ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

DOTTORATO DI RICERCA IN
DATA SCIENCE AND COMPUTATION

Ciclo 36

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

Settore Scientifico Disciplinare: CHEM-07/A - CHIMICA FARMACEUTICA

Application of advanced computational methods in drug discovery for targeting RNA

Presentata da: Riccardo Aguti

Coordinatore Dottorato
Prof. Daniele Bonacorsi

Supervisore
Prof. Andrea Cavalli

Co-supervisore
Dr. Mattia Bernetti
Prof. Matteo Masetti
Dr. Sergio Decherchi

Esame Finale Anno 2025

Who's afraid of little old me? You should be

Taylor Swift

Index

1 Abstract

This thesis presents a novel computational approach to RNA-targeted drug discovery, addressing the challenges posed by RNA's inherent flexibility and the limitations of traditional protein-docking protocols. The first part of the research focuses on two key aspects: druggability prediction and allosteric analysis. We introduce a one-class learning approach using the Import Vector Domain Description (IVDD) algorithm with customized DrugPred descriptors on pockets identified by NanoShaper. This method, validated on a dataset of 100 proteins from the Potential Drug Target Database (PDTD), offers a more nuanced and efficient approach to identifying druggable pockets compared to traditional binary classifications. While the investigation of allostery compares three computational methods – DyNet, DF, and Pocketron – across three pharmaceutical targets: the adenosine A2A receptor, androgen receptor, and EGFR kinase domain. Pocketron consistently demonstrates great performances in identifying known allosteric pockets with high correlation to the orthosteric site. Applying our refined protocols on proteins to the long non-coding RNA MALAT1, we used NanoShaper and Pocketron to identify potential target pockets that could disrupt the triple helix structure through long-range communication. Once the sites were defined we employ molecular dynamics simulations (unbiased and enhanced) to generate a comprehensive conformational ensemble. Having defined the ensembles we generated poses using two pose generation software (AutoDock GPU and rDock), we then evaluated various scoring functions (AutoDock, rDock, Vina, AnnapuRNA, and SPRank) for their ability to predict experimental binding affinities of diminazene-based ligands to MALAT1. While most scoring functions show limited correlation, AutoDock demonstrates promising results in (at least) distinguishing between high- and low-affinity ligands. Finally, we extend a non-equilibrium binding free energy estimation method to RNA molecules, focusing on the Riboswitch-preQ1 system. Using steered molecular dynamics and the Crooks Fluctuation Theorem, we calculate binding free energies for complexes with both cognate and synthetic ligands. Our results highlighted the importance of protonation states and unbinding pathways in these calculations for accurate results. This research contributes to the advancement of RNA-targeted drug discovery by providing novel computational tools and insights into the complex dynamics of RNA-ligand interactions.

2 Introduction

2.1 Biological and Pharmaceutical relevance of RNAs

Ribonucleic acid (RNA) is a crucial macro-molecule in biology, performing a variety of functions essential for life. Messenger RNA (mRNA) acts as a temporary carrier of genetic instructions [1], while transfer RNA (tRNA) serves as the molecular bridge between these instructions and the amino acid building blocks of proteins [2]. The ribosome, a complex molecular machine composed of both RNA and protein components, orchestrates the translation process, ensuring accurate matching of mRNA codons with the correct aminoacylated tRNAs and facilitating the formation of peptide bonds to build the resulting protein [3, 4, 5]. The discovery of catalytic RNAs, known as ribozymes, unveiled a revolutionary concept: RNA can not only store genetic information, akin to DNA, but also catalyze chemical reactions, much like protein enzymes. These findings reshaped our understanding of RNA's capabilities, establishing its unique dual role in both preserving genetic information and actively participating in crucial biochemical reactions [6]. As the genomic landscape has expanded over the past two decades, RNA's functional diversity has become increasingly apparent, far exceeding its initially recognized roles. Riboswitches in bacteria, for example, play a widespread role in gene regulation, responding to various physiological signals [8]. In more complex organisms, such as eukaryotes, RNA participates in a myriad of processes crucial for maintaining, regulating, and processing genetic information [9, 10]. Noncoding RNAs (ncRNAs), the vast majority of RNA transcripts in human cells, have emerged as key players in regulatory pathways even if they are not translated into proteins 1. Therefore, deciphering RNA function has far-reaching implications beyond fundamental research. For instance, mutations in ncRNAs have been associated with various diseases, notably cancer [11, 12]. Consequently, both riboswitches and ncRNAs are being investigated as promising drug targets, potentially leading to novel therapeutic strategies that could be particularly beneficial in addressing

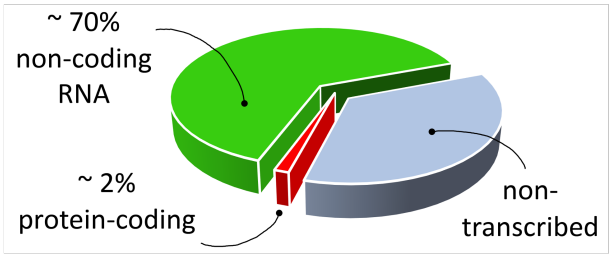

Figure 1: Distribution of RNA transcripts in the human genome. [7]

drug resistance or targeting traditionally undruggable proteins. However, the sheer number of RNA transcripts presents a challenge for experimental methods aiming to decipher the complex relationships between RNA structure, dynamics, and function. Computational tools offer a valuable solution by interpreting existing experimental data, bridging the gap between structure and function, and generating hypotheses for experimental validation. This synergistic approach promises to accelerate our understanding of RNA's diverse roles and therapeutic potential. [13].

2.2 Nucleic Acids: Structure and properties

RNA, a fundamental biological polymer, is constructed from four nucleotide monomers: adenine (A), cytosine (C), guanine (G), and uracil (U). Each nucleotide comprises a flat, aromatic base connected to a ribose sugar, which itself carries a 5'-phosphate group. RNA chains are formed through phosphodiester bonds linking the 3'-carbon of one ribose to the 5'-carbon of the next, resulting in a linear molecule with distinct 5' and 3' ends. The 5' end, possessing a free phosphate group, is considered the starting point of the chain, reflecting its role as the initiation site of RNA synthesis in biological systems (Figure 2).

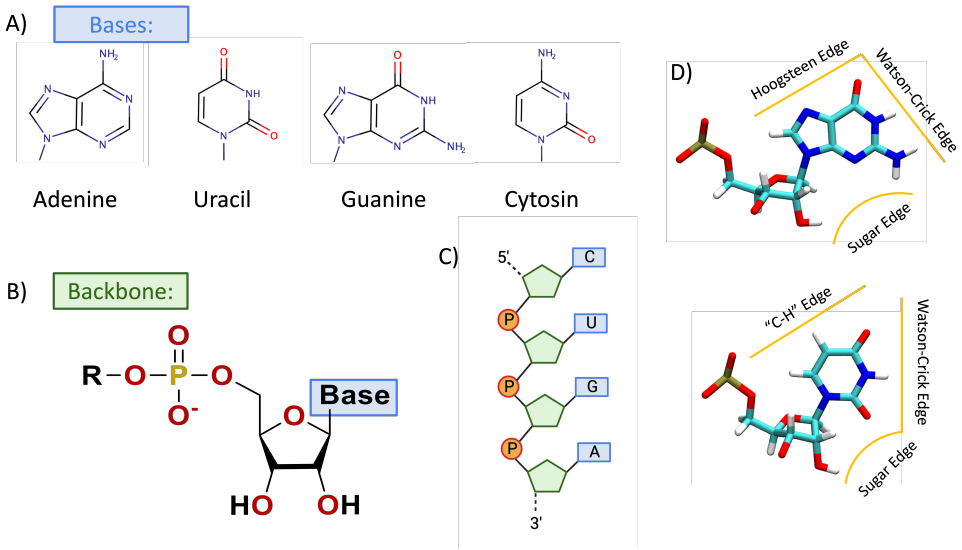

Figure 2: (A-C) 2D representations of the four nucleotides, showcasing the constituent bases (A, C, G, and U), ribose sugar, and phosphate group, and their connectivity in RNA formation. (D) Illustration of potential interaction faces of a nucleotide

RNA molecules can be described through a hierarchical organization of their structure, from the primary sequence of nucleotides to the complex three-dimensional (3D) folds they adopt. The primary structure, a linear sequence of the nucleotides, which serves as the foundation. The secondary structure, often depicted as a two-dimensional (2D) representation, arises from the formation of canonical Watson-Crick (WC) base pairs (A=U and G=C) between complementary nucleotide sequences within the RNA strand. These base pairs are stabilized by hydrogen bonds, contributing significantly to the overall stability of RNA structures (1-3 kcal/mol per base pair). [14, 15] Moreover, RNAs favor short helices, usually containing fewer than 12 consecutive WC base pairs, as longer stretches might be too rigid and stable for functional flexibility that is crucial for their biological role. [13] While RNA is typically single-stranded, it often folds back on itself, creating short antiparallel double helices interspersed with regions of unpaired nucleotides, forming loops in the 2D structure. Non-canonical base pairing plays a crucial role in RNA structure, as many nucleotides that appear "unpaired" actually participate in alternative interactions. These non-canonical interactions occur when hydrogen bonds are formed between different faces of the nucleotides beyond the standard Watson-Crick edge (Figure 2D). Examples include A-G Hoogsteen/Sugar-edge or G-G Watson-Crick/Hoogsteen. These alternative binding modes expand the structural repertoire of RNA beyond classical base pairing.

The tertiary structure, the full 3D conformation of the RNA molecule, is a product of a complex interplay between canonical Watson-Crick base pairs and a multitude of non-canonical base pairing interactions. The unique chemical properties of RNA, particularly the 2'-OH group on the ribose sugar, differentiate it from DNA. This 2'-OH group not only makes RNA more prone to self-cleavage but also acts as a versatile hydrogen bond donor and acceptor, this versatility enables RNA to form intricate and compact structures not observed in DNA. [16] The intricate interactions occur across the sugar, Watson-Crick, and Hoogsteen faces of the nucleotide bases, facilitating the formation of a vast array of base pairing combinations. From an evolutionary perspective, this geometric diversity is advantageous, as it enables RNA molecules to form highly specific interactions with a wide range of molecular partners, including other RNAs, proteins, DNA, small molecules, and ions.

Evolution has harnessed these self-interaction capabilities to generate a remarkable diversity of RNA structures, facilitating countless specific interactions with other molecules, including RNA, proteins, DNA, small molecules, and ions.

The overall shape, or topology, of an RNA molecule is determined by the path its

backbone takes, which in turn depends on the position and orientation of the backbone segments connected to the bases. [16] This intricate relationship links the local geometry of base pairs to the global topology, ultimately dictating the molecule's biological function, emphasizing the inherent flexibility of the target. Even an alteration in a base pair, such as a non-isosteric substitution (a substitution with a base that alters the base pair shape due to its differing geometry), can propagate changes throughout the RNA structure. Two-dimensional (2D) RNA structures typically consist of short canonical helices spaced with segments of nominally "unpaired" nucleotides, often depicted as "loops" in 2D diagrams. These loops, comprising one or more strand segments, can be classified into four main types (depicted in Figure 3):

1. Hairpin loops: Single, continuous strand segments folding back on themselves at the ends of helices.
2. Internal loops: Two strand segments located between two helices.
3. Bulge loops: Similar to internal loops, but with one strand containing unpaired nucleotides while the other is entirely base-paired.
4. Multi-helix junctions: Regions where three or more helices converge.
5. Pseudoknots: base pairing between a hairpin loop or another secondary structure element and a distal complementary strand.

The nucleotides within these structured loops often form numerous interactions, both with each other and with distant parts of the same RNA or other molecules. These interactions make loop regions particularly interesting and functionally important, often forming specific structural motifs crucial for RNA function, further broadening their functional capabilities.

2.3 Drug Discovery campaign: from proteins to RNAs

Rational drug discovery campaigns typically start with a rigorously preclinically validated biomolecular target, whose modulation is anticipated to yield therapeutic benefits due to its pivotal role in a pathological process. Computational strategies are now deeply embedded in the initial hit identification phase, particularly within the structure-based drug discovery paradigm. These approaches include fragment-based methods

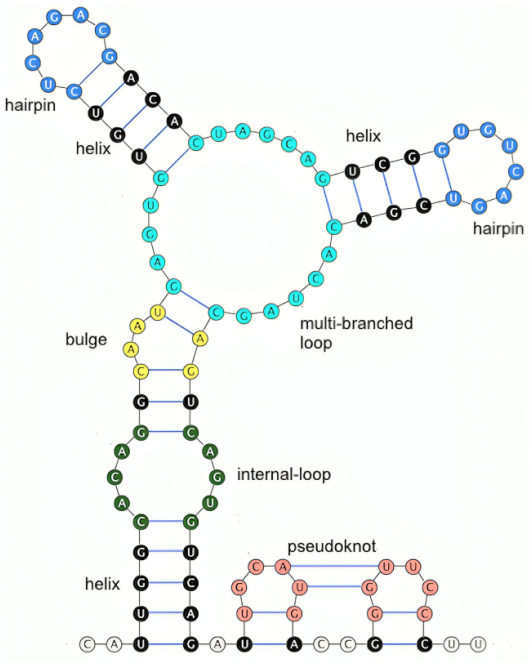

Figure 3: Secondary structure 2D representation of RNAs molecules. [17]

[18], de novo design [19] and virtual screening [20]. In a virtual screening campaign, a vast repository of small molecules is subjected to molecular docking simulations, aiming to predict the binding modes of ligands to the target. This computational methodology, characterized by its rapidity and ability to coarsely discriminate between binders and non-binders, has emerged as a well-established strategy for the identification of small-molecule hits with potential therapeutic relevance. However, adapting these protocols to RNA targets may pose unique challenges, primarily in two critical areas: accurately representing the inherently flexible and dynamic nature of RNA structures, and quantitatively evaluating the resulting binding poses.

These challenges, already significant in protein-ligand docking, are amplified in the context of RNA targets due to their increased conformational variability. Docking campaigns fundamentally rely on a structural representation of the target, such as experimental methods like X-ray crystallography, nuclear magnetic resonance (NMR), and cryo-electron microscopy [21] that can provide atomistic 3D structures. In case these are missing, modelling approaches offer varying degrees of accuracy when experimental data is unavailable [22]. Notably, AlphaFold's machine-learning-based approach has achieved remarkable success in predicting protein 3D structures, but a comparable tool for RNA remains elusive.

Given a target structure, the docking procedure aims to identify plausible binding modes, characterized by favorable interactions between the ligand and target. However, the intrinsic flexibility of RNA presents a challenge, as it needs considering a vast conformational space during the docking process. This, coupled with the need for accurate assessment of RNA-ligand interactions, underscores the complexity of extending docking protocols to RNA systems.

Various software packages have been developed to address these challenges. While tools like Glide [23], GOLD [24], and AutoDock Vina [25, 26] were initially designed for protein targets, they can be adapted for RNA docking when necessary. While, on the other hand, the growing interest in RNA-targeted drug discovery has led to the development of new software such as MORDOR [27], rDOCK [28], RLDOCK [29], and NLDock [30] designed specifically for RNAs. Validation studies often demonstrate that these RNA-specific methods outperform generic macromolecule docking tools in reproducing experimental binding poses, highlighting the importance of considering the unique characteristics of RNA-ligand interactions. The quality of the poses identified by the docking software is typically assessed using "scoring functions". Several standalone scoring functions have been developed, including the RNA specific ones such as the knowledge-based ITScore-NL [31] and the machine-learning-based RNAPosers [32], RNAmigos [33], and AnnapuRNA [34]. These scoring functions aim to quantify the favorability of a given binding pose, aiding in the prioritization of potential drug candidates.

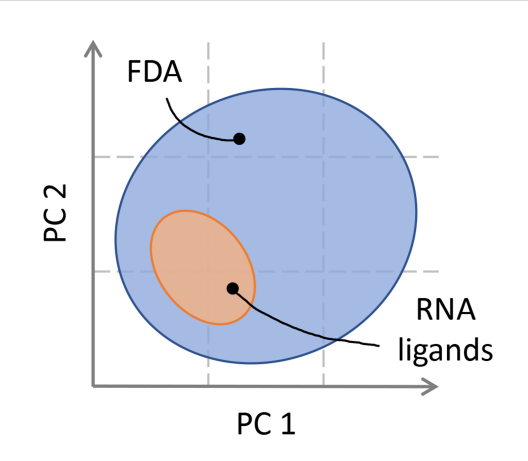

Figure 4: Schematic representation of the chemical space occupied by RNA-binding ligands in comparison to the chemical space of FDA-approved drugs. [7]

In drug discovery, understanding the physicochemical properties required for ligands to bind to specific biomolecular targets is crucial. For protein targets, the focus is on identifying small organic molecules with properties that align with the criteria for oral drugs, such as solubility, bioavailability, permeability, stability, and non-toxicity. Lipinski's rule of five [35] serves as a guiding principle in this endeavor, setting boundaries for parameters like molecular weight, lipophilicity, and hydrogen-bonding capacity, thus defining the chemical space characteristic of small-molecule protein binders. However, for RNA targets, the chemical space of small-molecule ligands remains less well-defined due to the limited number of

known binders, making it an active area of research. Early RNA ligands often exhibited a positive net charge and intercalated between bases [36, 37], resulting in non-specific binding and poor selectivity due to interactions with the negatively charged RNA backbone. However, researchers are gradually uncovering a distinct "RNA-privileged" chemical space defined by characteristic molecular features. The structural hallmarks of this space include an enriched nitrogen content coupled with reduced oxygen presence, alongside predominant aromatic ring systems. These compounds also typically display reduced molecular complexity, as evidenced by their limited number of stereocenters and sp^3-hybridized carbons.[38, 39, 40] Additionally, rod- and planar-like molecular shapes are prevalent among RNA-targeting ligands [41] (Figure 4). Interestingly, this RNA-privileged chemical space appears to be a subset of the broader chemical space occupied by protein-binding ligands, suggesting that some protein-targeting compounds may also possess RNA-binding potential [42].
Despite the valuable insights gained from current knowledge, the limited number of known RNA-binding ligands necessitates continued exploration and refinement of the chemical space. Thus, continued investigation in this area is expected to yield further insights and potentially reshape our current understanding in the foreseeable future.

2.4 Molecular Dynamics simulation on RNAs

While a reliable structure is a crucial starting point, alone it may not fully capture the complexity of the target. Biomolecules are inherently dynamic, exhibiting varying degrees of structural flexibility in solution. This dynamism can influence molecular interactions and even be triggered by the binding of small molecules. Incorporating information about the target's dynamics into docking protocols offers a more realistic depiction of molecular events and can potentially enhance prediction accuracy. Protein flexibility has been addressed in docking protocols through various approaches [43], including soft-docking [44] and induced-fit docking [45], however, these methods typically account for limited structural changes. An alternative strategy involves representing the target as an ensemble of multiple conformations [46, 47], allowing for greater chemical space search, this approach increases computational demands, as docking must be performed for each conformation in the ensemble. In practice, conformational ensembles are generated independently from docking calculations us-

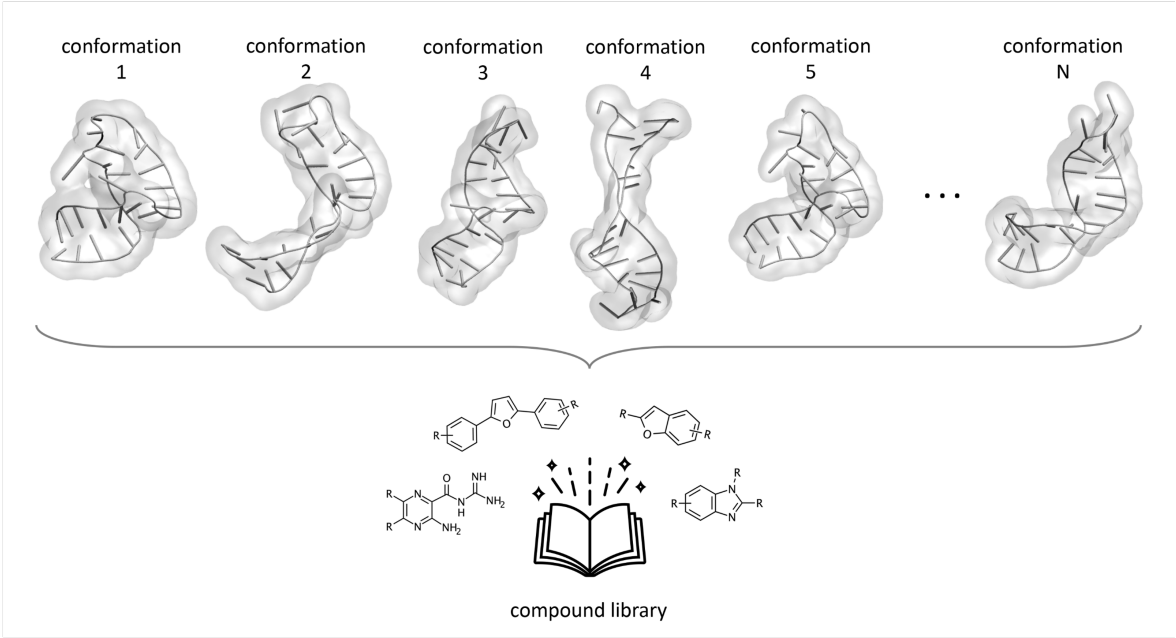

Figure 5: Ensemble docking approach for RNA-ligand interactions. Multiple conformations of the target RNA are included in the ensemble to account for its structural dynamics. Each conformation is subjected to the docking procedure. The RNA structures shown here represent the conformational ensemble of the transactivation response element (TAR) RNA from human immunodeficiency virus type-1 [48].

ing various computational methods. Due to the significant and complex structural dynamics exhibited by RNA molecules compared to proteins [49], ensemble docking, where ligand docking is performed against a collection of target conformations, is the preferred approach for nucleic acids targets. This strategy has demonstrated success in several studies [50, 51]. Molecular dynamics (MD) simulations, which model the dynamic behavior of biomolecules under realistic conditions, are a valuable tool for generating RNA conformational ensembles [13]. MD provides insights into atomistic mechanisms and has become indispensable in structural biology and drug discovery [52, 53]. However, the accuracy of MD simulations heavily relies on the underlying force field's ability to capture the complex physics of molecular interactions. Historically, force field development has been primarily focused on proteins, leaving RNA force fields lagging behind.

Among the Amber force fields, the ff99 force field, combined with the bsc0 and χ_{OL3} refinements, is widely regarded as the gold standard due to its extensive validation and widespread use [54, 55, 56]. Notably, ff99 [54] represents a major advancement in the Amber family, encompassing parameters for both proteins and nucleic acids, and builds upon the earlier ff94 version [57]. Key refinements in ff99 include modifications to the

sugar puckering and χ dihedral parameters for DNA and RNA. Further improvements were introduced with the bsc0 correction [55], which addressed the formation of nonnative backbone conformations and unrealistic helical twists in A-RNA by modifying the α and γ dihedral angles of the nucleic acid backbone. The subsequent χ_{OL3} refinement [56] focused on the χ dihedral to prevent ladder-like, untwisted RNA structures.

However a thorough conformational sampling remains a challenge in both protein and RNA modeling due to the inherent limitations in the timescales accessible through conventional MD simulations. Despite these limitations, significant progress has been made in overcoming the timescale barriers of conventional MD simulations, thanks to advances in hardware [58, 59] and the development of sophisticated enhanced sampling methods [60, 61] implemented in widely used tools such as PLUMED [62]. Additionally, clustering algorithms, now routinely used in MD analysis [63], can effectively select representative structures from MD-generated ensembles for subsequent docking and virtual screening studies. Notably, enhanced sampling methods have been successfully applied to challenging protein systems like intrinsically disordered proteins [64, 65, 66, 67, 68], paving the way for their increased uses in studying the complex dynamics of RNA molecules.

Despite recent advancements, limitations in current force fields, coupled with sampling challenges, can lead to RNA conformational ensembles that deviate from experimental observations. However, incorporating experimental data during MD simulations can guide sampling towards experimentally supported conformations, thereby improving ensemble accuracy. Experimental methods offering coarser structural information, such as small-angle X-ray scattering (SAXS), have proven valuable in this regard [69].

3 Theory

3.1 Molecular Dynamics (MD)

Broadly speaking, MD simulations serve two primary functions: they enable the exploration of theoretical models beyond specific approximations, and they provide valuable insights to experimentalists, guiding further investigations. A significant milestone was achieved in 1964 by A. Rahman [70], who performed the first MD simulation on atoms interacting through using a Lennard-Jones potential to model argon interactions and a finite difference scheme for integration. This pioneering work marked a substantial advancement in the calculation of various dynamic properties, laying the foundation for the widespread application of MD simulations in scientific research today.

3.1.1 Intro to MD simulations

While diverse approaches exist for gaining insights into complex systems, particle simulations necessitate a model to describe the dynamic interactions between the system's constituents. Such model must be rigorously tested against experimental data, ensuring its ability to reproduce or approximate key experimental observations like distribution functions or phase diagrams [71]. Moreover, the model should adhere to fundamental theoretical constraints and conditions, guaranteeing adherence with principles such as energy conservation. In essence, conducting MD simulations requires three key ingredients:

- A model to represent the forces at play between the components of the system under scrutiny, whether they are atoms, molecules, surfaces, or other entities.
- A numerical integrator, which employs a finite difference scheme to evolve trajectories discretely from time t to $t + \partial t$. The time step ∂t is meticulously chosen to balance the integrator's stability, accuracy, and computational efficiency.
- A statistical ensemble that governs the various thermodynamic conditions of the system under investigation, such as pressure (p), temperature (T), volume (V) or the number of particles (N).

This multifaceted approach empowers researchers to simulate and analyze the behavior of complex systems with remarkable precision and detail, offering a powerful tool for

unraveling the intricacies of molecular dynamics. A key factor in simulation studies is the range of time and length scales that can be effectively addressed. Quantum simulations (QM), capturing fast electron dynamics, operate on angstrom and picosecond scales. In contrast, classical MD simulations employ a simplified electronic representation, enabling them to cover longer timescales and larger length scales compared to QM-MD. In classical MD, processes such as intermolecular collisions, rotational motions, and intramolecular vibrations, being considerably slower than electron motions, dictate the dominant timescales. Consequently, trajectory lengths typically reach nanoseconds, and accessible length scales span up until tens of microseconds (depending on the available computer resources).

Today, MM-MD simulations are widely used to investigate diverse problems, including liquid properties, solid defects, fracture mechanics, surface phenomena, molecular clusters, and biomolecules. Their ability to model systems at relevant time and length scales, coupled with the continuous development of computational resources and algorithms, has established MM-MD as a powerful tool in various scientific disciplines. In classical molecular dynamics (MD) simulations, the complex dynamics of electrons are not explicitly modeled through the Schrödinger equation. Instead, electronic interactions are implicitly incorporated into a simplified model, where the nucleus and its surrounding electrons are treated as a single particle, essentially representing atoms as inert spheres. This simplification is grounded in the Born-Oppenheimer (BO) approximation, which leverages the timescales of nuclear and electronic motion to express the system's energy solely as a function of nuclear coordinates. The evolution of the system's state can then be described using classical Newtonian mechanics.

$$\overrightarrow{F_i} = m_i \frac{\partial^2 \overrightarrow{r_i}}{\partial t} \tag{1}$$

Where $\overrightarrow{F_i}$ denotes the force acting on each particle of the system, $\overrightarrow{r_i}$ accounts for the generalized positions of the N particles in the systems (i spans from 1 to N), t is the time and m_i represents the masses of the particles. Moreover, if the particles interact via a potential $U(\overrightarrow{r_i})$, the forces acting on the systems can be expressed as:

$$\overrightarrow{F_i} = -\frac{\partial U(\overrightarrow{r_i})}{\partial \overrightarrow{r_i}} \tag{2}$$

Over the course of the MD simulations, particle positions and velocities are integrated through Hamilton's equations of motion providing insights into system dynamics.

$$\dot{\overrightarrow{p_i}} = -\frac{d\hat{H}}{d\overrightarrow{r_i}} \qquad \dot{\overrightarrow{r_i}} = -\frac{d\hat{H}}{d\overrightarrow{p_i}} \tag{3}$$

In these equations, $\hat{H}$ represents the Hamiltonian function associated with the system. The vectors $\dot{\overrightarrow{r_i}}$ account for the generalized velocities within the n-dimensional system (where the index i ranges from 1 to n). Furthermore, $\overrightarrow{p_i}$ represents the generalized momentum coordinates, and $\dot{\overrightarrow{p_i}}$ denotes their corresponding time derivatives.

3.1.2 Force Fields (FF)

In classical MD simulations, forces are represented using a combination of mathematical functions and parameters, collectively known as a force field (FF). This FF includes information about the system's energy and the forces acting on each particle within various chemical environments regulated by Equation 1. The accuracy of commonly used FFs relies on two assumptions: additivity and transferability. Additivity implies that the system's potential energy is the sum of individual potential energy terms (e.g., bond stretching, angle beinging, electrostatics). Transferability suggests that FF models developed for small systems can be applied to larger systems with similar chemical groups [72], this allows for the simulation of a wide array of molecular systems.
To simplify both computational demands and implementation, the majority of force fields currently employed for molecular systems rely on a pairwise additive approach, decomposing the system's energy into contributions from intra- to inter-molecular forces.

$$V_{total} = V_{bonded} + V_{non-bonded} \tag{4}$$

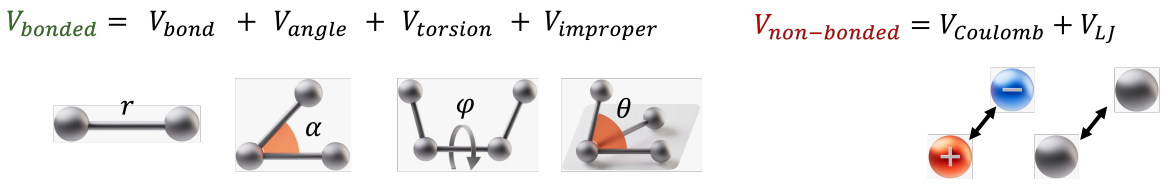

Figure 6: Schematic breakdown of the total potential energy (V_{total}) in a molecular dynamics simulation into its bonded and non-bonded components.

The total potential energy of the system, V_{total}, consists of two main contributions: a bonded component (V_{bonded}) and a non-bonded component ($V_{non-bonded}$), which components are schematized in Figure 6. The V_{bonded} focuses on pairwise interactions between directly bonded atoms.

$$\begin{aligned} V_{bonded} = V_{bond} + V_{angle} + V_{tortion} + V_{improper} = \\ = \sum_{bonds} k_b(b - b_0)^2 + \sum_{angle} k_\theta(\theta - \theta_0)^2 + \sum_{dihedral} k_\Phi[1 + cos(n\Phi - \delta)] + \\ + \sum_{impropers} k_\omega(\omega - \omega_0)^2 \end{aligned} \tag{5}$$

In this formula, the energy associated with bonded interactions (V_{bonded}) is calculated by summing the contributions from bond stretching (V_b), angle bending (V_θ), and proper/improper dihedral torsions (V_Φ and V_ω). Each term involves a force constant (k_b, k_Φ, k_ω) multiplied by the squared deviation of atom positions from their equilibrium. This model, based on the harmonic approximation, is suitable for small deviations. Larger deviations might necessitate more complex expressions like the Morse potential or anharmonic terms. Essentially, intra-molecular interactions are assessed by comparing current and reference nuclear positions, with energy changes reflecting conformational adjustments. Force constants are generally lower for angle bending and dihedral torsions as these deformations require less energy compared to bond stretching.
In contrast to the bonded term, the non-bonded term encompasses the factors that account for potential interactions that extend beyond directly connected atoms within the system. This includes both long-range interactions, as well as crucial electrostatic interactions between charged or polar groups.

$$\begin{aligned} V_{non-bonded} = V_{Coulomb} + V_{LJ} = \\ = \sum_{i,j} \frac{q_i q_j}{\epsilon_D r_{ij}} + \sum_{i,j} 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] \end{aligned} \tag{6}$$

The electrostatic component, $V_{non-bonded}$, within the equation is modeled using Coulomb's law, where q represents the charges on the interacting particles i and j, r_{ij} denotes the inter-molecular distance, and ϵ_D is the inverse of the Coulomb constant $k_e = \frac{1}{4\pi\epsilon_0}$, where ϵ_0 is the vacuum permittivity. However, accurately describing the electrostatic term in simulations poses a challenge due to the lack of experimental observables for atomic charges, which are crucial for representing the molecular elec-

tronic density. For practical implementation and computational efficiency, common approaches typically assign partial charges to atomic sites or nuclei. These charges are derived through various methods such as fitting to experimental data, using atom electronegativities, or performing ab-initio calculations. Coulomb's law is then employed to compute the electrostatic contribution to the total energy.
The Van der Waals term, V_{LJ}, follows a Lennard-Jones 12-6 potential. In this potential, ϵ represents the depth of the potential well (dispersion energy), occurring at a distance of $r = 2^{1/6}\sigma$, while σ corresponds to the distance at which the particle-particle potential energy V_{LJ} is zero. The potential becomes predominantly repulsive when $r < \sigma$, and attractive when $r > \sigma$. While the Lennard-Jones potential offers a reasonable approximation in many scenarios, it's worth noting that for neutral particles, these contributions can also encompass London forces arising from induced dipole interactions or fluctuations in electron distribution, which lead to transient dipole moments that average out to zero.

3.2 Enhanced sampling methods

Within this broader context, it is crucial to recognize the limitations of relying solely on conventional MD simulations to adequately sample the conformational space of complex systems, especially when slow events are involved. To address this challenge, a class of techniques known as enhanced sampling methods has been developed. The unifying principle of these methods, regardless of their specific strategies, is to facilitate a more efficient and statistically accurate exploration of the phase space. Enhanced sampling methods can be broadly categorized into two major subclasses:

- CV-based methods: These methods use reaction coordinates, also referred to as collective variables (CVs), to guide the exploration of phase space. By focusing on specific slow degrees of freedom, CV-based methods can efficiently sample relevant conformational transitions. Notable examples include steered molecular dynamics (SMD), and metadynamics.
- Non-CV-based methods: In contrast, these enhance sampling techniques, they move across all degrees of freedom within the system, without relying on predefined reaction coordinates. Examples include methods such as Hamiltonian replica exchange (HREX).

In this thesis, we strategically employed both CV-based and non-CV-based enhanced

sampling methods to overcome the timescale limitations of traditional MD and gain a deeper understanding of the systems under investigation. The following section provides a concise theoretical overview of the specific techniques employed.

3.2.1 Collective Variables (CV)

Achieving comprehensive sampling of a vast phase space, especially one with high energy barriers, poses a considerable computational challenge. Conventional simulations risk becoming trapped in local energy minima, hindering the exploration of other regions and making accurate free energy surface (FES) estimation difficult. To address this, numerous enhanced sampling techniques have been developed to accelerate sampling and optimize computational resources. Metadynamics and SMD are prime examples of methods that effectively reconstruct complex free energy surfaces (FESs) using appropriate estimators.

However, a significant limitation is the challenge of selecting appropriate collective variables (CVs) to guide the sampling process. On one hand, the chosen CVs need to encompass all relevant slow degrees of freedom to ensure the simulation accurately captures the system's behavior, while on the other hand, too many CVs can lead to impractically slow simulating times. Finding a balance between comprehensiveness and efficiency is crucial. This issue is particularly pronounced in protein-ligand binding simulations, where a multitude of slow processes, such as solute desolvation, ligand conformational changes, and protein residue rearrangements, can influence the process. To address this, the path CV formalism has been developed to streamline the management of high-dimensional phase space and minimize the need for manual CV selection.

Suppose we have a conceptual understanding of a complex reaction pathway and can represent it as a series of intermediate frames. We can then leverage this "guess path" to guide the sampling process using specific collective variables (CVs).

$$S(X) = \frac{1}{P-1}\left(\frac{\sum_{i=1}^{P}(i-1)e^{-\lambda(X-X(i))^2}}{\sum_{i=1}^{P}e^{-\lambda(X-X(i))^2}}\right) \tag{7}$$

$$Z(X) = -\frac{1}{\lambda}ln\left(\sum_{i=1}^{P}e^{-\lambda(X-X(i))^2}\right) \tag{8}$$

For a given microscopic configuration X during the simulation, the collective variable S varies between 1 and P, where P represents the total number of frames in the predefined

frameset. The summations encompass all frames i within the frameset, and for each frame, the squared difference $(X - X(i))^2$ quantifies the distance between the current configuration X and the configuration represented by frame i. Importantly, the choice of distance metric is flexible, but mean square deviation (MSD) is commonly employed, resulting in squared distance units. MSD is often preferred over root mean square deviation (RMSD) due to numerical considerations, although they are conceptually equivalent.

When the system's configuration aligns precisely with a specific frame i, all other terms in the summation vanish, resulting in $S(X) = i$. Therefore, S effectively tracks progress along the predefined pathway. The second parameter, Z, functions orthogonally to S, measuring the deviation from the guess path. As the system traverses the pathway, Z allows exploration of neighboring regions within the conformational space. To visualize this, imagine the accessible configurational space as a cylinder, with the axis representing the frameset and Z defining the radius. In the aforementioned functions, λ is a tunable parameter ensuring smooth progression along the path. It is inversely proportional to the average MSD between consecutive frames, with a suggested formula provided as a general guideline.

$$\lambda = \frac{2.3(P-1)}{\sum_{i=1}^{P-1} |X_i - X_{i+1}|} \tag{9}$$

The high-dimensional phase space is simplified into a 2D representation using CVs that track progression along a hypothesized pathway. Combining S and Z allows for flexible exploration around the guess path, enabling the identification of the minimum free energy pathway within the reconstructed FES.

Additionally, a suitable frameset for path collective variables must fulfill certain criteria. Primarily, it should depict a unidirectional progression towards the final state, avoiding any loops or back-and-forth movements. Secondly, equidistant spacing between consecutive frames is necessary, as defined by the metric used to parameterize the pathway. Finally, the number of frames should be carefully chosen to ensure that the distance between them is not too large, as the resolution of the reconstructed free energy surface (FES) will be directly influenced by this spacing.

3.2.2 Steered Molecular Dynamics

In the context of steered molecular dynamics simulations one effective strategy for applying external forces to a protein-ligand complex involves restraining the ligand to a

designated point in space through an external potential, often a harmonic potential. By systematically shifting this restraint point along a predetermined trajectory, the ligand is compelled to move away from its initial binding site within the target. This controlled movement allows the ligand to explore new interactions and potential binding sites along its unbinding pathway. [73].
So a parabolic potential (ΔU) is added to the standard MD potential (U) to bias the system towards exploring a specific region of the phase space. The center of this parabolic potential is dynamically shifted along the desired range of the reaction coordinate ξ.

$$\Delta U = \frac{1}{2} K(\xi - \xi_0(t))^2 \tag{10}$$

In this expression, the center of the parabolic potential, $\xi_0(t)$, is moved at a constant velocity, defined as:

$$\xi_0(t) = \xi_0(0) + vt \tag{11}$$

v represents the constant velocity at which the center of the parabolic potential moves within the collective variable space. Building upon this concept, Park and Schulten [74] developed a groundbreaking theory for extracting the potential of mean force (PMF), or free energy profile, from steered MD simulations. Their work established a crucial link between non-equilibrium processes like SMD and the equilibrium concept of the PMF. This theoretical framework is rooted in the Jarzynski equality [75], a fundamental principle in statistical mechanics, or alternatively, utilizes bidirectional non-equilibrium estimators based on the Crooks Fluctuation Theorem (CFT) [76]. By conducting multiple independent replicas of the same steering process, the free energy landscape can be reconstructed.

3.2.3 Metadynamics

In Metadynamics, a history-dependent bias potential is introduced along specific reaction coordinates, or collective variables (CVs), within the system. These CVs represent slow degrees of freedom and guide the sampling process towards relevant regions of the phase space. By employing CVs, the exploration of the system's high-dimensional phase space is effectively reduced to a lower-dimensional problem. The bias potential added during Metadynamics is expressed through the following function:

$$V_G(q,t) = \sum_{t=0,\tau,2\tau,\ldots} W e^{-(q-q(t))^2/2\sigma^2} \tag{12}$$

in practice, metadynamics involves periodically adding small Gaussian potentials (with width W and height σ) along the chosen collective variable (CV) at regular intervals throughout the simulation. The overall bias potential at a given time t is the sum of all the deposited Gaussians. These Gaussians act as repulsive forces, discouraging the system from revisiting previously explored regions of the CV space, thus promoting exploration of new areas.

For instance, if the simulation starts with the system in a local energy minimum, the bias potential will initially encourage exploration within that basin. This can be visualized as gradually filling the energy basin, as depicted in Figure 7.

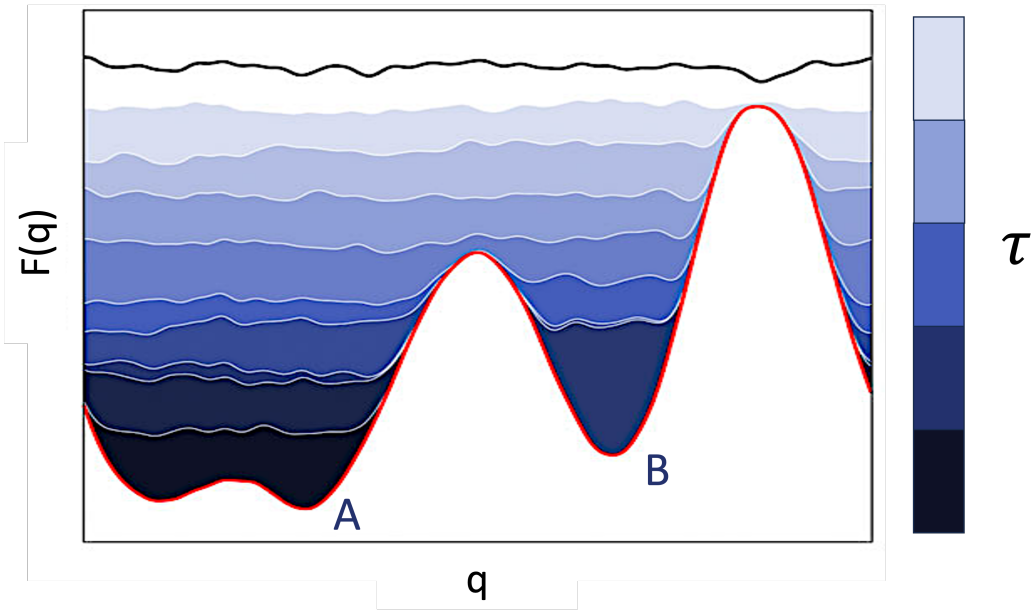

Figure 7: Schematic representation of a Metadynamics simulation. The red line depicts the free energy surface (FES) projected onto the collective variable (CV). Metadynamics progressively deposits Gaussian potentials (dark blue to white) along the CV, gradually filling the energy landscape until the system achieves diffusion across the CV space [77].

Initially, if the system resides in basin A, Metadynamics progressively fills this basin with bias potential, eventually prompting the system to transition into basin B. This process continues, with the bias potential sequentially filling each basin until the entire CV space is explored, enabling free diffusion. Once this state is achieved, the free energy surface (FES) in the CV space can be reconstructed using the accumulated bias potential.

$$V_G(q, t \to \infty) = -F(q) \tag{13}$$

Two major challenges arise when using Metadynamics: assessing simulation convergence and choosing appropriate collective variables (CVs). Determining when to stop the simulation is not straightforward, in standard metadynamics, continuous deposition of bias potential after all basins have been explored leads to overfilling of the FES and exploration of high-energy regions, potentially compromising the accuracy of the reconstructed FES. Well-tempered Metadynamics addresses this issue by gradually reducing the bias potential over time, mitigating the overfilling effect. The Gaussian height becomes a function of simulation time, according to the following equation:

$$W(t) = W_0 e^{-V_G(q,t)/k_B \Delta T} \tag{14}$$

where the initial Gaussian height (W_0) is gradually reduced based on the total bias potential deposited at a given time ($V_G(q,t)$). The parameter ΔT represents the upper limit of the temperature range within which the CVs are sampled. In essence, the gradual addition of bias potential in standard Metadynamics is analogous to sampling at progressively higher temperatures. With well-tempered Metadynamics, each time the system enters a new basin, the Gaussian height is reset to W_0, and the time-dependent scaling restarts. This approach leads to smoother convergence of the bias potential over time.

The second challenge lies in identifying a suitable set of collective variables (CVs), to ensure convergence and avoid non-physiological behavior, the chosen CVs must encompass all relevant slow degrees of freedom within the system. Failure to account for these crucial variables can hinder the simulation's ability to accurately explore the conformational space and converge.

3.2.4 Hamiltonian Replica Exchange

Let's consider a system with coordinates r and potential energy $U(r)$, constructed as a sum of few-body terms, as is typical in atomistic biomolecular modeling [78, 57]. This system is assumed to be in thermal equilibrium at temperature T, with the probability of exploring a specific configuration given by the Boltzmann distribution.

$$P(r) \propto e^{-\frac{U(r)}{k_B T}} \tag{15}$$

Replica exchange methods generally involve sampling one "cold" replica, from which unbiased statistics are extracted, and several "hot" replicas to accelerate sampling.

The 'hottest' replica should efficiently overcome barriers relevant to the process being studied, while intermediate replicas bridge the gap between the hottest and coldest ensembles. The number of replicas required depends on the temperature difference between the hottest and coldest ensembles. In traditional parallel tempering, "hot" and "cold" refer to physical temperature controlled by a thermostat. However, in the more general Hamiltonian replica exchange (HREX), "hot" replicas can be biased in any way that enhances sampling. It's worth noting that parallel tempering may be less effective for processes hindered by entropic barriers, as their transition rates might not increase with temperature [79]. In the most general formulation, each replica within the system is governed by a distinct Hamiltonian, resulting in simulations at different effective temperatures. The collective coordinates of all N replicas (represented as r_i for the $i-th$ replica) are considered to generate the final product ensemble.

$$P(r_1) \times \cdots \times P(r_N) \propto e^{\frac{-U_1(r_1 - \cdots - U_N(r_N))}{k_B T}} \tag{16}$$

Since ensemble probability depends solely on $U/(k_B T)$, doubling the temperature is equivalent to halving the energy. Scaling potential energy, instead of temperature, offers the advantage of selectively targeting specific system regions or Hamiltonian components for "heating", while however, scaling coupling terms involves some arbitrariness.
This approach partitions the system into designated "hot" and "cold" regions, with each atom assigned to one of these regions. Subsequently, a parameterized Hamiltonian dependent on λ is formulated, incorporating specific scaling factors to modulate interactions within and between these regions and enhance slow dynamic processes.

- Charges in the "hot" region are scaled by $\sqrt{\lambda}$.
- Lennard-Jones parameter ϵ in the "hot" region is scaled by λ.
- Proper dihedral potentials with both first and fourth atoms in the "hot" region are scaled by λ.
- Proper dihedral potentials with either the first or fourth atom in the "hot" region are scaled by $\sqrt{\lambda}$.

This approach focuses on scaling force-field terms contributing to energy barriers (electrostatics, Lennard-Jones, and proper dihedrals). Interactions within the "hot" region experience an effective temperature of T/λ, while those between "hot" and "cold" regions experience $T/\sqrt{\lambda}$. All interactions within the "cold" region remain at the original

temperature T.

It's important to emphasize that the effective temperature is not enforced by a thermostat; rather, simulations and replica exchanges occur under conditions of thermodynamic equilibrium. The scaling parameter λ can take any value between 0 and 1, with 1 representing the reference (unmodified) system. While the code permits setting λ to 0 (equivalent to infinite temperature in the "hot" region), this typically results in low acceptance rates and is therefore not recommended. Additionally, if the "hot" region carries a net charge, the 'hot' replicas will have a different total charge compared to the unbiased replica. This discrepancy is resolved in periodic calculations employing Ewald-like methods [80], as a neutralizing background is implicitly added. Finally, our approach to scaling dihedral parameters ensures consistent treatment of both dihedral potentials and their associated 1-4 interactions.

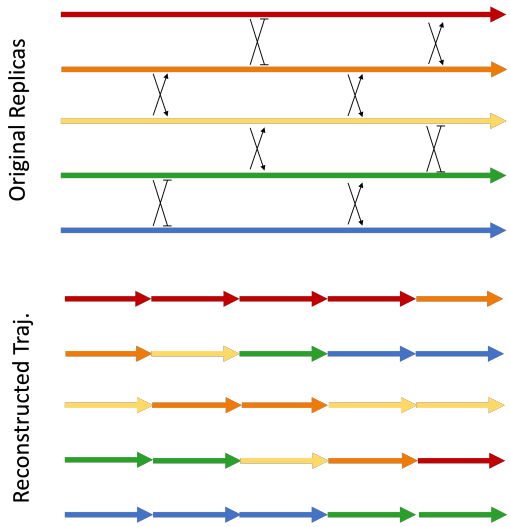

Figure 8: Schematic representation of the Hamiltonian replica exchange algorithm. Above: Arrows represent replicas, with potential exchanges (crosses) subject to acceptance criteria. Below: Reconstructed trajectories after all exchanges.

The Hamiltonian replica exchange algorithm, illustrated in Figure 8, begins with an unconditional exchange of coordinates between replicas at the start of a time step requiring an exchange. The total energy of the system is then calculated using the local force field for the swapped coordinates, this energy is then stored for future use, and the original (unswapped) coordinates are restored through another unconditional exchange. At the end of the number of steps specified, when the actual exchange is attempted, the previously stored energy is employed to determine acceptance or rejection of the exchange [81]. This acceptance criterion is evaluated in a generalized

manner, accommodating replicas with different Hamiltonians.

$$\alpha = min\left(1, \frac{P_i(x_j)P_j(x_i)}{P_i(x_i)P_j(x_j)}\right) \tag{17}$$

In this expression, P_i and Pj is the probability associated with the potential energy of the specific conformation x_i and x_j, which is calculated as the sum of the force-field potential and any additional potentials computed by PLUMED.

3.3 Molecular Docking

Molecular docking simulations aim to predict the binding mode and score of a ligand within a target receptor's binding site. This computationally driven approach plays a crucial role in drug discovery, aiding in the identification and optimization of potential therapeutic compounds [82].

The docking process typically involves two fundamental steps: pose prediction and scoring. Pose prediction focuses on determining the optimal conformation and orientation of a ligand within the binding site. This step is computationally challenging due to the inherent flexibility of both the ligand and the receptor. Efficient sampling algorithms are essential to explore the vast conformational space and identify the most favorable binding pose [47].

Scoring functions are then employed to evaluate the predicted poses and estimate their binding score. Early scoring functions relied on simplified models, primarily considering shape and electrostatic complementarity. While these methods remain valuable for initial screening, more sophisticated scoring functions have been developed to incorporate detailed energetic contributions, such as van der Waals interactions, solvation effects, and even entropic considerations. However, accurately capturing the delicate balance between enthalpic and entropic contributions to binding remains a significant challenge.

Beyond the complexities of scoring, several factors can further complicate accurate docking predictions. These include limitations in the resolution of experimental structures used to define the target, the inherent flexibility of both the ligand and receptor, induced-fit effects upon binding, and the often crucial role of water molecules in mediating ligand-receptor interactions [83, 84]. Addressing these challenges requires careful consideration of the limitations of current docking methodologies and interpretation of the results.

3.3.1 Poses generation

To assess docking methods, it's vital to consider how the target and ligand are represented. The three main representations for the receptor are atomic, surface, and grid [85]. Among these, the atomic representation, due to its computational demands in evaluating pairwise interactions, is typically employed in conjunction with a potential energy function [86] and often reserved for the final ranking stages of docking.
Surface-based docking programs, while often applied to protein-protein docking [87, 88], employ molecular surface representations, largely inspired by Connolly's pioneering work [89, 90], to align points on surfaces and minimize intermolecular angles [91]. However, many of these techniques still rely on a rigid body approximation, which may not fully capture the dynamic nature of the target interactions. Several docking programs employ grid representations for energy calculations. The core idea is to pre-compute and store information about the receptor's energetic contributions at grid points, enabling efficient ligand scoring during docking. Typically, these grid points contain information about electrostatic and van der Waals potentials, simplifying the energy evaluation process.
The treatment of ligand flexibility in docking software can be broadly categorized into three distinct approaches: systematic, random search, and simulation methods. Each of these approaches employs different strategies to account for the conformational flexibility of ligands during the docking process [92].

(i) Systematic search: systematic methods aim to comprehensively explore all conformational possibilities of a ligand, but they face the inherent challenge of combinatorial explosion due to the vast number of potential combinations [93]. To address this, ligands are often incrementally constructed within the active site. This can be achieved through fragment-based docking, where molecular fragments are docked and then linked together, or by dividing ligands into rigid cores and flexible side chains. In the latter approach, the rigid cores are docked first, followed by the flexible side chains, thereby reducing the complexity of the conformational search.

(ii) Random search: stochastic methods, such as Monte Carlo and genetic algorithms, introduce random changes to a single ligand or a population of ligands. The acceptance or rejection of a new conformation is determined by a pre-defined probability function. Tabu search (another stochastic method) keeps track of previously explored conformations, preventing the algorithm from revisiting them and promoting exploration of new areas of the conformational space [94, 95]. The

decision to accept a conformation is based on the root mean square deviation (RMSD) calculated between the current conformation and previously recorded ones.

(iii) Simulation methods: Molecular dynamics, a widely used simulation approach, often encounters limitations in crossing high-energy barriers within practical simulation times, potentially trapping ligands in local energy minima. To address this, researchers sometimes simulate different parts of the receptor-ligand system at varying temperatures [96]. Another strategy involves initiating molecular dynamics calculations from diverse ligand positions. Unlike molecular dynamics, energy minimization methods, which only reach local minima, are rarely used independently but frequently complement other search methods like Monte Carlo [97].

While progress in modeling ligand flexibility has advanced considerably, the treatment of receptor flexibility in docking remains less sophisticated. Nevertheless, various techniques have been employed to introduce flexibility into at least parts of the target [98]. These include molecular dynamics and Monte Carlo simulations [99, 100, 101], rotamer libraries [102, 103], and target conformational ensemble [104]. Rotamer libraries (used in case of protein complexes) model the receptor conformational space using a finite set of experimentally observed side-chain conformations. Another approach is ensemble docking (efficient also in case of nucleic acid complexes [105]), which uses multiple conformations of the target as starting points for docking.

3.3.2 Scoring Functions

Evaluating and ranking predicted ligand conformations is pivotal in structure-based virtual screening. Even with accurate binding pose predictions, the success of virtual screening hinges on the ability to distinguish true ligands from incorrect poses. While free-energy simulation techniques offer quantitative modeling of complexes interactions and binding affinity prediction [106, 107], their computational expense limits their applicability to large-scale virtual screening.

Current scoring functions employed in docking programs often incorporate assumptions and simplifications, neglecting certain physical phenomena that govern molecular recognition, such as entropic effects. Broadly, scoring functions can be classified into three categories: force-field-based, empirical, and knowledge-based, each with its own strengths and limitations.

(i) Force-field-based: molecular mechanics force fields typically quantify the energy landscape of a complex by summing two key components: the interaction energy between the receptor and ligand, and the internal energy of the ligand, which includes factors like steric strain induced upon binding. Ligand-receptor interactions are primarily described using van der Waals and electrostatic energy terms. The van der Waals term employs a Lennard-Jones potential, where the parameters influence the "hardness" of the potential, affecting how close receptor and ligand atoms can approach each other. Higher-order terms, like 12-6 Lennard-Jones, lead to more repulsive potentials, while lower-order terms, like 8-4, result in softer potentials. Electrostatic interactions are modeled using Coulomb's law, but with a distance-dependent dielectric function to dampen the contribution of charge-charge interactions. The internal ligand energy is typically calculated using a similar functional form to the target-ligand interaction energy, incorporating van der Waals and/or electrostatic terms. Conventional force-field scoring functions, while valuable, possess inherent limitations. These functions were initially designed to model enthalpic contributions in the gas phase, neglecting crucial solvation and entropic terms. Additionally, the need for arbitrarily chosen cut-off distances for non-bonded interactions complicates the accurate representation of long-range effects often critical in ligand binding.

(ii) Empirical: empirical scoring functions, first proposed by Böhm [108], aim to reproduce experimental data like binding energies and conformations by fitting a sum of parameterized functions. They operate under the assumption that binding energies can be approximated as the sum of independent terms. The coefficients associated with these terms are determined through regression analysis, employing experimentally measured binding energies and, in some cases, structural data from X-ray crystallography. Empirical scoring functions, although based on similar approximations to force-field functions, often apply simpler functional forms, making them computationally more efficient. Their terms are straightforward to evaluate, but they heavily rely on the specific molecular datasets used for regression and fitting. This reliance leads to varying weightings for different terms, making it difficult to combine them from disparate scoring functions. Consequently, the performance of empirical scoring functions can be highly dependent on the training data, potentially limiting their generalizability.

(iii) Knowledge-based: Knowledge-based scoring functions are primarily designed to re-

produce experimental structures rather than explicitly focusing on binding energies. These functions employ relatively simple atomic interaction-pair potentials, with a variety of atom-type interactions defined according to their molecular environment. Within the realm of knowledge-based scoring functions, a notable subset comprises ML-based approaches. These leverage machine learning algorithms to train scoring functions based on existing data, enabling the assessment of binding poses based on learned patterns of interactions and distances observed in a training set of receptor-ligand complexes. One of the main advantages of knowledge-based scoring functions is their computational efficiency, enabling the rapid screening of extensive compound databases. However, a notable drawback is that their derivation relies on the implicit information embedded within limited sets of complex structures, which can potentially introduce biases and limit their generalizability.

Given the inherent limitations of individual scoring functions, a recent trend in the field has been the adoption of consensus scoring schemes [109]. These schemes leverage information from multiple scoring functions to mitigate errors [110] associated with any single function, thereby enhancing the likelihood of identifying true ligands. However, the potential benefits of consensus scoring might be limited if terms within different scoring functions exhibit significant correlation. Such correlations could amplify calculation errors rather than balance them, potentially hindering the overall accuracy of ligand prioritization.

3.4 Machine Learning

Machine Learning (ML) is a field of artificial intelligence that focuses on developing algorithms that allow machines to learn from data without being explicitly programmed. Instead of following rigid instructions, ML algorithms use data to identify patterns, make predictions, and make decisions. There are several categories of ML algorithms, including:

1. Supervised learning: The algorithm learns from a labeled dataset, where each example has a corresponding label or output value. The goal is to learn a function that can predict the output for new examples. A common algorithms in this category, also used in this thesis, include k-Nearest Neighbors (k-NN), which is a non-parametric classification algorithm that assigns a label to a new sample based on the labels of its k nearest neighbors in the feature space. The k-NN algorithm

aims to classify a data point by finding its closest neighbors in the dataset and assigning it the predominant class label among those neighbors. To achieve this, k-NN relies on a distance metric, which measures how similar data points are, and the value of k, which determines the number of neighbors to consider. Selecting an appropriate distance metric and k value is crucial for the algorithm's effectiveness.

2. Unsupervised learning: The algorithm learns from an unlabeled dataset, where no labels or output values are available. The goal is to discover patterns or hidden structures in the data, such as clusters or associations. Principal Component Analysis (PCA) is a prime example of unsupervised leaning method. PCA is a powerful dimensionality reduction technique, effectively simplifying complex datasets by identifying and highlighting their most significant features. This dimension reduction is accomplished through a linear transformation of the original variables into a new coordinate system of uncorrelated variables called principal components (PCs). These PCs are ordered by the amount of variance they explain in the original data, with the first few components often capturing the majority of the information. This allows for a lower-dimensional representation of the data while preserving its salient characteristics, facilitating visualization, noise reduction, and subsequent analysis. In essence, PCA can be conceptualized as a process of "information compression," where a high-dimensional dataset is projected onto a lower-dimensional subspace spanned by the PCs. This projection maximizes the variance of the data in the new subspace, ensuring that the most important information is retained. By reducing the number of variables needed to represent the data, PCA simplifies analysis and can reveal hidden structures or relationships that may not be apparent in the original high-dimensional space.

3. Reinforcement learning: The algorithm learns by interacting with an environment. It receives feedback in the form of rewards or penalties, and the goal is to learn a strategy that maximizes rewards over time. In particular a one-class learning method learns from a dataset with examples from only one class. The goal is to identify patterns in the data that can be used to distinguish between samples belonging to that class and those that do not. Import Vector Domain Description (IVDD) is a one-class learning algorithm that uses a kernel function to map the data into a high-dimensional feature space, which used is detailed in Chapter 5.2.2.

4 Aim and Objectives

The goal of this project is to develop and validate novel computational methodologies for RNA-targeted drug discovery, with a particular focus on addressing the unique challenges presented by RNA's dynamic nature. Traditional protein-focused drug discovery approaches often fall short when applied to RNA targets, necessitating the development of specialized tools and protocols.

To address these challenges, this thesis is structured into three applications: we begin by developing and validating computational methods to identify druggable pockets and map allosteric communication patterns in protein systems. Building on these validated protein-based pocket trakers approaches, we then conduct a computational drug discovery study on RNA, integrating specialized RNA-specific methods. Finally, we implement non-equilibrium simulations to accurately calculate standard binding free energies in RNA-ligand complexes. The first phase focuses on validating and establishing a robust computational protocol for pocket detection, druggability and pocket communication analysis. This involves evaluating two complementary methodologies: NanoShaper for pocket detection, coupled with a sophisticated one-class learning approach using the Import Vector Domain Description (IVDD) algorithm and customized DrugPred descriptors to assess pocket druggability. In parallel, we validate Pocketron's ability to analyze communication networks between pockets by comparing its performance with other established methods (DyNet and DF) across three pharmaceutical targets. This comprehensive validation demonstrates Pocketron's great performance in identifying known allosteric pockets and their correlation with orthosteric sites. Building upon these validated tools for protein systems, our second objective focuses on their application to the long non-coding RNA MALAT1. Here, we employ the established NanoShaper-Pocketron protocol to identify potential pockets able to accommodate a ligand and their communication along the structure. This phase incorporates unbiased molecular dynamics simulations to capture RNA's conformational flexibility. We then incorporated biased simulation to enhance the conformational search of the MALAT1 in order to construct a conformational ensemble of the target sites identified by Pocketron. We then evaluate various docking approaches (AutoDock GPU and rDock) and scoring functions (AutoDock, rDock, Vina, AnnapuRNA, and SPRank) for their ability to predict experimental binding affinities of diminazene-based ligands. Finally, we aim to extend and validate non-equilibrium methods for binding free energy calculations to RNA systems, using two Riboswitch-preQ1 complexes as a model systems. This

includes investigating the critical role of protonation states and unbinding pathways in achieving accurate predictions through steered molecular dynamics and the Crooks Fluctuation Theorem.

Through these interconnected objectives, our project seeks to contribute significantly to the field of RNA-targeted drug discovery by providing validated computational tools and deepening our understanding of RNA-ligand interactions.

5 Application Case 1: Pocket druggability and allosteric communication [111][112]

5.1 Introduction

5.1.1 Drug Discovery and Target identification

Drug discovery is a complex and time-consuming process [113]. It involves a multistep pipeline from understanding biological mechanisms to fine-tuning the lead candidate (for small molecules), often utilizing computational methods [114, 115]. Over the past 20 years, computation has made significant contributions to many steps in drug discovery through physics-based simulations, machine learning modeling, and their combination [53, 116].

Computer-aided drug discovery and design (CADDD) use information technologies to aid in the identification and development of novel chemical structures with optimal physicochemical and biological properties. This process heavily depends on the structural information of the pharmacological target receptor (direct drug design) or known ligands that bind to these targets (indirect drug design). Computational approaches have become indispensable in contemporary drug discovery, offering essential tools for both the initial identification of promising compounds and the subsequent optimization of their pharmacological and biopharmaceutical properties. [117]. The selection, prioritization, and validation of drug targets pose critical challenges, which are now often addressed through the integration of computational methods in the initial phase of rational drug discovery projects. The process begins by selecting one or more drug targets. If no known ligand binds to the potential target, druggability prediction can be performed, which generally involves analyzing the target surface for binding sites or identifying similar proteins that have already been shown to be druggable [118]. Druggable binding pockets for small molecules in proteins can be identified using structural information, ranging from primary to quaternary structures. Sequence-based approaches, often called "evolutionary algorithms", analyze residue conservation, operating under the assumption that binding residues are essential for functionality and thus likely to be conserved through evolution [119][120]. While advantageous, these methods

often have low accuracy because non-binding residues can also be highly conserved due to other functions. Moreover, allosteric binding sites, which have recently gained significant interest in the drug discovery community [121], are less likely to be conserved across species.

Computational approaches for predicting protein druggability have evolved significantly in recent years. Modern structure-based prediction tools integrate two key components: automated detection of binding pockets and machine learning algorithms that assess the druggability of these sites. Building upon these traditional methods, newer template-based approaches employ sophisticated matching algorithms to compare microenvironments and physicochemical properties between protein pairs. This advancement enables the identification of similar binding sites even among proteins that lack evolutionary relationships [122]. Computational modeling plays a crucial role in identifying potentially druggable targets and pockets that can bind small molecules. A protein is considered druggable if it can be inhibited by a drug, although some experts argue that the term ligandability is more precise for describing a protein's ability to bind drug-like molecules without considering the complex pharmacokinetic and pharmacodynamic factors [123]. In this discussion, we use the term *druggable pocket* to denote a protein region likely to accept a small molecule. Identifying these pockets reliably through computational methods is vital for drug discovery. Discovering new druggable *hot spots* are particularly significant for finding allosteric binders and improving selectivity, which is critical when designing chemical entities such as PROTACs [124][125], where selectivity is often more important than the affinity of the warhead. Although researchers are often aware of a protein's orthosteric pocket, identifying alternative druggable pockets requires a deep understanding of both geometric and chemical properties, making it a challenging task. Thus, effective computational tools are necessary to help medicinal chemists detect and prioritize new pockets to design highly selective drugs.

5.1.2 Computational method for druggability prediction

Numerous studies have explored the computational estimation of druggability [126], offering a range of tools for this task. These tools include standalone software like P2Rank [127] and online platforms such as PockDrug [128]. Typically, prediction methods involve characterizing geometric and chemical features to train and apply

machine learning techniques [129], like for example DrugPred [130]. Additionally, recent deep learning approaches frequently use 3D voxel grids of physicochemical properties for their predictions. DoGSiteScorer [131] is an algorithm that identifies pockets and assesses their druggability by evaluating both global and local pocket properties, with the support of vector machines to construct a predictive model. PRANK [132] employs decision trees and random forests to re-rank and rescore pockets identified by other tools, such as ConCavity [133] and Fpocket [134], potentially enhancing the accuracy of existing prediction techniques. PRANK specifically focuses on predicting the ligandability of particular points near the pocket surface. Additionally, TRAPP [135], known for its molecular dynamics trajectory analysis, has recently been equipped with druggability assessment capabilities, allowing it to analyze entire ensembles of structures.

Our work addresses the challenge of estimating protein druggability with particular emphasis on mitigating the bias. The conventional approach to druggability assessment relies on a binary classification between druggable and less druggable (or nondruggable) pockets, suggesting machine learning classifiers as a natural solution. However, this binary classification presents an inherent challenge: while certain cases, such as extremely small pockets, can be definitively classified as "nondruggable", most binding sites fall into a more ambiguous middle ground making the classification less definitive. In this way, labeling a pocket as nondruggable introduces bias into the model, potentially obscuring the identification of potentially valuable targets. Therefore, we propose that druggability estimation should be treated as a one-class unsupervised learning task rather than a classification task. Building on this observation, we developed a protocol employing the Import Vector Domain Description method (IVDD), a probabilistic one-class nonlinear learner [136][137]. This method constructs a hypersphere that encompasses druggable pockets. To assist the learner, we employed a NanoShaper-based version of the DrugPred [130] descriptors, incorporating some minor adjustments, such as adding the entrance area computed by NanoShaper as an additional descriptor.

From a protein dataset perspective, specific datasets highlighted in the literature could be used as benchmarks. These datasets are frequently employed for training and validating machine learning algorithms, establishing a standardized framework for assessment. In this work, the Non-Redundant set of Druggable and Less Druggable (NRDLD) dataset, introduced by Hajduk et al. [130], was selected for both training and validation. This dataset comprises 113 unique proteins, which were subsequently divided into two subsets: one for training, consisting of 71 druggable proteins, and

another for testing, containing 42 less druggable proteins. Finally, we established a new dataset comprising 100 protein targets to test the method. This dataset is extracted from the Potential Drug Target Database (PDTD) [138].

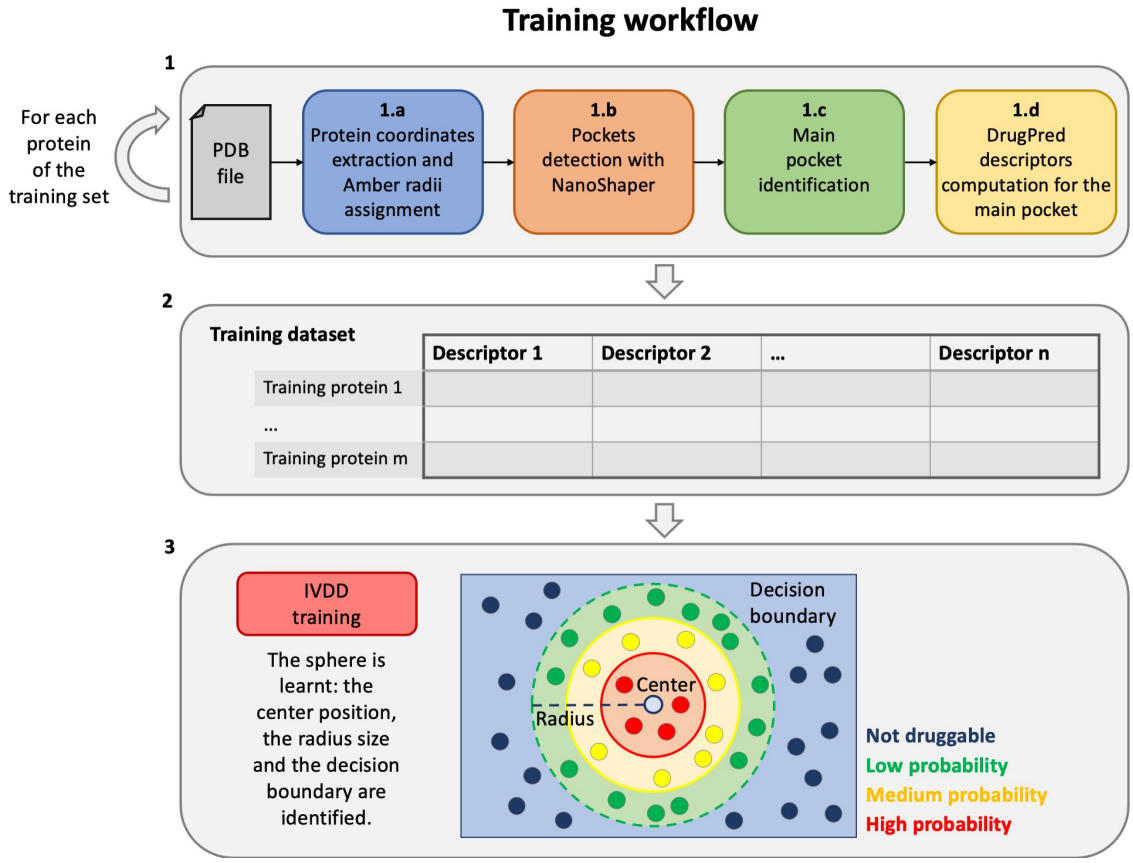

Figure 9: Training workflow [111].

In Figure 9, it is reported the workflow for druggability prediction, distinguishing clearly between the training and testing phases. The training phase, essential for development, consists of three primary steps:

1. Descriptor calculation for training for each protein:
   (a) The protein component is extracted from the input PDB file, and the Amber99SB-ildn force field radii are assigned it;
   (b) The PDB file is processed to obtain a .xyzr file and subsequently processed by NanoShaper to detect all pockets;

(c) A primary druggable pocket is identified for each training protein (the one with a docked small molecule);

(d) The geometric/chemical descriptors for those pockets are calculated.

2. All accumulated data from the previous steps is compiled to construct the training dataset, which includes descriptors specific to each primary druggable pocket identified across the training dataset.

3. Ultimately, the training dataset is employed to train the Import Vector Domain Description (IVDD) machine learning method. During this phase, the model learns a sphere that assigns probability value to each pocket, enabling the distinction between druggable (probability > 0.5) and nondruggable pocket (probability < 0.5).

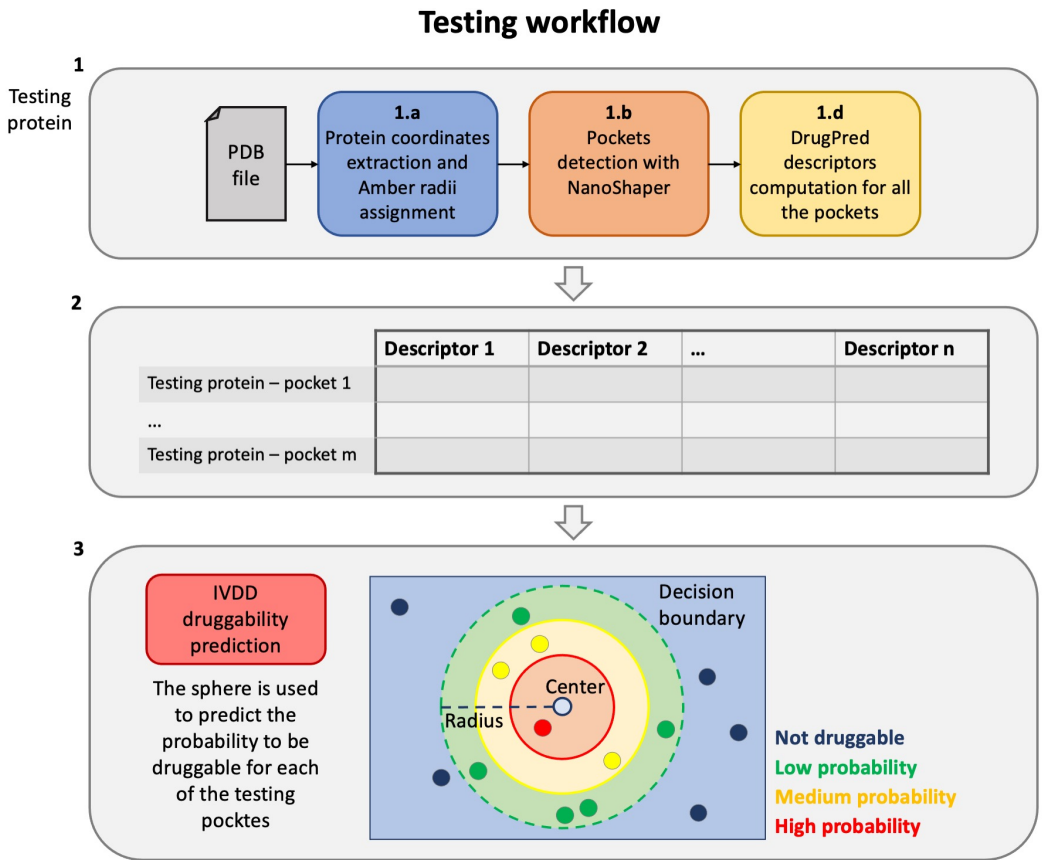

Figure 10: Testing workflow [111].

Conversely, the testing or operative protocol, during which the model is exclusively

used for predictions, involves three steps (Figure 10):

1. Initially, we calculate the descriptors specific to the current target protein. In particular:
   (a) The protein component is extracted from the input PDB file and the Amber99SB-ildn force field radii are assigned to it;
   (b) The PDB file is processed to obtain a .xyzr file and subsequently processed by NanoShaper to detect all pockets;
   (c) The geometric/chemical descriptors for ALL pockets are calculated.
2. All information from the preceding step is aggregated into single file containing the descriptors for each pocket of the current target.
3. Ultimately, the hypersphere estimated during the training phase is employed to predict the probability for each of the newly detected pockets. Pockets with the highest probability are considered most likely to be druggable.

In the Methods section 5.2.1 and 5.2.2, we provide more details regarding the above-mentioned steps.

5.1.3 Dynamic Communication Network

Biomolecular systems rely on coordinated conformational changes, both global and local, to execute their diverse functions. These changes establish intricate communication networks within the systems, collectively known as allosteric regulation [139] [140] [141]. The concept of allostery has broadened over time to encompass any long-range signaling triggered by a local perturbation at a distant site, including protein-protein interactions, covalent modifications, mutations, and the binding of small molecules to sites other than the main active site [140] [142] [143]. Allosteric regulation ultimately modulates the activity of biomolecules, making it a fundamental mechanism in biological processes [139] [143] [144]. Despite its importance, the landscape of allosteric pathways remains largely uncharted [145]. Elucidating these mechanisms holds considerable promise for advancing fundamental biological knowledge, particularly in drug discovery where, the development of allosteric modulators, for instance, can lead to increased selectivity, safer dosages, and reduced risks of toxicity or side effects [146] [147] [148].
However, investigating allostery experimentally poses significant challenges with commonly used methods like X-ray crystallography, NMR, and mutagenesis [139] [145] [149].

Additionally, studies on allosteric mechanisms often involve comparing the structural dynamics of biomolecules in the presence (holo state) and absence (apo state) of the effector that triggers the allosteric response. This scenario also applies to computational studies, where simulations and analysis of a biomolecular target are often conducted both in the presence and absence of ligands [150] [151] [152]. The outcomes of these simulations are then compared to identify the relevant allosteric sites. In a dynamics-based perspective of allostery, structural fluctuations within the biomolecule can facilitate intramolecular communication between different binding sites, even in the absence of ligands [140] [153]. This indicates that, in certain cases, characteristics of allostery may be inherent to biomolecules and can be observed even in their apo forms [154] [155] [156].

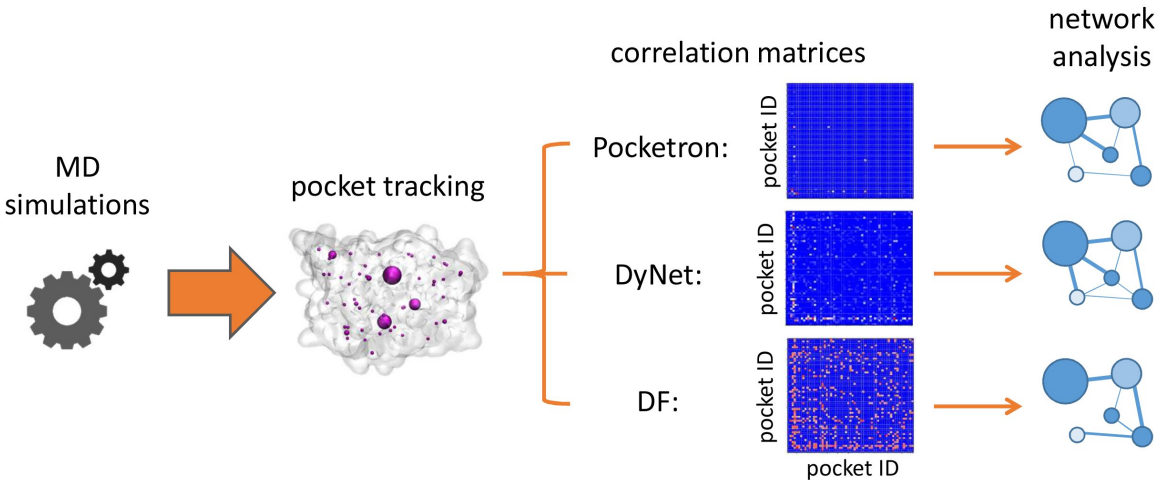

Figure 11: Allosteric pocket crosstalk pipeline [112].

In this work, computational approaches were employed to investigate intramolecular communication of three biomolecules of pharmaceutical interest, known for exhibiting allosteric behaviors in their apo state. Specifically, the effectiveness of three distinct methods was compared across the adenosine A2A receptor (A_{2A}), the androgen receptor (AR), and the epidermal growth factor receptor (EGFR) kinase domain. The adenosine A_{2A} receptor, a transmembrane protein, is implicated in diseases such as inflammation, cancer, and Parkinson's disease [157] [158]. The AR, a nuclear receptor, is essential in the progression of prostate cancer [159], whereas EGFR is involved in signaling pathways that regulate cell growth, differentiation, and survival [160]. Our aim is to assess the extent to which allosteric communication can be extracted from simulations of the apo form of these proteins using three different methods and to compare their

results.
Our methodology combines molecular dynamics (MD) simulations [52] with graph theory for a clearer interpretation of the data. As depicted in Figure 11, we begin with microsecond-long MD simulations. We then identify pockets in the MD-sampled structures using a pocket-tracking algorithm called Pocketron [150]. To further analyze the relationships between these identified pockets, the correlations between the identified pockets are further analyzed using three computational techniques: pocket crosstalk analysis [150], dynamic network analysis [161], and distance fluctuation analysis [152], detailed in the Methods. Finally, to facilitate comparison and understanding, we present the results as network diagrams, highlighting the most significant communication pathways.

5.2 Methods

5.2.1 Pockets Detection and Orthosteric Identification

Detecting all accessible pockets is crucial for assessing the druggability of each pocket in the target protein. To accomplish this, we employ the NanoShaper tool [162] [137], which efficiently identifies these pockets. NanoShaper was selected for its precise estimation of the molecular surface [163], and it triangulates the detected pockets using the same method used for the molecular surface, ensuring smooth triangulated meshes. The identified pockets are stored as mesh files in MSMS or .off format, facilitating easy parsing for subsequent descriptor generation. NanoShaper also computes key properties such as volume, surface area, and lists of constituent atoms for all internal cavities and pockets within the molecular system. This is achieved through a methodical approach involving the volumetric differentiation of regions enclosed by the system's solvent-excluded surfaces (SESs), using two probe radii: a large probe (radius R) and a small probe (radius r) [137]. These probe sizes influence the pocket shapes: a higher R value enhances detection of shallower pockets, while a higher r value smooths out inner surface gaps.
To build our training dataset, we used NanoShaper to automatically detect protein pockets, including the orthosteric one. Since NanoShaper identifies multiple pockets across the protein structure, we used the Jaccard index (J) to identify which can

automatically detected the pocket that best matches the known orthosteric site:

$$J(O, P_i) = \frac{|O \cap P_i|}{|O \cup P_i|} \tag{18}$$

Here, O represents the set of indices for the atoms of the orthosteric site, and P_i denotes the set of detected atom indices in the i^{th} pocket identified by NanoShaper. The orthosteric pocket was defined as the one with the highest Jaccard index relative to reference indices derived from the ligand's surrounding atoms. Moreover, the Jaccard index can be decomposed into two components by multiplying and dividing for O the numerator and denominator separately. In this case we obtain two new metric called in this work: J_{int} and J_{or}

$$J_{int}(O, P_i) = \frac{|O \cap P_i|}{|O|} \tag{19}$$

$$J_{or}(O, P_i) = \frac{|O|}{|O \cup P_i|} \tag{20}$$

These indices reach their maximum values when there is perfect overlap between the reference orthosteric pocket (O) and the detected pocket (P_i). J_{int} measures, in particular, the completeness of the pocket detection (all the atoms of O are identified in the pocket P_i). Meanwhile, J_{or} quantifies the precision of the detection relative to the reference pocket, decreasing when the sets contain non-shared indices (the indices identified in the pocket P_i from Nanoshaper match the one in the orthosteric one). In this work, all three metrics will be used to comprehensively evaluate the quality of pocket detection.

5.2.2 Descriptors Building and Druggability Prediction

Each pocket identified by NanoShaper were analyzed employing the descriptors outlined by Krasowski et al. [130], in conjunction with the data provided by the software itself (Table 1).

The size, shape, polarity, and amino acid composition were computed using NanoShaper output files as input for the descriptor builder. Specifically, to determine volume (vol), total surface area ($area_b$) and entrance area ($area_e$), NanoShaper's calculation were directly employed. The hydrogen-bond donor and acceptor properties (dsa_t and asa_t) were determined by considering the surface areas surrounding all polar atoms. The

Descriptor	**Abbreviation**
Binding site volume	vol
Total surface area	$area_b$
Entrance area	$area_e$
Binding site compactness	$cness$
Relative hydrogen-bond donor surface area	dsa_r
Hydrogen-bond donor surface area	dsa_t
Relative hydrogen-bond acceptor surface area	asa_r
Hydrogen-bond acceptor surface area	asa_t
Relative hydrophobic surface area	hsa_r
Hydrophobic surface area	hsa_t
Relative occurrence of polar amino acids	paa
Relative occurrence of non-polar amino acids	haa
Relative occurrence of multifunctional amino acids	maa
Relative occurrence of charged amino acids	caa
Relative polar surface area ($dsa_r + asa_r$)	psa_r
Incidence of amino acid X in binding site relative to the surface	in_X

Table 1: Descriptors for the characterization of the pockets identified by NanoShaper.

hydrophobic surface area (hsa_t) was calculated as the total surface area minus the combined surface areas of the hydrogen-bond donors and acceptors. Relative descriptors for hydrogen-bond donors (dsa_r), acceptors (asa_r), and hydrophobic surface areas (hsa_r) were obtained by dividing each respective surface area by the total surface area of the binding site. The relative polar surface area (psa_r) was defined as the sum of the relative hydrogen-bond donor and acceptor surface areas. Moreover, to describe the shape of different cavities, they were used the compactness descriptor defined by Krasowski et al. [130], which quantifies how closely a cavity's shape approaches a sphere.

$$cness = \frac{4\pi \left(\sqrt[3]{\frac{vol}{\frac{4}{3}\pi}} \right)^2}{area_b} \tag{21}$$

based on this equation, when compactness approaches 1, the pocket is more spherical. Other descriptors related to amino acid composition were determined by assessing the presence of various classes of amino acids categorized by their physicochemical properties. The proportion of each amino acid group relative to the total number of

amino acids in each cavity (*paa*, *haa*, *maa* and *caa*) were determined.
We employed a specialized one-class learning method called Import Vector Domain Description (IVDD) to identify druggable pockets. IVDD works by mapping training samples into a kernel space, where they are embedded within a hypersphere. In our study, this hypersphere exists in a 35-dimensional space, corresponding to our 35 pocket descriptors. Unlike traditional methods, this allows for the encapsulation of data within complex surfaces that may not be spherical in the original input space. This flexibility is enhanced by a probabilistic model associated with the enclosing surface, which assigns probabilities indicating whether a sample belongs inside or outside the sphere (Figure 12).

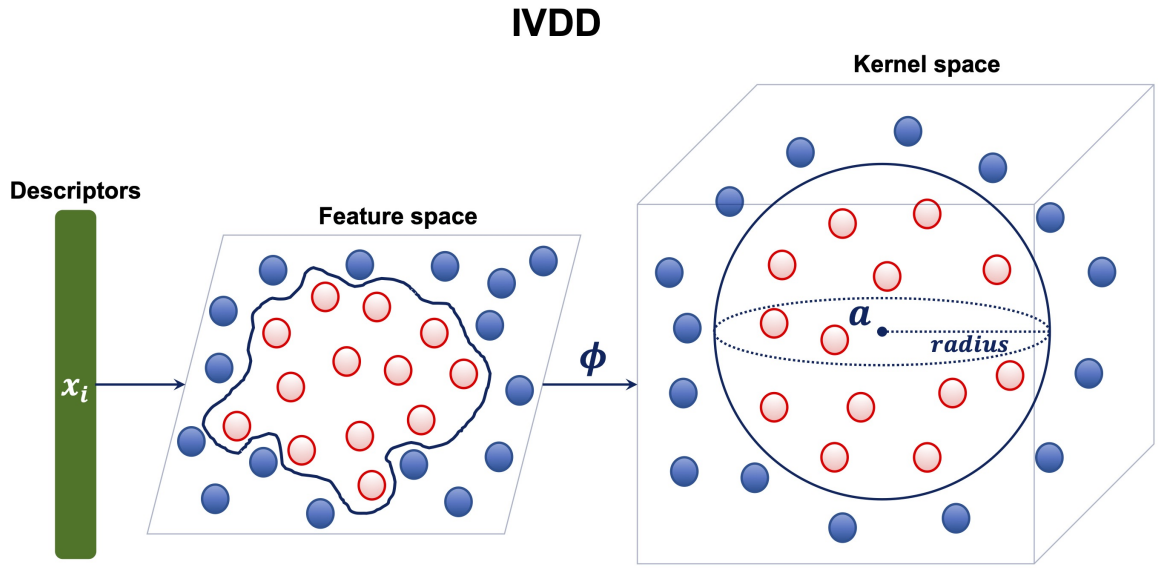

Figure 12: IVDD method: each pocket is a single point in a 35-dimensional space. The mapping between the feature space and the kernel space is given by the function ϕ [111].

During training, the descriptor are used by IVDD to optimizes the configuration of the hypersphere, its center position and radius minimizing a defined cost function.

$$\min_{\Gamma,\mathbf{a}} \Gamma^2 - \hat{C} \sum_{i=1}^{n} \log(p_i) \tag{22}$$

with Γ^2 is the radius of the hypersphere, $\hat{C}$ balances the relationship between the size of the radius and minimizing the error in the model (initially set at 0.5) and p_i the

probability calculated as:

$$p_i = \frac{1}{1 + e^{\beta f_i}} \tag{23}$$

where β is a fixed coefficient (equal to 25 for these analyses) and f_i is the decision function. Equation 22 balances the inclusion of as many samples as possible within the sphere while controlling the sphere's size, potentially allowing some samples to lie outside, with the optimal sphere configuration being unique due to the convex nature of the problem, where at least 80-90% of the data is included in the sphere.
In the testing phases, the determined sphere configuration guides the predictions. Non-druggable pockets are interpreted based on probability values; strictly speaking, a one-class learning assesses the adherence of a pocket to the druggability concept rather than providing a definitive classification. For crisp classification, a probability threshold can be applied: samples outside the sphere are predicted as non-druggable ($probability < 0.5$), whereas those inside are classified as druggable ($probability > 0.5$). Pockets closer to the sphere's center are associated with higher probabilities of being druggable, with probability decreasing towards the sphere's periphery.

5.2.3 MD simulations for pocket crosstalk analysis

Molecular dynamic simulations were performed to study three molecular systems: the adenosine A_{2A} receptor (modelled using PDB IDs *3uzc* and *4eiy* as reported in the literature [150]), the androgen receptor (PDB ID: *1t63*), and the EGFR kinase domain (PDB ID: *2jit*). In the last case the simulation was taken from a previous work [164]. The systems are prepared using BiKi Life Science [165], they were removed any small molecules that were bound to the receptors, but kept the water molecules and ions that were present in the original crystals. The simulations were set up using a standard force field ff99SB-ILDN [166] and the TIP3P model for water [167], each system ended up with around $60000 - 75000$ atoms then, ions (Na^+ and Cl^-) were added to balance the electrical charge of the systems. Finally, the systems were brought to a stable state at 300 K and 1 bar through simulations in the NVT and NPT ensembles. The A_{2A} system, within a POPC membrane, underwent a more extensive equilibration process in the NPT ensemble. This involved 50 ns of simulation with restraints on the protein's heavy atoms, followed by another 50 ns without restraints. Molecular dynamics simulations were then performed on all systems: 3 μs for A_{2A} and EGFR, and 2.6 μs for the AR. Using GROMACS 2021.4 [168] with a 2 fs time step, 50000 frames (approximately 48000 for the androgen receptor) were extracted for analysis

using the Pocketron algorithm [150] in BiKi Life Sciences.

5.2.4 Correlation Matrices

The simulations are then subjected to three distinct analysis methods to generate a correlation matrix. These matrices quantify the degree of communication or interaction between pockets, and subsequently, between different regions of the receptors.

1 Pocketron [150]: Pocketron is an algorithm that tracks the dynamic evolution of pockets in molecular trajectories. It uses NanoShaper to identify pockets frame by frame based on solvent SES as detailed in Section 5.2.1. Each pocket is assigned a unique identifier (pIDs), and the algorithm tracks changes in pocket composition over time. If the newly identified pockets differ from those previously recorded, the new pocket is added to the existing list and given a unique pID. Moreover, it monitors atom movement within and between pockets, identifying "merging" events (atoms from different pockets combining) and "splitting" events (atoms from a single pocket dividing). This analysis of "pocket crosstalk" provides insights into how pockets interact, the output includes pocket IDs with associated atom indices and split-merge matrices representing the propensity for atom exchange. A final, symmetric split-merge matrix is derived for subsequent network analysis.

2 Generalized correlation-based dynamical network analysis [161]: It was also employed a graph-based approach called Dynamic Network (DyNet) analysis to examine the dynamic movement and communication between residues in the MD simulation. The correlations between motions of residue pairs are quantified using a metric known as generalized mutual information, as described by Lange and Grubmüller [169].

$$I[x_1, x_2, \ldots, x_N] = \int p(x) \ln \frac{p(x)}{\prod_{i=1}^{N} p_i(x_i)}\, dx. \tag{24}$$

$$r_{MI}[x_i; x_j] = \left(1 - e^{-2I[x_i;x_j]/d}\right)^{\frac{1}{2}} \tag{25}$$

The mutual information (I) quantifies the pair-correlations between amino acid residues by measuring how their motions deviate from an uncorrelated distribution, while the generalized mutual information is denoted as r_{MI}, where d is the dimensionality of the variables x_i and x_j. It measures the correlation between the motions of the two residues throughout the MD simulation. Then to construct the protein

graph, each node represents the alpha-carbon (C_α) of an amino acid residue, an adjacency matrix (A) is defined as the connections (edges) between nodes based on a distance cutoff, specifically, $A_{ij} = 1$ if the distance between residues i and j is less than or equal to 5.5 Å, and $A_{ij} = 0$ otherwise. A dynamic weighted network is then constructed by assigning weights to the edges. For connected residue pairs i and j (i.e., where $A_{ij} = 1$), the edge weight w_{ij} is calculated as $-log(rMI[x_i, x_j])$, where $rMI[x_i, x_j]$ represents a measure of mutual information between the fluctuations of residues i and j. In this dynamic weighted network, the edge betweenness (EB) of an edge is defined as the number of shortest paths (SPs) that traverse that edge. Edges with high EB values are considered crucial for information flow within the protein, as they mediate a large number of communication pathways between different parts of the protein structure.

3 Distance fluctuation analysis [152]: This analysis method, adapted from Chennubhotla and Bahar's approach [153] by Colombo's lab, investigates signal propagation in all-atom molecular dynamics simulations. An elastic network model is used to analyze protein signal transduction by examining atomic fluctuations during molecular dynamics simulations. The method relies on the distance fluctuation, also known as communication propensity, which is the mean-square fluctuation of the inter-residue distance (d_{ij}) calculated between the alpha-carbons (C_α) of residues i and j.

$$DF = \langle (d_{ij} - \overline{d}_{ij})^2 \rangle \tag{26}$$

In this method, $\overline{d}_{ij}$ represents the average inter-residue distance over the entire simulation. By calculating the distance fluctuation (DF) for each $C_\alpha - C_\alpha$ pair, an NxN DF matrix is generated, where N is the number of C_α atoms. Low DF values often occur for residues within the same structural element (e.g., $\alpha-$helix) and are considered trivial. To avoid these, a double filtration is applied [152][170]: not only are low fluctuating pairs (below a threshold) identified, but a minimum distance threshold for $\overline{d}_{ij}$ is also imposed to ensure that only small fluctuations between distant residues are detected. The DF upper bound is estimated using the average DF value for consecutive amino acids along the sequence, considering neighbors within a range of $i-4$ and $i+4$. This system-specific value reflects local fluctuations, resulting in DF upper bounds of 0.06, 0.04, and 0.10 Å^2 for A_{2A}, AR and EGFR, respectively. The lower bound for $\overline{d}_{ij}$ is set to 8 Å to filter out trivially correlated distances.

5.2.5 Matrices Coarse-Graining and Network Diagrams

Comparing the results of the three different methods used in this work is challenging due to their varying output formats. Pocketron's crosstalk analysis yields an MxM matrix (M being the number of pockets), with each element representing the probability of crosstalk between two pockets. DyNet produces a weighted network with edge betweenness for C_α pairs, while the DF method generates an NxN matrix of pairwise distance fluctuations between all C_α. To unify these representations, we aggregated DyNet and DF data into a pocket-oriented format, emphasizing correlations between whole pockets. This involved dividing the DyNet network into communities corresponding to Pocketron-identified pockets, resulting in an MxM matrix where each element represents the total edge betweenness between residues connecting two pockets. Similarly, the DF matrix was condensed into an MxM matrix, with elements representing average DFs between residues within pocket pairs. When creating the MxM matrices, a challenge arises due to Pocketron setup that allows residues to belong to multiple pockets, reflecting the dynamic nature of pockets, which can merge and split.

As a result, summarizing DyNet and DF data into MxM matrices is not straightforward. To resolve this, when pockets i and j shared residues, the calculation was performed twice: once assigning all shared residues to pocket i, and once to pocket j. For DyNet, the highest value from these calculations was used, representing the strongest correlation. While for DF, the lowest value was chosen, indicating the smallest fluctuation (most coordinated movement) between the pockets. To make the MxM matrix values easier to understand, the DFs values were additionally transformed into a more intuitive distance fluctuation score (DFS), as the DF approach favors lower values

$$DFS = e^{-\beta DF} \tag{27}$$

spanning values in the range [0,1]. To achieve a standardized range of [0, 1], the distance fluctuation score (DFS) values were linearly rescaled to ensure that higher DFS values (closer to 1) indicate stronger correlations. The parameter β acts as the inverse of the distribution's variance, and a value of 10 was chosen to enhance the differentiation of values near 1 by shifting the distribution towards lower DFS values.

To maintain uniformity, the MxM matrices derived from Pocketron and DyNet were also normalized to fall within the range [0,1]. The correlations within the MxM

matrices were visually represented as network diagrams. In these diagrams, each node corresponds to a pocket (identified by its pID), with node size reflecting the average non-zero pocket volume throughout the simulations. The color of each node indicates pocket persistency, defined as the proportion of MD snapshots where the pocket had a non-zero volume. Both pocket volume and persistency data are derived from Pocketron analysis and the edges connecting nodes represent correlations between pockets (the ij element of the matrices), with thicker edges signifying stronger correlations. To enhance the distinction between strong and weak correlations, the edges weights in the network diagrams were also normalized to a consistent range across all three systems. The network diagrams were created using the NetworkX Python package [171] where nodes were placed in a 2D space derived from multidimensional scaling of the 3D pocket center coordinates, preserving pocket distances to aid interpretation.

5.3 Results and Discussions: Pocket druggability

5.3.1 Datasets

In this work, two distinct datasets were employed, each with two variations: one including hydrogen atoms and one excluding them. The first dataset, NRDLD [130], is a publicly available, non-redundant resource for developing and validating structure-based druggability assessment methods. It consists of 113 proteins (71 druggable and 42 less druggable). Each binding site was characterized by 35 descriptors as reported in Chapter 5.2.2.

Additionally, a second dataset was created, comprising 100 proteins sourced from the PDTD (Potential Drug Target Database) [138]. These targets encompass a diverse range of protein types, including enzymes, receptors, antibodies, signaling proteins, and lipid-binding proteins. From these structures, all pockets and their associated descriptors were extracted, resulting in 4807 binding sites without hydrogen atoms and 5692 binding sites with hydrogen atoms, respectively, including 100 orthosteric sites (one per target). The orthosteric site is defined as the pocket hosting the drug or substrate (excluding cofactors), these pockets were characterized as orthosteric (or main) pockets throughout the text, additionally each binding site and pockets were also characterized by the same 35 descriptors used for the NRDLD dataset.

Regarding the pocket detection probes radii, various values for the small and large NanoShaper probes were tested to optimize pocket detection. The small probe was readily set at 1.4 Å, approximating the size of a water molecule however, selecting the optimal large probe value it proved to be more challenging. It was found that A probe size value of 3.5 Å was found to be more effective than 3 Å in identifying both shallow and buried pockets. Larger probe sizes generally led to less accurate pocket shapes, as evidenced by lower Jaccard index values.
For further details regarding the targets in both the NRDLD and PDTD datasets refer to Table 6 and Table 7, respectively.

5.3.2 Train IVDD with NRDLD

IVDD was trained on 71 druggable protein structures from the NRDLD dataset. An RBF kernel was used, with initial $\hat{C} = 0.5$, $\beta = 25$ and an accepted sample range of 80-90% to avoid overfitting. The learning phase adjusted $\hat{C}$ until 90% of samples were within the sphere, resulting in a final $\hat{C}$ values of 0.1 (without hydrogens) and 0.12 (with hydrogens).
In the analysis without hydrogen atoms (Figure 13a), sample 1udt, with a compact and well-defined pocket, received the highest probability score and is closest to the center of the sphere, since IVDD performs optimally when the pocket tightly encloses the bound ligand. In contrast, samples outside the sphere (10% of the dataset) exhibited low scores due to their pocket shapes, especially structures like *1kvo*, *4cox*, and *1k7f* appear as merged pockets, resulting in descriptors that deviate from the druggable reference learned by the algorithm. These structures are then classified as outliers, emphasizing the power that a pre-processing segmentation could have to avoid this limitation. However, IVDD can manage such cases by excluding or marginalizing percolating pockets. Another outlier, *2aa2*, highlights a limitation of NanoShaper, as the identified pocket is too shallow to accurately represent the binding site, leading to a low probability score. This is expected, as NanoShaper's ability to detect shallow pockets is limited by the probe size, which is primarily optimized for deep, buried pockets.
On the other hand, in the analysis with hydrogen atoms (Figure 13b), sample *1xm6* received the highest probability score so, the inclusion of hydrogens made the pocket structure more compact around the ligand, leading to a higher J_{int} value. This improved NanoShaper's accuracy in identifying the orthosteric pocket, resulting in a higher IVDD

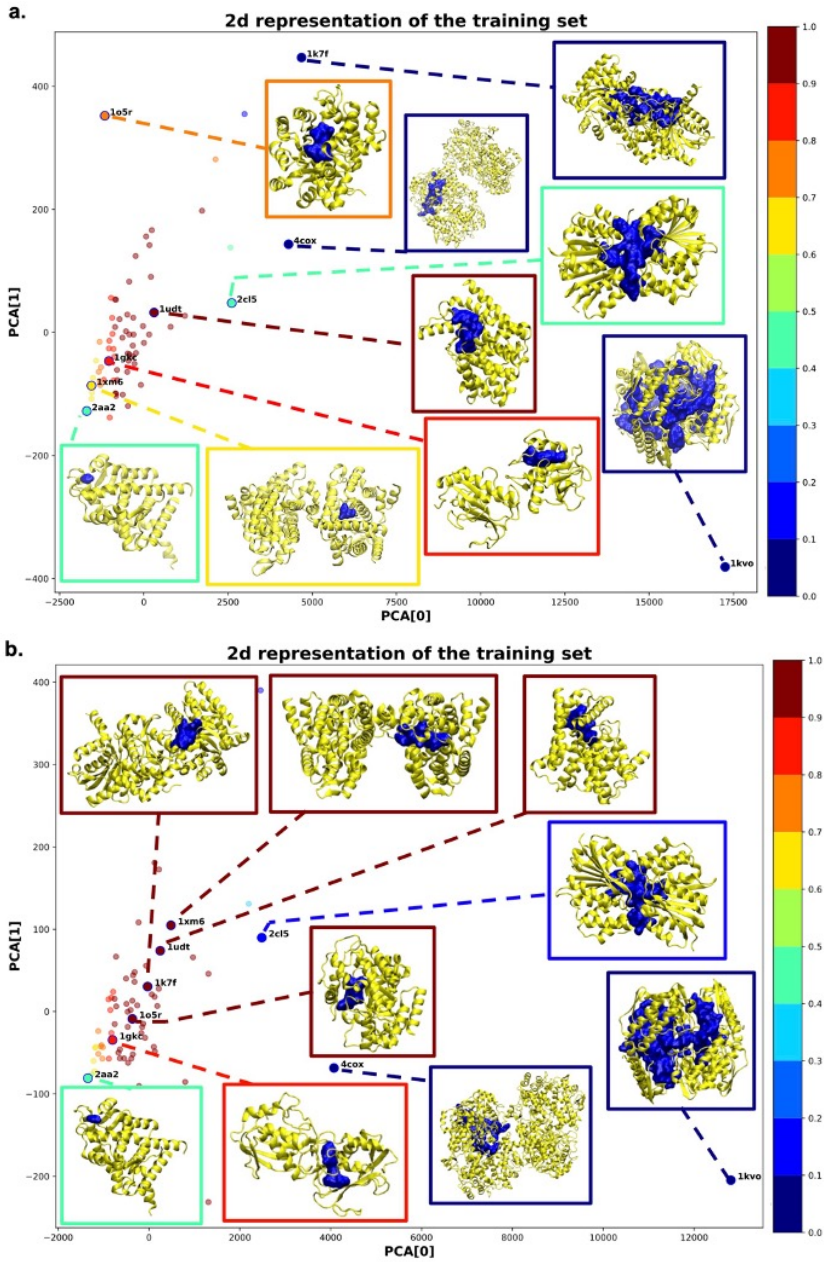

Figure 13: Training results are visualized through dimensionality reduction using PCA. Each point represents a main protein pocket, with color-coding indicating the probability assigned by IVDD. Some 3D structures corresponding to select training samples are displayed. Panel a) shows results without hydrogen atoms, while panel b) includes hydrogens [111].

probability. Similarly, for *1k7f*, the addition of hydrogens closed a channel that previously led to an overly large pocket, significantly improving the Jaccard index for the pocket identification. However, some NanoShaper errors persisted, such as with *1kvo*, *4cox*,

and *2aa2*, where pockets remained either too large or too shallow.

5.3.3 Validation IVDD with NRDLD

Next, the previously trained model was validated using the 42 less druggable structures from Krasowski et al. (2011) [130] to predict their druggability. Figures 14 display the probabilities assigned by IVDD to each structure, for the models without and with hydrogen atoms, respectively.

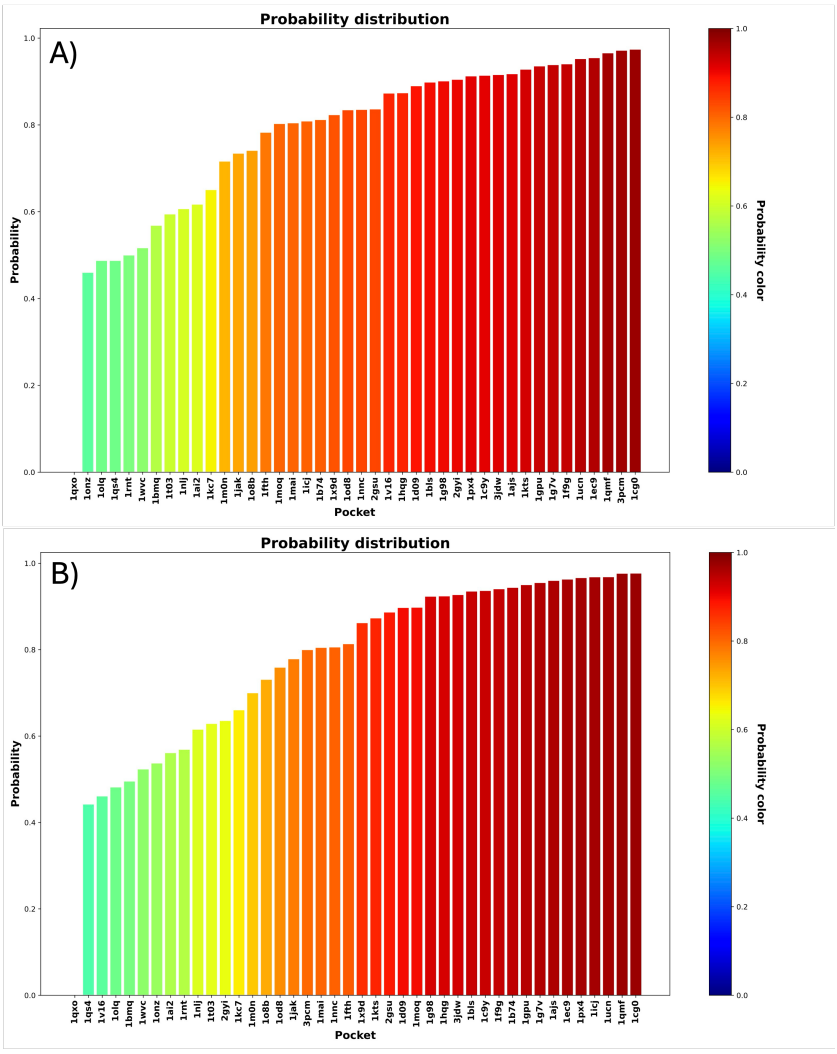

Figure 14: Druggability prediction of the less druggable subset of NRDLD using our model, both without A) and with B) hydrogens. For each protein binding site (shown on the x-axis), the predicted druggability probability is displayed on the y-axis and indicated by the color of the bar [111].

Overall, the results from both models are comparable, in particular, the IVDD model

predicted a probability exceeding 0.8 for roughly half of the "less druggable" set. This suggests a potential bias in the dataset classification (classified as undruggable proteins), indeed this unsupervised approach doesn't impose a pre-defined distinction. Notably, over half of the pockets received high probability scores where, for the remaining cases (low probability), the "less druggable" classification can be attributed to shallow pockets, where NanoShaper's large probe size (3.5 Å) can still allow (in roughly half the cases) for their correct detection.

This finding seems to contradicts the "less druggable" labeling of this dataset, particularly, in the NRDLD the proteins are classified as less druggable if none of its ligands satisfied specific criteria: oral bioavailability (assessed by Lipinski's rule of five), lipophilicity ($clogP \geq -2$), and ligand efficiency (≥ 0.3 $kcal/mol$ per heavy atom). Defining druggability based on specific ligands is problematic because it requires sampling the entire chemical space, which is impractical. Due to limited sampling and biases in the drug discovery process, this classification it could be unreliable while, a pocket's true druggability may be defined by the activity of its most effective ligand within the entire chemical environment. So, this approach avoids pre-labeling and focuses on learning from known druggable pockets.

A systematic analysis reveals a trend where lower probability pockets tend to be smaller, shallower, and have more solvent-exposed ligands compared to higher probability pockets, which are deeper and more compact. This shift is illustrated in Figure 15, showcasing orthosteric pockets of a sub-sample of 1 every 5 receptors of the less druggable proteins (without hydrogens).

For example, *1onz*, with a shallow pocket and partially exposed ligand, scores a low probability of 0.46 conversely, *1cg0*, featuring a well-defined and accommodating pocket, is assigned a high probability of 0.97. This pattern, where lower scores correlate with smaller, less enclosed pockets, holds true for most cases, except for *1qxo*, where the detected pocket is unrealistically large. Moreover, the less druggable set includes interesting cases like *1kts*, *1gpu*, *1ucn*, and *1cg0*, where the ligands are small molecules / small molecule-like ligands where, missing these pockets in drug discovery campaign would be detrimental, but our method scores them highly. This is important beyond the traditional small molecule drug paradigm, as even moderately active warheads can be effective for example in PROTACs or molecular glue degraders.

A technical comparison of pocket probabilities with and without hydrogen atoms reveals interesting findings such as, the inclusion or exclusion of hydrogen atoms does not affect NanoShaper's identification of the main pocket (as indicated by the highest Jaccard

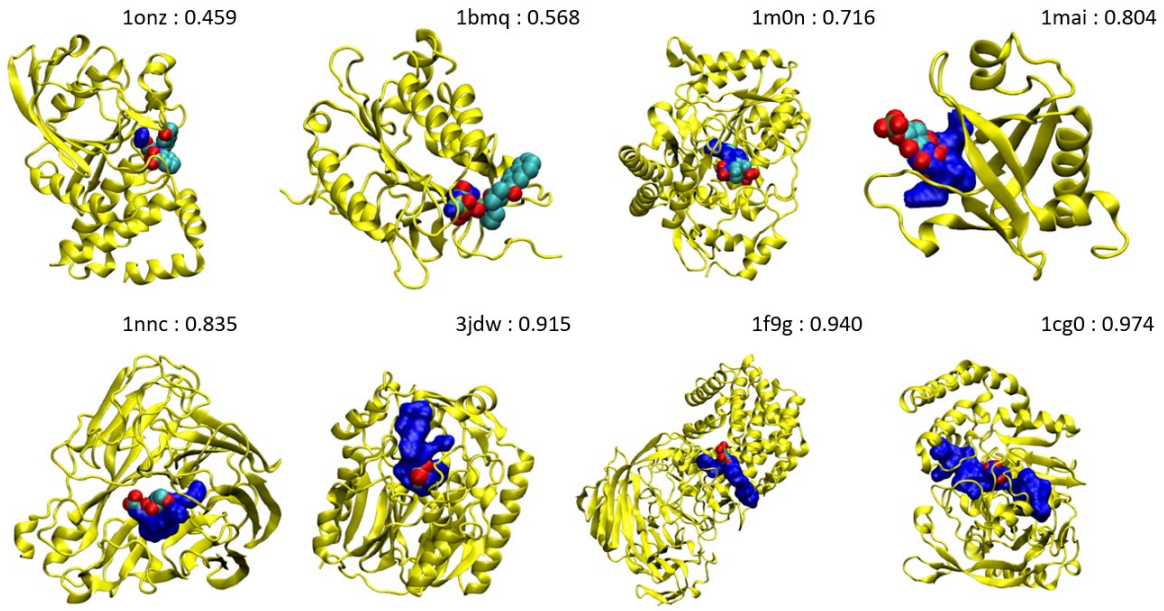

Figure 15: Main pockets (computed without hydrogens) of 1onz, 1bmq, 1m0n, 1mai, 1nnc, 3jdw, 1f9g and 1cg0. The pocket surface is depicted in blue, while the bound ligand from the PDB file is represented in Van der Waals (VdW) style. The accompanying number indicates the calculated probability of the pocket being druggable [111].

index). However, it does influence the pocket's shape and its relative ranking in terms of probability for example, adding hydrogens doesn't necessarily solve NanoShaper's percolation issues, as seen in *1qxo*. Another interesting case is *1icj* (Figure 16) here, the main pocket detection remains consistent but shifts between monomers depending on hydrogen presence.

Importantly, this pocket is always identified as druggable, though with varying probabilities. This suggests that while druggability is a consistent feature, the specific pocket conformation and hydrogen inclusion influence the probability score, therefore, considering dynamic aspects and conformational probabilities is crucial for accurately assessing overall druggability, treating it as a quantifiable characteristic.

5.3.4 Test experiment on PDTD

At this point, the method was tested on a curated 100-protein dataset, a subset of the PDTD, to assess classification accuracy and investigate potential biases. While pocket volume is known to be crucial for druggability, as demonstrated by Nayal et al. [172] using SCREEN obtaining a success rate of 64%. But, relying solely on volume

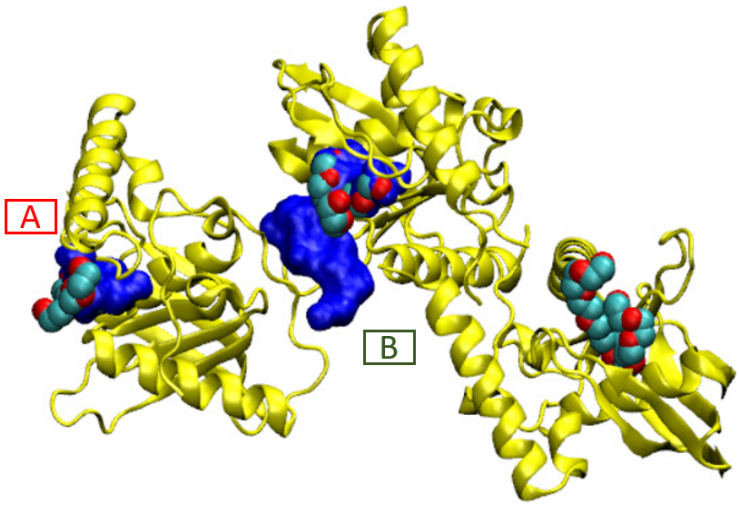

Figure 16: Shift of the main pocket in 1icj with the co-crystallized ligand. (A) Main pocket identified with hydrogen atoms. (B) Main pocket without hydrogen atoms. The orthosteric pocket remains functionally the same but shifts between monomers. All three ligands from the PDB structure are shown [111].

can introduce biases due to various factors, for example, overly large pockets might be mistakenly identified as main sites if they only partially contain the true binding site. This phenomenon that can occur due to percolation effects of the pocket detection engine.

We compared IVDD's performance with a simple ranking based on pocket volume, considering both cases with and without hydrogen atoms and with IVDD algorithm (Table 2 and Figure 17 show the results). While volume ranking achieved a higher top 5 accuracy (97% without hydrogens, 89% with hydrogens), IVDD identified 90% and 92% of orthosteric pockets within the top 5 for the respective cases. This suggests that IVDD, although slightly less accurate in terms of raw top-5 accuracy, excels at identifying high-quality pockets, considering both their correct identification by Nanoshaper and physicochemical properties.

Pocket quality is crucial in both scenarios, adding hydrogen atoms can sometimes fragment overly large pockets, improving accuracy in druggability estimation but potentially leading to tighter shapes due to NanoShaper's behavior. This is demonstrated in Figures 18 and 19, Figure 18 reports the distribution of the three Jaccard indexes (J, J_{int} and J_{or}) for the main pockets of the PDTD subset. While, Figure 19 presents cumulative scores for the top-ranked pockets using the three metrics, comparing volume and IVDD ranking. The results show that IVDD without hydrogen atoms consistently yields higher values across all three metrics. Interestingly, despite lower accuracy than

Description	IVDD	Volume	IVDD+H	Volume+H
Top 1	50	60	50	50
Top 2	67	76	69	65
Top 3	81	87	81	79
Top 5	89	97	92	89
Top 10%	90	97	87	86

Table 2: Outcomes for identifying orthosteric/main sites in the PDTD subset using the IVDD method and a simple descending ranking of pocket volumes, with and without hydrogen atoms.

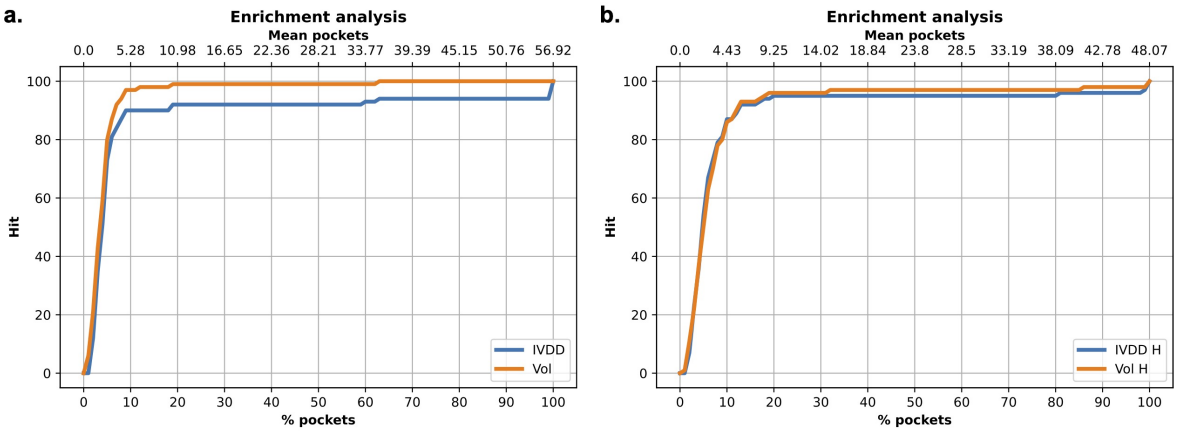

Figure 17: Enrichment analysis of the top 10% highest-probability pockets in the 100-protein dataset, comparing IVDD and volume ranking methods for identifying orthosteric sites. (a) Results without hydrogens show IVDD accuracy of 90% (avg. 5.28 pockets/protein) and volume ranking accuracy of 97%. (b) Results with hydrogens show IVDD accuracy of 87% (avg. 4.43 pockets/protein) and volume ranking accuracy of 86% [111].

volume ranking in this case, the remaining pockets identified by IVDD exhibit higher quality scores. This highlights the risk of over-fitting to volume-based ranking, where a percolating volume may lead (paradoxically) to a misleading 100% accuracy, by not over-fitting, IVDD mitigates this bias and prioritizes pocket quality over quantity. Comparing IVDD results with and without hydrogen atoms provides further insights, many structures not ranked in the top 5 share common features: they either have large pockets with low/intermediate J and low J_{or} values (e.g., *1vkg*, *1qpb*, and *1ht8*), or they are shallow pockets with high J_{or} but low J_{int} scores. In some cases, the presence of hydrogen atoms improves the results, reducing the number of structures outside the top five from 11 to 8. However, for some shared structures (e.g., *1ht8*, *1h9u*, and *1v8b*), the inclusion of hydrogen atoms doesn't significantly change the shape of the orthosteric pocket, leading to only minor changes in probability. These structures exhibit similar characteristics with or without hydrogens, maintaining either their large pocket-size or

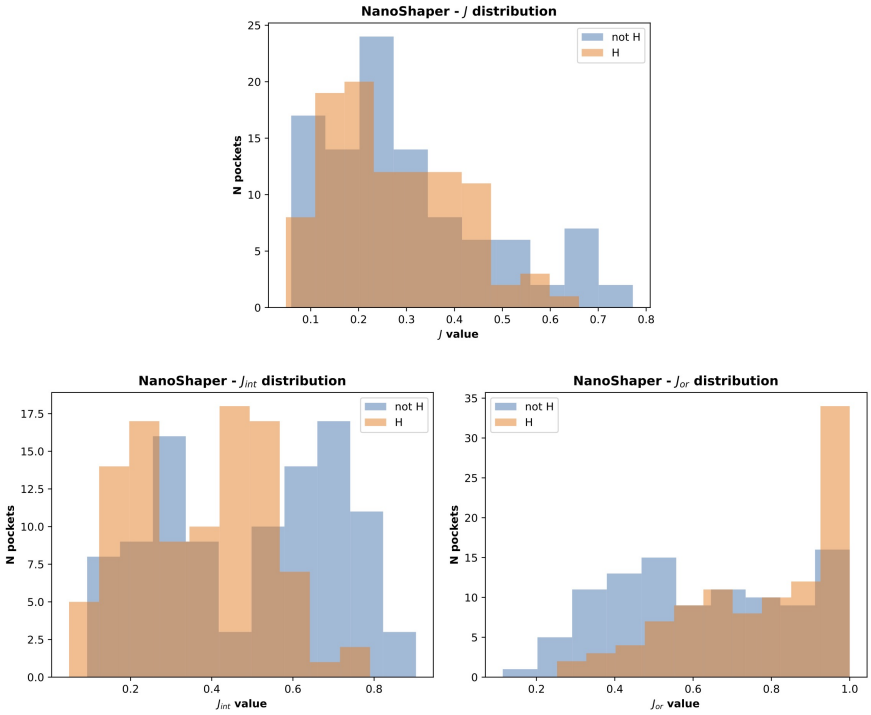

Figure 18: Distribution of NanoShaper scores (J, J_{int} and J_{or}) with and without hydrogen atoms [111].

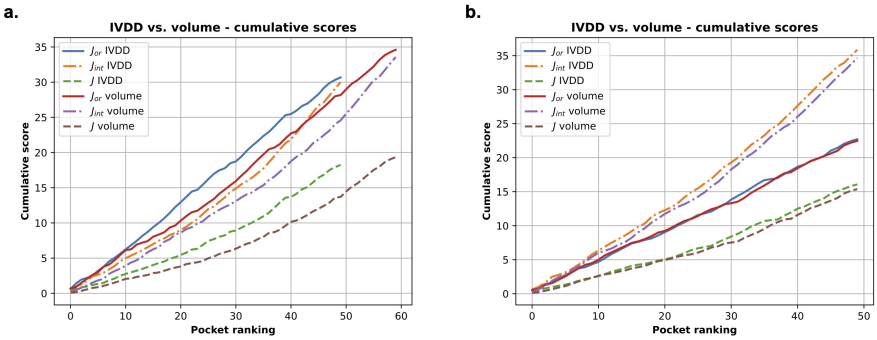

Figure 19: Cumulative scores (J, J_{int} and J_{or}) for the top-ranked orthosteric sites identified by both IVDD and volume ranking. Inset (a) displays results without hydrogen atoms, while inset (b) shows results with hydrogen atoms [111].

shallowness, despite the presence of hydrogen atoms.

The proposed methodology, designed to mitigate biases and enhance robustness, yields comparable accuracy to existing methods in druggability prediction. Specifically, excluding hydrogen atoms from the model, we achieve 81% and 89% accuracy in identifying druggable pockets within the top 3 and top 5 predictions, respectively. Some misclassifications can be attributed to limitations in the pocket detection algorithm,

NanoShaper. Notably, our approach achieves comparable accuracy to established methods such as DoGSiteScorer (88% accuracy) [131], DrugPred (91% accuracy) [130], and fpocket (83% top 3 accuracy) [134]. Overall, the method achieves comparable accuracy to existing methods, but with the added benefit of inherent precautions against biases related to labels and volume.

To better understand the IVDD results, it was investigated the impact of each feature on predictions. Since IVDD doesn't have built-in feature selection, a post-labeling strategy was required. The average probability for orthosteric sites (0.852 without hydrogens, 0.877 with hydrogens) were calculated and used these as thresholds to label binding sites (0 for lower, 1 for higher probability). Moreover, a random forest classifier was employed to estimate feature importance, the classifier employed 100 decision trees (estimators) [173] and utilized the Gini index as the criterion for splitting the data.

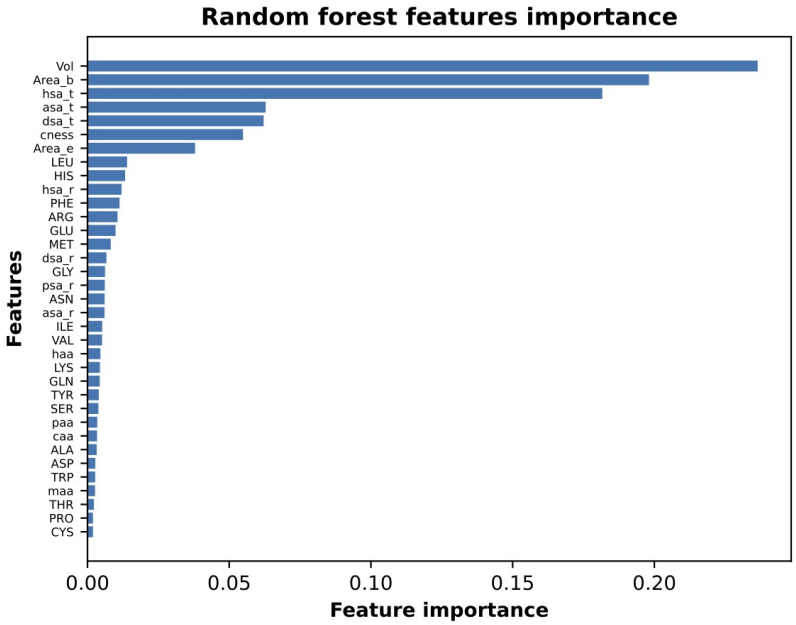

Figure 20: Random forest feature importance in descending order by assigning ex post labels to the IVDD predictions. Results shown are without hydrogens; similar results are obtained with hydrogens [111].

As shown in Figure 12, volume (vol) was the most influential feature, followed by pocket surface area ($area_b$), hydrophobic surface area (hsa_r), hydrogen-bond acceptor/donor surface areas (asa_t and dsa_t), binding site compactness ($cness$), and entrance surface area ($Area_e$). Less influential, but still notable, is the slightly higher ranking of hydrophobic residues (LEU, PHE, MET, and GLY) and some charged residues (GLU

and ARG). The importance of hydrophobic residues and volume is consistent with the chemical understanding of binding pockets.
Finally, looking ahead, it is planned to enhance our methodology through several refinements. Implementing a dedicated volume segmentation algorithm could improve accuracy, especially in selecting the optimal large probe value, offering greater flexibility for this parameter. Segmentation, as demonstrated in Aggarwal et al. [174], can help identify the real pockets (without the percolation effect) that are better suited for virtual screening and docking. Additionally, developing a user-friendly web server would make this tool more accessible to the broader scientific community.

5.4 Results and Discussions: Pocket communication

5.4.1 Convergence of the methods

Molecular dynamics simulations were run on three systems: the adenosine A2A receptor (A2A), the androgen receptor (AR), and the epidermal growth factor receptor (EGFR) kinase domain, generating 3 μs-long trajectories (2.6 μs for AR) without ligands. Pocketron was used to identify pockets within these trajectories, providing a common baseline for subsequent analysis. The residue composition of these pockets was used to generate correlation matrices spanning [0,1] for Pocketron, DyNet, and DF analysis, with 1 indicating maximum correlation. For each system, three distinct correlation matrices were derived, quantifying the inter-pocket communication observed in the MD simulations (Figures S1-S3). Higher matrix elements were assigned to pocket pairs that exhibited: (i) frequent merge and split events (Pocketron), (ii) high edge betweenness, indicative of significant information exchange (DyNet), or (iii) minimal distance fluctuations (DF).
As a preliminary assessment, it was evaluated the convergence of matrix entries for the three methods using the EGFR system as a reference. The Frobenius norm [175] was calculated to measure the difference between matrices at various time points (1.0, 1.5, 2.0, 2.5, and 3.0 μs) and the initial 0.5 μs matrix. The analysis, detailed in the supplementary material (Figures S4-S6), reveals that DyNet converges fastest, stabilizing its values after 1.0 μs. While DF and Pocketron, they require longer simulation times, reaching convergence approximately at 2.5 μs. Notably, Pocketron's convergence is

influenced by the rate of pocket emergence throughout the trajectory, whereas the other methods benefit from the final pocket set identified by Pocketron.
Upon comparing the final results of the three methodologies, a notable disparity in the sparseness of the correlation matrices emerged. The DF method exhibited a generally uniform distribution of signals, with distinct areas of heightened correlation. Conversely, the DyNet approach produced fewer signals, primarily concentrated in specific matrix regions. Notably, Pocketron yielded the sparsest matrices with the fewest signals. This variance in sparsity stems from the different mechanisms employed by each method to assign values to the correlation matrices. Pocketron, for instance, only identifies pocket communication if residues are exchanged between pockets, a stringent criterion that naturally results in sparser matrices. In the DyNet method, betweenness is calculated exclusively for pocket pairs with inter-residue distances below a predetermined cutoff (5.5 Å in this case). This inherently limits the detection of direct communication for spatially distant pockets. Conversely, the DF method directly computes distance fluctuations for every pocket pair, resulting in non-zero values and a denser matrix, even after applying a DF upper bound filter allowing the identification of signals between pockets located further apart.
In the following network diagrams, each node represents a pocket, with size indicating its average non-zero volume during simulation and color representing pocket persistency. The thickness of edges between nodes reflects the strength of correlation between pockets, as determined by the three different methods. For clarity, only the top 30 most relevant correlations are displayed. To facilitate interpretation, the nodes are projected onto a 2D plane using multidimensional scaling based on the 3D coordinates of pocket centers of mass, thus preserving spatial relationships. These results are presented for each protein system individually.

5.4.2 Adenosine A_{2A} Receptor

Following the assessment of general principles and convergence behavior, a detailed analysis of the results obtained for each individual protein system is presented.
Figure 21 illustrates the analysis results for the A2A receptor. The largest identified pocket (pID 4) corresponds to the orthosteric site, deeply embedded within the seven transmembrane helices near the extracellular region. This site, along with pIDs 3, 5, and 6, exhibits high persistency, indicating their consistent presence throughout the simulation. The resulting pocket correlation network offers valuable insights into the

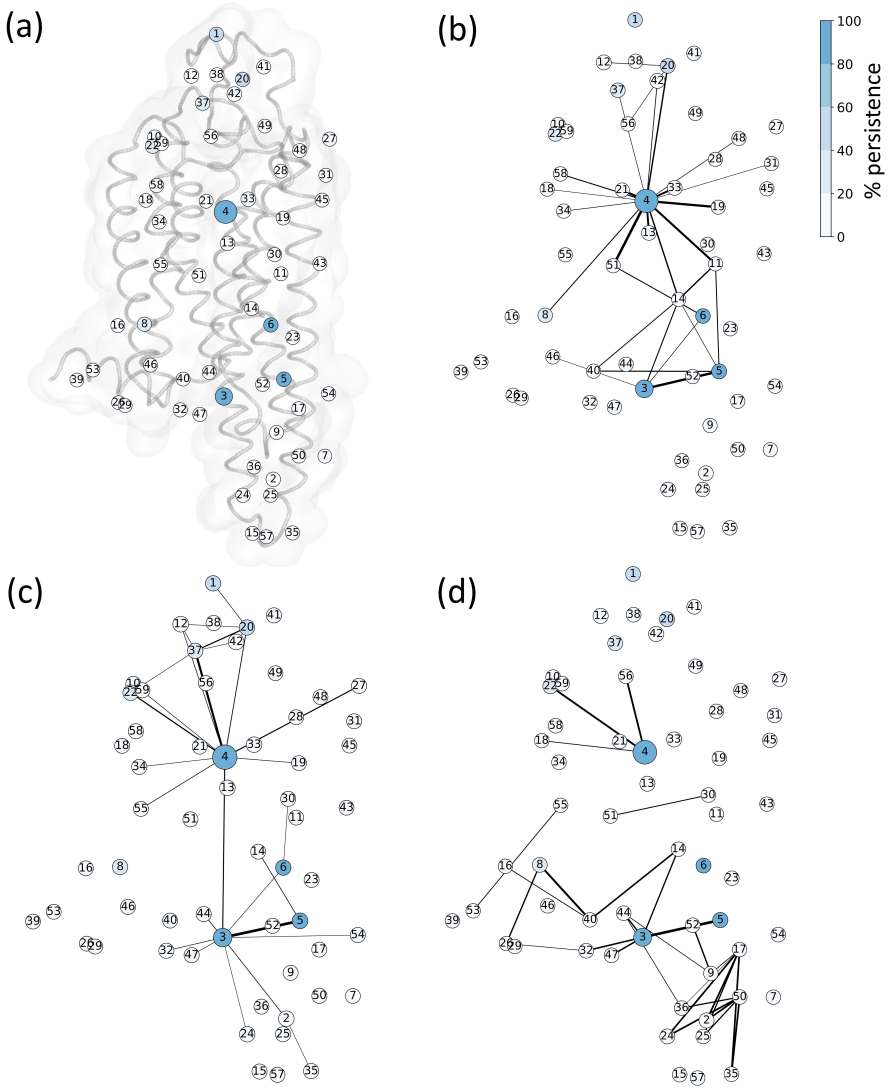

Figure 21: Network representation of pocket communication in the A2A system. Node size correlates with average non-zero pocket volume, color indicates pocket persistency, and edge thickness denotes correlation strength (top 30 shown) as determined by (b) DyNet, (c) Pocketron, and (d) DF analyzes. (a) 2D node topology, obtained via multidimensional scaling, is superimposed on the A2A receptor structure to visualize pocket locations [112].

interconnectivity of different protein regions, revealing potential pathways of communication that may be relevant to protein function and dynamics.

DyNet analysis (Figure 21(b)) reveals extensive connections between the orthosteric site and neighboring pockets, with the strongest communication observed with pIDs 11, 13, 19, 33, 37, and 58. Pocketron results (Figure 21(c)) also validate these findings, highlighting connections between the orthosteric site (pID 4) and surrounding pockets. Notably, Pocketron uniquely identifies a direct connection with pID 3, not mediated

by intermediate pockets, and maintains the connection between pIDs 3 and 5. The DF approach (Figure 21(d)) also identifies connections involving the most persistent pockets 3, 4, and 5, including the conserved connection between pIDs 3 and 5, but does not detect any direct or indirect pathways between pIDs 4 and 3. Additionally, DF analysis reveals a dense network of connections in the lower intracellular region of the receptor.

The DF method, as mentioned, prioritizes connections with smaller distance fluctuations, often within highly structured regions. To focus on longer-range correlations, we filtered out local fluctuations by considering only inter-pocket residue distances exceeding specific cutoffs (Figures S7-S9). For the A2A receptor, filtering out distances < 10 Å revealed long-range correlations between the orthosteric site and surrounding pockets, including a pathway from pID 4 to pID 3 via pID 14, mirroring a pathway identified with DyNet. Increasing the cutoff to 15 Å kept the connection between pID 4 and 3 but made it direct, eliminating intermediate pockets 13 and 14 with closer inter-pocket residue distances. This approach effectively highlights long-range correlations hidden by local fluctuations. Importantly, while the choice of distance cutoffs may appear arbitrary, it was carefully considered in relation to the system size. For example, while AR and EGFR exhibit maximum inter-pocket residue distances of 46 Å and 55 Å respectively, the A2A receptor reaches 76 Å. This wider range in the A2A receptor allows for the investigation of long-range correlations using larger distance cutoffs, which would not be feasible for the smaller systems.

To validate and contextualize our findings, we compared the identified pockets with known hotspots reported in the literature. For the A2A receptor, we observed a striking agreement between the locations of these hotspots according to literature [176] [177] [178] and the pockets identified by Pocketron (Figure 22).

Specifically, we detected pockets in the regions corresponding to the "G Protein-Coupling Site" (pID 3), the "Lipid interface" (pID 6), the "C-Terminus Cleft" (pID 8), and the "Extracellular Cleft" (pID 33). While the "Intracellular Crevice" is known to vary according to the conformation (active, intermediate 1, intermediate 2, inactive), but still it was identified a small pocket (pID 5) within a portion of this region. Notably, the pockets in the "C-Terminus Cleft" (pID 8) and "Extracellular Cleft" (pID 33) regions, visible only in the active conformation, displayed low persistency in our analysis, but still demonstrating a degree of correlation with the orthosteric pocket.

Regarding the sodium ion site, a known hotspot often displaced by allosteric drugs (Figure 22, yellow residues), was also considered in our analysis. Studies suggest that

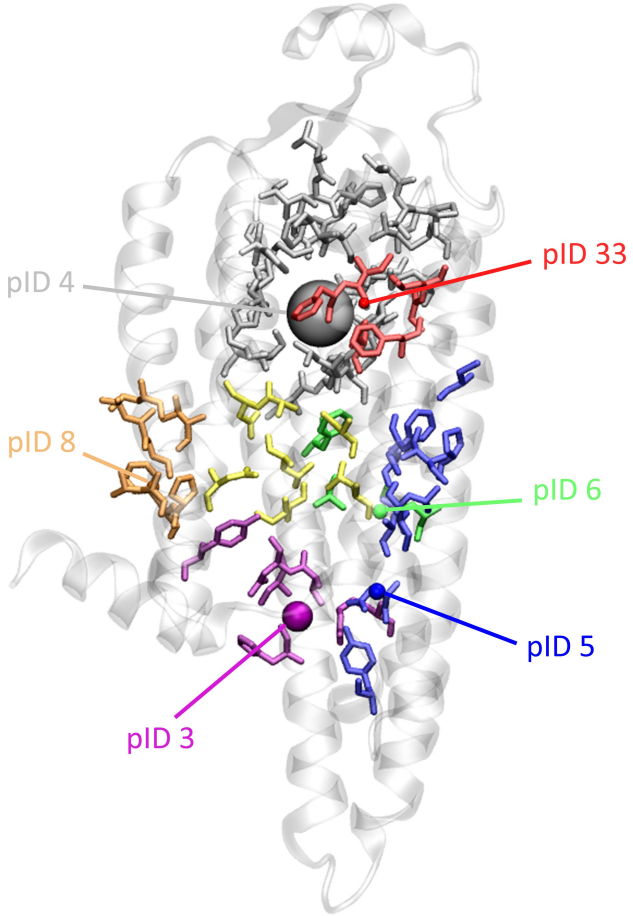

Figure 22: Comparative analysis of literature-reported hotspots (colored licorice) and pockets identified by Pocketron (colored spheres) for the A2A system. Sphere size reflects average non-zero pocket volume, with color consistency between corresponding hotspots and pockets. Sphere placement corresponds to the center of mass of pocket residues. Color code: Orthosteric site (gray), Intracellular Crevice (blue), G Protein-Coupling Site (purple), Lipid Interface (green), C-Terminus Cleft (orange), Extracellular Cleft (red), and Na^+ site (yellow) [112].

sodium binding deactivates the receptor, with the ion remaining stable in the inactive form [178] [179]. In contrast, the active receptor conformation constricts the ion pocket, preventing ion binding. In this case, MD simulations, initiated from a crystal structure (PDB ID 3uzc) with sodium in the ion pocket, maintained this configuration throughout. However, the ion pocket was not identified as an independent pocket in the final analysis. This indicates the presence of a channel connecting the ion pocket to the neighboring orthosteric site, resulting in the ion pocket residues being incorporated into the orthosteric pocket definition.

5.4.3 Androgen Receptor

Pocket analysis of the AR revealed three large pockets (pID 7, 8, and 13) and a medium-sized one (pID 3), all showing high persistency throughout the MD simulation. However, comparing the results from the three analysis methods (DyNet, Pocketron, and DF) revealed inconsistencies (Figure 23).

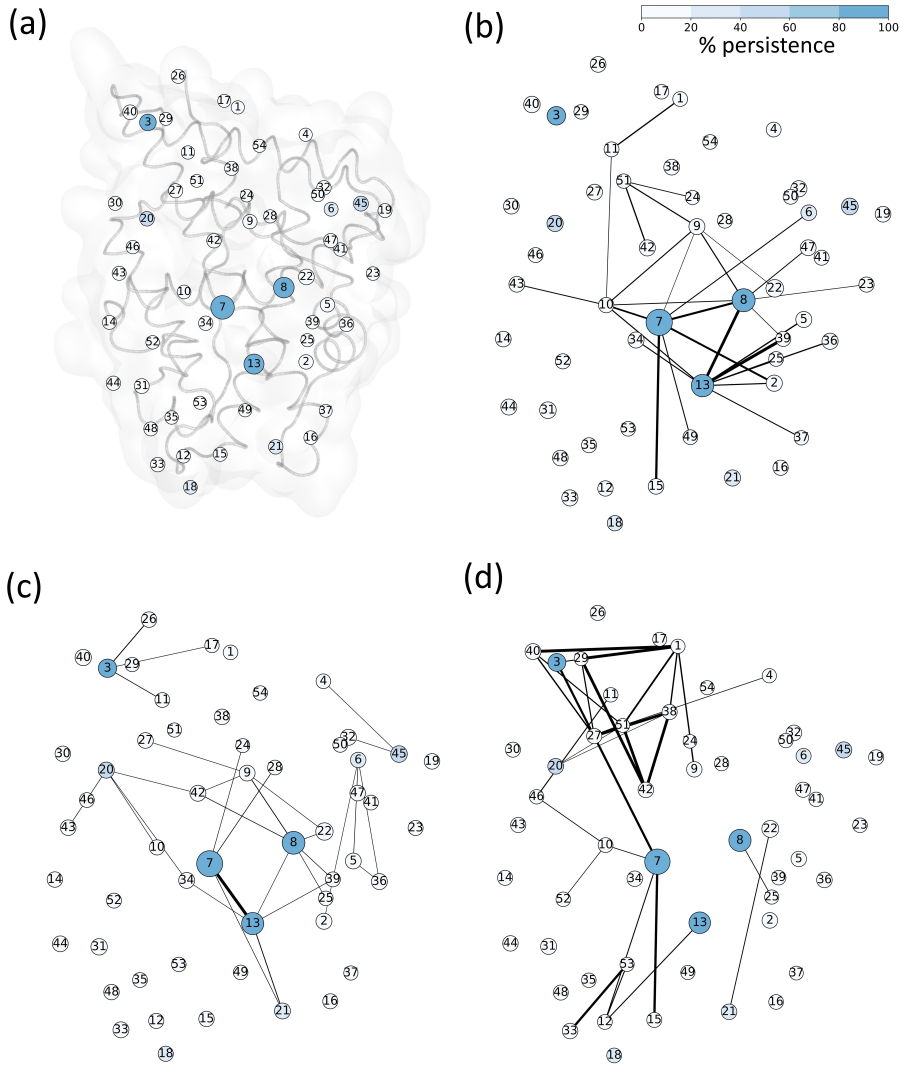

Figure 23: Network representation of pocket communication in the AR system as determined by (b) DyNet, (c) Pocketron, and (d) DF analyzes. (a) 2D node topology, obtained via multidimensional scaling, is superimposed on the AR receptor structure to visualize pocket locations [112].

DyNet (Figure 23b) indicated strong correlations between the orthosteric pocket (pID 13) and several neighboring pockets, most notably pIDs 8 and 39. While no direct

connection with the large, persistent pID 7 was observed, indirect connections via other pockets were identified. Additionally, a pathway between pID 7 and pID 1 was found, passing through pIDs 10 and 11. In contrast, Pocketron (Figure 23c) highlighted the central role of the orthosteric site (pID 13), which was connected to multiple surrounding pockets, including the large and persistent pIDs 7 and 8. However, Pocketron did not detect any connections between the persistent pID 3 and the larger pockets. This may be attributed to the protein conformation, where helix 4 separates these regions, potentially hindering residue exchange and thus communication as detected by Pocketron. Finally, the DF analysis (Figure 23d) identified two key clusters with strong correlations: one around pID 3 and 1, and another surrounding the orthosteric pocket (pID 13), primarily connected through pIDs 7 and 27. This suggests that communication pathways within the AR may be more complex than those revealed by DyNet or Pocketron alone.
The AR possesses three primary hotspots: the orthosteric binding pocket (Figure 24, gray residues), the BF-3 pocket (cyan residues), and the AF-2 pocket (green residues).

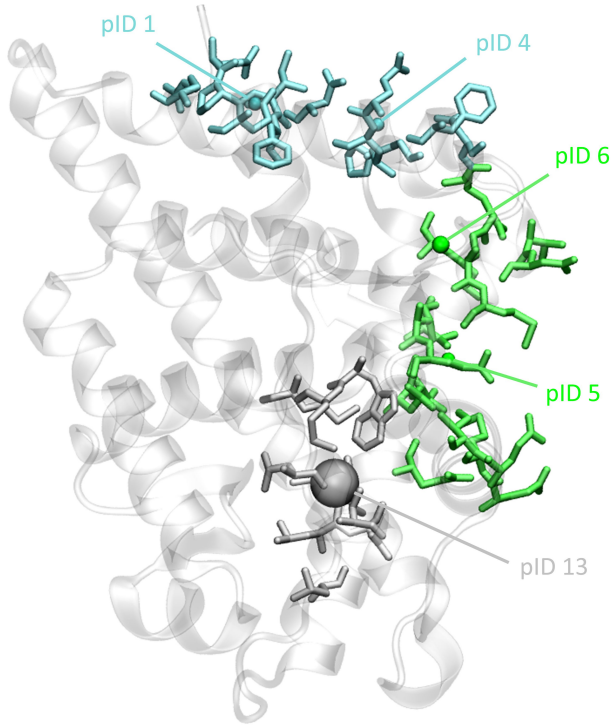

Figure 24: Comparative analysis of literature-reported hotspots (colored licorice) and pockets identified by Pocketron (colored spheres) for the AR system. Sphere colors correspond to the respective hotspots: orthosteric site (gray), BF3 site (cyan), and AF2 site (green) [112].

The AF-2 site [180] is crucial for the binding of co-activator proteins, essential for

AR-regulated gene transcription. Conversely, BF-3 site [181] functions as a potential target for antagonists, preventing co-activator binding and inhibiting AR activation. Antagonist binding to the AF-2 or BF-3 sites can modify AR activity by inducing conformational changes that modify the affinity with co-activator. Our simulations identified several pockets within these hotspot regions, including pIDs 5 and 6 in the AF-2 region, even if with low persistence. Despite the known communication between the orthosteric site and the AF-2 region, our analyses did not reveal significant communication between them. Only Pocketron's crosstalk analysis showed a weak connection between pID 13 (orthosteric site) and the small cluster comprising pockets 6, 5, and 36, via pID 39. Regarding the BF-3 pocket (pIDs 1 and 4), Pocketron detected no relevant communication with other pockets. However, DyNet revealed a pathway from pID 1 to pID 5, passing through pIDs 11, 10, and the orthosteric pID 13. Importantly, Pocketron correctly identified a notable communication between pID 13 and pID 21, located at the terminal part of H11, which undergoes conformational changes upon agonist or antagonist binding. This finding supports the possibility of information exchange between these two regions.

5.4.4 Epidermal Growth Factor Receptor kinase domain

Analysis of the EGFR system (Figure 25) consistently identified pID 1, corresponding to the hinge region, as a central hub connecting two distinct areas of the structure involved in inter-pocket communication.

DyNet analysis [Figure 25(b)] revealed numerous connections between the hinge region (pID 1) and nearby pockets (pIDs 30, 44, 53, and 57) as well as pockets formed by the A loop (pIDs 7, 8, 37, and 45). Additionally, a clear connection between pID 57 and pID 44 was observed. Pocketron results [Figure 25(c)] largely corroborated the DyNet findings, showing consistent connections between pID 1 and pIDs 57 (located between beta-strands and the alpha C-helix) and 44 (a deep interfacial groove). Both methods also identified a communication network within the A loop region, involving pID 1 and pIDs 7 and 8. The DF analysis [Figure 25(d)] further supported these observations, highlighting strong correlations within two opposing regions of the EGFR system. These regions, connected through the persistent pID 1, could potentially exchange structural dynamics information, underscoring the importance of the hinge region as a central communication hub.

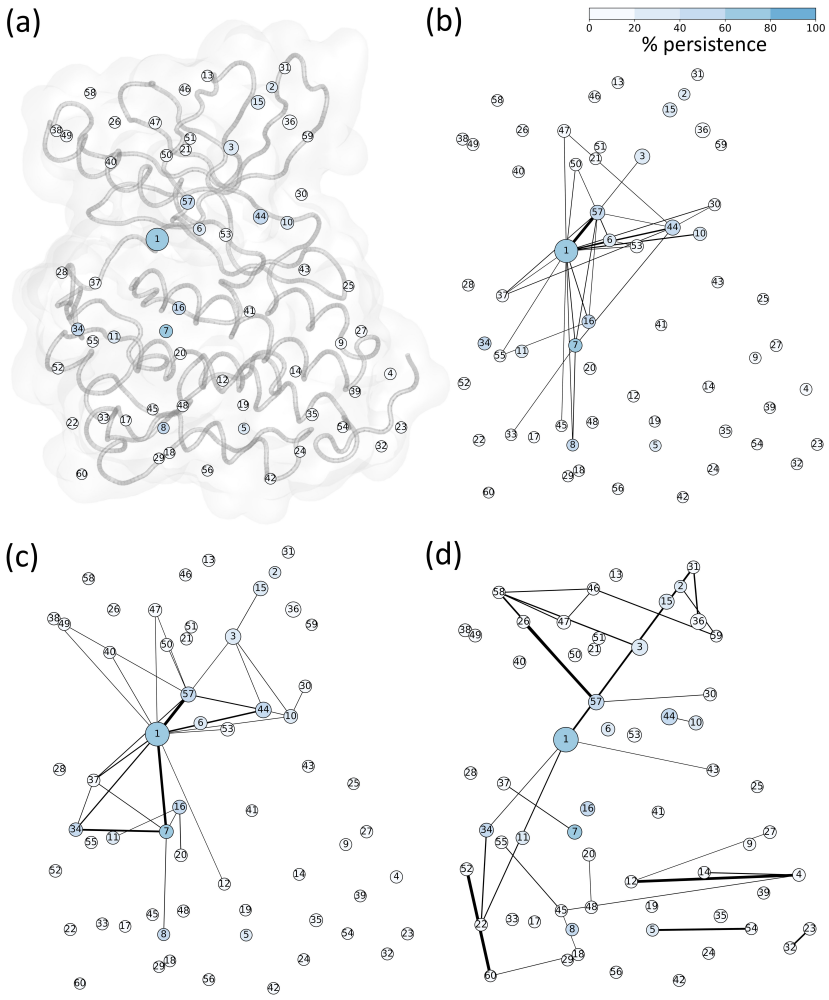

Figure 25: Network representation of pocket communication in the EGFR system as determined by (b) DyNet, (c) Pocketron, and (d) DF analyzes. (a) 2D node topology, obtained via multidimensional scaling, is superimposed on the EGFR receptor structure to visualize pocket locations [112].

The literature highlights several key structural elements within the receptor that govern its activation and deactivation. A hinge region, composed of flexible amino acids, enables bending and rotation without compromising the overall structure. In proximity to the hinge, we find the αC-helix, five β-strands, the A loop, and the orthosteric ATP binding pocket containing the "DFG" motif (residues D855, F856, G857). These elements collectively contribute to the dynamic regulation of the receptor's functional state.

A detailed study by Qiu et al. investigated the conformational changes of the EGFR kinase in response to ligand binding. They found that Osimertinib, when bound to

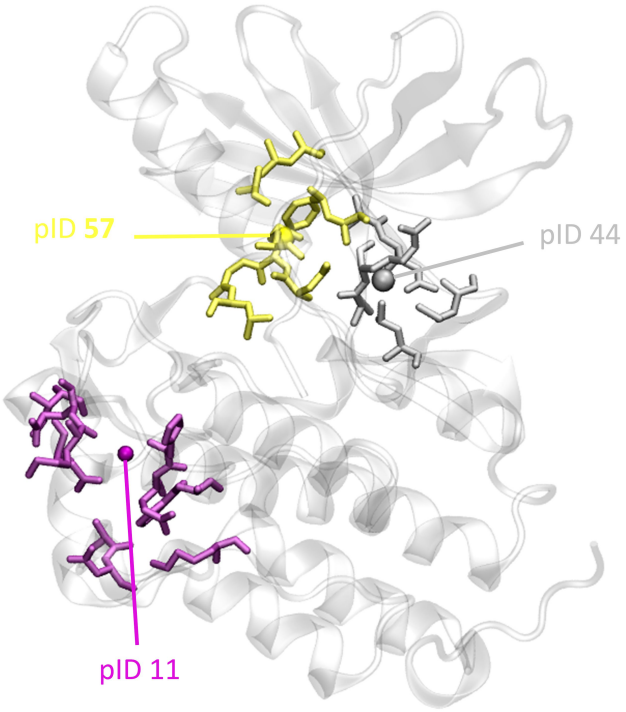

Figure 26: Comparative analysis of literature-reported hotspots (colored licorice) and pockets identified by Pocketron (colored spheres) for the EGFR system. Sphere colors correspond to the respective hotspots: orthosteric site (gray), MEK site (yellow), and pocket X (purple) [112].

the ATP binding pocket, induced an αC-out conformation with an open A loop. In contrast, the presence of the allosteric inhibitor JBJ in the MEK pocket led to an inactive "Src-like" state, characterized by a closed A loop interacting with the out-rotated αC-helix. Figure 26 highlights key regions within the EGFR structure: the ATP binding pocket (gray residues), the MEK pocket (red residues), and a potential allosteric site (purple residues) identified by Qiu et al. [182]. Our pocket detection analysis largely corroborates these findings, with pIDs 44 and 57 corresponding to the ATP and MEK pockets, respectively. Furthermore, our analysis identified the A loop (around pID 7) as another significant region, frequently showing correlations with the hinge region (pID 1) and pID 57 in both DyNet and Pocketron analyzes. This observation aligns with the literature's view of the hinge as a central communication hub within the EGFR structure. The connections between the hinge, the A loop, and pID 57 suggest an intricate communication network within this region, potentially playing a crucial role in the receptor's activation and function.

6 Application Case 2: Application of established computational methods in drug discovery to target RNAs

6.1 Introduction

6.1.1 Long non-coding RNAs

The organization of the eukaryotic genome is intricate, with nearly 98% of the human genome not encoding proteins [183]. This non-coding DNA, once dismissed as "junk DNA" [184, 185, 186], is now recognized as a repository of valuable information, encompassing diverse nucleotide elements and various non-coding RNAs (ncRNAs). The extent of functionality within these non-coding sequences remains a subject of ongoing debate. However, the Encyclopedia of DNA Elements (ENCODE) project has revealed that approximately 80.4% of the genome participates in biochemical activities, including chromatin structure regulation, histone modification, and RNA transcription [187].

Non-coding RNAs, which lack protein-coding capacity, are further classified based on size such as: short ncRNAs, less than 200 bases in length, that include tRNAs, rRNAs, miRNAs, snoRNAs [189]. In contrast to their shorter counterparts, long non-coding RNAs (lncRNAs), exceeding 200 nucleotides in length, constitute a diverse class of regulatory molecules [190]. lncRNAs can be categorized based on various factors, including their structure, function, localization, metabolism, and interaction with protein-coding genes or other DNA elements [191]. Interestingly, lncRNAs exhibit greater conservation in their secondary and tertiary structures compared to their primary sequences. However, due to the high flexibility, investigating the structure-function relationship of these large molecules remains challenging due to difficulties in crystallization, even with new technique such as cryo-EM [192, 193]. Transcribed from intergenic, exonic, or distal protein-coding regions by RNA polymerase II, lncRNA precursors undergo a maturation process that includes a 3'-polyadenylation and a 5'-capping with methyl-guanosine [194]. They also frequently undergo alternative splicing,

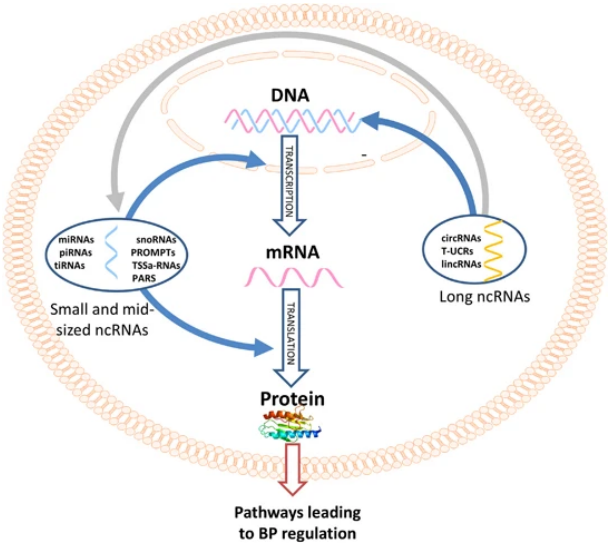

Figure 27: Non-coding RNAs (ncRNAs) play a diverse role in gene expression, expanding upon the traditional "central dogma." Long ncRNAs primarily interact with chromatin, influencing DNA accessibility and thus transcription levels (blue arrow to the right). Small and mid-size ncRNAs directly regulate various stages of gene expression, including transcription, RNA processing, and translation (blue arrows to the left). Additionally, certain long ncRNAs can control the abundance of miRNAs, adding another layer of complexity to gene regulation (grey arrow). [188]

a mechanism that generates protein diversity and can be broadly categorized into three modes of action: interaction with splicing factors, formation of RNA-RNA duplexes with pre-mRNAs, and modulation of chromatin remodeling [195].
lncRNAs, being the most prevalent class of ncRNAs, are involved in critical biological processes such as epigenetic regulation, transcriptional control [196], X chromosome inactivation [197, 198], genomic imprinting [199, 200], and cell development [201]. Consequently, their dysregulation is linked to various diseases, including cancer [202].

6.1.2 Pharmacological relevance of MALAT1

A prime example of the pharmacological relevance of lncRNAs is the metastasis-associated lung adenocarcinoma transcript 1 (MALAT1). Its overexpression has been linked to various cancers [203], while its knockdown has been shown to reduce oncogenic processes [204, 205], among other phenotypic effects [206, 207, 208]. However, the precise mechanisms and interactions underlying MALAT1's regulatory functions remain an active area of research [209]. MALAT1 represents a promising target for small-molecule-based therapeutics due to the presence of a well-characterized 3'-triple helix, implicated in transcript stability. This structure, functionally assessed in cell-based assays [210] and structurally resolved through X-ray crystallography [211], protects

MALAT1 from degradation by sequestering its adenine-rich 3'-tail through base pairing with a uridine-rich stem loop.

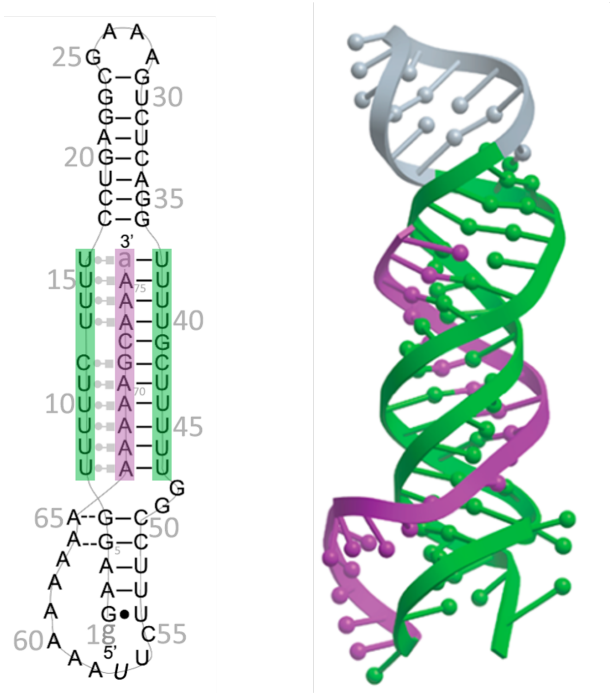

Figure 28: Structural Insights into MALAT1. The left panel shows a 2D representation with the distinctive triple helix (in green) and poly(A) tail (in purple). The right panel presents the 3D conformation of Malat1 in its apo state, color-coded to match the 2D structure. (PDB ID: 4PLX)

Beyond overall stability, recent research suggests that local conformational dynamics play a crucial role in triplex-mediated transcript protection [212]. Notably, small molecules can influence both the global stability and the conformational landscapes of RNA molecules [213, 50, 214, 215], further highlighting the potential for targeted therapeutic intervention.

Although the complete functional and structural landscape of MALAT1 remains to be fully elucidated, the formation of a 3'-terminal A-U rich RNA triple helix is known to be instrumental in its cellular accumulation [210]. This triplex structure sequesters the poly(A) tail within the major groove of a uridine-rich stem loop, effectively shielding the transcript from RNase degradation and extending its half-life [211]. Consequently, the MALAT1 triple helix has emerged as a promising target for small molecules that modulate its stability [216, 217, 213]. However, the lack of knowledge regarding the molecular recognition features that differentiate stabilizers and destabilizers of the MALAT1 triplex poses a challenge for targeted drug development. Additionally, the

limited availability of in vitro functional assays has hindered the understanding of the mechanisms governing RNA triplex modulation. A deeper comprehension of these mechanisms is crucial for improving the efficiency of future efforts aimed at targeting the MALAT1 triple helix and exploiting its therapeutic potential.

To assess the proficiency of current docking software and scoring functions in accurately predicting ligand experimental trend within RNA structures, a dataset of 21 small molecules (from the work of Hargrove et al. [218]) was employed. The molecules in this dataset share a common diminazene scaffold, but are differentiated by various substituents at the ortho, meta, and para positions, denoted by numeric codes and the prefixes "o-", "m-", and "p-", respectively.

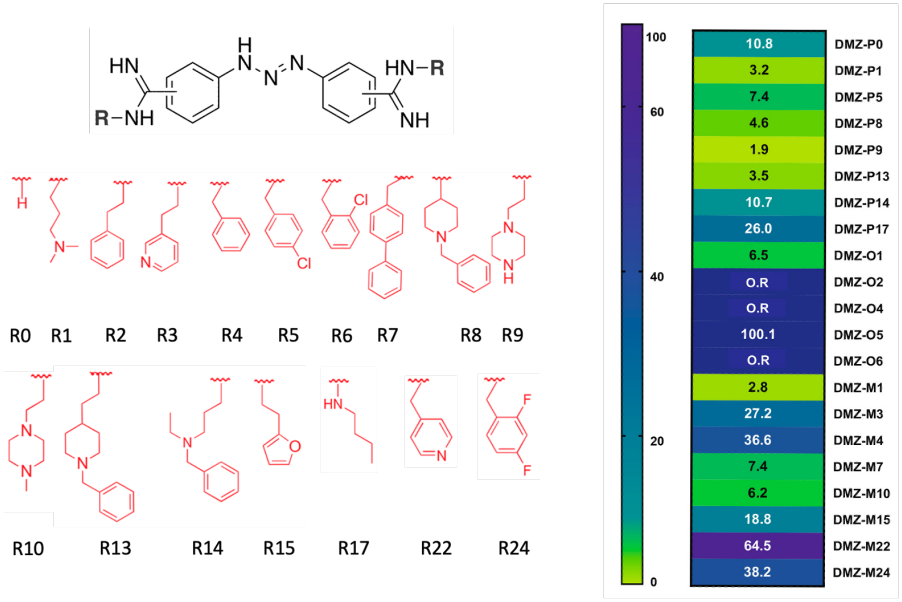

Figure 29: Dataset employed in the project, featuring diminazene as the central scaffold. Substitutions are indicated by codes and can occur at ortho, meta, or para positions. [218], the results on the right of the IDA experimental binding affinities are expressed in μM

Diminazene is an FDA-approved [219] antiparasitic agent and a known nucleic acid binder [220, 221, 222]. It has also been shown to preferentially bind A-U rich RNA triple helices [221], making it an attractive case study possible triple helix disruption by targeting the A-U rich MALAT1 triple helix. This dataset was chosen due to the availability of experimental binding affinity data for these ligands with the target, enabling us to evaluate the ability of the scoring functions to identify the correct ranking trends. The experimental binding affinities shown in Figure 29 were determined by the group of Hargrove [218] using Indicator Displacement Assay (IDA), a colorimetric

sensing technique [223] employing RiboGreen as dye. This method provides both qualitative and quantitative measurements by analyzing the different signals between the apo and complex states [224].

6.1.3 Methodology Workflow

The primary objective of this investigation is to critically assess the ability of current docking pipeline methodologies, predominantly developed for protein complexes, to identify possible ligand rankings when applied to RNA targets. As highlighted in Chapter 2.4, the inherent structural flexibility of RNA presents a formidable challenge in such endeavors. To address this, our pipeline starts with the generation of a conformational ensemble focusing on the receptor's triple helix, employing enhanced sampling techniques to capture its conformational landscape comprehensively. Subsequent docking calculations are then performed on each conformation within this ensemble to elucidate ligand binding modes and, crucially, to estimate the binding scores of the ligands under investigation. The selection of a suitable ligand library for this study needs a scaffold known to exhibit affinity for the target, with experimental validation documented in the literature. This ensures a meaningful comparison between computational predictions and established experimental data, facilitating a robust evaluation of the docking pipeline's performance in the context of RNA-ligand interactions.
A focused library based on the diminazene scaffold was constructed by Hargrove's group [218]. This choice was motivated by them considering the scaffold's synthetic accessibility and established nucleic acid binding properties, enabling an investigation into the fundamental recognition principles governing triplex modulators. The library incorporated variations from the central scaffold and subunit chemical groups to maximize diversity in the 3D shapes of the analogues. A comprehensive assessment of this 21-member library was conducted using four in vitro assays designed and/or optimized to specifically probe small molecule-RNA triplex interactions. These assays revealed that the previously observed trend for RNAs of rod-like 3D shapes correlation held true also for the MALAT1 triplex for the library compounds. However, while affinity also correlated with the impact of small molecules on triplex thermal stability, strikingly different trends emerged in enzymatic degradation assays.
In this study, we started our investigation with the crystallographic structure of the apo form of the MALAT1. Subsequently, both unbiased and biased (Hamiltonian replica exchange) molecular dynamics (MD) simulations were performed. The unbiased

simulations were subjected to Pocketron dynamic analysis to identify potential binding pockets for subsequent virtual screening.

A clustering analysis was then performed on these selected pockets, and centroids of clusters with an eRMSD distance above a predefined cutoff were extracted to generate an ensemble of target conformations. This ensemble was used for docking simulations, where ligand poses were generated for each conformation using AutoDock GPU [225] and rDock [28] software. Finally, the generated poses were evaluated using a diverse set of scoring functions, including AutoDock [226], rDock [28], AutoDock Vina [25, 26], AnnapuRNA [34], and SPRank [227]. This multi-faceted approach allowed us to comprehensively explore the conformational landscape of the MALAT1 structure and evaluate the ability of current scoring functions to accurately predict experimental ligand binding affinities for RNA systems.

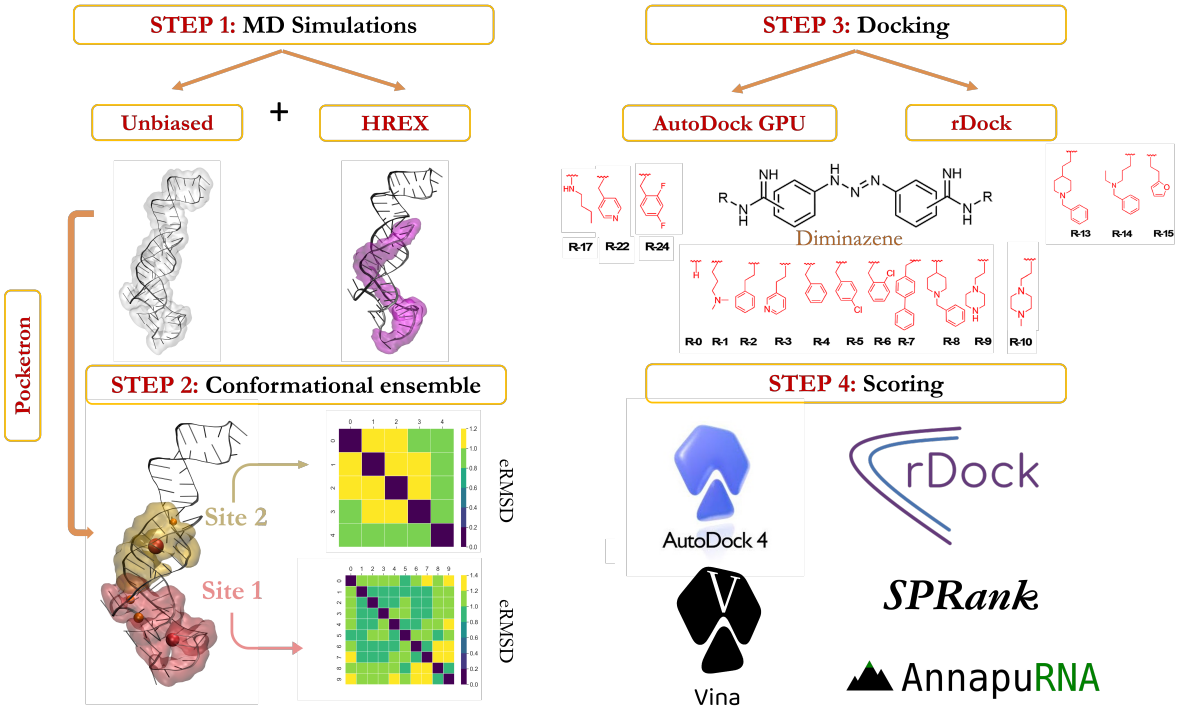

Figure 30: Computational pipeline for evaluating scoring functions in RNA ligand docking, using an MD generated conformational ensemble and a diminazene based ligand library

6.2 Methods

6.2.1 System preparation

To conduct molecular dynamics simulations, the PDB target file (PDB ID 4PLX) was processed using GROMACS 2021.4 [168], adhering to established protocols. Initially, the system was converted into a GROMACS-compatible input file, employing the state-of-the-art Amber ff99 force field with the *bsc*0 and χ_{OL3} refinements, specifically devised for RNA simulations. The system was solvated using the four-point OPC water model, ensuring an appropriate representation of the aqueous environment. To maintain charge neutrality, potassium (K^+) and chloride (Cl^-) ions were added to achieve a concentration of 0.15 M. Subsequently, the system underwent energy minimization followed by equilibration in both the NVT and NPT ensembles, with a total simulation time of 1.2 ns (600 ps for each ensemble). This meticulous preparation ensured a stable and well-equilibrated system for subsequent production MD simulations.
The equilibrated system was subsequently employed as the starting point for the unbiased molecular dynamics simulations, which were conducted within the NPT ensemble. In contrast, for the Hamiltonian replica exchange simulations, an additional step was incorporated into the equilibration process. Specifically, the fluctuations in the simulation box volume were analyzed, and the frame exhibiting a volume closest to the mean value (preferably in the second half of the equilibration) was carefully chosen as the initial configuration for subsequent biased MD simulations. This selection was made to mitigate potential artifacts caused by fluctuations in box dimensions, which could result in artificially high or low pressures. By choosing a starting frame with a volume close to the average, we aimed to ensure stable pressure conditions for the subsequent biased simulations performed in the NVT ensemble.

6.2.2 MD simulations

Following the established protocol, two distinct molecular dynamics (MD) simulations were conducted on the MALAT1 system: an unbiased MD simulation and a Hamiltonian replica exchange simulation. The unbiased MD simulation, performed in the NPT ensemble, was initiated from the pre-equilibrated system and comprised three independent replicas, each with a simulative time of 500 ns. These simulations were carried out under constant temperature and pressure conditions of 300 K and 1 bar,

respectively, maintained using the V-rescale thermostat and the C-rescale barostat. The resulting trajectories from these three replicas were then concatenated for subsequent analysis.

To enhance conformational sampling and overcome potential energy barriers, Hamiltonian replica exchange (HREX) MD simulations were also employed. Sixteen replicas were created, with the scaling parameter λ ranging from 0.7 to 1, specifically targeting the poly(A) triple helix residues (residues 55 to 76 highlighted in purple in the Step 1 of Figure 30). As detailed in Section 3.2.4, Coulomb interactions, Lennard-Jones potentials, and dihedral parameters within the "hot" region were scaled by factors of λ or $\sqrt{\lambda}$. This approach facilitates exploration of conformational space beyond local minima, particularly compared to unbiased simulations. As outlined in Chapter 6.2.1, simulations were initiated from the frame with the volume closest to the average value, ensuring consistent pressure conditions. These biased simulations were conducted in the NVT ensemble using the V-rescale thermostat, with each replica running for approximately 100 ns (precisely 96 ns). During the simulations, exchanges between replicas were attempted every 240 steps based on the Metropolis criterion (Equation 17). The unscaled replica ($\lambda = 1$) was subsequently extracted, processed and subjected to further analysis to generate the conformational ensemble.

6.2.3 Conformational ensemble preparation

Upon termination of the simulations, we started on the trajectory analysis to extract representative frames for subsequent docking campaigns. Initially, the combined unbiased trajectory was analyzed using Pocketron to identify dynamic pocket formation and communication patterns. A communication matrix was constructed to pinpoint residues belonging to pockets exhibiting the largest volume, highest persistency, and significant long-range communication. Subsequently, we examined the conformational variability of these residues across both the unbiased and biased ($\lambda = 1$ replica) trajectories, employing two distinct metrics: RMSD and eRMSD [228]. While RMSD is a common measure of structural deviation, eRMSD, specifically designed for nucleic acids, offers a more nuanced understanding of base pairing alterations. High RMSD values may not necessarily indicate substantial conformational changes in RNA (in comparison to proteins), whereas elevated eRMSD values (typically exceeding 0.7-0.8) signify a more significant base pairing rearrangements.

Figure 31 shows the distance r_{ij} between two bases that can be also expressed in its

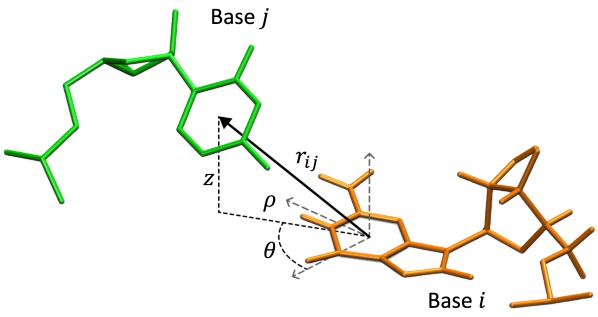

Figure 31: Graphical representation of the resultant g-vector between base i and base j, considering their 3D spatial distances and relative orientations.

cylindrical coordinates (ρ, θ and z). In this way it can be introduced the anisotropic position vector (Equation 28):

$$\tilde{\mathbf{r}} = \left(\frac{r_x}{a}, \frac{r_y}{a}, \frac{r_z}{b}\right) = \left(\frac{\rho}{a}\cos\theta, \frac{\rho}{a}\sin\theta, \frac{z}{b}\right) \tag{28}$$

where $a = 5\mathring{A}$ and $b = 3\mathring{A}$ are spatial scaling parameters that define the ellipsoidal interaction shell. These parameters are chosen to ensure that the normalized distance $\tilde{r}$ between interacting base pairs falls within the range $1 < \tilde{r} < \sqrt{2.5}$.
To compute eRMSD for trajectory frames, a 4-dimensional vector known as the g-vector is calculated (Equation 29) [228]. This vector includes information regarding the relative positions and orientations of base pairs, enabling a more accurate assessment of conformational changes in RNA structures.

$$G(\tilde{r}) = \begin{pmatrix} \sin(\gamma\tilde{r})\tilde{r_x}/\tilde{r} \\ \sin(\gamma\tilde{r})\tilde{r_y}/\tilde{r} \\ \sin(\gamma\tilde{r})\tilde{r_z}/\tilde{r} \\ 1 + \cos(\gamma\tilde{r}) \end{pmatrix} \times \frac{\Theta(\tilde{r}_{cutoff} - \tilde{r})}{\gamma} \tag{29}$$

$$eRMSD = \sqrt{\frac{1}{N}\sum_{i,j} |G(\tilde{r}^{\alpha}) - G(\tilde{r}^{\beta})|^2} \tag{30}$$

where $\tilde{r}_{cutoff}$ represents the distance threshold beyond which g-vectors are set to zero, set to 2.4. The parameter $\tilde{r}$ denotes the anisotropic position vector between the two bases pair expressed in Equation 28, while γ is defined as $\pi/\tilde{r}_{cutoff}$. The resulting g-vectors, when applied to Equation 30, enable the calculation of eRMSD values between conformations or their selected regions.

Following the calculation of g-vectors, Principal Component Analysis (PCA) was performed to reduce the dimensionality of the simulation data with particular attention to the residues within the selected pockets and verify the adequate sparseness of the points (conformations) in the reduced space. Subsequently, the Quality Threshold cluster method [229] was implemented to extract cluster centroids, enabling the construction of a total eRMSD matrix for the corresponding frames. To generate the final (filtered) ensemble, frames with eRMSD values below 0.7 were excluded, ensuring the retention of only those conformations exhibiting significant structural variations.

6.2.4 Poses generation

For the docking campaign, a dataset of 21 small molecules, comprising analogues of the diminazene scaffold with a diverse set of substituents in ortho-, meta-, and para-, was employed. Docking poses for each ligand were generated within every conformation of the ensemble using two distinct software packages: AutoDock GPU and rDock.
The molecular docking protocol comprised two key preparation stages: target and ligand processing. Preparing the target RNA structure required specific steps tailored to the requirements of each docking software. In both cases, we used a united-atom approach, explicitly adding hydrogen atoms only to polar atoms. However, the methodology for assigning partial charges varied significantly between the softwares. For AutoDock GPU, the target RNA structure was first converted from PDB to PDBQT format, with partial charges automatically assigned using AMBER force field parameters. In contrast, rDock start from a MOL2 file format for the coordinates while it adopts a different method to assign charges. Instead of directly reading charges from the MOL2 file, rDock assigns partial charges based on predefined substructure patterns and standard atom nomenclature.
For the ligands, we started from SMILES strings and processed them through the Epik tool [230] within Maestro (Schrodinger 2022.2) [231] for protonation state prediction at specific pH values, particularly 7.3 $\pm$ 0.1, converting them in SDF format files. The resulting ligands exhibited net charges of +2 or +4. The subsequent processing differed between the two docking software: For AutoDock GPU, the SDF files were converted to PDBQT format, retaining hydrogen atoms only on polar groups and assigning Gasteiger charges to the atoms. In contrast, rDock, which does not incorporate partial charges in its scoring function, processed the SDF files directly using a united-atom representation. The grid calculation phase also differed between the two docking software. For both

methods, the grid was centered on the center of mass (excluding hydrogens) of the pocket-defining residues of the site. AutoDock GPU required a rectangular grid box configuration. We employed dimensions of 110 × 110 × 130 points for the first site and 110 × 80 × 80 points for the second site, with a consistent spacing of 0.3 Å between points. These dimensions were chosen to ensure complete coverage of the binding pockets identified by Pocketron.
Conversely, the rDock algorithm employ a two-sphere approach for grid file generation. A radius of 17 Å was employed for both sites, with a small probe of 1 Å and a large probe of 17 Å. This strategy aimed to create a spherical grid closely approximating the dimensions of the AutoDock grid box, thus ensuring comparability between the results obtained from the two docking software.

6.2.5 Poses rescoring

For scoring the generated poses, we employed a variety of functions encompassing both force field-based (AutoDock, rDock and Vina) and machine learning-based (AnnapuRNA and SPRank) approaches. AutoDock evaluates binding affinity through a two-step process. First, it estimates the intramolecular energetics required for the ligand and target to transition from their unbound conformations to their bound states. In the second step, it calculates the intermolecular energetics of the ligand-target complex in the bound conformation. The AutoDock force field accounts for six pairwise potential components ($V^{X-X}Y$) and an entropy loss term associated with binding ($\Delta Sconf$), leading to the total binding free energy:

$$\Delta G = (V^{L-L}_{bound} - V^{L-L}_{unbound}) + (V^{P-P}_{bound} - V^{P-P}_{unbound}) + (V^{P-L}_{bound} - V^{P-L}_{unbound} + \Delta S_{conf}) \quad (31)$$

where L represents the ligand and P the protein. Each pairwise energy term includes contributions from dispersion/repulsion forces, hydrogen bonding, electrostatics, and desolvation effects. Conversely, Vina's scoring function follows a different approach, as it does not explicitly include electrostatics or solvation terms [26]. Instead, it relies on a van der Waals-like potential constructed from a repulsion term and two attractive Gaussian functions, a non-directional hydrogen bond term, a hydrophobic interaction term, and a conformational entropy penalty [25]. Similarly, the rDock scoring function (S_{total}) is computed as a weighted sum of multiple contributions, including intermolecular interactions (S_{inter}), ligand intramolecular energy (S_{intra}), site intramolecular energy (S_{site}), and external restraints ($S_{restraint}$). The key component, S_{inter}, quantifies the

protein-ligand (or RNA-ligand) interaction strength. S_{intra} represents the relative energy of the ligand conformation, while S_{site} reflects the relative energy associated with flexible regions within the active site. Finally, $S_{restraint}$ encompasses a collection of non-physical restraint functions, which can be strategically employed to bias the docking calculation in several beneficial ways.

$$S_{total} = S_{inter} + S_{intra} + S_{site} + S_{restraint} \tag{32}$$

Furthermore, AnnapuRNA, an ML-based scoring function, was employed. To derive a general model of RNA-small molecule interactions, this approach uses a coarse-grained representation of both binding partners based on a vast set of RNA-small molecule complex structures available in the literature. For RNA molecules, a coarse-grained representation similar to that used in the SimRNA simulation method is adopted [232], featuring five "beads" per ribonucleotide residue, strategically positioned at the locations of real atoms. Small-molecule ligands are represented using the concept of pharmacophores, as implemented by Taminau et al. [233]. Six pharmacophore types are employed, each with associated Euclidean vectors indicating their direction (Figure 32). The centers of HDON (hydrogen bond donor), HACC (hydrogen bond acceptor), POSC (positive charge), and NEGC (negative charge) points match with the corresponding heavy atoms. The AROM (aromatic ring) point is located at the center of the represented aromatic ring. The LIPO (lipophilic) point is generated through a multi-step process where adjacent lipophilic regions are averaged into a single pharmacophore, weighted by their lipophilic contributions. This coarse-grained representation facilitates the training of a machine learning model capable of predicting RNA-small molecule binding affinities.
Therefore, the total score for an RNA-ligand complex is calculated as the sum of two terms:

$$E = E_{RNA-Ligand} + E_{Ligand} \tag{33}$$

In this model, the final score (E) of the complex is calculated as the sum of two components: the RNA-Ligand interaction score ($E_{RNA-Ligand}$) and the ligand's internal conformation score (E_{Ligand}). The RNA-Ligand interaction score is determined by summing the probabilities (p) of each interaction between RNA atoms and ligand pseudoatoms within a 10 Å cutoff distance. These interaction probabilities are derived from a machine learning model, representing the likelihood of a specific interaction

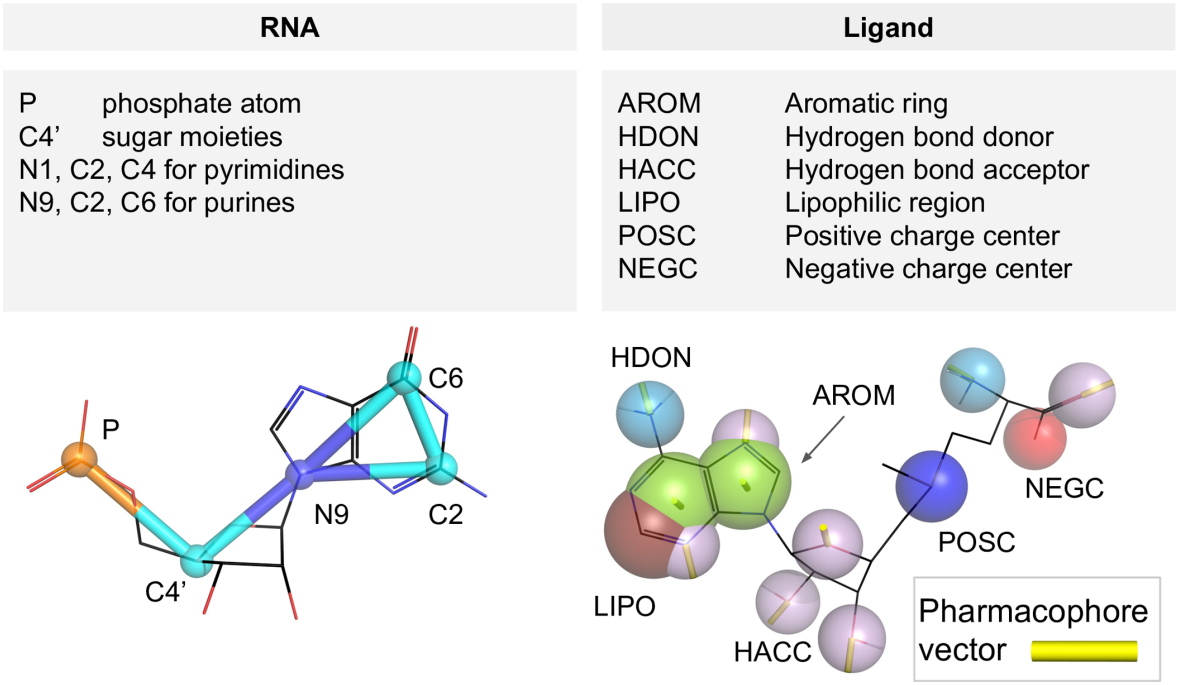

Figure 32: Coarse-grained representations used for RNA and ligand molecules. The upper left panel shows the simplified building blocks (atoms and pseudoatoms) for RNA. The upper right panel shows the corresponding representation for ligand molecules. The bottom left provides an example of a ribonucleotide as depicted in the SimRNA model. The bottom right displays a small molecule represented using pharmacophore features [34].

belonging to the "positive" (favorable) class.

$$E_{RNA-Ligand} = -1 * \sum_{interactions} p_{interactions} \tag{34}$$

The score for the internal energy of the ligand, denoted as E_{Ligand}, is derived from the GAFF (General Amber Force Field) internal energy of the ligand [234] and is computed as follows:

$$E_{Ligand} = (E_{GAFF} - b) * w \tag{35}$$

The ligand internal energy contribution is scaled by the factor b, effectively shifting positive GAFF energy values to negative values within the scoring function.
Lastly, the SPRank knowledge-based and ML scoring function operates on two fundamental principles [235, 236, 237]. Firstly, it assumes that the overall strength of interactions within an RNA-ligand complex (both between the RNA and ligand, and within each molecule) can be determined by summing up the individual interaction

energies between all pairs of atoms.

$$U_{pair} = \sum_{ij} \mu_{ij} \tag{36}$$

This equation calculates the total interaction energy (U_{pair}) within an RNA-ligand complex by summing the individual interaction energies (μ_{ij}) for every pair of atoms. Moreover, SPRank also relies on the principle that the statistical distribution of interatomic distances observed in experimental RNA-ligand structures should mirror the distribution predicted by the scoring function. In simpler terms, the scoring function must effectively differentiate between correct binding poses—those that closely resemble native or near-native conformations—and incorrect ones, known as decoys. This distinction is based on how well the predicted poses replicate the patterns of interatomic distances observed in real-world experimental data. [31, 238, 235, 236]. This is achieved through an iterative process that refines the scoring parameters until they effectively capture these observed distance distributions. Equation 37 provides a mathematical expression for deriving the pairwise potentials.

$$u_{ij}^{k+1}(r) = u_{ij}^{k}(r) + \lambda k_B T \ln[g_{ij}^{k}(r)/g_{ij}^{\text{obs}}(r)] \tag{37}$$

where k, k_B, and T denote the iteration step, Boltzmann constant, and temperature, respectively. This method considers the inherent properties of the RNA and ligand, such as their shape and flexibility, by incorporating simulated decoy structures. These decoys account for factors like excluded volume and chain connectivity. The scoring function is refined by comparing the predicted interatomic distance distribution with the one observed in experimental data. Ideally, these distributions should align perfectly. However, to ensure numerical stability and prevent potential issues during the refinement process, a modified equation is used instead of Equation 37, avoiding the logarithmic ratio. This adjustment helps maintain robustness in the calculations and avoids numerical errors.

$$u_{ij}^{k+1}(r) = u_{ij}^{k}(r) + \lambda k_B T(g_{ij}^{k}(r) - g_{ij}^{\text{obs}}(r)). \tag{38}$$

The calculations can be simplify by setting the thermal energy ($k_B T$) to 1 and using a parameter called λ to control how quickly the scoring function is refined set to 1.0. The variable $g_{ij}^{\text{obs}}(r)$ represents the experimentally observed distribution of distances between atom pairs, weighted by factors that account for the structural variability of the RNA. While variable $g_{ij}^{k}(r)$ represents the predicted distribution of distances between atom

pairs at each step of the refinement process. This prediction is calculated by averaging the Boltzmann-weighted distances for each RNA-ligand complex and then weighting them further based on the structural variability of the RNA.

6.3 Results and Discussions

6.3.1 Comparative Analysis of Unbiased and Biased Simulations

As outlined in the Methods (Chapter 6.2.2), three independent, unbiased MD simulations of 500 ns each were conducted to investigate the dynamic behavior of the MALAT1 structure.

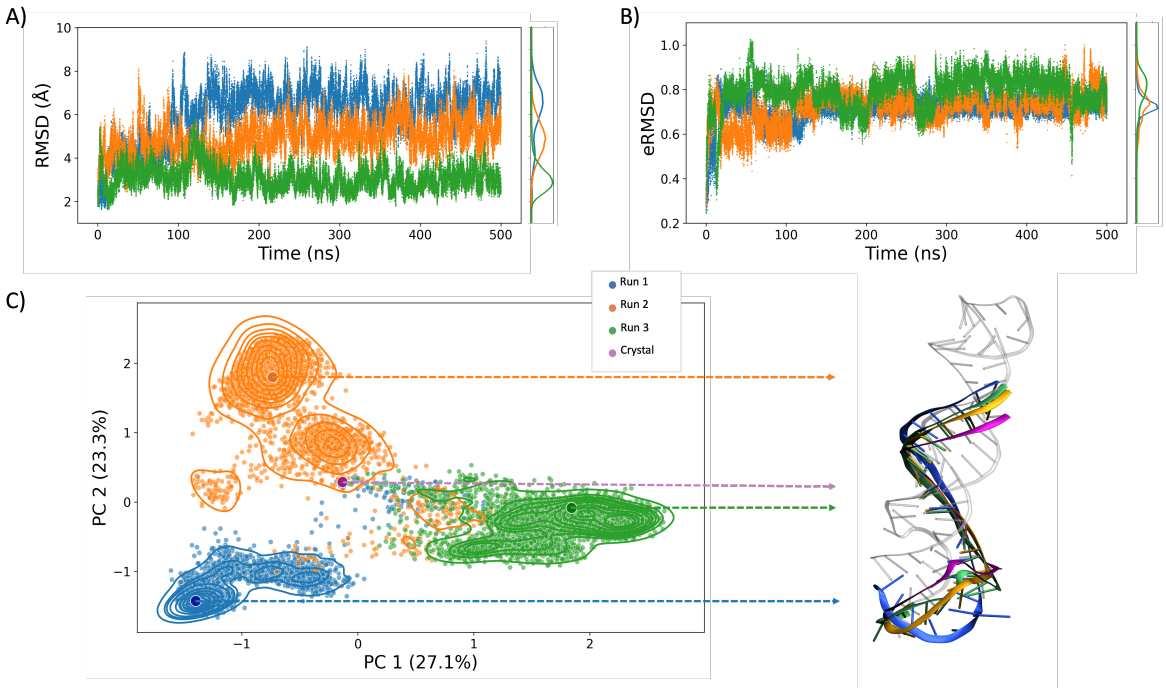

Figure 33: Conformational Analysis of MALAT1 Unbiased Simulations. A-B) RMSD and eRMSD of polyA residues along with their respective probability distributions, across three independent unbiased simulations. The minimized crystallographic structure is used as the reference point for these calculations. C) PCA of all conformations sampled during the unbiased simulations, color-coded by replica with crystallographic reference in purple. The medoids representing the three main clusters from each run are shown superimposed with the crystal structure, using the same color scheme.

Figure 33 provides insights into the conformational dynamics of the poly(A) tail during unbiased simulations. The RMSD analysis (Figure 33A) indicates significant structural

fluctuations, particularly in Run 1 (blue line), where RMSD values reach approximately 8 Å. In contrast, the eRMSD (Figure 33B) indicates much greater stability, with values for all three replicates fluctuating around 0.7 and only Run 3 (green line) briefly approaching 1. This suggests that the high RMSD values observed in Figure 33A do not necessarily correspond to significant base pairing rearrangements, emphasizing the importance of eRMSD as a metric specifically designed to capture changes in the relative arrangement of nucleic acids. Furthermore, the PCA plot, color-coded by replica, demonstrates the accessibility of diverse conformational clusters within a short timeframe, resulting in three distinct clusters representing each replica.

To assess the conformational flexibility of the system during the simulations, a principal component analysis (PCA) was performed on the combined trajectory. Specifically, the dimensionality reduction was achieved using g-vectors calculated for the poly(A) residues (residues 55 to 76 highlighted in purple in the Step 1 of Figure 30) within the triplex helix region. This approach allowed to test if the metric used gave us a comprehensive characterization of the conformational landscape explored by the MALAT1 during the unbiased simulations. Furthermore, we combined the unbiased and biased ($\lambda = 1$ replica) trajectories, using a stride of 10 and 2, respectively, to investigate the expanded conformational space captured by incorporating the hREX simulations. This approach allowed us to assess the impact of enhancing sampling specifically within the triple helix region.

Figure 34 illustrates the effectiveness of the replica exchange method in sampling a similar range of RMSD and eRMSD values compared to the unbiased simulations. Notably, the hREX simulation reaches eRMSD values near 1.2, indicating broader exploration of conformational space. This expansion is clearly visualized also in the PCA plot, where the hREX conformations (gray points) not only overlap with the regions explored by unbiased simulations (blue, orange, and green dots) but also form distinct clusters, especially one located further away, highlighting the greater conformational sampling achieved with hREX.

6.3.2 Pocket analysis and selection

After confirming our simulation setup's capability to accurately represent conformational dynamics, we employed the Pocketron algorithm to analyze inter-pocket communication patterns. We processed the joined unbiased MD trajectories through the BiKi Life Sciences platform, selecting non-hydrogen atoms of the MALAT1 as our input. To

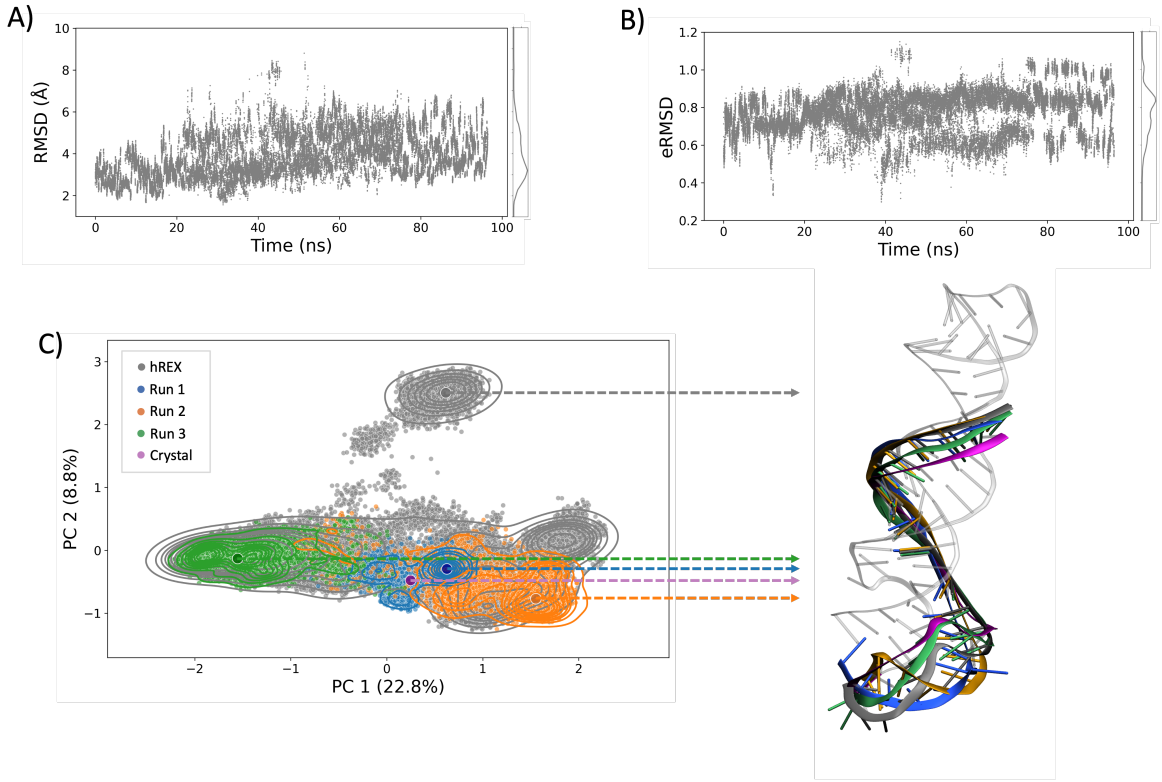

Figure 34: Conformational Analysis of MALAT1 unbiased plus biased simulations. A-B) RMSD and eRMSD of polyA residues along with their respective probability distributions, across three independent unbiased and the biased hREX simulations. The minimized crystallographic structure is used as the reference point for these calculations. C) PCA of all conformations sampled, color-coded by replica with crystallographic reference in purple. The medoids representing the four main clusters from each run (unbiased or biased) are shown superimposed with the crystal structure, using the same color scheme.

focus on potentially druggable pockets capable of mediating long-range communication across the structure, only pockets with a volume greater than three water molecules were tracked using Pocketron. This criterion ensured our analysis targeted pockets with the potential to accommodate a small molecule and by induced fit, potentially, disrupt the triple helix structure through allosteric mechanisms.

The analysis yielded a total of 37 pockets, encompassing those present in the initial structure and those dynamically formed during the trajectory. Figure 35A offers a comprehensive view of pocket communication within the MALAT1 structure. It displays the $37x37$ correlation matrix derived from Pocketron analysis, where each element quantifies the communication between corresponding pocket IDs (pIDs). These pockets are visually represented as spheres in Figure 35B, where the sphere radius correlates with pocket volume, and color denotes pocket persistency (percentage of frames with

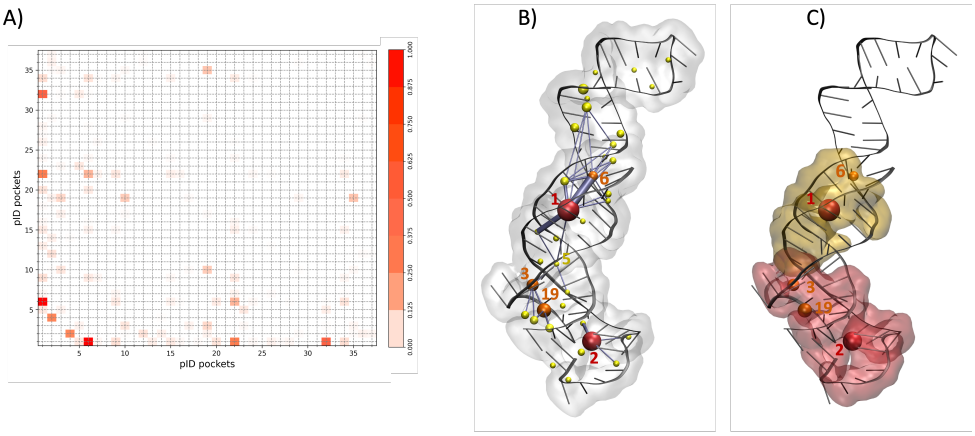

Figure 35: Inter-pocket communication analysis of the MALAT1 structure. A) Pocketron-derived correlation matrix, where rows and columns represent pocket IDs. B) 3D visualization of pockets and their connections. Edge thickness indicates correlation strength C) Residues selected for ensemble generation and docking campaign named Site 1 (red) and Site 2 (yellow).

non-zero volume). The color scheme reflects the following pockets persistency ranges: yellow (<33%), orange (33-66%), and red (>66%). As shown in Figure 35B, the most persistent pockets (red and orange) also exhibit larger volumes, suggesting their potential suitability for further analysis and targeting in drug discovery efforts.

To further assess the suitability of these pockets for ligand binding, we employed Pocketron to track residue exchange between pockets and quantify the extent of communication across different target regions. This involved calculating the average number of "merge" and "split" events, which was then visualized in a 3D graph. This analysis provided crucial insights into the dynamic interplay of pockets within the MALAT1 structure, highlighting promising sites for the docking campaign and potentially uncovering key allosteric pathways (ideally able to disrupt the triple helix). Moreover, Figure 35B provides a spatial context for these correlations, with edge thickness visually representing the strength of communication between pockets. The analysis reveals two promising regions which harbor the largest and most persistent pockets. These pockets not only engage in frequent residue exchange with neighboring pockets, fundamental to be a suited docking site, but also exhibit communication across the two sites, albeit to a lesser extent. This inter-site communication hints a potential allosteric effects that could propagate through the structure and ultimately influence the stability of the crucial triple helix. A notable short communication pathway is observed, originating from pocket 19 at the base of the triple helix and traversing through pockets 3, 5, 1, and 6 towards the upper portion of the structure. Additionally

to those communications, a localized exchange of residues around the high volume and high persistence pocket 2 can be seen.
Based on these insights, we next focused our analysis on the residues within these persistent pockets (Table 3), the targeted regions, visually highlighted in red and yellow in Figure 35C as Site 1 (pIDs 2, 3 and 19) and Site 2 (pIDs 1 and 6), respectively, represent promising areas for further exploration of ligand binding sites, in perfect agreement with the work of Khanna et al. [239].

pID	Av. Volume (Å^3)	Residues	Persistency
1	455	9, 10, 11, 39, 40, 68, 69, 70	96%
6	99	6, 7, 64, 65	42%
2	297	1, 2, 3, 4, 50, 51, 52, 54, 55, 56, 57, 58, 59, 60, 61, 62	69%
3	179	45, 47, 48, 64, 65	62%
19	294	47, 48, 49, 50	35%

Table 3: Pocket characteristics identified in the MALAT1 structure, including pocket IDs (pID), average volumes, constituent residues, and pocket persistency values across the simulation trajectory.

6.3.3 Conformational ensembles

With the pocket definitions established, we proceeded with dimensionality reduction analysis to assess the conformational space explored for each of the two identified sites. Initially, we analyzed the combined unbiased simulation trajectory, following a similar approach to the poly(A) residue analysis in Chapter 6.3.1. Subsequently, we incorporated the $\lambda = 1$ replica from the 96 ns Hamiltonian replica exchange simulations to evaluate the effect of biased sampling on the conformational landscape, particularly with the triple helix region designated as the "hot" region.
For both sites, g-vectors were calculated and PCA plots were generated, focusing solely on the pre-selected residues. This approach allowed us to visualize and quantify the conformational variability captured within the entire simulation dataset, encompassing both unbiased and biased sampling.
Figure 36 illustrates the eRMSD values for both sites, ranging from 0.6 to approximately 1.3, with the hREX simulations effectively expanding the conformational space sampled by the unbiased simulations. This increase in conformational diversity is particularly evident in the PCA analysis, where the gray points (representing hREX-derived conformations) not only encompass the regions explored by the unbiased simulations but also venture into new clusters, revealing additional conformations not captured by the unbiased approach alone. Comparing the two sites, we observe that site 1 exhibits

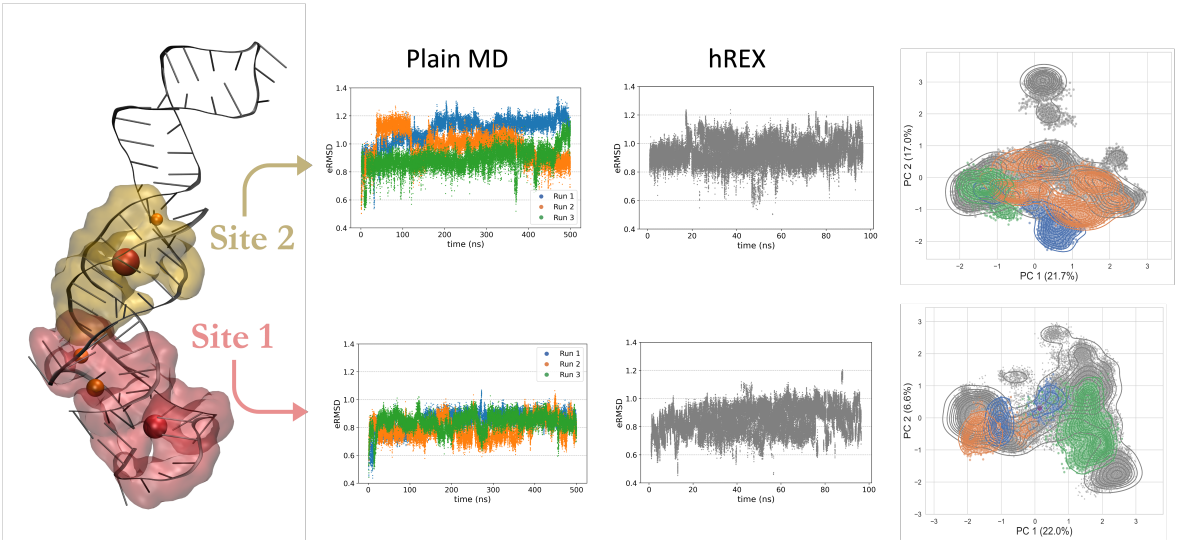

Figure 36: Structural dynamics investigation: eRMSD-based comparison of plain and hREX simulations, complemented by PCA analysis of the joined trajectory for both binding sites.

greater stability in base pairing compared to site 2, whose dynamics are regulated by the triple helix. This is highlighted in the eRMSD time series (referenced against the crystallographic structure), where site 1 spans values between 0.7 and 1.2, whereas site 2 shows higher eRMSD values, excluding replica 3 (green line), reaching up until approximately 1.3. This indicates a greater degree of base pairing rearrangements within the triple helix region of site 2.

Subsequently, the Quality Threshold clustering algorithm was employed to cluster the conformations defined by the g-vectors. Centroids of clusters with an eRMSD cutoff of 0.7 were extracted to generate two initial ensembles. From these, two matrices were constructed for both sites, with each element representing the eRMSD value between the corresponding centroid pairs. This approach enabled us to systematically assess the conformational diversity captured within each centroid and help in the selection of representative structures for subsequent docking studies.

These matrices were subsequently filtered to reduce redundancy among cluster centroids, thereby generating a more manageable set of conformations for downstream docking analysis. An eRMSD cutoff of 0.7 was employed to eliminate frames with lower eRMSD values between each other. This step is necessary because the QT-clustering algorithm generates clusters where points within the same cluster do not have eRMSD values exceeding the cutoff. However, this doesn't guarantee that the eRMSD between centroids of different clusters will also be above the cutoff. This resulted in two refined

matrices, comprising 10 and 5 elements respectively, with eRMSD values spanning a range from 0.7 to 1.4 for Site 1 and form 0.7 to 1.2 for Site 2.

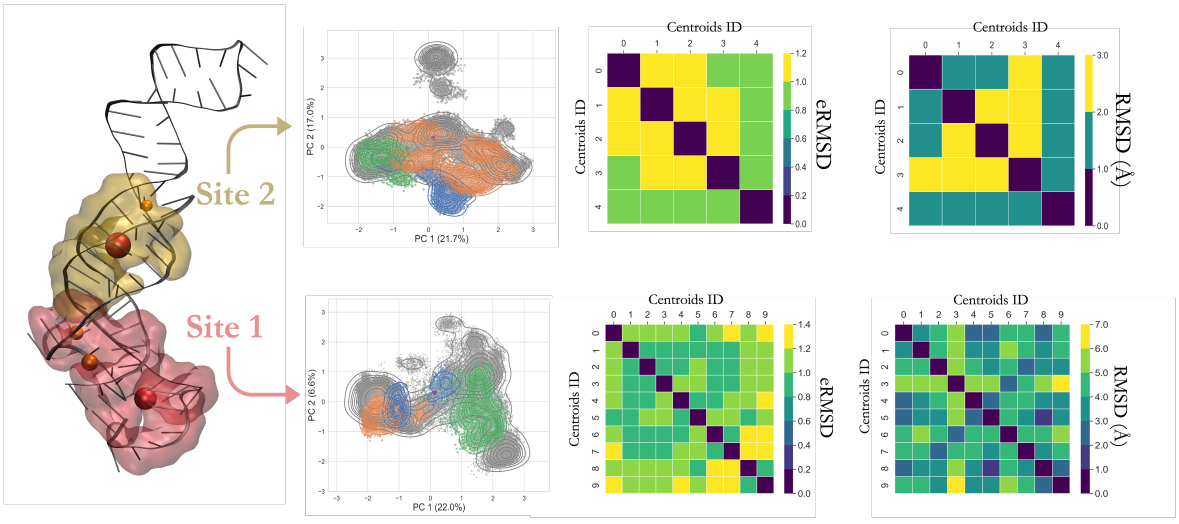

Figure 37: Principal Component Analysis (PCA) of the combined trajectory, supplemented with eRMSD and RMSD matrices calculated from filtered centroids for both binding sites.

Comparing the eRMSD and RMSD matrices reveals a higher flexibility for Site 1, with a wider range of distinct structures found during the analysis and selected. Specifically, Site 1 exhibits eRMSD values reaching up to 1.4 and RMSD values extending to 7 Å. In contrast, the other sites show a lesser conformational diversity, with eRMSD values reaching 1.2 and RMSD values reaching only 3 Å. This reduced flexibility is likely attributed to the inherent stability of the triple helix, conferring greater rigidity to this region compared to Site 1. The refined ensemble effectively captures the full range of conformational dynamics observed in our simulations, while optimizing computational efficiency for subsequent docking studies through careful selection of a minimal, yet representative, set of conformations.

6.3.4 Docking: Pose generation

The second phase of this study involves a docking campaign employing the previously identified conformational ensembles. We then proceeded to dock all 21 small molecules into each of the 16 conformations comprising the two ensembles (10 for Site 1, 5 for Site 2 and the crystallographic structure for both sites). Two docking software were employed: AutoDock GPU (designed for proteins) and rDock (designed for nucleic acids). For each ligand-target pair, 250 poses were generated per software, resulting in

a total of 500 poses for each case.

To accommodate the docking process within the specific sites, different input parameters were required for each software. In AutoDock GPU, a grid box was generated, centered on the center of mass of each site. For Site 1, a box of 110, 110, and 130 points along the x, y, and z dimensions, respectively, was used with a 0.3 Å spacing between points. For Site 2, a box of 110, 80, and 80 points was employed, maintaining the same spacing. The dimensions of the grid box were tailored to each site to ensure comprehensive coverage of the binding pocket identified by Pocketron. For instance, Site 1 required a larger number of points, particularly along the z-dimension, to encompass pocket pID 2 in addition to pIDs 3 and 19.

In contrast, for rDock, the two-sphere method was employed to define the docking cavity. The center of mass of the residues within each site was used as the center point, and a small probe radius of 1 Å and a large probe radius of 17 Å were employed for cavity detection. The large probe radius was carefully selected to ensure that the target regions encompassed were comparable to those used in the AutoDock GPU simulations. Notably the same probe radii were applied to both sites, particularly, the 17 Å large probe radius was chosen based on the size of the AutoDock grid box, effectively treating the sides of the box as the diameter of the large search sphere. This approach was used to prevent overlap between the spheres generated for the two sites, ensuring accurate and independent docking calculations for each site, while being able to explore docking poses for both software in the same regions (Figure 38). To investigate the ability of

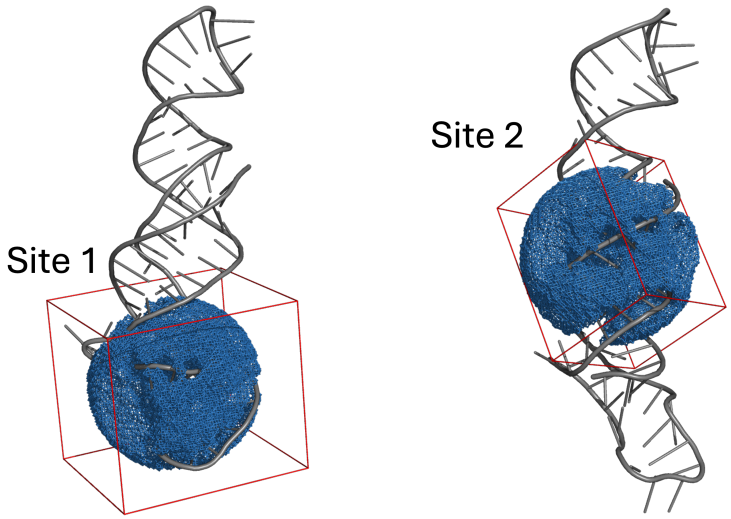

Figure 38: Comparison of AutoDock grid box (in red) and rDock cavity search (in blue) results for both binding sites.

the two docking software to generate diverse and potentially complementary poses, we performed dimensionality reduction on the generated poses for all conformations within each site. Specifically, we focused on the ligand with the highest number of heavy atoms (p13), as its larger size and steric hindrance would pose the greatest challenge for successful accommodation within the target.

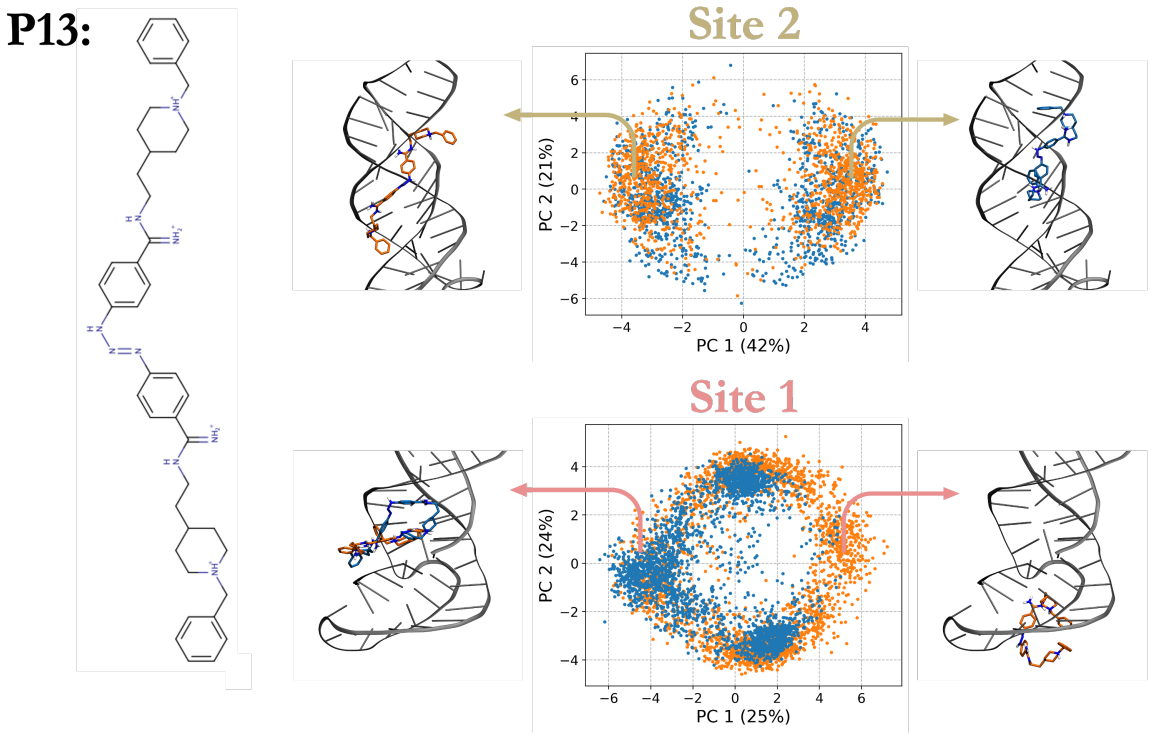

Figure 39: PCA analysis of ligand p13 poses at each site. Representative poses from opposing regions of the PCA plot are shown.

Figure 39 presents a PCA analysis of 5000 poses for ligand p13 across 10 RNA conformations for Site 1 and 2500 poses across 5 RNA conformations for Site 2. While AutoDock GPU (blue dots) and rDock (orange dots) show some overlap in pose generation, rDock (for Site 1) uniquely samples regions of the target not accessed by AutoDock GPU (right snapshot). In contrast, the left snapshot showcases poses from both software that are in good agreement with each other, derived from the same region of the PCA plot. This observation highlights the importance of employing multiple docking software to enhance the ability to capture the full spectrum of potential binding poses.
In contrast, for the more confined Site 2, the generated poses occupy a similar region in the PCA plot, suggesting less diversity between the poses generated by the two software.

The two snapshots from opposite PCA regions highlight distinct binding modes: the left pose is elongated along the major groove, whereas the right pose adopts a more compact, sphere-like conformation within the upper part of the site.

6.3.5 Docking: Rescoring

Having assessed the conformational diversity of the generated poses, we proceeded to rescore them using various scoring functions: AutoDock, rDock, AutoDock Vina, AnnapuRNA and SPRank. The aim was to compare the performance of these scoring functions, particularly those capable of evaluating poses generated by other software, in accurately predicting the experimental affinity trends among the ligands.
For Vina, AnnapuRNA and SPRank, the output poses from AutoDock GPU and rDock were directly used after input format conversion. However, in the case of AutoDock GPU, some poses generated by rDock using the original grid box (used for the pose generation) fell outside the box, leading to artifacts (high score values) due to the specific algorithm employed by rDock in order to generate poses. To address this, the grid box for both sites was extended to 130, 130, and 130 points in the three dimensions. To ensure comparability, not only were rDock poses rescored with this extended grid box, but the AutoDock GPU poses were also rescored using the new grid.
Conversely, for rDock, simply increasing the large sphere radius did not necessarily result in a larger cavity. This is because rDock defines the cavity as grid points accessible to the small probe but not the large probe, meaning solvent-exposed target regions are always excluded. Therefore, we rescored the AutoDock GPU poses using the same probe radii as those employed during pose generation. This ensured consistency in the cavity definition and allowed for a fair comparison of scoring function performance across different docking software.
Upon plotting the scoring function results Figure S10-S13, it becomes evident that the different scoring functions span a wide range of values, with the ML-based ones reaching as low as -500. However, our primary focus lies not in comparing the absolute scores of different scoring functions. Our analysis revealed a trend for poses have more favorable scores when evaluated by their native scoring function. This trend is clearly visible in Figures S10 and S11, where poses generated by AutoDock GPU scored better under AutoDock's scoring function (blue and orange lines), while poses from rDock performed better when evaluated by rDock's scoring function (red and green lines). While both software exhibit this trend, the difference in scores between the two sets is

more pronounced for the rDock scoring function.
Vina (Figure S10 and Figure S11), similar to AutoDock, displays a comparable pattern, albeit less pronounced, with greater overlap between the two lines. This suggests that Vina identifies stable poses derived from both software more frequently. Furthermore, AnnapuRNA (Figure S12 and Figure S13) presents two distinct patterns, in some cases, the lines representing poses from both software overlap significantly within the same score range. However, in other cases, particularly for hindered ligands, it shows numerous unstable poses, generated by AutoDock GPU. This observation highlights that AnnapuRNA, a tool specifically designed for RNA structures, showed perfect compatibility with poses generated by rDock, another RNA-focused software, finding no structural clashes. However, when evaluating poses from AutoDock GPU, AnnapuRNA frequently detected problematic interactions. While, SPRank exhibited the opposite behavior, it consistently assigned higher scores to poses generated by AutoDock GPU rather than rDock.
Then we proceeded to plot the best score for each combination ligand-target coloring the point accordingly to the pose generation software from which are generated. Plotting the top-scoring poses for each software-scoring function combination (Figure 40) confirms our previous observations. Notably, for both Site 1 and Site 2, AutoDock and rDock scoring functions favor poses generated by their respective software, while AnnapuRNA scoring function consistently scores rDock-generated poses more favorably. This suggests that RNA-specific scoring functions, like AnnapuRNA, are more compatible with poses generated by RNA-specific docking software, such as rDock. In contrast, Vina exhibits a site-dependent preference. For Site 1, Vina generally favors rDock-generated poses, with a few instances where AutoDock GPU poses are preferred (blue dots). Conversely, for Site 2, Vina primarily prefers AutoDock GPU poses, with occasional instances of favoring rDock poses. Similarly, SPRank favored mainly poses generated by AutoDock GPU for site 2, while for site 1, the top-scoring poses are originated from both AutoDock GPU and rDock.
To investigate the influence of different target conformations within the ensemble on the docking results, we generated plots similar to Figure 40, but with points color-coded according to the specific conformation from which they were derived (with the crystallographic structure results in black). This allowed us to visually assess the impact of conformational variability on the predicted binding poses and scoring function performance for both sites. By examining these plots, we aimed to determine if any specific scoring function, in conjunction with any particular conformation for either

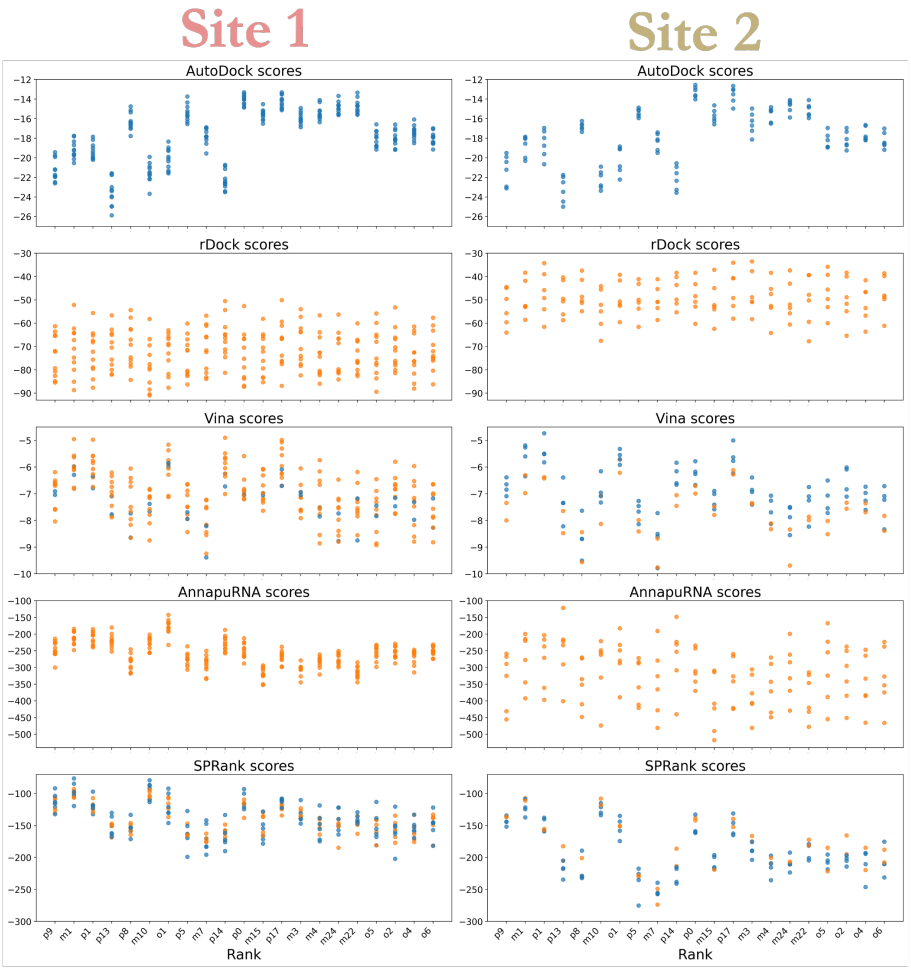

Figure 40: Evaluation of ligand-target binding using multiple scoring functions for both sites. Top-ranked poses for each conformer are color-coded by docking software: AutoDock GPU (blue), rDock (orange).

site, could reproduce the experimental affinity trends. Analyzing the best scores for each ligand, color-coded according to the target conformation from which they were derived (with the results from the crystallographic structure in black), reveals distinct patterns for Site 1 and Site 2. In the case of Site 1, the best scores originate from various conformations within the ensemble, however, conformer 10 (cyan in Figure 41) consistently yields top scores across multiple scoring functions.

Conversely, Site 2 demonstrates a higher degree of consistency in terms of the conformations associated with the best scores. For both AutoDock and Vina, a mix of conformations 1, 3 and the crystallographic predominantly produce the most favorable scores. Conformer 4 consistently ranked as the optimal conformation across all ligands

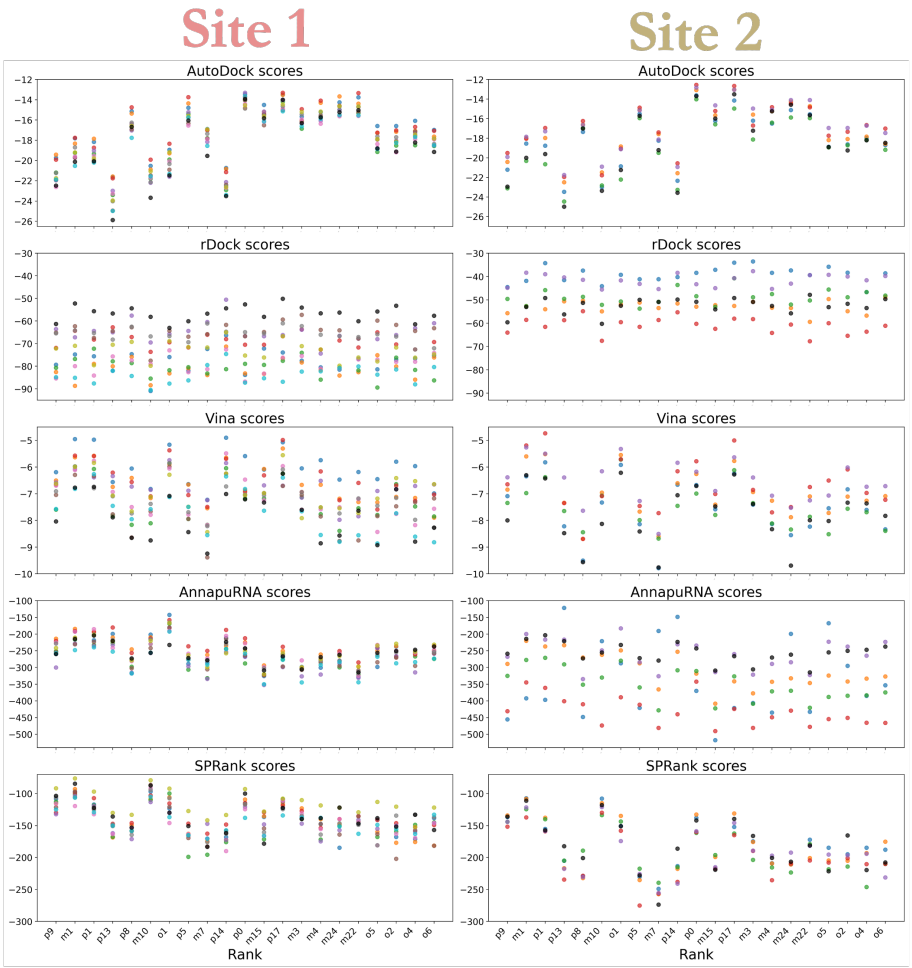

Figure 41: Evaluation of ligand-target binding using multiple scoring functions. Top-ranked poses for each target's conformation with colors indicating their respective ensemble element for both sites.

when scored with rDock. This preference for conformer 4 was also largely observed with AnnapuRNA, although occasionally interspersed with conformer 1. On the other hand, SPRank favored conformer 4 and 3, appearing among the top-ranked conformations in the majority of the cases but with also some cases derived from the other conformations not observed anywhere else.

Figure 42 visualizes the previously discussed observations by highlighting the best score for each ligand in each site across different scoring functions, ensuring a consistent y-axis range for comparisons between sites. Although a direct comparison between the two sites doesn't provide additional information, it is evident that, except for AutoDock, no discernible trend between (at least) the best- and worst-affinity ligands

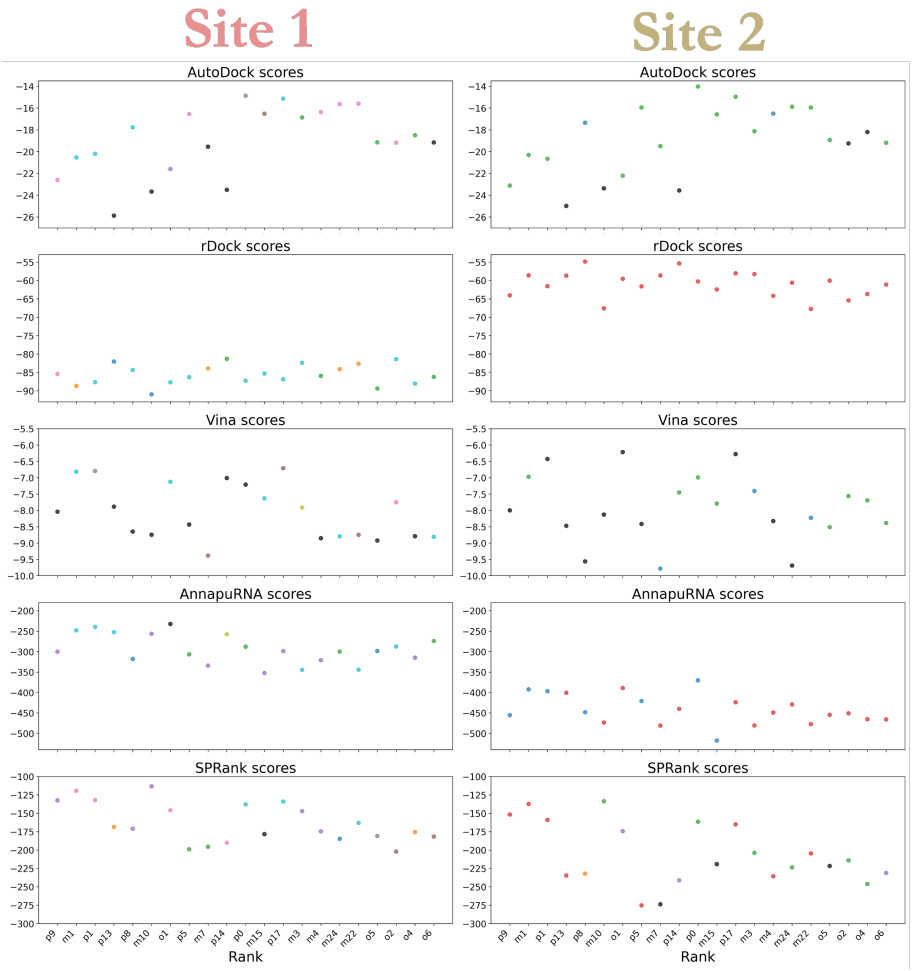

Figure 42: Evaluation of ligand-target binding using multiple scoring functions. Top-ranked pose for each conformation is extracted with colors indicating their respective element in the ensemble for both sites.

is observed. In the case of AutoDock, a slight distinction emerges between the left and right regions of the plot, suggesting potential ligand affinity differences, though not a direct reproduction of experimental results. This observation suggests that AutoDock scores could be used to distinguish ligands with potentially higher or lower affinities for the target. Additionally, the AutoDock results for both binding sites exhibit a remarkably similar scoring trend across ligands, a pattern not observed with any other scoring function. This implies that the target's contribution to the AutoDock score may be largely independent of the specific binding site, highlighting a potential limitation of this scoring function in accurately capturing site-specific interactions. The similarity in scoring results across different sites may be rationalized by examining the

chemical composition of the binding regions. Both sites exhibit a similar nucleotide architecture, dominated by A-U base pairs with two G-C pairs in the middle. This compositional similarity is particularly relevant for force-field based scoring functions, which rely heavily on atomic partial charges and van der Waals parameters. While AutoDock showed the strongest site-independent behavior, similar scoring patterns emerged, though less pronounced, when comparing Vina and SPRank results across different binding sites. Particularly interesting is the correlation observed within Site 1 between AnnapuRNA and SPRank scores. This partial reproducibility of scoring trends across different methods could suggest that certain ligand properties might dominate the scoring outcomes, regardless of the specific algorithm used. However, the more moderate nature of these correlations, compared to AutoDock, indicates that these scoring functions maintain some degree of site-specificity in their evaluations.

In conclusion, our analysis is consistent with previous findings [240], reinforcing the known limitations in accurately generating and evaluating docking poses for nucleic acid-ligand complexes. The sensitivity of scoring functions to initial target conformations, along with the challenges associated with ligand and target flexibility, highlights the need for further methodological improvements in this area. Moreover, the scarcity of high-quality experimental binding affinity data for nucleic acid-ligand complexes remains a significant hurdle in the development and validation of robust scoring functions.

Our work contributes with addressing these challenges by emphasizing the importance of incorporating conformational dynamics and employing diverse scoring approaches in RNA-ligand docking studies. Future efforts in this field should prioritize the generation of more comprehensive and diverse experimental datasets, as well as the development of novel scoring functions that explicitly account for the unique characteristics of nucleic acid-ligand interactions.

7 Application Case 3: Non equilibrium binding free energy estimation

7.1 Introduction

7.1.1 Biological relevance of the target

The binding free energy (ΔF_b) quantifies the affinity of a potential drug for its biological target, making ΔF_b a central thermodynamic observable in drug discovery campaigns [53]. Over the past few decades, numerous computational methods have been developed to estimate ΔF_b, yet accurately predicting this parameter remains challenging in many cases. The difficulties stems from various aspects, including the high degrees of freedom of the systems and the inherent flexibility of both the receptor and ligand.

Motivated by the goal of developing a systematic protocol for computing binding free energy with path-based methods, a semi-automatic computational workflow successful on systems of moderate size was designed by our group. [241] However, challenges emerged when dealing with intricate systems, such as the RNAs complex, where a very high dissipation may occur and, hence poor convergence of free energy estimates. This prompted the development of a refined strategy to mitigate dissipation, particularly relevant for improving convergence in complex systems. The possibility of targeting RNAs with small molecules to achieve therapeutic effects has recently gained increasing interest. Nevertheless, peculiar features of RNA molecules, such as the complex structural dynamics or the high charge density that hinders the capability of forming hydrophobic binding sites, make the transition from protein to RNA targets particularly challenging from both the experimental and computational standpoint. Thus, further investigations in this respect are highly desirable and may promote tangible advancements towards the design of RNA-targeting small-molecule drugs.

Within this context, Riboswitches represent a compelling class of RNA molecules (FMN [242], TPP [243] and preQ1 [244]), since they are able to bind metabolites and modulate gene expression as a result. In particular, riboswitches typically display complex three-dimensional structure by adopting non-trivial folding, and are able to form binding pockets with varying levels of complexity to host the native ligands. This

natural disposition to bind small molecules make them suitable targets for the rational design of small molecules binders. A concrete example in this respect is the campaign against the bacterial FMN riboswitch, leading to a compound at the preclinical stage. Herein, we test our approach on the preQ1 riboswitch. Given the availability of crystal structures in complex with different ligands and experimental data quantifying their affinity, this system represents an optimal test case. In particular, the available data refer to two different ligands: the cognate ligand PreQ1 (7-aminomethyl-7-deazaguanine) and a synthetic ligand (2-[(dibenzo[b,d]furan-2-yl)oxy]-N,N-dimethylethan-1-amine), [245, 244] here referred to as cognate and dibenzofuran for conciseness, respectively.

7.1.2 Function and binding mode of Riboswitch-PreQ1

Riboswitches, typically found in the 5'-UTR (untranslated regions) of mRNA, are genetic regulatory elements prevalent in bacteria [246, 247, 248]. These RNA elements function by binding specific small molecules, independent of protein factors, and subsequently control gene expression by inducing conformational changes in the mRNA. Primarily located upstream of bacterial mRNAs encoding biosynthetic enzymes or metabolite transporters, riboswitches consist of two domains: an aptamer domain responsible for ligand binding and an expression platform that regulates downstream gene expression (Figure 43) [244].

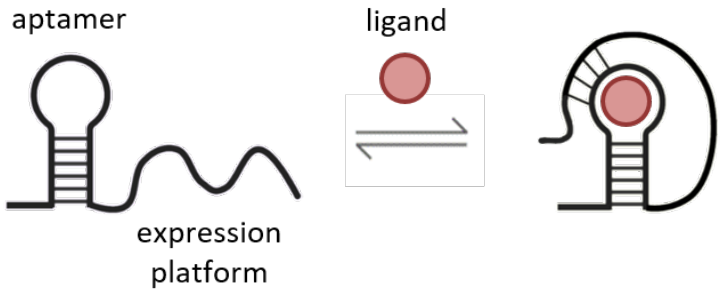

Figure 43: Schematic representation of a generic riboswitch in its unbound and ligand-bound conformations.

Upon ligand binding to the aptamer domain, the riboswitch undergoes a significant conformational change (as illustrated in Figure 43). This structural rearrangement stabilizes specific elements within the expression platform, triggering regulatory effects on both transcription and translation mechanisms. The versatility of riboswitches

is demonstrated by their ability to detect and respond to a diverse array of cellular metabolites, such as pre-quenosine (preQ1) - a molecule synthesized from GTP through a complex multienzyme pathway [249, 250].
The preQ1 riboswitch, characterized by its three loops and two stems (3D structure depicted in Figure 44A, color-coded by secondary structure), creates a binding pocket organized into three distinct layers.

- Ceiling: The top layer where a G-C base pair at the bottom of stem 2 forms an C-G-C triple above the binding core (Figure 44B).
- Binding core: The central layer where preQ1 forms hydrogen bonds with one residue from each of the three loops, creating a preQ1-C-A-U base quartet (Figure 44C).
- Floor: The bottom layer where a C-G base pair at the top of stem 1 forms an A-C-G-A base quartet with two adjacent adenine from loop 3 (Figure 44D).

Within the binding core, preQ1 specificity is maintained by hydrogen bonds involving all of its proton donors and acceptors. C15 in loop 2 forms a standard Watson-Crick base pair, while A29 in loop 3 and U6 in loop 1 interact with the sugar edge of preQ1. Although a bound sulfate ion sterically hinders the interaction between preQ1's exocyclic amine and G5 in WT crystal structure, substitution of G5 significantly reduces binding affinity, suggesting its importance. This interaction, along with the base quartet and triple above and below the binding core, stabilizes preQ1 through stacking interactions and limits its exit to the major groove side.

7.1.3 Methodology's workflow

Our research introduces an enhanced computational approach incorporating advanced sampling techniques and principal component vectors (PCVs). This novel method employs S(X) (explained in detail in chapter 3.2.1) as collective variable to avoid inherently sequential algorithms like Metadynamics, opting for steered molecular dynamics (SMD) for sampling instead. As SMD is a non-equilibrium method, the Crooks Fluctuation Theorem (CFT) [251] is employed to reconstruct the potential of mean force (PMF) and estimate binding free energy. This integration of PCVs with bi-directional SMD and CFT significantly improves convergence speed and enables easy parallelization of simulations, consistently improving the computational cost.
Our recent computational pipeline is followed to run non-equilibrium SMD simulations,

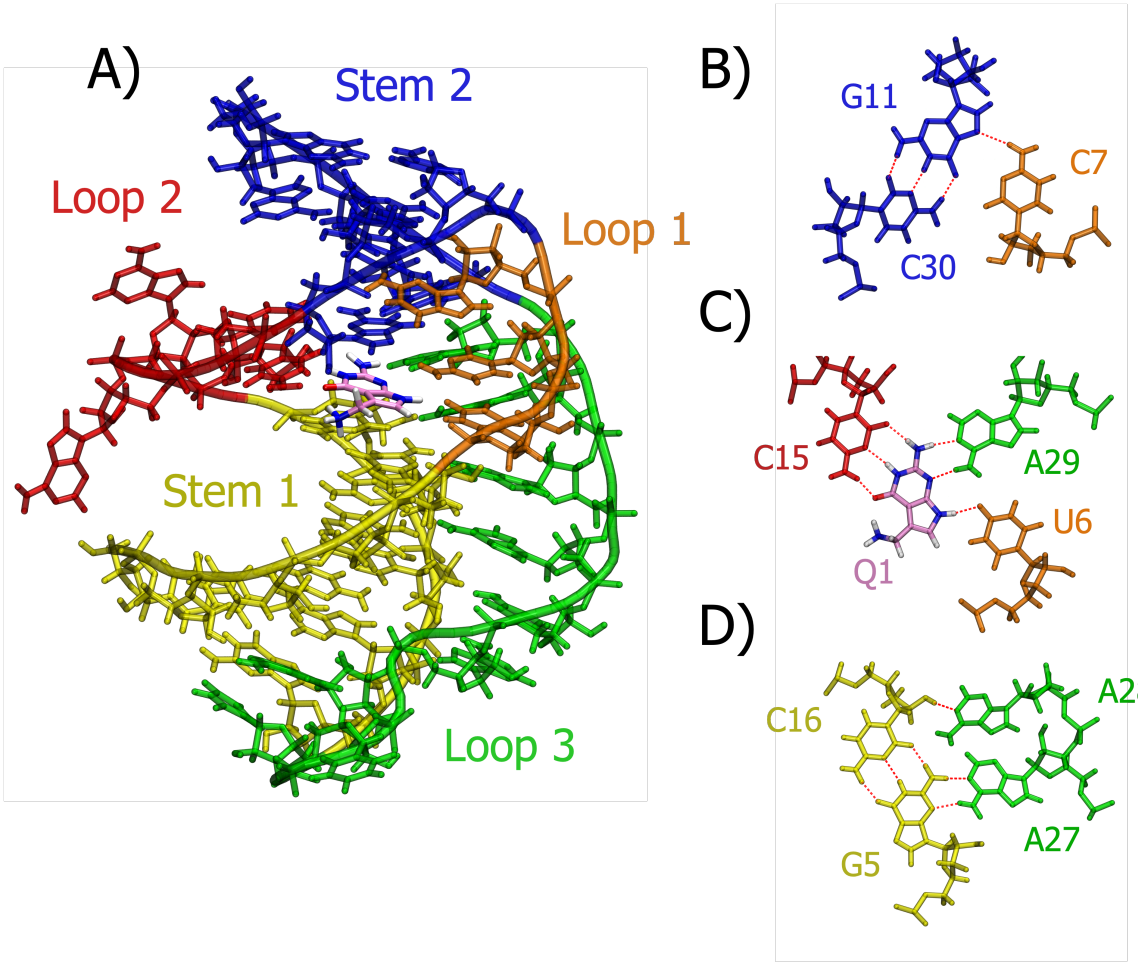

Figure 44: Binding pocket and binding mode of Riboswitch preQ1 with its cognate ligand (Q1). A) 3D structure with residues colored according to their secondary structure. B-D) Stick representation of the three layers (ceiling, binding core, and floor) of the binding pocket.

reconstruct the PMF, i.e. the free energy as a function of the collective variable used to perform the enhanced sampling simulation, and finally estimate the binding free energy difference of each complex studied.

1. Generation of unbinding MD trajectories starting from a bound complex through Adiabatic Bias MD (ABMD) [252] to promote target-ligand dissociation. Among these trajectories, the one with the smoothest unbinding path (with no rolling around the structure or going back) with a plausible mechanism is selected as guess path.

2. The guess path is optimized using two path algorithms, the principal path algorithm [253] and the equidistant waypoints algorithm. Consequently, the reference path

for PCVs is obtained.

3. Non-equilibrium bidirectional (binding and unbinding) SMD simulations are performed following the reference path with PCVs, collecting the Jarzynski work (W_J) performed.

4. By applying the CFT to the W_J values collected during binding and unbinding SMD simulations, the PMF is reconstructed.

5. The binding free energy is calculated by applying the Crooks Fluctuation Theorem and adding the correction for the desolvatation of the ligands and the results are then compared with the one extracted via Metadynamics simulation.

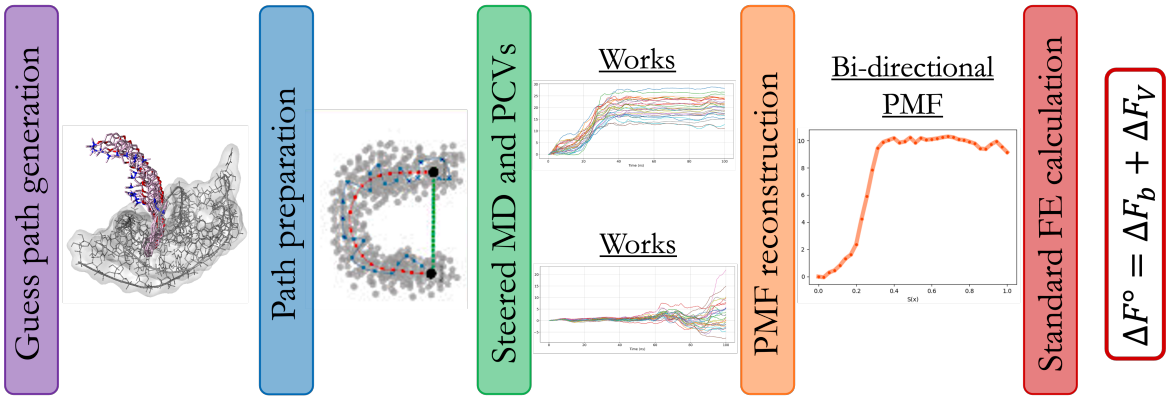

Figure 45: Schematic representation of the workflow employed for this work

7.2 Methods

7.2.1 System setup

The RNA-ligand complexes were modelled from PDB IDs 6E1W and 6E1U for the cognate and synthetic ligand, respectively. Ligands were parametrized according to the GAFF force field and AM1-BCC charges, considering a total charge of +1, due to presence of a quaternary nitrogen in both. The systems were then solvated with TIP4P-D water model and charges neutralized with Na^+ and Cl^- ions, using a concentration of 0.15 M. The solvated systems were then equilibrated in two steps. First, 100 ps of MD were conducted at 300K in the NVT ensemble, with positional restraints of 1000 kJ $\cdot$ mol^{-1} $\cdot$ nm^{-2} on all heavy atoms were and temperature control via the Bussi-

Parrinello [254] velocity rescaling thermostat. Second, a 100 ps MD was conducted in the NPT ensemble applying the same restraints, with the Berendsen barostat for pressure control[255].
Topology and starting structure for the PreQ1 riboswitch in complex with the two ligands, based on the PDB IDs 6E1U and 6E1W, were taken from a recent work.[244] For the cognate ligand, the complex with the alternative tautomeric state was constructed by replacing the newly parametrized ligand in such form with the other one in the original structure. Further re-parameterization of the cognate ligand with Restrained Electrostatic Potential (RESP) charges and optimized dihedral parameters was achieved via the PlayMolecule web server [256]. Specifically, RESP charges were derived via quantum mechanism using the $6-311{+}{+}G^{**}$ basis set at the wB97X-D level of theory, while optimization of dihedral angle parameters via neural network potentials using the xTB fitting method.

7.2.2 Adaptively Biased Molecular Dynamics

This phase focused on generating ligand unbinding trajectories from the RNA binding pocket using enhanced sampling molecular dynamics (MD) simulations, specifically the ABMD method. We relied on the intrinsic electrostatics of the RNA-ligand system, using the Debye-Huckel interaction energy as a collective variable (CV) to guide the dissociation process. This strategy steered the simulations towards a state of zero electrostatic interaction between the RNA and ligand, effectively representing complete unbinding.
Ten ABMD replicate simulations, each lasting 10 ns, were performed for each system. Analysis of these simulations revealed consistent unbinding pathways for both ligands, with dissociation occurring through the solvent-exposed side of the binding pocket between loop 2 and stem 1. The force constant for ABMD was carefully tuned to ensure a smooth and controlled ligand dissociation, minimizing undesirable rolling motions on the RNA surface (approximately in the order of 10^{-1} and 10^{-2} kcal/mol*C^2 for the cognate and synthetic ligand, respectively). The final ABMD trajectories were truncated to include only frames where the ligand remained within 20 Å of the RNA, guaranteeing a fully solvated unbound state.
From these replicates, a representative ABMD trajectory was selected as an initial guess for the unbinding pathway, considering both the unbinding time and the mechanistic plausibility. From these replicates, a representative ABMD trajectory was selected as

an initial guess for the unbinding pathway, considering both the unbinding time and the mechanistic plausibility. The initial path underwent a two-step refinement process. First, we applied a principal path algorithm to optimize the trajectory configurational space [253]. Second, we implemented an equidistant waypoint algorithm to maintain uniform spacing between consecutive conformations [257]. This refined approach produced a smooth unbinding pathway with consistently spaced frames, characterized by an average RMSD of approximately 1 Å between consecutive conformations.

7.2.3 SMD Simulations and PMF reconstruction

According to the protocol, the reference minimum free energy path is followed with PCVs during bidirectional non-equilibrium SMD simulations. As mentioned (Chapter 3.2.2), in SMD a time-dependent harmonic restraint $R(x,t)$ is added to the potential of the system:

$$R(x,t) \equiv \frac{1}{2}k(S(x) - \hat{S}(t))^2 \tag{39}$$

In detail, the harmonic restraint is applied along the pulling coordinate, representing the progress along the reference path. Additionally, harmonic walls are employed to confine the system along $Z(x)$, activating only when a $Z(x)$ value surpasses the threshold of 0.05.

The amount of work performed on the system during the transformation is known as the Jarzynski work W_J [258, 259] as is the path integral of $\dot{\xi}\partial H_\xi/\partial\xi$ along the trajectory Γ_t:

$$W_J = \int_0^{t_s} \mathrm{d}t\ \dot{\xi}\frac{\partial H_\xi}{\partial \xi}(\Gamma_t) \tag{40}$$

where H_ξ represents the time-dependent part of Hamiltonian that is added to the regular potential in SMD and ξ is the time-varying variable. The total Jarzynski work (W_J) of an irreversible simulation accounts for the energy required to transition the system from the initial to the final state (or vice versa).

According to the second law of thermodynamics, in a quasi-static transformation, this amount of work corresponds to the free energy difference ΔF_{AB}. However, for non-equilibrium irreversible transformations, the total Jarzynski work will, on average, exceed the free energy difference by a quantity known as the dissipated work:

$$W_J^{diss} = \langle W_J \rangle - \Delta F_{AB} \geq 0 \tag{41}$$

The higher the pulling speed of SMDs of a system, the greater the total W_J, and in turn the higher the dissipated work. Multiple binding and unbinding simulations with a constant pulling speed are necessary to achieve convergence in free energy estimation, in particular 30 replicas of binding and 30 of unbinding simulations with a pulling constant setted at 20 kJ/mol for a total simulation time of 100 ns each.
Free energy profiles along the path collective variable (PCV) $S(x)$ are reconstructed from the work values obtained in binding and/or unbinding simulations. Two estimators are employed for this reconstruction: the unidirectional Jarzynski estimator (applied to both forward and reverse transformations) [75] and the Bennett Acceptance Ratio (BAR) estimator. In detail, we rely on the automatic procedure we devised leveraging the CFT in a maximum likelihood form [260].

$$e^{-\beta \Delta F_{AB}} = \langle e^{-\beta W_J^x} \rangle \tag{42}$$

$$\frac{P_f(W_J)}{P_b(-W_J)} = e^{[\beta (W_J \ - \ \Delta F_{AB})]} \tag{43}$$

The equations above illustrate the conversion of work values into a potential of mean force (PMF) as a function of the collective variable S(x). They are solved iteratively for each interval S_i, S_{i+1}, where the PMF point $F(S_{i+1})$ is calculated as $F(S_i) + \Delta F(S_i, Si + 1)$. Here, the index i denotes the i-th configuration along the reference pathway
Equation 42 represents the unidirectional PMF using the Jarzynski estimator, where x can correspond to either forward or backward simulations. Equation 43, on the other hand, represents the bidirectional non-equilibrium estimator derived from the minimum of the maximum likelihood of the Bennett acceptance ratio method, employing both forward and backward work profiles for PMF reconstruction.

7.2.4 Standard Binding Free Energy Estimation

Standard binding free energies can be determined as the sum of the binding free energy (ΔF_b) and the standard volume correction term (ΔF_v) as described by Doudou *et al.* [261]:

$$\Delta F_b^\circ = \Delta F_b - \Delta F_v = -RTln\frac{Q_{site}}{Q_{bulk}} - RTln\frac{V_{bulk}}{V^\circ} \tag{44}$$

ΔF_b is the ratio between the probabilities of the bound and unbound ligand states, i.e. of the canonical partition functions of the bound (Q_{site}) and unbound (Q_{bulk}) states.

To compute this, the PMF is integrated in both the bound and unbound region:

$$\frac{Q_{site}}{Q_{bulk}} = \frac{\int_{site} \exp\left(-\frac{F(S)}{RT}\right) dS}{\int_{bulk} \exp\left(-\frac{F(S)}{RT}\right) dS} \tag{45}$$

where $F(S)$ is the PMF along the $S(x)$ PCV. The value of $S(x)$ discriminating between the bound and unbound region required to perform this integration is determined via visual inspection of the PMF (when it starts to decrease right after the maximum) and of the predicted path. The standard volume correction ΔF_v quantifies the variation of the free energy due to considering the standard-state volume V° corresponding to 1661 Å^3 (concentration of 1 M) instead of the effectively sampled unbound volume V_{bulk}. This contribution is computed through NanoShaper as in our previous work [165].
By adding this correction term to ΔF_b, standard binding free energy differences directly comparable with experimental values can be determined. The same procedure was used to estimate ΔF_b from the MetaD simulations, after reconstructing the free energy profile along the $S(x)$ PCV. Finally, errors associated with standard binding free energy are calculated via bootstrapping [262] employing 500 bootstrap iterations.

7.2.5 Metadynamics setup

The deposition time of Gaussians was set to 250 MD steps, and a bias factor of 15 was used. Gaussian height of 1.0 kcal/mol was used, while Gaussian width was set to 0.2 nm^2 along $S(x)$ and 0.01 nm^2 along $Z(x)$. Moreover, the available space along the $Z(x)$ PCV was restricted by placing a wall at $Z(x)$ equal to 0.1 nm^2 with force constant of 40000000.0 kJ/mol.
Convergence was determined based on two conditions: the system's full diffusivity along the PCV $S(x)$, and the residual Gaussian height being less than 10 % of the initial height, aligning with previous methodologies [257]. Post-processing analysis involved reconstructing the PMFs, with a free energy of zero defined at their lowest point corresponding to the ligand bound state. After reconstructing the free energy profile along the $S(x)$ PCV, the same procedure used to estimate ΔF_b from the SMD simulations was used.
Statistical errors associated with standard binding free energy are calculated via bootstrapping [262] employing 500 bootstrap iterations.

7.3 Results and Discussions

7.3.1 Path Definition

A critical major difference between protein and RNA systems is the greater structural flexibility displayed by the latter. Given such peculiar features of the RNA target, the definition of the pulling pathway for the RNA-ligand complexes required dedicated adjustments. The optimal alignment of the target structure on the reference pathway is a crucial prerequisite for meaningful mapping of trajectory frames along the S(x) PCV. This can be compromised when including highly flexible regions of the target structure in the alignment. Thus, this can be particularly critical in the case of the RNA systems. In order to mitigate this effect, we only included the most stable residues in the RNA structure. To identify these regions, we ran three independent, unbiased MD simulations of 100 ns each. The trajectories were then combined to compute the residue-wise root-mean-square-fluctuation (RMSF) (Figure 46).

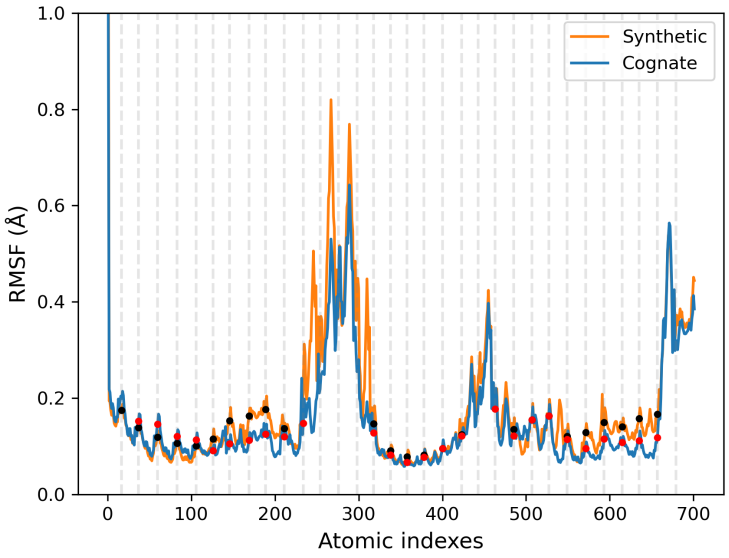

Figure 46: RMSF (Root Mean Square Fluctuation) analysis for cognate (blue) and synthetic (orange) ligands calculated across three combined 100 ns unbiased trajectories. Black and red dots represent phosphate atoms of residues with RMSF < 0.18 Å, included in the unbinding pathway.

This allowed us identifying low-fluctuating residues, specifically, those displaying RMSF below a 1.8 Å. We used these low-mobility residues as alignment references for the final path, systematically excluding both high-fluctuation regions and their neighboring

residues to ensure robust structural comparisons, thus minimizing potential artifacts in the mapping of the S(x) path-CV arising from regions of the RNA displaying higher structural flexibility.
Additionally, to remove possible noise associated with a large number of atoms contributing to both the alignment and calculation of the S(x) path-CV, we reduced the number of atoms included in this task. To this end, we applied a coarse-grained like selection of RNA residue atoms (Figure 47).

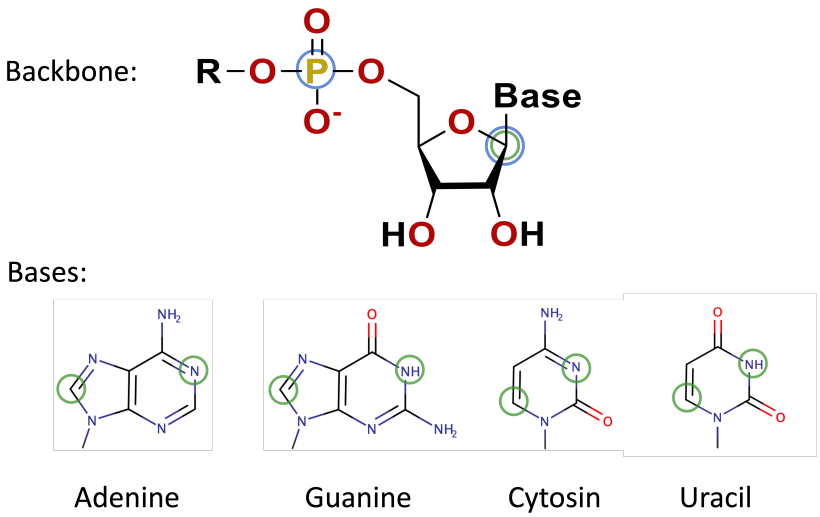

Figure 47: Schematic representation of the selected atoms for the coursed grain like unbinding path. It is represented in orange the atom selection for the alignment and in yellow the ones for the RMSD calculation

Specifically, for alignment we included phosphate and C1' carbon atoms from the RNA backbone. Notably, this is also conceptually consistent with the selection used in a previous work of our group for the protein system [241], where only C_α carbon atoms from the protein backbone are used for alignment. Moreover, the unbinding simulations via ABMD revealed notable structural rearrangements of RNA residues during the dissociation process, which needed to be carefully taken into account in the pathway definition. Therefore, in the guess path we also included all the RNA residues displaying a distance within a 6 Å from the ligand in the ABMD unbinding trajectories. In particular, concerning the calculation of the RMSD, N1 and C8 atoms of purines and C6 and N3 atoms of the pyrimidines were considered, in addition to the C1' carbon backbone atom used for the alignment (Figure 47).
This process yielded an equidistant unbinding pathway, comprising 31 to 37 frames in total, with an average RMSD of approximately 1 Å between consecutive frames. These

pathways are illustrated in Figure 48.

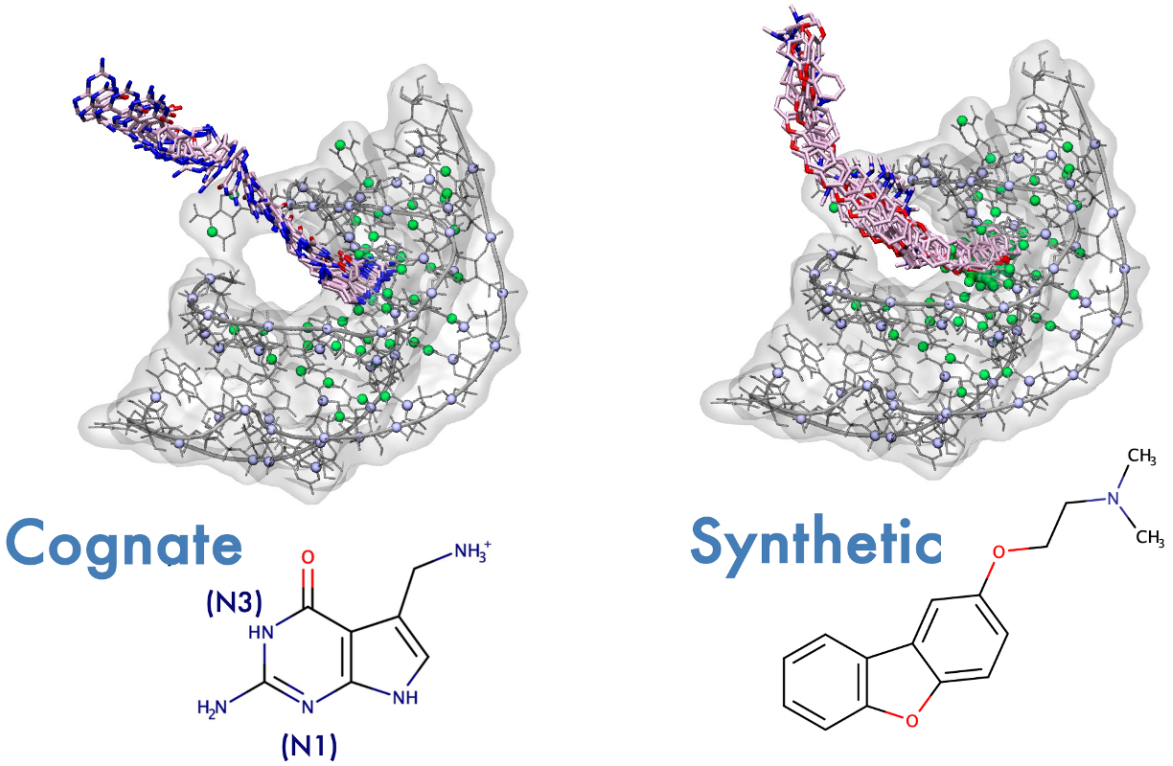

Figure 48: Unbinding pathways for both ligands, visualized with licorice representations (hydrogens omitted). Cyan dots indicate alignment atoms, green dots represent atoms used for RMSD calculation.

7.3.2 Water model effect on binding free energy estimates

After optimizing the path for the RNA-ligand system, we performed 30 replicates of SMD binding and unbinding simulations. The path definition we devised proved to be suitable for the RNA-ligand systems. This could be appreciated by the higher degree of consistency between work profiles in the replicates, demonstrating adherence to the reference path, in comparison with the initial attempts. Accurately capturing the structural flexibility of the target is a critical aspect for correctly describing the process, and in turn obtain reliable results, evidenced by smooth work profiles without abrupt increases. This becomes particularly relevant when complex conformational rearrangements upon (un)binding are involved. Thus, while taking this into account becomes essential for the highly flexible RNA molecules, it is particularly important

in the case of the RNA-ligand systems considered herein, were ligand (un)binding is associated with rearrangement in the stem 1 comprising residues 12 to 16. Such region establishes an intricate network of hydrogen bonds with the cognate ligand and adopts specific conformations of residue C15 to accommodate the bulkier synthetic ligand in the binding site (Figure S14).

As typically observed, inspection of the work profiles revealed that work in the unbinding simulations was smoother and increasing nearly monotonically (Figure 49). This indicates presence of a significant energy barrier that needs to be overcome by the ligand to disrupt the interactions of the bound state and escape the binding pocket, and is higher in the case of the cognate ligand. Once complete detachment of the ligand from the target is achieved, the work profile reaches a plateau at $S(x)$ around 0.5, indicating the ligand has reached the unbound state and is fully solvated.

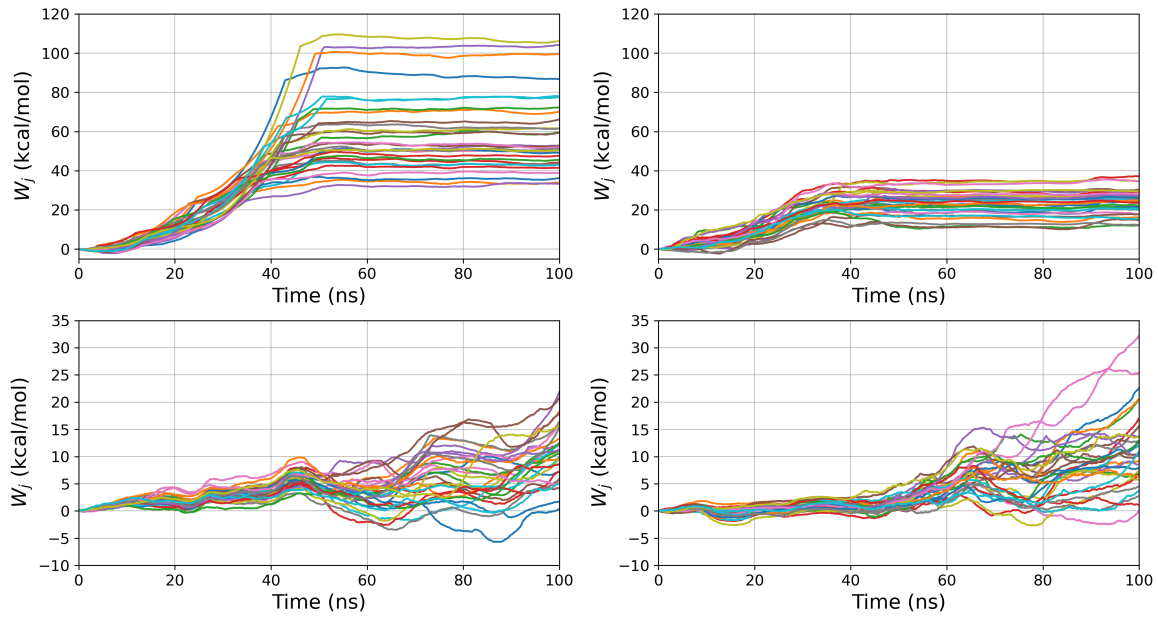

Figure 49: Jarzynski work profiles for the cognate ligand (A) and synthetic ligand (B) regarding unbinding (upper section) and binding (lower section) simulations measured over a simulation time of 100 ns in TIP4P-D.

Conversely, the work profiles obtained from the SMD binding simulations displayed a rather consistent trend for both ligands at the beginning of the simulations, which tended to diverge in the different replicates as the ligand approached the RNA target and eventually the bound state, specifically from simulation time 40 ns, corresponding to $S(x)$ of about 0.4 (Figure 49). Such work fluctuations in the second part of the binding simulations reflects the complexity of the ligand association process, which

includes remarkable conformational rearrangements required by the target for properly accommodating the ligand. Notably, the final stages of the binding simulations displayed in general only a slight increase in the work, indicating minor difficulties for the ligand to precisely adopt the known binding mode in the crystal structure once the binding site is reached. Comparing the binding work profiles between the two ligands, the process appeared slightly smoother for the synthetic ligand as it approached the RNA target (Figure 49B, lower panel), with few replicates terminating with higher work values, suggesting a deviation from the reference path and, particularly, the crystallographic bound state of this ligand. Indeed, this indicates the ligand preference for a bound state that is slightly different than the reference one included in the predicted pathway.

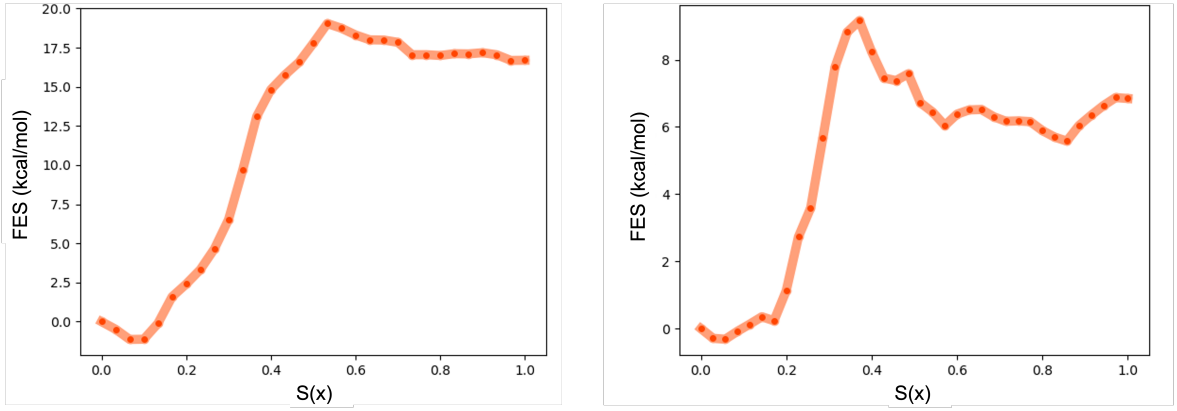

Figure 50: Free energy profiles along S(x) obtained by applying CFT to SMD (simulation time of 100 ns in TIP4P-D) for the cognate (A) and synthetic (B) ligands

Figure 50 shows the FE profile calculated applying the CFT on the SMD simulations of both ligands in TIP4P-D waters. Basing on the profile of the free energy, we selected a discriminating region corresponding to values of S(x) = 0.6 and 0.4 for the cognate and synthetic ligands, respectively. Our results were compared with experimental affinity data from recent works in the literature [244, 245].

The binding free energy calculations showed varying levels of agreement with experimental values. For the synthetic ligand, we calculated a ΔF of -5.6 ± 1.1 kcal/mol, deviating by approximately 2 kcal/mol from the experimental measurement of -7.9 kcal/mol. The cognate ligand showed a more substantial discrepancy: our calculated value of -17.2 ± 1.2 kcal/mol differed by about 6 kcal/mol from the experimental reference of -10.9 kcal/mol. Although the CFT-derived free energy profile for the cognate ligand appeared reasonable, further analysis of the binding and unbinding PMFs (calculated via the Jarzynski equality, Figure S15A) revealed problematic features. Specifically, the binding

PMF showed considerable scatter and poorly defined bound and unbound states, likely explaining the inaccurate free energy estimate.
Notably, all simulations were conducted with the TIP4P-D water model, the four-site model typically used in combination with the DESRES force field for RNA. To investigate the possibility of an effect in the estimate of the binding free energy due to the use of this water model in our non-equilibrium SMD simulations, we repeated all simulations with the most popular and less computationally demanding TIP3P model [263], using the same reference path. Thus, the systems were re-solvated with TIP3P waters, equilibrated and, subsequently, production SMD simulations were conducted using the same setup employed for the TIP4P-D case.
Interestingly, this revealed a non-negligible effect on the obtained work profiles (Figure 51) and free-energy values (Figure 52), with a different impact for the two ligands. The results for the synthetic ligand (Figure 51B) revealed no significant differences compared to the TIP4P-D results (Figure 49B). In particular, we observed a highly similar trend in the unbinding simulations, while the binding simulation replicates displayed a slightly more consistent profile between them, with no relevant barrier identified until the final 10 ns, i.e. S(x)=0.9, where the binding pocket residues need to rearrange in order to accommodate the bulky ligand and the necessary work rises as a result.
Concerning the cognate ligand (Figure 51A), also here the unbinding replicates exhibited a pattern consistent with the simulations in TIP4P-D, with an small barrier at about 40 ns, i.e. S=0.4, followed by a plateau region. Conversely, a more marked difference was observed in the binding simulations. Specifically, the simulations showed that lower work was necessary to reach the bound state if compared with the results in TIP4P-D (see Figure 49A), with some replicate undergoing notable stabilization close the bound state. Furthermore, within this fluctuating region, the majority of the replicate demonstrated consistent trend, suggesting a stricter adherence to the reference pathway. This led to an improvement in the trend of the corresponding PMF derived from JE (Figure S16A).
Interestingly, simulations with different water models revealed conformational differences in the final state of the RNA target. Specifically, stem 1 (residues 12-16) failed to adopt the reference pathway and crystal binding mode when using the TIP4P-D model.
Accordingly, the CFT-estimated binding free energy differed from those obtained with the two models. In particular and interestingly, the improved work profiles for the synthetic ligand translated into a calculated ΔF of -8.7 $\pm$ 0.7 kcal/mol (discriminating frame: S(x)=0.4), around 3 kcal/mol lower than the value obtained in TIP4P-D, and showing a remarkably improved agreement with the reference experimental data (-7.9

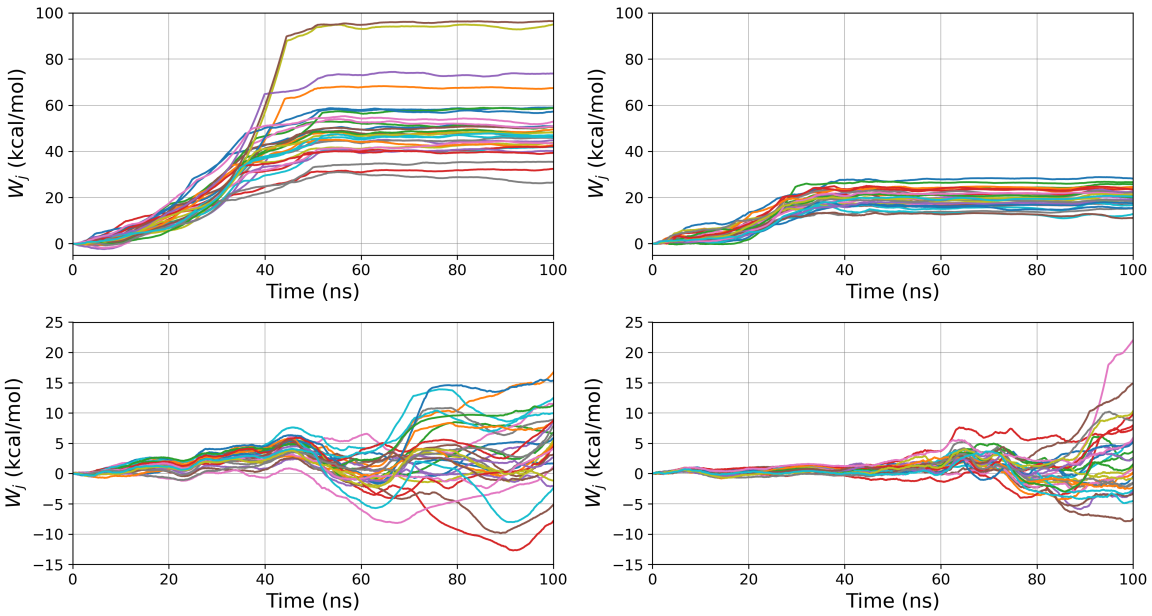

Figure 51: Jarzynski work profiles for the cognate ligand (A) and synthetic ligand (B) regarding unbinding (upper section) and binding (lower section) simulations measured over a simulation time of 100 ns in TIP3P.

kcal/mol). Differently, the value obtained for the cognate ligand only showed a slightly decreased value compared to TIP4P-D results, with a computed value of -18.5 ± 2.9 kcal/mol, still remarkably deviating from the reference experimental value of -10.9 kcal/mol.

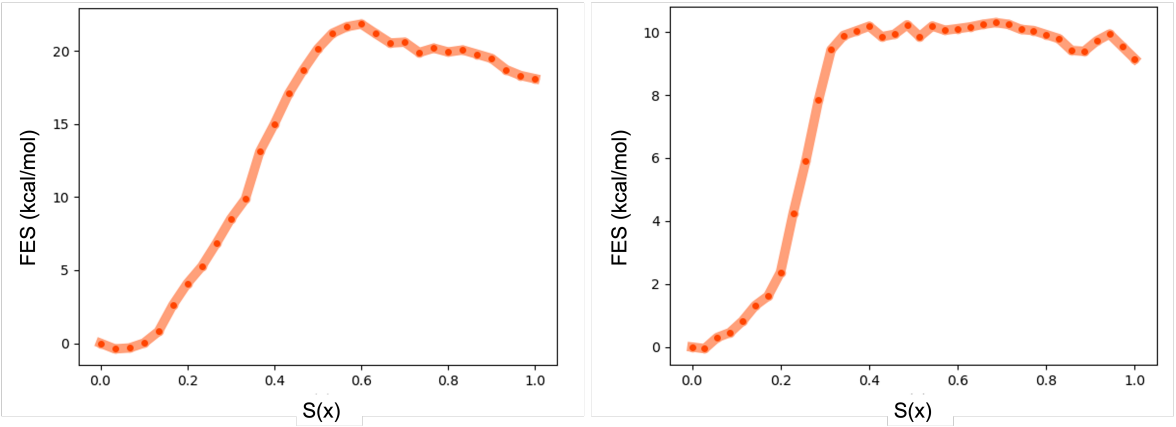

Figure 52: Free energy profiles along S(x) obtained by applying CFT to SMD (simulation time of 100 ns in TIP3P) for the cognate (A) and synthetic (B) ligands

Our simulations showed that the choice of water model significantly impacted the results, especially for the synthetic ligand. TIP4P-D substantially underestimated

the binding free energy (ΔF), while TIP3P values were within the experimental error range. In contrast, the water model had less effect on the cognate ligand, which is less hindered and has higher affinity. Notably, SMD results for the cognate ligand showed a marked discrepancy with experimental data, regardless of the water model. To further validate these results, we used well-tempered metadynamics with the path-CV (using the same reference pathway as the SMD runs and the TIP4P-D model) as an independent sampling method.

The metad simulations were able to adequately sample the path-CV, allowing to observe multiple (un)binding events in each simulation, as the ligands were able to transition multiple times between associated and dissociated states (Figure S19). Interestingly, the cognate ligand populated the bound state for a significant fraction of the simulation time. This is reasonable, given the higher amount of RNA-ligand interactions found for the cognate ligand in the bound state. After reconstructing the free-energy profile (Figure S20), we estimated the binding free energy, considering the same discriminating region, to distinguish the bound and unbound states, used to analyze the corresponding SMD analysis. The results obtained from the metadynamics simulations are reported in Table 4 and compared with the SMD results. Considering the totally independent

	SMD	**Metad**
Cognate	-17.2 ± 1.2	-19.4 ± 0.8
Synthetic	-5.6 ± 1.1	-7.4 ± 0.3

Table 4: Comparison of binding free energies (kcal/mol) for the cognate and synthetic ligands using TIP4P-D water model in Steered Molecular Dynamics and Metadynamics simulations.

nature of the estimation, relying on a remarkably different simulative approaches, it is worth noting how the binding free energy values from the metad were consistent with those obtained with the SMD simulations, within statistical errors. Additionally, the metad results, obtained from equilibrium simulations and the TIP4P-D water model, demonstrated higher agreement with the out-of-equilibrium SMD conducted with the TIP3P waters. For the synthetic ligand, this also corresponded with improved agreement with respect to the reference experimental data. This is particularly interesting, as it indicates that a more accurate, and computationally expensive, water model can provide results in agreement with experiments when used in equilibrium simulations. Indeed, the force field we used for the RNA in our simulations is typically employed in conjunction with the TIP4P-D model for waters. Conversely, a water model that is less accurate, though less expensive, such as the popular TIP3P, may be preferred for

conducting non-equilibrium simulations.

7.3.3 Effect of ligand parametrization on binding free energy estimates

The binding free energy values obtained for the cognate ligand deviated remarkably from the reference experimental value, consistently in both the SMD and metad simulations. To gain further insights into possible determinants underlying such disagreement, we investigated potential sources of discrepancy. Despite still being imperfect, force fields for macromolecules have significantly been improved over the years, becoming progressively more reliable and increasingly capable of quantitative prediction of experimental observables. While this is true for macromolecular species such as proteins and RNAs, small molecule parameterization is still lagging behind, mainly due to the heterogeneity and wide size of the chemical space associated with the small organic molecules that can potentially be designed/assembled. In particular, inaccuracies in charge and dihedral angle parameters can impact remarkably simulations results. In this respect, nowadays it is rather standard to use the AM1-BCC topological model to assign charges to ligands for MD simulations. Despite these may results in high-quality description of electrostatic properties of ligands in most cases, they may be insufficient in more complex scenarios, where substantial differences in charges for certain atom types, particularly the ammonium nitrogen (atom type nh) and certain aliphatic carbon atoms (atom type c3) may influence the results [264]. In these cases, resorting to a more accurate description, such as using the Restrained Electrostatic Potential (RESP) charges, may be more appropriate (charges differences are shown in Figure S24. Therefore, considering the higher amount of heteroatoms concurring to formation of hydrogen bond interactions with the RNA target in the bound state, we pursued full re-parametrization of the cognate ligand with RESP charges and optimized dihedrals. We then applied the already described protocol to generate a guess unbinding pathways for the new complex with RESP charges for the ligand via ABMD simulations, and performed the production SMD runs. Interestingly, the preferred unbinding pathway, passing through Loop 2 and Stem 1 (Figure 44), was consistent with the one observed using the previous ligand parametrization and was similar across all ABMD runs. To assess the effect of the different water models in combination with the newly parametrized ligand, both the TIP4P-D and TIP3P water models were considered (Table 5). This allowed achieving a more comprehensive picture for systematically comparing results

	TIP4P-D	TIP3P
RESP	-21.2 ± 1.0	-21.4 ± 1.2
AM1-BCC	-17.2 ± 1.2	-18.6 ± 2.9

Table 5: Binding free energies (kcal/mol) for the cognate ligand with RESP and AM1-BCC charges in different water models.

under different force-field variants. Consistently with what observed using the first set of parameters, the different water models returned highly similar results for this ligand, compatible within statistical error (Figure 5). Interestingly and unexpectedly, the values deviated further from the experimental binding free energy compared to the AM1-BCC parametrized ligand, suggesting that the newly devised set of parameters increased stabilization of the bound state through hydrogen bond interactions.

Inspecting the crystallographic pose of the two ligands reveals how the binding mode of the cognate ligand is driven by a higher number of interactions with the RNA target. In particular, while both ligands form a comparable amount of stacking interactions with the RNA, the cognate ligand is able to establish a remarkably higher number of hydrogen bonds (7 vs 2, see Figure 53B, Figure S23). On the one hand, this undoubtedly supports the experimentally-reported higher affinity of the cognate ligand for the RNA target. On the other hand, one would expect remarkable discrepancy in the experimental affinities, larger than the reported about 3 kcal/mol.

Since we obtained results in remarkable agreement with experiments for the synthetic ligand, we further investigated possible sources of inaccuracy for the cognate ligand. An extensive exploration of ligand conformational and tautomeric states in solution revealed the possibility of having an N1-N2 tautomer, i.e. where the proton is on the N1 instead of N2 (Figure 53A). In particular, such alternative state appeared to be promoted in solution by the intramolecular hydrogen bond between the carbonyl and the protonated amine groups.

Note that PreQ1 is structurally similar to the Guanine nucleobase, with relatively subtle modifications. Therefore, such tautomeric state was counterintuitive and unexpected, since the typical expected hydrogen bonding network for PreQ1 in the binding site involves an interaction with nucleobase C15 via their Watson-Crick edges, as expected for a Guanine in such three-dimensional arrangement (Figure 53).

Indeed, the cognate ligand looses two hydrogen bond interactions, with the C15 and A29, respectively (Figure 53B). Nevertheless, we decided to included it in our simulation panel by repeating the entire pipeline with the cognate ligand in the alternative

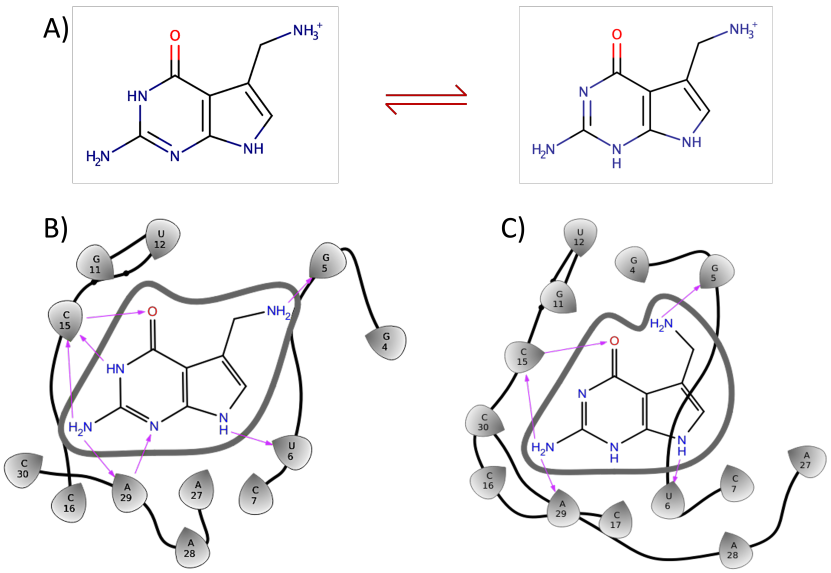

Figure 53: The two tautomeric forms considered for the cognate ligand (A) and the corresponding hydrogen bonding networks (B).

tauromeric form. To obtain a comprehensive picture and compare with the other results, we conducted SMD runs with the TIP4P-D and TIP3P water models and performed metad simulations.

The work profiles do not exhibit any unexpected behavior, showcasing a fluctuation region after approximately 60 ns, followed by good stabilization within the binding pocket, that is more pronounced in the TIP3P simulations compared to the TIP4P-D simulations (Figure S17-S18).

Figure 54 summarizes the results for this ligand. Surprisingly, we observed a remarkably improved agreement with experimental data. Specifically, in TIP4P-D simulations, the binding free energy amounted to -13.2 ± 1.4 kcal/mol, while in TIP3P, it was -11.9 ± 1.0 kcal/mol. Both estimates of the binding free energy align well with the experimental value of -10.9 kcal/mol (especially the TIP3P results), with again the validation of the metad simulation in TIP4P-D water with a free energy of -13.9 ± 0.4 kcal/mol.

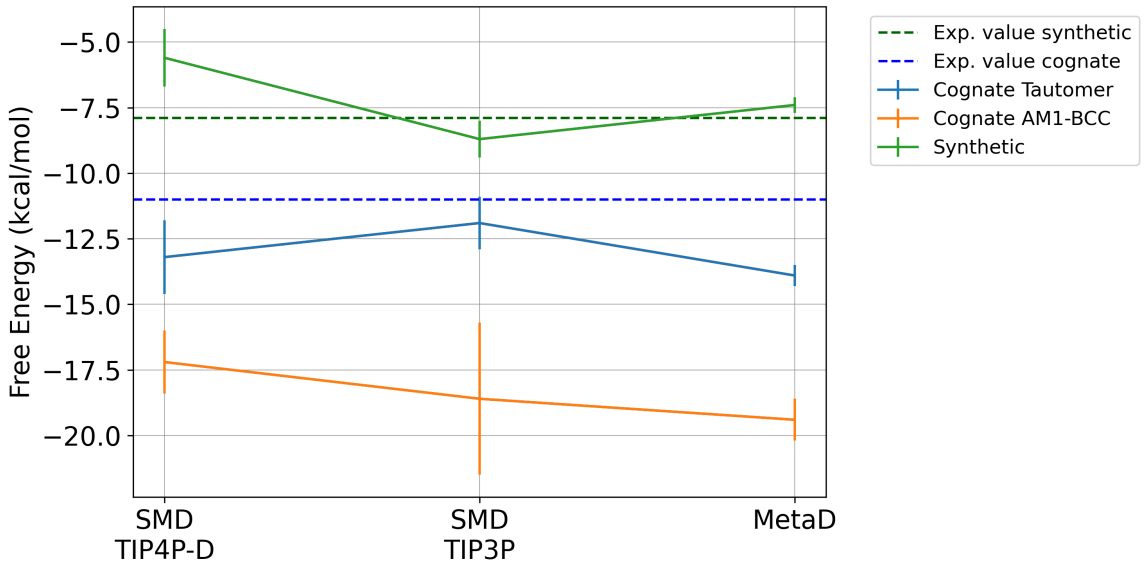

Figure 54: Binding free energies for all complexes tested. Dashed lines indicate experimental values (green: synthetic, blue: cognate). The x-axis represents the calculation method (SMD in TIP4P-D and TIP3P or Metadynamics), while the y-axis shows the corresponding binding free energy with associated errors. Ligands are color-coded: synthetic (green), cognate with AM1-BCC charges (orange) and cognate tautomer (light blue).

8 Conclusions

To sum up, this thesis has explored the intricate landscape of RNA-targeted drug discovery, focusing on the development and application of computational methods to address the unique challenges posed by RNA's dynamic nature and the limitations of traditional docking protocols. Through a multi-faceted approach encompassing druggability prediction, allosteric analysis, conformational ensemble generation, docking, and scoring, we have sought to advance our understanding of RNA-ligand interactions and pave the way for the rational design of novel RNA-targeting therapeutics.

Our investigation into druggability prediction has challenged the conventional binary classification of "druggable" vs. "non-druggable" targets, proposing instead a one-class learning approach that leverages the unambiguous information derived solely from known druggable pockets. This strategy, implemented through the Import Vector Domain Description (IVDD) algorithm with customized DrugPred descriptors, has demonstrated promising results in identifying potential druggable pockets with enhanced focus and efficiency.

Complementing this effort, our exploration of allostery has highlighted the complex interplay of factors that govern long-range communication within biomolecular systems. By comparing different computational methods for correlation estimation, we have gained valuable insight into the identification and characterization of allosteric sites, underscoring the importance of considering different methods of correlation in drug discovery endeavors. Notably, our analysis has revealed the great performance of Pocketron in consistently identifying known allosteric pockets with high correlation.

These methodologies for druggability prediction and allosteric analysis (NanoShaper and Pocketron in particular) have been implemented in the development of a refined computational protocol, which we have applied to investigate the long non-coding RNA MALAT1, a promising therapeutic target implicated in various cancers. Through the strategic application of NanoShaper and Pocketron, we have identified potential druggable pockets within MALAT1, focusing on those that could (potentially) disrupt the functionally critical triple helix structure of the RNA.

To account for RNA's inherent flexibility, we have employed unbiased and biased molecular dynamics (MD) simulations, specifically Hamiltonian replica exchange, to generate a comprehensive conformational ensemble of MALAT1. This ensemble has served as the foundation for our docking campaign, enabling us to evaluate the ability of different scoring functions to accurately predict experimental binding affinities for a

library of diminazene-based ligands with known activity against MALAT1.
Our findings have revealed limitations in the performance of most scoring functions, highlighting the need for further methodological refinements in RNA-ligand docking. Nevertheless, AutoDock has demonstrated a promising capacity to distinguish between high- and low-affinity ligands, suggesting its potential utility in RNA-targeted drug discovery.
Finally, we have extended our investigation beyond equilibrium methods by adapting a non-equilibrium binding free energy estimation method, originally designed for proteins, to RNA molecules. Our work on the Riboswitch-preQ1 system, employing steered molecular dynamics (SMD) simulations and the Crooks Fluctuation Theorem, has demonstrated the feasibility of extending this approach to RNA-ligand systems, while also underscoring the critical influence of protonation states and unbinding pathways on the accuracy of the calculations.
In conclusion, this thesis has contributed to the advancement of RNA-targeted drug discovery by introducing novel computational tools and insights into the complex dynamics of RNA-ligand interactions. Our work has addressed key challenges associated with targeting RNA, paving the way for the development of innovative therapeutic strategies against challenging diseases.

Bibliography

[1] Francis Crick. "Central dogma of molecular biology". In: *Nature* 227.5258 (1970), pp. 561–563.

[2] Stephen Jefferson Sharp et al. "Structure and transcription of eukaryotic tRNA gene". In: *Critical Reviews in Biochemistry* 19.2 (1985), pp. 107–144.

[3] James M Ogle and VJARB Ramakrishnan. "Structural insights into translational fidelity". In: *Annu. Rev. Biochem.* 74.1 (2005), pp. 129–177.

[4] James M Ogle, Andrew P Carter, and V Ramakrishnan. "Insights into the decoding mechanism from recent ribosome structures". In: *Trends in biochemical sciences* 28.5 (2003), pp. 259–266.

[5] Miguel JB Pereira et al. "Reaction pathway of the trans-acting hepatitis delta virus ribozyme: a conformational change accompanies catalysis". In: *Biochemistry* 41.3 (2002), pp. 730–740.

[6] Kelly Kruger et al. "Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of Tetrahymena". In: *cell* 31.1 (1982), pp. 147–157.

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

[8] Alexander Serganov and Evgeny Nudler. "A decade of riboswitches". In: *Cell* 152.1 (2013), pp. 17–24.

[9] Sethuramasundaram Pitchiaya et al. "Single molecule fluorescence approaches shed light on intracellular RNAs". In: *Chemical reviews* 114.6 (2014), pp. 3224–3265.

[10] Sven Diederichs et al. "The dark matter of the cancer genome: aberrations in regulatory elements, untranslated regions, splice sites, non-coding RNA and synonymous mutations". In: *EMBO molecular medicine* 8.5 (2016), pp. 442–457.

[11] SW Cheetham et al. "Long noncoding RNAs and the genetics of cancer". In: *British journal of cancer* 108.12 (2013), pp. 2419–2425.

[12] Ekta Khurana et al. "Role of non-coding sequence variants in cancer". In: *Nature Reviews Genetics* 17.2 (2016), pp. 93–108.

[13] Jiri Sponer et al. "RNA structural dynamics as captured by molecular simulations: a comprehensive overview". In: *Chemical reviews* 118.8 (2018), pp. 4177–4338.

[14] David H Mathews et al. "Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure". In: *Journal of molecular biology* 288.5 (1999), pp. 911–940.

[15] David H Mathews and Douglas H Turner. "Prediction of RNA secondary structure by free energy minimization". In: *Current opinion in structural biology* 16.3 (2006), pp. 270–278.

[16] Neocles B Leontis, Jesse Stombaugh, and Eric Westhof. "The non-Watson–Crick base pairs and their associated isostericity matrices". In: *Nucleic acids research* 30.16 (2002), pp. 3497–3531.

[17] Adane Letta Mamuye, Emanuela Merelli, and Luca Tesei. "A graph grammar for modelling RNA folding". In: *arXiv preprint arXiv:1612.01639* (2016).

[18] Daniel A Erlanson et al. "Twenty years on: the impact of fragments on drug discovery". In: *Nature reviews Drug discovery* 15.9 (2016), pp. 605–619.

[19] Gisbert Schneider and Uli Fechner. "Computer-based de novo design of drug-like molecules". In: *Nature Reviews Drug Discovery* 4.8 (2005), pp. 649–663.

[20] Eduardo Habib Bechelane Maia et al. "Structure-based virtual screening: from classical to artificial intelligence". In: *Frontiers in chemistry* 8 (2020), p. 343.

[21] Martina Palamini, Anselmo Canciani, and Federico Forneris. "Identifying and visualizing macromolecular flexibility in structural biology". In: *Frontiers in molecular biosciences* 3 (2016), p. 47.

[22] Tareq Hameduh et al. "Homology modeling in the time of collective and artificial intelligence". In: *Computational and Structural Biotechnology Journal* 18 (2020), pp. 3494–3506.

[23] Richard A Friesner et al. "Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy". In: *Journal of medicinal chemistry* 47.7 (2004), pp. 1739–1749.

[24] Gareth Jones et al. "Development and validation of a genetic algorithm for flexible docking". In: *Journal of molecular biology* 267.3 (1997), pp. 727–748.

[25] Jerome Eberhardt et al. "AutoDock Vina 1.2. 0: New docking methods, expanded force field, and python bindings". In: *Journal of chemical information and modeling* 61.8 (2021), pp. 3891–3898.

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

[27] Christophe Guilbert and Thomas L James. "Docking to RNA via root-mean-square-deviation-driven energy minimization with flexible ligands and flexible targets". In: *Journal of chemical information and modeling* 48.6 (2008), pp. 1257–1268.

[28] Sergio Ruiz-Carmona et al. "rDock: a fast, versatile and open source program for docking ligands to proteins and nucleic acids". In: *PLoS computational biology* 10.4 (2014), e1003571.

[29] Li-Zhen Sun et al. "RLDOCK: a new method for predicting RNA–ligand interactions". In: *Journal of chemical theory and computation* 16.11 (2020), pp. 7173–7183.

[30] Yuyu Feng et al. "NLDock: A fast nucleic acid–ligand docking algorithm for modeling RNA/DNA–ligand complexes". In: *Journal of Chemical Information and Modeling* 61.9 (2021), pp. 4771–4782.

[31] Yuyu Feng and Sheng-You Huang. "Itscore-nl: an iterative knowledge-based scoring function for nucleic acid–ligand interactions". In: *Journal of chemical information and modeling* 60.12 (2020), pp. 6698–6708.

[32] Sahil Chhabra, Jingru Xie, and Aaron T Frank. "RNAPosers: machine learning classifiers for ribonucleic acid–ligand poses". In: *The Journal of Physical Chemistry B* 124.22 (2020), pp. 4436–4445.

[33] Carlos Oliver et al. "Augmented base pairing networks encode RNA-small molecule binding preferences". In: *Nucleic acids research* 48.14 (2020), pp. 7690–7699.

[34] Filip Stefaniak and Janusz M Bujnicki. "AnnapuRNA: A scoring function for predicting RNA-small molecule binding poses". In: *PLoS computational biology* 17.2 (2021), e1008309.

[35] Christopher A Lipinski. "Lead-and drug-like compounds: the rule-of-five revolution". In: *Drug discovery today: Technologies* 1.4 (2004), pp. 337–341.

[36] Jason R Thomas and Paul J Hergenrother. "Targeting RNA with small molecules". In: *Chemical reviews* 108.4 (2008), pp. 1171–1224.

[37] Lirui Guan and Matthew D Disney. "Recent advances in developing small molecules targeting RNA". In: *ACS chemical biology* 7.1 (2012), pp. 73–86.

[38] Brittany S Morgan et al. "R-BIND: an interactive database for exploring and developing RNA-targeted chemical probes". In: *ACS chemical biology* 14.12 (2019), pp. 2691–2700.

[39] Hafeez S Haniff et al. "Design of a small molecule that stimulates vascular endothelial growth factor A enabled by screening RNA fold–small molecule interactions". In: *Nature chemistry* 12.10 (2020), pp. 952–961.

[40] Noreen F Rizvi et al. "Targeting RNA with small molecules: identification of selective, RNA-binding small molecules occupying drug-like chemical space". In: *SLAS DISCOVERY: Advancing the Science of Drug Discovery* 25.4 (2020), pp. 384–396.

[41] Matthias Wirth and Wolfgang HB Sauer. "Bioactive molecules: perfectly shaped for their target?" In: *Molecular Informatics* 30.8 (2011), pp. 677–688.

[42] Brittany S Morgan et al. "Discovery of key physicochemical, structural, and spatial properties of RNA-targeted bioactive ligands". In: *Angewandte Chemie International Edition* 56.43 (2017), pp. 13498–13502.

[43] Rosa Buonfiglio, Maurizio Recanatini, and Matteo Masetti. "Protein flexibility in drug discovery: from theory to computation". In: *ChemMedChem* 10.7 (2015), pp. 1141–1148.

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

[45] Woody Sherman et al. "Novel procedure for modeling ligand/receptor induced fit effects". In: *Journal of medicinal chemistry* 49.2 (2006), pp. 534–553.

[46] Sheng-You Huang and Xiaoqin Zou. "Ensemble docking of multiple protein structures: considering protein structural variations in molecular docking". In: *Proteins: Structure, Function, and Bioinformatics* 66.2 (2007), pp. 399–421.

[47] Rommie E Amaro et al. "Ensemble docking in drug discovery". In: *Biophysical journal* 114.10 (2018), pp. 2271–2278.

[48] Loic Salmon et al. “A general method for constructing atomic resolution RNA ensembles using NMR residual dipolar couplings: the basis for interhelical motions revealed”. In: *Journal of the American Chemical Society* 135.14 (2013), pp. 5457–5466.

[49] Laura R Ganser et al. “The roles of structural dynamics in the cellular functions of RNAs”. In: *Nature reviews Molecular cell biology* 20.8 (2019), pp. 474–489.

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

[51] Laura R Ganser et al. “High-performance virtual screening by targeting a high-resolution RNA dynamic ensemble”. In: *Nature structural & molecular biology* 25.5 (2018), pp. 425–434.

[52] Marco De Vivo et al. “Role of molecular dynamics and related methods in drug discovery”. In: *Journal of medicinal chemistry* 59.9 (2016), pp. 4035–4061.

[53] Sergio Decherchi and Andrea Cavalli. “Thermodynamics and kinetics of drug-target binding by molecular simulation”. In: *Chemical Reviews* 120.23 (2020), pp. 12788–12833.

[54] Junmei Wang, Piotr Cieplak, and Peter A Kollman. “How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules?” In: *Journal of computational chemistry* 21.12 (2000), pp. 1049–1074.

[55] Alberto Pérez et al. “Refinement of the AMBER force field for nucleic acids: improving the description of α/γ conformers”. In: *Biophysical journal* 92.11 (2007), pp. 3817–3829.

[56] Marie Zgarbová et al. “Refinement of the Cornell et al. nucleic acids force field based on reference quantum chemical calculations of glycosidic torsion profiles”. In: *Journal of chemical theory and computation* 7.9 (2011), pp. 2886–2902.

[57] Wendy D Cornell et al. “A second generation force field for the simulation of proteins, nucleic acids, and organic molecules”. In: *Journal of the American Chemical Society* 117.19 (1995), pp. 5179–5197.

[58] Vijay S Pande et al. “Atomistic protein folding simulations on the submillisecond time scale using worldwide distributed computing”. In: *Biopolymers: Original Research on Biomolecules* 68.1 (2003), pp. 91–109.

[59] David E Shaw et al. “Anton 2: raising the bar for performance and programmability in a special-purpose molecular dynamics supercomputer”. In: *SC’14: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. IEEE. 2014, pp. 41–53.

[60] Cameron Abrams and Giovanni Bussi. “Enhanced sampling in molecular dynamics using metadynamics, replica-exchange, and temperature-acceleration”. In: *Entropy* 16.1 (2013), pp. 163–199.

[61] Vojtěch Mlỳnskỳ and Giovanni Bussi. “Exploring RNA structure and dynamics through enhanced sampling simulations”. In: *Current opinion in structural biology* 49 (2018), pp. 63–71.

[62] Gareth A Tribello et al. “PLUMED 2: New feathers for an old bird”. In: *Computer physics communications* 185.2 (2014), pp. 604–613.

[63] Mattia Bernetti, Martina Bertazzo, and Matteo Masetti. "Data-driven molecular dynamics: a multifaceted challenge". In: *Pharmaceuticals* 13.9 (2020), p. 253.

[64] Johnny Habchi et al. "Introducing protein intrinsic disorder". In: *Chemical reviews* 114.13 (2014), pp. 6561–6588.

[65] Daniele Granata et al. "The inverted free energy landscape of an intrinsically disordered peptide by simulations and experiments". In: *Scientific reports* 5.1 (2015), p. 15449.

[66] Ferruccio Palazzesi et al. "Accuracy of current all-atom force-fields in modeling protein disordered states". In: *Journal of chemical theory and computation* 11.1 (2015), pp. 2–7.

[67] Mattia Bernetti et al. "Structural and kinetic characterization of the intrinsically disordered protein SeV NTAIL through enhanced sampling simulations". In: *The Journal of Physical Chemistry B* 121.41 (2017), pp. 9572–9582.

[68] Matteo Masetti, Mattia Bernetti, and Andrea Cavalli. "Enhanced molecular dynamics simulations of intrinsically disordered proteins". In: *Intrinsically Disordered Proteins: Methods and Protocols* (2020), pp. 391–411.

[69] Mattia Bernetti, Kathleen B Hall, and Giovanni Bussi. "Reweighting of molecular simulations with explicit-solvent SAXS restraints elucidates ion-dependent RNA ensembles". In: *Nucleic acids research* 49.14 (2021), e84–e84.

[70] Aneesur Rahman. "Correlations in the motion of atoms in liquid argon". In: *Physical review* 136.2A (1964), A405.

[71] Johannes Grotendorst. "Quantum simulations of complex many-body systems: from theory to algorithms: winter school, 25 February-1 March 2002, Rolduc Conference Centre, Kerkrade, the Netherlands; poster presentations/ed. by Johannes Grotendorst". In: (2002).

[72] Luca Monticelli and D Peter Tieleman. "Force fields for classical molecular dynamics". In: *Biomolecular simulations: Methods and protocols* (2013), pp. 197–213.

[73] Sergei Izrailev et al. "Steered molecular dynamics". In: *Computational Molecular Dynamics: Challenges, Methods, Ideas: Proceedings of the 2nd International Symposium on Algorithms for Macromolecular Modelling, Berlin, May 21–24, 1997*. Springer. 1999, pp. 39–65.

[74] Sanghyun Park and Klaus Schulten. "Calculating potentials of mean force from steered molecular dynamics simulations". In: *The Journal of chemical physics* 120.13 (2004), pp. 5946–5961.

[75] Christopher Jarzynski. "Nonequilibrium equality for free energy differences". In: *Physical Review Letters* 78.14 (1997), p. 2690.

[76] Gavin E Crooks. "Entropy production fluctuation theorem and the nonequilibrium work relation for free energy differences". In: *Physical Review E* 60.3 (1999), p. 2721.

[77] Katya Ahmad et al. "Enhanced-sampling simulations for the estimation of ligand binding kinetics: current status and perspective". In: *Frontiers in molecular biosciences* 9 (2022), p. 899805.

[78] Thomas E Cheatham III, Piotr Cieplak, and Peter A Kollman. "A modified version of the Cornell et al. force field with improved sugar pucker phases and helical repeat". In: *Journal of Biomolecular Structure and Dynamics* 16.4 (1999), pp. 845–862.

[79] Hugh Nymeyer. "How efficient is replica exchange molecular dynamics? An analytic approach". In: *Journal of chemical theory and computation* 4.4 (2008), pp. 626–636.

[80] Tom Darden, Darrin York, and Lee Pedersen. "Particle mesh Ewald: An N log (N) method for Ewald sums in large systems". In: *The Journal of chemical physics* 98.12 (1993), pp. 10089–10092.

[81] Giovanni Bussi. "Hamiltonian replica exchange in GROMACS: a flexible implementation". In: *Molecular Physics* 112.3-4 (2014), pp. 379–384.

[82] Xuan-Yu Meng et al. "Molecular docking: a powerful approach for structure-based drug discovery". In: *Current computer-aided drug design* 7.2 (2011), pp. 146–157.

[83] Pietro Cozzini et al. "How protein flexibility can influence docking/scoring simulations". In: *In silico drug discovery and design: theory, methods, challenges, and applications. CRC Press, Taylor & Francis Group, Boca Raton, FL* (2015), pp. 411–440.

[84] Alfonso T García-Sosa. "Hydration properties of ligands and drugs in protein binding sites: tightly-bound, bridging water molecules and their effects and consequences on molecular design strategies". In: *Journal of chemical information and modeling* 53.6 (2013), pp. 1388–1405.

[85] Inbal Halperin et al. "Principles of docking: An overview of search algorithms and a guide to scoring functions". In: *Proteins: Structure, Function, and Bioinformatics* 47.4 (2002), pp. 409–443.

[86] Jeffrey S Taylor and Roger M Burnett. "DARWIN: a program for docking flexible molecules". In: *Proteins: Structure, Function, and Bioinformatics* 41.2 (2000), pp. 173–191.

[87] Raquel Norel et al. "Shape complementarity at protein–protein interfaces". In: *Biopolymers: Original Research on Biomolecules* 34.7 (1994), pp. 933–940.

[88] Raquel Norel et al. "Examination of shape complementarity in docking of unbound proteins". In: *Proteins: Structure, Function, and Bioinformatics* 36.3 (1999), pp. 307–317.

[89] Michael L Connolly. "Analytical molecular surface calculation". In: *Journal of applied crystallography* 16.5 (1983), pp. 548–558.

[90] Michael L Connolly. "Solvent-accessible surfaces of proteins and nucleic acids". In: *Science* 221.4612 (1983), pp. 709–713.

[91] R Norel, HJ Wolfson, and R Nussinov. "Small molecule recognition: solid angles surface representation and molecular shape complementarity". In: *Combinatorial Chemistry and High Throughput Screening* 2 (1999), pp. 223–236.

[92] Douglas B Kitchen et al. "Docking and scoring in virtual screening for drug discovery: methods and applications". In: *Nature reviews Drug discovery* 3.11 (2004), pp. 935–949.

[93] Andrew R Leach. *Molecular modelling: principles and applications*. Pearson education, 2001.

[94] David R Westhead, David E Clark, and Christopher W Murray. "A comparison of heuristic search algorithms for molecular docking". In: *Journal of Computer-Aided Molecular Design* 11 (1997), pp. 209–228.

[95] Carol A Baxter et al. "Flexible docking using Tabu search and an empirical estimate of binding affinity". In: *Proteins: Structure, Function, and Bioinformatics* 33.3 (1998), pp. 367–382.

[96] Alfredo Di Nola, Danilo Roccatano, and Herman JC Berendsen. "Molecular dynamics simulation of the docking of substrates to proteins". In: *Proteins: Structure, Function, and Bioinformatics* 19.3 (1994), pp. 174–182.

[97] Jean-Yves Trosset and Harold A Scheraga. "Reaching the global minimum in docking simulations: a Monte Carlo energy minimization approach using Bezier splines". In: *Proceedings of the National Academy of Sciences* 95.14 (1998), pp. 8011–8015.

[98] Heather A Carlson and J Andrew McCammon. "Accommodating protein flexibility in computational drug design". In: *Molecular pharmacology* 57.2 (2000), pp. 213–218.

[99] Trevor N Hart and Randy J Read. "A multiple-start Monte Carlo docking method". In: *Proteins: Structure, Function, and Bioinformatics* 13.3 (1992), pp. 206–222.

[100] Connie M Oshiro, Irwin D Kuntz, and J Scott Dixon. "Flexible ligand docking using a genetic algorithm". In: *Journal of computer-aided molecular design* 9 (1995), pp. 113–130.

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

[102] Andrew R Leach. "Ligand docking to proteins with discrete side-chain flexibility". In: *Journal of molecular biology* 235.1 (1994), pp. 345–356.

[103] Johan Desmet et al. "The dead-end elimination theorem and its use in protein side-chain positioning". In: *Nature* 356.6369 (1992), pp. 539–542.

[104] Ronald MA Knegtel, Irwin D Kuntz, and CM Oshiro. "Molecular docking to ensembles of protein structures". In: *Journal of molecular biology* 266.2 (1997), pp. 424–440.

[105] Yuanzhe Zhou, Yangwei Jiang, and Shi-Jie Chen. "RNA–ligand molecular docking: Advances and challenges". In: *Wiley Interdisciplinary Reviews: Computational Molecular Science* 12.3 (2022), e1571.

[106] Peter Kollman. "Free energy calculations: applications to chemical and biochemical phenomena". In: *Chemical reviews* 93.7 (1993), pp. 2395–2417.

[107] Thomas Simonson, Georgios Archontis, and Martin Karplus. "Free energy simulations come of age: Protein- ligand recognition". In: *Accounts of chemical research* 35.6 (2002), pp. 430–437.

[108] Hans-Joachim Böhm. "LUDI: rule-based automatic design of new substituents for enzyme inhibitor leads". In: *Journal of computer-aided molecular design* 6 (1992), pp. 593–606.

[109] Paul S Charifson et al. "Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins". In: *Journal of medicinal chemistry* 42.25 (1999), pp. 5100–5109.

[110] Renxiao Wang, Yipin Lu, and Shaomeng Wang. "Comparative evaluation of 11 scoring functions for molecular docking". In: *Journal of medicinal chemistry* 46.12 (2003), pp. 2287–2303.

[111] Riccardo Aguti et al. "Probabilistic pocket druggability prediction via one-class learning". In: *Frontiers in Pharmacology* 13 (2022), p. 870479.

[112] Riccardo Aguti et al. "On the allosteric puzzle and pocket crosstalk through computational means". In: *The Journal of Chemical Physics* 158.16 (2023).

[113] KC Nicolaou. "Advancing the drug discovery and development process". In: *Angewandte Chemie* 126.35 (2014), pp. 9280–9292.

[114] Peter Csermely et al. "Structure and dynamics of molecular networks: a novel paradigm of drug discovery: a comprehensive review". In: *Pharmacology & therapeutics* 138.3 (2013), pp. 333–408.

[115] Ali Akbar Jamali et al. "DrugMiner: comparative analysis of machine learning algorithms for prediction of potential druggable proteins". In: *Drug discovery today* 21.5 (2016), pp. 718–724.

[116] Sergio Decherchi et al. *molecular dynamics and machine learning in drug discovery*. 2021.

[117] Mohini Gore and Umesh B Jagtap. *Computational drug discovery and design*. Springer, 2018.

[118] Roweida Mohammed, Jumanah Rawashdeh, and Malak Abdullah. "Machine learning with oversampling and undersampling techniques: overview study and experimental results". In: *2020 11th international conference on information and communication systems (ICICS)*. IEEE. 2020, pp. 243–248.

[119] Dong-Sheng Cao, Qing-Song Xu, and Yi-Zeng Liang. "propy: a tool to generate various modes of Chou's PseAAC". In: *Bioinformatics* 29.7 (2013), pp. 960–962.

[120] Nan Xiao et al. "protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences". In: *Bioinformatics* 31.11 (2015), pp. 1857–1859.

[121] Yasser B Ruiz-Blanco et al. "ProtDCal: A program to compute general-purpose-numerical descriptors for sequences and 3D-structures of proteins". In: *BMC bioinformatics* 16 (2015), pp. 1–15.

[122] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[123] Fredrik NB Edfeldt, Rutger HA Folmer, and Alexander L Breeze. "Fragment screening to predict druggability (ligandability) and lead discovery success". In: *Drug discovery today* 16.7-8 (2011), pp. 284–287.

[124] Si-Min Qi et al. "PROTAC: an effective targeted protein degradation strategy for cancer therapy". In: *Frontiers in Pharmacology* 12 (2021), p. 692574.

[125] Kenichiro Shimokawa et al. "Targeting the allosteric site of oncoprotein BCR-ABL as an alternative strategy for effective target protein degradation". In: *ACS medicinal chemistry letters* 8.10 (2017), pp. 1042–1047.

[126] Clement Agoni et al. "Druggability and drug-likeness concepts in drug design: are biomodelling and predictive tools having their say?" In: *Journal of molecular modeling* 26 (2020), pp. 1–11.

[127] Radoslav Krivák and David Hoksza. "P2Rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure". In: *Journal of cheminformatics* 10 (2018), pp. 1–12.

[128] Hiba Abi Hussein et al. "PockDrug-Server: a new web server for predicting pocket druggability on holo and apo proteins". In: *Nucleic acids research* 43.W1 (2015), W436–W442.

[129] Li Xie et al. "Drug discovery using chemical systems biology: identification of the protein-ligand binding network to explain the side effects of CETP inhibitors". In: *PLoS computational biology* 5.5 (2009), e1000387.

[130] Agata Krasowski et al. "DrugPred: a structure-based approach to predict protein druggability developed using an extensive nonredundant data set". In: *Journal of chemical information and modeling* 51.11 (2011), pp. 2829–2842.

[131] Andrea Volkamer et al. "DoGSiteScorer: a web server for automatic binding site prediction, analysis and druggability assessment". In: *Bioinformatics* 28.15 (2012), pp. 2074–2075.

[132] Radoslav Krivák and David Hoksza. "Improving protein-ligand binding site prediction accuracy by classification of inner pocket points using local features". In: *Journal of cheminformatics* 7 (2015), pp. 1–13.

[133] John A Capra et al. "Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure". In: *PLoS computational biology* 5.12 (2009), e1000585.

[134] Vincent Le Guilloux, Peter Schmidtke, and Pierre Tuffery. "Fpocket: an open source platform for ligand pocket detection". In: *BMC bioinformatics* 10 (2009), pp. 1–11.

[135] Jui-Hung Yuan et al. "Druggability assessment in TRAPP using machine learning approaches". In: *Journal of Chemical Information and Modeling* 60.3 (2020), pp. 1685–1699.

[136] Sergio Decherchi and Walter Rocchia. "Import vector domain description: a kernel logistic one-class learning algorithm". In: *IEEE transactions on neural networks and learning systems* 28.7 (2016), pp. 1722–1729.

[137] Sergio Decherchi et al. "NanoShaper–VMD interface: computing and visualizing surfaces, pockets and channels in molecular systems". In: *Bioinformatics* 35.7 (2019), pp. 1241–1243.

[138] Zhenting Gao et al. "PDTD: a web-accessible protein database for drug target identification". In: *BMC bioinformatics* 9 (2008), pp. 1–7.

[139] Pablo R Arantes, Amun C Patel, and Giulia Palermo. "Emerging methods and applications to decrypt allostery in proteins and nucleic acids". In: *Journal of molecular biology* 434.17 (2022), p. 167518.

[140] Lalima G Ahuja, Susan S Taylor, and Alexandr P Kornev. "Tuning the "violin" of protein kinases: The role of dynamics-based allostery". In: *IUBMB life* 71.6 (2019), pp. 685–696.

[141] Jingjing Guo and Huan-Xiang Zhou. "Protein allostery and conformational dynamics". In: *Chemical reviews* 116.11 (2016), pp. 6503–6515.

[142] Alexandr P Kornev and Susan S Taylor. “Dynamics-driven allostery in protein kinases”. In: *Trends in biochemical sciences* 40.11 (2015), pp. 628–647.

[143] Jin Liu and Ruth Nussinov. “Allostery: an overview of its history, concepts, methods, and applications”. In: *PLoS computational biology* 12.6 (2016), e1004966.

[144] Shoshana J Wodak et al. “Allostery in its many disguises: from theory to applications”. In: *Structure* 27.4 (2019), pp. 566–578.

[145] Kyle W East et al. “NMR and computational methods for molecular resolution of allosteric pathways in enzyme complexes”. In: *Biophysical reviews* 12.1 (2020), pp. 155–174.

[146] Rommie E Amaro. *Toward understanding “the ways” of allosteric drugs*. 2017.

[147] Ruth Nussinov and Chung-Jung Tsai. “Allostery in disease and in drug discovery”. In: *Cell* 153.2 (2013), pp. 293–305.

[148] Alice Triveri et al. “Protein allostery and ligand design: Computational design meets experiments to discover novel chemical probes”. In: *Journal of Molecular Biology* 434.17 (2022), p. 167468.

[149] Stefano Gianni and Per Jemth. “Allostery frustrates the experimentalist”. In: *Journal of Molecular Biology* 435.4 (2023), p. 167934.

[150] Giuseppina La Sala et al. “Allosteric communication networks in proteins revealed through pocket crosstalk analysis”. In: *ACS central science* 3.9 (2017), pp. 949–960.

[151] Ivan Rivalta et al. “Allosteric pathways in imidazole glycerol phosphate synthase”. In: *Proceedings of the National Academy of Sciences* 109.22 (2012), E1428–E1436.

[152] Giulia Morra, Gennady Verkhivker, and Giorgio Colombo. “Modeling signal propagation mechanisms and ligand-based conformational dynamics of the Hsp90 molecular chaperone full-length dimer”. In: *PLoS computational biology* 5.3 (2009), e1000323.

[153] Chakra Chennubhotla and Ivet Bahar. “Signal propagation in proteins and relation to equilibrium fluctuations”. In: *PLoS computational biology* 3.9 (2007), e172.

[154] Alexios Chatzigoulas and Zoe Cournia. “Rational design of allosteric modulators: Challenges and successes”. In: *Wiley Interdisciplinary Reviews: Computational Molecular Science* 11.6 (2021), e1529.

[155] K Gunasekaran, Buyong Ma, and Ruth Nussinov. “Is allostery an intrinsic property of all dynamic proteins?” In: *Proteins: Structure, Function, and Bioinformatics* 57.3 (2004), pp. 433–443.

[156] Yan Zhang et al. “Intrinsic dynamics is evolutionarily optimized to enable allosteric behavior”. In: *Current opinion in structural biology* 62 (2020), pp. 14–21.

[157] Akio Ohta et al. “A2A adenosine receptor protects tumors from antitumor T cells”. In: *Proceedings of the National Academy of Sciences* 103.35 (2006), pp. 13132–13137.

[158] KV Sivak et al. “Adenosine A 2A receptor as a drug target for treatment of sepsis”. In: *Molecular Biology* 50 (2016), pp. 200–212.

[159] Takahiro Matsumoto et al. “The androgen receptor in health and disease”. In: *Annual review of physiology* 75 (2013), pp. 201–224.

[160] Yuki Yuza et al. “Allele-dependent variation in the relative cellular potency of distinct EGFR inhibitors”. In: *Cancer biology & therapy* 6.5 (2007), pp. 661–667.

[161] Marcelo CR Melo et al. “Generalized correlation-based dynamical network analysis: a new high-performance approach for identifying allosteric communications in molecular dynamics trajectories”. In: *The Journal of Chemical Physics* 153.13 (2020).

[162] Sergio Decherchi and Walter Rocchia. “A general and robust ray-casting-based algorithm for triangulating surfaces at the nanoscale”. In: *PloS one* 8.4 (2013), e59744.

[163] Leighton Wilson and Robert Krasny. “Comparison of the MSMS and NanoShaper molecular surface triangulation codes in the TABI Poisson–Boltzmann solver”. In: *Journal of Computational Chemistry* 42.22 (2021), pp. 1552–1560.

[164] Sarmistha Majumdar et al. “Molecular Dynamics and Machine Learning Give Insights on the Flexibility–Activity Relationships in Tyrosine Kinome”. In: *Journal of Chemical Information and Modeling* 63.15 (2023), pp. 4814–4826.

[165] Sergio Decherchi et al. “BiKi life sciences: a new suite for molecular dynamics and related methods in drug discovery”. In: *Journal of chemical information and modeling* 58.2 (2018), pp. 219–224.

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

[167] William L Jorgensen et al. “Comparison of simple potential functions for simulating liquid water”. In: *The Journal of chemical physics* 79.2 (1983), pp. 926–935.

[168] Mark James Abraham et al. “GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers”. In: *SoftwareX* 1 (2015), pp. 19–25.

[169] Oliver F Lange and Helmut Grubmüller. “Generalized correlation for biomolecular dynamics”. In: *Proteins: Structure, Function, and Bioinformatics* 62.4 (2006), pp. 1053–1061.

[170] Carlos Sanchez-Martin et al. “Rational design of allosteric and selective inhibitors of the molecular chaperone TRAP1”. In: *Cell reports* 31.3 (2020).

[171] A Hagberg, P Swart, and DS Chult. *Exploring Network Structure, Dynamics, and Function Using NetworkX; Los Alamos National Lab.(LANL): Los Alamos, NM, USA, 2008.*

[172] Murad Nayal and Barry Honig. “On the nature of cavities on protein surfaces: application to the identification of drug-binding sites”. In: *Proteins: Structure, Function, and Bioinformatics* 63.4 (2006), pp. 892–906.

[173] Leo Breiman. “Random forests”. In: *Machine learning* 45 (2001), pp. 5–32.

[174] R Aggarwal et al. "Deeppocket: Ligand binding site detection and segmentation using 3d convolutional neural networks, 2021". In: *Ovchinnikov, S., Lee, GR, Wang, J., Cong, Q., Kinch, LN, Schaeffer, RD, et al. Accurate prediction of protein structures and interactions using a three-track neural network. Science* 373.6557 (2021), pp. 871–876.

[175] Charles F Van Loan and Gene H Golub. *Matrix computations*. Vol. 3. Johns Hopkins University Press Baltimore, 1983.

[176] Dima Kozakov et al. "The FTMap family of web servers for determining and characterizing ligand-binding hot spots of proteins". In: *Nature protocols* 10.5 (2015), pp. 733–755.

[177] Alisha D Caliman, Yinglong Miao, and James A McCammon. *Mapping the allosteric sites of the A2A adenosine receptor*. 2018.

[178] Wei Liu et al. "Structural basis for allosteric regulation of GPCRs by sodium ions". In: *Science* 337.6091 (2012), pp. 232–236.

[179] Hugo Gutiérrez-de-Terán et al. "The role of a sodium ion binding site in the allosteric modulation of the A2A adenosine G protein-coupled receptor". In: *Structure* 21.12 (2013), pp. 2175–2185.

[180] Peter Axerio-Cilies et al. "Inhibitors of androgen receptor activation function-2 (AF2) site identified through virtual screening". In: *Journal of medicinal chemistry* 54.18 (2011), pp. 6197–6205.

[181] Eva Estébanez-Perpiñá et al. "A surface on the androgen receptor that allosterically regulates coactivator binding". In: *Proceedings of the National Academy of Sciences* 104.41 (2007), pp. 16074–16079.

[182] Yuran Qiu et al. "Untangling dual-targeting therapeutic mechanism of epidermal growth factor receptor (EGFR) based on reversed allosteric communication". In: *Pharmaceutics* 13.5 (2021), p. 747.

[183] Jeremy E Wilusz, Hongjae Sunwoo, and David L Spector. "Long noncoding RNAs: functional surprises from the RNA world". In: *Genes & development* 23.13 (2009), pp. 1494–1504.

[184] Johnny TY Kung, David Colognori, and Jeannie T Lee. "Long noncoding RNAs: past, present, and future". In: *Genetics* 193.3 (2013), pp. 651–669.

[185] Susumu Ohno. "So much" junk" DNA in our genome. In" Evolution of Genetic Systems"". In: *Brookhaven symposium in biology*. Vol. 23. 1972, pp. 366–370.

[186] Gary J Olsen and Carl R Woese. "Archaeal genomics: an overview". In: *Cell* 89.7 (1997), pp. 991–994.

[187] Deng-Ke Niu and Li Jiang. "Can ENCODE tell us how much junk DNA we carry in our genome?" In: *Biochemical and biophysical research communications* 430.4 (2013), pp. 1340–1343.

[188] FZ Marques, SA Booth, and FJ Charchar. "The emerging role of non-coding RNA in essential hypertension and blood pressure regulation". In: *Journal of human hypertension* 29.8 (2015), pp. 459–467.

[189] Tamás Kiss. "Biogenesis of small nuclear RNPs". In: *Journal of cell science* 117.25 (2004), pp. 5949–5951.

[190] John R Prensner and Arul M Chinnaiyan. "The emergence of lncRNAs in cancer biology". In: *Cancer discovery* 1.5 (2011), pp. 391–407.

[191] Luka Bolha, Metka Ravnik-Glavač, and Damjan Glavač. "Long noncoding RNAs as biomarkers in cancer". In: *Disease markers* 2017.1 (2017), p. 7243968.

[192] L Li et al. "A long non-coding RNA interacts with Gfra1 and maintains survival of mouse spermatogonial stem cells". In: *Cell death & disease* 7.3 (2016), e2140–e2140.

[193] Rui Li, Hongliang Zhu, and Yunbo Luo. "Understanding the functions of long non-coding RNAs through their higher-order structures". In: *International journal of molecular sciences* 17.5 (2016), p. 702.

[194] Magdalena Losko, Jerzy Kotlinowski, and Jolanta Jura. "Long noncoding RNAs in metabolic syndrome related disorders". In: *Mediators of inflammation* 2016.1 (2016), p. 5365209.

[195] Natali Romero-Barrios et al. "Splicing regulation by long noncoding RNAs". In: *Nucleic acids research* 46.5 (2018), pp. 2169–2184.

[196] Igor Martianov et al. "Repression of the human dihydrofolate reductase gene by a non-coding interfering transcript". In: *Nature* 445.7128 (2007), pp. 666–670.

[197] Di Tian, Sha Sun, and Jeannie T Lee. "The long noncoding RNA, Jpx, is a molecular switch for X chromosome inactivation". In: *Cell* 143.3 (2010), pp. 390–403.

[198] Jeannie T Lee and Marisa S Bartolomei. "X-inactivation, imprinting, and long noncoding RNAs in health and disease". In: *Cell* 152.6 (2013), pp. 1308–1323.

[199] Marisa S Bartolomei, Sharon Zemel, and Shirley M Tilghman. "Parental imprinting of the mouse H19 gene". In: *Nature* 351.6322 (1991), pp. 153–155.

[200] Frank Sleutels, Ronald Zwart, and Denise P Barlow. "The non-coding Air RNA is required for silencing autosomal imprinted genes". In: *Nature* 415.6873 (2002), pp. 810–813.

[201] Alessandro Fatica and Irene Bozzoni. "Long non-coding RNAs: new players in cell differentiation and development". In: *Nature Reviews Genetics* 15.1 (2014), pp. 7–21.

[202] Pedro J Batista and Howard Y Chang. "Long noncoding RNAs: cellular address codes in development and disease". In: *Cell* 152.6 (2013), pp. 1298–1307.

[203] Nicola Amodio et al. "MALAT1: a druggable long non-coding RNA for targeted anti-cancer approaches". In: *Journal of hematology & oncology* 11 (2018), pp. 1–19.

[204] Tony Gutschner et al. "The noncoding RNA MALAT1 is a critical regulator of the metastasis phenotype of lung cancer cells". In: *Cancer research* 73.3 (2013), pp. 1180–1189.

[205] Gayatri Arun et al. "Differentiation of mammary tumors and reduction in metastasis upon Malat1 lncRNA loss". In: *Genes & development* 30.1 (2016), pp. 34–51.

[206] Gayatri Arun and David L Spector. "MALAT1 long non-coding RNA and breast cancer". In: *RNA biology* 16.6 (2019), pp. 860–863.

[207] Jongchan Kim et al. “Long noncoding RNA MALAT1 suppresses breast cancer metastasis”. In: *Nature genetics* 50.12 (2018), pp. 1705–1715.

[208] Zhi Hao Kwok et al. “A non-canonical tumor suppressive role for the long non-coding RNA MALAT1 in colon and breast cancers”. In: *International journal of cancer* 143.3 (2018), pp. 668–678.

[209] Zhi-Xing Li et al. “MALAT1: a potential biomarker in cancer”. In: *Cancer management and research* (2018), pp. 6757–6768.

[210] Jeremy E Wilusz et al. “A triple helix stabilizes the 3’ ends of long noncoding RNAs that lack poly (A) tails”. In: *Genes & development* 26.21 (2012), pp. 2392–2407.

[211] Jessica A Brown et al. “Structural insights into the stabilization of MALAT1 noncoding RNA by a bipartite triple helix”. In: *Nature structural & molecular biology* 21.7 (2014), pp. 633–640.

[212] Abeer A Ageeli et al. “Finely tuned conformational dynamics regulate the protective function of the lncRNA MALAT1 triple helix”. In: *Nucleic acids research* 47.3 (2019), pp. 1468–1481.

[213] Fardokht A Abulwerdi et al. “Selective small-molecule targeting of a triple helix encoded by the long noncoding RNA, MALAT1”. In: *ACS chemical biology* 14.2 (2019), pp. 223–235.

[214] Ben Davis et al. “Rational design of inhibitors of HIV-1 TAR RNA through the stabilisation of electrostatic “hot spots””. In: *Journal of molecular biology* 336.2 (2004), pp. 343–356.

[215] Saki Matsumoto et al. “Small synthetic molecule-stabilized RNA pseudoknot as an activator for–1 ribosomal frameshifting”. In: *Nucleic acids research* 46.16 (2018), pp. 8079–8089.

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

[217] Anita Donlic et al. “Discovery of small molecule ligands for MALAT1 by tuning an RNA-binding scaffold”. In: *Angewandte Chemie* 130.40 (2018), pp. 13426–13431.

[218] Martina Zafferani et al. “Multiassay profiling of a focused small molecule library reveals predictive bidirectional modulation of the lncRNA MALAT1 triplex stability in vitro”. In: *ACS chemical biology* 17.9 (2022), pp. 2437–2447.

[219] Indu G Rajapaksha et al. “The small molecule drug diminazene aceturate inhibits liver injury and biliary fibrosis in mice”. In: *Scientific reports* 8.1 (2018), p. 10175.

[220] Binh Nguyen et al. “Characterization of a novel DNA minor-groove complex”. In: *Biophysical journal* 86.2 (2004), pp. 1028–1041.

[221] Daniel S Pilch, Michael A Kirolos, and Kenneth J Breslauer. “Berenil binding to higher ordered nucleic acid structures: complexation with a DNA and RNA triple helix”. In: *Biochemistry* 34.49 (1995), pp. 16107–16124.

[222] Binh Nguyen, Stephen Neidle, and W David Wilson. “A role for water molecules in DNA- ligand minor groove recognition”. In: *Accounts of chemical research* 42.1 (2009), pp. 11–21.

[223] Zheng Li, Jon R Askim, and Kenneth S Suslick. “The optoelectronic nose: colorimetric and fluorometric sensor arrays”. In: *Chemical reviews* 119.1 (2018), pp. 231–292.

[224] Junjie Li et al. “Discrimination of Chinese teas according to major amino acid composition by a colorimetric IDA sensor”. In: *Sensors and Actuators B: Chemical* 240 (2017), pp. 770–778.

[225] Diogo Santos-Martins et al. “Accelerating AutoDock4 with GPUs and gradient-based local search”. In: *Journal of chemical theory and computation* 17.2 (2021), pp. 1060–1073.

[226] Garrett M Morris et al. “AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility”. In: *Journal of computational chemistry* 30.16 (2009), pp. 2785–2791.

[227] Yuanzhe Zhou, Yangwei Jiang, and Shi-Jie Chen. “SPRank- A Knowledge-Based Scoring Function for RNA-Ligand Pose Prediction and Virtual Screening”. In: *Journal of Chemical Theory and Computation* (2024).

[228] Sandro Bottaro, Francesco Di Palma, and Giovanni Bussi. “The role of nucleobase interactions in RNA structure and dynamics”. In: *Nucleic acids research* 42.21 (2014), pp. 13306–13314.

[229] Roy González-Alemán et al. “Quality threshold clustering of molecular dynamics: a word of caution”. In: *Journal of chemical information and modeling* 60.2 (2019), pp. 467–472.

[230] Ryne C Johnston et al. “Epik: p K a and Protonation State Prediction through Machine Learning”. In: *Journal of chemical theory and computation* 19.8 (2023), pp. 2380–2388.

[231] Kannan Sankar et al. “A descriptor set for quantitative structure-property relationship prediction in biologics”. In: *Molecular Informatics* 41.9 (2022), p. 2100240.

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

[233] Jonatan Taminau, Gert Thijs, and Hans De Winter. “Pharao: pharmacophore alignment and optimization”. In: *Journal of Molecular Graphics and Modelling* 27.2 (2008), pp. 161–169.

[234] Junmei Wang et al. “Development and testing of a general amber force field”. In: *Journal of computational chemistry* 25.9 (2004), pp. 1157–1174.

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

[236] Dirk Reith, Mathias Pütz, and Florian Müller-Plathe. “Deriving effective mesoscale potentials from atomistic simulations”. In: *Journal of computational chemistry* 24.13 (2003), pp. 1624–1636.

[237] Manfred J Sippl. “Boltzmann’s principle, knowledge-based mean fields and protein folding. An approach to the computational determination of protein structures”. In: *Journal of computer-aided molecular design* 7 (1993), pp. 473–501.

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

[239] Liberty François-Moutal et al. “In silico targeting of the long noncoding RNA MALAT1”. In: *ACS Medicinal Chemistry Letters* 12.6 (2021), pp. 915–921.

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

[241] Alessia Ghidini et al. “Bidirectional path-based non-equilibrium simulations for binding free energy”. In: *Molecular Physics* (2024), e2374465.

[242] Elaine R Lee, Kenneth F Blount, and Ronald R Breaker. “Roseoflavin is a natural antibacterial compound that binds to FMN riboswitches and regulates gene expression”. In: *RNA biology* 6.2 (2009), pp. 187–194.

[243] Meredith J Zeller et al. “Subsite ligand recognition and cooperativity in the TPP riboswitch: implications for fragment-linking in RNA ligand discovery”. In: *ACS chemical biology* 17.2 (2022), pp. 438–448.

[244] Yihang Wang et al. “Interrogating RNA–small molecule interactions with structure probing and artificial intelligence-augmented molecular simulations”. In: *ACS Central Science* 8.6 (2022), pp. 741–748.

[245] Colleen M Connelly et al. “Synthetic ligands for PreQ1 riboswitches provide structural and mechanistic insights into targeting RNA tertiary structure”. In: *Nature communications* 10.1 (2019), p. 1501.

[246] Mikhail S Gelfand et al. “A conserved RNA structure element involved in the regulation of bacterial riboflavin synthesis genes”. In: *Trends in Genetics* 15.11 (1999), pp. 439–442.

[247] Tina M Henkin. “Transcription termination control in bacteria”. In: *Current opinion in microbiology* 3.2 (2000), pp. 149–153.

[248] Maumita Mandal and Ronald R Breaker. “Gene regulation by riboswitches”. In: *Nature reviews Molecular cell biology* 5.6 (2004), pp. 451–463.

[249] Jeffrey E Barrick and Ronald R Breaker. “The power of riboswitches”. In: *Scientific American* 296.1 (2007), pp. 50–57.

[250] N Sudarsan et al. “Riboswitches in eubacteria sense the second messenger cyclic di-GMP”. In: *Science* 321.5887 (2008), pp. 411–413.

[251] Gavin E Crooks. “Nonequilibrium measurements of free energy differences for microscopically reversible Markovian systems”. In: *Journal of Statistical Physics* 90 (1998), pp. 1481–1487.

[252] Massimo Marchi and Pietro Ballone. “Adiabatic bias molecular dynamics: a method to navigate the conformational space of complex molecular systems”. In: *The Journal of chemical physics* 110.8 (1999), pp. 3697–3702.

[253] Marco Jacopo Ferrarotti, Walter Rocchia, and Sergio Decherchi. “Finding principal paths in data space”. In: *IEEE transactions on neural networks and learning systems* 30.8 (2018), pp. 2449–2462.

[254] Giovanni Bussi and Michele Parrinello. “Accurate sampling using Langevin dynamics”. In: *Physical Review E* 75.5 (2007), p. 056707.

[255] Herman JC Berendsen et al. “Molecular dynamics with coupling to an external bath”. In: *The Journal of chemical physics* 81.8 (1984), pp. 3684–3690.

[256] Gerard Martínez-Rosell, Toni Giorgino, and Gianni De Fabritiis. “PlayMolecule ProteinPrepare: a web application for protein preparation for molecular dynamics simulations”. In: *Journal of chemical information and modeling* 57.7 (2017), pp. 1511–1516.

[257] Martina Bertazzo et al. “Machine learning and enhanced sampling simulations for computing the potential of mean force and standard binding free energy”. In: *Journal of chemical theory and computation* 17.8 (2021), pp. 5287–5300.

[258] Giovanni Ciccotti and Lamberto Rondoni. “Jarzynski on work and free energy relations: The case of variable volume”. In: *AIChE Journal* 67.1 (2021), e17082.

[259] Christopher Jarzynski. “Equilibrium free-energy differences from nonequilibrium measurements: A master-equation approach”. In: *Physical Review E* 56.5 (1997), p. 5018.

[260] Michael R Shirts et al. “Equilibrium Free Energies from Nonequilibrium Measurements¡? format?¿ Using Maximum-Likelihood Methods”. In: *Physical review letters* 91.14 (2003), p. 140601.

[261] Slimane Doudou, Neil A Burton, and Richard H Henchman. “Standard free energy of binding from a one-dimensional potential of mean force”. In: *Journal of chemical theory and computation* 5.4 (2009), pp. 909–918.

[262] Bradley Efron. “Bootstrap methods: another look at the jackknife”. In: *Breakthroughs in statistics: Methodology and distribution*. Springer, 1992, pp. 569–593.

[263] William L Jorgensen. “Quantum and statistical mechanical studies of liquids. 10. Transferable intermolecular potential functions for water, alcohols, and ethers. Application to liquid water”. In: *Journal of the American Chemical Society* 103.2 (1981), pp. 335–340.

[264] Niel M Henriksen and Michael K Gilson. “Evaluating force field performance in thermodynamic calculations of cyclodextrin host–guest binding: Water models, partial charges, and host force field parameters”. In: *Journal of chemical theory and computation* 13.9 (2017), pp. 4253–4269.

9 Supplementary Materials

Application Case 1: Pocket druggability

Table 6 describes all the proteins included in the druggable (training) and the less druggable datasets. Druggable proteins are marked with d (training set), less druggable proteins are marked with n.

PDB code	**Name**	**Category**
1pwm	Aldose reductase	d
1lox	15-lipoxygenase	d
3etr	Xanthine oxidase	d
3f1q	Dihydroorotate dehydrogenase	d
3ia4	Dihydrofolate reductase	d
2cl5	Catechol-O-methyltransferase	d
1uou	Human thymidine phosphorylase	d
1t46	c-Kit kinase	d
1unl	cyclin-dependent kinase5	d
1q41	Glycogen synthase kinase 3	d
2i1m	FMS kinase	d
1pmn	c-Jun kinases	d
1fk9	HIV reverse transcriptase (nonnucleoside reverse transcriptase inhibitor binding site)	d
1e66	Acetylcholinesterase	d
1xoz	Phosphodiesterase 5A	d

1owe	Urokinase plasminogen activator	d
1r55	A disintegrin and metalloprotease	d
3f0r	Histone Deacetylase 8	d
1oq5	Carbonic anhydrase II	d
1kzn	DNA gyrase	d
2aa2	Mineralocorticoid receptor	d
3b68	Androgen receptor	d
1sqn	Progesterone receptor	d
1v16	Branched-chain alpha-keto acid dehydrogenase	n
3jdw	Arginine:glycine amidinotransferase	n
1ajs	Aspartate aminotransferase	n
1wvc	CDP-D-glucose synthase	n
1kc7	Pyruvate phosphate dikinase	n
1mai	Phospholipase C	n
1px4	Beta-galactosidase	n
1od8	Xylanase	n
1bmq	Interleukin-1 beta-converting enzyme 1	n
1bls	Beta-lactamase	n
1m0n	Dialkylglycine Decarboxylase	n
1ec9	D-glucarate dehydratase	n
1b74	Glutamate racemase	n

1g98	Phosphoglucose isomerase	n
1e9x	Cytochrome P450 14alpha -sterol demethylase	d
1hw8	3-hydroxy-3-methylglutaryl-CoA	d
1sqi	4-hydroxyphenylpyruvate dioxygenase	d
1r9o	Cytochrome P450 2C9	d
4cox	Cyclooxygenase 2	d
1c14	Enoyl reductase	d
2bxr	Monoamine oxidase A	d
2gh5	Glutathione reductase	d
1hvy	Thymidylate synthase	d
1rsz	Purine nucleoside phosphorylase	d
1n2v	tRNA-guanine transglycosylase	d
1v4s	Hexokinase	d
1u4d	ACK1 kinase	d
1m17	Epidermal growth factor receptor kinase	d
2dq7	Fyn kinase	d
1qpe	Lck kinase	d
1qhi	Thymidine kinase	d
2fb8	B-Raf kinase	d
1ke6	cyclin-dependent kinase2	d
2br1	Chk1 kinase	d

1ywr	p38 Mitogen-activated protein kinases	d
2ivu	RET kinase	d
2hiw	Abl tyrosin kinase	d
2i0e	Protein kinase C	d
1ywn	Vascular endothelial growth factor receptor-2	d
1ig3	Thiamin pyrophosphokinase	d
1yvf	Hepatitis C virus polymerase NS5B	d
1k8q	Gastric lipase	d
1kvo	Phospholipase A 2	d
1xm6	Phosphodiesterase 4B	d
1udt	Phosphodiesterase 5	d
1u30	Amylase	d
1r58	Methionine aminopeptidase-2	d
1rwq	Dipeptidyl peptidase-IV	d
1lpz	Factor Xa	d
2g24	Renin	d
1hvr	HIV protease	d
1gkc	Matrix metalloproteinase-9	d
1yqy	Lethal factor	d
1o5r	Adenosine deaminase	d
1js3	DOPA decarboxylase	d
1k7f	Tryptophan synthase	d

1j4i	FKBP13	d
1vbm	Tyrosyl-tRNA synthetase	d
1rv1	Ubiquitin-protein ligase E3 Mdm2	d
1gwr	Estrogen receptor	d
1m2z	Glucocorticoid receptor	d
3d4s	Beta-2-adrenergic receptor	d
1ai2	Isocitrate dehydrogenase	n
3pcm	3,4-dioxygenas	n
1d09	Aspartate transcarbamoylase	n
1c9y	Ornithine carbamoyltransferase	n
1gpu	Transketolase	n
1qmf	Penicillin binding protein-2X	n
1moq	Glucosamine 6-phosphate synthas	n
1ucn	Nucleoside diphosphate kinase	n
1t03	HIV reverse transcriptase (nucleoside binding site)	n
1qs4	HIV integrase	n
1fth	Acyl carrier protein synthase	n
1rnt	Ribonuclease T2	n
1onz	Protein-tyrosine phosphatase 1B	n
1x9d	Mannosidase	n
1nnc	Neuraminidase	n
1olq	Endo-beta-1,4-glucanase	n

1jak	Beta-N-Acetylhexosaminidases	n
1kts	Thrombin	n
1nlj	Cathepsin K	n
1icj	Peptide deformylase	n
1hqg	Arginase	n
2gsu	Phosphodiesterase-nucleotide Pyrophosphatase	n
1g7v	3-deoxy-D-manno-2-octulosonate-8-phosphate synthase	n
1f9g	Hyaluronate lyase	n
1qxo	Chorismate synthase	n
2gyi	D-xylose isomerase	n
1o8b	Ribose-5-phosphate isomerase	n
1cg0	Adenylosuccinate synthetase	n

Table 6: Proteins description of the NRDLD dataset.

Table 7 describes all the proteins included in the PDTD (100-proteins) dataset.

PDB code	**Name**	**Category**
1a28	Progesterone receptor	d
1acj	Acetylcholine esterase	d
1aco	Aconite with transaconitate bound	d
1adc	NAD analogues bound to alcohol dehydrogenase	d

1coy	Cholesterol oxidases	d
1cqe	Prostaglandin H2 synthase-1	d
1d3g	Dihydroorotate dehydrogenase	d
1d6u	E. Coli amine oxidase	d
1db1	Nuclear receptor for vitamin D	d
1dht	Estrogenic 17-beta hydroxysteroid dehydrogenase	d
1diy	Cyclooxygenase active site of PGHS-1	d
1dkf	Heterodimeric complex of RAR and RXR	d
1e1f	Beta-glucosidase	d
1e3g	Androgen receptor	d
1e3k	Progesteron receptor	d
1e55	Mutant Monocut beta-glucosidase	d
1eet	HIV-1 reverse transcriptase	d
1efh	Hydroxysteroid sulfotransferase	d
1f2a	Cruzain hydrolase	d
1fm6	Heterodimer of the RXR-α and PPAR-γ	d
1gii	Cyclin dependent kinase	d
1gos	Monoamine oxidase B	d
1gp6	Anthocyanidin synthase	d
1gpk	Acetylcholinesterase	d

1gqs	Acetylcholinesterase complexed with NAP	d
1gs4	Androgen receptor ARccr	d
1h5u	Glycogen phosphorylase B	d
1h9u	Retinoid X receptor beta	d
1hb2	Isopenicillin N synthase	d
1hdy	Alcohol dehydrogenase variant	d
1hfc	Fibroblast collagenase	d
1hj1	Estrogen receptor beta	d
1hld	Liver alcohol dehydrogenase	d
1ho4	Pyridoxine 5-phosphate	d
1ht8	Oxidoreductase COX-1	d
1hy3	Estrogen sulfotransferase V269E	d
1hzx	Bovine Rhodopsin	d
1i7g	Human PPAR-α	d
1ie9	Nuclear receptor for vitamin D	d
1iiu	Plasma retinol-binding protein	d
1j90	Deoxyribonuclease kinase	d
1jbp	Catalytic subunit of c-AMP dependent protein kinase	d
1jkh	HIV-1 reverse transcriptase	d
1js3	Dopa decarboxylase	d
1k3u	Tryptophan synthase	d
1k4w	Nuclear receptor ROR-β	d

1k74	Heterodimer of PPAR-γ and RXR-α	d
1k7l	Human PPAR-α	d
1lde	Liver alcohol dehydrogenase	d
1ldy	Liver alcohol dehydrogenase complexed to NADH and cyclohexyl formamide	d
1mup	Pheromone binding to two urinary proteins	d
1n7i	Phenylethanolamine N-metyltransferase	d
1nwk	Monomeric actin in the ATP state	d
1og5	Human cytochrome P450 CYP2C9	d
1oi9	Human thr160-phospho CDK2/cyclin A	d
1p1n	GluR2 ligand binding core (S1S2J) mutant	d
1p2d	Glycogen phosphorylase B	d
1p4g	Glycogen phosphorylase B in complex with C-(1-azido-alpha-D-glucopyranosyl) formamide	d
1p93	Glucocorticoid receptor	d
1pcg	Helix-stabilized cyclic peptides	d
1pha	Cytochrome P450-CAM	d
1pig	Pancreatic alpha-amylase	d
1ppl	Aspartyl proteinases	d

1qab	Retinol binding protein	d
1kvo	Phospholipase A 2	d
1qkm	Estrogen receptor β	d
1qkt	Mutant estrogen nuclear receptor	d
1qpb	Pyruvate decarboxylase	d
1r18	Isoaspartyl methyltransferase	d
1r1k	Heterodimer EcR/USP bound to ponasterone A	d
1rbp	Serum retinol binding protein	d
1rlb	Retinol binding protein complexed with transthyretin	d
1rt6	HIV-1 reverse transcriptase	d
1tvr	HIV-1 RT/9-CL TIBO	d
1uhl	LXRα-RXRβ LBD heterodimer	d
1ulb	Purine nucleoside phosphorylase	d
1uom	Estrogen receptor complexed with Tetrahydroisochiolin	d
1upv	Liver X receptor β	d
1v8b	Hydrolase	d
1vkg	HDAC8	d
1vlb	Aldehyde oxidoreductase	d
1vot	Acetylcholine esterase	d
1w6k	Human OSC	d

1x07	Undecaprenyl pyrophosphate synthase	d
1xnx	Androstane receptor	d
1y0s	PPAR-γ	d
1zhy	Oxysterol binding protein Osh4	d
2a3i	Mineralocorticoid receptor	d
2a3l	Adenosine 5'-Monophosphate deaminase	d
2ack	Acetylcholinesterase	d
2ae2	Tropinone reductase-II	d
2bx8	Human serum albumin	d
2dln	D-alanine ligase	d
2mas	Purine nucleoside hydrolase	d
3bto	Liver alcohol dehydrogenase	d
3ert	Estrogen receptor-α	d
3hvt	Human immunodeficiency virus type 1 reverse transcriptase heterodimer	d
4thi	Thiaminase I	d
6cox	Cyclooxygenase-2	d
8cat	Liver catalase	d

Table 7: Proteins description of the PDTD dataset.

Application Case 1: Pocket communication

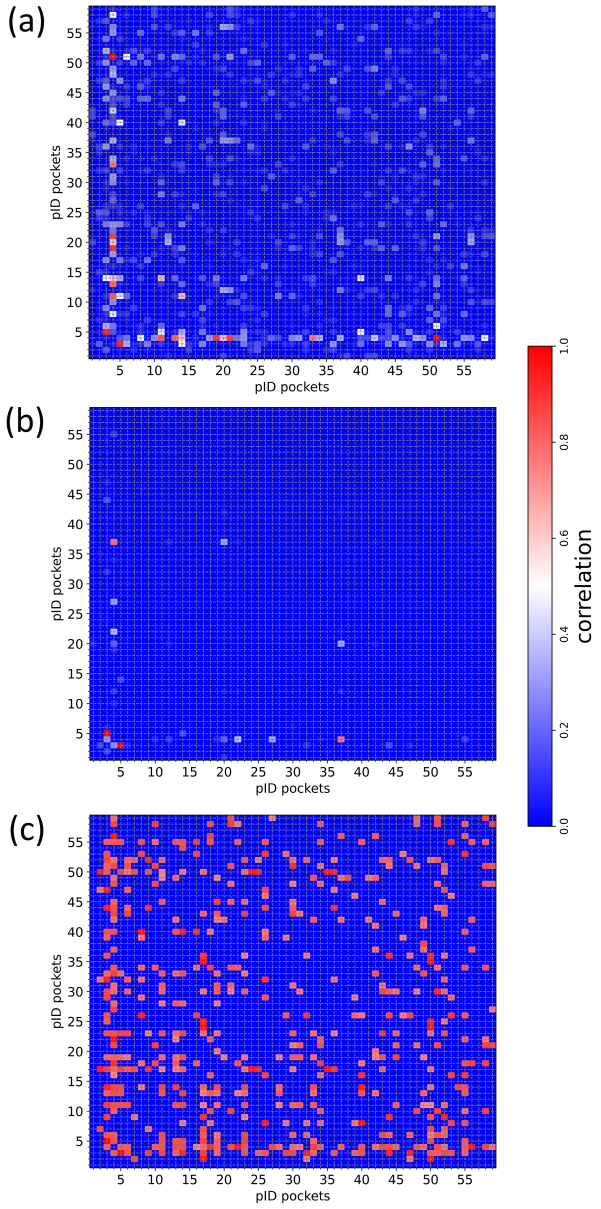

Figure S1: Correlation matrices for the A2A systems. Results are from the DyNet (a), Pocketron (b), and DF (c) approaches are displayed. The colour bar for the matrix's elements is shown next to the middle panel.

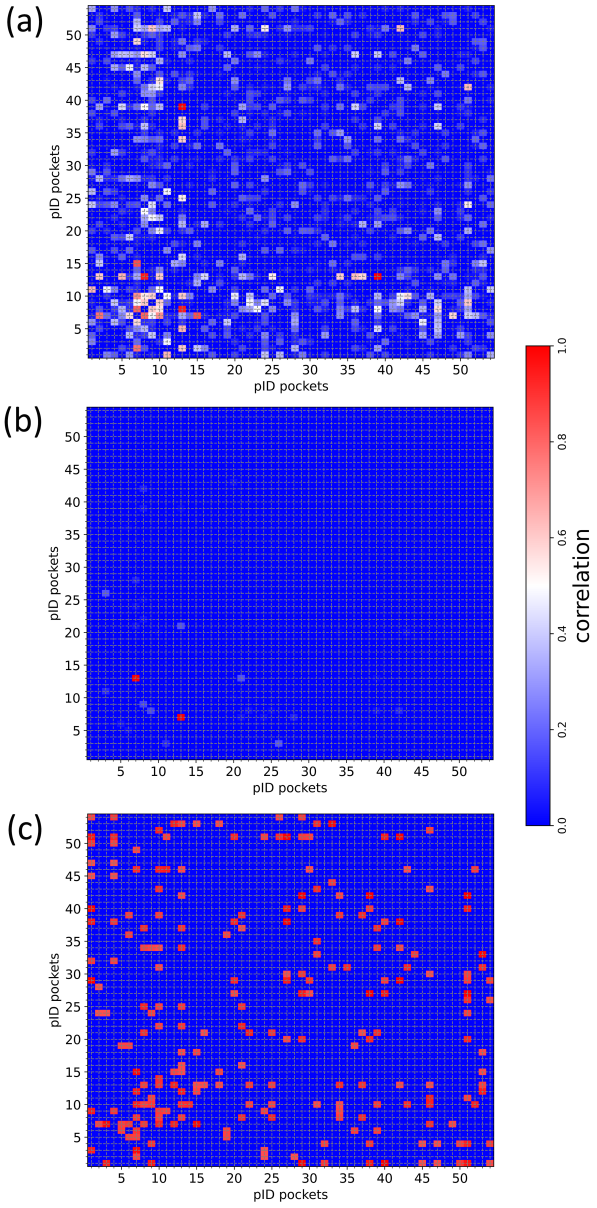

Figure S2: Correlation matrices for the AR systems. Results are from the DyNet (a), Pocketron (b), and DF (c) approaches are displayed. The colour bar for the matrix's elements is shown next to the middle panel.

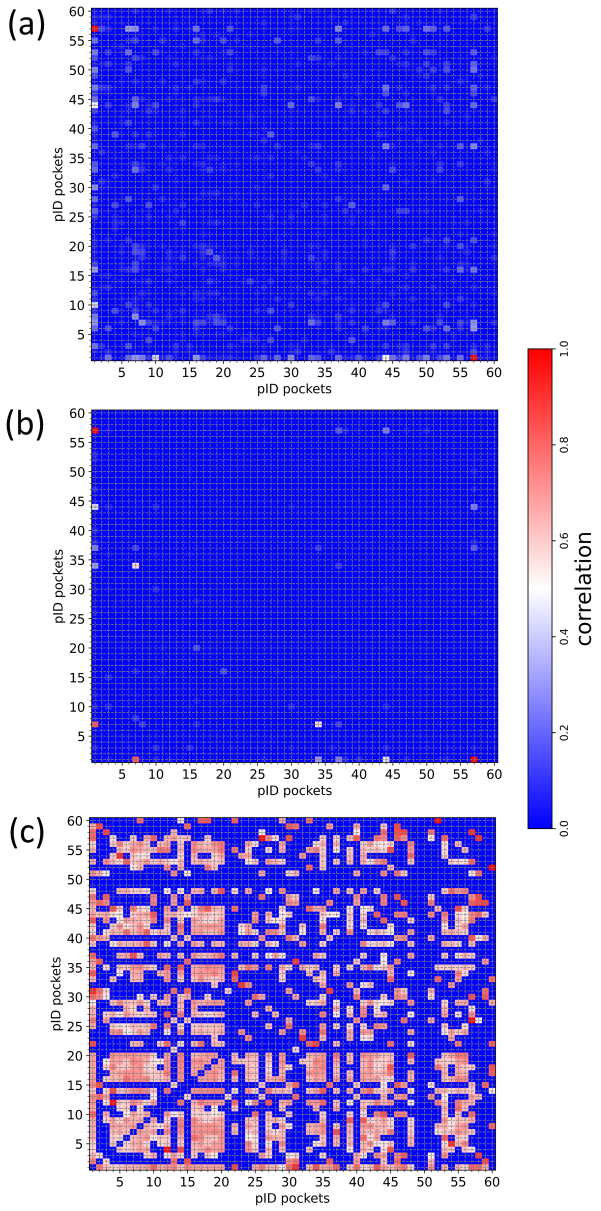

Figure S3: Correlation matrices for the EGFR systems. Results are from the DyNet (a), Pocketron (b), and DF (c) approaches are displayed. The colour bar for the matrix's elements is shown next to the middle panel.

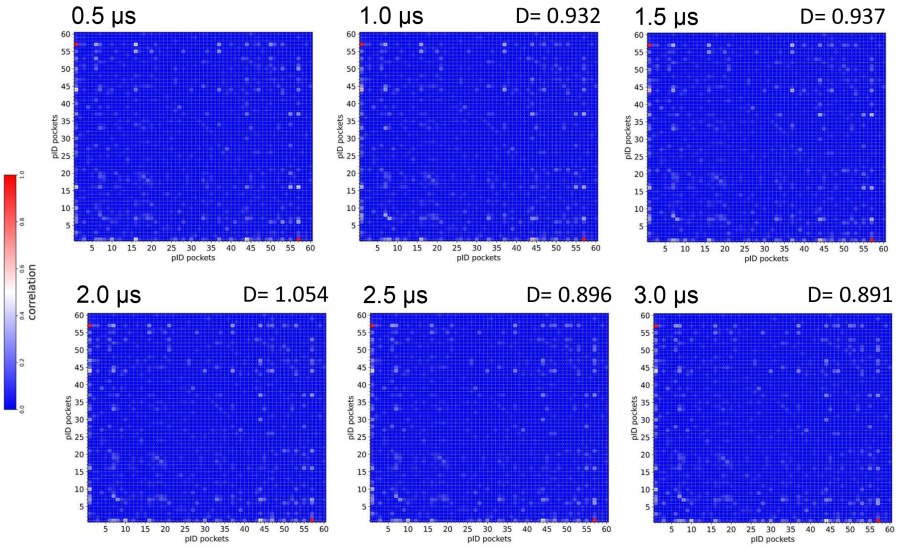

Figure S4: Correlation matrices obtained with the DyNet analysis for the EGFR system, at increasing simulation lengths (0.5 ms, 1.0 ms, 1.5 ms, 2.0 ms, 2.5 ms, 3.0 ms). For each correlation matrix, the Frobenius distance with respect to the matrix obtained at simulation time 0.5 ms is also reported in the upper right corner. The analysed pockets are those obtained with Pocketron from the total simulation time.

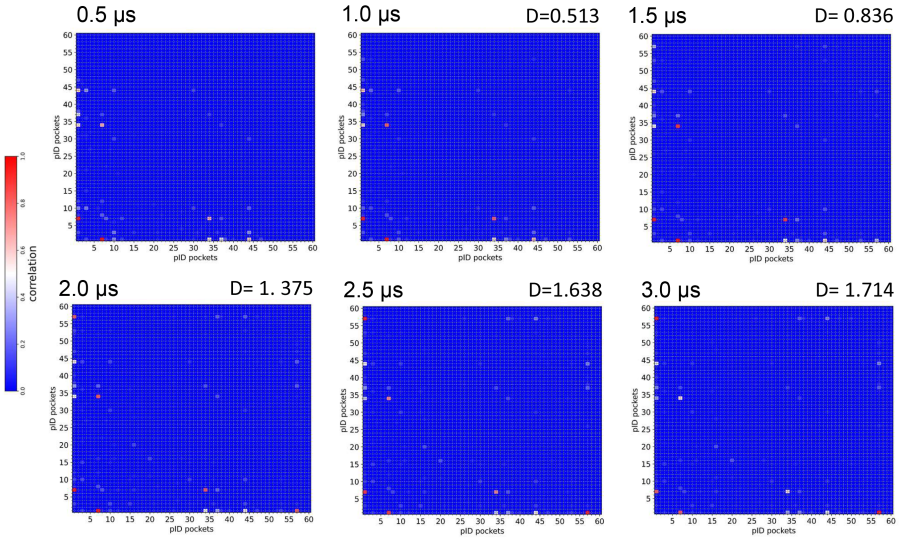

Figure S5: Correlation matrices obtained with Pocketron for the EGFR system, at increasing simulation lengths (0.5 ms, 1.0 ms, 1.5 ms, 2.0 ms, 2.5 ms, 3.0 ms). For each correlation matrix, the Frobenius distance with respect to the matrix obtained at simulation time 0.5 ms is also reported in the upper right corner. Values of 0 have been assigned to correlations with pockets that were not discovered yet in the initial fractions of simulations.

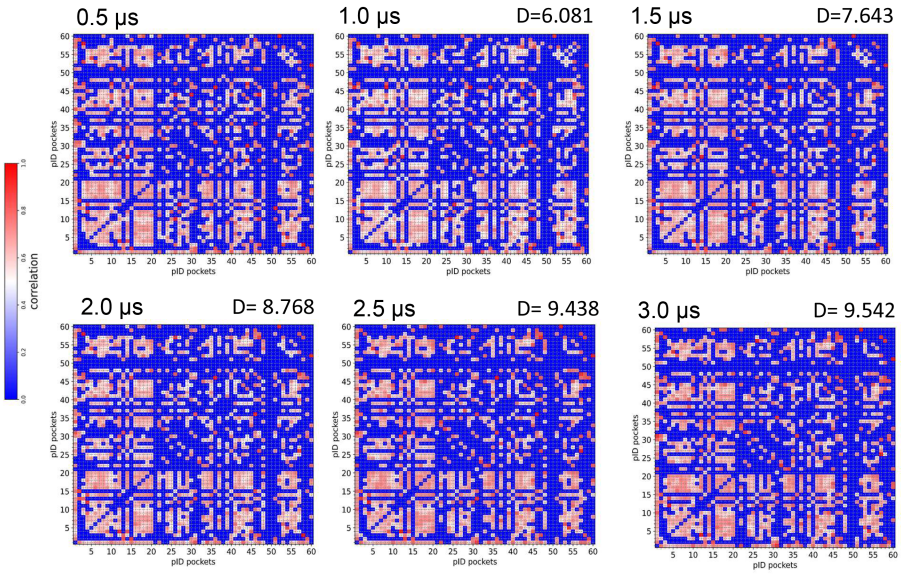

Figure S6: Correlation matrices obtained with the distance fluctuation analysis for the EGFR system, at increasing simulation lengths (0.5 ms, 1.0 ms, 1.5 ms, 2.0 ms, 2.5 ms, 3.0 ms). For each correlation matrix, the Frobenius distance with respect to the matrix obtained at simulation time 0.5 ms is also reported in the upper right corner. The analysed pockets are those obtained with Pocketron from the total simulation time.

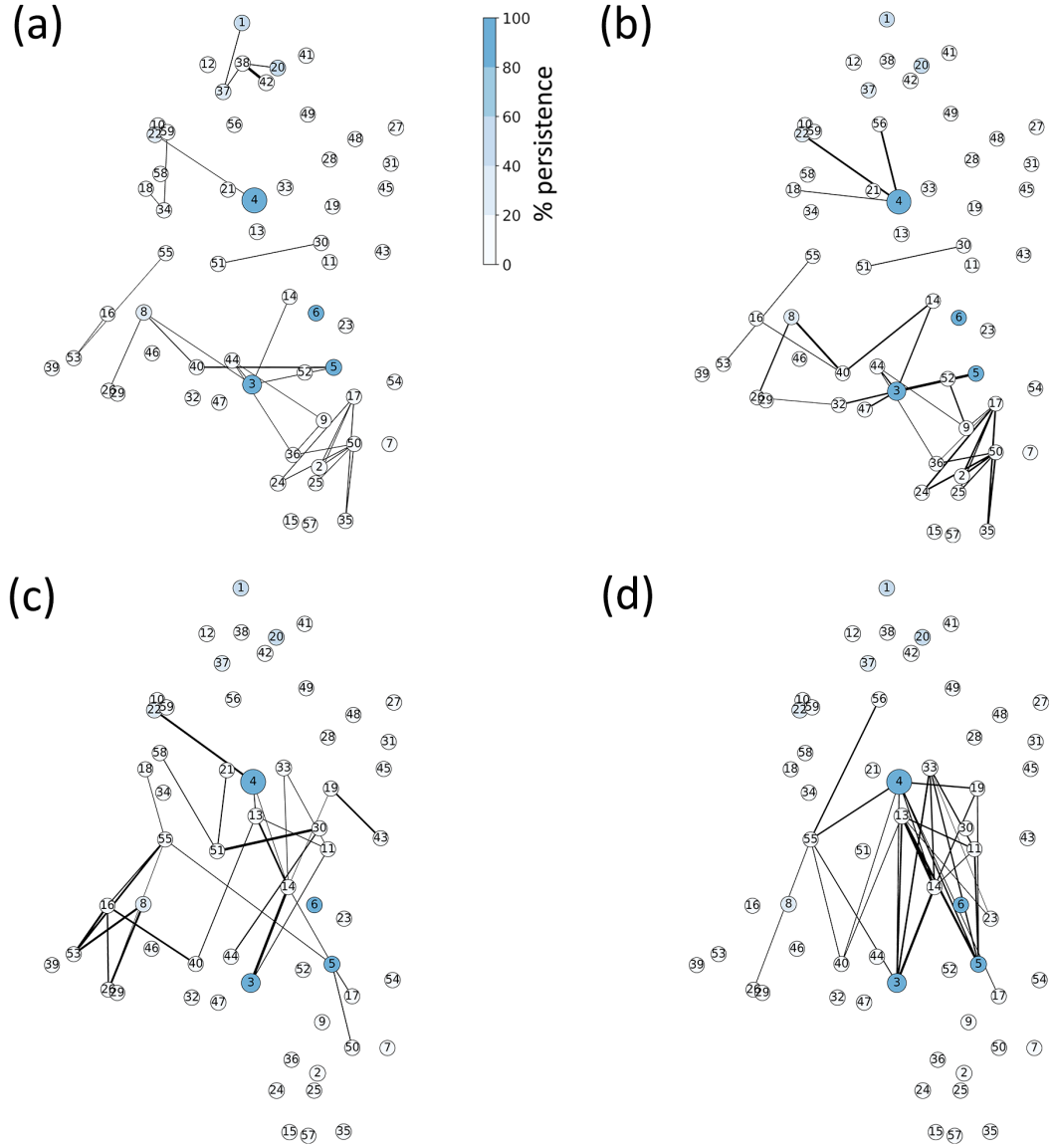

Figure S7: Correlation matrices obtained with the distance fluctuation analysis for the EGFR system, at increasing simulation lengths (0.5 ms, 1.0 ms, 1.5 ms, 2.0 ms, 2.5 ms, 3.0 ms). For each correlation matrix, the Frobenius distance with respect to the matrix obtained at simulation time 0.5 ms is also reported in the upper right corner. The analysed pockets are those obtained with Pocketron from the total simulation time.

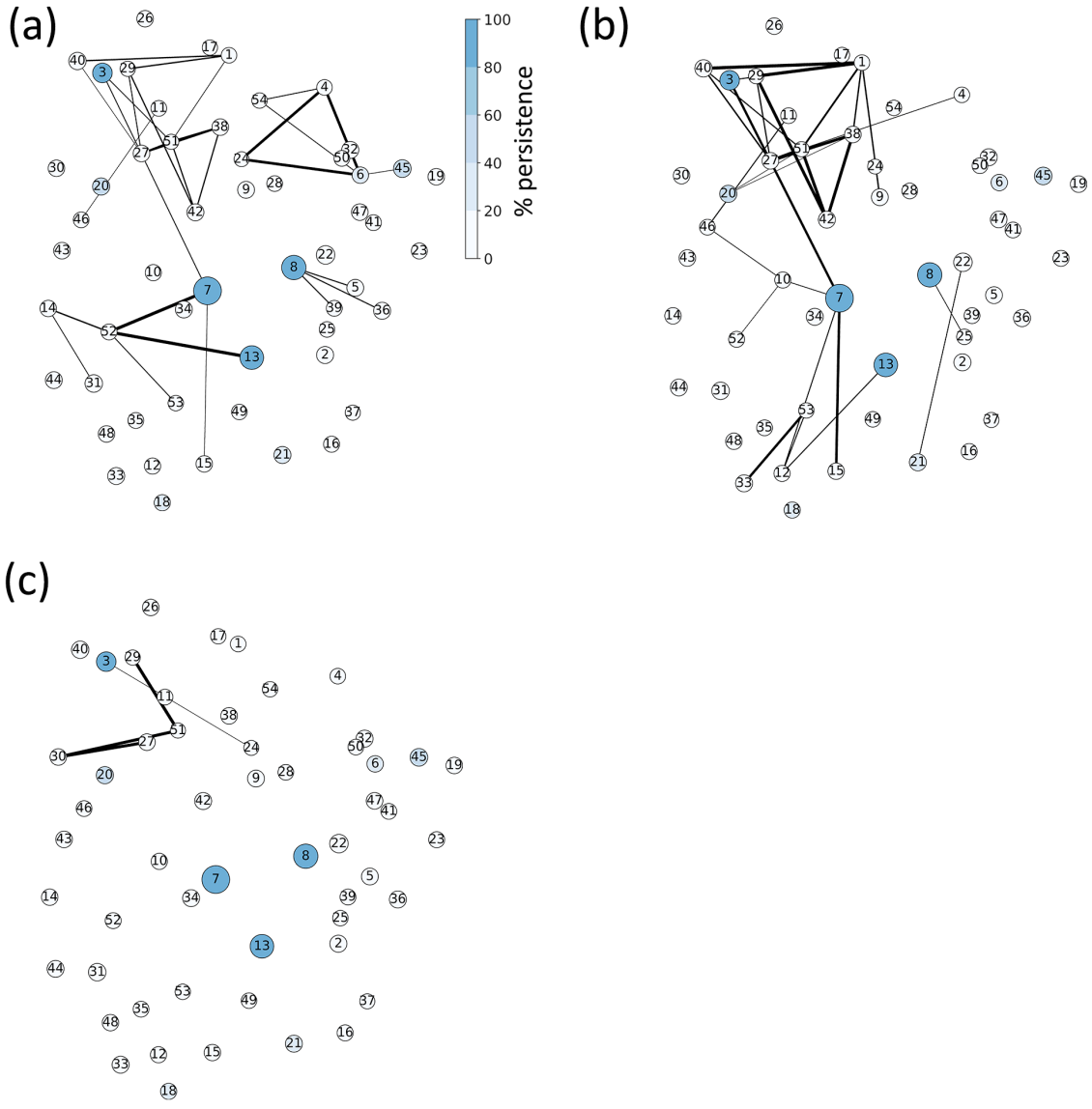

Figure S8: Correlation matrices obtained with the distance fluctuation analysis for the EGFR system, at increasing simulation lengths (0.5 ms, 1.0 ms, 1.5 ms, 2.0 ms, 2.5 ms, 3.0 ms). For each correlation matrix, the Frobenius distance with respect to the matrix obtained at simulation time 0.5 ms is also reported in the upper right corner. The analysed pockets are those obtained with Pocketron from the total simulation time.

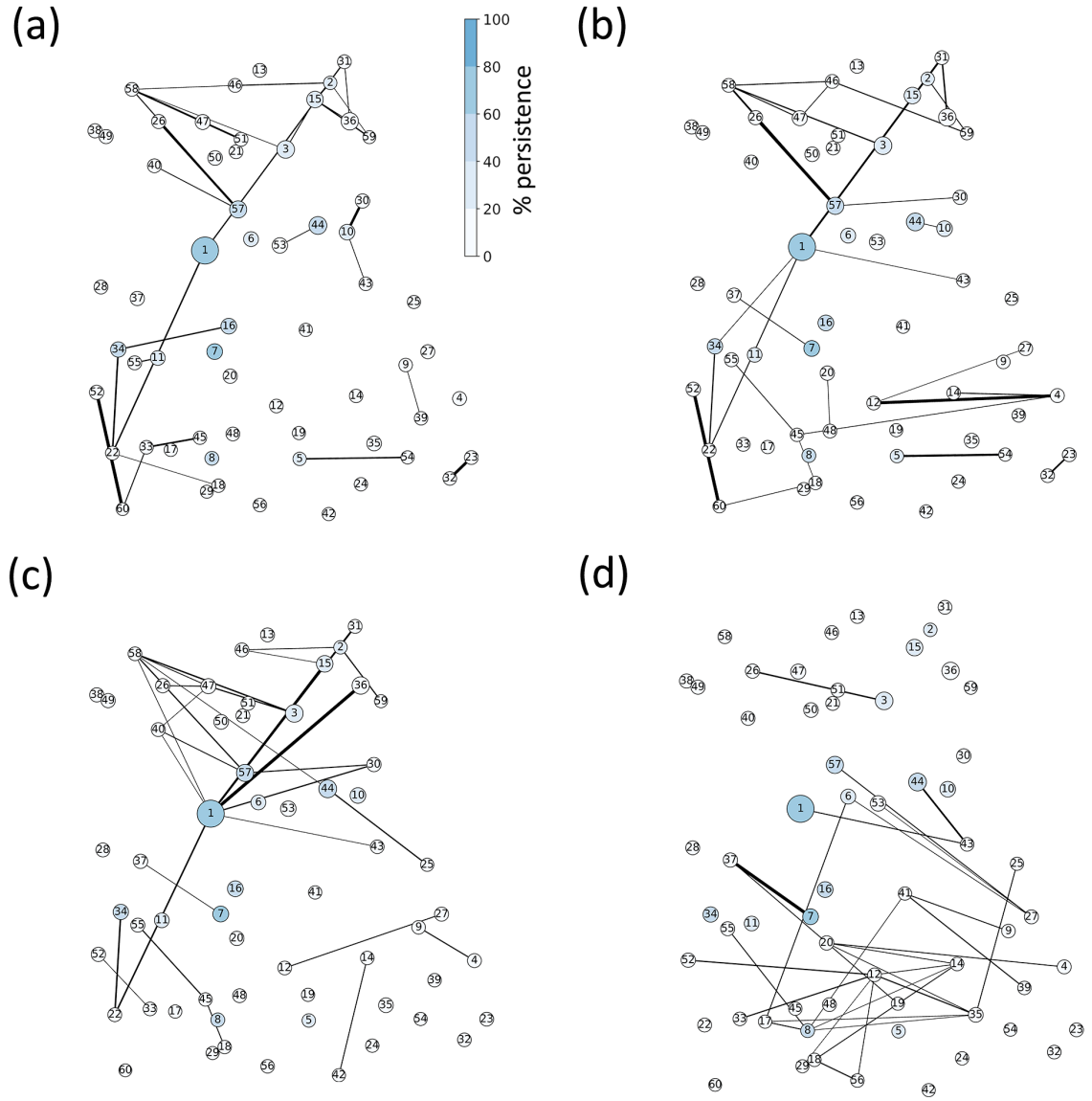

Figure S9: Correlation matrices obtained with the distance fluctuation analysis for the EGFR system, at increasing simulation lengths (0.5 ms, 1.0 ms, 1.5 ms, 2.0 ms, 2.5 ms, 3.0 ms). For each correlation matrix, the Frobenius distance with respect to the matrix obtained at simulation time 0.5 ms is also reported in the upper right corner. The analysed pockets are those obtained with Pocketron from the total simulation time.

Application Case 2: Application of established computational methods in drug discovery to target RNAs

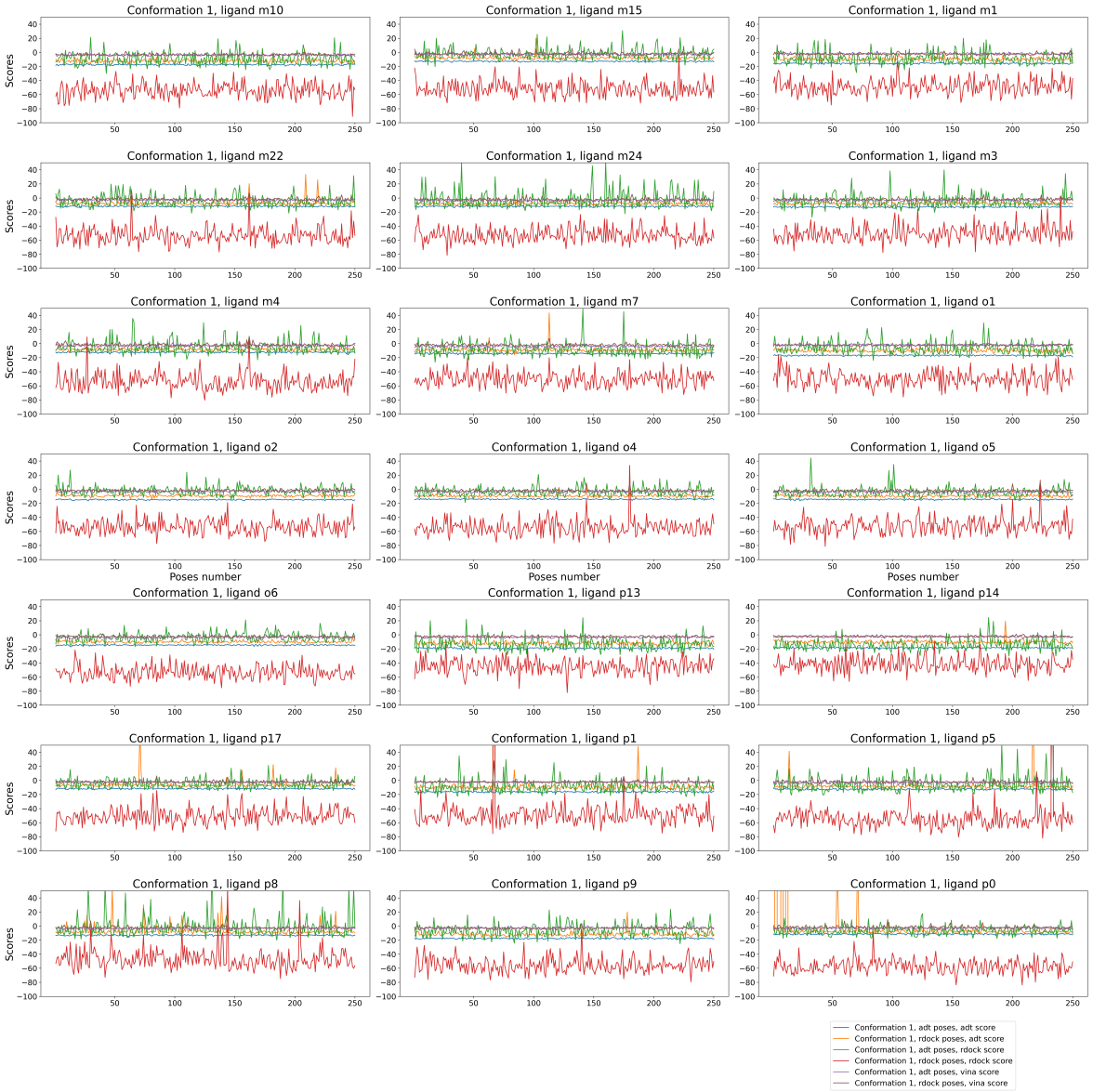

Figure S10: Docking scores for all the generated poses on Site 1, color-coded by scoring function and docking software: AutoDock (blue/orange), rDock (green/red), Vina (purple/gray). Blue/green/purple lines represent poses generated with AutoDock GPU, while orange/red/gray lines represent those from rDock.

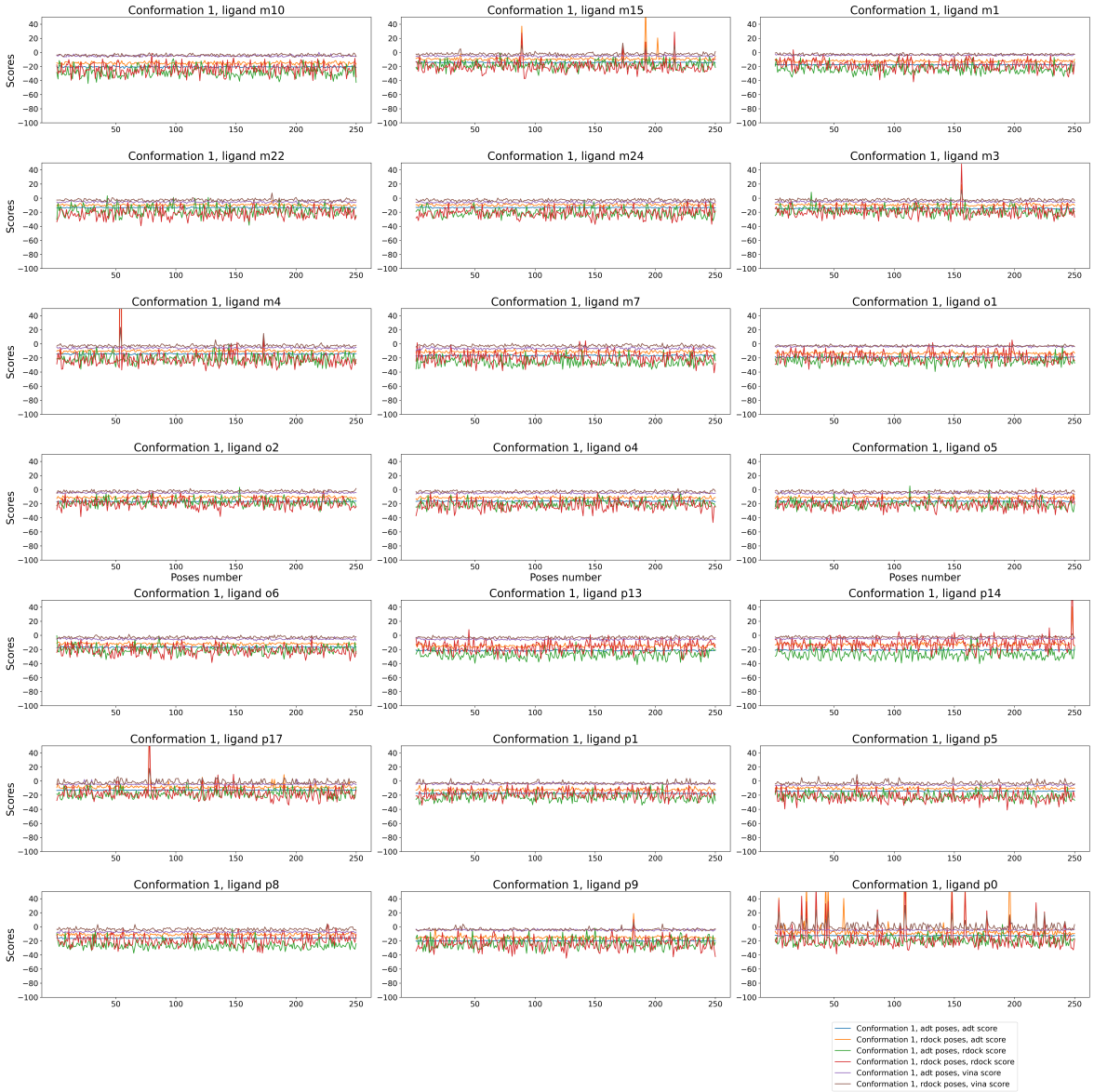

Figure S11: Docking scores for all the generated poses on Site 2, color-coded by scoring function and docking software: AutoDock (blue/orange), rDock (green/red), Vina (purple/gray). Blue/green/purple lines represent poses generated with AutoDock GPU, while orange/red/gray lines represent those from rDock.

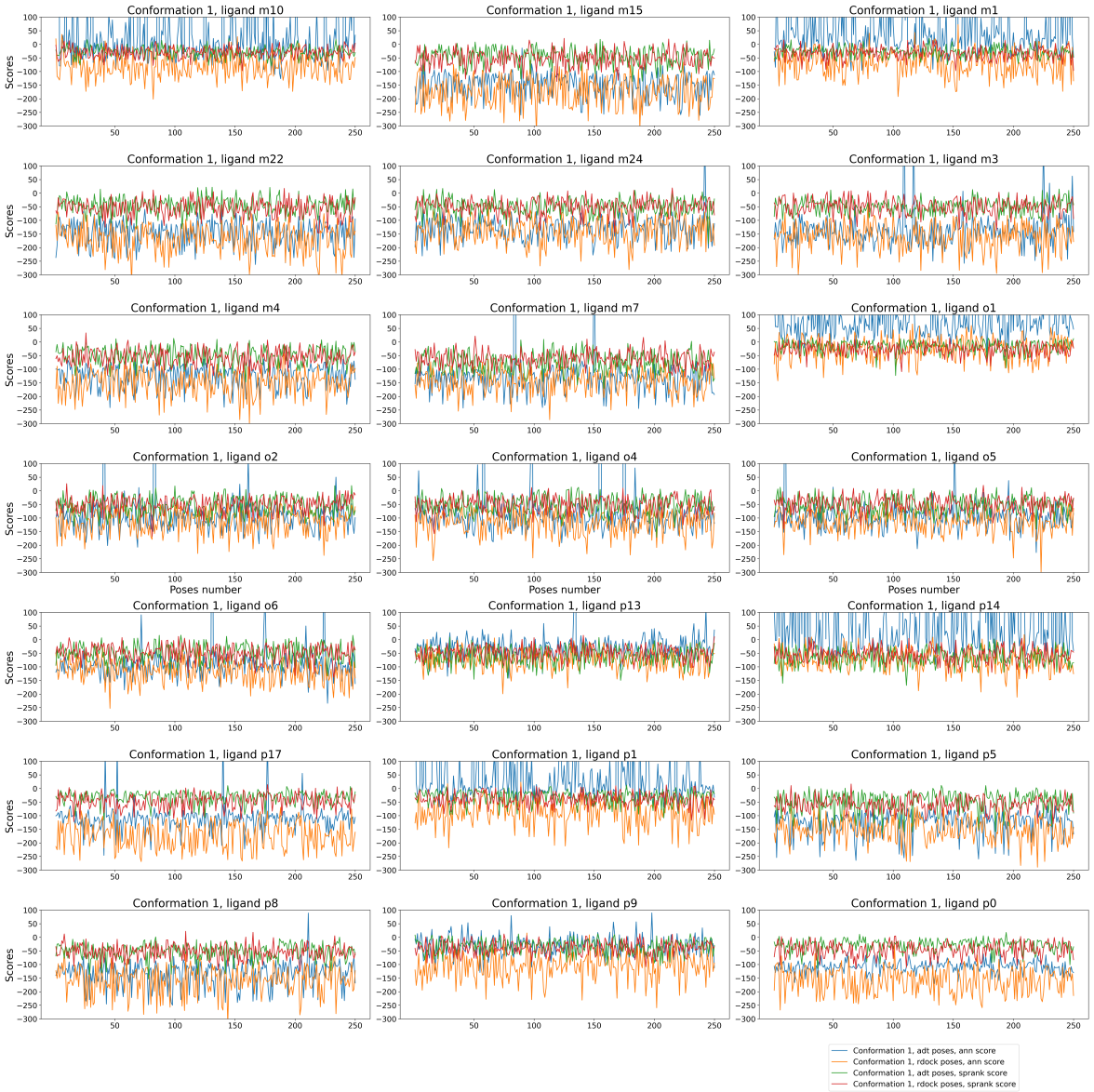

Figure S12: Docking scores for all the generated poses on Site 1, color-coded by scoring function and docking software: AnnapuRNA (blue/orange) and SPRank (green/red). Blue and green lines represent poses generated with AutoDock GPU, while orange and red lines represent those from rDock.

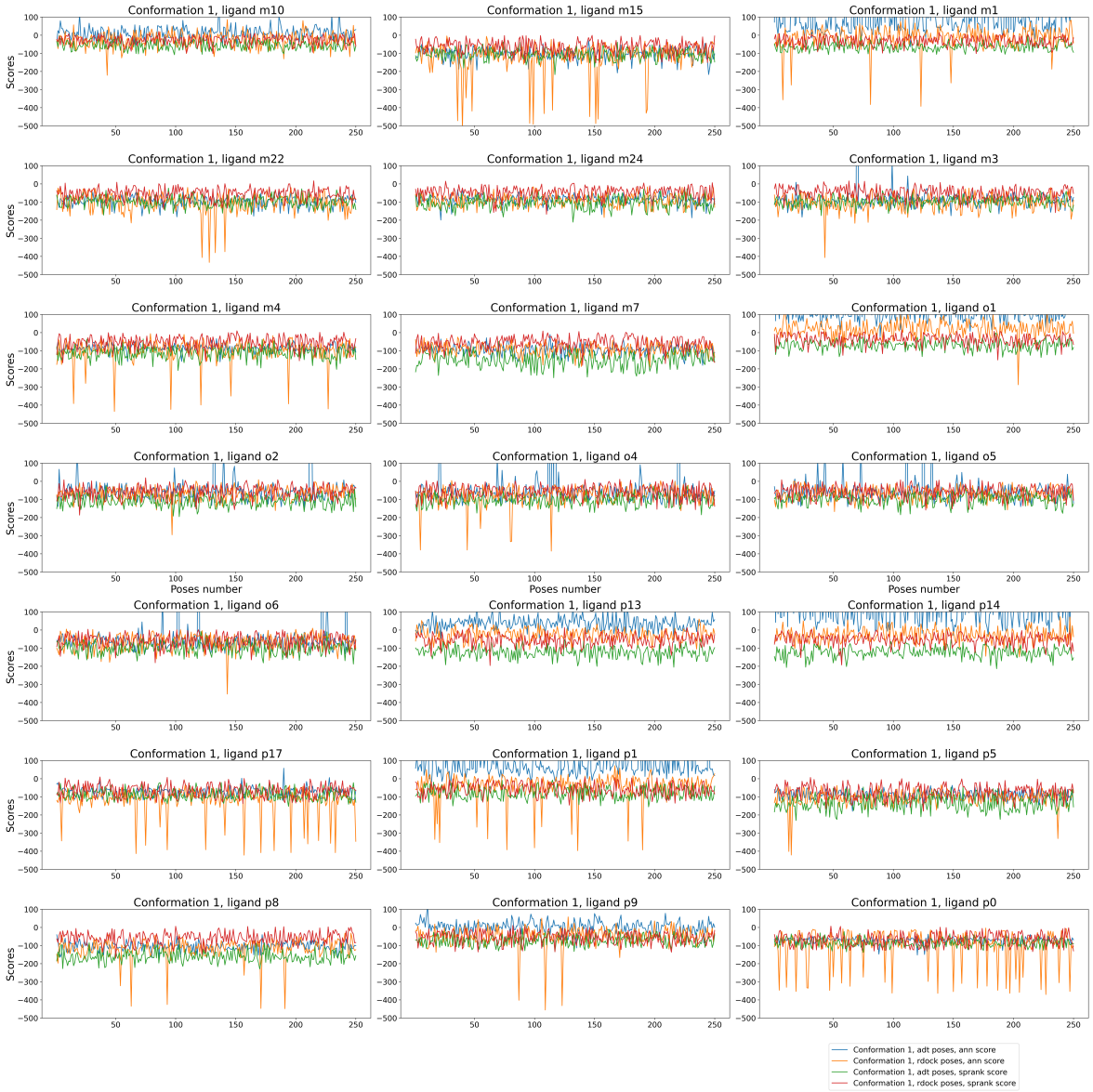

Figure S13: Docking scores for all the generated poses on Site 2, color-coded by scoring function and docking software: AnnapuRNA (blue/orange) and SPRank (green/red). Blue and green lines represent poses generated with AutoDock GPU, while orange and red lines represent those from rDock.

Application Case 3:
Non equilibrium binding free energy estimation

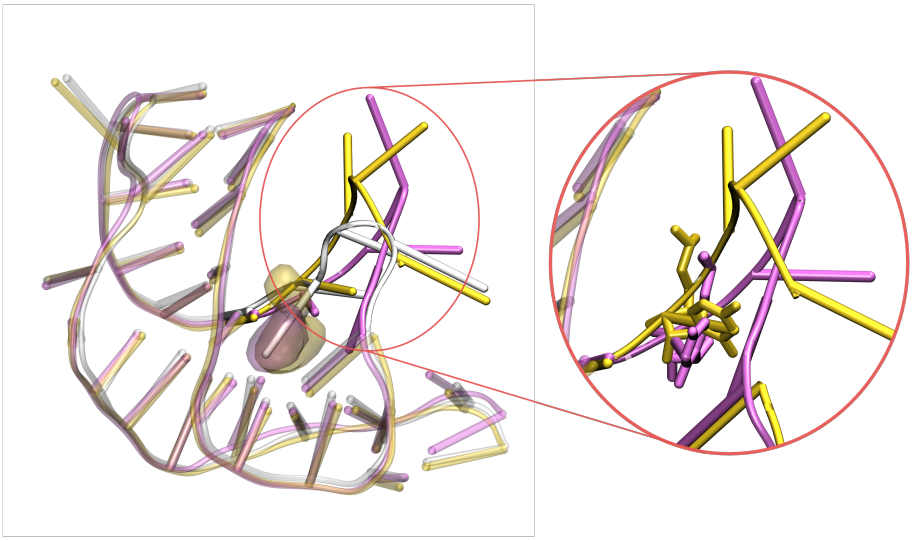

Figure S14: Structural superposition of the apo (white), cognate ligand-bound (purple), and synthetic ligand-bound (yellow) states of the receptor, highlighting the conformational rearrangement of Stem 1. A close-up view, on the left, reveals the distinct orientations of residue C15 induced by the different ligands.

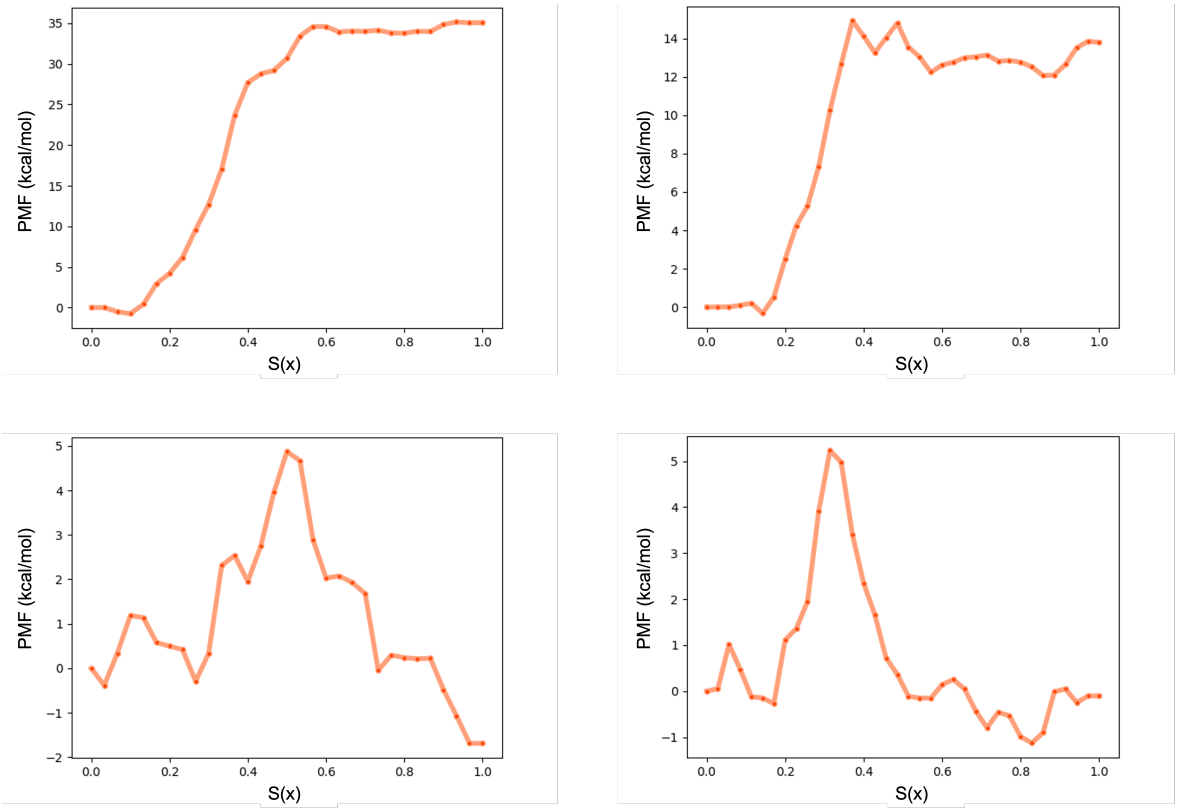

Figure S15: PMF reconstruction using Jarzynski Equality (JE) for both unbinding (upper section) and binding (lower section) simulations in TIP4P-D water model for the cognate (A) and synthetic (B) ligands.

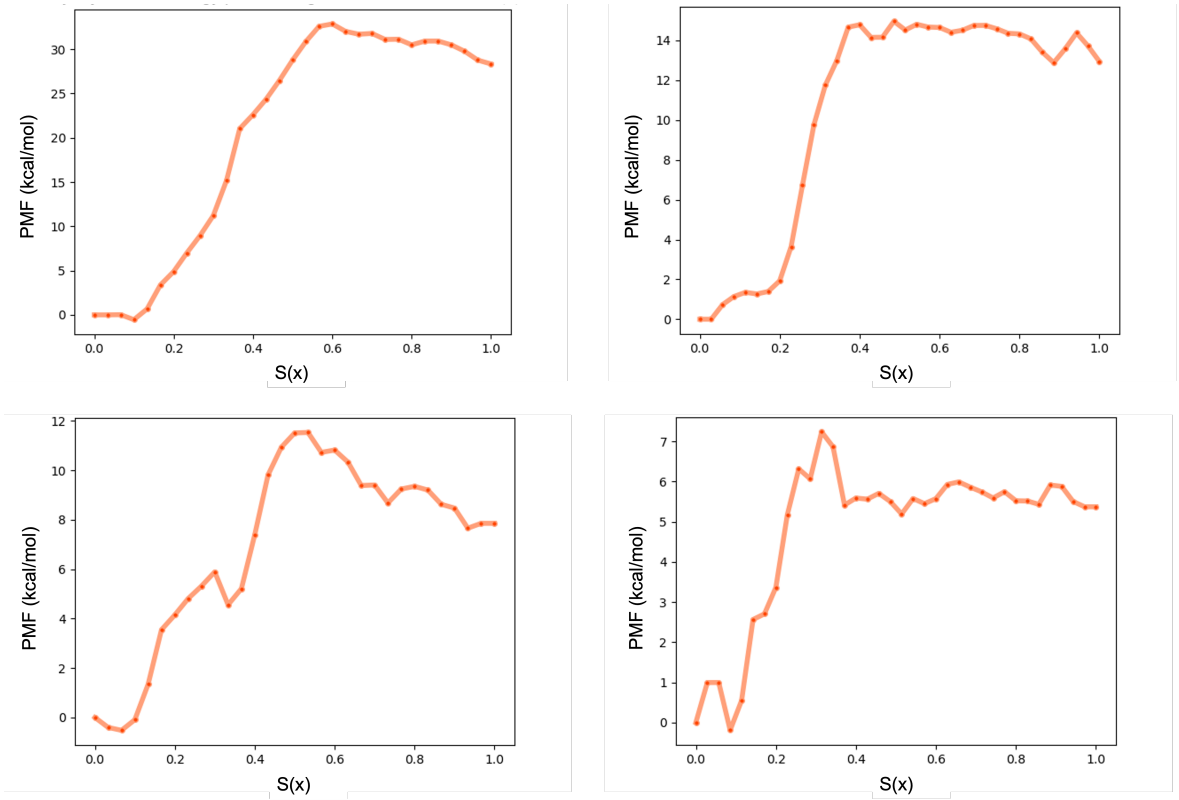

Figure S16: PMF reconstruction using Jarzynski Equality (JE) for both unbinding (upper section) and binding (lower section) simulations in TIP3P water model for the cognate (A) and synthetic (B) ligands.

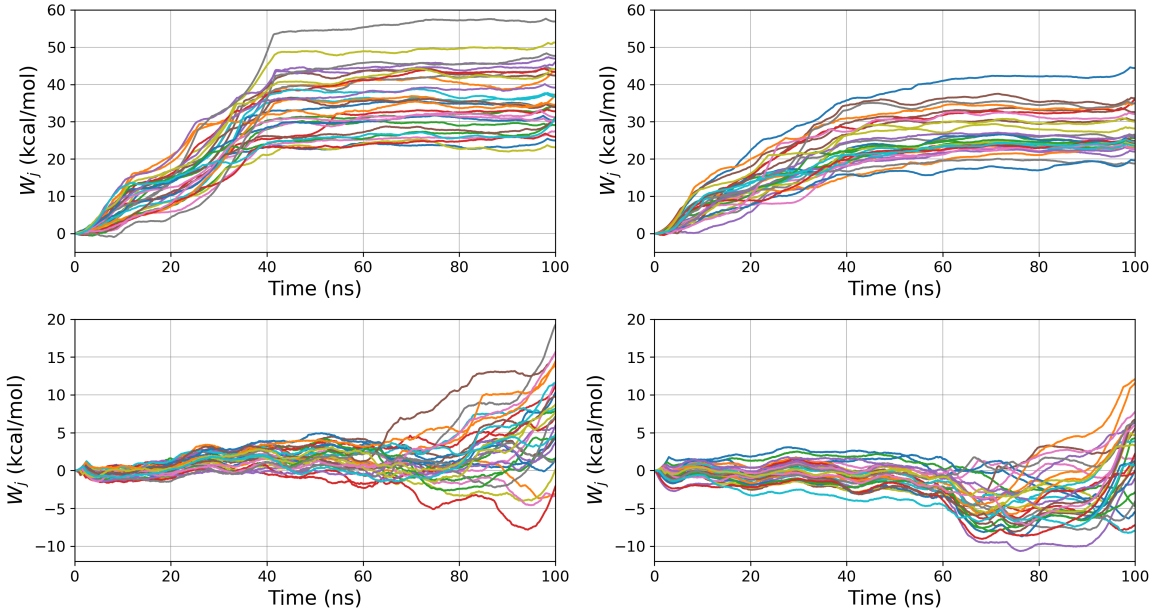

Figure S17: Jarzynski work profiles for the tautomer of the cognate ligand regarding unbinding (upper section) and binding (lower section) simulations measured over a simulation time of 100 ns in TIP4P-D (A) and TIP3P (B).

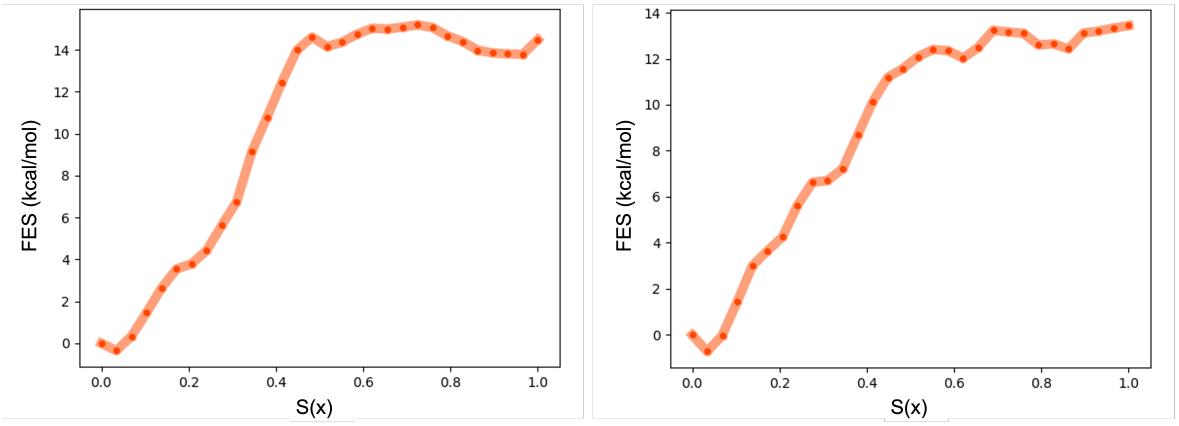

Figure S18: Free energy profiles along S(x) obtained by applying CFT to SMD simulations in TIP4P-D (A) and TIP3P (B) for the tautomer of the cognate ligand.

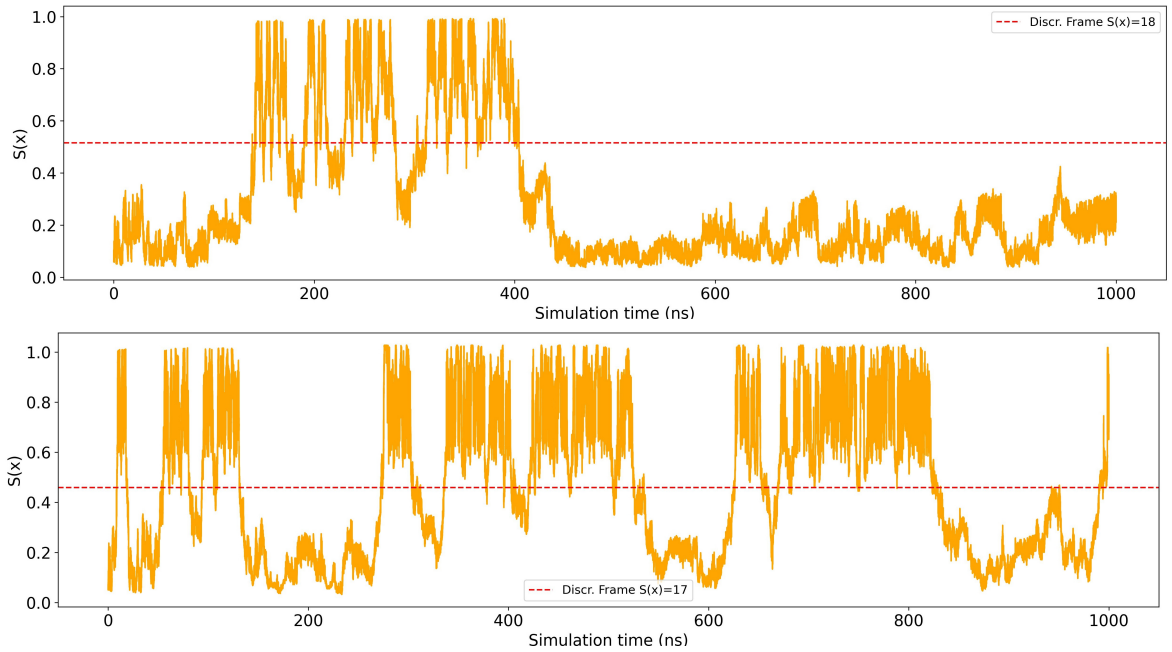

Figure S19: Time series of the collective variable S(x) along the metadynamics trajectories for the cognate (top) and synthetic (bottom) ligands. The red dotted line represents the discriminating frame used for free energy calculations in both cases.

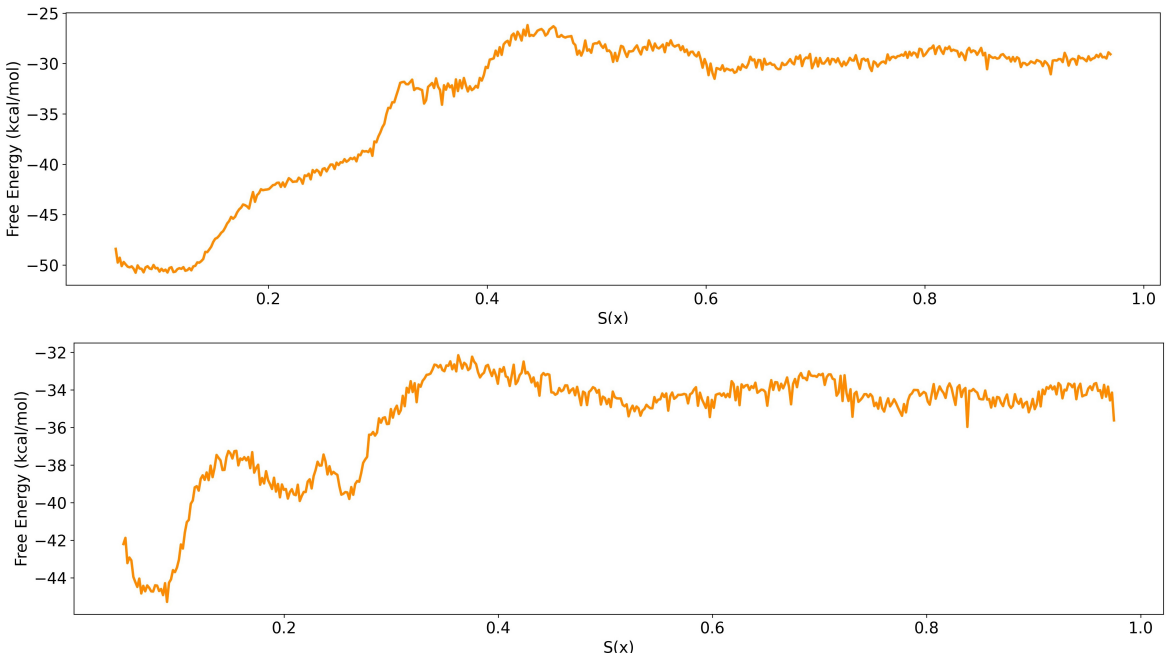

Figure S20: Reweighted free energy profile as a function of the collective variable S(x) from metadynamics simulation for the cognate (top) and synthetic (bottom) ligands.

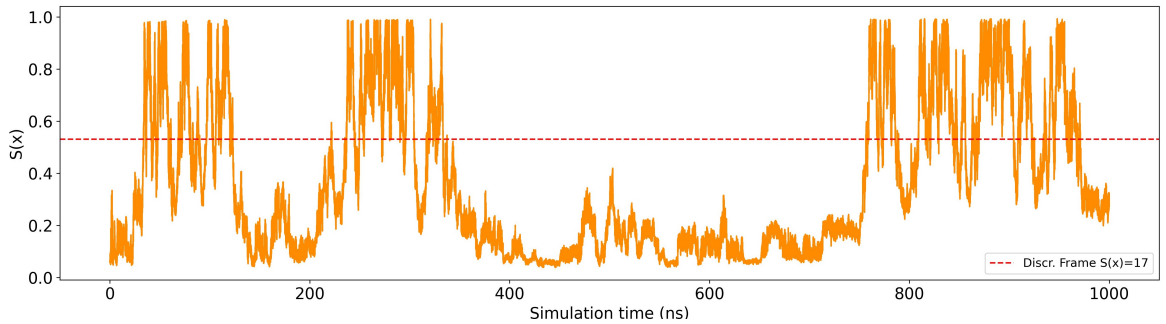

Figure S21: Time series of the collective variable S(x) along the metadynamics trajectory for the tautomer. The red dotted line indicates the discriminating frame used for free energy calculations.

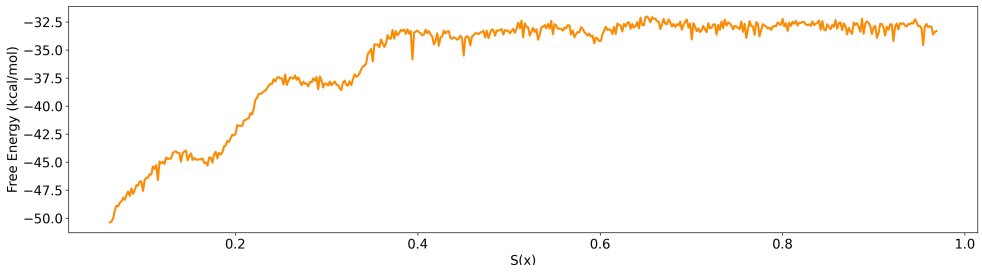

Figure S22: Reweighted free energy profile as a function of the collective variable S(x) from metadynamics simulation for the tautomer.

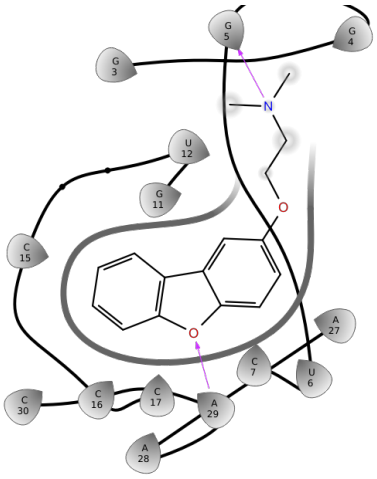

Figure S23: 2D representation of the hydrogen bonds established by the synthetic ligand within the riboswitch-preQ1 complex.

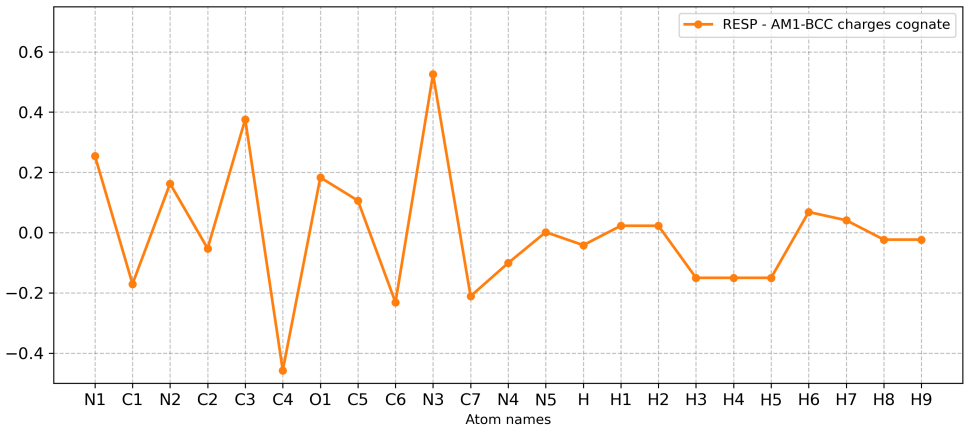

Figure S24: Atomic charge differences between RESP charges (calculated by PlayMolecule) and AM1-BCC charges (calculated by antechamber) for the cognate ligand.

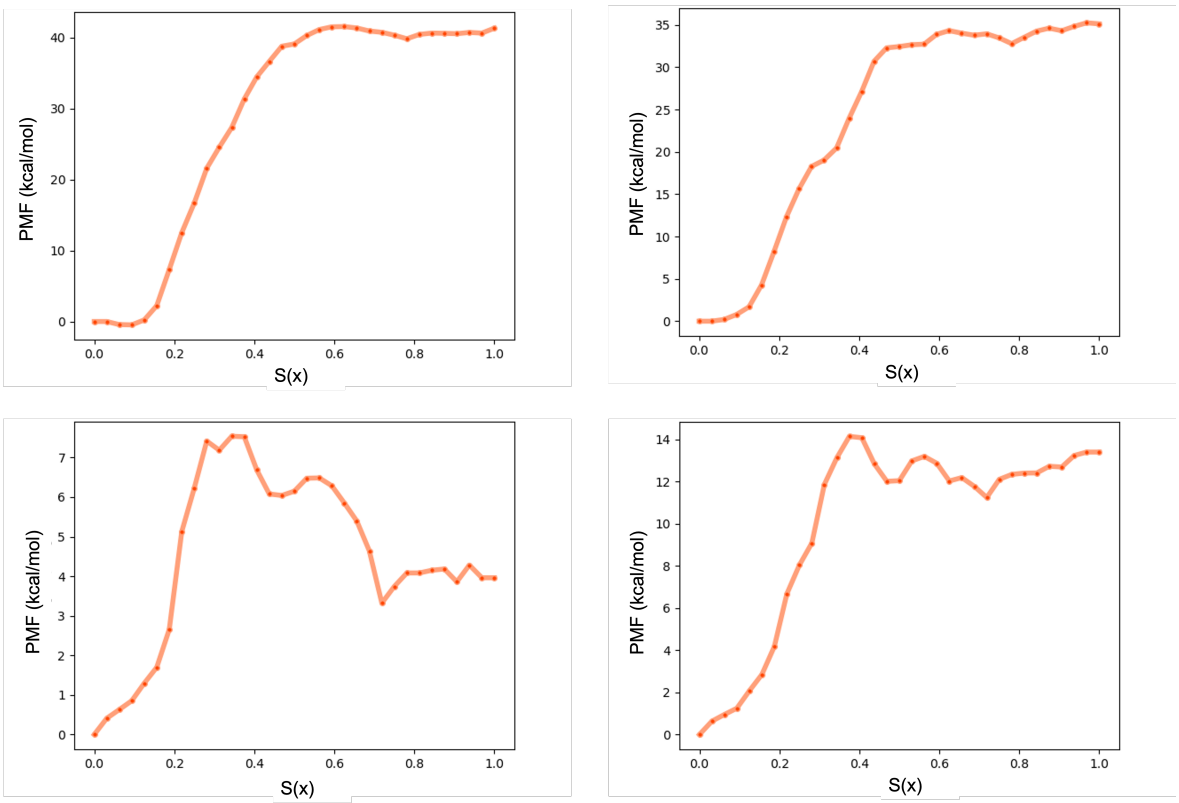

Figure S25: PMF reconstruction using Jarzynski Equality (JE) for both unbinding (upper section) and binding (lower section) simulations in TIP4P-D (A) and TIP3P (B) water model for the cognate ligand with RESP charges

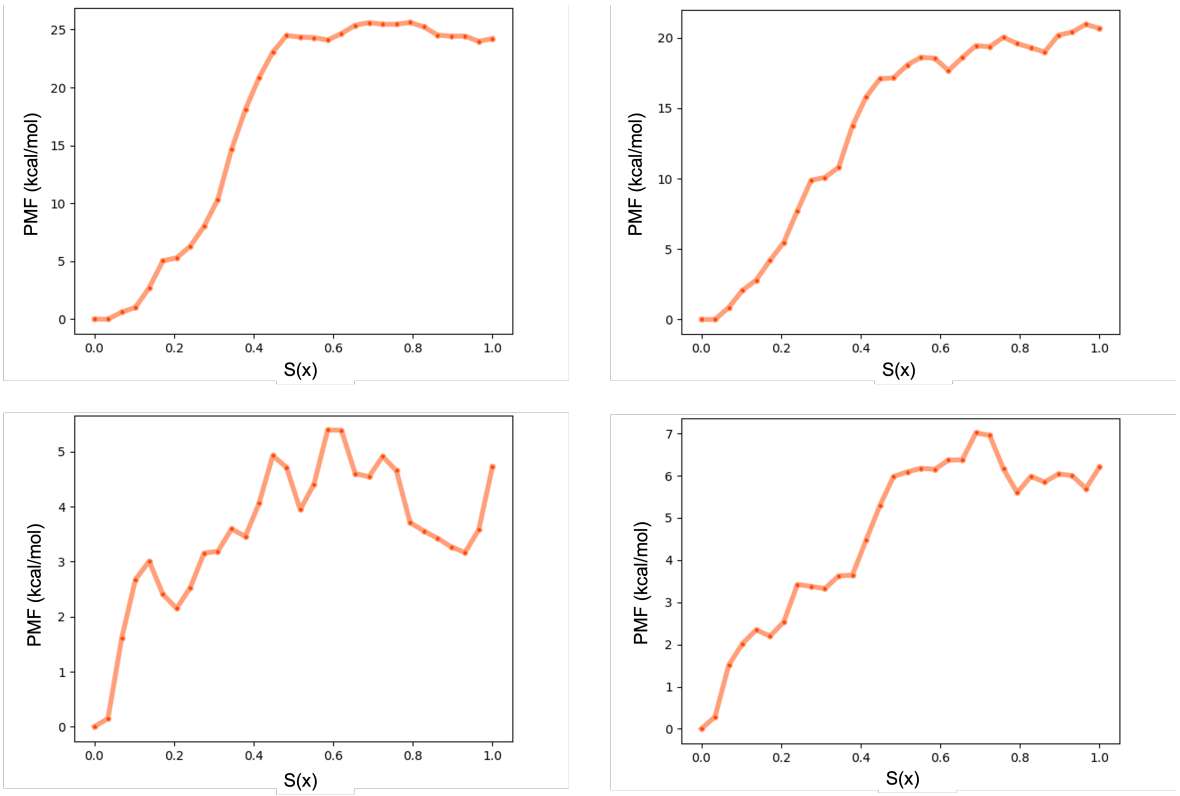

Figure S26: PMF reconstruction using Jarzynski Equality (JE) for both unbinding (upper section) and binding (lower section) simulations in TIP4P-D (A) and TIP3P (B) water model for the tautomer of the cognate ligand.