing up to hundreds of thousands of molecules within second. [75] Because RNAmigos2 operates in a pose-free manner, its predictions can be combined with downstream docking to refine the top-ranked compounds, thereby reducing the overall computational cost, or used in parallel to explore complementary regions of chemical space.

3.3 Unsupervised Machine Learning for MD simulations

3.3.1 Dimensionality Reduction

MD simulations produce high-dimensional time-resolved data that describe the atomic motions of biomolecular systems. While such simulations provide atomistic detail, the large number of degrees of freedom often obscures the identification of collective behaviors, long-timescale processes, and functionally relevant conformational changes. To address this challenge, dimensionality reduction techniques have become essential tools for extracting interpretable features and simplifying the analysis of complex trajectory data.[149, 150]

The general goal of dimensionality reduction is to map the original data, defined in a high-dimensional space, onto a low-dimensional manifold while retaining as much of the relevant structural or dynamical information as possible. In the context of MD simulations, this allows the visualization and clustering of conformational ensembles, the detection of slow collective variables, and the construction of reaction coordinates.

Numerous dimensionality reduction algorithms have been proposed, each grounded in distinct theoretical formulations, ranging from linear algebraic pro-

jections to probabilistic embedding schemes, designed to capture relevant features of high-dimensional data. In the present thesis, and for methodological focus, only the techniques directly employed in the following projects are described. These include Principal Component Analysis (PCA), a linear method that projects the data onto orthogonal directions of maximal variance, and t-Distributed Stochastic Neighbor Embedding (t-SNE), a nonlinear approach that constructs a low-dimensional representation by preserving local neighborhood probabilities derived from the high-dimensional geometry.

3.3.1.1 Principal Component Analysis (PCA)

PCA aims to find a set of orthogonal vectors, referred to as principal components (PCs), that define directions along which the variance of the data is maximized. Given a dataset $\mathbf{X} \in \mathbb{R}^{N \times D}$, where N is the number of frames and D the number of features (e.g., Cartesian coordinates, internal distances, or dihedral angles), the procedure consists of the following steps:

1. The data matrix is centered by subtracting the mean of each feature, resulting in a zero-centered matrix $\hat{\mathbf{X}}$.
2. The empirical covariance matrix is computed as

$$\mathbf{C} = \frac{1}{N} \hat{\mathbf{X}}^\top \hat{\mathbf{X}}. \quad (3.27)$$

3. The eigenvectors and eigenvalues of $\mathbf{C}$ are then obtained by solving the associated eigenvalue problem.

By retaining only the top $d \ll D$ eigenvectors corresponding to the largest eigenvalues, one obtains a projection matrix $\mathbf{V} \in \mathbb{R}^{D \times d}$ that allows the dimensionality-reduced representation of the trajectory as

$$\mathbf{T} = \mathbf{XV}. \quad (3.28)$$

This transformation allows the identification of dominant motions and facilitates the visualization of conformational landscapes in reduced dimensions.

3.3.1.2 t-Distributed Stochastic Neighbor Embedding (t-SNE)

Unlike PCA, which captures global variance, t-SNE is a nonlinear method that aims to preserve the local structure of the data. It constructs a probability distribution over pairs of points in the high-dimensional space, such that similar configurations have high conditional probabilities of being neighbors. A corresponding distribution is then defined in the low-dimensional space, and the embedding is optimized by minimizing their divergence.

Formally, the conditional probability that point j is a neighbor of point i in the original space is

$$P_{j|i} = \frac{K_{ij}}{\sum_{k \neq i} K_{ik}}, \quad (3.29)$$

where K_{ij} is a Gaussian kernel applied to the distance between x_i and x_j . The perplexity parameter controls the effective number of neighbors by tuning the Gaussian variance.

A symmetric joint distribution is then defined as

$$\bar{P}_{ij} = \frac{P_{j|i} + P_{i|j}}{2N}. \quad (3.30)$$

In the low-dimensional space, similarities between points y_i and y_j are computed using a heavy-tailed Student's t-distribution:

$$Q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq l} (1 + \|y_k - y_l\|^2)^{-1}}. \quad (3.31)$$

The embedding minimizes the Kullback–Leibler divergence between the two distributions:

$$\text{KL}(\bar{P}||Q) = \sum_{i \neq j} \bar{P}_{ij} \log \frac{\bar{P}_{ij}}{Q_{ij}}, \quad (3.32)$$

using gradient descent.

3.3.2 Clustering

Clustering of MD trajectories provides a fundamental strategy to transform the very large set of raw configurations into an ensemble of structurally representative

states. This reduction is essential for two complementary reasons. First, it reduces the computational burden associated with downstream analyses, whose complexity increases with the number of frames. Second, it enhances statistical interpretability by assigning explicit populations to clusters that can be related to metastable regions on the underlying free energy surface. In the following, three widely used clustering algorithms in MD are reviewed: K-means, the RMSD cutoff clustering introduced by Daura *et al.*, [151] and Quality Threshold clustering. [152, 153]

K-means K-means partitions the set $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^N$ of molecular conformation, where each of them is represent by a d -dimensional vector, into k disjoint clusters $\mathcal{C}_j$, $j = 1, \dots, k$. The objective is to minimize the within-cluster sum of squared Euclidean distances

$$J = \sum_{j=1}^k \sum_{\mathbf{x}_i \in \mathcal{C}_j} \|\mathbf{x}_i - \boldsymbol{\mu}_j\|^2, \quad (3.33)$$

where $\boldsymbol{\mu}_j$ denotes the centroid of cluster $\mathcal{C}_j$. The Lloyd–MacQueen procedure alternates assignment and update steps until J converges within a tolerance on its relative variation. Centroid initialization strongly influences convergence. Algorithmic complexity scales as $\mathcal{O}(Nkd)$ per iteration, which remains tractable after prior dimensionality reduction through time-lagged independent component analysis.

Cutoff clustering by Daura The clustering method introduced by Daura *et al.* [151] is based on a cutoff in the distance matrix, originally defined on the backbone RMSD. Let $\mathbf{D}$ denote the pairwise distance matrix, with elements $D_{ij} = \text{RMSD}(\mathbf{x}_i, \mathbf{x}_j)$. For a defined cutoff distance d_c , a binary adjacency matrix is introduced as

$$A_{ij} = \begin{cases} 1 & \text{if } D_{ij} \leq d_c, \\ 0 & \text{otherwise,} \end{cases} \quad (3.34)$$

where A_{ij} indicates whether structures i and j are neighbors. For each configuration $\mathbf{x}_i$, the neighbor count is

$$n_i = \sum_{j=1}^N A_{ij}. \quad (3.35)$$

At each iteration, the configuration with maximal neighbor count is chosen as the cluster center,

$$i^* = \arg \max_i n_i, \quad (3.36)$$

and together with its neighbors defines a cluster. These frames are removed from the pool and the procedure repeats until exhaustion. Frames with no neighbor within d_c form singleton clusters. Computational demand scales as $\mathcal{O}(N^2)$ to build $\mathbf{D}$. The method does not require predefining the number of clusters, but instead relies on a physically interpretable threshold. This preserves the intrinsic heterogeneity of the conformational ensemble, although kinetically distinct states may be merged when their structural dissimilarity lies below d_c .

Quality Threshold clustering Quality Threshold (QT) clustering[152, 153] seeks clusters of maximal size whose internal diameter does not exceed a given cutoff $d_{\max}$. Given the pairwise distance matrix $\mathbf{D}$, the diameter of a candidate cluster $\mathcal{C}$ is defined as

$$\text{diam}(\mathcal{C}) = \max_{i,j \in \mathcal{C}} D_{ij}. \quad (3.37)$$

For each seed i , the set of neighbors within $d_{\max}$ is enumerated:

$$\mathcal{N}_i = \{j \mid D_{ij} \leq d_{\max}\}. \quad (3.38)$$

Among all candidate sets $\mathcal{N}_i$ satisfying $\text{diam}(\mathcal{N}_i) \leq d_{\max}$, the one with largest cardinality is selected,

$$\mathcal{C}^* = \arg \max_{\mathcal{N}_i} |\mathcal{N}_i|, \quad (3.39)$$

and its members are removed from further consideration. The process iterates until no further cluster larger than a singleton can be formed. By construction, every cluster has maximal diameter bounded by $d_{\max}$, and the method is deterministic for a fixed $d_{\max}$. The principal drawback of QT is its high computational cost, though pruning heuristics reduce the average complexity. The method avoids imposing an arbitrary number of clusters and ensures uniform structural quality, at the expense of reduced scalability.

4 From Proteins to RNA:

Comparative Analysis of Ligand Binding Determinants

As discussed in the Introduction, Chapter 1, RNA performs several functional and regulatory roles across cellular processes, which motivates therapeutic strategies that act directly on RNA (See Section 1.2). Here we deliberately narrow the focus to small molecule approaches,[33] which leverage medicinal chemistry optimization to tune pharmacodynamic and pharmacokinetic properties.[65] Recent advances and case studies, among them the approval of Risdiplam[57] and earlier precedents such as Linezolid,[154] sustain the interest in this route.[4, 33, 53, 155–161] At the same time, most RNA targeting small molecules reported so far originate from phenotypic screening rather than from structure based design,[33, 57] which underscores the need for a deeper mechanistic understanding of RNA recognition by drug like ligands.

Several features of RNA complicate the discovery of selective binders. The dense anionic backbone limits the formation of deeply hydrophobic sites,[158] and the chemical variability is less enhanced than in proteins.[162] Many RNAs are best represented as conformational ensembles with pronounced dynamics,[67] which challenges the classical techniques to obtain three dimensional structures of RNA ligand complexes and the direct transfer of structure based design practices from the protein field. Nevertheless, highly folded RNAs can form cavities reminiscent of protein pockets and can undergo ligand induced rearrangements.[33, 158, 163, 164] Concentrating on these classes, including riboswitches and ribozymes, is therefore a promising strategy. In this context, the integration of experiments and molecular simulations has proved effective both for characterizing ensembles and for diagnosing force field limitations that impact predictive modeling.[23, 84]

These constraints do not preclude rational design, but they require approaches that can accommodate flexibility, solvent exposure, and limited chemical diversity. Within this context, computational methods constitute a central component of modern discovery programs.[61] Molecular docking provides rapid exploration of putative binding

modes and prioritization, originally developed for proteins and increasingly adapted to RNA targets through revised protocols and scoring functions.[34, 50, 165–167] Molecular dynamics complements docking by resolving conformational heterogeneity and by offering mechanistic insight into binding, site discovery, and population shifts.[86, 122, 168–179] In the RNA domain, applications have mainly addressed ensemble reconstruction and force field assessment, with fewer examples directly aimed at ligand discovery.[68, 180, 181] This gap motivates a careful comparison between protein and RNA systems under controlled conditions.

In this chapter we therefore investigate how a single ligand engages two molecular contexts, a protein target and an RNA target. We select riboflavin as the common ligand, and study riboflavin kinase (RK) and the flavin mononucleotide riboswitch (FMN) as representative systems that bind the same chemical entity.[182, 183] This design controls for the intrinsic properties of the ligand and highlights the contribution of the binding environment. Our analysis addresses three interrelated questions. First, how do the conformational ensembles of the two targets differ in the regions relevant for recognition, and how do these differences modulate accessible binding modes. Second, what is the quantitative role of electrostatics in shaping the interaction pattern and the effective complementarity in the binding site. Third, how do ligand dynamics and water dynamics within the pockets contribute to stability and selectivity, including possible signatures of solvent reorganization concurrent with ligand binding.[184]

The remainder of the chapter is organized around these questions. We begin by presenting the systems and the structural context of recognition, then analyze the conformational space explored by each target, and finally examine interaction patterns and solvent behavior in the binding sites. This progression allows us to delineate common principles and system specific features that are relevant for structure based discovery on RNA, while maintaining a direct comparison with the protein benchmark. The systems considered are shown in Figure 4.1, with riboflavin highlighted for reference.

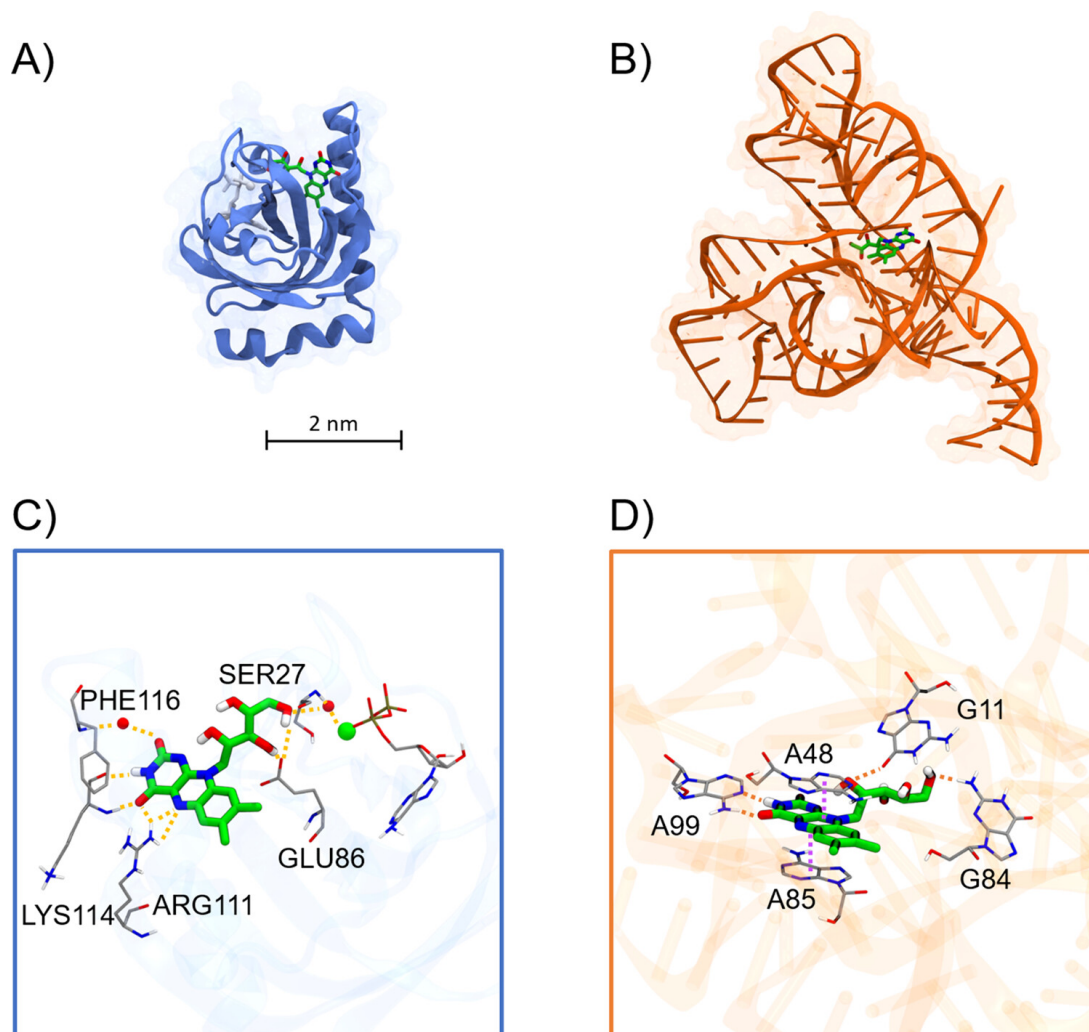


Figure 4.1. Systems considered in this study: (A) the riboflavin kinase (PDB ID 1NB9 [182]) and (B) the FMN riboswitch (PDB ID 3F4G [183]). The protein and RNA molecules are shown in the NewCartoon representation of the Visual Molecular Dynamics (VMD) software, [185] while the riboflavin ligand is displayed in licorice (with ligand carbon atoms highlighted in green). The two biomolecules are shown in scale. In the corresponding panels below, we report details of the (C) protein- and (D) RNA-ligand interactions in the crystal structures, with hydrogen bonds and $\pi - \pi$ stackings shown as dashed lines in yellow and purple, respectively. The interacting residues from the target are labeled and are shown in licorice. In the RF kinase catalytic site (C), the ADP is shown in licorice, while the magnesium ion is shown in green.

4.1 Methods

4.1.1 System Setup and Simulation Protocol

Initial structures for each biomolecule were obtained from the protein data bank (PDB ID 1NB9 [182] and 3F4G [183] for the kinase and riboswitch structure, respectively). The unbound states were obtained by removing the original ligand. All the structures were preliminarily processed using the Protein Preparation Wizard module of the Schrodinger suite.[186, 187] In particular, missing side chains were filled with Prime,[188] and the protonation and tautomeric states of residues at pH 7.0 were assigned with Epik. [187, 189] Simulation boxes were constructed using the standard Ambergtools [190] and GROMACS modules by solvating the systems with OPC waters[191] and adding a 100 mM concentration of KCl. The KCl concentration was chosen according to the experimental conditions that are often used to study riboswitches [52, 165, 192, 193] and that were used to measure the affinity (KD) of the riboflavin ligand for the FMN riboswitch. [183] The same was used for the RF kinase for consistency. The AMBER force field was used for protein (ff19sb)[194] and RNA (OL3), [195] the OPC [191] model was used for water molecules, and the parameters for magnesium were taken from Villa et al. [196] The RESP scheme [197] was applied to parametrize both the riboflavin in its neutral state, as predicted by Epik, and the ADP, using Gaussian 16 [198] for computing the molecular electrostatic potential. van der Waals and electrostatic interactions were truncated at 12 Å, while long-range electrostatic interactions were treated with the particle-mesh Ewald method using a grid spacing of 1.6 Å. Minimization of each system was performed with 5000 steps of steepest descent. The equilibration was carried out for each of the systems for a total of 1.15 ns, comprising gradual increase of the temperature to 300 K and equilibration of pressure at 1 bar. Production runs were performed in the NPT ensemble. The GROMACS 2021.3 [199–201] MD engine was used for all the simulations with a time step of 2 fs and the LINCS algorithm to constrain bonds. [202] The Nosé–Hoover thermostat [203] was used to control the temperature at the reference value with a time constant for coupling of 1 ps, while constant pressure was maintained at 1 bar using the C-rescale barostat [204] with a time constant of 0.5 ps. For each of the two systems, three independent runs of 1 μs were performed in the bound and unbound states, amounting to a total simulation time of 12 μs.

4.1.1.1 Water Relaxation Time

We estimated water residence time in the riboflavin binding sites using a survival probability (SP) function, which quantifies the relaxation time of the water molecules inside binding pockets. Specifically, SP measures the probability that a group of particles, herein the water molecules, will remain in a specified region, herein the hydration shell of the binding cavity, for a certain amount of time t , and is defined as

$$SP(t; r) = \frac{1}{T - t} \sum_{j=1}^{N_w(r)} \sum_{t_0=0}^{T-t} P_j(t_0, t_0 + t; t^*) \quad (4.1)$$

In the above formula, T is the total simulation time, while $N_w(r)$ is the total number of water molecules, which is in turn a function of the radius of the hydration shell r [205–208] centered on the binding site. $P_j(t_0, t_0 + t; t^*)$ is an indicator function that equals 1 if the j th water molecule found within the hydration shell at time t_0 has not left after a certain lag time t (i.e., in the time period between t_0 and $t_0 + t$), except for a time $t^* \ll (t_0 + t)$, and equals 0 otherwise. In our analyses, we considered a hydration shell with radius 10 Å centered in the binding pockets of both systems. To capture fine water dynamics with high statistical significance, we computed SP on MD trajectories saved every 5 ps. The calculation of SP was achieved via the Survival Probability module of the MDAnalysis package. For the calculation, we explored lag time values up to 1000 ps, and set t^* , referred to as intermittency in the MDAnalysis implementation, to 10 ps. The trend of the SP function gives information about the local dynamics of water in the macromolecules' hydration shell, which herein encloses the binding sites. In particular, fitting the data with exponential functions allows computing the residence time τ of the water molecules. [205, 209, 210] Depending on the nature of water dynamics in the hydration shell, the time decay of the SP can be better approximated by a single exponential

$$SP(t; r) = Ae^{-t/\tau} \quad (4.2)$$

or a double exponential function:

$$SP(t; r) = Ae^{-t/\tau_{fast}} + Be^{-t/\tau_{slow}} \quad (4.3)$$

In the latter case, τ_{fast} and τ_{slow} are short and long time decay constants, indicating a dual lifetime regime for water molecules, such as a fraction of solvent molecules staying within the hydration shell for prolonged periods of time and another entering and egressing rapidly.[205, 206] A stretched exponential function (or Kohlrausch–Williams–Watts, KWW, function) can also be used[205, 210]

$$SP(t; r) = Ae^{-(t/\tau)^\beta} \quad (4.4)$$

where the stretching parameter β reflects the extent of nonexponential nature of the phenomenon. However, the interpretation is not straightforward as that of a single or double exponential. Nevertheless, it can provide a valuable estimate of the time scale over which the process evolves. Indeed, the diffusion of solvent molecules can be more complex than the description via single or double exponentials, as it may include multiple relaxation rates and may be better characterized by a distribution of rate constants.

4.1.1.2 Electrostatic Contribution to the Binding Free Energy

We estimated the electrostatic contribution to the binding free energy of the riboflavin ligand in both the RF kinase and FMN riboswitch systems. The electrostatic component to the binding free energy for a target-ligand system in a water buffer ($\Delta G_{binding}^{ele, solv}$) can be computed according to a thermodynamic cycle as shown in Figure 4.2.

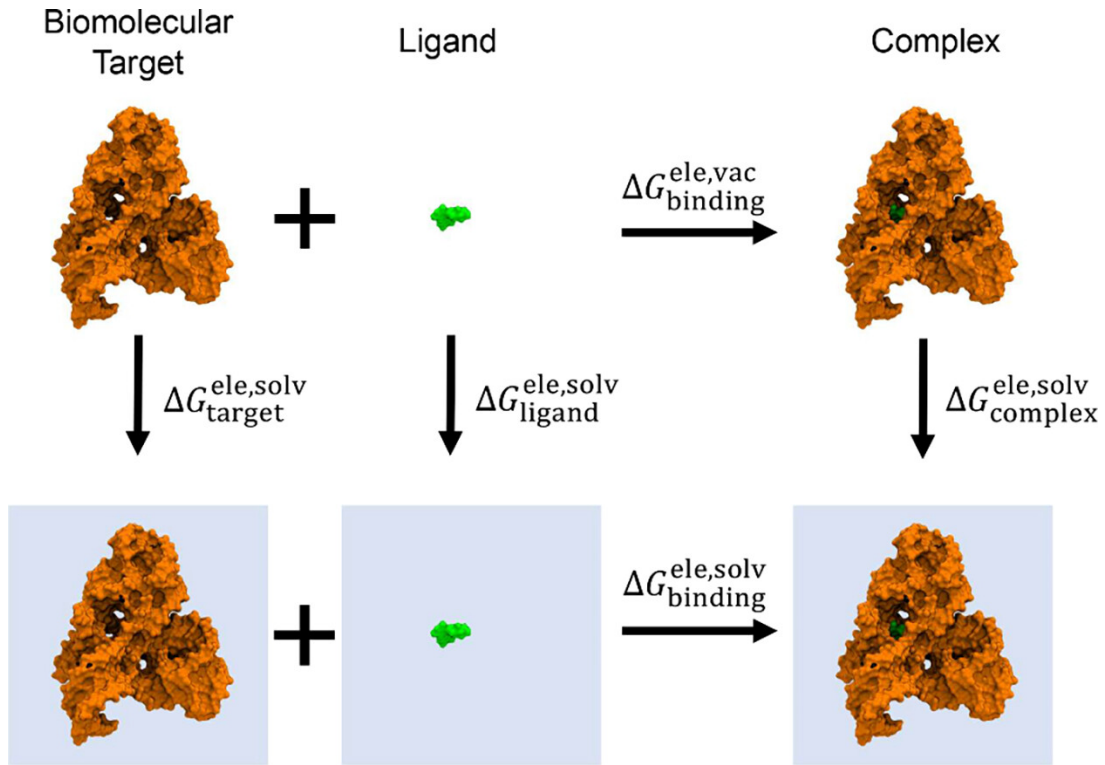


Figure 4.2. Schematic representation of the thermodynamic cycle considered for the calculation of the electrostatic contribution to the binding free energy. A generic ligand and biomolecular target are shown, with the top row indicating the binding process in vacuum, while the bottom row represents the event in solution. The arrows are labeled with the corresponding term.

Following the scheme of the thermodynamic cycle, we can write

$$\Delta G_{\text{target}}^{\text{ele,solv}} + \Delta G_{\text{ligand}}^{\text{ele,solv}} + \Delta G_{\text{binding}}^{\text{ele,solv}} - \Delta G_{\text{complex}}^{\text{ele,solv}} - \Delta G_{\text{binding}}^{\text{ele,vac}} = 0 \quad (4.5)$$

from which we can determine the electrostatic component to the binding free energy for a target-ligand system in a water buffer as

$$\Delta G_{\text{binding}}^{\text{ele,solv}} = \Delta G_{\text{complex}}^{\text{ele,solv}} + \Delta G_{\text{binding}}^{\text{ele,vac}} - (\Delta G_{\text{target}}^{\text{ele,solv}} + \Delta G_{\text{ligand}}^{\text{ele,solv}}) \quad (4.6)$$

In this way, the calculation of $\Delta G_{\text{binding}}^{\text{ele,solv}}$ is decomposed in three main steps, where (i) the isolated target and ligand in vacuum are transferred in solution ($\Delta G_{\text{target}}^{\text{ele,solv}} + \Delta G_{\text{ligand}}^{\text{ele,solv}}$), (ii) the formed target-ligand complex in vacuum is transferred in solution

($\Delta G_{\text{complex}}^{\text{ele,solv}}$), and (iii) the target-ligand complex is formed in vacuum ($\Delta G_{\text{binding}}^{\text{ele,vac}}$). Each of the ΔG s in the right-hand side of the above equation can be determined from differences between electrostatic contributions in the different media (vacuum or water buffer)

$$\Delta G_{\text{binding}}^{\text{ele,solv}} = (G_{\text{complex}}^{\text{ele,solv}} - G_{\text{complex}}^{\text{ele,vac}}) + (G_{\text{complex}}^{\text{ele,vac}} - (G_{\text{target}}^{\text{ele,vac}} + G_{\text{ligand}}^{\text{ele,vac}})) - \left[(G_{\text{target}}^{\text{ele,solv}} - G_{\text{target}}^{\text{ele,vac}}) + (G_{\text{ligand}}^{\text{ele,solv}} - G_{\text{ligand}}^{\text{ele,vac}}) \right] \quad (4.7)$$

where $G^{\text{ele,solv}}$ is the electrostatic energy of the species (target, ligand, or complex) evaluated in a water medium, and $G^{\text{ele,vac}}$ is the electrostatic energy of the species (target, ligand, or complex) evaluated in vacuum. Simplifying the equation, $\Delta G_{\text{binding}}^{\text{ele,solv}}$ can be obtained as

$$\Delta G_{\text{binding}}^{\text{ele,solv}} = G_{\text{complex}}^{\text{ele,solv}} - (G_{\text{target}}^{\text{ele,solv}} + G_{\text{ligand}}^{\text{ele,solv}}) \quad (4.8)$$

Each of these three terms was estimated on a pool of structures sampled from the MD simulations, selected through a cluster analysis procedure to represent as much as possible the conformational space explored. To this end, the k-means clustering algorithm was performed on a low dimensional embedding (see Section 3.3.1) derived from the distances between residues in the binding pockets (selected from being in close proximity to the ligands). Thus, the centroids from the first 15 most populated clusters were used in the estimation of the electrostatic contribution to the binding free energy. All the electrostatic calculations were performed by solving the linearized and nonlinearized Poisson–Boltzmann equation, for the protein and RNA systems respectively, through the software Delphi version V5.1 [211] patched with NanoShaper. [212] Delphi takes as input PQR files (i.e., modified pdb files, that contain also information about atom partial charge, Q, and radius, R), which we prepared using the radii and the charges reported in the force fields used in the MD simulations (i.e., ff19sb for the protein, OL3 for RNA, and GAFF2 for the ligand). The dielectric constant describing the water medium was set to 80, while a value of 2 was used to describe the electronic polarizability of the protein interior. An ionic atmosphere

of 100 mM of monovalent salt was considered, and the linearized and nonlinearized Poisson–Boltzmann equations were solved at the temperature of 298 K. Averages for the two systems were computed considering the respective cluster populations, and associated errors were estimated via bootstrap over the 15 values with 500 iterations. The energy extracted from DelPhi calculation includes the polarization contribution, due to the jump in dielectric constant from the solute and the solvent, and the ionic contribution, due to the ionic atmosphere surrounding the solute. In the case of the nonlinearized equation, also the electrostatic stress and the osmotic pressure terms are accounted for. In the Poisson–Boltzmann model, the energy is a free energy since it derives from averaging over all the solvent degrees of freedom assuming the canonical ensemble. A further averaging is then done over different conformations coming from the molecular dynamics simulation. A 2 Å Stern layer was used in the calculations.

4.2 Results and Discussions

In this study, we compare central features for target-ligand binding in protein and RNA targets of pharmaceutical relevance. To this end, we perform MD simulations and apply computational analyses that are established in MD simulations of protein–ligand systems but have found less to no application in the emerging realm of RNA molecules. We focus on major issues in the drug discovery context, including the conformational dynamics of the biomolecular targets, the role of electrostatics in target-ligand binding, and the dynamics of ligand and water molecules within the binding sites. We take the riboflavin (RF) kinase enzyme and the flavin mononucleotide (FMN) riboswitch as case studies of protein and RNA targets, respectively (Figure 4.1). The RF kinase belongs to the well-known class of protein kinases, a prototypical and historical protein family within the drug discovery community, which has long been approached to identify potent and selective small molecule modulators. Conversely, the FMN riboswitch is a highly conserved RNA motif in bacteria, which has gained remarkable attention as a promising pharmaceutical target for the development of antibiotics. [213] Given its highly complex and stable folding, that allowed characterization of the 3D structure via X-ray crystallography, [183] and its natural disposition to bind metabolites, this riboswitch displays ideal prerequisites for establishing rational drug discovery campaigns. Indeed, Merck has designed a small

molecule targeting the riboflavin binding site in the FMN riboswitch, and the compound is currently at the preclinical stage. [4, 213] Riboflavin is the natural ligand of both RF kinase and FMN riboswitch, thus providing a common ground for comparing relevant features underlying target-ligand binding in these diverse macromolecules.

4.2.1 Enhanced Conformational Flexibility in RNA without Appreciable Pocket Volume Variation

A thorough characterization of the structural dynamics of biomolecules is of paramount importance to understand their functioning. Furthermore, in the case of pharmaceutical targets, this can be crucial to identify fundamental features associated with ligand binding. For instance, assessing structural changes, e.g. by means of the RMSD, between bound and unbound states of a drug-target system can be informative about stabilizing effects induced by the presence of the ligand. Herein, for both the RF kinase and FMN riboswitch, we observed an average RMSD on the whole structure ranging between 2.5 and 3 Å in both the ligand-bound and unbound states (Figure 4.3A, see Figure B.1 for timeseries). This indicates that the overall folding was maintained by both systems during the MD simulations in the μ s time scale, independently from the presence of the riboflavin ligand. Nevertheless, the individual profiles in the four simulated systems demonstrated some noticeable differences. In the case of riboflavin-bound RF kinase, the RMSD shows a narrow distribution, which becomes slightly wider and tends to bimodality in the unbound protein. This suggests that the presence of the ligand may promote stability of the overall structure, as also confirmed by inspecting the root-mean-square fluctuation (RMSF, Figure B.2). Interestingly, the same behavior is conserved and more marked in the FMN riboswitch. However, regardless of the presence of the riboflavin, the distributions of the RMSD in the FMN riboswitch are appreciably broader, demonstrating a global conformational flexibility that is enhanced in the RNA molecule with respect to RF kinase. Looking at the timeseries, a peak could be observed in the bound FMN riboswitch between 400 and 500 ns, which occurred once in only a single replicate. This was due to a transient interaction between the ends of the strands (see Figure B.1), further demonstrating the significant structural flexibility allowed for the RNA structure despite a high level of folding achieved through its complex tertiary structure

(see 3D structure in Figure 4.1B).

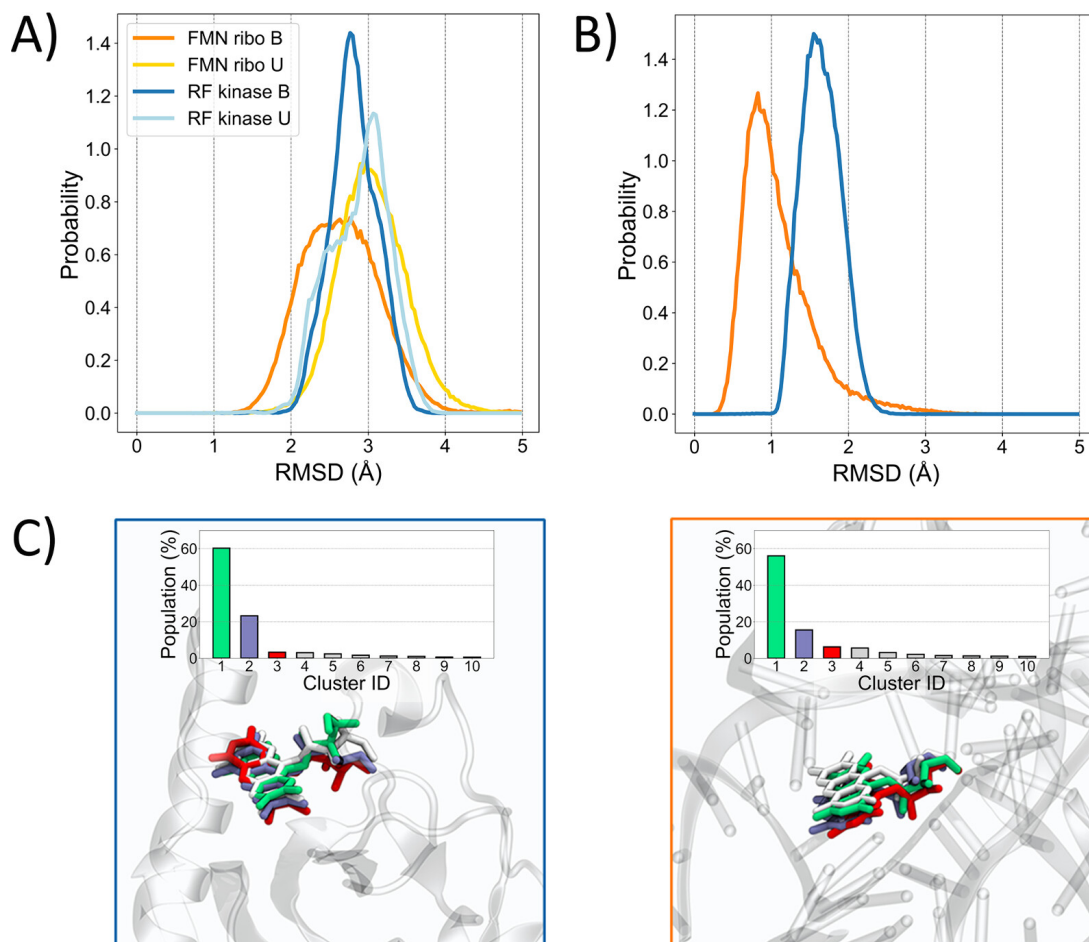


Figure 4.3. (A) RMSD of the RF kinase and the FMN riboswitch with respect to the starting structure of the MD simulations. The RMSDs were computed on the macromolecule heavy atoms and are shown as probability distributions by aggregating data from the three MD-simulation replicates performed on both ligand-bound (RF kinase B, FMN ribo B) and unbound systems (RF kinase U, FMN ribo U). (B) RMSD of riboflavin in the ligand-bound simulations with respect to the starting structure of the MD simulations. The RMSD, shown as probability density distribution, was computed after optimal alignment on binding pocket atoms with lowest root-mean-square fluctuation during the MD simulations. (C) Superposition of the most populated configurations in the MD simulations (colored structures in green, purple and red, in decreasing cluster population order) with the corresponding one in the crystal (white structure, including the surrounding targets in transparent New-Cartoon representation), for the riboflavin ligand bound to the RF kinase (left panel) and the FMN riboswitch (right panel).

Focusing on riboflavin in the target-ligand complexes, we observed a variable configurational space accessible to the ligands in the two biomolecular targets. In particular, we found remarkably higher RMSD values for the ligand in complex with the

RF kinase compared to the complex with the FMN riboswitch (Figure 4.3B, see Figure B.3 for timeseries). Interestingly, the RMSD distribution was broader in the RNA target, suggesting that a wider variety of configurations were accessible, but all closer to the reference, if compared with the RF kinase complex. In particular, looking at the timeseries, we noticed that the higher RMSD displayed in the protein complex was subsequent to the loss of interactions in the hydroxylic tail of the ligand with the magnesium atom in the binding pocket. This can also be noticed by visually comparing the most populated clusters (Figure 4.3C), where the centroid in the RF kinase complex exhibits major difference in the orientation of the hydroxylated tail, while the one in the FMN riboswitch has most of the difference due to rigid traslation of the ligand inside the pocket. The structural dynamics of the residues comprised in the binding pocket can also play a remarkable role in the binding process. To highlight local conformational rearrangements in proximity of the riboflavin binding site, we focused on a subset of distances, namely those between residues belonging to the pockets (see Section 4.1). Thus, we applied dimensionality reduction approaches to characterize the states visited in the MD simulations. We tested both linear and nonlinear methods, namely the well-known principal component analysis (PCA) and the more recent t-distributed stochastic neighbor embedding (t-SNE), respectively. While PCA has found historical application in the analysis of MD trajectory data and has become of routine use nowadays,[214] the exploitation of t-SNE on MD data is rather recent. The latter approach is particularly suited to disentangle complex manifolds associated with high-dimensional data, where correlations possibly vary in a nonlinear manner. Thus, t-SNE might be particularly suited to deal with MD trajectory data, such as MD simulations of biomolecules displaying high flexibility and complex structural dynamics. Indeed, very recently t-SNE has demonstrated to be effective in characterizing the conformational space explored by a selected set of intrinsically disordered proteins (IDPs), a notoriously challenging class of proteins in this respect.[215] Thus, we envisage promising potential in the study of RNA molecules, which pose comparable challenges from the computational standpoint.[34] We performed the dimensionality reductions on the MD simulation data of all four systems (RF kinase and FMN riboswitch, both in the ligand-bound and unbound states), using as input features the distances between residues comprised in the binding pockets. In the RF kinase, the distribution in the low dimensional space obtained with PCA was highly crowded and

partially overlapped between the ligand-bound and unbound simulations (Figure B.4). While this indicated that the conformational space sampled by the residues in the two cases was comparable to a certain extent, it was not possible to discriminate distinct conformations of the pocket. Conversely, in the FMN riboswitch there was a clearer separation between the simulations in the two states, and few clusters were visible (Figure B.4). Interestingly, this indicated that for both the protein and RNA targets, the binding pockets explored distinct regions of configurational space in the ligand-bound and unbound states, and this behavior was consistent in the three replicates of simulation for each case. This suggested that the presence of the riboflavin in the binding sites has the effect of noticeably reshaping the conformational preference of the pocket, with a sharper effect in the FMN riboswitch. Besides these considerations, on the whole, the results from the PCA were not particularly informative, as also quantified by the relatively low cumulative explained variance by the first two PCs (about 40% and 50% for the kinase and the riboswitch, respectively). Conversely, the crowding effect observed from the PCA analyses was remarkably attenuated by t-SNE, which had a better performance in separating clusters of conformations (Figure 4.4).

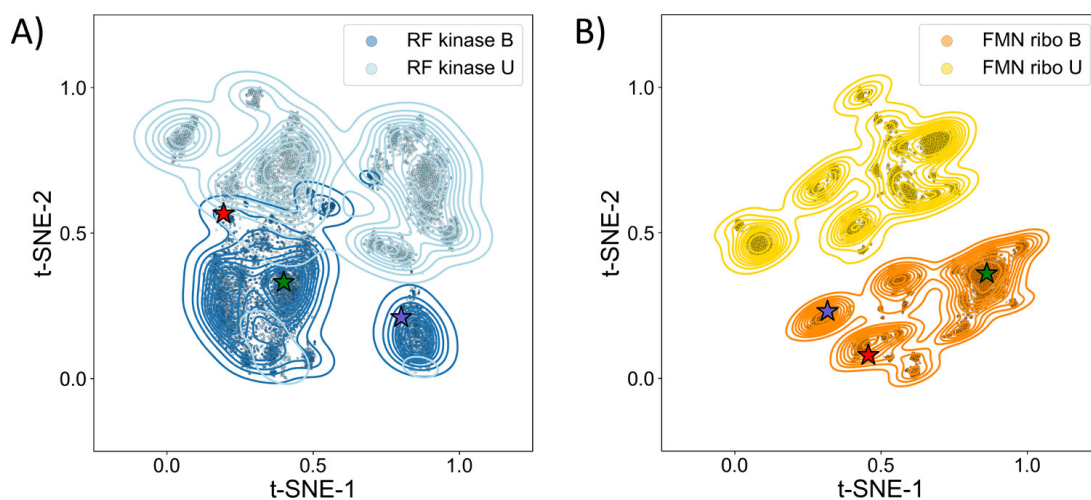


Figure 4.4. t-SNE space for (A) the RF kinase and (B) the FMN riboswitch, both in the ligand-bound and unbound states. The low dimensionality spaces were generated using as input features the minimum distances between residues within the protein and RNA binding sites. Contour lines are the kernel density estimation of the point distributions and better highlight regions of overlap between the ligand-bound and unbound systems. Colored stars indicate the location of ligand configurations in the clusters shown in Figure 4.3, using the same color code.

The projections obtained with t-SNE have the property to preserve the local neighborhood information, thus favoring a more straightforward interpretation of the conformational heterogeneity in each cluster and possibly highlighting variations associated with ligand binding. Indeed, the overall picture reported by PCA was maintained, but the separation between different conformations of the pockets was remarkably improved through t-SNE. Interestingly, by highlighting the three most populated clusters of the ligand configurations in the binding site reveals that these belong to clearly distinct regions in the t-SNE space. In other words, different ligand configurations stabilize diverse pocket conformations, and the t-SNE approach was sufficiently sensitive to the input features (distances between residues) to detect such different states of the pocket. It is interesting to note that the sharp separation in the t-SNE space between the ligand-bound and unbound conformations of the pocket in the FMN riboswitch does not correspond to significant variation in the pocket volume (Figure 4.5B,D), which has a rather peaked and overlapping distribution (mean = $637 \text{ \AA}^3$ and $590 \text{ \AA}^3$, and stdev = $78 \text{ \AA}^3$ and $104 \text{ \AA}^3$ for the bound and unbound states, respectively).

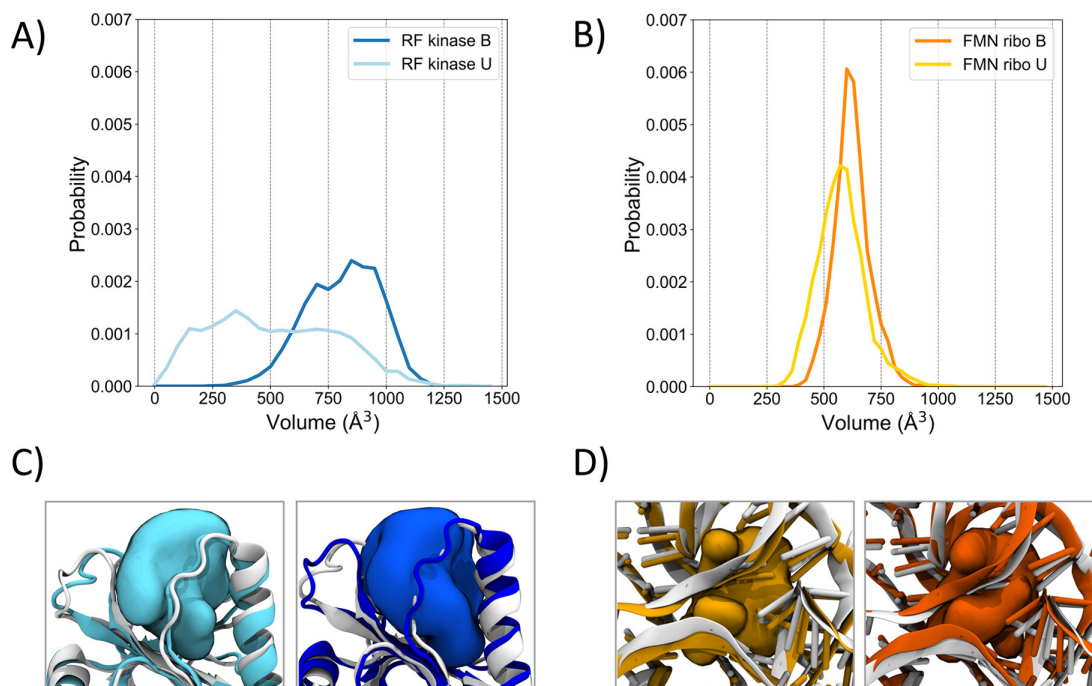


Figure 4.5. Distributions of the pocket volumes explored during the MD simulations of the (A) RF kinase and (B) FMN riboswitch, both in the bound and unbound states. In the bottom panels C) and D), the surfaces of the pockets for the different systems are also shown, color coded consistently with the corresponding panels A) and B). In C), the left panel has colored target structure and pocket surface for the unbound state of the protein, with the target structure in the bound state in white for comparison, while the right panel has colored target structure and pocket surface for the bound state, with the target structure in the unbound state in white. The same scheme is followed in D) for the RNA target. The displayed snapshots correspond to trajectory frames with pocket volumes closest to the average values observed in the simulations.

This indicates that the volume of the pocket is poorly sensitive to the conformations accessible by the binding site, and the RNA molecule is able to experience appreciable conformational rearrangements within the binding site while preserving the size and overall architecture of the cavity. Additionally, this is intuitively in contrast with the enhanced flexibility displayed by the structure of the whole RNA molecule (see RMSD distribution in Figure 4.3). Conversely, the volume range covered by the binding site in the RF kinase was significantly wider. In particular, the distribution of the volume was remarkably different in the ligand-bound and unbound cases, with the former adopting larger values on average (mean = 836 Å³ and 533 Å³ and stdev = 157 Å³ and 269 Å³ for the bound and unbound states, respectively). While compared to the FMN riboswitch, the volume distribution was broader in both cases; in the unbound RF kinase, it was almost uniform over the spanned range, highlighting a

conformational exploration with no particular preference. Differently, a well-defined, narrower and Gaussian-like distribution was observed in the simulations of the ligand-bound system. This comparison is interesting, as it suggests how the presence of the ligand is able to restrict the conformational dynamics of the binding site and stabilize specific conformations. Such a divergent behavior between RF kinase and the FMN riboswitch can be attributed to both the different location and the biological role of the pockets in the two investigated systems. On the one hand, the more conserved volume of the pocket observed in the case of FMN riboswitch with respect to RF kinase underscores the ability of RNA structures to form sizable pockets that might be druggable. On the other hand, the enhanced conformational space might represent a challenge for conventional computational methods used in drug discovery that take into account only partially the conformational flexibility of the target. The known flexibility of the RNA structure is seen as a major challenge for RNA-targeted drug discovery, as it may be difficult to take into account, for instance by selecting representative RNA conformations. This reasonably holds true for smaller RNA molecules, such as single stranded RNAs or short stem-loop motifs, which are indeed highly challenging to treat with conventional rational drug design approaches (mostly because they typically do not form clearly identifiable and stable binding sites). Nevertheless, the case of the FMN riboswitch studied herein showed how, notwithstanding a clearly appreciable structural flexibility of the RNA as a whole, the architecture of the ligand binding site was overall rather maintained, while experiencing noticeable conformational rearrangements between the ligand-bound and unbound states. We note that in this case the FMN riboswitch binding site demonstrated even a higher stability with respect to RF kinase, and protein kinases represent well-known and tractable systems, though challenging in terms of selectivity. While these aspects are not sufficient, they nevertheless represent necessary prerequisites to establish rational drug discovery campaigns. The considerations herein may be highly system specific; however, they provide a first indication on the feasibility of RNA-targeted drug discovery via standard computational means.

4.2.2 Comparable Electrostatic Contribution to Ligand Binding

Having investigated the structural and dynamical behavior of the two systems, looking both at the overall structure and with emphasis on the binding site, we then

inspect the electrostatic properties of target-ligand binding. In particular, we focus both on the electrostatic features of the targets and on the electrostatic contribution to the binding free energy. Notably, electrostatics is a key factor in shaping short- and long-ranged interactions and has been proven to provide a significant contribution to drug binding in several cases. [216–219] RNA molecules are rich in negatively charged phosphate groups, and RNA nucleotides display relatively limited chemical variability compared to protein amino acids. This may result in nonobvious electrostatic profiles when it comes to ligand binding. Thus, since the RF kinase and the FMN riboswitch are bound by the same ligand, riboflavin, we investigated the differences in their binding to the same small molecule. We therefore first computed the electrostatic potential of the targets in implicit solvent (Figure 4.6, upper panels), as implemented in the DelPhi software (see Section 4.1). As the figure shows, we observed a more pronounced negative electrostatic potential, broadly distributed over the FMN riboswitch surface. This was expected, and it is primarily due to the presence of the phosphate groups on the RNA backbone. In contrast, the RF kinase exhibited an overall electrostatic potential that was more balanced, showing regions of both locally concentrated and depleted charge density. Interestingly and consistently, these considerations on the two systems held true also in the two binding-site regions.

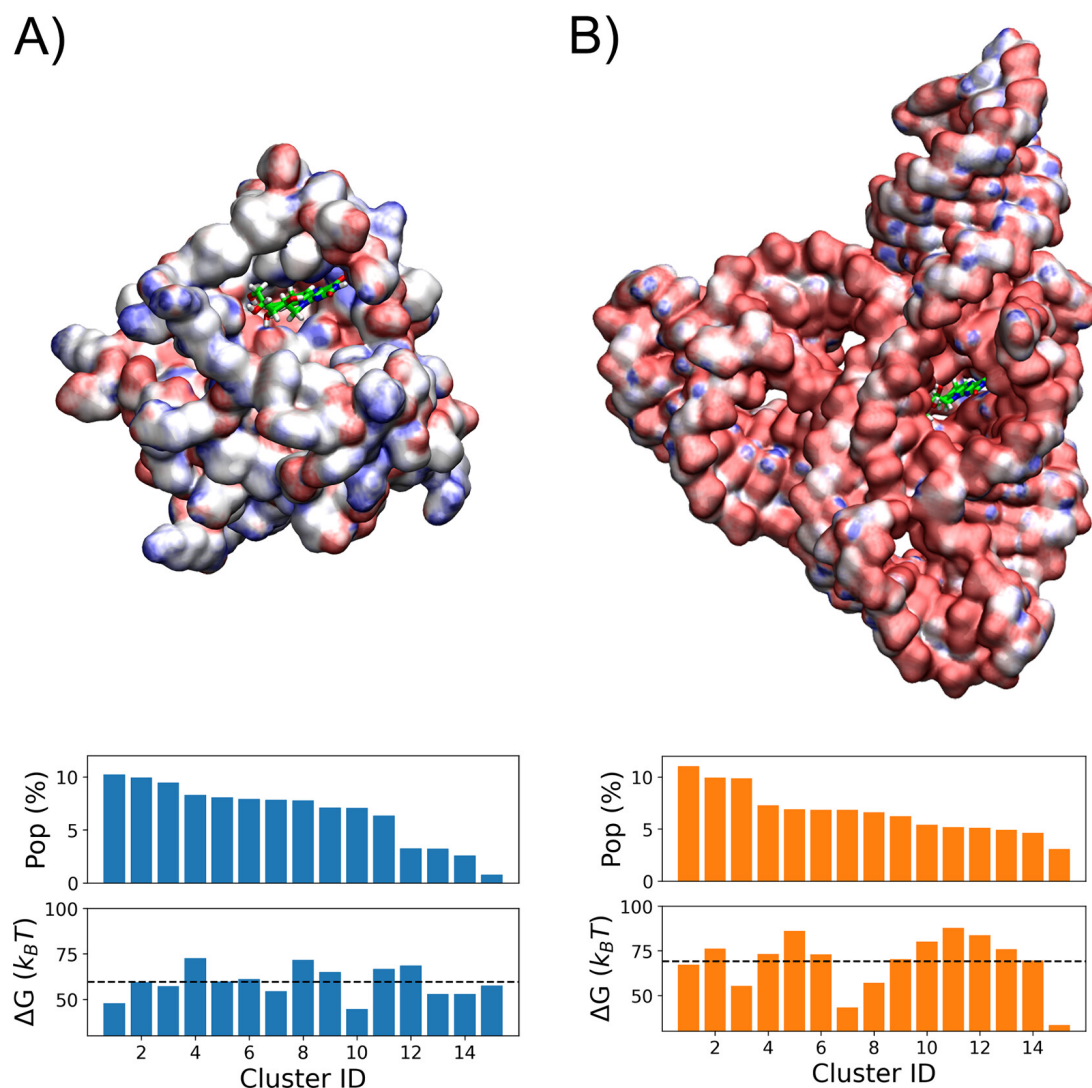


Figure 4.6. Electrostatic analysis on the RF kinase and the FMN riboswitch systems. The surface representation of the most populated cluster for the (A) protein and (B) RNA complexes is colored according to their electrostatic potential. The color scale is consistent for the two systems and ranges from $-10 k_B T/e$ (red) to $+10 k_B T/e$ (blue). The bound riboflavin ligand is shown in licorice, with its carbon atoms highlighted in green. Below each system, we report the cluster population (upper panel) and the corresponding electrostatic component to the binding free energy in an implicit water buffer (lower panel, expressed in $k_B T/e$).

To further obtain a more quantitative picture relative to ligand binding, we computed the electrostatic contribution to the binding free energy for riboflavin in complex with the two biomolecular targets (Figure 4.6, lower panels). To include possible local rearrangements of both the ligand and the surrounding residues in the targets' binding pockets, we repeated the calculation on representative conformations from our MD

simulations, which we selected through a cluster analysis procedure. The results of our calculations averaged over these representative conformations are $(59.6 \pm 2.2) k_B T$ and $(69.1 \pm 3.0) k_B T$ for the RF kinase and the FMN riboswitch, respectively, considering the respective cluster populations. Notably, the two systems displayed an overall similar contribution. This was unexpected, particularly considering the different electrostatic potentials displayed by the two target macromolecules. The value of the electrostatic contribution was slightly lower for the RF kinase system. Interestingly, this can be reconciled with the different affinities reported for the ligand toward the two biomolecular targets. Previous studies in the literature report $K_D \approx 40 \mu M$ for riboflavin binding to the FMN riboswitch [183] and a $K_D \approx 1 \mu M$ for binding to RF kinase, [220] indicating that the ligand displays higher affinity for the protein. We note that the latter K_D was measured on a bacterial RF kinase, while the present study focused on the human enzyme. Nevertheless, the two forms display high similarity in the RF kinase binding site (74%, see the B.1), suggesting that, in first approximation, we can reasonably consider this affinity approximately valid also for the human enzyme. Another remarkable aspect was that the electrostatic contribution was unfavorable in both cases. This suggests that the driving force for riboflavin binding to the two targets should be found elsewhere, for instance in nonpolar interactions and/or indirect polar interactions.

4.2.3 Distinct Pattern of Hydrogen Bonds and Ligand Conformations

The conformational space accessible to ligands inside binding sites and the possibility to establish multiple interactions with pockets' residues are remarkable aspects that can affect both thermodynamics and kinetics of the binding process. Within such a picture, the role played by water molecules can be of paramount importance, ranging from ligand and pocket (de)solvation to mediation of target-ligand specific interactions. As already discussed, a comparison of the conformational flexibility displayed by the riboflavin ligand in the RF kinase and FMN riboswitch binding sites from our MD simulations revealed that the ligand was less stable inside the FMN riboswitch binding pocket (broader RMSD distribution, see Figure 4.3B), despite remaining closer to its initial configuration with respect to the RNA. This observation

can be further corroborated by inspecting the detailed network of ligand-target contacts and the internal degrees of freedom explored by the riboflavin ligand. Among the possible interactions, we focus in particular on hydrogen bonds, which explain most of the enthalpic contribution of drug-target binding in many cases.

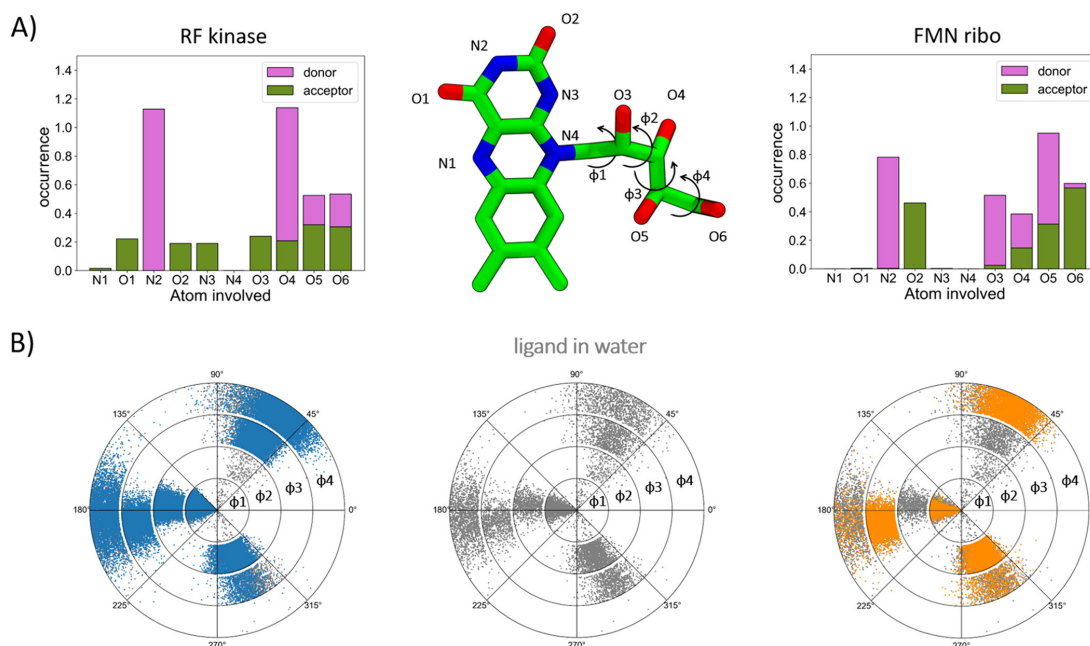


Figure 4.7. (A) Occurrence of hydrogen bonds formed by the riboflavin ligand with the RF kinase (leftmost panel) and the FMN riboswitch (rightmost panel) during the MD simulations. For each riboflavin heteroatom (central structure, only heavy atoms displayed and heteroatoms annotated with atom names), the two plots report the percentage occurrence of hydrogen bonds, highlighting the fraction where the atom acts as a donor or acceptor. (B) Polar plots of the dihedrals in the riboflavin flexible tail (see the annotated structure in (A) for schematics). The central panel reports dihedral values computed from the MD simulation of riboflavin in water (gray points). In the leftmost and rightmost panels, these values are overlapped with the ones computed from the simulations of riboflavin in complex with the RF kinase (blue points) and FMN riboswitch (orange points), respectively.

In the RF kinase, the formation of hydrogen bonds was distributed over the whole structure of the ligand and occurred for a significant fraction of simulation time, with most of the ligand heteroatoms mainly acting as acceptors (Figure 4.7A, leftmost panel). Besides N2 in the aromatic moiety and O4 in the hydroxylated tail, that acted as donors for a significant fraction of the simulation, the other heteroatoms appear to be transiently involved in hydrogen bonds. Comparing the results from the RF kinase to the ones of the RNA revealed that the occurrence of direct hydrogen bonds with the binding pocket was comparable on average (0.42 and 0.37, respectively). A

closer inspection showed that the number of ligand heteroatoms involved in hydrogen bonding was lower when bound to the FMN riboswitch (Figure 4.7A: 6 atoms in the RNA, 8 in the protein). However, they all displayed significant occurrence, rather balanced between donors and acceptors. Interestingly, most of these hydrogen bonds involved the hydroxylic tail of the molecule (Figure 4.7A, right panel). While this can be highly case specific, it can be related to the different nature of the targets. The FMN riboswitch binding site is enriched with nucleobases, able to establish $\pi - \pi$ interactions, and through which the ligand is able to intercalate its aromatic moiety (Figure 4.1D). Such intercalating behavior is common in known ligands binding to RNA targets, [51, 221] and the current space of RNA ligands is enriched with aromatic rings.[66, 222]

Interestingly, these interaction patterns observed in the two systems led to differential conformational preference of the ligand in the binding pockets. This behavior can be further inspected in terms of the different values of relevant dihedral angles explored by the ligand in the bound states (Figure 4.7B, blue and orange points for RF kinase and FMN riboswitch, respectively) during the simulations compared to those sampled in water (Figure 4.7B, gray points). Dihedral ϕ_1 , which is the main determinant in the orientation of the tail with respect to the flavin moiety, shows only minor deviations, with the same dominant rotameric state displayed in all the simulations. Conversely, dihedral ϕ_2 undergoes a rotameric restriction upon binding, resulting in two almost equipopulated rotamers in the case of RF kinase and a single one in the FMN riboswitch, thus more marked for the RNA. Finally, in the RF kinase, ϕ_3 and ϕ_4 dihedrals retain the possibility to cover the whole dihedral space even in the bound state, indicating an overall wider conformational accessibility at the end of the tail. This can also be appreciated observing the most populated clusters of ligand configurations inside the pocket (Figure 4.3C, left panel). Thus, notwithstanding the higher degree of conformational variability of the riboflavin's tail in the RF kinase pocket, this did not preclude the establishment of hydrogen bonds. Differently, in the FMN riboswitch, while ϕ_4 retains the possibility to cover the entire dihedral space, although to a lower extent compared to RF kinase, the ϕ_3 dihedral experiences a significant rotameric restriction (Figure 4.7B, rightmost panel). Thus, the internal conformational space accessible to the ligand was more limited in the binding site of the FMN riboswitch. Here, a closer inspection of the ligand interaction network in the simulations

revealed an interaction between the O4 in the ligand tail and a Mg^{2+} ion, mediated by two bridging water molecules, which was transiently maintained in the simulations (Figure B.5), narrowing down even further the conformational space accessible to the ligand hydroxylated tail. Finally, this picture in the RNA system was compatible with the ligand RMSD with respect to the binding pocket, showing a broad distribution but closer to the reference (Figure 4.3B), demonstrating a semirigid rototranslational motion of the riboflavin in the FMN riboswitch binding site. Here, topological constraints associated with different size and shape of the pockets undoubtedly play a role. The FMN riboswitch adopts a highly complex tertiary structure, with an elaborated folding that is comparable to what can be observed in globular proteins. In this RNA, the ligand-occupied binding site is rather deep and of relatively low volume ($637 \text{ \AA}^3$), while the one in the RF kinase is more solvent-exposed and adopts on average a larger volume ($836 \text{ \AA}^3$). This may be related to the biological role and functioning of the RF kinase, where the binding pocket is a catalytic site, and substrate binding needs to occur with high turnover. Given the above considerations, in the case of the FMN riboswitch, it is reasonable to expect a larger penalty contribution to the binding free energy due to a higher decrease of conformational entropy upon binding compared to the RF kinase. At the same time, the riboflavin as a whole experiences larger fluctuations in terms of conformational variability within the FMN riboswitch binding pocket (Figure 4.3B, $\text{stdev} = 0.45 \text{ \AA}$ in the FMN riboswitch vs $\text{stdev} = 0.25 \text{ \AA}$ in the RF kinase). We expect that the entropic conformational penalty can be partially or entirely compensated by the higher residual rototranslational mobility (lower rototranslational entropic penalty), leading to a comparable net entropic contribution to binding for the two investigated systems.

4.2.4 Faster Exchange Rate of Water in the Bound Protein Binding Site

Water molecules are central characters in drug-target binding and can affect both thermodynamics and kinetic aspects of this process. They play a key role in both ligand and binding site (de)solvation, and their contribution to favor ligand binding can be decisive in some cases. [223–225] We investigated the dynamics of the solvent molecules both in the bound and unbound states of the two biomolecular targets. In

particular, we monitored the survival probability (SP) of waters in the binding sites, i.e. the amount of simulation time spent by water molecules in close proximity to the pockets' residues prior to being exchanged with the external buffer. The trend of the SP function provides information about the local dynamics of water in the considered region and allows estimating water residence time. Comparing the SP curves for the RF kinase and FMN riboswitch systems (Figure 4.8), we can clearly distinguish two main trends.

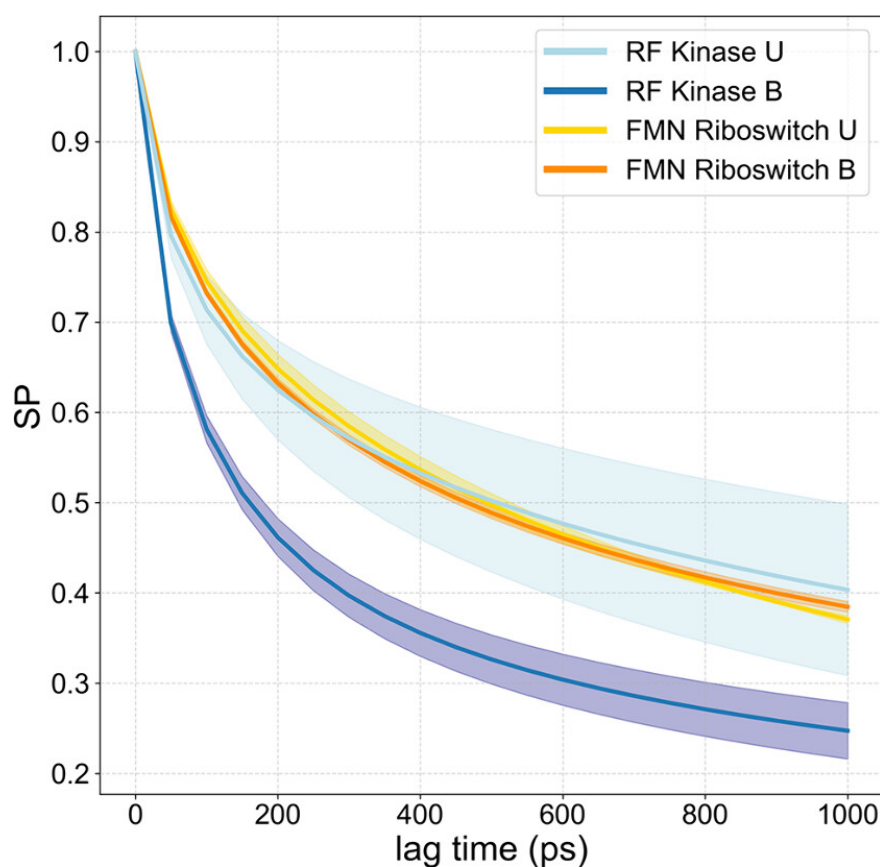


Figure 4.8. Survival probability of the water molecules in the binding site region of the RF kinase and FMN riboswitch systems from the MD simulations. Different colors are used for the two macromolecular systems and to distinguish simulations in complex with riboflavin (B, bound) or without the riboflavin ligand (U, unbound).

In particular, the RF kinase in the bound state displayed the fastest time decay, indicating a rapid and efficient exchange of water molecules in the binding region with the outer solvent buffer. Interestingly, compared to its unbound state, the riboflavin-bound RF kinase displayed a larger pocket volume on average (Figure 4.5A), which

was also larger than the FMN riboswitch's one. Thus, the ligand presence stabilizes the binding pocket in conformations (see peaked distribution of the volume in Figure 4.5A) that offer a wider pocket-solvent interface for exchange of water molecules (Figure B.6A), which may result in the observed faster decay. Conversely, the FMN riboswitch showed a similar time decay for water SP in the bound and unbound cases, which is consistent with the similar pocket volume observed in the two forms (Figure 4.5B). Indeed, as previously discussed, we have observed internal rearrangements of the pocket residues in the RNA (Figure 4.4B), while maintaining the overall volume of the binding site (Figure 4.5B). The mean residence time of the water molecules can be estimated through fitting of the SP curves. We initially explored a single exponential function, which gave acceptable results (Figure B.7A). However, when using a double exponential, the agreement was remarkably improved in all cases (Figure B.7B). This is indicative of complex water dynamics, in a dual lifetime regime, for the water molecules within or in close proximity of the binding site. In particular, we could distinguish fast (τ_{fast}) and slow (τ_{slow}) residence times (Table 4.1).

Table 4.1. Residence times estimated through double exponential fitting of the SP function for the RF kinase and FMN riboswitch, both in the riboflavin-bound and unbound states.

System	τ_{fast} (ps)	τ_{slow} (ps)
RF kinase U	51.0	531.8
RF kinase B	48.1	344.2
FMN ribo U	56.3	582.4
FMN ribo B	55.6	469.8

The faster regime was characterized by values of τ_{fast} of about 50 ps in all cases, thus independent of the presence of the riboflavin ligand in the binding site. This observation suggests a rapid exchange between interfacial waters of the solvation shell considered with the surrounding buffer. Indeed, such time scale is compatible with residence time of mobile water molecules close to the biomolecule surface.[226] Conversely, the slower regime ranged approximately the 350–600 ps time scale, approaching residence time values ascribed to more buried waters in protein binding sites. [226, 227] Notably, the RNA in the bound and unbound and the protein in the unbound states displayed τ_{slow} around 500 ps, while the bound protein displayed around 350

ps. This was unexpected as, intuitively, the presence of the ligand would retard water exchanges from the binding pocket, due to possible hydrogen bonding with the ligand, or to ligand steric hindrance reducing the mobility of other species within the cavity. However, as already mentioned, this faster exchange rate in the slow regime, i.e. lower τ_{slow} compared to the other systems, may be associated with the widest volume explored on average by the RF kinase in the presence of the ligand, that favors exchange of water molecules with the surrounding solvent buffer. A further inspection of τ_{slow} revealed that the same trend can be observed between the protein and the RNA, with larger values in the unbound systems ($\tau_{slow} = 344.2$ and 531.8 ps for the RF kinase and $\tau_{slow} = 469.8$ and 582.4 ps for the FMN riboswitch, in the bound and unbound states, respectively), although the discrepancy was lower in the RNA case. For the latter, a visual inspection of the trajectories revealed that the pocket in the unbound state explored conformations where the access to the cavity was narrower (Figure B.6B), partially hindering access and egress of water molecules and possibly causing the estimated lower exchange rates. Such modest flexibility of the pocket is compatible with a slightly broadened distribution of the sampled pocket volumes (Figure 4.5B) and, in particular, with the shift to lower volume values, indicating that also in the RNA case the presence of the ligand had a stabilizing effect on the pocket conformations. Finally, we note that compared to the riboflavin-bound RF kinase, water molecules in the FMN riboswitch exhibited in general higher persistency, independently from the presence of the ligand (Figure 4.8). Therefore, we can speculate that a ligand might display higher difficulty in displacing them, and this can have important implications for ligand binding free energies and association/dissociation rate constants, which should be taken into account in drug design endeavors.

4.2.5 Summary of Main Findings

In the following, we provide a brief summary of the major observations from each section of our study:

- The RF kinase and FMN riboswitch maintained the global folding in the MD simulations, with the RNA displaying higher conformational flexibility. The presence of the ligand stabilized the overall structure, particularly in the FMN riboswitch. Focusing on the binding site, the ligand in the RF kinase accessed

distinct configurations, mainly rearranging the hydroxylated tail, while in the FMN riboswitch, the variety of configurations visited, closer to the native pose, was primarily due to semirigid roto-translation. Additionally, we observed different conformations for the binding pockets, particularly between the ligand-bound and unbound states. Notably, in the FMN riboswitch, such conformational variations did not correspond to appreciable differences in pocket volume between bound and unbound states. Interestingly, this was intuitively in contrast with the global flexibility displayed by the RNA structure. Differently, in the RF kinase, the presence of the ligand restricted the conformational dynamics of the binding site. Finally, in the bound state of both protein and RNA, we observed how different ligand configurations stabilized different pocket conformations.

- The electrostatic potential on the surface of the FMN riboswitch was mainly negative, as expected given the presence of the phosphate groups in the RNA. Conversely, a more balanced picture was found for the electrostatic potential of the RF kinase. Unexpectedly, particularly considering such discrepancy, we found a similar electrostatic contribution to the binding free energy for the protein and RNA systems. Notably, we observed that this contribution was unfavorable in sign for both systems, suggesting that the driving force for riboflavin ligand binding to these two targets should be found elsewhere.
- We observed different patterns of hydrogen bonding and ligand conformations in the protein and RNA targets. Hydrogen-bonding interactions with the RF kinase binding site were rather uniformly distributed over the whole structure of the riboflavin ligand, with most of its heteroatoms acting as acceptors. Conversely, in the FMN riboswitch, hydrogen bonds were mainly formed through the hydroxylated tail of the ligand, with heteroatoms balanced between donors and acceptors, while the aromatic moiety was involved in stacking interactions with the RNA target. Inspection of dihedral angle distributions of the ligand hydroxylated tail inside the RF kinase binding site revealed how riboflavin was able to explore a wide internal conformational space, comparable to being in the unbound state in solution. Differently, when bound to the FMN riboswitch, the ligand displayed semirigid roto-translational motion, and its internal conformational space was more restricted.
- The bound RF kinase displayed rapid and efficient exchange of water molecules

between the binding site and the buffer. Conversely, the time decay for water SP was comparable in the bound and unbound states of the FMN riboswitch. Inspection of the mean residence times of the waters in the binding pockets revealed two regimes. The faster one, with similar residence times of about 50 ps for all systems, was associated with rapid exchange between interfacial waters of the solvation shell and the surrounding buffer. The slower regime, in the 500 ps time scale, was ascribed to exchanges of more buried waters in the binding sites. In the case of the bound RF kinase, the exchange in this slower regime was the fastest among the four systems and was attributed to the widest volume explored on average by the RF kinase in the presence of the ligand.

These results have been obtained by applying computational approaches that are established to analyze MD trajectory data of protein targets and that have been less applied, if at all, to RNA molecules. Therefore, from a pragmatic standpoint, our work demonstrates that comparable analyses can be performed on protein and RNA targets. Despite focusing on a limited set of systems, we have initially explored the applicability of these methods to an RNA molecule with highly folded structure and pharmaceutical relevance. In this sense, our study may serve as a practical guide in the selection of suitable computational approaches that demonstrated to be able to catch relevant features in RNA-ligand binding. In light of the relatively limited knowledge on the topic, this may prove particularly valuable to accelerate the setup of innovative RNA-targeted drug discovery campaigns. Additionally, we provided further insights into RNA-ligand recognition, at the atomistic level. In this respect, the use of MD simulations to characterize specific interaction patterns and conformational preferences of small molecules inside RNA binding sites is of outstanding relevance in light of optimizing ligand binding. For instance, here we found how the RNA-ligand system was able to establish polar and noncovalent interactions as much as in the protein–ligand system, suggesting that rational drug design frameworks can possibly be applied also for RNA targets. Alike, we were able to explore and characterize the pocket conformational dynamics, which may reveal peculiar conformations that can be exploited for RNA ligand design. We believe that advancing the knowledge on these aspects is of fundamental importance to accelerate concrete RNA-targeted drug discovery.

4.2.6 Limitations of the study

Some methodological limitations of the present study should be acknowledged.

- Magnesium ions were modeled using the parametrization proposed by Allnér and Villa, which is known to reproduce slow exchange kinetics and an enhanced affinity for phosphate groups. While this choice may partially influence conformational fluctuations in the riboflavin kinase system, such fluctuations are intrinsically associated with the solvent exposure of the binding site. On the other hand, in the FMN riboswitch simulations magnesium ions remained predominantly solvated and only partially coordinated to phosphate groups.
- Electrostatic contributions were estimated using Poisson Boltzmann calculations, which provide an efficient continuum description of solvent mediated electrostatics. In this framework, the computed quantity includes polarization and ionic atmosphere contributions, whereas the total binding free energy additionally contains nonelectrostatic terms and entropic contributions associated with conformational and solvent degrees of freedom. As a consequence, a difference of approximately $10 k_B T$ between electrostatic components should not be interpreted as implying a comparable fold change in K_D , since partial cancellation with nonpolar and entropic terms is expected and systematic uncertainties intrinsic to the Poisson Boltzmann approximation may affect absolute magnitudes. These calculations are therefore used primarily to compare electrostatic trends between the two targets rather than to predict experimental affinities quantitatively. More rigorous alchemical free energy methods could in principle offer an explicit treatment of ions and solvent, but at a substantially higher computational cost and with demanding convergence requirements, particularly for flexible and highly charged RNA systems
- The same internal dielectric constant was applied to both proteins and RNA in the Poisson Boltzmann calculations. Although RNA differs from proteins in terms of chemical composition and charge density, there is currently no consensus on an optimal dielectric value specific to RNA.[228, 229] The use of the same dielectric constant across proteins and RNA therefore represents a commonly accepted and pragmatic approximation.[229]

4.3 Conclusions

RNA molecules are gradually emerging as promising pharmaceutical targets, opening the way to novel therapeutic avenues. Nevertheless, compared to more established protein targets, the current knowledge about their interaction with small molecule partners is still lagging behind. In an effort to provide further insight in this respect, we compare relevant aspects for target-ligand binding in pharmaceutically relevant protein and RNA molecules, namely the RF kinase and the FMN riboswitch. Leveraging MD simulations and computational analyses, we focus on central issues including the conformational dynamics of the biomolecular targets, the role of electrostatic interactions, and the dynamics of ligand and water molecules in the binding sites. First, our study demonstrated that comparable computational analyses can be performed on the protein and RNA targets. Furthermore, by discussing similarities and differences in the results from the two systems, we provided, to the best of our knowledge, a first comparison of key ligand-target binding properties in protein and RNA targets. These not only highlighted the applicability of the employed approaches but also allowed us to make valuable observations and draw conclusions, in a comparable way as it can be done for protein targets. Interestingly, our analyses shed further light on the capability of the RNA-ligand system to establish polar, noncovalent interactions to an extent comparable to ligand-binding protein targets. We stress that the observed properties and the domain of applicability were explored on specific test cases and may be highly system dependent. For instance, we are aware that different RNA molecules can display different degrees of structural complexity, and here we picked an example possessing a high level of folding with a structurally complex and stable binding site, representing a favorable prerequisite to engage structure-based rational drug design efforts. While this is certainly not representative of the entire class of RNA molecules, nevertheless others of this kind may exist in the wide realm of noncoding, functional RNAs that are increasingly being identified. Thus, our study represents a first resource reporting on the feasibility of applying relatively standard computational approaches in computer-aided drug design to treat suitable RNA targets, laying the groundwork for the use of molecular simulations toward successful RNA-targeted drug discovery.

5 Exploring RNA Druggability through SHAMAN

The previous chapter establishes a common analytical ground for proteins and RNA in a specific test case. Here, we adopt a drug discovery perspective and focus on a paradigmatic viral RNA target, the HIV-1 trans activation response (TAR) element. TAR is a short noncoding RNA located at the 5' end of nascent viral transcripts, where it nucleates transcriptional activation by coordinating the viral protein Tat with the host cofactor cyclin T1 and the kinase CDK9, the catalytic unit of P_TEFb. This assembly relieves promoter proximal pausing of RNA polymerase II and promotes productive elongation. When the interaction is disrupted, transcription stalls and abortive short RNAs accumulate, with consequent impairment of viral replication. [230, 231] Evidence further indicates that HIV-1 integrase (IN) engages TAR at the RNA level through its C terminal domain, with a preference for structural motifs in the apical loop and contributions from the central bulge. (Figure 5.1)

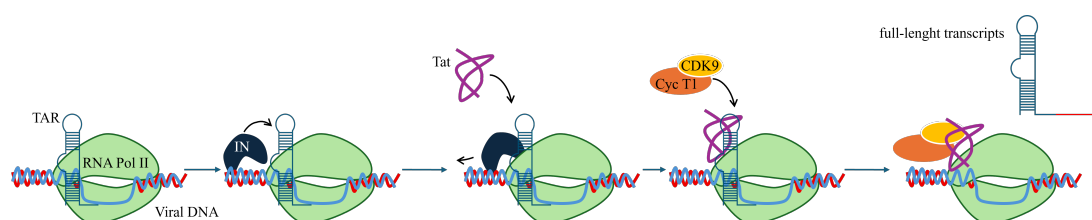


Figure 5.1. Schematic representation of the interplay among Integrase (IN), Tat, and TAR during the early phase of HIV-1 proviral transcription. After integration, Integrase binds the nascent TAR RNA and promotes conformational rearrangements that enable Tat recognition. Tat then displaces Integrase from TAR and recruits the catalytic P_TEFb complex, composed of cyclin T1 and CDK9, which stimulates RNA polymerase II elongation and yields full length transcripts.

Biochemical and cell based studies suggest that integrase contacts can transiently remodel the hairpin and assist early transcriptional activation by facilitating subsequent Tat loading, after which Tat predominates on TAR.[232] In this view, both bulge and loop act as cooperative regulatory modules that integrate integrase assisted remodeling with high affinity Tat recruitment and P_TEFb engagement. [232, 233]

Structurally, TAR adopts a stem loop architecture that features a central pyrimidine rich bulge and an apical loop (Figure 5.2), both indispensable for transactivation.

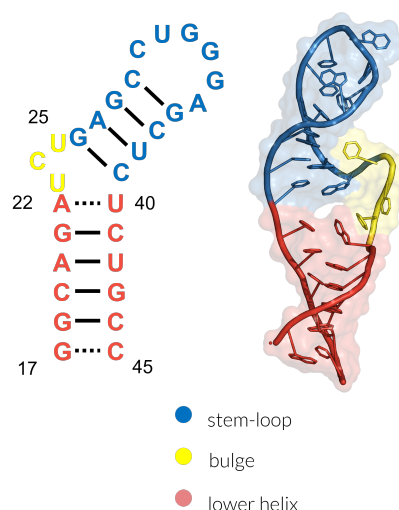


Figure 5.2. Structure of HIV-1 TAR. The secondary structure with residue numbering (left) and a representative conformer (right) illustrate the stem-loop (blue), the bulge (yellow), and the lower helix (red)

These motifs provide recognition elements for Tat and accessory factors, and they also encode conformational plasticity that couples structure and function. Upon effector binding in the major groove, the central bulge can reorganize cooperatively, the lower and upper stems can coaxially stack. In the ligand free state, the helix is more acutely bent, the major groove is wide and shallow.[230] Recent integrative studies suggest the co-existence of two states. In the ground state, the loop samples a broad distribution of open conformations in which most residues are unpaired and preorganized for recognition by the Tat super elongation complex. In the excited subensemble, the loop zips up into a narrower set of closed conformations that sequester key recognition residues, consistent with a regulatory role and with measurable effects on transactivation when this subensemble is stabilized.[234]

Despite decades of efforts, the development of clinically relevant small molecule agents that target TAR to block its biological functions has been largely unsuccessful. Early strategies aimed at the TAR Tat and TAR Tat T1 complexes, often based on

aminoglycosides[235] and intercalating agents,[236] lacked the required specificity and pharmacological profiles.

In this context, advanced computational approaches provide a natural framework to capture conformational heterogeneity and to identify elusive binding pockets. The SHAMAN protocol (Figure 5.3), see Paragraph 3.1.4.1, was introduced to detect small molecule binding sites in RNA ensembles by combining atomistic molecular dynamics with enhanced sampling and probe based mapping. [135]

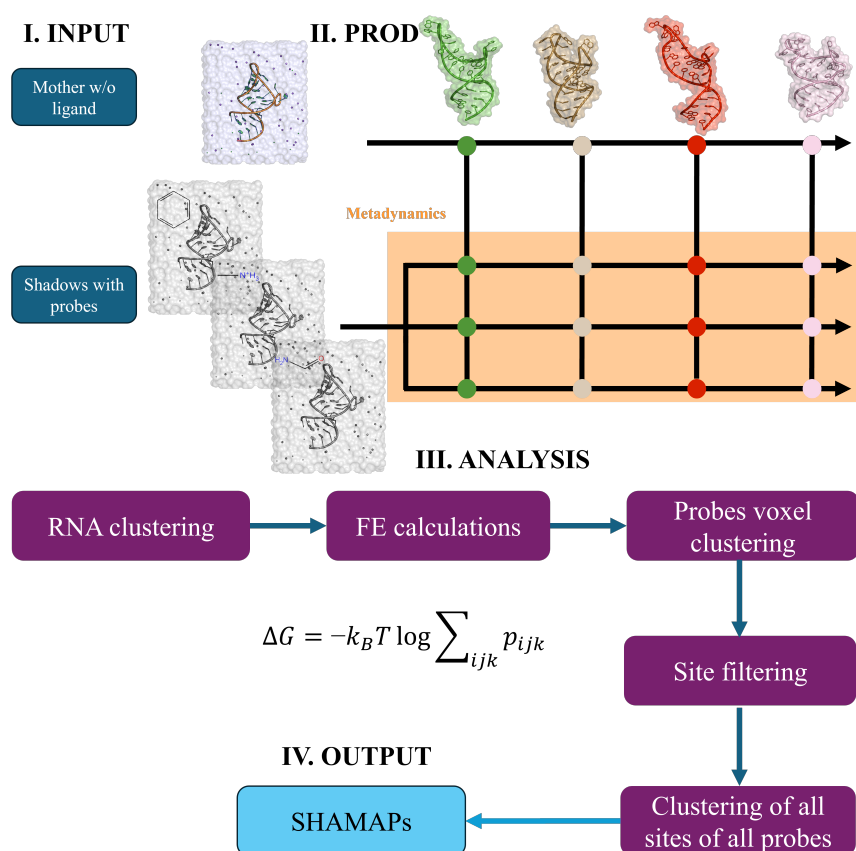


Figure 5.3. Workflow of SHAMAN for mixed solvent mapping on RNA. Input: select the RNA target and a panel of probe molecules, set up a mother system without ligands and multiple replica systems, each replica carries a different probe. Production: run an unbiased and unrestrained molecular dynamics simulation of the mother, broadcast the positions of RNA backbone atoms to the replicas that are restrained to follow the mother like shadows, accelerate probe exploration with metadynamics. Analysis: first, cluster the sampled RNA ensemble into representative conformations; second, for each RNA cluster and each probe, compute a free energy map from probe occupancy; third, merge contiguous voxels into interacting sites; fourth, assign to each site a free energy and a buriedness score and discard solvent exposed sites; fifth, within each RNA cluster, combine the interacting sites from all probes to obtain SHAMAPs. The procedure is detailed in Paragraph 3.1.4.1

In this chapter we apply SHAMAN to the HIV-1 TAR to characterize to identify druggable sites suitable for small molecule targeting in order to prevent or disrupt HIV-1-IN interface (Figure 5.4). We coupled SHAMAN derived sites to docking and to machine learning based scoring and selection, and we map a chemically diverse set to extract design implications for RNA pockets.

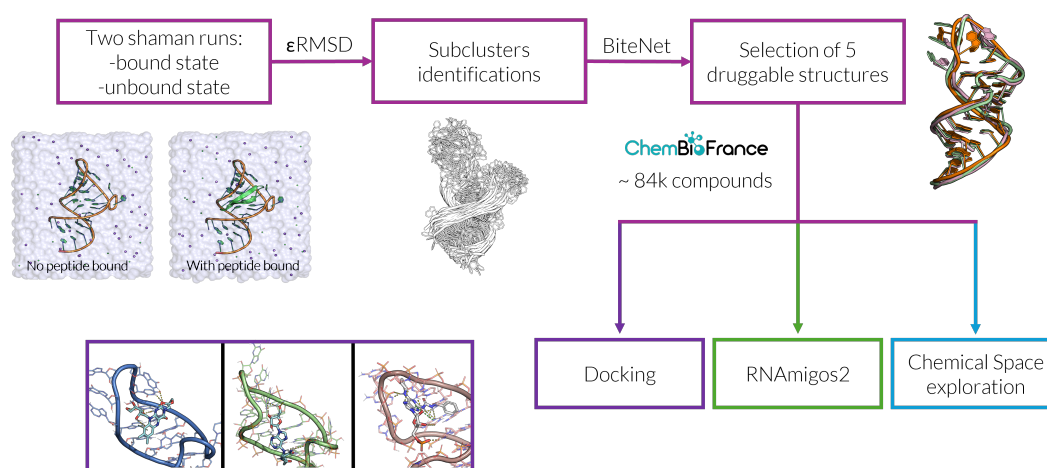


Figure 5.4. Schematic representation of the workflow employed in this study. The protocol begins with two parallel SHAMAN simulations in the presence and absence of the cyclic peptide. Ensemble refinement was performed using ϵ RMSD to obtain more compact subclusters, which were subsequently evaluated with BiteNet to prioritize druggable conformations. Finally, docking and complementary machine learning-based screening were applied to select candidate molecules.

By integrating conformational sampling, probe-based mapping, and virtual screening, we seek to advance the identification of candidate inhibitors of the TAR–IN interaction.

5.1 Methods

5.1.1 System setup and simulation protocol

The initial coordinates were obtained from the NMR structure of the HIV-1 TAR RNA in complex with a cyclic peptide that mimics the Tat protein (PDB ID: 2KXD)[237]. Two different simulation setups were designed to investigate the role of the peptide in modulating RNA conformational dynamics: (i) a system containing

only the RNA and (ii) a system including both the RNA and the peptide within the simulation box.

All molecular dynamics simulations were performed with GROMACS v.2021.5 patched with a development version of PLUMED.[137–139] The AMBER force field was employed to describe both the RNA and the cyclic peptide. In particular, the χ_{OL3} correction was used for RNA,[195] while the ff99SB parametrization was adopted for the peptide [238] and Joung and Cheatham parameters were adopted for the ions.[239] Parameters for the probe were derived using SAGE.[240] Simulation boxes were constructed in triclinic geometry and solvated with explicit OPC waters,[191] ensuring that each edge of the box extended by at least 10 Å from the closest RNA atom. Potassium and chloride ions were added to reproduce a bulk concentration of 0.15 M and to maintain charge neutrality of the systems. Long-range electrostatics were treated with the smooth particle mesh Ewald (PME) method, using a cutoff of 9 Å. van der Waals interactions were gradually switched off at 8 Å and fully truncated at 9 Å. The equations of motion were propagated with the leap-frog algorithm using a time step of 2 fs.

Each system underwent a multi-stage equilibration protocol prior to production. Energy minimization was first performed using the steepest descent algorithm. Subsequently, the systems were equilibrated for 10 ns in the NPT ensemble at 1 atm using the C-rescale barostat,[204] followed by an additional 10 ns in the NVT ensemble at 300 K using the V-rescale thermostat.[241] During both equilibration phases, harmonic positional restraints with a force constant of $400 \text{ kJ mol}^{-1} \text{ nm}^{-2}$ were applied to the RNA backbone and probe heavy atoms.

Production simulations were carried out for 3 μs per replica. To ensure consistency in the conformational space across replicas, the positions of the RNA backbone atoms from the mother system were communicated every 0.2 ps, and the corresponding atoms in each replica were restrained by an upper-wall harmonic potential with a force constant of $10000 \text{ kJ mol}^{-1} \text{ nm}^{-2}$, allowing a maximum RMSD deviation of 2 Å from the reference configuration. To enhance sampling of probe–RNA interactions, a well-tempered metadynamics bias was applied to the three Cartesian coordinates of the probe center of mass after alignment of the RNA backbone to the initial reference conformation. A bias factor of 10 was employed, with Gaussians of initial height

1.2 kJ mol⁻¹ and width 0.1 nm deposited every 1 ps. Finally, to prevent detachment of the probe from the RNA surface, the probe center of mass was confined within 10 nm of the closest RNA atom using an additional upper-wall harmonic restraint of 10000 kJ mol⁻¹ nm⁻².

5.1.1.1 Occupancy analysis and representative selection

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. We first concatenated all the trajectories of the mother and replica simulations, after removal of probes, water and ions, and fixed the discontinuities due to the periodic boundary conditions. We then clustered all the RNA conformations with the gromos implementation of Daura et Al.[151] using as metrics the RMSD calculated on the RNA backbone atoms with a cutoff of 3 Å. To reduce memory requirements, the clustering was first performed on a subset of frames (1 every 10) and then the excluded frames were assigned to the closest cluster using a python script based on the MDAnalysis library v. 2.2.0.[242] The cluster center was taken as the representative structure for each state. The cluster populations were calculated independently for the mother and each replica simulation and clusters populated <10% were discarded in the following analysis.

5.1.2 Virtual screening database preparation

A 3D molecular database was constructed using the Schrödinger Suite (Schrödinger Release 2023-2), starting from 2D structures obtained from the Chimiothèque Nationale (ChemBioFrance, <https://chembiofrance.cn.cnrs.fr/>). The 2D structures were converted into 3D representations, and stereoisomers were generated using LigPrep (Schrödinger Release 2023-2: LigPrep, Schrödinger, LLC, New York, NY, 2023). All possible ionization states at physiological pH (7 ± 1), together with relevant tautomeric forms, were generated using Epik.[243] To control for possible liabilities in the pharmaceutical profile of the database, compounds containing PAINS (Pan-Assay Interference Compounds) were flagged, and molecules containing metal ions were excluded from further analysis.

5.1.3 Structure preparation and virtual screening

Virtual screenings were conducted using Maestro Schrödinger version 2023-2. The selected RNA structures were prepared using the Protein Preparation Wizard (Schrödinger, LLC, New York, NY, 2023). For each representative structure, grids were generated using inner and outer boxes of 10 Å and 30 Å, respectively. These grids were centered on the geometric center of the identified binding sites, as identified by BiteNet,[76] ensuring comprehensive coverage and sufficient flexibility for ligand sampling during docking.

A hierarchical docking workflow was employed to efficiently screen and refine potential binders.[244] The process began with an initial screening using the High Throughput Virtual Screening (HTVS) scoring function as implemented in Glide to evaluate the entire database.[146, 168, 245] Top-ranking compounds were selected for further analysis and re-docked using the Extra Precision (XP) scoring function available in Glide, allowing for a more accurate consideration of solvation effects and structure–ligand interactions.[146, 168, 245] A final visual inspection was performed to assess binding-site compatibility and binding-mode consistency, with particular attention to the placement of key chemical functional groups.

In parallel with this workflow, additional molecules were identified through machine-learning predictions with RNAmigos2[75] and by exploring the chemical diversity space using the Canvas module implemented in the Schrödinger Suite.[186, 246]

5.1.4 Diversity-based compound selection

To generate a chemically diverse set of candidate molecules while minimizing redundancy with the pre-identified set of molecules from Docking and RNAmigos2, we performed diversity-based compound selection using Canvas module in Schrödinger 2023-2 (Schrödinger, LLC, New York, NY, 2023).

Molecular fingerprints were computed for all compounds using the MolPrint2D.[247] Fingerprints were stored with 64-bit precision to reduce bit-index collision. MolPrint2D encodes each molecule as a collection of atom-centered environments, where each heavy atom serves as a reference and all neighboring atoms

up to a defined topological depth ($D = 2$) are enumerated. Each environment is described by the combination of atom type, bond order, and topological distance, producing a non-hashed representation that preserves chemically meaningful features. This fingerprinting scheme captures local connectivity patterns and functional-group arrangements, allowing robust comparison across chemically diverse scaffolds.

Pairwise molecular similarity was quantified through the Tanimoto coefficient (τ), defined as:

$$\tau = \frac{c}{a + b - c} \quad (5.1)$$

where a and b are the number of bits set to one in fingerprints A and B , respectively, and c is the number of shared features between them. The resulting coefficient ranges from 0 (no shared environments) to 1 (identical fingerprints). Similarity matrices based on these coefficients were used to guide compound selection and clustering.

Compound selection was then performed using the Sphere diversity-selection algorithm implemented in Canvas. This algorithm operates directly in the molecular fingerprint space and iteratively selects compounds that maximize pairwise dissimilarity according to the Tanimoto distance ($d = 1 - \tau$). Starting from the pre-existing library the algorithm defines an exclusion sphere of radius r around each selected compound. All molecules within this exclusion radius are removed from further consideration, ensuring that the final selection is composed of mutually dissimilar compounds. The iterative process continues until either the target number of molecules is reached or the chemical space is fully covered.

5.1.5 Scaffold and Chemical Space Network Analyses

To characterize the structural composition and internal diversity of the selected compound set, we performed a two-step analysis combining scaffold decomposition and network-based representation of the chemical space.

Scaffold networks were constructed with RDKit.[248, 249] Prior to analysis, input structures underwent a standardized preprocessing workflow to minimize representation-dependent artifacts:

- retention of the largest fragment;
- normalization of formal charges;
- canonicalization of tautomeric states;
- removal of explicit hydrogens;
- perception of aromaticity.

From each standardized molecule, the Bemis–Murcko (BM) framework was computed by excising all side chains and retaining only ring systems and the linker atoms connecting them.[249] An RDKit scaffold network was then generated from the BM set via iterative ring–removal and ring–retention operations, allowing each compound to contribute multiple parent–child nodes along the resulting hierarchy.[250] Node identities were recorded as canonical SMILES prior to aggregation. For frequency analysis, substructure matching ignored stereochemistry and atom–mapping information, whereas differences in saturation pattern and heteroatom composition defined distinct nodes. Counts therefore refer to unique nodes in the generated network and include single–ring scaffolds (e.g., benzene) when produced by the algorithm.

To complement the scaffold-based view with a similarity-driven representation, we constructed chemical space networks based on pairwise Tanimoto coefficients (Eq. 5.1). For each molecule, circular fingerprints were generated using the Morgan algorithm (radius 2, 2048 bits). Nodes in the network correspond to individual molecules, and edges connect pairs whose fingerprint similarity exceeds a given threshold τ . Three thresholds were considered ($\tau = 0.40, 0.50, 0.60$) to capture the progressive tightening of structural relationships. Communities were identified using the greedy modularity optimization algorithm implemented in NetworkX.[251] Node colors represent community membership, while isolated nodes correspond to compounds lacking close analogues at the selected threshold. The networks were visualized interactively using the PyVis.[252]

5.2 Results and Discussion

5.2.1 Simulations and Pocket Characterization

The final set of probes used in this project was adapted from a previous study,[135] with the addition of a dibenzothiazine scaffold present in a known ligand (Figure 5.5).[253] The rationale was to exploit the local induced fit response elicited by this tricyclic scaffold during the SHAMAN daughter simulations, in order to promote enhanced rearrangements of nearby bases. This is enabled by the application of positional restraints limited to the nucleic acid backbone at prospective binding sites, while allowing base orientations and local conformational adjustments.

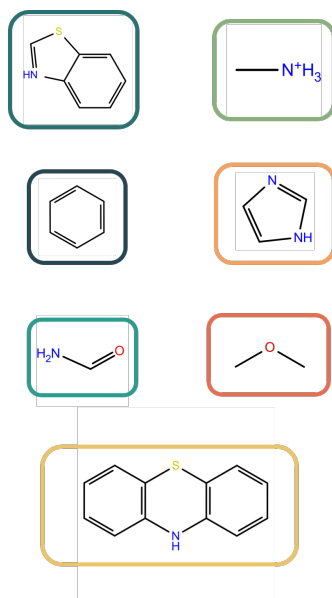


Figure 5.5. Chemical structures of the probes considered in this work. Teal, Benzothiazole, a bicyclic aromatic heteroaromatic system containing nitrogen and sulfur. Green, Primary ammonium, a positively charged amine. Dark blue, Benzene, the prototypical aromatic carbocycle. Orange, Imidazole, a five-membered aromatic heterocycle with two nitrogen atoms. Light teal, Formamide, the simplest amide functional group. Red, Ether group, an oxygen linker. Yellow, Dibenzothiazine, a tricyclic heteroaromatic scaffold containing nitrogen and sulfur.

Simulations were performed for 3 μ s. After the production phase, the analysis indicated a stable behavior of the RNA backbone. Clustering was carried out through the radial threshold algorithm from Daura et al.(Section 3.3.2) on the RMSD of nucleotide heavy atoms. Each run was dominated by a single cluster. In the apo system,

the first cluster accounted for 89.7% of frames, while all the others remained below 3%. In the presence of the cyclic peptide, this trend was even more pronounced, with the first cluster covering 98.6% of the simulation.

Focusing on the hotspot mapping, both simulations revealed two relevant binding regions, one near the apical loop, hereafter Pocket 1, and a second one adjacent to the bulge, hereafter Pocket 2. Probe density in the unbound (apo) system showed a clear preference for the apical loop pocket, mainly sampled by benzothiazole, formamide, and imidazole (Figure 5.6A). The second region near the bulge displayed more diffuse density and was visited by the same probes with an additional contribution from the ether fragment. In the peptide-bound (holo) system, dibenzothiazine played a central role in defining the density of both pockets, together with benzothiazole and imidazole. In the unbound (apo) state, Pocket 1 shows a $\Delta\Delta G$ of 1.07 kJ mol^{-1} relative to Pocket 2. In the holo state, a similar $\Delta\Delta G$ of 1.48 kJ mol^{-1} is found, with Pocket 1 becoming the most favorable site.

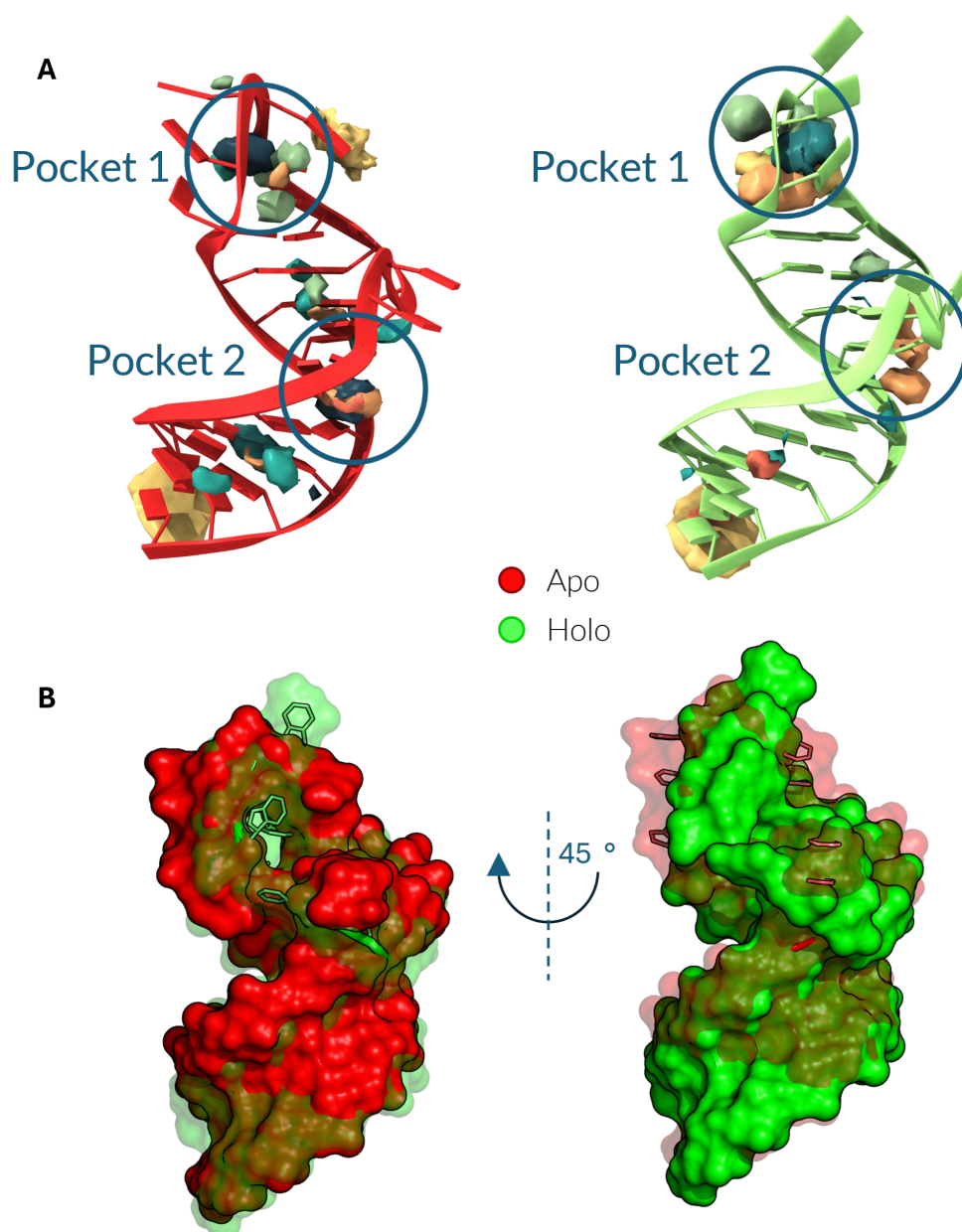


Figure 5.6. Probability maps for probe localization and pocket definition. Probe densities are encoded following Figure 5.5: teal, benzothiazole; green, primary ammonium; dark blue, benzene; orange, imidazole; light teal, formamide; yellow, dibenzothiazine. (A) Apo system, probe densities highlighting two main binding hotspot, Pocket 1 and Pocket 2. (B) Holo system, same representation, highlighting the strong contribution of the dibenzothiazine scaffold.

5.2.2 From Ensemble Clustering to Docking Integration

Inspection of cluster representatives revealed a critical limitation for downstream docking. In the apo system, the representative of the dominant cluster aligned well with the free-energy density clouds of the probes. In the holo system, however, we observed a barrier formed by nucleobases at the entrance of the main density region, which would impede access for the searching phase in docking, even though the pocket is thermodynamically favored. This highlights a potential limitation for directly integrating this approach into classical drug discovery workflows. In standard docking protocols the receptor is treated as rigid, which makes outcomes sensitive to the orientation of side chain groups in proteins and of bases in nucleic acids.

To address this problem, we introduced two additional steps into the SHAMAN-based workflow (Figure 5.4). First, we subclustered the dominant backbone cluster in regions of interest, apical loop and bulge, using ϵ -RMSD as defined by Bottaro and co-workers,[254] and QT clustering to ensure within-cluster homogeneity.[153] With a threshold of 0.7 Å we obtained multiple subclusters with non-negligible populations.

To select the most promising subconformations, we employed BiteNet,[76] a machine learning method that voxelizes local structural and chemical features into three-dimensional grids and evaluates the probability of ligand binding centers through a convolutional neural network. BiteNet produced probability maps for druggable centers, which we cross-referenced with SHAMAN density to select five structures with high scores and consistent interacting residues. From this procedure, 8 different pockets, located in 5 structural representative were selected for the next steps.

A first docking screen was carried out with Glide HTVS precision. The top scoring ten percent of molecules were retained for an additional step with Glide XP scoring function. The XP scoring function applies more stringent geometric and energetic filters, penalizes steric clashes, and accounts in greater detail for hydrogen bonding and hydrophobic enclosure, thereby reducing false positives and providing a more reliable ranking of potential binders. The final set of docked candidates was filtered by visual inspection to remove obviously strained poses and redundancies in scaffolds, given the limited chemical diversity of the library. A total of 90 molecules were selected.

In parallel, we performed an independent selection from RNAmigos2. For each representative conformation, we ranked compounds by model score, excluded those

already retained from docking, and chose additional candidates from the top two hundred predictions. This yielded 60 further molecules, for a combined selection of 150 compounds subjected to biological evaluation.

Finally, we completed the set of molecules by ensuring broad coverage of the available chemical space, with a reduced number of entries in order to mitigate the influence of scoring functions in preventing diverse scaffolds from being selected (See Section 5.1.4). From this procedure we ultimately selected 50 additional compounds, thereby bringing the final set of candidates to 200.

5.2.3 Analysis of the Final Set Properties

We characterized the distribution of common descriptors across the selected molecules (Figure 5.7). Overall, the majority of compounds exhibit molecular weights between 300 and 500 Da, consistent with typical drug-like space.

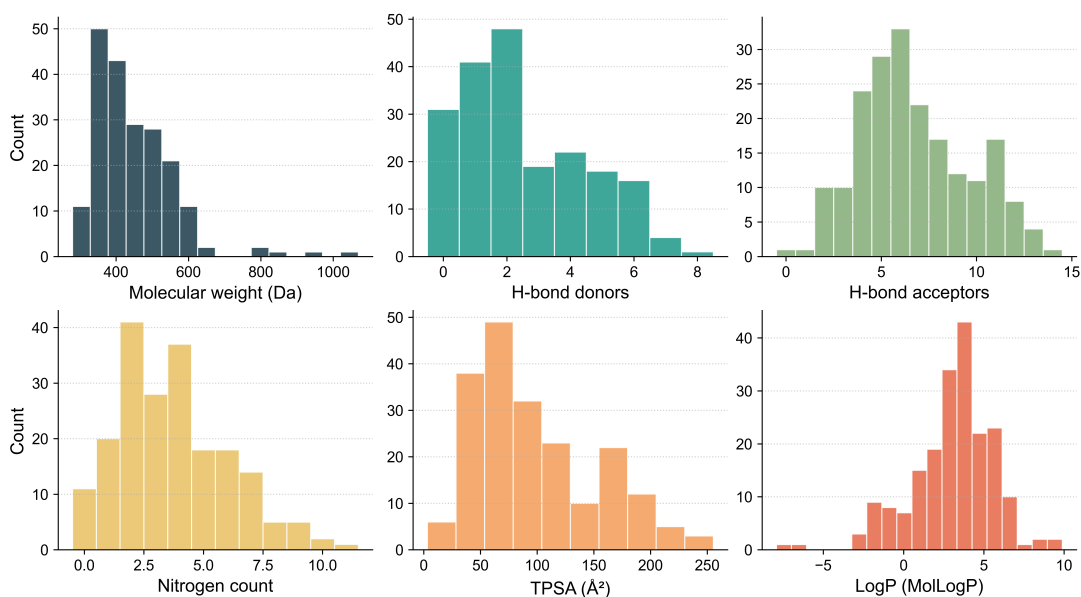


Figure 5.7. Histogram distributions of molecular descriptors for the selected molecules. Each panel reports the normalized density curve for a different property: molecular weight (dark blue), hydrogen bond donors (teal), hydrogen bond acceptors (green), nitrogen atom count (yellow), topological polar surface area (TPSA; orange), and calculated LogP (red).

The number of hydrogen bond donors is mostly below five, and the number of hydrogen bond acceptors predominantly falls between five and ten, consistent with

Lipinski’s rule of five. Although this rule is often regarded as outdated when applied rigidly, it remains a useful and non prescriptive heuristic to interpret the overall composition of the library. Nitrogen atoms are commonly present, with distributions peaking between two and four per molecule, reflecting the widespread inclusion of heteroaromatic and heteroaliphatic scaffolds. TPSA values concentrate below 150 Å², suggesting that most compounds are likely to maintain favorable permeability and oral bioavailability. Finally, the LogP distribution is centered between 0 and 5, indicating a balanced hydrophilic–hydrophobic profile. Together, these distributions confirm that the screened library largely populates the drug-like chemical space, while still maintaining a degree of structural diversity across key physicochemical properties.

We next built a Bemis–Murcko–derived scaffold network, in which each compound can contribute multiple parent and ring–retained scaffolds. The resulting network comprised a total of 1,762 distinct nodes (Figure C.1). The distribution of scaffold frequencies was highly skewed, with a small number of recurrent motifs dominating the library and a long tail of rarely observed cores.

The most frequent node was the benzene ring (175 occurrences), reflecting the prevalence of aromatic motifs in the dataset. Other common nodes included pyridine (51), tetrahydrofuran (46), pyridazine (45), and imidazole (45). Additional privileged structures such as pyrrole (41), purine (33), and indole (25) were also identified.

The dominance of aromatic and heteroaromatic scaffolds was expected, due to the possibility of interacting with the nucleobasis through $\pi - \pi$ stacking. Accordingly, these scaffold classes were prioritized during the scoring stage.

Beyond the most abundant motifs, over 1000 scaffolds were observed only once or twice, contributing to the overall chemical diversity of the dataset. This distribution suggested that, while the library is enriched in a limited number of recurrent privileged structures, it also retains a considerable level of structural variety.

Thanks to results on the scaffold network analysis, which highlighted the recurrent substructures present in the library, we extended our investigation to a chemical space network (CNS) representation. This approach enabled us to place the identified scaffolds into a broader structural context, capturing not only the most frequent cores but also their interconnections and relative positioning in chemical space. To further probe

these relationships, we constructed chemical space networks at increasing Tanimoto similarity thresholds (Figure C.2). At $\tau = 0.40$, the network already reveals the presence of a limited number of congeneric clusters, consistent with the overall chemical diversity of the set. Increasing the threshold to $\tau = 0.50$ produces a finer partitioning, with small communities reflecting closer structural relationships. At $\tau = 0.60$, only highly similar compounds remain connected, highlighting the presence of a few true analogue series (Figure 5.8), which are particularly useful to assess how limited structural variations may influence ligand binding.

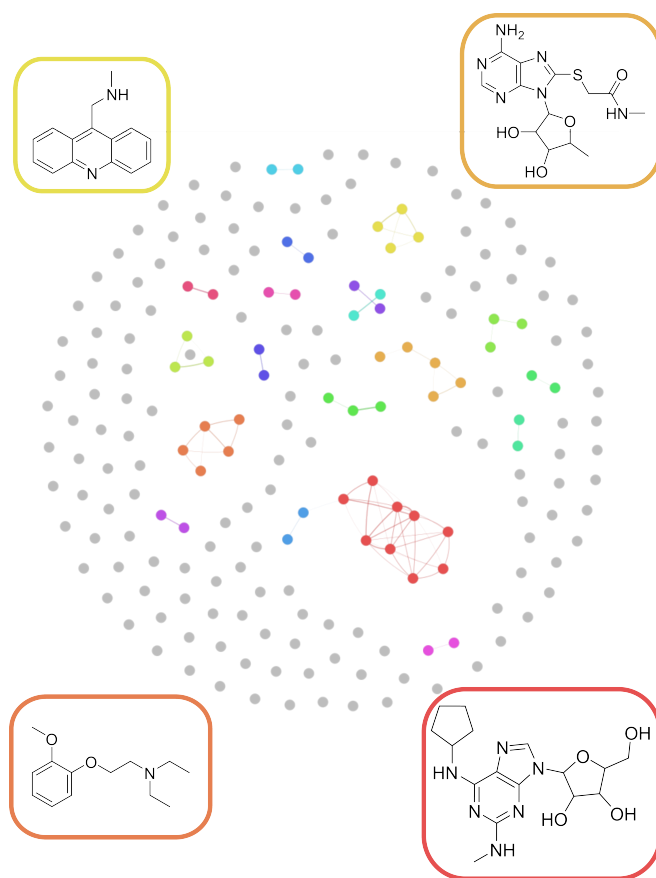


Figure 5.8. Chemical Space Network at Tanimoto similarity threshold $\tau = 0.60$. Only highly similar molecules remain connected, forming small communities of structural analogues (colored), while the majority of compounds appear as isolated nodes (grey), reflecting the limited presence of true congeneric series within the selected compounds. The boxes display the common scaffolds for the most populated communities, with box outlines matching the corresponding community colors.

The network metrics corroborate this picture (Table 5.1). As τ increases from 0.40 to 0.60, the graphs become sparser and more fragmented (density $0.0101 \rightarrow 0.0033$, av-

average degree $2.00 \rightarrow 0.65$; components $121 \rightarrow 162$, GCC fraction $0.17 \rightarrow 0.06$). The average clustering coefficient decreases ($0.264 \rightarrow 0.099$), while the global transitivity remains high ($0.67\text{--}0.79$), indicating that the few groups surviving at higher thresholds are triad-rich but fewer in number. Degree-entropy decreases ($1.74 \rightarrow 1.01$), consistent with the disappearance of hub-like nodes and the predominance of singletons or dimers. Modularity peaks at $\tau = 0.50$ ($Q = 0.81$), revealing the sharpest community structure at intermediate sparsification; at $\tau = 0.60$ only highly similar compounds remain connected, delineating a handful of analogue series.

Metric	$\tau = 0.40$	$\tau = 0.50$	$\tau = 0.60$
Nodes	200	200	200
Edges	200	112	65
Components	121	141	162
GCC fraction	0.17	0.09	0.06
Density	0.0101	0.0056	0.0033
Average degree	2.00	1.12	0.65
Clustering (avg)	0.264	0.229	0.099
Transitivity	0.671	0.789	0.766
ASPL (GCC)	3.11	2.36	1.75
Degree entropy	1.737	1.326	1.009
Modularity Q	0.722	0.810	0.761

Table 5.1. Graph metrics for the chemical space networks at increasing Tanimoto thresholds. Clust.: average clustering coefficient; Trans.: transitivity; ASPL(GCC): average shortest path length on the largest connected component; H(deg.): degree entropy; Q : weighted modularity.

This progressive analysis supports the conclusion that the dataset contains only a few congeneric chemotypes, with relatively limited analogue expansion.

5.3 Summary of Main Findings

- Probe driven mapping revealed two recurrent pockets on TAR, at the apical loop and at the bulge, with three microseconds simulations showing a stable backbone and a highly dominant conformational cluster. In the apo system densities

avored the apical loop, mainly for benzothiazole, formamide, and imidazole; in the holo system dibenzothiazine contributed strongly to both pockets.

- Inspection of cluster representatives exposed a steric barrier at the entrance of a thermodynamically favored region in the holo ensemble, which challenges direct coupling to rigid receptor docking. To address this limitation we refined the clustering procedure focusing on the apical loop and bulge, then intersected BiteNet probability maps with SHAMAN densities. This yielded five structural representatives and eight pockets that preserved SHAMAN supported interactions while improving accessibility for docking.
- Virtual screening adopted a two stage protocol with Glide HTVS followed by XP rescoring. Visual curation removed strained poses and scaffold redundancies, leading to 90 docked candidates; an independent RNAmigos2 ranking added 60 non overlapping compounds; a final diversity oriented completion contributed 50 further entries. The integrated workflow produced a prospective set of 200 candidates.
- Cheminformatics analyses indicated that the set populates drug like physico-chemical space, with most molecular weights between 300 and 500 Da, donors largely below five, acceptors predominantly between five and ten, TPSA concentrated below $150 \text{ \AA}^2$, and LogP centered between zero and five. Scaffold decomposition identified 1762 distinct Bemis Murcko frameworks with a marked skew toward a few recurrent motifs, including benzene in 175 molecules and frequent heteroaromatics such as pyridine, tetrahydrofuran, pyridazine, imidazole, pyrrole, purine, and indole, consistent with expected pi pi contacts with nucleobases. Chemical space networks at increasing Tanimoto thresholds showed few true analogue series at $\tau = 0.60$, suggesting ample room for structure activity expansion around the most promising chemotypes.

5.4 Conclusions

In this work we combined mixed solvent methods, virtual screening and cheminformatic strategies to characterize RNA pockets and guide ligand prioritization. The simulations revealed stable conformational ensembles dominated by a few clusters, while probe density highlighted recurrent pockets at the apical loop and bulge regions.

Integration with BiteNet and docking allowed us to refine subclusters and identify a limited number of chemotypes, with only a few true analogue series emerging from the chemical space network analysis. This confirms that the screened library, while diverse, contains privileged motifs recurrently associated with RNA binding.

The selected molecules are currently undergoing biological testing. Their validation will provide a critical benchmark to assess the predictive power of the proposed workflow. Beyond this specific case study, the methodology described here offers a generalizable framework to explore RNA ligandability and prioritize candidates for experimental follow-up. Future extensions may include the integration of orthogonal experimental data to further refine ensemble weighting, and the application of this workflow to larger compound libraries or other therapeutically relevant RNA targets.

6 Integrative Refinement of RNA Conformational Ensembles via the Maximum Entropy Principle

Starting from the biological and structural insights outlined in the previous chapter, we now transition from drug discovery campaigns to quantitative ensemble modeling of the HIV-1 TAR. Our central aim is to find an agreement between extensive molecular dynamics sampling with experimental observables using Maximum Entropy reweighting. In doing so, we seek to obtain an ensemble that is both thermodynamically consistent with the simulation protocol and statistically consistent with independent measurements, thereby bridging the gap between force field based dynamics and experiment driven inference.

MD simulations of small and flexible RNAs such as TAR, even when accelerated by enhanced sampling, tend to generate broad and heterogeneous ensembles that encompass both highly populated conformations and low probability states with potential functional relevance.[23] This behavior reflects several intertwined factors. The conformational landscape of TAR is high dimensional and rugged, with multiple basins separated by barriers that may be modest in free energy yet separated by slow collective coordinates.[234, 255] Moreover, functionally meaningful substates can be defined by subtle variations in non canonical interactions, sugar puckers, and backbone torsions, which are individually fast but collectively produce rare transitions when coupled across the molecule.[234, 255, 256]

In this setting, unbiased molecular dynamics often requires extensive trajectories and multiple independent replicas to achieve statistical convergence of observables and to sample transitions between substates with acceptable frequency.[115, 257] Enhanced sampling methods can mitigate these limitations by flattening selected components of the free energy landscape or by tempering the distribution to promote barrier recrossing, however, the finite accuracy of the force field can become more evident in these cases.[115] For these reasons, integrating extensive sampling with experimental restraints is essential to refine the prior ensemble into a posterior that remains

faithful to the simulation protocol while agreeing with independent measurements; in doing so, the joint use of simulations and experiments offsets their respective limitations and enables accurate reconstruction of atomic-scale conformational ensembles in solution.[88, 258, 259]

In this respect, we leverage the combination of Parallel Tempering, the Well Tempered Ensemble, and the theoretical formalism of Maximum Entropy (see Section 3.1.2) to enforce agreement with residual dipolar couplings while preserving the entropy of the prior. In this case, the prior is the ensemble generated by molecular dynamics.

We first generate an extensive ensemble for TAR using the Well Tempered Ensemble (see Paragraph 3.1.2.4). We then focused on the reweighted space, with emphasis on states that carry non negligible statistical weight. Building on the previous chapter, we incorporate druggability assessments and examine how pocket level scores vary as a function of the weight of each frame. Finally, we analyze available complexes of small molecules bound to TAR to determine whether the local pocket conformations observed in the complexes are accessible in the unbound simulations, which would support a conformational selection component, or whether they are largely absent, which would indicate a predominant induced fit contribution (Figure 6.1).

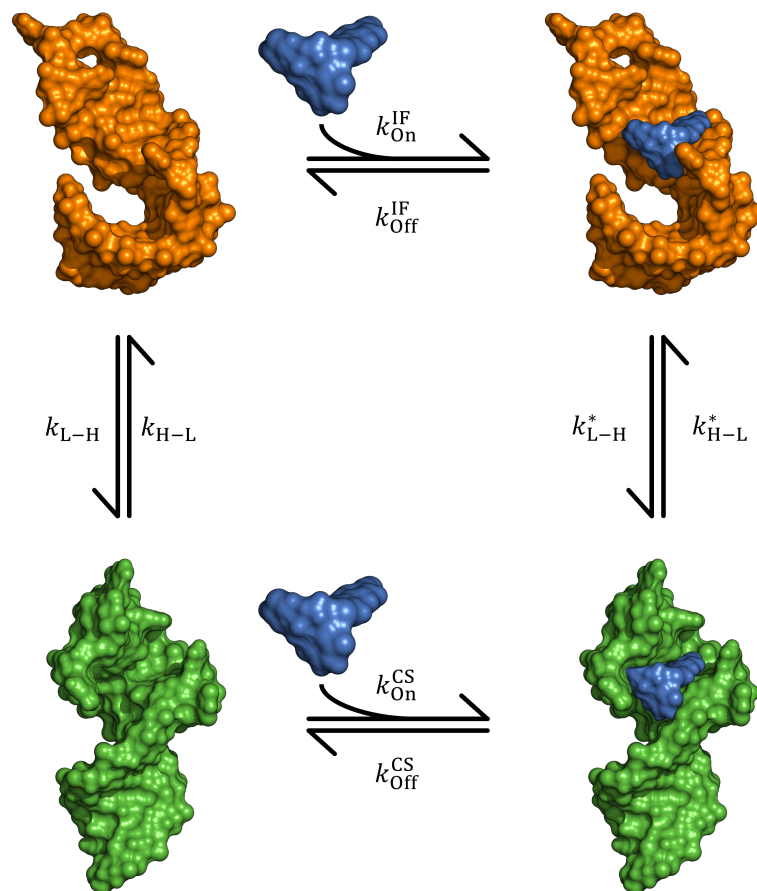


Figure 6.1. General mechanism of ligand binding coupled to conformational change. The macromolecule interconverts between a low-affinity conformation (L, orange) and a high-affinity conformation (H, green), while the ligand is represented in blue. The kinetic constants k_{On}^{IF} , k_{Off}^{IF} , k_{On}^{CS} , and k_{Off}^{CS} describe association and dissociation along the induced fit and conformational selection pathways, respectively. The interconversion rates between the unbound states (k_{L-H} , k_{H-L}) and between the ligand-bound states (k_{L-H}^* , k_{H-L}^*) represent the transitions between low- and high-affinity conformations in the absence and presence of the ligand.

6.1 Methods

6.1.1 System setup and simulation protocol

The initial coordinates were obtained from the NMR structure of the HIV-1 TAR RNA in complex with a cyclic peptide that mimics the Tat protein (PDB ID: 2KX5). [237] The AMBER force field was used to parametrize the RNA. In particular, the χ_{OL3} correction was used for RNA [195], the Joung and Cheatham parameters for ions, [239] and the OPC water model. [191] The simulation box was a truncated

octahedron filled with explicit water molecules, with a minimum distance of 20 Å between any RNA atom and the box boundary to allow substantial conformational rearrangements. Potassium and chloride ions were added to neutralize the system and to reproduce a bulk salt concentration of 0.15 M.

Long-range electrostatics were treated with the smooth particle mesh Ewald (PME) method, using a cutoff of 9 Å. van der Waals interactions were smoothly switched off between 8 Å and 9 Å, consistently with the simulation protocol adopted in the previous chapter. The equations of motion were integrated with the leap-frog algorithm using a time step of 2 fs.

For the short parallel tempering (PT) simulations, the temperature range 300–450 K was covered with 24 replicas, geometrically distributed. [260] Each system was equilibrated at its target temperature following a sequential protocol.

The Parallel Tempering in the Well-Tempered Ensemble (PT-WTE) simulation was run for a total of 700 ns using the potential energy as the only collective variable (CV). Gaussians with a height $W = 2.4 \text{ kJ mol}^{-1}$ and a width $\sigma = 500 \text{ kJ mol}^{-1}$ were deposited every 500 steps. A bias factor of 25 was applied to ensure sufficient overlap between the energy distributions of neighboring replicas, and replica exchanges were attempted every 100 steps (Figure 6.2). No additional restraints were applied during the PT-WTE simulations.

All molecular dynamics simulations were performed with GROMACS v. 2023.5 patched with PLUMED 2.9. [137–139]

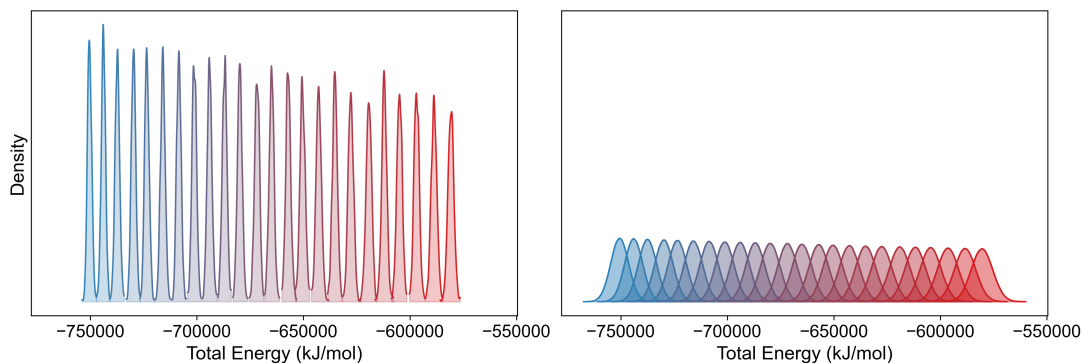


Figure 6.2. Energy distributions of replicas at increasing reference temperatures (color-coded from blue to red). Left: PT. Right: PT-WTE, both employing the same number of replicas. In standard PT, only negligible overlap between adjacent replicas was observed, implying that a substantially larger number of replicas would be required to ensure efficient exchanges. In contrast, the PT-WTE protocol significantly broadened the energy distributions, thereby enhancing overlap across replicas and allowing the same temperature range to be covered with fewer replicas.

6.1.2 Residual Dipolar Couplings (RDCs)

Residual dipolar couplings (RDCs) arise when molecules in solution experience a partial alignment with respect to the external magnetic field, such that dipolar interactions are not completely averaged out by isotropic tumbling. As a result, RDCs provide long-range orientational restraints that are highly complementary to short-range contacts such as NOEs, and are particularly informative for RNA.

The dipolar interaction between two nuclei i and j can be expressed as

$$D_{ij} = -\frac{\mu_0}{4\pi} \frac{\gamma_i \gamma_j \hbar}{2\pi r_{ij}^3} \left\langle \frac{3 \cos^2 \theta_{ij} - 1}{2} \right\rangle, \quad (6.1)$$

where γ_i and γ_j are the gyromagnetic ratios of the two nuclei, r_{ij} is their internuclear distance, and θ_{ij} is the angle between the internuclear vector and the external magnetic field. In isotropic solution the angular average vanishes, whereas in an alignment medium a residual contribution persists.[261]

In practical applications, RDCs are commonly expressed in terms of the alignment tensor. In this case, the coupling can be written as

$$D_{ij} = -S \frac{\gamma_i \gamma_j \mu_0 \hbar}{16\pi^3 r_{ij}^3} \left[A_a (3 \cos^2 \theta_{ij} - 1) + \frac{3}{2} A_r \sin^2 \theta_{ij} \cos 2\varphi \right], \quad (6.2)$$

where S is a scaling factor reflecting the degree of alignment, A_a and A_r are the axial and rhombic components of the Saupe order tensor, and (θ, φ) are the polar coordinates of the internuclear vector with respect to the alignment frame. This formulation highlights how RDCs report on the average orientation of bond vectors rather than on local distances, making them sensitive probes of global RNA architecture.

Since experimental and back-calculated RDCs may differ by an overall scaling factor related to the effective degree of alignment, the two datasets are compared through a linear rescaling procedure, such that

$$\langle F_{\text{calc}} \rangle \approx \alpha F_{\text{exp}}, \quad (6.3)$$

where α is an optimized scaling parameter obtained by linear regression.

In this work, RDCs were back-calculated from the simulated ensembles using PALES with a global alignment setting.[262] For each dataset, the alignment scaling factor was optimized against the corresponding experimental values to account for differences in the effective degree of alignment between experiment and simulation.

6.1.3 SHAPE reactivity

Selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE) provides, at single-nucleotide resolution, an experimental proxy for local flexibility and accessibility: unpaired or dynamically disordered regions display higher reactivity. In SHAPE, 2'-hydroxyl groups are modified by electrophilic reagents via acylation of the ribose 2'-OH, generating reverse-transcription stops whose frequency reports the per-nucleotide reactivity profile.[263] For back-calculation from molecular dynamics, per-nucleotide SHAPE reactivity can be obtained by ensemble-averaging structural descriptors.[264] In this work, we compared experimental SHAPE reactivity profiles with an MD-derived flexibility metric based on inter-nucleotide distance fluctuations between consecutive residues, with the per-residue value obtained by averaging the fluctuations of the adjacent pairs ($i - 1/i$ and $i/i + 1$).

6.1.4 SiteMap

SiteMap evaluates potential ligand-binding sites by computing several descriptors: the number of site points (n), the enclosure score (e), and the hydrophobic/hydrophilic balance (p). From these, two composite indices are defined, the SiteScore and the Dscore, which are widely used to quantify pocket quality and druggability.[265]

The SiteScore is calculated as

$$\text{SiteScore} = 0.0733 \sqrt{n} + 0.6688 e - 0.20 p. \quad (6.4)$$

The druggability score (Dscore) has a similar form, but places more emphasis on enclosure and penalizes hydrophilicity more strongly:

$$\text{Dscore} = 0.094 \sqrt{n} + 0.60 e - 0.324 p. \quad (6.5)$$

In the recently developed RNA-specific mode of SiteMap, the hydrophobic/hydrophilic balance term is explicitly included in the scoring function. This modification reflects the more polar and solvent-exposed nature of RNA binding sites and improves SiteMap performance in identifying true ligand-binding pockets in nucleic acid receptors. The modified SiteScore takes the form [69]:

$$\text{SiteScore}_{RNA} = 0.0733 \sqrt{n} + 0.6688 e - 0.20 p + b, \quad (6.6)$$

where b is the balance term. It is important to note that the coefficients appearing in Eq. 6.6 were obtained through an empirical parametrization guided by benchmarking against experimentally validated RNA ligand binding sites, with the objective of improving the identification and ranking of true ligand occupied pockets.

6.2 Results and Discussion

We characterized the conformational landscape of the highly flexible HIV-1 TAR RNA, leveraging Parallel Tempering in the Well-Tempered Ensemble (PT-WTE) for broad exploration of rugged free-energy surfaces and subsequent refinement against solution observables (Figure 6.3). This integrative framework reconciles extensive

6 Integrative Refinement of RNA Conformational Ensembles via the Maximum Entropy Principle

sampling with experimental information, mitigating force-field bias and exposing functionally relevant substates. In the following, we (i) quantify ensemble heterogeneity and its agreement with experiment, (ii) delineate the dominant conformational substates, and (iii) evaluate druggability-relevant features emerging from the refined ensembles.

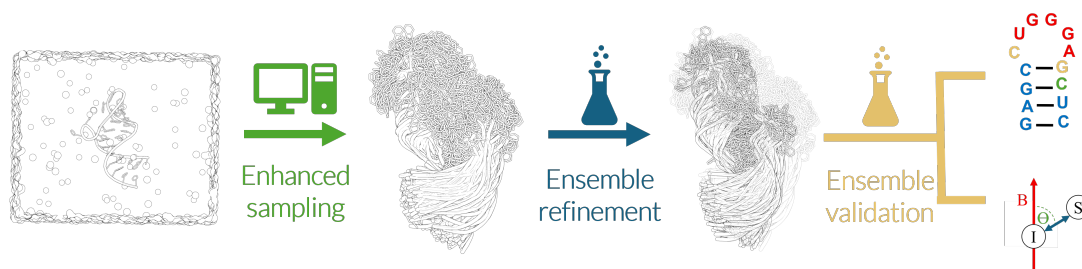


Figure 6.3. Schematic representation of the workflow followed in this study. The protocol begins with molecular dynamics simulations of TAR within the PT-WTE framework, proceeds through ensemble refinement using the maximum entropy principle, and culminates with validation against independent experimental observables. This integrative strategy aims to reconcile extensive conformational sampling with experimental accuracy.

A preliminary assessment of sampling efficiency was obtained by monitoring structural descriptors along the 700 ns PT-WTE trajectory. We focused on the root-mean-square deviation (RMSD), ε RMSD and the radius of gyration (R_{gyr}), which provide complementary views of TAR’s global and local dynamics (Figure 6.4 and Figure D.1).

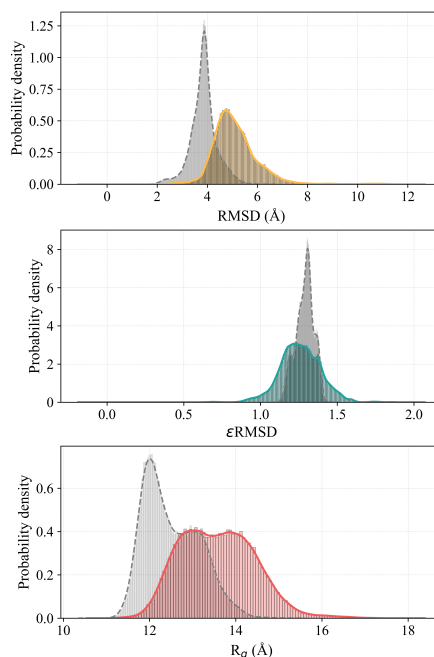


Figure 6.4. Probability density distributions of RMSD (top), ϵ RMSD (middle), and radius of gyration R_g (bottom) computed over the same 700 ns trajectory. Bars report normalized histograms, solid curves show Gaussian kernel density estimates. For each metric, the distribution obtained from a 700 ns plain MD simulation of the same system is shown in grey for comparison.

RMSD quantifies the overall deviation from a reference structure and reveals whether the system undergoes large-scale rearrangements or stabilizes around specific conformers. In contrast, ϵ RMSD is an RNA-aware metric that emphasizes the geometry of base pairing and stacking, thereby increasing sensitivity to local rearrangements within helices and loops. Finally, R_{gyr} reports on global compactness, distinguishing expanded conformations from compact states. RMSD values were computed after least-squares alignment on all heavy atoms; (ϵ RMSD) was evaluated on the all the nucleobases; R_{gyr} was computed over all heavy atoms.

The distributions indicate that TAR does not remain confined to a single conformational basin but undergoes continuous fluctuations spanning both compact and extended geometries. The RMSD rises rapidly within the first tens of nanoseconds, reflecting relaxation away from the initial reference structure, and subsequently oscillates between 4 and 10 Å, consistent with transitions among multiple substates. The ϵ RMSD values are broadly distributed between 0.8 and 1.5, confirming persistent local rearrangements likely occurring in the apical loop and bulge regions, which are

known for their intrinsic flexibility. The R_{gyr} exhibits frequent excursions between 12 and 17 Å, underscoring the ability of TAR to sample both compact and extended states within a single trajectory.

These preliminary observations confirm that the PT-WTE protocol effectively promotes barrier crossing: the system does not remain trapped in local minima but explores a broad ensemble of conformations. Moreover, the observed structural variability aligns with the view that small RNAs populate a dynamic equilibrium of interconverting states rather than a single stable structure.

While the PT-WTE simulations provided extensive coverage of the conformational space of TAR, the raw ensembles obtained from MD may not reproduce experimental observables with quantitative accuracy, this discrepancy arises from limitations of the force field and from finite-sampling effects. To overcome these issues and to reconcile simulation with experiment, we refined the PT-WTE ensemble by reweighting conformations against NMR residual dipolar couplings (RDCs) using the maximum entropy principle with the formulation from Cesari et Al.[266]

6.2.1 Maximum entropy reweighting and validation

We applied MaxEnt reweighting to minimally adjust the statistical weights of PT-WTE conformations so that ensemble-averaged observables match experimental data within their uncertainties (See Paragraph 3.1.3). This refinement preserves ensemble diversity and keeps the posterior distribution close to the WTE prior, thereby mitigating overfitting. RDCs were chosen as target observables. RDCs report on average internuclear vector orientations relative to the magnetic field and thus provide long-range constraints that are highly sensitive to inter-helical motions and relative domain orientations. Given that TAR dynamics are dominated by such collective rearrangements, incorporating RDCs is expected to sharpen the distribution of inter-helical geometries and improve the quantitative agreement between simulations and experiment.

We next optimized the regularization strength σ to balance accuracy and generalization. To this end, we performed 10-fold cross-validation with randomized splits: in each fold, 90% of RDCs were used to fit the MaxEnt multipliers and the remaining 10% were held out to compute the root-mean-square error (RMSE) (Figure 6.5E).

For each σ , we averaged the RMSE across folds to obtain a robust estimate of predictive performance, and selected the σ minimizing this metric. Using the selected σ ($\sigma = 7$ Hz), we recalculated the weights and recomputed the RMSE over all observables. We also monitored the RMSE and the Kish effective sample size to diagnose potential overfitting. The presence of a plateau in the proximity of the minimum, at $\sigma = 7$ Hz, indicates the absence of overfitting and suggests that the reweighting procedure remains stable in this region.

As shown in Figure 6.5A, back-calculated RDCs from the raw PT-WTE ensemble (green) deviate systematically from experiment (grey), with larger discrepancies in the apical loop and bulge. Quantitatively, the initial agreement was modest ($R \approx 0.68$, RMSE = 8.4 Hz), motivating MaxEnt refinement. The cross-validated selection of σ yielded a posterior that improved agreement on held-out RDCs while preserving a sizable effective sample size, indicating that the ensemble was sharpened without collapsing onto a few frames.

6 Integrative Refinement of RNA Conformational Ensembles via the Maximum Entropy Principle

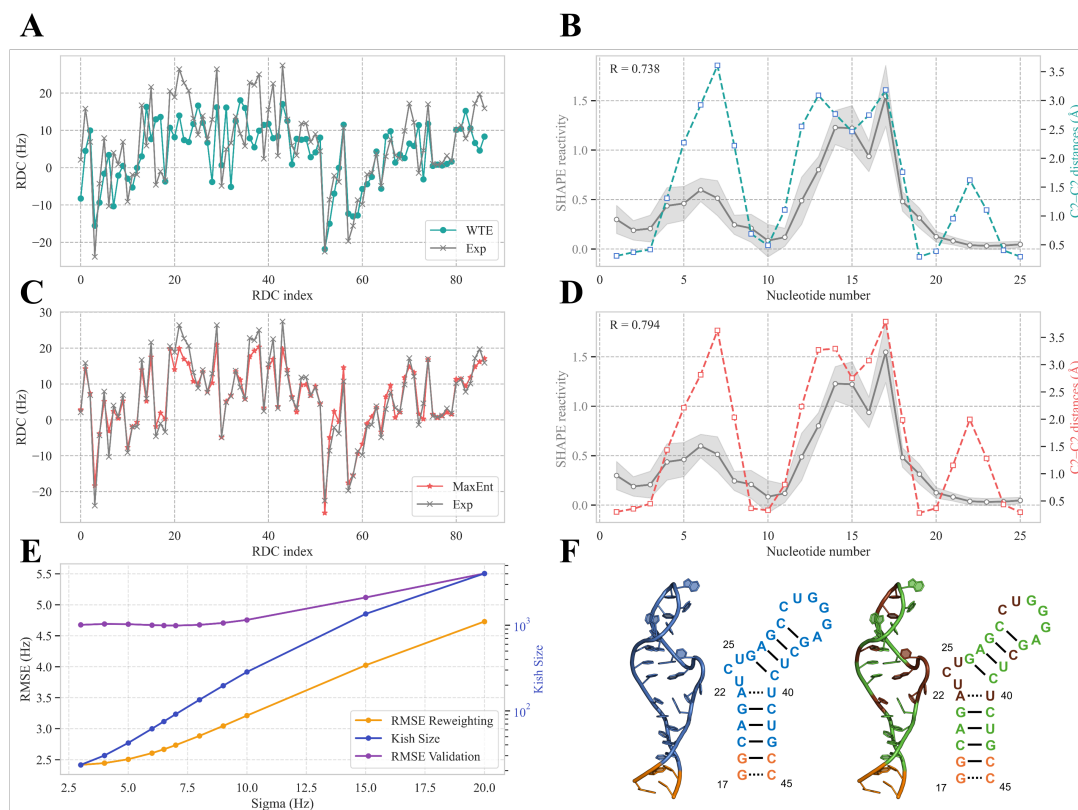


Figure 6.5. Maximum entropy reweighting of TAR ensembles against RDCs and validation with independent data. (A) Comparison between experimental RDC values (grey) from previous NMR studies and back-calculated RDCs from the raw WTE ensemble (green). (B) Comparison between experimental SHAPE reactivity profiles (grey) and computed fluctuations in the mean inter-nucleotide distances between consecutive residues ($i-1/i$ and $i/i+1$). (C) Comparison between experimental RDC values (grey) from previous NMR studies and back-calculated RDCs from reweighted back-calculated RDCs (red). (D) Comparison between experimental SHAPE reactivity profiles (grey) and computed inter-nucleotide distance fluctuations between consecutive residues, with the per-residue metric obtained by averaging the fluctuations of adjacent pairs ($i-1/i$ and $i/i+1$) after reweighting (red). (E) Cross-validation scheme for the optimization of the σ parameter. The root mean square error (RMSE) obtained from 10-fold validation of the RDC dataset (violet) is plotted as a function of σ . The associated Kish size is reported in blue, whereas the RMSE computed over the full dataset is shown in yellow. (F) Structural representation of TAR highlighting the residues included (blue) or excluded (orange) from SHAPE-based validation (right); residues correctly reproduced by the reweighted ensemble are shown in green, while those with the largest deviations are marked in brown (right).

After maximum-entropy reweighting, agreement improved markedly (Figure 6.5C): the correlation with experimental RDCs increased to $R \approx 0.97$, while the RMSE decreased to 2.8 Hz. Cross-validated test errors displayed a minimum at intermediate σ , indicating that the selected regularization balances fitting accuracy with preservation of ensemble diversity. Despite this good agreement, the Kish effective

sample size remained limited ($K = 74.2$), implying that the posterior concentrates weight on a relatively small subset of frames. This highlights the intrinsic difficulty of capturing the dynamics of TAR.

While SHAPE data proved valuable for independent validation, their direct use as target observables in MaxEnt refinement would not be straightforward. Unlike RDCs, which are ensemble-averaged quantities that can be computed for each conformation as functions of internuclear vectors, SHAPE reactivities are indirect measures of local flexibility and solvent accessibility, influenced by multiple chemical and structural factors. Consequently, a per-frame SHAPE observable $s_i(\mathbf{q})$ cannot be unambiguously defined, preventing a rigorous inclusion in the MaxEnt functional. In principle, one could attempt a reweighting using structural proxies correlated with SHAPE (such as fluctuations in local stacking distances or base-pair opening), yet this would represent an approximate mapping rather than a direct refinement against experimental data. For this reason, we used the two available SHAPE datasets to further validate the reweighted ensemble.

As an independent validation, we compared the simulated ensemble against SHAPE reactivity data from *ex virio* and transcript samples (Figure 6.5B-D, Figure D.2), neither of which were used for fitting. In both cases, the reweighted ensemble reproduced the experimental profiles more closely than the raw PT-WTE ensemble. Notably, the correlation with SHAPE reactivities increased from $R = 0.73$ before reweighting to $R = 0.79$ after reweighting in the transcripts data and a similar trend is observed in the *ex virio* dataset. While the absolute increase in the correlation coefficient is moderate, its consistency across datasets supports an improved agreement between simulation and experiment for observables that primarily report on local backbone flexibility and base stacking propensities.

6.2.2 Redistribution of conformational ensembles after reweighting

To characterize the conformational space sampled in our simulations, we characterized structures using G-vectors, a feature set well suited to RNA because it is sensitive to subtle rearrangements of base pairing/stacking that can nucleate or disrupt ligandable pockets. As a reference, we collected experimental HIV-1 TAR structures

from the PDB in the unbound (red) and ligand-bound (blue) states and projected them onto the same reduced space. Dimensionality reduction by PCA on the $\mathbf{G}$ -vector representation allows us to visualize the high-dimensional manifold in two dimensions. The projection (Figure 6.6A right) shows that the raw PT-WTE ensemble spans a broad range of conformations that envelop the regions populated by both unbound and ligand-bound PDB structures, indicating that the simulation recapitulates the diversity of the unbound RNA as well as states compatible with ligand binding.

After maximum-entropy reweighting (Figure 6.6A, left), the geometric support of the ensemble changes little, but the probability mass is markedly redistributed. Populations shift toward unbound-like states, including regions not sampled by available PDB structures, while overextended conformations that were overrepresented in the prior are down-weighted, as in the case of cluster 3 (yellow) and cluster 4 (orange). At the same time, a few discrete microstates that overlap with ligand-bound crystallographic references gain statistical weight, suggesting a selective emphasis on experimentally supported configurations without collapsing the ensemble.

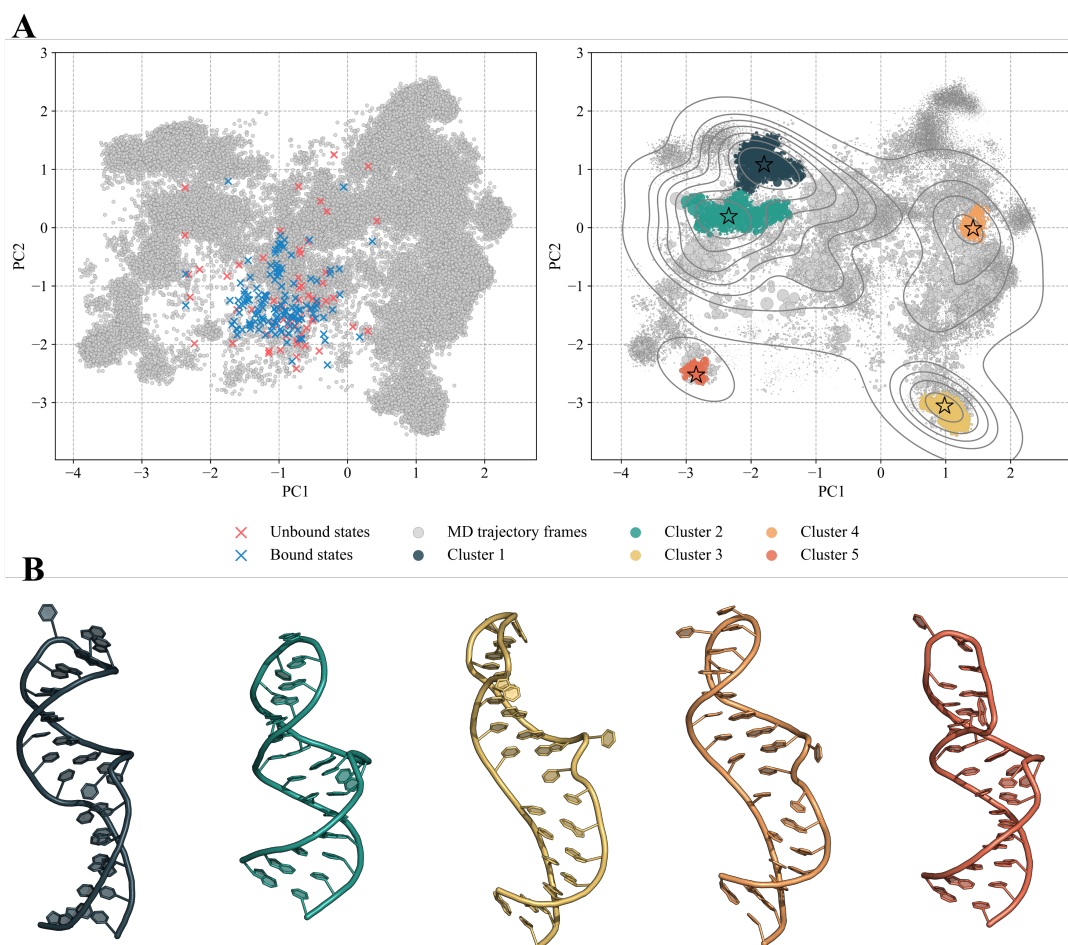


Figure 6.6. Conformational space of TAR before and after maximum entropy reweighting. (A) Two-dimensional projection of the conformational space sampled during the PT-WTE simulation (grey points), compared with reference PDB structures of TAR in the unbound (red) and bound (blue) states. The left panel shows the points distribution before reweighting, while the right panel displays the same projection after applying MaxEnt reweighting. Grey contour lines trace isodensity levels from a kernel density estimate computed with the MaxEnt weights, highlighting high probability basins and their spatial extent. Eight levels are shown with a bandwidth adjustment of 0.65 and a density threshold of 0.02. PC1 and PC2 explain 16.72% and 11.25% of the total variance, respectively. Representative structures of the main clusters are indicated by star markers. (B) Representative structures of the main clusters identified in the reweighted ensemble.

Representative post-reweighting conformations reveal a spectrum of inter-helical arrangements. Cluster 1 is compact and near-coaxial, with a modestly splayed bulge and the apical loop folded toward the stem. Cluster 2 remains compact but exhibits a mild inter-helical twist that extrudes the bulge and opens a narrow cleft along the minor groove, suggesting a pocket-competent intermediate. Clusters 3 and 4 are more extended, with the helical arms splayed and the loop projected outward, increasing

solvent exposure but yielding a diffuse, poorly defined cavity. Cluster 5 is compact yet asymmetric, with the loop bending toward the bulge to form transient tertiary contacts that may stabilize a lateral microcavity or modulate access to the principal pocket.

From a drug-discovery perspective, cluster 1 (primary) and clusters 2 and 5 (secondary) are promising targets for structure-based design, as their geometries can accommodate a structured, albeit transient, pocket.

We next mapped SiteMap druggability scores onto the PCA landscape to relate geometry with pocket quality (Figure 6.7 and Figure D.3). In the PT-WTE prior, scores span a broad range, with a dense basin of low SiteScore corresponding to overextended conformations that lack well-defined hotspots for small-molecule recognition. Outside this basin, most regions exhibit medium-to-high SiteScore values, typically associated with compact or kinked inter-helical arrangements. Interestingly, after Max-Ent reweighting, which only account for experimental observables, the probability of low-scoring states decreases markedly, with population shifting toward medium-high scoring basins. Nonetheless, a small subpopulation of elongated conformations persists with non-negligible statistical weight. This could suggest the co-existence of ligand accessible structure in the ensemble in equilibrium, with non-ligandable ones.

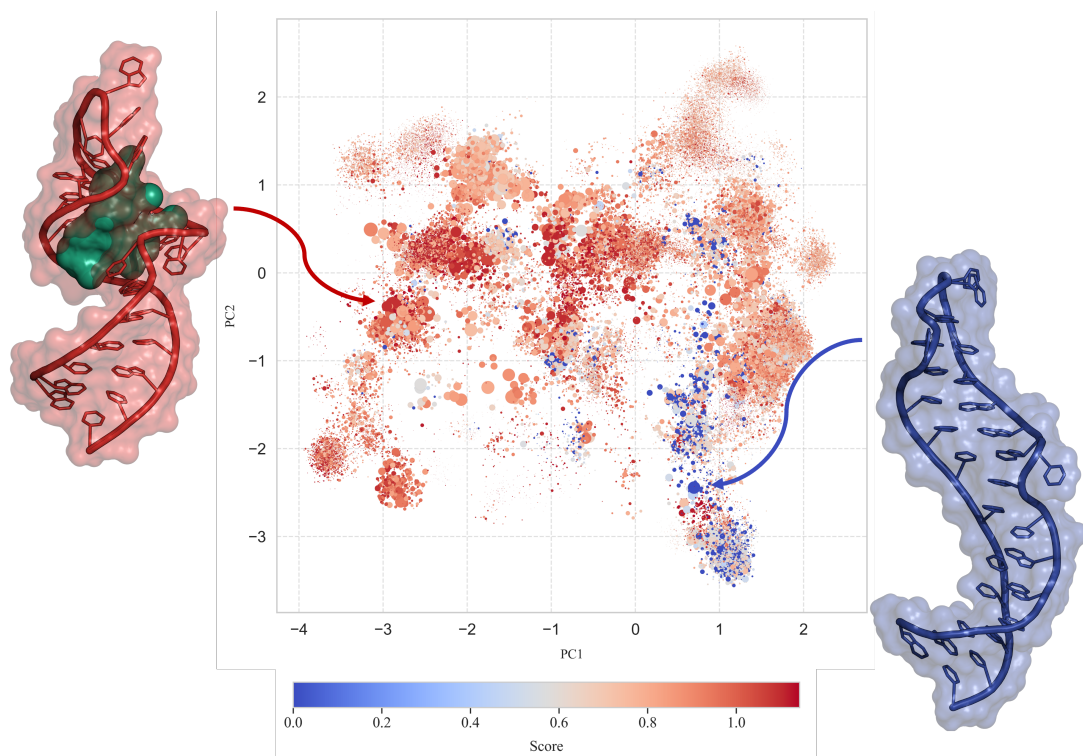


Figure 6.7. Two dimensional projection of the conformational space sampled during the PT-WTE simulation. Colors are mapped to SiteScore values, from blue for low scores to red for high scores. Point sizes are adjusted according to their statistical weight. Two prototypical examples of low and high scoring regions are highlighted. In red, the best scoring conformation within its the basin is shown, with the corresponding binding pocket in green; in blue, the worst performing conformation.

6.2.3 Comparison with available ligand-binding pockets

Mechanisms of molecular recognition in proteins are often framed as conformational selection versus induced fit: in the former, pre-existing substates are selectively stabilized upon binding; in the latter, binding causes rearrangements that are rare or absent in the unbound ensemble.[267] In practice, recognition typically reflects a continuum between these limits. Although less explored for RNA targets, analogous phenomena are expected given RNA's pronounced conformational plasticity.[268–270]

Guided by this premise, we compared small-molecule-bound HIV-1 TAR structures in the PDB (Figure 6.8) and their ligand-binding pockets to those present in our simulations. As in previous analyses, conformations were represented by G-vectors. Pocket boundaries in PDB complexes were defined by a distance cutoff of 4.5 Å between the ligand and at least one RNA heavy atom.

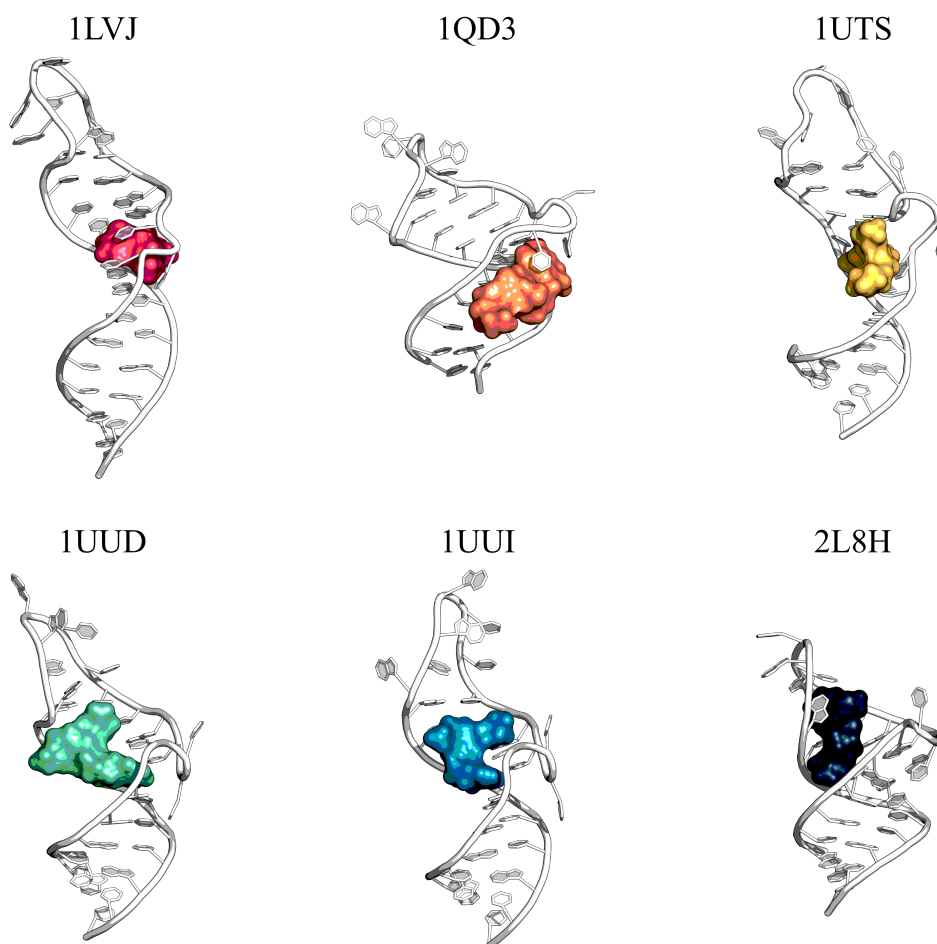


Figure 6.8. Three dimensional structures of the examined HIV-1 TAR complexes with small molecules. Ligands are colored by PDB identifier, using a scheme consistent with Figures 6.10 and D.4. The corresponding PDB code is displayed above each structure.

We first examined the region of conformational space proximal to the crystal structures, to quantify the relative populations of conformers that could convert to comparable structures with minor rearrangements.

To quantify the balance between these mechanisms, we performed QT clustering (See Paragraph 3.3.2) using the same metrics as for dimensionality reduction, and we evaluated cluster populations as a function of increasing ε RMSD from the crystallographic references and the cluster representatives (Figure 6.9).

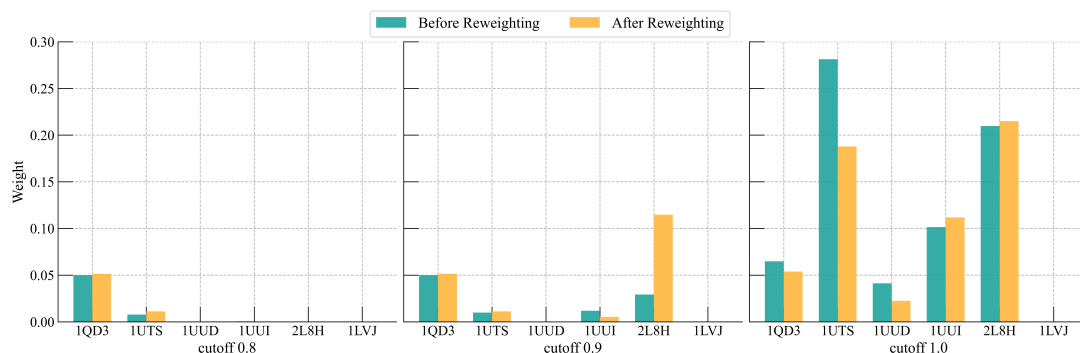


Figure 6.9. Population of TAR conformations within increasing ϵ RMSD thresholds (0.8, 0.9, and 1.0) relative to reference ligand-bound structures. The green bars correspond to the pre-reweighting ensemble, while the yellow bars represent the reweighted ensemble.

Comparing the prior and the reweighted ensembles shows that 1QD3 [271] remains consistently represented across all thresholds, with only minor changes in weight. At the lowest threshold, approximately 5% of the ensemble populates conformations whose pocket is compatible with the 1QD3 bound site. A similar pattern holds for 1UTS [236], although populations below 0.9 are limited. At the highest threshold, the cumulative weight for 1UTS decreases after reweighting, indicating a redistribution away from structures compatible with that reference. Increasing the cutoff to 0.9 brings the 1UUI [272] and 2L8H [273] references into the ensemble, with 2L8H showing a marked increase after reweighting. At a cutoff of 1.0, 1UUD [272] also becomes populated, whereas 1LVJ [253] remains poorly represented, with the closest cluster lying beyond the considered range. Overall, the analysis of the reweighted ensemble reveals that conformations compatible with all the investigated structures, except for 1LVJ, remain represented, indicating that these states are consistent with the experimental observables. Conversely, 1LVJ-like conformations are scarcely populated, suggesting a binding process dominated by induced-fit rearrangements.

To systematize this comparison, we matched each crystallographic pocket to the nearest cluster differing by no more than 5 kcal mol⁻¹ in free energy from the global minimum and inspect the main differences that characterize, prevent or allow ligand binding. In this respect, We projected these pocket-bearing references onto the principal-component space derived from the PT-WTE ensemble (Figure D.4). This

projection enables a direct assessment of whether experimentally observed pocket geometries lie, thereby informing the relative contributions of conformational selection versus induced fit in TAR.

Among the six experimental cases examined, three (1QD3, 2L8H, 1UUD) showed substantial overlap between PT-WTE conformations and the ligand-bound pockets resolved in the PDB, indicating that pocket geometries compatible with binding are sampled even in the absence of ligands.

For 1UTS, the agreement was partial: crystallographic conformations fall in low density regions of the simulated distribution, suggesting that the relevant pocket is accessible but underpopulated in the unbound ensemble. By contrast, 1LVJ and 1UUI exhibited limited overlap with the sampled space, consistent with a larger contribution of ligand-induced stabilization and/or selection of rare substates.

An illustrative example is 1QD3 (Figure 6.10A). The nearest unbound representative falls within the same basin as the crystal structure in the pocket PCA space, with a crystal–centroid ϵ RMSD of 0.72. The overall geometry matches the bound state closely, with minor deviations localized to U23–U24 at the solvent-exposed entrance of the groove. These observations are consistent with binding to a pre-existing conformation followed by a small induced-fit adjustment.

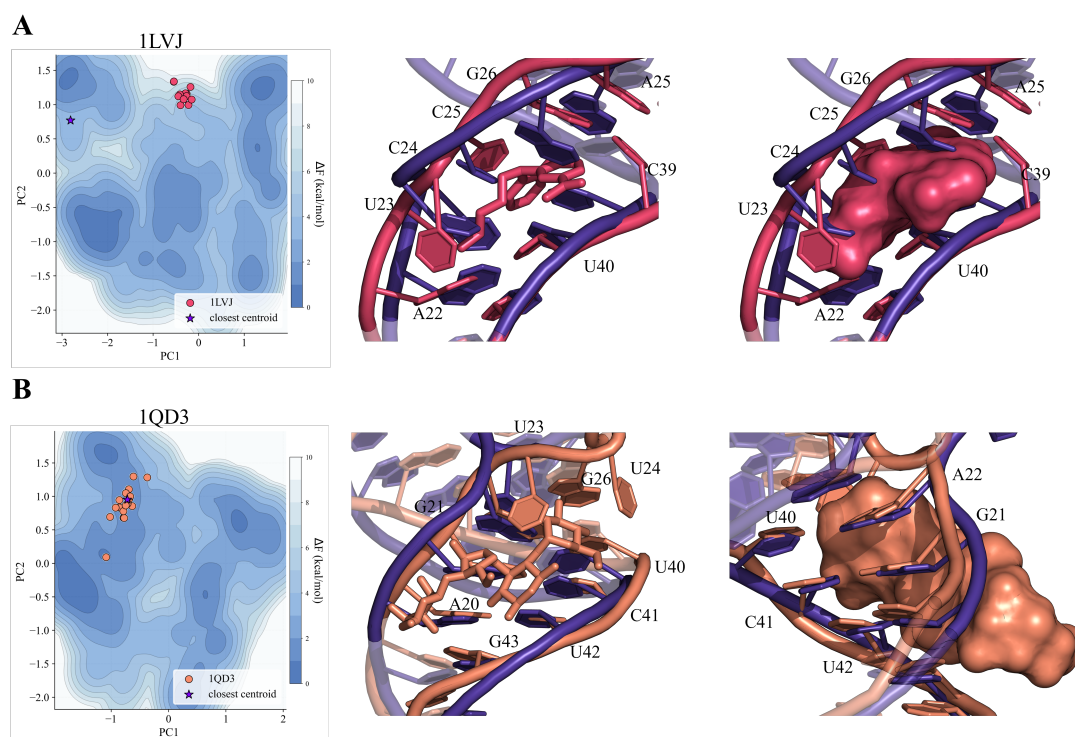


Figure 6.10. (A) Two-dimensional projection of the conformational free-energy landscape sampled around the binding mode observed in the 1LVJ structure, shown along the first two principal components (PC1 and PC2). Colored circles correspond to the crystallographic poses of the ligand in 1LVJ, and the purple star indicates the centroid of the closest simulated cluster. In the center, the three-dimensional comparison between the crystallographic binding mode and the closest simulated conformation is shown. The ligand is represented in stick form, and HIV-1 TAR is displayed as a cartoon and stick representation. Residue numbers are indicated. On the right, the same view and color scheme are used, but the ligand is represented as a molecular surface to illustrate the occupied binding cavity. (B) Two-dimensional projection of the conformational free-energy landscape sampled around the binding mode observed in the 1QD3 structure, shown along PC1 and PC2. Colored circles correspond to the crystallographic poses of the ligand in 1QD3, and the purple star indicates the centroid of the closest simulated cluster. In the center, the three-dimensional comparison between the crystallographic binding mode and the closest simulated conformation is shown. The ligand is displayed in stick representation from the front view, and HIV-1 TAR is shown as a cartoon and stick representation. On the right, the ligand is represented as a molecular surface after a 180° rotation to highlight shape complementarity.

A similar scenario is observed for 1UUD and 1UUI (Figure D.5C and Figure D.6A). In both cases, the ligands occupy solvent-exposed regions where modest changes in base pairing/stacking do not substantially alter the interacting surface. The principal rearrangement involves U23: in the unbound ensemble its orientation occludes key contacts, whereas a local reorientation is sufficient to permit proper ligand recognition.

For 1UTS, the closest representative lies at an ε RMSD 0.77 from the crys-

tallographic structure. Here the ligand binds in the bulge, and two nucleotides, U23 and U40, must reposition to create space and establish the correct interactions (Figure D.5C). The G10–C23 base pair is preserved but slightly reorients, enabling a π – π contact between the purine and the indole group present in the ligand.

By contrast, 2L8H features a pocket within the apical loop, where intrinsic flexibility is enhanced. In the closest unbound representative, G36 partially blocks the cavity and must shift by ~ 3.5 Å to accommodate the probe. A35 is extrahelical, leaving access for entry but requiring reinsertion to form π – π stacking with the aromatic system; similarly, G34 must transition from extrahelical to intrahelical to further stabilize the complex via aromatic interactions. Moreover, we observe a broad detachment of the nucleic backbone at the arginine contacting segment, suggesting an induced-fit mechanism during ligand binding (Figure D.6C).

Finally, 1LVJ exemplifies an induced-fit–dominated mechanism. The ligand wedges between the two helical arms, disrupting or reshaping several pairing and stacking interactions. The nearest non-negligible representative is found at an ϵ RMSD of 1.11 from the crystal. Notably, the G26–C13 base pair is retained but its relative orientation would clash sterically with the ligand; C24 must lose stacking with U25 and rotate to create room, while the U23–U40 base pair must be disrupted to permit binding (Figure 6.10).

6.2.4 Summary of main findings

- The well-tempered ensemble (PT-WTE) protocol enabled extensive exploration of the rugged conformational landscape of HIV-1 TAR, revealing a dynamic equilibrium of interconverting substates with broad variability across global and local metrics.
- Maximum-entropy reweighting against RDCs improved agreement with experiment while preserving ensemble diversity, primarily by redistributing populations within already sampled regions and enriching substates consistent with the data.
- Low-dimensional projections showed that sampling covers the regions spanned by available PDB structures, both unbound and ligand-bound. Comparisons

with known complexes support a predominant role of conformational selection with occasional induced-fit adjustments.

- Mapping per-frame druggability scores onto the low-dimensional landscape revealed a separation between highly druggable and poorly druggable regions. Reweighting decreased the weight of low-scoring areas and maintained or enriched clusters with medium-to-high druggability that are most relevant for ligand engagement.

6.3 Conclusions

HIV-1 TAR is a challenging benchmark for integrative strategies that combine enhanced sampling with experimental restraints. Here we show that well-tempered ensemble (PT-WTE) sampling, followed by maximum-entropy refinement against RDCs, reconciles broad conformational exploration with quantitative agreement to experiment. The refined ensembles retain heterogeneity and recover bound-like substates at increased populations, a prerequisite for a realistic description of molecular recognition.

Analysis in the $\mathbf{G}$ -vector space and associated clustering shows that conformations compatible with ligand binding are already accessible in the apo ensemble; reweighting acts primarily by redistributing probabilities across pre-existing substates. Case-by-case comparisons with crystallographic complexes span a continuum from conformational selection-dominated scenarios (1QD3, 1UTS) to an induced-fit-dominated case (1LVJ), which requires larger local rearrangements in pairing and stacking. These findings rationalize why small-molecule binding to TAR can proceed by stabilizing rare pre-existing conformations, while some complexes demand additional ligand-induced stabilization.

This study also highlights limitations and avenues for improvement. The effective sample size after reweighting remains modest, reflecting the difficulty of capturing slow collective motions and local rearrangements. Future work could leverage longer PT-WTE replicas, additional orthogonal restraints (e.g., SAXS, NOEs), and pocket-specific observables, refining the balance between selection and induction and prioritizing druggable substates for prospective screening on TAR and related RNAs. More-

over, running simulations of comparable length from multiple crystallographic/NMR starting structures may better characterize the dynamic landscape and reduce dependence on the initial reference.

From a drug-discovery perspective, we identified pocket-competent substates not yet exploited in design that warrant prioritization for prospective screening. In particular, compact or mildly twisted conformations represent targetable cavities for ensemble docking and fragment growth, whereas diffuse, overextended conformations are deprioritized. More broadly, this integrative workflow offers a general route to expose experimentally supported, pocket-competent RNA substates for structure-based design.

7 Conclusions and Outlook

This thesis set out to advance RNA targeted small molecule discovery through computational means, developing a computational framework that progresses from comparative case studies to a focused application on a pharmaceutically relevant RNA. The guiding premise is that flexible RNAs do not present a single lock and key structure; rather, they populate heterogeneous ensembles in which ligandable states are transient and state specific, so discovery must reason on populations of substates and on their redistribution upon binding. In Chapter 4 we addressed a simple yet revealing question: if the ligand is held fixed, how do a protein pocket and an RNA pocket differ in what they require for recognition. Using riboflavin as the common ligand across riboflavin kinase and the FMN riboswitch, we showed that much of the modern protein oriented simulation toolkit transfers to suitably folded RNA sites, including conformational analyses, electrostatics, water kinetics, and ligand dynamics. Despite a reduced chemical alphabet, RNA pockets can sustain rich polar and noncovalent interaction patterns and can exploit structured waters in ways that are directly analyzable with standard workflows. At the same time, the study highlighted domain specific caveats, namely pronounced conformational heterogeneity, soft pocket boundaries, and a dominant role of electrostatics in shaping effective complementarity. Importantly, these conclusions are system dependent, because the selected riboswitch represents a favorable case with a well formed site, and they therefore motivate ensemble aware strategies for more flexible RNAs. Although the present work characterizes a specific and well studied system, a broader and systematically curated benchmark would enable more general inferences and the formulation of transferable principles that translate established protein based discovery concepts to RNA and other flexible targets, including intrinsically disordered proteins, in order to delineate shared rules of ligand recognition for flexible biomolecular ensembles. Chapter 5 advanced an explicitly ensemble centric perspective. We introduced an exploration and prioritization workflow that couples SHadow Mixed solvent metadynamics with cheminformatics and docking. Mixed solvent maps delineated recurrent hotspots in HIV-1 TAR, most prominently around the apical loop and the bulge, while BiteNet refined these maps into substate level hypotheses. Using this pipeline, we identified and selected two hundred candidate inhibitors of the interaction between HIV-1 TAR and the viral

integrase. Cheminformatic analysis confirmed scaffold heterogeneity, which is desirable at the hit identification stage in a combined virtual and experimental screening campaign. These compounds are undergoing biological evaluation that will provide quantitative feedback to validate the procedure and to guide, in a data driven manner, the further development of new inhibitors. In the first stage of the project we deliberately restricted the analysis to a specific molecular library. At the same time, many commercial vendors now provide RNA focused libraries, which makes it feasible to broaden the prospective screen and to uncover additional chemotypes and alternative binding hypotheses across diverse RNA topologies. As experimental results become available, the campaign can transition to a more strongly data driven strategy in which assay readouts refine hypotheses, retrain prioritization models, and iteratively enrich for active series. In Chapter 6 we addressed a central bottleneck, namely the construction of an experimentally compliant ensemble for HIV-1 TAR that is suitable for design. Parallel Tempering in the Well Tempered Ensemble afforded broad exploration, yet the unrefined molecular dynamics ensemble did not reproduce solution data quantitatively. Maximum Entropy reweighting against residual dipolar couplings reconciled simulation with experiment, reducing the root mean square error from 8.4 to 2.8 Hz while preserving the essential structural heterogeneity of this small RNA. Independent validation with SHAPE profiles supported improved prediction of local flexibility. Pocket level comparisons to Protein Data Bank complexes revealed a continuum from selection dominated cases to induced fit dominated binding, rationalizing why certain ligands stabilize rare pre existing conformations whereas others require additional local rearrangements in pairing and stacking. Three broader insights emerge across projects. First, ensemble awareness is not an optional garnish for RNA design, it is the organizing principle. Whether contrasting protein and RNA targets, mapping hotspots, or integrating experimental data, the relevant units of inference are substates and their populations rather than a single average structure. Second, Maximum Entropy reweighting provides a practical bridge between physics based sampling and solution measurements, it improves quantitative agreement, sharpens inter helical geometry, and remains interpretable for medicinal chemistry decision making. Third, pocket quality depends on the substate, with clear separations between compact or mildly twisted basins that score highly and diffuse overextended regions that do not. Since druggability predictors are assumption laden, orthogonal validation with com-

plementary tools, together with expert judgment, is essential. These caveats do not undercut the conclusions; rather, they indicate where to invest effort in the next steps of the research. Taken together, these results advance a reproducible and data integrated strategy for discovery on RNA. By turning RNA conformational plasticity from an obstacle into an organizing principle, the framework developed here provides a general foundation for rational, structure based design on flexible RNA targets.

References

1. Cech, T. R. & Steitz, J. A. The noncoding RNA revolution—trashing old rules to forge new ones. *Cell* **157**, 77–94 (2014).
2. Sharp, P. A. The centrality of RNA. *Cell* **136**, 577–580 (2009).
3. Crick, F. Central dogma of molecular biology. *Nature* **227**, 561–563 (1970).
4. Childs-Disney, J. L. *et al.* Targeting RNA structures with small molecules. *Nature Reviews Drug Discovery* **21**, 736–762 (2022).
5. Wang, S., Weissman, D. & Dong, Y. RNA chemistry and therapeutics. *Nature Reviews Drug Discovery*, 1–24 (2025).
6. Bolland, C. R. Non-coding RNA: it's not junk. *Digestive diseases and sciences* **62**, 1107 (2017).
7. Fire, A. *et al.* Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *nature* **391**, 806–811 (1998).
8. Lee, R. C., Feinbaum, R. L. & Ambros, V. The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. *cell* **75**, 843–854 (1993).
9. Ma, L., Bajic, V. B. & Zhang, Z. On the classification of long non-coding RNAs. *RNA biology* **10**, 924–933 (2013).
10. Storz, G. An RNA thermometer. *Genes & development* **13**, 633–636 (1999).
11. Gilbert, W. Origin of life: The RNA world. *nature* **319**, 618–618 (1986).
12. Tanner, N. K. Ribozymes: the characteristics and properties of catalytic RNAs. *FEMS microbiology reviews* **23**, 257–275 (1999).
13. Kim, V. N. Small RNAs: classification, biogenesis, and function. *Molecules and cells* **19**, 1–15 (2005).
14. Hombach, S. & Kretz, M. Non-coding RNAs: classification, biology and functioning. *Non-coding RNAs in colorectal cancer*, 3–17 (2016).
15. Yao, R.-W., Wang, Y. & Chen, L.-L. Cellular functions of long noncoding RNAs. *Nature cell biology* **21**, 542–551 (2019).
16. Chen, L.-L. & Kim, V. N. Small and long non-coding RNAs: Past, present, and future. *Cell* **187**, 6451–6485 (2024).
17. Rinn, J. L. & Chang, H. Y. Genome regulation by long noncoding RNAs. *Annual review of biochemistry* **81**, 145–166 (2012).
18. Li, W., Notani, D. & Rosenfeld, M. G. Enhancers as non-coding RNA transcription units: recent insights and future perspectives. *Nature Reviews Genetics* **17**, 207–223. doi:10.1038/nrg.2016.4 (2016).
19. Sartorelli, V. & Lauberth, S. M. Enhancer RNAs are an important regulatory layer of the epigenome. *Nature structural & molecular biology* **27**, 521–528 (2020).
20. Dawoud, A. *et al.* Circular RNAs: New layer of complexity evading breast cancer heterogeneity. *Non-coding RNA research* **8**, 60–74 (2023).
21. Hsu, P. D., Lander, E. S. & Zhang, F. Development and Applications of CRISPR-Cas9 for Genome Engineering. *Cell* **157**, 1262–1278. doi:10.1016/j.cell.2014.05.010 (2014).
22. Lönnberg, T. A. Can Phosphate-Branched RNA Persist under Physiological Conditions? *ChemBioChem* **13**, 2690–2700 (2012).
23. Sponer, J. *et al.* RNA structural dynamics as captured by molecular simulations: a comprehensive overview. *Chemical reviews* **118**, 4177–4338 (2018).
24. Tinoco Jr, I. & Bustamante, C. How RNA folds. *Journal of molecular biology* **293**, 271–281 (1999).

REFERENCES

25. Brion, P. & Westhof, E. Hierarchy and dynamics of RNA folding. *Annual review of biophysics and biomolecular structure* **26**, 113–137 (1997).
26. Holbrook, S. R. Structural principles from large RNAs. *Annu. Rev. Biophys.* **37**, 445–464 (2008).
27. Lilley, D. M. Structures of helical junctions in nucleic acids. *Quarterly reviews of biophysics* **33**, 109–159 (2000).
28. Neidle, S. & Sanderson, M. *Principles of nucleic acid structure* (Academic Press, 2021).
29. Ruszkowska, A. *et al.* Molecular structure of a U• AU-rich RNA triple helix with 11 consecutive base triples. *Nucleic acids research* **48**, 3304–3314 (2020).
30. Batey, R. T., Rambo, R. P. & Doudna, J. A. Tertiary motifs in RNA structure and folding. *Angewandte Chemie International Edition* **38**, 2326–2343 (1999).
31. Butcher, S. E. & Pyle, A. M. The molecular interactions that stabilize RNA tertiary structure: RNA motifs, patterns, and networks. *Accounts of chemical research* **44**, 1302–1311 (2011).
32. Staple, D. W. & Butcher, S. E. Pseudoknots: RNA structures with diverse functions. *PLoS biology* **3**, e213 (2005).
33. Warner, K. D., Hajdin, C. E. & Weeks, K. M. Principles for targeting RNA with drug-like small molecules. *Nature reviews Drug discovery* **17**, 547–558 (2018).
34. Bernetti, M. *et al.* Computational drug discovery under RNA times. *QRB discovery* **3**, e22 (2022).
35. Damase, T. R. *et al.* The limitless future of RNA therapeutics. *Frontiers in bioengineering and biotechnology* **9**, 628137 (2021).
36. Wang, Y. *et al.* End-to-end differentiable construction of molecular mechanics force fields. *Chemical Science* **13**, 12016–12033 (2022).
37. DeVos, S. L. & Miller, T. M. Antisense oligonucleotides: treating neurodegeneration at the level of RNA. *Neurotherapeutics* **10**, 486–497 (2013).
38. Hoy, S. M. Nusinersen: first global approval. *Drugs* **77**, 473–479 (2017).
39. Neil, E. E. & Bisaccia, E. K. Nusinersen: a novel antisense oligonucleotide for the treatment of spinal muscular atrophy. *The Journal of Pediatric Pharmacology and Therapeutics* **24**, 194–203 (2019).
40. Lim, K. R. Q., Maruyama, R. & Yokota, T. Eteplirsen in the treatment of Duchenne muscular dystrophy. *Drug design, development and therapy*, 533–545 (2017).
41. Syed, Y. Y. Eteplirsen: first global approval. *Drugs* **76**, 1699–1704 (2016).
42. Adams, D. *et al.* Patisiran, an RNAi Therapeutic, for Hereditary Transthyretin Amyloidosis. *New England Journal of Medicine* **379**, 11–21. doi:10.1056/NEJMoa1716153 (2018).
43. Ledford, H. Gene-silencing technology gets first drug approval after 20-year wait. *Nature* **560**, 291–293 (2018).
44. Balwani, M. *et al.* Phase 3 Trial of Givosiran for Acute Intermittent Porphyria. *New England Journal of Medicine* **382**, 2289–2301. doi:10.1056/NEJMoa1913147 (2020).
45. Garrelfs, S. F. *et al.* Lumasiran, an RNAi Therapeutic for Primary Hyperoxaluria Type 1. *New England Journal of Medicine* **384**, 1216–1226. doi:10.1056/NEJMoa2021712 (2021).
46. Seyhan, A. A. Trials and tribulations of MicroRNA therapeutics. *International journal of molecular sciences* **25**, 1469 (2024).
47. Polack, F. P. *et al.* Safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine. *New England Journal of Medicine* **383**, 2603–2615. doi:10.1056/NEJMoa2034577 (2020).
48. Baden, L. R. *et al.* Efficacy and Safety of the mRNA-1273 SARS-CoV-2 Vaccine. *New England Journal of Medicine* **384**, 403–416. doi:10.1056/NEJMoa2035389 (2021).

49. Shi, Y. *et al.* Progress and prospects of mRNA-based drugs in pre-clinical and clinical applications. *Signal Transduction and Targeted Therapy* **9**, 322 (2024).
50. Manigrasso, J., Marcia, M. & De Vivo, M. Computer-aided design of RNA-targeted small molecules: A growing need in drug discovery. *Chem* **7**, 2965–2988 (2021).
51. Guan, L. & Disney, M. D. Recent advances in developing small molecules targeting RNA. *ACS chemical biology* **7**, 73–86 (2012).
52. Connelly, C. M. *et al.* Synthetic ligands for PreQ1 riboswitches provide structural and mechanistic insights into targeting RNA tertiary structure. *Nature communications* **10**, 1501 (2019).
53. Fedorova, O. *et al.* Small molecules that target group II introns are potent antifungal agents. *Nature chemical biology* **14**, 1073–1078 (2018).
54. Ratni, H. *et al.* Discovery of Risdiplam, a Selective Survival of Motor Neuron-2 (SMN2) Gene Splicing Modifier for the Treatment of Spinal Muscular Atrophy (SMA). *Journal of Medicinal Chemistry* **61**, 6501–6517. doi:10.1021/acs.jmedchem.8b00741 (2018).
55. Ratni, H., Scalco, R. S. & Stephan, A. H. Risdiplam, the first approved small molecule splicing modifier drug as a blueprint for future transformative medicines. *ACS medicinal chemistry letters* **12**, 874–877 (2021).
56. Paik, J. Risdiplam: a review in spinal muscular atrophy. *CNS drugs* **36**, 401–410 (2022).
57. Sheridan, C. *First small-molecule drug targeting RNA gains momentum* 2021.
58. Jorgensen, W. L. The many roles of computation in drug discovery. *Science* **303**, 1813–1818 (2004).
59. Macalino, S. J. Y. *et al.* Role of computer-aided drug design in modern drug discovery. *Archives of pharmacol research* **38**, 1686–1701 (2015).
60. Sadybekov, A. V. & Katritch, V. Computational approaches streamlining drug discovery. *Nature* **616**, 673–685 (2023).
61. De Vivo, M. *et al.* Role of molecular dynamics and related methods in drug discovery. *Journal of medicinal chemistry* **59**, 4035–4061 (2016).
62. Decherchi, S. & Cavalli, A. Thermodynamics and kinetics of drug-target binding by molecular simulation. *Chemical Reviews* **120**, 12788–12833 (2020).
63. Vamathevan, J. *et al.* Applications of machine learning in drug discovery and development. *Nature reviews Drug discovery* **18**, 463–477 (2019).
64. Ghidini, A. *et al.* Bidirectional path-based non-equilibrium simulations for binding free energy. *Molecular Physics* **123**, e2374465 (2025).
65. Juru, A. U. & Hargrove, A. E. Frameworks for targeting RNA with small molecules. *Journal of Biological Chemistry* **296**, 100191 (2021).
66. Morgan, B. S. *et al.* R-BIND: an interactive database for exploring and developing RNA-targeted chemical probes. *ACS chemical biology* **14**, 2691–2700 (2019).
67. Ganser, L. R. *et al.* The roles of structural dynamics in the cellular functions of RNAs. *Nature reviews Molecular cell biology* **20**, 474–489 (2019).
68. Stelzer, A. C. *et al.* Discovery of selective bioactive small molecules by targeting an RNA dynamic ensemble. *Nature chemical biology* **7**, 553–559 (2011).
69. Chopra, N. *et al.* Enabling In-Silico Hit Discovery Workflows Targeting RNA with Small Molecules. *Journal of Chemical Information and Modeling* **65**, 7393–7398 (2025).
70. Bottaro, S. *et al.* Conformational ensembles of RNA oligonucleotides from integrating NMR and molecular simulations. *Science advances* **4**, eaar8521 (2018).

REFERENCES

71. Feng, Y. *et al.* NLDock: A fast nucleic acid–ligand docking algorithm for modeling RNA/DNA–ligand complexes. *Journal of Chemical Information and Modeling* **61**, 4771–4782 (2021).
72. Chhabra, S., Xie, J. & Frank, A. T. RNAPosers: machine learning classifiers for ribonucleic acid–ligand poses. *The Journal of Physical Chemistry B* **124**, 4436–4445 (2020).
73. Agarwal, R., T, R. R. & Smith, J. C. Comparative assessment of pose prediction accuracy in rna–ligand docking. *Journal of Chemical Information and Modeling* **63**, 7444–7452 (2023).
74. Jiang, D. *et al.* How good are current docking programs at nucleic acid–ligand docking? a comprehensive evaluation. *Journal of Chemical Theory and Computation* **19**, 5633–5647 (2023).
75. Carvajal-Patiño, J. G. *et al.* RNAmigos2: accelerated structure-based RNA virtual screening with deep graph learning. *Nature Communications* **16**, 1–12 (2025).
76. Kozlovskii, I. & Popov, P. Structure-based deep learning for binding site detection in nucleic acid macromolecules. *NAR Genomics and Bioinformatics* **3**, lqab111 (2021).
77. Wei, H. & McCammon, J. A. Structure and dynamics in drug discovery. *npj Drug Discovery* **1**, 1 (2024).
78. Ahmed, M., Maldonado, A. M. & Durrant, J. D. From byte to bench to bedside: molecular dynamics simulations and drug discovery. *BMC biology* **21**, 299 (2023).
79. Frenkel, D. & Smit, B. *Understanding molecular simulation* (Elsevier, 1957).
80. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *nature* **596**, 583–589 (2021).
81. Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).
82. Passaro, S. *et al.* Boltz-2: Towards accurate and efficient binding affinity prediction. *BioRxiv*, 2025–06 (2025).
83. Wohlwend, J. *et al.* Boltz-1 democratizing biomolecular interaction modeling. *BioRxiv* (2024).
84. Bernetti, M. & Bussi, G. Integrating experimental data with molecular simulations to investigate RNA structural dynamics. *Current Opinion in Structural Biology* **78**, 102503 (2023).
85. Bernetti, M. *et al.* Probing allosteric communication with combined molecular dynamics simulations and network analysis. *Current Opinion in Structural Biology* **86**, 102820 (2024).
86. Bernetti, M. *et al.* An integrated Markov state model and path metadynamics approach to characterize drug binding processes. *Journal of chemical theory and computation* **15**, 5689–5702 (2019).
87. Bellucci, L. *et al.* The interaction with gold suppresses fiber-like conformations of the amyloid β (16–22) peptide. *Nanoscale* **8**, 8737–8748 (2016).
88. Borthakur, K. *et al.* Determining accurate conformational ensembles of intrinsically disordered proteins at atomic resolution. *Nature Communications* **16**, 9036 (2025).
89. Wang, J., Cieplak, P. & Kollman, P. A. How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *Journal of computational chemistry* **21**, 1049–1074 (2000).
90. Cornell, W. D. *et al.* A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *Journal of the American Chemical Society* **117**, 5179–5197 (1995).
91. Cheatham III, T. E., Cieplak, P. & Kollman, P. A. A modified version of the Cornell *et al.* force field with improved sugar pucker phases and helical repeat. *Journal of Biomolecular Structure and Dynamics* **16**, 845–862 (1999).
92. Pérez, A. *et al.* Refinement of the AMBER Force Field for Nucleic Acids: Improving the Description of α/γ Conformers. *Biophysical Journal* **92**, 3817–3829. doi:10.1529/biophysj.106.097782 (2007).

-
93. Yildirim, I. *et al.* Reparameterization of RNA χ Torsion Parameters for the AMBER Force Field and Comparison to NMR Spectra for Cytidine and Uridine. *Journal of Chemical Theory and Computation* **6**, 1520–1531. doi:10.1021/ct900604a (2010).
94. Zgarbová, M. *et al.* Refinement of the Cornell *et al.* Nucleic Acids Force Field Based on Reference Quantum Chemical Calculations of Glycosidic Torsion Profiles. *Journal of Chemical Theory and Computation* **7**, 2886–2902. doi:10.1021/ct200162x (2011).
95. Yildirim, I. *et al.* Revision of AMBER torsional parameters for RNA improves free energy predictions for tetramer duplexes with GC and iGiC base pairs. *Journal of chemical theory and computation* **8**, 172–181 (2012).
96. Bergonzo, C. *et al.* Highly Sampled Tetranucleotide and Tetraloop Motifs Enable Evaluation of Common RNA Force Fields. *RNA* **21**, 1578–1590. doi:10.1261/rna.051888.115 (2015).
97. Kuhrova, P. *et al.* Improving the Performance of the AMBER RNA Force Field by Tuning the Hydrogen-Bonding Interactions. *Journal of Chemical Theory and Computation* **15**, 3288–3305. doi:10.1021/acs.jctc.9b00160 (2019).
98. Fröhling, T. *et al.* Toward Empirical RNA Force Fields That Match Experimental Data. *Journal of Chemical Theory and Computation* **18**, 5716–5730. doi:10.1021/acs.jctc.2c00258 (2022).
99. Chen, H.-F. & Garcia, A. E. High-Resolution RNA Force Fields with CMAP Corrections. *Journal of Chemical Theory and Computation* **18**, 3653–3666. doi:10.1021/acs.jctc.1c01208 (2022).
100. Mlynsky, V. *et al.* Simple adjustment of intranucleotide base-phosphate interaction in the OL3 AMBER force field improves RNA simulations. *Journal of chemical theory and computation* **19**, 8423–8433 (2023).
101. Raguet, L. E. *et al.* Adjusting the energy profile for CH–O interactions leads to improved stability of RNA stem-loop structures in MD simulations. *The Journal of Physical Chemistry B* **128**, 7921–7933 (2024).
102. Maier, J. A. *et al.* ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB. *Journal of chemical theory and computation* **11**, 3696–3713 (2015).
103. Tan, D. *et al.* RNA Force Field with Accuracy Comparable to State-of-the-Art Protein Force Fields. *Proceedings of the National Academy of Sciences* **115**, E1346–E1355. doi:10.1073/pnas.1713027115 (2018).
104. MacKerell Jr, A. D., Wiorkiewicz-Kuczera, J. & Karplus, M. An all-atom empirical energy function for the simulation of nucleic acids. *Journal of the American Chemical Society* **117**, 11946–11975 (1995).
105. MacKerell Jr, A. D., Banavali, N. & Foloppe, N. Development and current status of the CHARMM force field for nucleic acids. *Biopolymers: original Research on biomolecules* **56**, 257–265 (2000).
106. Denning, E. J. *et al.* Impact of 2'-Hydroxyl Sampling on the Conformational Properties of RNA: Update of the CHARMM All-Atom Additive Force Field for RNA. *Journal of Computational Chemistry* **32**, 1929–1943. doi:10.1002/jcc.21777 (2011).
107. Huang, J. *et al.* CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nature methods* **14**, 71–73 (2017).
108. Lemkul, J. A. & MacKerell Jr, A. D. Polarizable force field for RNA based on the classical drude oscillator. *Journal of computational chemistry* **39**, 2624–2646 (2018).
109. Torrie, G. M. & Valleau, J. P. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *Journal of computational physics* **23**, 187–199 (1977).
110. Laio, A. & Parrinello, M. Escaping free-energy minima. *Proceedings of the national academy of sciences* **99**, 12562–12566 (2002).
111. Barducci, A., Bussi, G. & Parrinello, M. Well-tempered metadynamics: a smoothly converging and tunable free-energy method. *Physical review letters* **100**, 020603 (2008).
112. Sugita, Y. & Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chemical physics letters* **314**, 141–151 (1999).

REFERENCES

113. Wang, L., Friesner, R. A. & Berne, B. Replica exchange with solute scaling: a more efficient version of replica exchange with solute tempering (REST2). *The Journal of Physical Chemistry B* **115**, 9431–9438 (2011).
114. Bonomi, M., Barducci, A. & Parrinello, M. Reconstructing the equilibrium Boltzmann distribution from well-tempered metadynamics. *Journal of computational chemistry* **30**, 1615–1621 (2009).
115. Hénin, J. *et al.* Enhanced Sampling Methods for Molecular Dynamics Simulations. *Living Journal of Computational Molecular Science* **4**, 1583 (2022).
116. Hoff, S. E. *et al.* Accurate model and ensemble refinement using cryo-electron microscopy maps and Bayesian inference. *PLOS Computational Biology* **20**, e1012180 (2024).
117. Bernetti, M., Hall, K. B. & Bussi, G. Reweighting of molecular simulations with explicit-solvent SAXS restraints elucidates ion-dependent RNA ensembles. *Nucleic acids research* **49**, e84–e84 (2021).
118. Bresciani, V. *et al.* Integrating computational and experimental biophysics reveals novel insights into the RAD51-BRC4 interaction. *bioRxiv*, 2025–03 (2025).
119. Seco, J., Luque, F. J. & Barril, X. Binding site detection and druggability index from first principles. *Journal of medicinal chemistry* **52**, 2363–2371 (2009).
120. Ghanakota, P. & Carlson, H. A. Driving structure-based drug discovery through cosolvent molecular dynamics: Miniperspective. *Journal of medicinal chemistry* **59**, 10383–10399 (2016).
121. Bemelmans, M. P. *et al.* Computational advances in discovering cryptic pockets for drug discovery. *Current Opinion in Structural Biology* **90**, 102975 (2025).
122. Kuzmanic, A. *et al.* Investigating cryptic binding sites by molecular dynamics simulations. *Accounts of chemical research* **53**, 654–661 (2020).
123. Yu, W. *et al.* Pharmacophore modeling using site-identification by ligand competitive saturation (SILCS) with multiple probe molecules. *Journal of chemical information and modeling* **55**, 407–420 (2015).
124. Lal Gupta, P. & Carlson, H. A. Cosolvent Simulations with Fragment-Bound Proteins Identify Hot Spots to Direct Lead Growth. *Journal of chemical theory and computation* **18**, 3829–3844 (2022).
125. Lee, J. Y. *et al.* Pharmmaker: Pharmacophore modeling and hit identification based on druggability simulations. *Protein Science* **29**, 76–86 (2020).
126. Goodford, P. J. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *Journal of medicinal chemistry* **28**, 849–857 (1985).
127. Artese, A. *et al.* Molecular interaction fields in drug discovery: recent advances and future perspectives. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **3**, 594–613 (2013).
128. Greer, J. *et al.* Application of the three-dimensional structures of protein target molecules in structure-based drug design. *Journal of medicinal chemistry* **37**, 1035–1054 (1994).
129. Cruciani, G. *et al.* MetaSite: understanding metabolism in human cytochromes from the perspective of the chemist. *Journal of medicinal chemistry* **48**, 6970–6979 (2005).
130. Storch, L., Cruciani, G. & Cross, S. DeepGRID: Deep Learning Using GRID Descriptors for BBB Prediction. *Journal of chemical information and modeling* **63**, 5496–5512 (2023).
131. Feixas, F. *et al.* Exploring the role of receptor flexibility in structure-based drug discovery. *Biophysical chemistry* **186**, 31–45 (2014).
132. Miranker, A. & Karplus, M. Functionality maps of binding sites: a multiple copy simultaneous search method. *Proteins: Structure, Function, and Bioinformatics* **11**, 29–34 (1991).
133. Elber, R. & Karplus, M. Enhanced sampling in molecular dynamics: use of the time-dependent Hartree approximation for a simulation of carbon monoxide diffusion through myoglobin. *Journal of the American Chemical Society* **112**, 9161–9175 (1990).

-
134. Stultz, C. M. & Karplus, M. MCSS functionality maps for a flexible protein. *Proteins: Structure, Function, and Bioinformatics* **37**, 512–529 (1999).
135. Panei, F., Gkeka, P. & Bonomi, M. Identifying small-molecules binding sites in RNA conformational ensembles with SHAMAN. *Nature Communications* **15**, 5725 (2024).
136. Panei, F. P. *Advanced computational techniques to aid the rational design of small molecules targeting RNA* PhD thesis (Sorbonne Université, 2024).
137. Bonomi, M. *et al.* PLUMED: A portable plugin for free-energy calculations with molecular dynamics. *Computer Physics Communications* **180**, 1961–1972 (2009).
138. Tribello, G. A. *et al.* PLUMED 2: New feathers for an old bird. *Computer physics communications* **185**, 604–613 (2014).
139. Tribello, G. A. *et al.* PLUMED Tutorials: A collaborative, community-driven learning ecosystem. *The Journal of Chemical Physics* **162** (2025).
140. Meng, X. *et al.* Molecular docking: a powerful approach for structure-based drug discovery. *Current Computer-Aided Drug Design* **7**, 146–157 (2011).
141. Paggi, J. M., Pandit, A. & Dror, R. O. The art and science of molecular docking. *Annual review of biochemistry* **93**, 389–410 (2024).
142. Agu, P. C. *et al.* Molecular docking as a tool for the discovery of molecular targets of nutraceuticals in diseases management. *Scientific reports* **13**, 13398 (2023).
143. Jiménez, J. *et al.* KDEEP: protein–ligand absolute binding affinity prediction via 3D-convolutional neural networks. *Journal of Chemical Information and Modeling* **58**, 287–296 (2018).
144. Vittorio, S. *et al.* Addressing docking pose selection with structure-based deep learning: recent advances, challenges and opportunities. *Computational and Structural Biotechnology Journal* **23**, 2141–2151 (2024).
145. Singh, N. & Villoutreix, B. O. A hybrid docking and machine learning approach to enhance the performance of virtual screening carried out on protein–protein interfaces. *International Journal of Molecular Sciences* **23**, 14364 (2022).
146. Halgren, T. A. *et al.* Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *Journal of Medicinal Chemistry* **47**, 1750–1759 (2004).
147. Oliver, C. *et al.* Augmented base pairing networks encode RNA-small molecule binding preferences. *Nucleic acids research* **48**, 7690–7699 (2020).
148. Leontis, N. B. & Westhof, E. Conserved geometrical base-pairing patterns in RNA. *Quarterly reviews of biophysics* **31**, 399–455 (1998).
149. Glielmo, A. *et al.* Unsupervised learning methods for molecular simulation data. *Chemical Reviews* **121**, 9722–9758 (2021).
150. Ceriotti, M. Unsupervised machine learning in atomistic simulations, between predictions and understanding. *The Journal of chemical physics* **150** (2019).
151. Daura, X. *et al.* Peptide folding: when simulation meets experiment. *Angewandte Chemie International Edition* **38**, 236–240 (1999).
152. Heyer, L. J., Kruglyak, S. & Yooseph, S. Exploring expression data: identification and analysis of coexpressed genes. *Genome research* **9**, 1106–1115 (1999).
153. González-Alemán, R. *et al.* Quality threshold clustering of molecular dynamics: a word of caution. *Journal of chemical information and modeling* **60**, 467–472 (2019).
154. Leach, K. L. *et al.* Linezolid, the first oxazolidinone antibacterial agent. *Annals of the New York Academy of Sciences* **1222**, 49–54 (2011).

REFERENCES

155. Yu, A.-M., Choi, Y. H. & Tu, M.-J. RNA drugs and RNA targets for small molecules: principles, progress, and challenges. *Pharmacological reviews* **72**, 862–898 (2020).
156. Connelly, C. M., Moon, M. H. & Schneekloth, J. S. The emerging role of RNA as a therapeutic target for small molecules. *Cell chemical biology* **23**, 1077–1090 (2016).
157. Falese, J. P., Donlic, A. & Hargrove, A. E. Targeting RNA with small molecules: from fundamental principles towards the clinic. *Chemical Society Reviews* **50**, 2224–2243 (2021).
158. Menichelli, E. *et al.* Discovery of small molecules that target a tertiary-structured RNA. *Proceedings of the National Academy of Sciences* **119**, e2213117119 (2022).
159. Suresh, B. M. *et al.* Fragment-based approaches to identify RNA binders. *Journal of medicinal chemistry* **66**, 6523–6541 (2023).
160. Costales, M. G. *et al.* How we think about targeting RNA with small molecules. *Journal of medicinal chemistry* **63**, 8880–8900 (2020).
161. Di Giorgio, A. & Duca, M. Synthetic small-molecule RNA ligands: future prospects as therapeutic agents. *MedChem-Comm* **10**, 1242–1255 (2019).
162. Bagnolini, G., Luu, T. B. & Hargrove, A. E. Recognizing the power of machine learning and other computational methods to accelerate progress in small molecule targeting of RNA. *RNA* **29**, 473–488 (2023).
163. Hewitt, W. M., Calabrese, D. R. & Schneekloth Jr, J. S. Evidence for ligandable sites in structured RNA throughout the Protein Data Bank. *Bioorganic & Medicinal Chemistry* **27**, 2253–2260 (2019).
164. Campagne, S. *et al.* Structural basis of a small molecule targeting RNA for a specific splicing correction. *Nature chemical biology* **15**, 1191–1198 (2019).
165. Daldrop, P. *et al.* Novel ligands for a purine riboswitch discovered by RNA-ligand docking. *Chemistry & biology* **18**, 324–335 (2011).
166. Colizzi, F. *et al.* Probing riboswitch binding sites with molecular docking, focused libraries, and in-line probing assays 2013.
167. Lafontaine, D. & Dubé, A. *Therapeutic Applications of Ribozymes and Riboswitches* (Springer, 2016).
168. Wang, L. *et al.* Accurate and reliable prediction of relative ligand binding potency in prospective drug discovery by way of a modern free-energy calculation protocol and force field. *Journal of the American Chemical Society* **137**, 2695–2703 (2015).
169. Gervasio, F. L., Laio, A. & Parrinello, M. Flexible docking in solution using metadynamics. *Journal of the American Chemical Society* **127**, 2600–2607 (2005).
170. Limongelli, V. *et al.* Molecular basis of cyclooxygenase enzymes (COXs) selective inhibition. *Proceedings of the National Academy of Sciences* **107**, 5411–5416 (2010).
171. Favia, A. D. *et al.* Substrate binding process and mechanistic functioning of type 1 11 β -hydroxysteroid dehydrogenase from enhanced sampling methods. *PLoS One* **6**, e25375 (2011).
172. Provati, D., Bortolato, A. & Filizola, M. Exploring molecular mechanisms of ligand recognition by opioid receptors with metadynamics. *Biochemistry* **48**, 10020–10029 (2009).
173. Di Leva, F. S. *et al.* Mechanistic insight into ligand binding to G-quadruplex DNA. *Nucleic acids research* **42**, 5447–5455 (2014).
174. Bertazzo, M. *et al.* Fully flexible docking via reaction-coordinate-independent molecular dynamics simulations. *Journal of Chemical Information and Modeling* **58**, 490–500 (2018).
175. Aguti, R. *et al.* On the allosteric puzzle and pocket crosstalk through computational means. *The Journal of Chemical Physics* **158** (2023).

176. Schmidtke, P. *et al.* MDpocket: open-source cavity detection and characterization on molecular dynamics trajectories. *Bioinformatics* **27**, 3276–3285 (2011).
177. Durrant, J. D., de Oliveira, C. A. F. & McCammon, J. A. POVME: an algorithm for measuring binding-pocket volumes. *Journal of Molecular Graphics and Modelling* **29**, 773–776 (2011).
178. La Sala, G. *et al.* Allosteric communication networks in proteins revealed through pocket crosstalk analysis. *ACS central science* **3**, 949–960 (2017).
179. Limongelli, V., Bonomi, M. & Parrinello, M. Funnel metadynamics as accurate binding free-energy method. *Proceedings of the National Academy of Sciences* **110**, 6358–6363 (2013).
180. Ganser, L. R. *et al.* High-performance virtual screening by targeting a high-resolution RNA dynamic ensemble. *Nature structural & molecular biology* **25**, 425–434 (2018).
181. Levintov, L. & Vashisth, H. Ligand recognition in viral RNA necessitates rare conformational transitions. *The Journal of Physical Chemistry Letters* **11**, 5426–5432 (2020).
182. Karthikeyan, S. *et al.* Crystal structure of human riboflavin kinase reveals a β barrel fold and a novel active site arch. *Structure* **11**, 265–273 (2003).
183. Serganov, A., Huang, L. & Patel, D. J. Coenzyme recognition and gene regulation by a flavin mononucleotide riboswitch. *nature* **458**, 233–237 (2009).
184. Bosio, S. *et al.* Similarities and Differences in Ligand Binding to Protein and RNA Targets: The Case of Riboflavin. *Journal of Chemical Information and Modeling* **64**, 4570–4586 (2024).
185. Humphrey, W., Dalke, A. & Schulten, K. VMD: visual molecular dynamics. *Journal of molecular graphics* **14**, 33–38 (1996).
186. Madhavi Sastry, G. *et al.* Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. *Journal of computer-aided molecular design* **27**, 221–234 (2013).
187. Shelley, J. C. *et al.* Epik: a software program for pK_a prediction and protonation state generation for drug-like molecules. *Journal of computer-aided molecular design* **21**, 681–691 (2007).
188. Jacobson, M. P. *et al.* On the role of the crystal environment in determining protein side-chain conformations. *Journal of molecular biology* **320**, 597–608 (2002).
189. Greenwood, J. R. *et al.* Towards the comprehensive, rapid, and accurate prediction of the favorable tautomeric states of drug-like molecules in aqueous solution. *Journal of computer-aided molecular design* **24**, 591–604 (2010).
190. Case, D. A. *et al.* AmberTools. *Journal of chemical information and modeling* **63**, 6183–6191 (2023).
191. Izadi, S., Anandakrishnan, R. & Onufriev, A. V. Building water models: a different approach. *The journal of physical chemistry letters* **5**, 3863–3871 (2014).
192. White, N. *et al.* Na⁺ riboswitches regulate genes for diverse physiological processes in bacteria. *Nature chemical biology* **18**, 878–885 (2022).
193. White, N. *et al.* Lithium-sensing riboswitch classes regulate expression of bacterial cation transporter genes. *Scientific Reports* **12**, 19145 (2022).
194. Tian, C. *et al.* ff19SB: amino-acid-specific protein backbone parameters trained against quantum mechanics energy surfaces in solution. *Journal of chemical theory and computation* **16**, 528–552 (2019).
195. Zgarbová, M. *et al.* Refinement of the Cornell *et al.* nucleic acids force field based on reference quantum chemical calculations of glycosidic torsion profiles. *Journal of chemical theory and computation* **7**, 2886–2902 (2011).
196. Allnér, O., Nilsson, L. & Villa, A. Magnesium ion–water coordination and exchange in biomolecular simulations. *Journal of chemical theory and computation* **8**, 1493–1502 (2012).

REFERENCES

197. Bayly, C. I. *et al.* A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. *The Journal of Physical Chemistry* **97**, 10269–10280 (1993).
198. Frisch, M. e. *et al.* *Gaussian 16* 2016.
199. Bekker, H. *et al.* *Gromacs-a parallel computer for molecular-dynamics simulations in 4th international conference on computational physics (PC 92)* (1993), 252–256.
200. Berendsen, H. J., van der Spoel, D. & van Drunen, R. GROMACS: A message-passing parallel molecular dynamics implementation. *Computer physics communications* **91**, 43–56 (1995).
201. Abraham, M. J. *et al.* GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1**, 19–25 (2015).
202. Hess, B. *et al.* LINCS: A linear constraint solver for molecular simulations. *Journal of computational chemistry* **18**, 1463–1472 (1997).
203. Evans, D. J. & Holian, B. L. The Nose–Hoover thermostat. *The Journal of Chemical Physics* **83**, 4069–4074. doi:10.1063/1.449071. <https://doi.org/10.1063/1.449071> (Oct. 1985).
204. Bernetti, M. & Bussi, G. Pressure control using stochastic cell rescaling. *The Journal of Chemical Physics* **153** (2020).
205. Bizzarri, A. R. & Cannistraro, S. *Molecular dynamics of water at the protein- solvent interface* 2002.
206. Masetti, M., Falchi, F. & Recanatini, M. Protein dynamics of the HIF-2 α PAS-B domain upon heterodimerization and ligand binding. *PLoS One* **9**, e94986 (2014).
207. Halle, B. & Persson, F. Analysis of protein dynamics simulations by a stochastic point process approach. *Journal of Chemical Theory and Computation* **9**, 2838–2848 (2013).
208. Sánchez, H. R. Residence times from molecular dynamics simulations. *The journal of physical chemistry b* **126**, 8804–8812 (2022).
209. Bizzarri, A. & Cannistraro, S. Anomalous and anisotropic diffusion of plastocyanin hydration water. *Europhysics Letters* **37**, 201 (1997).
210. Makarov, V. A. *et al.* Residence times of water molecules in the hydration sites of myoglobin. *Biophysical journal* **79**, 2966–2974 (2000).
211. Rocchia, W., Alexov, E. & Honig, B. Extending the applicability of the nonlinear Poisson- Boltzmann equation: multiple dielectric constants and multivalent ions. *The Journal of Physical Chemistry B* **105**, 6507–6514 (2001).
212. Decherchi, S. & Rocchia, W. A general and robust ray-casting-based algorithm for triangulating surfaces at the nanoscale. *PloS one* **8**, e59744 (2013).
213. Howe, J. A. *et al.* Selective small-molecule inhibition of an RNA structural element. *Nature* **526**, 672–677 (2015).
214. Bernetti, M., Bertazzo, M. & Masetti, M. Data-driven molecular dynamics: a multifaceted challenge. *Pharmaceuticals* **13**, 253 (2020).
215. Appadurai, R. *et al.* Clustering heterogeneous conformational ensembles of intrinsically disordered proteins with t-distributed stochastic neighbor embedding. *Journal of chemical theory and computation* **19**, 4711–4727 (2023).
216. Sivanesan, D., Basu, G. & Go, N. The role of electrostatics in discrimination of adenine and guanine by proteins. *Genome Informatics* **13**, 316–317 (2002).
217. Radic, Z. *et al.* Electrostatic influence on the kinetics of ligand binding to acetylcholinesterase: Distinctions between active center ligands and fasciculin. *Journal of Biological Chemistry* **272**, 23265–23277 (1997).
218. Mehler, E. & Gerhards, J. Electronic determinants of the anti-inflammatory action of benzoic and salicylic acids. *Molecular pharmacology* **31**, 284–293 (1987).
219. Wade, R. C. *et al.* Electrostatic steering and ionic tethering in enzyme–ligand binding: Insights from simulations. *Proceedings of the National Academy of Sciences* **95**, 5942–5949 (1998).

220. Sebastián, M. *et al.* Kinetics and thermodynamics of the protein-ligand interactions in the riboflavin kinase activity of the FAD synthetase from *Corynebacterium ammoniagenes*. *Scientific reports* **7**, 7281 (2017).
221. Thomas, J. R. & Hergenrother, P. J. Targeting RNA with small molecules. *Chemical reviews* **108**, 1171–1224 (2008).
222. Haniff, H. S. *et al.* Design of a small molecule that stimulates vascular endothelial growth factor A enabled by screening RNA fold–small molecule interactions. *Nature Chemistry* **12**, 952–961 (2020).
223. Ladbury, J. E. Just add water! The effect of water on the specificity of protein-ligand binding sites and its potential application to drug design. *Chemistry & biology* **3**, 973–980 (1996).
224. Hummer, G. Under water's influence. *Nature Chemistry* **2**, 906–907 (2010).
225. Baron, R., Setny, P. & McCammon, J. A. Water in cavity- ligand recognition. *Journal of the American Chemical Society* **132**, 12091–12097 (2010).
226. Henchman, R. H. *et al.* Properties of water molecules in the active site gorge of acetylcholinesterase from computer simulation. *Biophysical journal* **82**, 2671–2682 (2002).
227. García, A. E. & Stillér, L. Computation of the mean residence time of water in the hydration shells of biomolecules. *Journal of computational chemistry* **14**, 1396–1406 (1993).
228. Cuervo, A. *et al.* Direct measurement of the dielectric polarization properties of DNA. *Proceedings of the National Academy of Sciences* **111**, E3624–E3630 (2014).
229. Rohs, R. *et al.* The role of DNA shape in protein–DNA recognition. *Nature* **461**, 1248–1253 (2009).
230. Davidson, A. *et al.* Simultaneous recognition of HIV-1 TAR RNA bulge and loop sequences by cyclic peptide mimics of Tat protein. *Proceedings of the National Academy of Sciences* **106**, 11931–11936 (2009).
231. Watts, J. M. *et al.* Architecture and secondary structure of an entire HIV-1 RNA genome. *Nature* **460**, 711–716. doi:10.1038/nature08237 (2009).
232. Rocchi, C. *et al.* The HIV-1 integrase C-terminal domain induces TAR RNA structural changes promoting Tat binding. *International Journal of Molecular Sciences* **23**, 13742 (2022).
233. Kessl, J. J. *et al.* HIV-1 integrase binds the viral RNA genome and is essential during virion morphogenesis. *Cell* **166**, 1257–1268 (2016).
234. Roy, R. *et al.* Kinetic resolution of the atomic 3D structures formed by ground and excited conformational states in an RNA dynamic ensemble. *Journal of the American Chemical Society* **145**, 22964–22978 (2023).
235. Raghunathan, D. *et al.* TAR-RNA recognition by a novel cyclic aminoglycoside analogue. *Nucleic acids research* **34**, 3599–3608 (2006).
236. Murchie, A. I. *et al.* Structure-based drug design targeting an inactive RNA conformation: exploiting the flexibility of HIV-1 TAR RNA. *Journal of molecular biology* **336**, 625–638 (2004).
237. Davidson, A. *et al.* Essential structural requirements for specific recognition of HIV TAR RNA by peptide mimetics of Tat protein. *Nucleic acids research* **39**, 248–256 (2011).
238. Hornak, V. *et al.* Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins: Structure, Function, and Bioinformatics* **65**, 712–725 (2006).
239. Joung, I. S. & Cheatham III, T. E. Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations. *The journal of physical chemistry B* **112**, 9020–9041 (2008).
240. Boothroyd, S. *et al.* Development and benchmarking of open force field 2.0. 0: the Sage small molecule force field. *Journal of chemical theory and computation* **19**, 3251–3275 (2023).
241. Bussi, G., Donadio, D. & Parrinello, M. Canonical sampling through velocity rescaling. *The Journal of chemical physics* **126** (2007).

REFERENCES

242. Gowers, R. J. *et al.* *MDAnalysis: a Python package for the rapid analysis of molecular dynamics simulations* tech. rep. (Los Alamos National Laboratory (LANL), Los Alamos, NM (United States), 2019).
243. Johnston, R. C. *et al.* Epik: p K a and protonation state prediction through machine learning. *Journal of chemical theory and computation* **19**, 2380–2388 (2023).
244. Bosio, S. *et al.* Morphological inhibitors of aggregation-prone amyloid- β conformers: A computational exploration. *Computers in Biology and Medicine* **196**, 110545 (2025).
245. Friesner, R. A. *et al.* Extra precision glide: Docking and scoring incorporating a model of hydrophobic enclosure for protein- ligand complexes. *Journal of medicinal chemistry* **49**, 6177–6196 (2006).
246. Duan, J. *et al.* Analysis and comparison of 2D fingerprints: insights into database screening performance using eight fingerprint methods. *Journal of Molecular Graphics and Modelling* **29**, 157–170 (2010).
247. Bender, A. & Glen, R. C. Molecular similarity: a key technique in molecular informatics. *Organic & biomolecular chemistry* **2**, 3204–3218 (2004).
248. Landrum, G. *et al.* *RDKit: Open-source cheminformatics* 2006.
249. Bemis, G. W. & Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *Journal of medicinal chemistry* **39**, 2887–2893 (1996).
250. Kruger, F., Stiefl, N. & Landrum, G. A. rdScaffoldNetwork: the scaffold network implementation in RDKit. *Journal of chemical information and modeling* **60**, 3331–3335 (2020).
251. Hagberg, A., Swart, P. J. & Schult, D. A. *Exploring network structure, dynamics, and function using NetworkX* tech. rep. (Los Alamos National Laboratory (LANL), Los Alamos, NM (United States), 2008).
252. Perrone, G., Unpingco, J. & Lu, H.-m. Network visualizations with Pyvis and VisJS. *arXiv preprint arXiv:2006.04951* (2020).
253. Du, Z., Lind, K. E. & James, T. L. Structure of TAR RNA complexed with a Tat-TAR interaction nanomolar inhibitor that was identified by computational screening. *Chemistry & biology* **9**, 707–712 (2002).
254. Bottaro, S., Di Palma, F. & Bussi, G. The role of nucleobase interactions in RNA structure and dynamics. *Nucleic acids research* **42**, 13306–13314 (2014).
255. Musiani, F. *et al.* Molecular dynamics simulations identify time scale of conformational changes responsible for conformational selection in molecular recognition of HIV-1 transactivation responsive RNA. *Journal of the American Chemical Society* **136**, 15631–15637 (2014).
256. Geng, A. *et al.* An RNA excited conformational state at atomic resolution. *Nature Communications* **14**, 8432 (2023).
257. Grossfield, A. & Zuckerman, D. M. Quantifying uncertainty and sampling quality in biomolecular simulations. *Annual reports in computational chemistry* **5**, 23–48 (2009).
258. Bottaro, S. & Lindorff-Larsen, K. Biophysical experiments and biomolecular simulations: A perfect match? *Science* **361**, 355–360 (2018).
259. Bonomi, M. *et al.* Principles of protein structural ensemble determination. *Current opinion in structural biology* **42**, 106–116 (2017).
260. Nadler, W. & Hansmann, U. H. Optimizing replica exchange moves for molecular dynamics. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics* **76**, 057102 (2007).
261. Prestegard, J., Bougault, C. & Kishore, A. Residual dipolar couplings in structure determination of biomolecules. *Chemical reviews* **104**, 3519–3540 (2004).
262. Zweckstetter, M. NMR: prediction of molecular alignment from structure using the PALES software. *Nature protocols* **3**, 679–690 (2008).

-
263. McGinnis, J. L. *et al.* The mechanisms of RNA SHAPE chemistry. *Journal of the American Chemical Society* **134**, 6617–6624 (2012).
264. Mlýnský, V. & Bussi, G. Molecular dynamics simulations reveal an interplay between SHAPE reagent binding and RNA flexibility. *The journal of physical chemistry letters* **9**, 313–318 (2018).
265. Halgren, T. A. Identifying and characterizing binding sites and assessing druggability. *Journal of Chemical Information and Modeling* **49**, 377–389. doi:10.1021/ci800324m (2009).
266. Cesari, A., Reißer, S. & Bussi, G. Using the maximum entropy principle to combine simulations and solution experiments. *Computation* **6**, 15 (2018).
267. Hammes, G. G., Chang, Y.-C. & Oas, T. G. Conformational selection or induced fit: a flux description of reaction mechanism. *Proceedings of the National Academy of Sciences* **106**, 13737–13741 (2009).
268. Liu, T. *et al.* Molecular insights into de novo small-molecule recognition by an intron RNA structure. *Proceedings of the National Academy of Sciences* **122**, e2502425122 (2025).
269. McCluskey, K. & Penedo, J. C. An integrated perspective on RNA aptamer ligand-recognition models: clearing muddy waters. *Physical Chemistry Chemical Physics* **19**, 6921–6932 (2017).
270. Olenginski, L. T., Spradlin, S. F. & Batey, R. T. Flipping the script: Understanding riboswitches from an alternative perspective. *Journal of Biological Chemistry* **300**, 105730 (2024).
271. Faber, C. *et al.* Structural rearrangements of HIV-1 Tat-responsive RNA upon binding of neomycin B. *Journal of Biological Chemistry* **275**, 20660–20666 (2000).
272. Davis, B. *et al.* Rational design of inhibitors of HIV-1 TAR RNA through the stabilisation of electrostatic “hot spots”. *Journal of molecular biology* **336**, 343–356 (2004).
273. Davidson, A. *et al.* A small-molecule probe induces a conformation in HIV TAR RNA capable of binding drug-like fragments. *Journal of molecular biology* **410**, 984–996 (2011).
274. Aytenfisu, A. H. *et al.* Revised RNA dihedral parameters for the amber force field improve RNA molecular dynamics. *Journal of chemical theory and computation* **13**, 900–915 (2017).
275. Chen, A. A. & García, A. E. High-resolution reversible folding of hyperstable RNA tetraloops using molecular dynamics simulations. *Proceedings of the National Academy of Sciences* **110**, 16820–16825 (2013).
276. Pranata, J., Wierschke, S. G. & Jorgensen, W. L. OPLS potential functions for nucleotide bases. Relative association constants of hydrogen-bonded base pairs in chloroform. *Journal of the American Chemical Society* **113**, 2810–2819 (1991).
277. Robertson, M. J. *et al.* Development and Testing of the OPLS-AA/M Force Field for RNA. *Journal of chemical theory and computation* **15**, 2734–2742 (2019).
278. The PLUMED consortium. Promoting transparency and reproducibility in enhanced molecular simulations. *Nature Methods* **16**, 670–673. doi:10.1038/s41592-019-0506-8. <https://doi.org/10.1038/s41592-019-0506-8> (Aug. 2019).

Appendices

List of Figures

1.1	Schematic representation of the Central Dogma, illustrating the flow of genetic information from DNA to RNA to proteins.	2
1.2	Structural features of RNA nucleotides and base pairing. (A) Chemical structures of the four canonical RNA nucleobases: purines (adenine, A; guanine, G) and pyrimidines (cytosine, C; uracil, U). (B) Structure of an RNA nucleotide, highlighting the ribose sugar, phosphate group, and nitrogenous base (adenine shown as an example). The principal torsional angles of the RNA backbone (α , β , γ , δ , ϵ , ζ , and χ) are indicated. (C) Base-pairing schemes: canonical Watson–Crick pairs (A–U and G–C) and the non-canonical wobble G–U pair. Hydrogen bonds and the orientation of major and minor grooves are shown. (D) Schematic representation of a nucleotide within the RNA backbone, showing the phosphate group (yellow), ribose sugar (green), and nitrogenous base (red), and their connectivity to adjacent nucleotides.	6
1.3	Schematic representations of common RNA secondary structure motifs, including duplexes, internal loops, bulges, hairpins, multiloops, multi-way junctions, and RNA triplexes. Nucleotides are color-coded as follows: adenine (A) in yellow, uracil (U) in green, guanine (G) in blue, and cytosine (C) in red. Light grey lines indicate phosphodiester backbone connections. Black solid lines denote canonical base pairs (Watson–Crick), while dotted lines indicate non-canonical interactions.	8
3.1	(A) Classical GRID approach: a cubic lattice surrounds the protein surface and nodes within 6 Å are colored according to probe type (DRY: hydrophobic, orange; N1: H-bond donor, blue; O: H-bond acceptor, red). (B) Multiple-copy simultaneous search (MCSS) with benzene probes placed on the protein surface.	27
3.2	Mixed-solvent MD simulations: small organic cosolvents are added to explicit solvent and monitored during MD to map binding hotspots and cryptic pockets on biomolecules.	28

LIST OF FIGURES

- 4.1 Systems considered in this study: (A) the riboflavin kinase (PDB ID 1NB9 [182]) and (B) the FMN riboswitch (PDB ID 3F4G [183]). The protein and RNA molecules are shown in the NewCartoon representation of the Visual Molecular Dynamics (VMD) software, [185] while the riboflavin ligand is displayed in licorice (with ligand carbon atoms highlighted in green). The two biomolecules are shown in scale. In the corresponding panels below, we report details of the (C) protein- and (D) RNA-ligand interactions in the crystal structures, with hydrogen bonds and $\pi - \pi$ stackings shown as dashed lines in yellow and purple, respectively. The interacting residues from the target are labeled and are shown in licorice. In the RF kinase catalytic site (C), the ADP is shown in licorice, while the magnesium ion is shown in green. 39
- 4.2 Schematic representation of the thermodynamic cycle considered for the calculation of the electrostatic contribution to the binding free energy. A generic ligand and biomolecular target are shown, with the top row indicating the binding process in vacuum, while the bottom row represents the event in solution. The arrows are labeled with the corresponding term. 43

- 4.3 (A) RMSD of the RF kinase and the FMN riboswitch with respect to the starting structure of the MD simulations. The RMSDs were computed on the macromolecule heavy atoms and are shown as probability distributions by aggregating data from the three MD-simulation replicates performed on both ligand-bound (RF kinase B, FMN ribo B) and unbound systems (RF kinase U, FMN ribo U). (B) RMSD of riboflavin in the ligand-bound simulations with respect to the starting structure of the MD simulations. The RMSD, shown as probability density distribution, was computed after optimal alignment on binding pocket atoms with lowest root-mean-square fluctuation during the MD simulations. C) Superposition of the most populated configurations in the MD simulations (colored structures in green, purple and red, in decreasing cluster population order) with the corresponding one in the crystal (white structure, including the surrounding targets in transparent NewCartoon representation), for the riboflavin ligand bound to the RF kinase (left panel) and the FMN riboswitch (right panel). 47
- 4.4 t-SNE space for (A) the RF kinase and (B) the FMN riboswitch, both in the ligand-bound and unbound states. The low dimensionality spaces were generated using as input features the minimum distances between residues within the protein and RNA binding sites. Contour lines are the kernel density estimation of the point distributions and better highlight regions of overlap between the ligand-bound and unbound systems. Colored stars indicate the location of ligand configurations in the clusters shown in Figure 4.3, using the same color code. 49

- 4.5 Distributions of the pocket volumes explored during the MD simulations of the (A) RF kinase and (B) FMN riboswitch, both in the bound and unbound states. In the bottom panels C) and D), the surfaces of the pockets for the different systems are also shown, color coded consistently with the corresponding panels A) and B). In C), the left panel has colored target structure and pocket surface for the unbound state of the protein, with the target structure in the bound state in white for comparison, while the right panel has colored target structure and pocket surface for the bound state, with the target structure in the unbound state in white. The same scheme is followed in D) for the RNA target. The displayed snapshots correspond to trajectory frames with pocket volumes closest to the average values observed in the simulations. 51
- 4.6 Electrostatic analysis on the RF kinase and the FMN riboswitch systems. The surface representation of the most populated cluster for the (A) protein and (B) RNA complexes is colored according to their electrostatic potential. The color scale is consistent for the two systems and ranges from $-10 k_B T/e$ (red) to $+10 k_B T/e$ (blue). The bound riboflavin ligand is shown in licorice, with its carbon atoms highlighted in green. Below each system, we report the cluster population (upper panel) and the corresponding electrostatic component to the binding free energy in an implicit water buffer (lower panel, expressed in $k_B T/e$). 54

4.7	(A) Occurrence of hydrogen bonds formed by the riboflavin ligand with the RF kinase (leftmost panel) and the FMN riboswitch (rightmost panel) during the MD simulations. For each riboflavin heteroatom (central structure, only heavy atoms displayed and heteroatoms annotated with atom names), the two plots report the percentage occurrence of hydrogen bonds, highlighting the fraction where the atom acts as a donor or acceptor. (B) Polar plots of the dihedrals in the riboflavin flexible tail (see the annotated structure in (A) for schematics). The central panel reports dihedral values computed from the MD simulation of riboflavin in water (gray points). In the leftmost and rightmost panels, these values are overlapped with the ones computed from the simulations of riboflavin in complex with the RF kinase (blue points) and FMN riboswitch (orange points), respectively.	56
4.8	Survival probability of the water molecules in the binding site region of the RF kinase and FMN riboswitch systems from the MD simulations. Different colors are used for the two macromolecular systems and to distinguish simulations in complex with riboflavin (B, bound) or without the riboflavin ligand (U, unbound).	59
5.1	Schematic representation of the interplay among Integrase (IN), Tat, and TAR during the early phase of HIV-1 proviral transcription. After integration, Integrase binds the nascent TAR RNA and promotes conformational rearrangements that enable Tat recognition. Tat then displaces Integrase from TAR and recruits the catalytic P _{TEFb} complex, composed of cyclin T1 and CDK9, which stimulates RNA polymerase II elongation and yields full length transcripts.	67
5.2	Structure of HIV-1 TAR. The secondary structure with residue numbering (left) and a representative conformer (right) illustrate the stem-loop (blue), the bulge (yellow), and the lower helix (red)	68

- 5.3 Workflow of SHAMAN for mixed solvent mapping on RNA. Input: select the RNA target and a panel of probe molecules, set up a mother system without ligands and multiple replica systems, each replica carries a different probe. Production: run an unbiased and unrestrained molecular dynamics simulation of the mother, broadcast the positions of RNA backbone atoms to the replicas that are restrained to follow the mother like shadows, accelerate probe exploration with metadynamics. Analysis: first, cluster the sampled RNA ensemble into representative conformations; second, for each RNA cluster and each probe, compute a free energy map from probe occupancy; third, merge contiguous voxels into interacting sites; fourth, assign to each site a free energy and a buriedness score and discard solvent exposed sites; fifth, within each RNA cluster, combine the interacting sites from all probes to obtain SHAMAPs. The procedure is detailed in Paragraph 3.1.4.1 . 69
- 5.4 Schematic representation of the workflow employed in this study. The protocol begins with two parallel SHAMAN simulations in the presence and absence of the cyclic peptide. Ensemble refinement was performed using ϵ RMSD to obtain more compact subclusters, which were subsequently evaluated with BiteNet to prioritize druggable conformations. Finally, docking and complementary machine learning-based screening were applied to select candidate molecules. 70
- 5.5 Chemical structures of the probes considered in this work. Teal, Benzothiazole, a bicyclic aromatic heteroaromatic system containing nitrogen and sulfur. Green, Primary ammonium, a positively charged amine. Dark blue, Benzene, the prototypical aromatic carbocycle. Orange, Imidazole, a five-membered aromatic heterocycle with two nitrogen atoms. Light teal, Formamide, the simplest amide functional group. Red, Ether group, an oxygen linker. Yellow, Dibenzothiazine, a tricyclic heteroaromatic scaffold containing nitrogen and sulfur. . 76

5.6	Probability maps for probe localization and pocket definition. Probe densities are encoded following Figure 5.5: teal, benzothiazole; green, primary ammonium; dark blue, benzene; orange, imidazole; light teal, formamide; yellow, dibenzothiazine. (A) Apo system, probe densities highlighting two main binding hotspot, Pocket 1 and Pocket 2. (B) Holo system, same representation, highlighting the strong contribution of the dibenzothiazine scaffold.	78
5.7	Histogram distributions of molecular descriptors for the selected molecules. Each panel reports the normalized density curve for a different property: molecular weight (dark blue), hydrogen bond donors (teal), hydrogen bond acceptors (green), nitrogen atom count (yellow), topological polar surface area (TPSA; orange), and calculated LogP (red).	80
5.8	Chemical Space Network at Tanimoto similarity threshold $\tau = 0.60$. Only highly similar molecules remain connected, forming small communities of structural analogues (colored), while the majority of compounds appear as isolated nodes (grey), reflecting the limited presence of true congeneric series within the selected compounds. The boxes display the common scaffolds for the most populated communities, with box outlines matching the corresponding community colors. . . .	82
6.1	General mechanism of ligand binding coupled to conformational change. The macromolecule interconverts between a low-affinity conformation (L, orange) and a high-affinity conformation (H, green), while the ligand is represented in blue. The kinetic constants $k_{\text{on}}^{\text{IF}}$, $k_{\text{off}}^{\text{IF}}$, $k_{\text{on}}^{\text{CS}}$, and $k_{\text{off}}^{\text{CS}}$ describe association and dissociation along the induced fit and conformational selection pathways, respectively. The interconversion rates between the unbound states ($k_{\text{L-H}}$, $k_{\text{H-L}}$) and between the ligand-bound states ($k_{\text{L-H}}^*$, $k_{\text{H-L}}^*$) represent the transitions between low- and high-affinity conformations in the absence and presence of the ligand.	89

6.2	Energy distributions of replicas at increasing reference temperatures (color-coded from blue to red). Left: PT. Right: PT-WTE, both employing the same number of replicas. In standard PT, only negligible overlap between adjacent replicas was observed, implying that a substantially larger number of replicas would be required to ensure efficient exchanges. In contrast, the PT-WTE protocol significantly broadened the energy distributions, thereby enhancing overlap across replicas and allowing the same temperature range to be covered with fewer replicas.	91
6.3	Schematic representation of the workflow followed in this study. The protocol begins with molecular dynamics simulations of TAR within the PT-WTE framework, proceeds through ensemble refinement using the maximum entropy principle, and culminates with validation against independent experimental observables. This integrative strategy aims to reconcile extensive conformational sampling with experimental accuracy.	94
6.4	Probability density distributions of RMSD (top), ϵ RMSD (middle), and radius of gyration R_g (bottom) computed over the same 700 ns trajectory. Bars report normalized histograms, solid curves show Gaussian kernel density estimates. For each metric, the distribution obtained from a 700 ns plain MD simulation of the same system is shown in grey for comparison.	95

- 6.5 Maximum entropy reweighting of TAR ensembles against RDCs and validation with independent data. (A) Comparison between experimental RDC values (grey) from previous NMR studies and back-calculated RDCs from the raw WTE ensemble (green). (B) Comparison between experimental SHAPE reactivity profiles (grey) and computed fluctuations in the mean inter-nucleotide distances between consecutive residues ($i-1/i$ and $i/i+1$). (C) Comparison between experimental RDC values (grey) from previous NMR studies and back-calculated RDCs from reweighted back-calculated RDCs (red). (D) Comparison between experimental SHAPE reactivity profiles (grey) and computed inter-nucleotide distance fluctuations between consecutive residues, with the per-residue metric obtained by averaging the fluctuations of adjacent pairs ($i-1/i$ and $i/i+1$) after reweighting (red). (E) Cross-validation scheme for the optimization of the σ parameter. The root mean square error (RMSE) obtained from 10-fold validation of the RDC dataset (violet) is plotted as a function of σ . The associated Kish size is reported in blue, whereas the RMSE computed over the full dataset is shown in yellow. (F) Structural representation of TAR highlighting the residues included (blue) or excluded (orange) from SHAPE-based validation(right); residues correctly reproduced by the reweighted ensemble are shown in green, while those with the largest deviations are marked in brown(right). 98

LIST OF FIGURES

- 6.6 Conformational space of TAR before and after maximum entropy reweighting. (A) Two-dimensional projection of the conformational space sampled during the PT-WTE simulation (grey points), compared with reference PDB structures of TAR in the unbound (red) and bound (blue) states. The left panel shows the points distribution before reweighting, while the right panel displays the same projection after applying MaxEnt reweighting. Grey contour lines trace isodensity levels from a kernel density estimate computed with the MaxEnt weights, highlighting high probability basins and their spatial extent. Eight levels are shown with a bandwidth adjustment of 0.65 and a density threshold of 0.02. PC1 and PC2 explain 16.72% and 11.25% of the total variance, respectively. Representative structures of the main clusters are indicated by star markers. (B) Representative structures of the main clusters identified in the reweighted ensemble. . 101
- 6.7 Two dimensional projection of the conformational space sampled during the PT-WTE simulation. Colors are mapped to SiteScore values, from blue for low scores to red for high scores. Point sizes are adjusted according to their statistical weight. Two prototypical examples of low and high scoring regions are highlighted. In red, the best scoring conformation within its the basin is shown, with the corresponding binding pocket in green; in blue, the worst performing conformation. 103
- 6.8 Three dimensional structures of the examined HIV-1 TAR complexes with small molecules. Ligands are colored by PDB identifier, using a scheme consistent with Figures 6.10 and D.4. The corresponding PDB code is displayed above each structure. 104
- 6.9 Population of TAR conformations within increasing ϵ RMSD thresholds (0.8, 0.9, and 1.0) relative to reference ligand-bound structures. The green bars correspond to the pre-reweighting ensemble, while the yellow bars represent the reweighted ensemble. 105

6.10	(A) Two-dimensional projection of the conformational free-energy landscape sampled around the binding mode observed in the 1LVJ structure, shown along the first two principal components (PC1 and PC2). Colored circles correspond to the crystallographic poses of the ligand in 1LVJ, and the purple star indicates the centroid of the closest simulated cluster. In the center, the three-dimensional comparison between the crystallographic binding mode and the closest simulated conformation is shown. The ligand is represented in stick form, and HIV-1 TAR is displayed as a cartoon and stick representation. Residue numbers are indicated. On the right, the same view and color scheme are used, but the ligand is represented as a molecular surface to illustrate the occupied binding cavity.(B) Two-dimensional projection of the conformational free-energy landscape sampled around the binding mode observed in the 1QD3 structure, shown along PC1 and PC2. Colored circles correspond to the crystallographic poses of the ligand in 1QD3, and the purple star indicates the centroid of the closest simulated cluster. In the center, the three-dimensional comparison between the crystallographic binding mode and the closest simulated conformation is shown. The ligand is displayed in stick representation from the front view, and HIV-1 TAR is shown as a cartoon and stick representation. On the right, the ligand is represented as a molecular surface after a 180° rotation to highlight shape complementarity. . . .	107
B.1	Timeseries and distribution of the RMSD of the biomolecular targets. The RMSD was calculated aligning and calculating on all the heavy atoms of the structures.	151

LIST OF FIGURES

- B.2 Root mean square fluctuation (RMSF) of the biomolecular targets, namely the RF kinase (left column) and the FMN riboswitch (right column). In A), the RMSF for the bound and unbound states are computed using the C α and P atoms for the protein and RNA, respectively, with dashed vertical lines indicating the pocket residues. In B), we compare different features for computing the RMSF, including the center of mass (COM) of entire residues for both systems, the C α for the protein, and the N1 or N9 (for pyrimidine or purine) of the nucleobase for the RNA. Being this metric less standard for RNAs, we compared the effect of using atoms from different regions of the nucleotide; interestingly, the plot shows how, for the RNA, fluctuations are more enhanced in the backbone (P atoms) than in the nucleobase (N1/N9 atoms), opposite to what is usually expected for proteins, where the backbone is typically more stable. 152
- B.3 Timeseries and distribution of the RMSD of the Riboflavin in complex with the RF kinase on the left and the FMN riboswitch on the right. The RMSD was calculated aligning and calculating on the most stable heavy atom in each residue that showed a RMSF values smaller than 2. 153
- B.4 PCA spaces generated by initiating from minimum distances within the protein (on the left) and RNA molecule (on the right). The scatter plot displays the different systems using the same color code used for the figures in the main text. 153
- B.5 A) Timeseries and distributions of the distance between the riboflavin O4 oxygen atoms and the Mg²⁺ ion in the three replicates of MD simulations of the ligand-bound FMN riboswitch. B) Snapshots with O4-Mg²⁺ distance value closest to their average in replicates 1 and 2 (blue and orange distributions in the above panel A)). The O4-Mg²⁺ distance is highlighted with blue dashed lines, while interactions, particularly those involving bridging water molecules, are shown in orange. 154

B.6	View of the main entrance to the A) RF kinase and B) FMN riboswitch binding site. For each system, the three panels show the target conformation in presence (leftmost panel) and absence (central panel) of the riboflavin ligand, and an overlap of these two states (rightmost panel). The protein and RNA surfaces, as long as the ligand, follow the same color code used in the main text. The snapshots for the bound and unbound states of each system are MD trajectory frames with pocket volume closest to the average value observed in the MD simulations. .	155
B.7	Fitting of the SP data with A) single exponential and B) double exponential functions.	156
B.8	Aligned sequences of the bacterial (top) and human (bottom) sequences of the RF kinase.	156
C.1	Most frequent Bemis–Murcko scaffolds identified in the selected compounds. Each structure is reported together with the number of molecules in which the scaffold occurs, highlighting the prevalence of aromatic and heteroaromatic motifs across the dataset.	158
C.2	Chemical Space Networks constructed at increasing Tanimoto similarity thresholds ($\tau = 0.40, 0.50$, and 0.60). At lower threshold ($\tau = 0.40$) the network reveals the presence of a limited number of congeneric clusters, while higher thresholds progressively fragment the graph into smaller communities, reflecting increasingly stringent structural similarity. At $\tau = 0.60$, only highly similar compounds remain connected, corresponding to true analogue series.	159
D.1	Time evolution of RMSD (top), ϵ RMSD (middle), and radius of gyration R_g (bottom) during the 700 ns well tempered metadynamics simulation of HIV-1 TAR RNA. The series display marked fluctuations in R_g and in ϵ RMSD, consistent with dynamic rearrangements of helical arms and loop regions across the trajectory.	161

LIST OF FIGURES

- D.2 Comparison between experimental SHAPE reactivity profiles (grey) and computed fluctuations in the mean inter-nucleotide distances between consecutive residues ($i-1/i$ and $i/i+1$) calculated on the ex virio dataset. Pre-reweighting data are shown in green, and reweighted data in red. 162
- D.3 Two dimensional projection of the conformational space sampled during the PT-WTE simulation. Colors indicate SiteScore values, with blue corresponding to low values and red corresponding to high values. On the right, point sizes are proportional to the statistical weight of each conformation. On the left, point sizes are constant because all weights are equal. 162
- D.4 Two-dimensional projection of the conformational free-energy landscape of TAR binding pockets. Free-energy surfaces were computed for pocket conformations sampled during the PT-WTE simulations along the first two principal components (PC1 and PC2). Colors indicate the relative free-energy values, from low-energy (dark blue) to high-energy (light blue) regions. Experimentally determined ligand-bound pocket conformations are overlaid as colored dots, with colors corresponding to the respective PDB identifiers. Purple stars indicate the centroids of the simulated clusters closest to each experimental conformation. 163
- D.5 Three dimensional comparison between the closest ensemble representative and the corresponding crystal structures. Crystal models are colored by PDB identifier. For each pocket, the left panel shows the ligand as a molecular surface to assess steric clashes and shape complementarity; the right panel shows the ligand in sticks to highlight atomic contacts. A, 1LVJ pocket; B, 1QD3 pocket; C, 1UTS pocket. Residue labels follow the original PDB numbering. 164

- D.6 Three dimensional comparison between the closest ensemble representative and the corresponding crystal structures. Crystal models are colored by PDB identifier. For each pocket, the left panel shows the ligand as a molecular surface to assess steric clashes and shape complementarity; the right panel shows the ligand in sticks to highlight atomic contacts. A, 1UUD pocket; B, 1UUI pocket; C, 2L8H pocket. Residue labels follow the original PDB numbering. 165

List of Tables

4.1	Residence times estimated through double exponential fitting of the SP function for the RF kinase and FMN riboswitch, both in the riboflavin-bound and unbound states.	60
5.1	Graph metrics for the chemical space networks at increasing Tanimoto thresholds. Clust.: average clustering coefficient; Trans.: transitivity; ASPL(GCC): average shortest path length on the largest connected component; H(deg.): degree entropy; Q : weighted modularity.	83
A.1	Main classes of RNA force fields and their major variants	149

A Theoretical Background

Table A.1. Main classes of RNA force fields and their major variants

Year	FF name	Composition	Main features
Amber			
1995	ff94[90]	ff94	
1999	ff98[91]	ff94 + P + χ	Improves pucker and twist, comparable to ff94
2000	ff99[89]	ff98 + P	Improves pucker, small changes
2007	parmbsc0[92]	ff99 + $\alpha\gamma_{\text{bsc0}}$	Avoids nonnative α/γ conformations but penalises native γ ones
2011	ff99bsc0 χ_{OL3} [195]	ff99 + $\alpha\gamma_{\text{bsc0}}$ + χ_{OL3}	Prevents high- <i>anti</i> g shifts, state of the art force field for Amber
2010	Amber99 γ [93]	ff99 + χ_{YIL}	Reduces ladder like structures and the A form inclination
2012	Amber99TOR[95]	ff99 + χ_{YIL} + $\beta_{\epsilon,\zeta\text{YIL}}$ + $\alpha\gamma_{\text{bsc0}}$	Improves the description of cytidine and uridine in solution, performs suboptimally for canonical RNA
2017	Aytenfisu, Spasic, Stern, Mathews[274]	ff99 + $\alpha\beta\gamma\epsilon\zeta_{\text{Mathews}}$	Performs well for tetraloops, reduces intercalation events
2013	Chen, García[275]	ff99 + vdW _{cc} + χ_{GC}	Reduces stacking, but overstabilises hydrogen bond interactions between bases
2018	Tan, Piana, Dirks, Shaw[103]	ff99bsc0 χ_{OL3} + $\gamma'_{\chi_{\text{Shaw}}}$ + vdW _{Shaw} + electrostatics _{Shaw}	Focused on non bonded interactions, the conformational ensembles closely reproduce experimental ones
2022	gHbfix21[98]	ff99bsc0 χ_{OL3} + HbFix	Improves hydrogen bond interactions, stabilises native structures

Continued on next page

Table A.1 – *Continued from previous page*

Year	FF name	Composition	Main features
2022	ff99OL3_CMAP1[99]	ff99bsc0 χ_{OL3} + ζ/α_{CMAP}	Decreases population of incorrect structures, improves stability of tetranucleotides
2023	NBfx _{0BPh} [100]	ff99bsc0 χ_{OL3} + NBfx _{0BPh}	Tunes base–phosphate H6/H8...O5' contacts, stabilising native <i>anti</i> states and limiting unfolding
2024	HR [101]	ff99bsc0 χ_{OL3} + HR	Weakens CH...O repulsion, improving hydrogen bonding and loop stability
CHARMM			
1995	CHARMM22[104]	CHARMM22	
2000	CHARMM27[105]	CHARMM27	Significant improvements over CHARMM22, known issues with pair opening
2011	CHARMM36[106]	CHARMM27 + 2' OH dihe-dral	Partial stabilisation of the structure, state of the art force field for CHARMM
OPLS			
1991	OPLS-AA[276]	OPLS-AA	
2019	OPLS-AA/M[277]	OPLS-AA + $\alpha\gamma\chi$ + P	Reduces intercalation events and is well suited to describing non canonical motifs

B Similarities and Differences in Ligand Binding to Protein and RNA Targets: The Case of Riboflavin

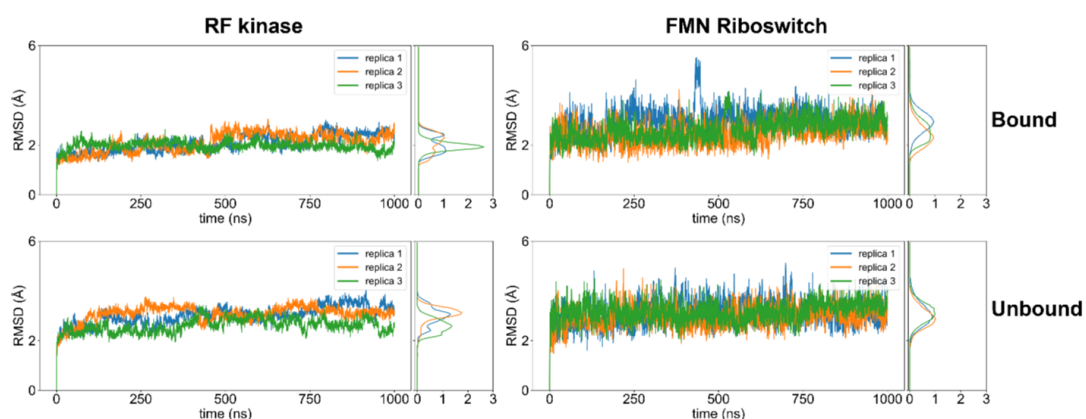


Figure B.1. Timeseries and distribution of the RMSD of the biomolecular targets. The RMSD was calculated aligning and calculating on all the heavy atoms of the structures.

B Similarities and Differences in Ligand Binding to Protein and RNA Targets: The Case of Riboflavin

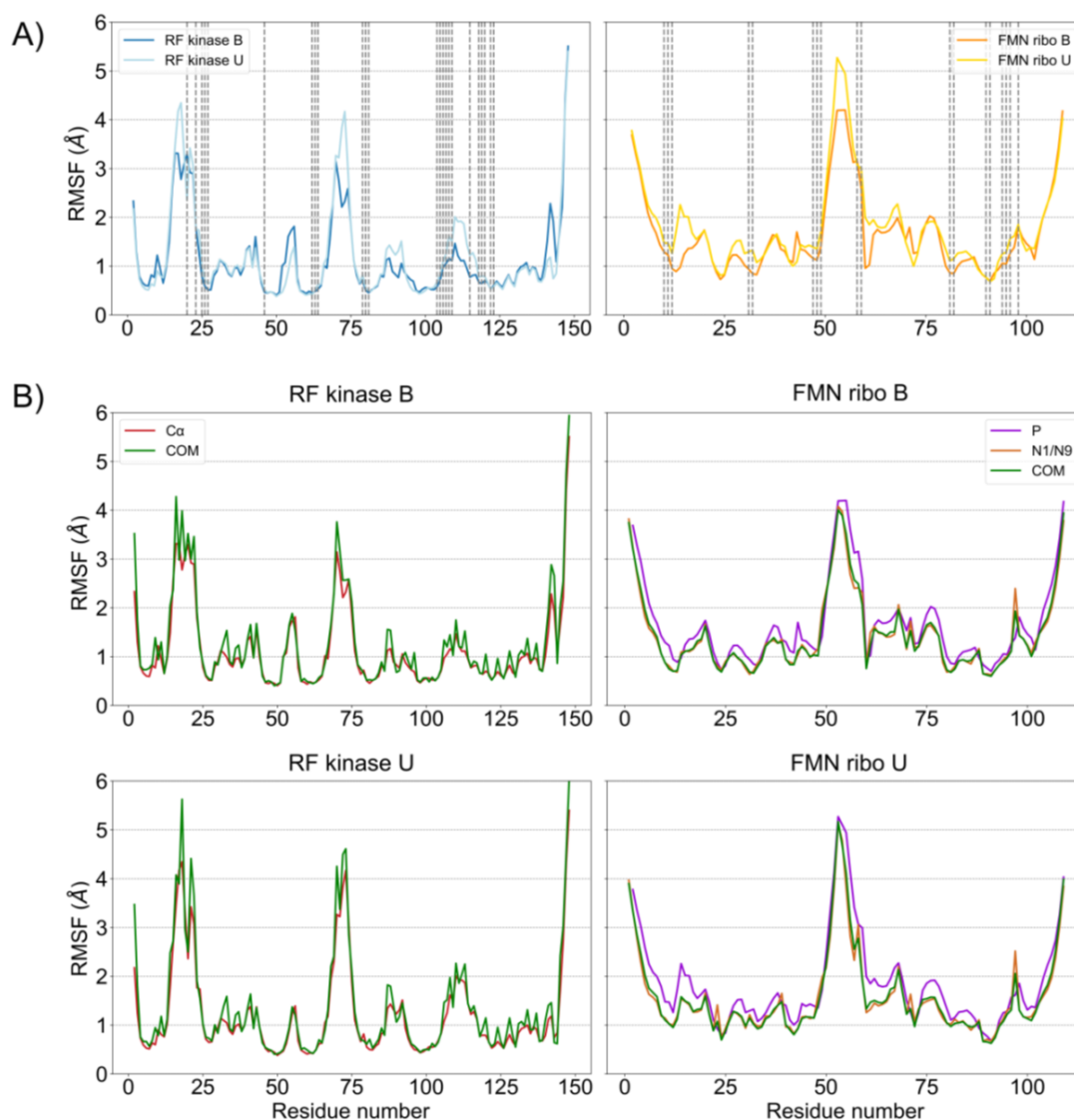


Figure B.2. Root mean square fluctuation (RMSF) of the biomolecular targets, namely the RF kinase (left column) and the FMN riboswitch (right column). In A), the RMSF for the bound and unbound states are computed using the C α and P atoms for the protein and RNA, respectively, with dashed vertical lines indicating the pocket residues. In B), we compare different features for computing the RMSF, including the center of mass (COM) of entire residues for both systems, the C α for the protein, and the N1 or N9 (for pyrimidine or purine) of the nucleobase for the RNA. Being this metric less standard for RNAs, we compared the effect of using atoms from different regions of the nucleotide; interestingly, the plot shows how, for the RNA, fluctuations are more enhanced in the backbone (P atoms) than in the nucleobase (N1/N9 atoms), opposite to what is usually expected for proteins, where the backbone is typically more stable.

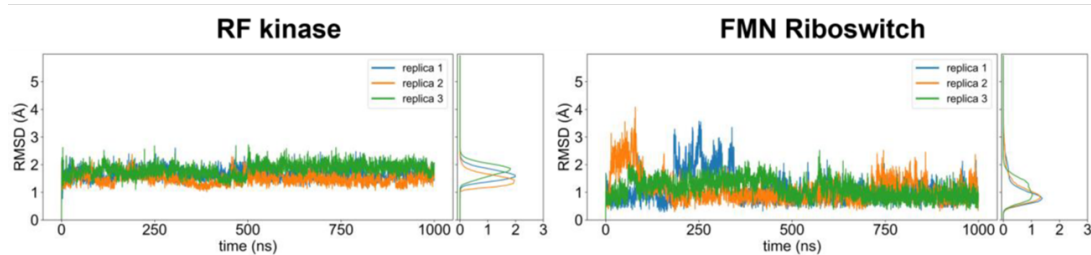


Figure B.3. Timeseries and distribution of the RMSD of the Riboflavin in complex with the RF kinase on the left and the FMN riboswitch on the right. The RMSD was calculated aligning and calculating on the most stable heavy atom in each residue that showed a RMSF values smaller than 2.

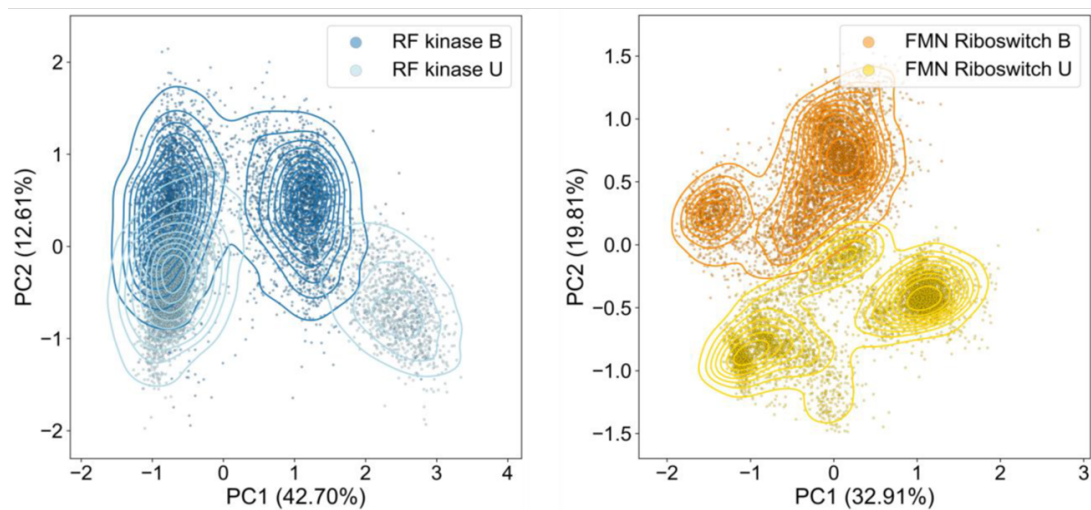


Figure B.4. PCA spaces generated by initiating from minimum distances within the protein (on the left) and RNA molecule (on the right). The scatter plot displays the different systems using the same color code used for the figures in the main text.

B Similarities and Differences in Ligand Binding to Protein and RNA Targets: The Case of Riboflavin

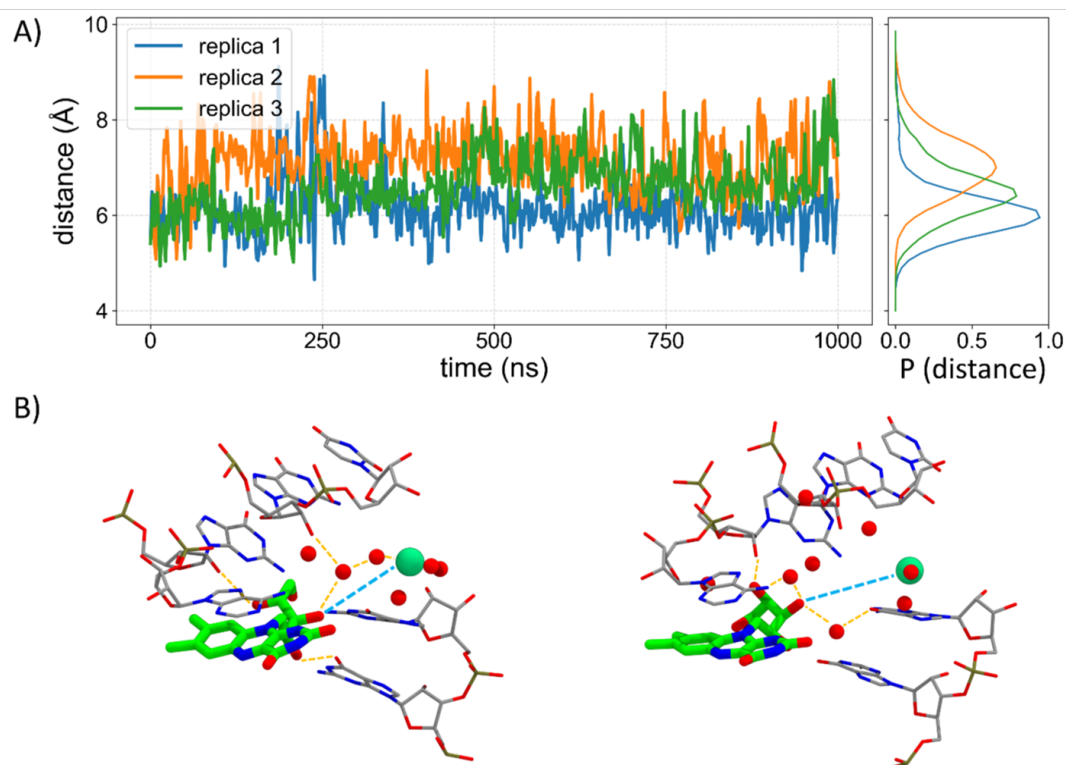


Figure B.5. A) Timeseries and distributions of the distance between the riboflavin O4 oxygen atoms and the Mg²⁺ ion in the three replicates of MD simulations of the ligand-bound FMN riboswitch. B) Snapshots with O4-Mg²⁺ distance value closest to their average in replicates 1 and 2 (blue and orange distributions in the above panel A)). The O4-Mg²⁺ distance is highlighted with blue dashed lines, while interactions, particularly those involving bridging water molecules, are shown in orange.

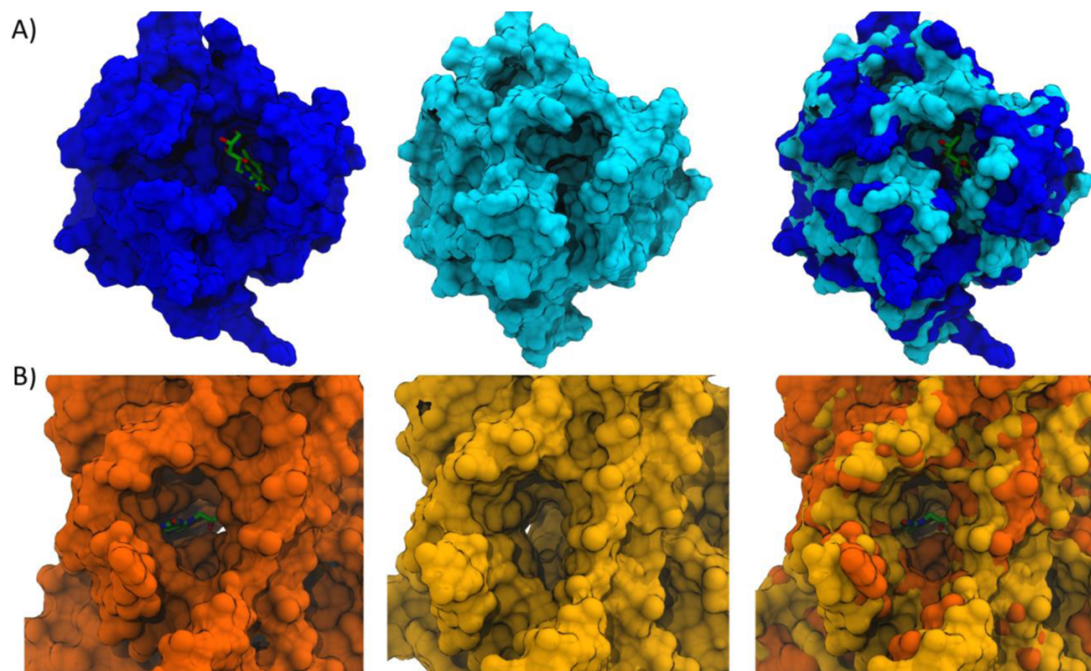


Figure B.6. View of the main entrance to the A) RF kinase and B) FMN riboswitch binding site. For each system, the three panels show the target conformation in presence (leftmost panel) and absence (central panel) of the riboflavin ligand, and an overlap of these two states (rightmost panel). The protein and RNA surfaces, as long as the ligand, follow the same color code used in the main text. The snapshots for the bound and unbound states of each system are MD trajectory frames with pocket volume closest to the average value observed in the MD simulations.

B Similarities and Differences in Ligand Binding to Protein and RNA Targets: The Case of Riboflavin

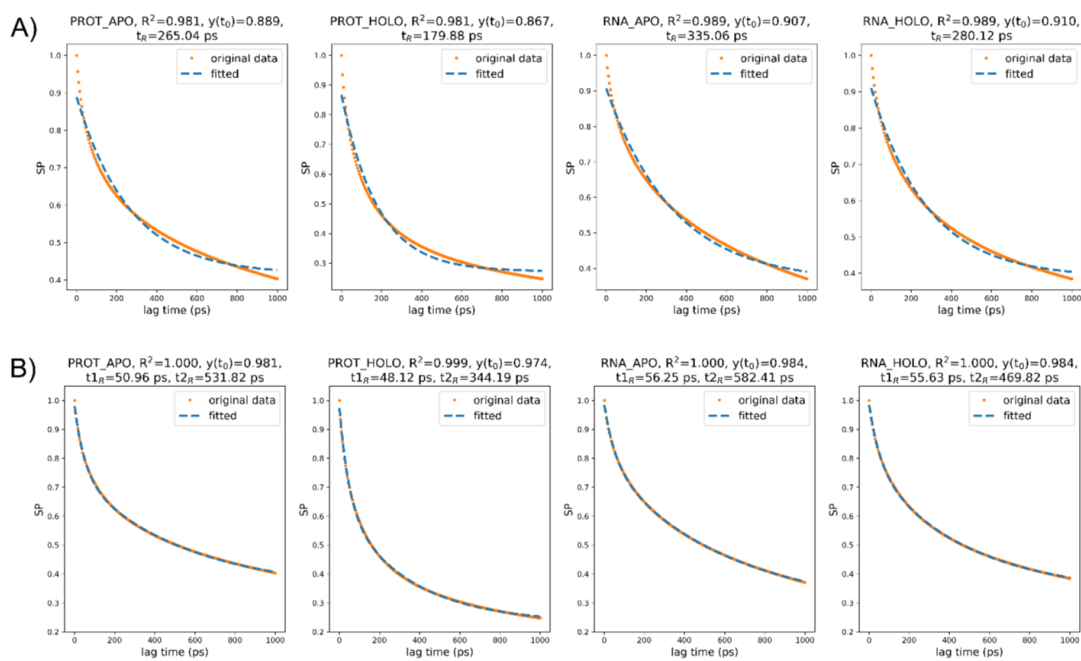


Figure B.7. Fitting of the SP data with A) single exponential and B) double exponential functions.

B.1 Calculation of sequence identity and similarity between human and bacterial RF kinase

Identities and similarities in the human and bacterial sequences of the RF kinases were computed through the Multiple Sequence Viewer Panel of the Schrödinger suite, using the FASTA sequences taken from the 1nb9 and 5a89 PDBs. Using default parameters and optimal alignment over the whole sequence, we obtained the following values (%): Identity: whole sequence=37, binding site=49 Similarity: whole sequence=61, binding site=74



Figure B.8. Aligned sequences of the bacterial (top) and human (bottom) sequences of the RF kinase.

C Exploring RNA Druggability through SHAMAN

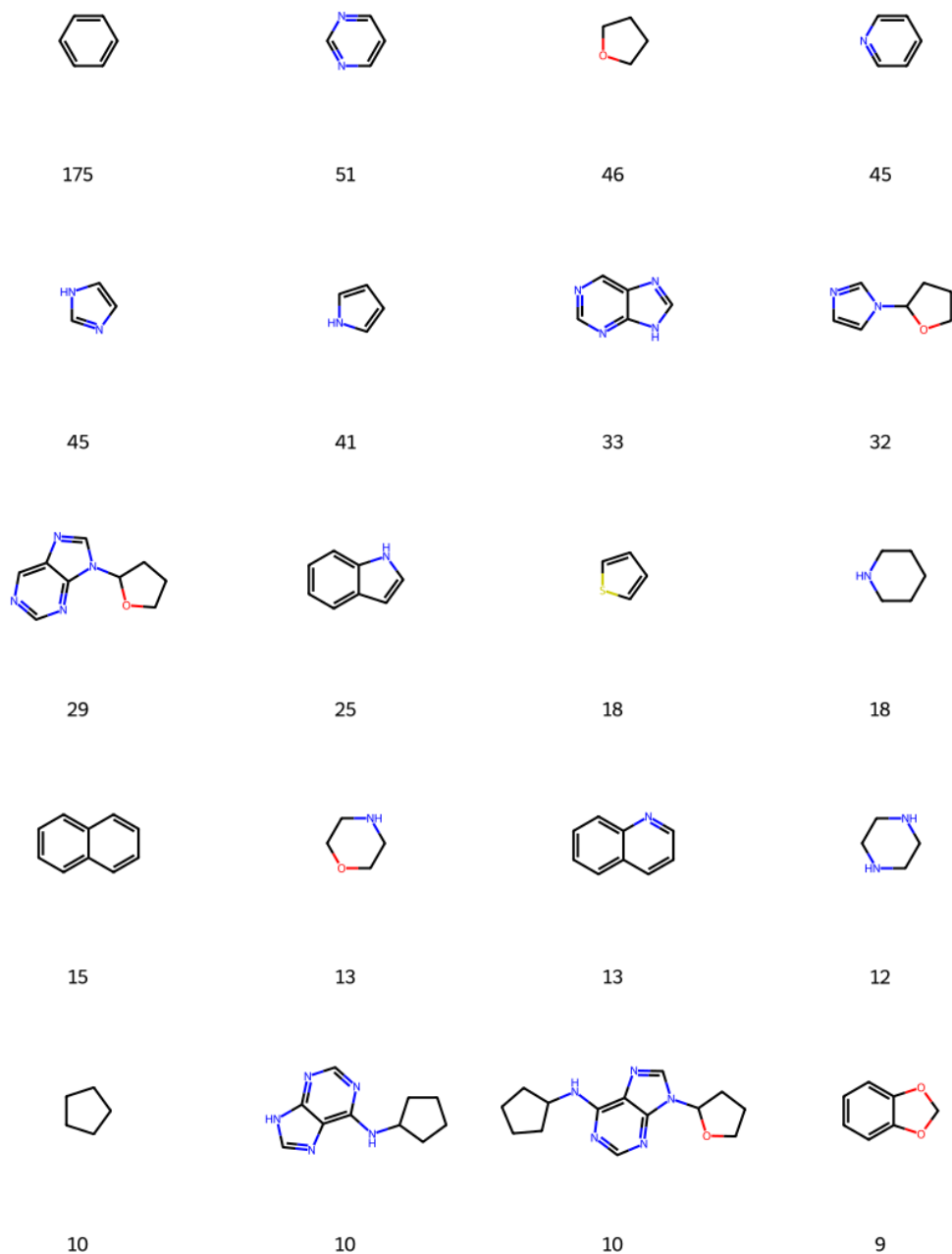


Figure C.1. Most frequent Bemis–Murcko scaffolds identified in the selected compounds. Each structure is reported together with the number of molecules in which the scaffold occurs, highlighting the prevalence of aromatic and heteroaromatic motifs across the dataset.

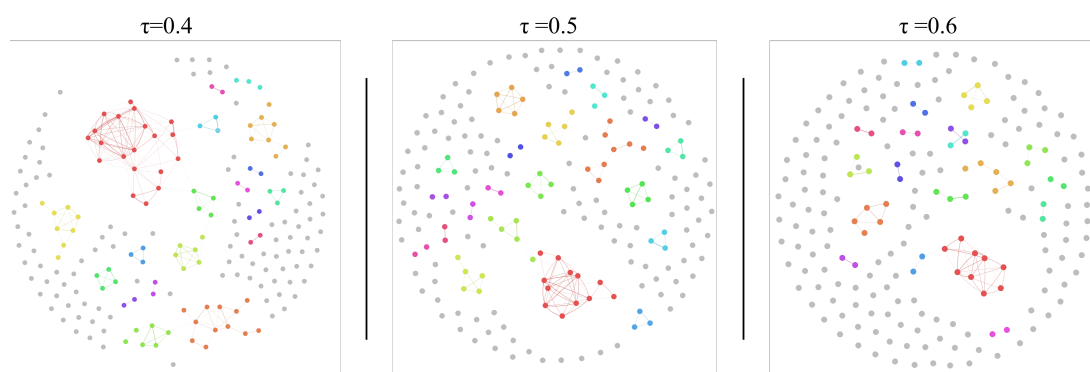


Figure C.2. Chemical Space Networks constructed at increasing Tanimoto similarity thresholds ($\tau = 0.40$, 0.50 , and 0.60). At lower threshold ($\tau = 0.40$) the network reveals the presence of a limited number of congeneric clusters, while higher thresholds progressively fragment the graph into smaller communities, reflecting increasingly stringent structural similarity. At $\tau = 0.60$, only highly similar compounds remain connected, corresponding to true analogue series.

D Integrative Refinement of RNA Conformational Ensembles via the Maximum Entropy Principle

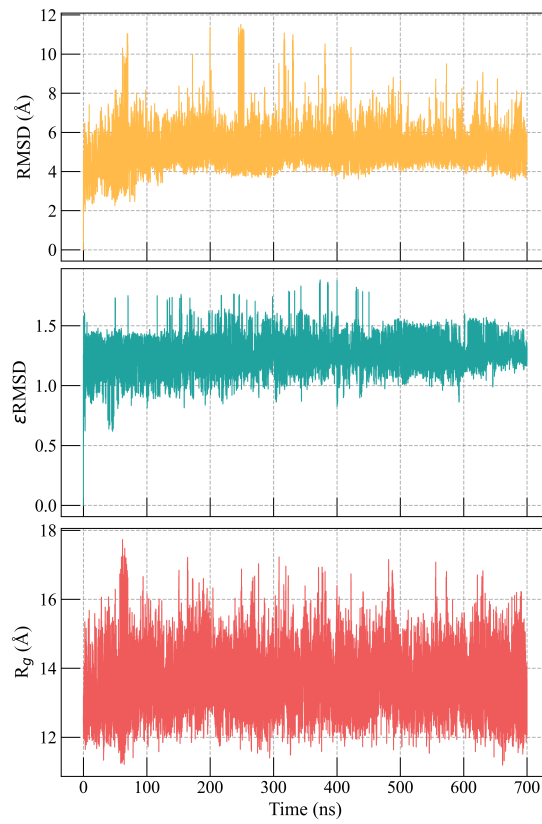


Figure D.1. Time evolution of RMSD (top), ϵ RMSD (middle), and radius of gyration R_g (bottom) during the 700 ns well tempered metadynamics simulation of HIV-1 TAR RNA. The series display marked fluctuations in R_g and in ϵ RMSD, consistent with dynamic rearrangements of helical arms and loop regions across the trajectory.

D Integrative Refinement of RNA Conformational Ensembles via the Maximum Entropy Principle

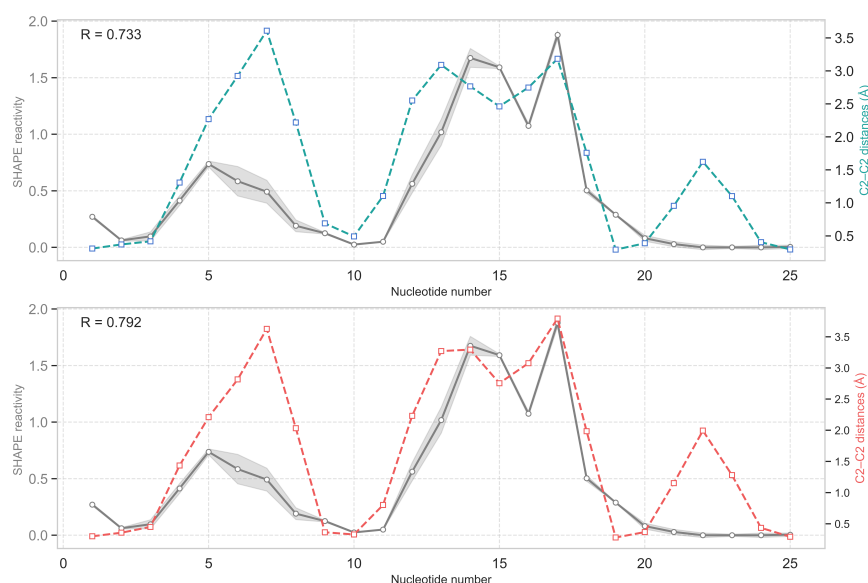


Figure D.2. Comparison between experimental SHAPE reactivity profiles (grey) and computed fluctuations in the mean inter-nucleotide distances between consecutive residues ($i-1/i$ and $i/i+1$) calculated on the ex virio dataset. Pre-reweighting data are shown in green, and reweighted data in red.

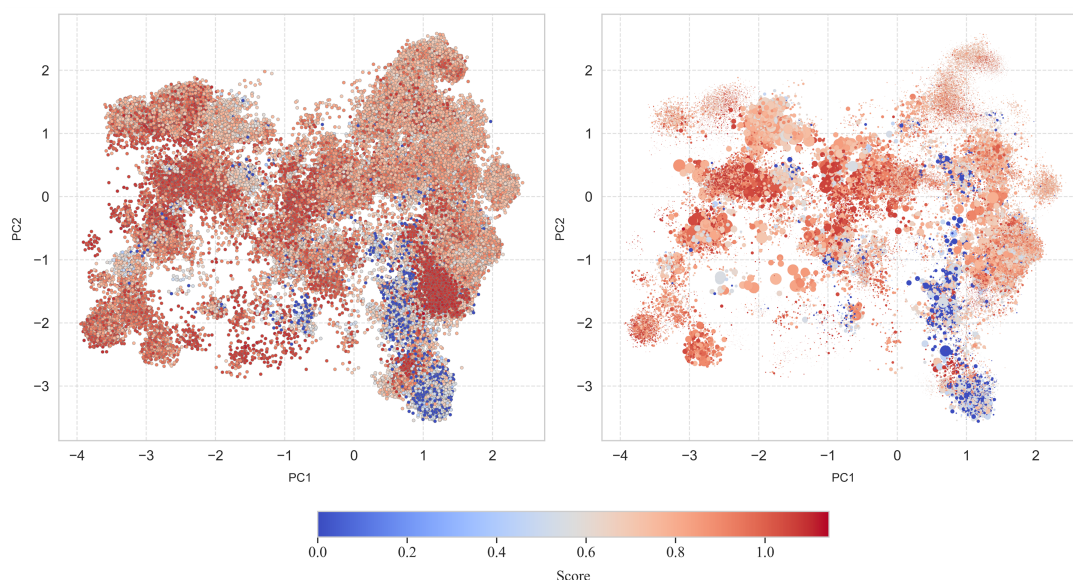


Figure D.3. Two dimensional projection of the conformational space sampled during the PT-WTE simulation. Colors indicate SiteScore values, with blue corresponding to low values and red corresponding to high values. On the right, point sizes are proportional to the statistical weight of each conformation. On the left, point sizes are constant because all weights are equal.

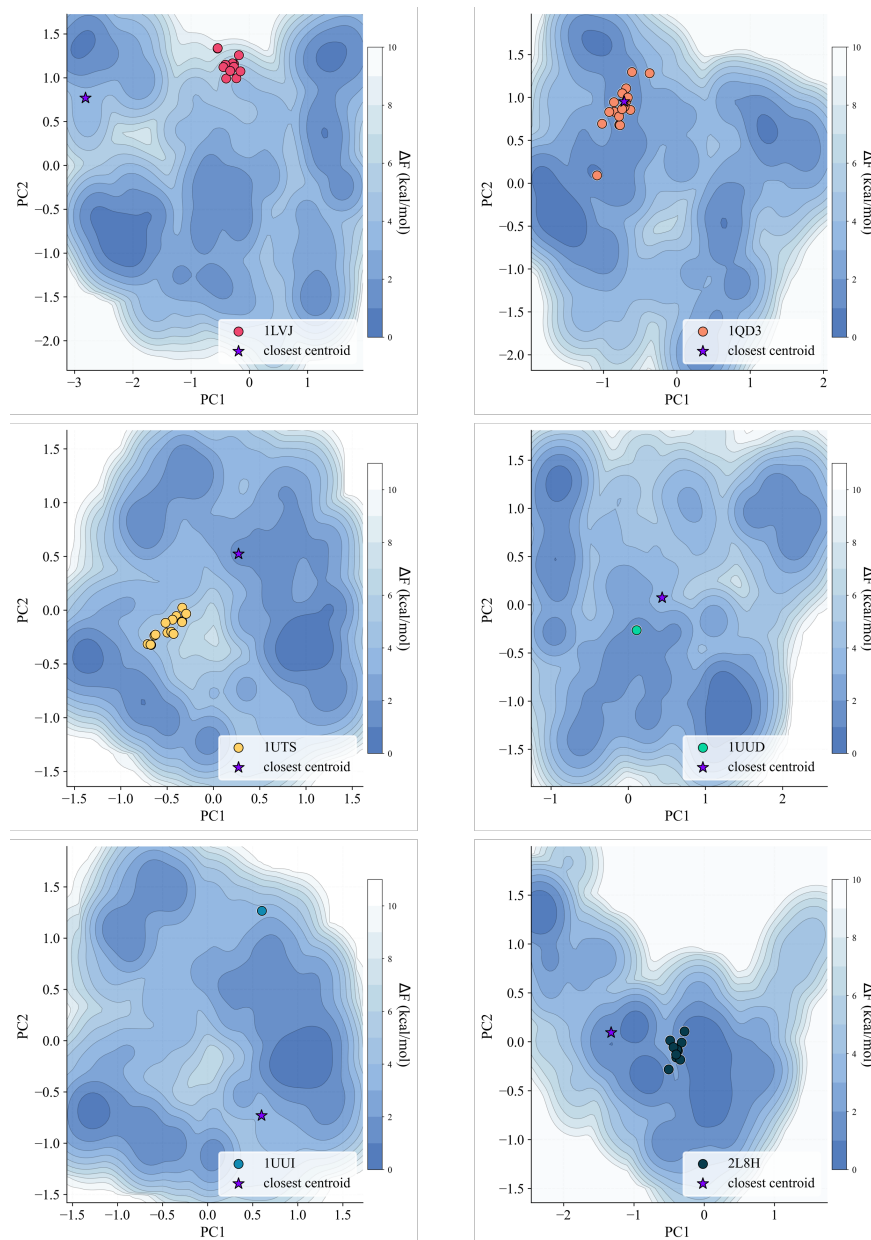


Figure D.4. Two-dimensional projection of the conformational free-energy landscape of TAR binding pockets. Free-energy surfaces were computed for pocket conformations sampled during the PT-WTE simulations along the first two principal components (PC1 and PC2). Colors indicate the relative free-energy values, from low-energy (dark blue) to high-energy (light blue) regions. Experimentally determined ligand-bound pocket conformations are overlaid as colored dots, with colors corresponding to the respective PDB identifiers. Purple stars indicate the centroids of the simulated clusters closest to each experimental conformation.

D Integrative Refinement of RNA Conformational Ensembles via the Maximum Entropy Principle

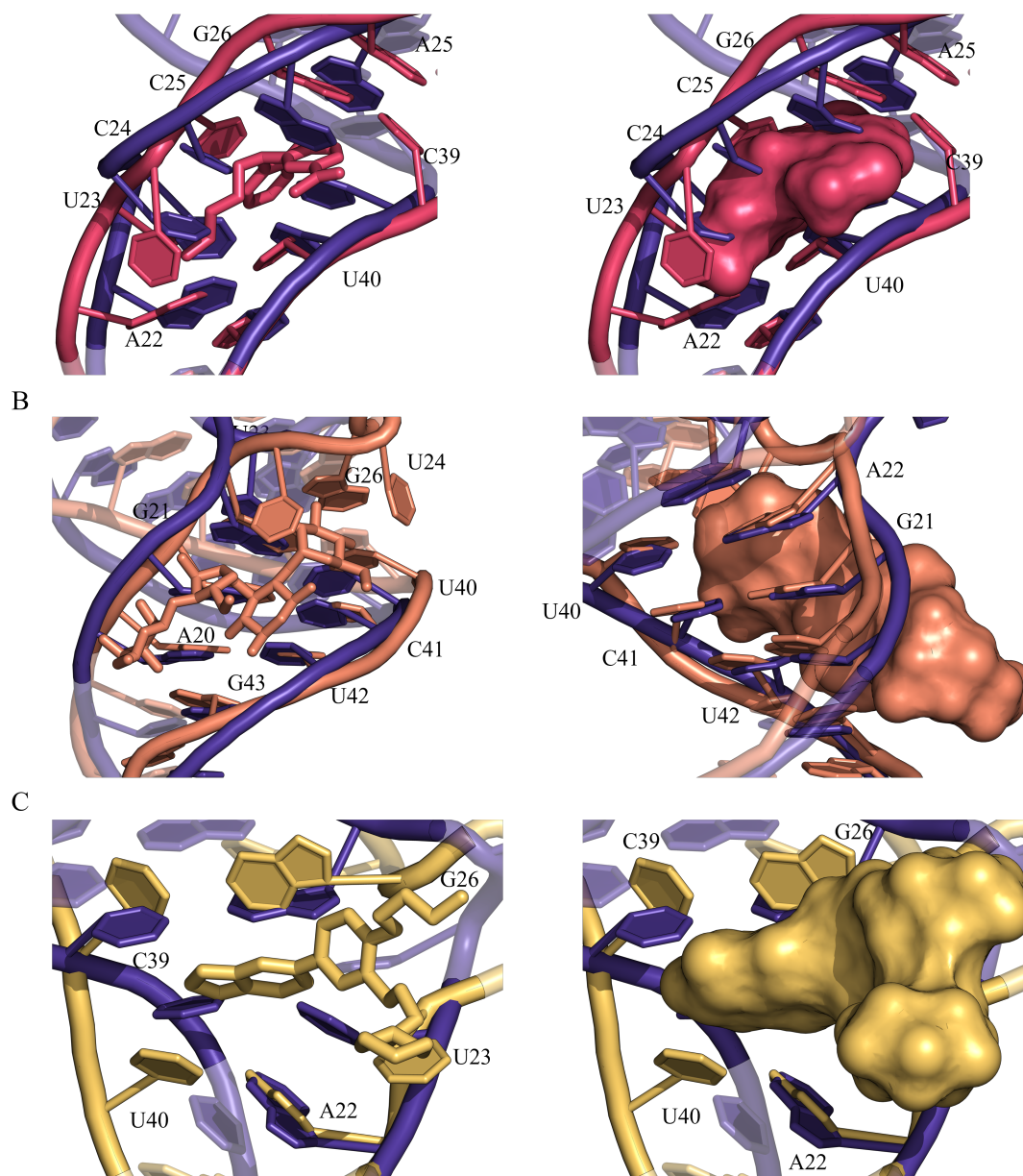


Figure D.5. Three dimensional comparison between the closest ensemble representative and the corresponding crystal structures. Crystal models are colored by PDB identifier. For each pocket, the left panel shows the ligand as a molecular surface to assess steric clashes and shape complementarity; the right panel shows the ligand in sticks to highlight atomic contacts. A, 1LVJ pocket; B, 1QD3 pocket; C, 1UTS pocket. Residue labels follow the original PDB numbering.

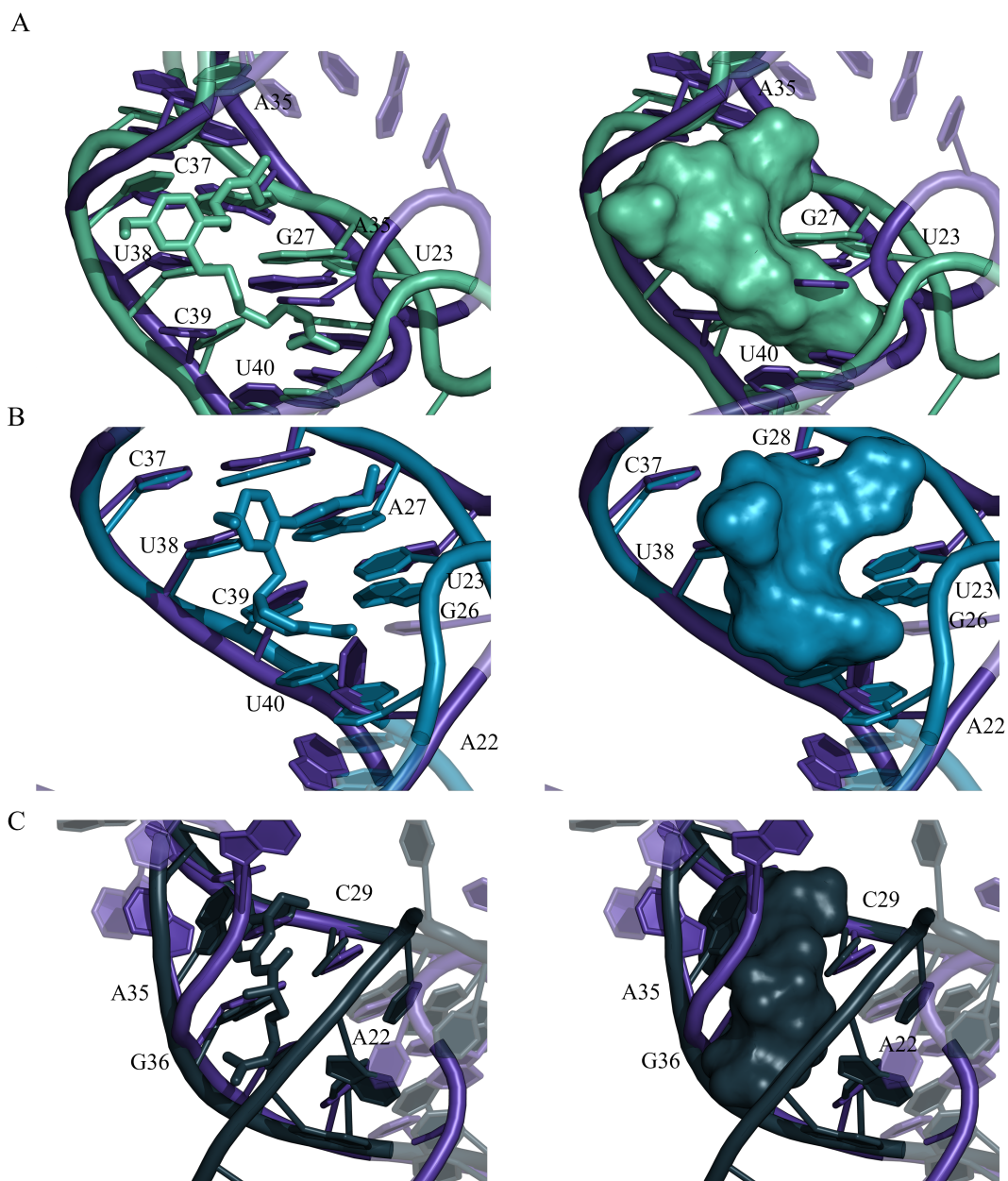


Figure D.6. Three dimensional comparison between the closest ensemble representative and the corresponding crystal structures. Crystal models are colored by PDB identifier. For each pocket, the left panel shows the ligand as a molecular surface to assess steric clashes and shape complementarity; the right panel shows the ligand in sticks to highlight atomic contacts. A, 1UUD pocket; B, 1UUI pocket; C, 2L8H pocket. Residue labels follow the original PDB numbering.

E Data availability

In line with the principles of Open Data and Open Science, all primary data and computational material generated during this doctoral work are or will be made freely available through Zenodo and GitHub. For the first project, we have deposited the initial coordinates, topologies, and GROMACS mdp parameter files, together with a strided version of the molecular dynamics trajectories that we generated, and the cluster centroids derived from the analysis. A Jupyter notebook is provided to reproduce all analyses and numerical results. These resources are available on Zenodo under the accession number 10.5281/zenodo.10797206, and the analysis code can be directly consulted and downloaded from the CompMedChemLab repository on GitHub at github.com/CompMedChemLab/project_riboflavin. The other two chapters describe projects that have only recently reached completion. For this reason, the corresponding input data and analysis workflows are not yet publicly accessible, since the related manuscripts are in preparation. After publication, these materials will be released on Zenodo and GitHub following the same policy described above. Input files for enhanced sampling and reweighting, including PLUMED input files, will also be deposited on PLUMED Nest.[278]

Acknowledgements

Even though the title of this chapter is in English, as is the rest of the manuscript, the following section will be written in Italian as it is the mother tongue of all the people mentioned below.

Voglio iniziare con un sincero ringraziamento al Prof. Matteo Masetti, che in questi anni mi ha sempre accordato fiducia e garantito un costante supporto scientifico. Senza la sua guida questo percorso non sarebbe stato possibile. La libertà che mi ha concesso, unita alla sua presenza nei momenti chiave, ha rappresentato per me un punto di riferimento fondamentale. Insieme a lui desidero ringraziare tutti i membri ed ex membri del CompMedChemLab e, in particolare, il mio co-supervisor, Dr. Mattia Bernetti, il Prof. Federico Falchi e il Dr. Riccardo Ocello. Ognuno di loro mi ha insegnato moltissimo, non soltanto dal punto di vista scientifico, ma anche umano. Il confronto quotidiano, le discussioni e la disponibilità costante hanno contribuito in modo determinante alla mia crescita.

Una delle cose che racconto spesso è la fortuna che ho avuto durante il mio periodo all'estero. Una fortuna che, in realtà, non è stata casuale, ma è dipesa da chi mi ha accolto e sostenuto, affrontando insieme a me anche le complessità della burocrazia. Per questo desidero ringraziare il Dr. Massimiliano Bonomi. I cinque mesi trascorsi nel suo laboratorio sono stati fondamentali per la mia formazione scientifica, ma soprattutto mi hanno aiutato a comprendere quale tipo di ricercatore vorrei diventare. Se nel mondo della ricerca ci fossero più Matteo e più Max, sarebbe sicuramente un ambiente migliore, non solo dal punto di vista scientifico.

Il percorso di dottorato è stato lungo e non sempre semplice. Fin dai primi mesi ho avuto la fortuna di dividerlo con persone eccezionali che, prendendo in prestito un'espressione dai ringraziamenti nella tesi di uno di loro, voglio definire la mia "compagnia dell'anello" nella ricerca. Luca, Riccardo, Stefano e Carlo hanno rappresentato un punto fermo in mezzo agli scossoni che inevitabilmente accompagnano questo percorso. Il vostro impegno, la passione per la scienza e per la conoscenza, l'immancabile

sostegno davanti alle difficoltà, così come la sincera felicità per i traguardi di ognuno, hanno reso questo cammino più leggero e bello.

Il periodo a Parigi ha rappresentato un'esperienza straordinaria sotto molti punti di vista. Tante persone hanno avuto un ruolo importante, dai colleghi in laboratorio alle persone con cui ho condiviso la quotidianità fuori dal lavoro. Sarebbe impossibile fare un elenco completo, ma voglio ringraziare in particolare i due Francesco, per le serate passate a parlare, le birrette sulla Senna e le esperienze incredibili che hanno reso possibili.

Bologna, in questi anni, mi ha dato moltissimo, così come le persone che ho incontrato lì. In primis voglio ringraziare Paola, che mi ha accompagnato lungo l'intero cammino, condividendo incertezze, paure e fatiche. Non credo che queste siano finite, anzi forse sono appena iniziate, ma insieme ogni difficoltà diventa affrontabile. Tra le tante persone che hanno reso questi anni incredibili, alcuni non possono non essere citati. Stefano, Jarno, Ludovica, Enrico e Santo sono tasselli fondamentali della mia vita. Ognuno di loro ha contribuito a costruire ciò che sono oggi.

Infine, arrivo alle persone più importanti, che aprono e chiudono simbolicamente questa tesi. Grazie infinite alla mia famiglia. A mio fratello Fabrizio, probabilmente la persona più diversa da me che riesca a immaginare, eppure una fonte costante di ispirazione per la sua determinazione. A mia mamma e a mio papà va tutta la mia riconoscenza. Non esistono parole sufficienti per descrivere ciò che mi avete dato, né per esprimere la gratitudine che provo nei vostri confronti, né per il sostegno che non mi avete mai fatto mancare. Siete, e sarete sempre, un modello a cui aspirare.

Grazie a tutti,

Stefano

:wq