Probing Local Atomic Environments to Model RNA Energetics and Structure

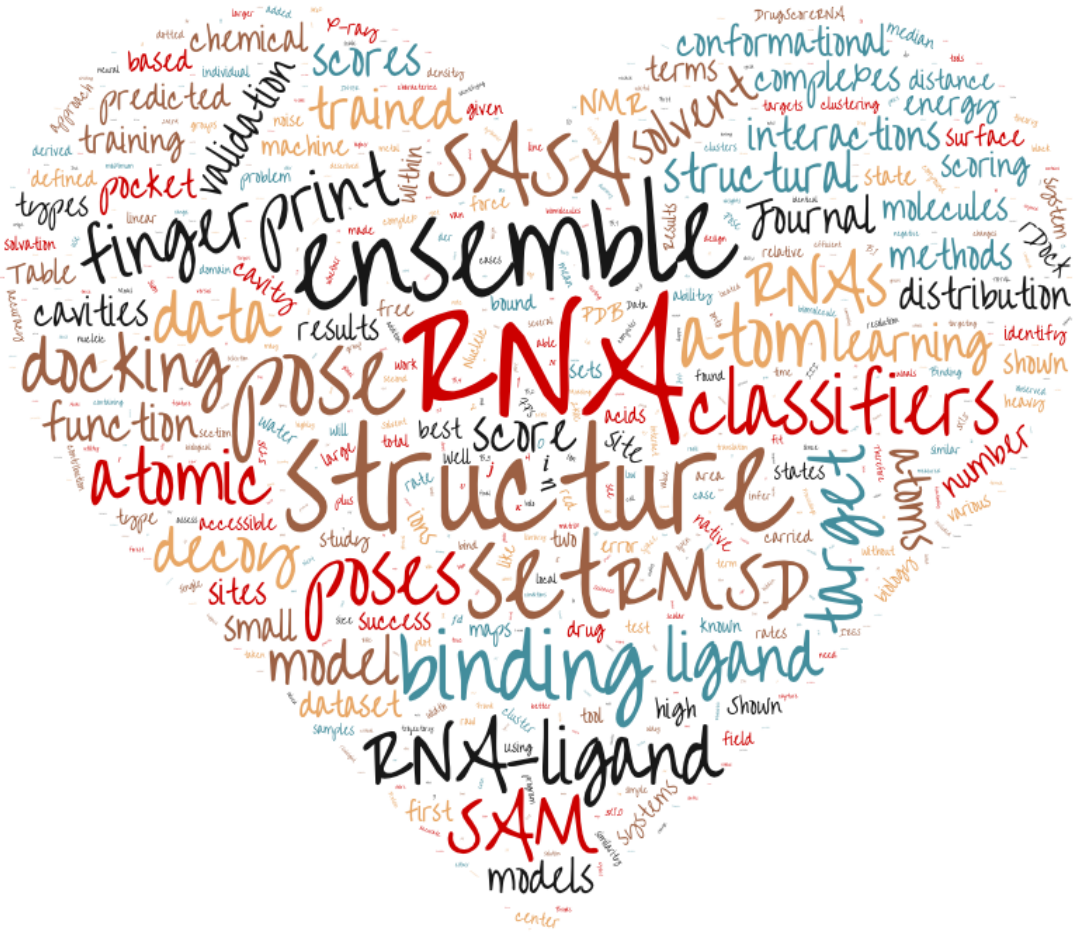
by

Jingru Xie

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Physics and Scientific Computing)
in the University of Michigan
2020

Doctoral Committee:

Assistant Professor Aaron T. Frank, Co-Chair
Associate Professor Kevin Wood, Co-Chair
Assistant Professor Yuekai Sun
Assistant Professor Qiong Yang
Professor Michal R. Zochowski



Word cloud in the shape of LOVE generated from all the texts of this thesis.

Jingru Xie
jingrux@umich.edu
ORCID iD: 0000-0002-6988-1630

Acknowledgments

First, I would like to thank my advisor, Aaron Frank, for being a great advisor and a friend. Thank you for admitting me into the lab and guiding me into the area of computational biophysics research. I learned a lot of valuable knowledge from you, from programming to problem-solving skills. Throughout our numerous meetings, you were always so patient in discussing all aspects of each project in detail, and listening to my results, thoughts and concerns. I am grateful to have been working with you for the past years, gaining research skills, and exploring exciting new ideas. You have enlightened me in my times of depression and anxiety, and I would like to thank you for your always support on my career decisions.

I would like to thank Professor Kevin Wood and Professor Yuekai Sun, who have provided invaluable support and advice since the preliminary exam. To my other two committee members, Professor Michal R. Zochowski and Professor Qiong Yang, thanks for being my committee and suggesting insightful next steps.

Thanks to all my lab mates, Dr. Indrajit Deb, Dr. Sahil Chhabra, Dr. Kexin Zhang, Ciara Witt, Yichen Liu, Ziqiao Xu. Thank you, Indrajit, for all the afternoon conversations we had together on research, career, and life. Thank you for your help on the SAM-I riboswitch project, and your valuable suggestions when proofing my thesis. Thanks to our alumnus, Sahil, who had the longest overlap with my time, thank you for working together on the RNAPosers project. Thank you, Kexin, for sharing with me your experience in the thesis writing and defense preparation process, and for all your effort on pushing forward and

making our manuscript ready for publication during the time I was off campus. Thank you, Ciara, for fixing language problems in my writing, taking notes for my practice presentation, sharing your stories with me, and being the environment regulator of the lab before and during this pandemic. And to the two youngest members in the lab, Yichen and Ziqiao, thank you for your help in the past year. I enjoy so much your company.

I would like to thank the Physics department for providing me with the opportunity and support, both mentally and financially, for me to join the Ph.D. program and for the full freedom to explore the research area. Thank you, Lauren Segall, Dr. Finn Larsen, Christina Zigulis, and Dr. Vanessa Sih for making the graduate program an active yet pleasant working and learning environment. Thanks to all the professors and instructors I have taken classes with, from whom I received instructions and learned valuable knowledge. Although I have not been around in the department in the last couple of years, I received continuous, timely support from the department.

Finally, I would like to thanks all my mentors, friends, and family. Thanks to my internship mentors, Dr. Ji Liu and Dr. Lei Yuan at Kwai Seattle AI Lab. The internship taught me great lessons on both technical and soft skill aspects that I continued to benefit from in my last year of research and thesis writing. Thanks to my colleagues in Seattle and Beijing. The time I spent with you left me lots of precious memories. Thanks to all my friends in Ann Arbor, for your warm accompany since the time I first came to the States. Thank you, Hao Huang, for always being on my side over the good and bad times. Looking back on the past years, I can not imagine how much we have gone through together, and I could not have done this without you. Finally, thanks to my mom and dad for your unconditional love, for making me who I am.

Contents

List of Figures

List of Tables

List of Appendices

List of Acronyms

AUC Area Under Curve.

MD Molecular Dynamics.

ML Machine Learning.

MLP Multi-Layer Perceptron.

PDB Protein Data Bank.

RF Random Forest.

RMSD Root-Mean-Square Deviation.

RNA Ribonucleic acid.

ROC Receiver Operating Characteristic.

SASA Solvent Accessible Surface Area.

t-SNE T-distributed Stochastic Neighbor Embedding.

XGB Extreme Gradient Boosting.

Abstract

Ribonucleic acids (RNA) are critical components of living systems. Understanding RNA structure and its interaction with other molecules is an essential step in understanding RNA-driven processes within the cell. Experimental techniques like X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, and chemical probing methods have provided insights into RNA structures on the atomic scale. To effectively exploit experimental data and characterize features of an RNA structure, quantitative descriptors of local atomic environments are required. Here, I investigated different ways to describe RNA local atomic environments. First, I investigated the solvent-accessible surface area (SASA) as a probe of RNA local atomic environment. SASA contains information on the level of exposure of an RNA atom to solvents and, in some cases, is highly correlated to reactivity profiles derived from chemical probing experiments. Using Bayesian/maximum entropy (BME), I was able to reweight RNA structure models based on the agreement between SASA and chemical reactivities. Next, I developed a numerical descriptor (the atomic fingerprint), that is capable of discriminating different atomic environments. Using atomic fingerprints as features enable the prediction of RNA structure and structure-related properties. Two detailed examples are discussed. Firstly, a classification model was developed to predict Mg^{2+} ion binding sites. Results indicate that the model could predict Mg^{2+} binding sites with reasonable accuracy, and it appears to outperform existing methods. Secondly, a set of models were developed to identify cavities in RNA that are likely to accommodate small-molecule ligands. The models

were also used to identify bound-like conformations from an ensemble of RNA structures. The frameworks presented here provide paths to connect the local atomic environment to RNA structure, and I envision they will provide opportunities to develop novel RNA modeling tools.

Chapter 1.

Introduction

Ribonucleic acids (RNAs) represent one of the fundamental building blocks of all domains of life. RNA molecules have crucial functions, most notably, as messengers in gene expression, transcribing genetic information from deoxyribonucleic acid (DNA), and translating it into proteins. Not all transcribed RNAs, however, are translated into proteins. Many classes of non-coding RNAs (ncRNAs) are now known to act as independent catalytic or regulatory molecules. With the increased awareness of the importance of RNA in critical biological functions, a rapidly growing number of RNAs, especially ncRNAs, have been discovered to be disease-relevant and were targeted to treat infectious diseases and cancer. To better understand the functionality of RNAs and improve the *in silico* RNA targeting drug design pipeline, there is a need for novel theoretical and experimental models that offer a more guided and rational explanation of RNA structure as dynamics. This work is aimed at the need, mainly from a theoretical perspective. In this chapter, I will introduce the fundamental elements of the RNA structure, discuss the interaction of RNA with other molecules, and survey the computational tools used in the rest of this thesis.

1.1. RNA structure and the structural ensemble

Like other biomolecules, RNA is a large organic molecule composed of a set of organized atoms, typically ranging from hundreds to hundreds of thousands of atoms. The arrangement of atoms in three-dimensional space and the inter-atomic interactions collectively define an RNA structure, or the RNA conformation or conformational state. RNA structure is key to understanding its stability and function. Below I review the basics of an RNA structure, structural ensemble, and how to determine RNA structures experimentally.

1.1.1. Three levels of RNA structure

RNA structure is typically described at three levels: primary, secondary, and tertiary (Figure 1.1a-c). On the first level, RNA is a biomolecule consisting of ribonucleotides. Each ribonucleotide contains a nitrogenous base, a five-carbon sugar (ribose), and a phosphate group. There are four natural ribonucleotides classified into two groups, purines (Adenine and Guanine, denoted by A or ADE and G or GUA, respectively) and pyrimidines (Cytosine and Uracil, denoted by C or CYT and U or URA, respectively), which differ only on the nitrogenous base. Two ribonucleotides can be linked together by phosphodiester bonds that connect the 5'-phosphate group of one ribonucleotide to the 3'-hydroxyl group of the other. Sequentially linked ribonucleotides form the RNA strand. An RNA molecule consists of one or more RNA strands, and the sequence of ribonucleotides in each strand collectively form the RNA primary structure (Figure 1.1a). At the second level, RNA secondary structure describes the base-pairing status between each pair of ribonucleotides (Figure 1.1b). In addition to the covalent phosphodiester bonds, ribonucleotides also interact with each other through non-covalent hydrogen bonds (H-bonds) and form base-pairs. The base-paring between ribonucleotides is the essential first step for a stable RNA structure and is also the

critical ingredient for gene replication and transcription. Base-paring between nucleotides generally follows the Watson-Crick (WC) base-pairing rules. WC rules state that purines are base-paired to their corresponding pyrimidines (C:G/G:C, A:U/U:A), with an additional weaker non-WC base-pairing (G:U/U:G). The third level, RNA tertiary structure, or 3D structure, describes the spatial arrangements of nucleotides in 3D space (Figure 1.1c). Stacking interactions between bases, electrostatic interactions, and other long-range interactions are the primary driving force of the formation of tertiary structures [1]. Non-canonical base-pairing interactions that do not follow the WC rule could also be induced by tertiary interactions [2, 3]. Ion interactions also come in at this stage to neutralize negative charges on RNA backbone and stabilize its tertiary structure. The presence of RNA tertiary structure and the ability of RNA folding into compact 3D structures make RNAs well suited for shape-based functionality. *In this thesis, unless otherwise specified, the term "structure" will refer to the 3D structure.*

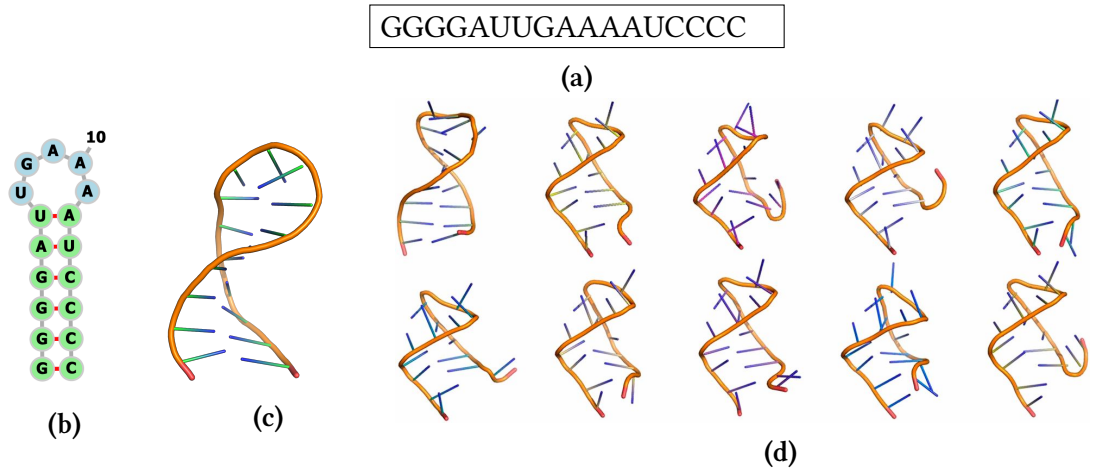


Figure 1.1: (a-c) Illustration of the three levels of RNA structure: (a) primary, (b) secondary and (c) tertiary structures of the anticodon stem-loop from E. coli tRNA (PDB ID: 1KKA) [4]. The secondary structure was generated using Forna [5] web server and tertiary structure was rendered in PyMOL [6]. (d) RNA "Dance": the diverse structures of a simple RNA hairpin. (d) was taken from [7].

1.1.2. RNA structural ensembles

Instead of maintaining a single rigid structure, most RNAs dynamically sample a variety of conformations and can have many excited or transient states (Figure 1.1d). The set of structures that an RNA can occupy under a set of environmental conditions collectively form a structural ensemble. Conformers within the ensemble can have various probabilities. Correctly folded structures with the lowest free energy typically have the highest probability and account for the largest proportion of the population, and are also known as "native" structures. In contrast, higher free energy (or transient) structures are rarer under normal physiological conditions.

The RNA ensemble changes in response to a diverse array of environmental factors, including temperature, pH, binding to other macromolecules, etc. These changes can be viewed as a redistribution of the conformations within the ensemble rather than creating new conformations. For example, when a free RNA molecule is placed in contact with a small-molecule ligand and ultimately form an RNA-ligand complex, the relative population of the *apo* (ligand-free) state, which dominates the free-state RNA ensemble, is decreased and correspondingly, the *holo* (ligand-bound) state population is increased. The transient state kinetics and the transition to low-populated states are essential components to understanding RNA function. For example, one study of excited states of HIV-1 TAR RNA shows that changes in base-pairing status potentially activate the transcription of the HIV-1 genome, which ultimately allowed for the discovery of new drug targets [8].

1.1.3. RNA structure descriptors

Describing RNA structures is one of the primary steps to analyze their interaction with other molecules and to rationalize their cellular function. Among all basic RNA 3D structure de-

scriptors, Root Mean Square Deviation (RMSD) and Solvent Accessible Surface Area (SASA) stand out as the most common ones.

Root Mean Square Deviation (RMSD)

Comparing RNA 3D structures yields valuable information on their functional conservation and structure-function relationships. The most commonly used criteria to discriminate RNA 3D structures is Root-Mean-Square Deviation (RMSD). RMSD is a measure of distances between the atoms in a structure and another reference structure with superimposed conformations (Equation 1.1).

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\delta_i^2} \tag{1.1}$$

Here N is the number of atoms in the structure, and δ_i is the euclidean distance between atom i in the structure and the corresponding atom in the reference structure. In practice, the two structures are aligned with each other to obtain optimal superposition before the calculating RMSD to eliminate the contribution from pure translation and rotation movements needed to align the two structures. For example, if two identical structures were only offset by a distance, they are still considered identical and should result in 0 RMSD.

In practice, the RMSD has diverse applications in both RNA structure and energetics analyses. Firstly, the RMSD is used to quantitatively measure the similarity between two structures. For example, in RNA structure predictions, the RMSD from the known experimental structure is vital in assessing the quality of the predicted structure [9]. Secondly, the RMSD is used to infer the energetics of an RNA ensemble. For example, when studying the folding kinetics of an RNA structure [10, 11] or RNA-ligand binding pathways [12], the RMSD from the native, starting structure could serve as a reaction coordinate to quantify whether the

RNA is in folded form or as constraints to guide the ligand-binding process.

Solvent Accessible Surface Area (SASA)

In cells, RNAs exist in the aqueous environment and interact with other molecules through their solvent accessible surface area (SASA). The concept of SASA was introduced for proteins to estimate the relative changes in solvent accessible surface upon folding in 1971 by Lee & Richards [13]. Like many other biomolecular methods and properties first introduced for proteins, SASA was introduced for nucleic acids not long after in 1979 by Charles & Sung-Hou [14] as a way to describe molecular, residuewise and atomwise exposure of nucleic acids to solvents and to study their conformational transitions.

By definition, SASA is the exposed surface of a molecule or an atom accessible to the solvent. Since the surface of a biomolecule is formed by interlocking spheres of each atom, this irregular shape hinders an accurate analytical formula for SASA calculation, while numerical approximation is employed. SASA was traditionally computed by the "rolling probe" method. A spherical ball with radius similar to a solvent molecule is used as a probe and rolled over the whole van der Waals surface of a biomolecule, and SASA of the biomolecule is thus obtained by numerical integration over all possible positions of the center of the probe. The numeral integration is performed by slicing the biomolecule into small, uniformly separated planes with separation h, and the surface in each plane is considered a 2D circle. The problem of computing surface of a 3D object (the biomolecule) is therefore converted into the sum of perimeters of 2D objects (circles) multiplied by separation h.

SASA is an important structural feature of RNA. Firstly, Solvent Accessible Surface Area (SASA) is related to chemical reactivity, RNA-protein interaction [15, 16] and RNA-ligand interaction. For example, SASA is an important term to include in computer docking programs [17] when predicting ligand-binding sites. Secondly, SASA is often used to calculate

solvation energy, the transfer free energy of moving the RNA from vacuum to aqueous environment, and assist in computational modeling of RNA dynamics. In many current solvation models for biomolecules, the nonpolar contribution to the solvation free energy is estimated from the SASA scaled by an effective surface tension parameter [18, 19].

1.1.4. Experimental determination of RNA structures

Tertiary interactions in RNA result in significant diversity in RNA structures [2]. Various stacking arrangements and sugar conformations, in addition to unusual base-pairing interactions caused by tertiary interactions, lead to the heterogeneity in RNA tertiary structures. Biophysical experimental techniques such as X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, and other methods like cryo-electron microscopy (cryo-EM) have revealed this diversity in RNA 3D structures, and they are still the most important and accurate tools used to characterize RNA structures [20]. Chemical probing methods like DMS [21], SHAPE (Selective 2'-hydroxyl acylation analyzed by primer extension) [22] and LASER (Light-Activated Structural Examination of RNA) [23] can also be used to analyze RNA local structures by letting the RNA interact with chemical probing reagents and then infer structural information from the reactivity pattern. Chemical probing methods typically provide indirect information about the structure and require more subsequent computational analysis to produce structure models.

X-ray crystallography and NMR spectroscopy are the two major classes of physics-based experimental techniques used for RNA structure determination. X-ray crystallography solves an RNA structure by examining the electron density map of heavy (non-hydrogen) atoms in an RNA crystal. The type and position of an atom in an RNA molecule can be determined based on the fact that different types of atoms diffract X-ray photons in different directions.

X-ray crystallography is known for its capability to solve structures at a high resolution, and the majority of experimental structures available for RNA are X-ray structures. X-ray crystallography could solve RNA structures of various sizes, ranging from simple stem-loops of 10 nucleotides (nts) to the ribozymes and riboswitch functional groups of hundreds to thousands of nts. NMR spectroscopy is also a powerful tool for studying RNA 3D structures, especially in solution. NMR spectroscopy determines RNA structures by placing the RNA molecule in a magnetic field and analyzing the resonance patterns found when varying the frequency of the magnetic wave to. In contrast to X-ray crystallography, in NMR experiments, RNA molecules are in solution and, therefore, can capture RNA dynamics. NMR experiments also yield several structural-relevant measurements, including chemical shifts and interatomic distances restraints that are valuable inputs for computational modeling.

However, experimental techniques also have some limitations. Firstly, high-resolution experiments are time-consuming and require significant expertise. The preparation of RNA crystals used for X-ray crystallography vary from structure to structure, and the crystals are extremely delicate since RNA structures are highly flexible and are sensitive to stabilizing compounds used in crystallography [24, 25]. On the other hand, although NMR spectroscopy could directly determine RNA structures in solution and less preparation is required than X-ray crystallography, the high-molecular weight of the RNA molecule limits the accuracy and suitability of standard solution NMR techniques [26, 27]. Secondly, experimental techniques, especially X-ray crystallography, lack the ability to characterize the dynamics or ensembles of RNA. Crystallography can only show the most probable structure that dominates the ensemble. Recently methods have been proposed to experimentally characterize RNA transient states [28] using NMR spectroscopy, but the framework is not yet as well established. Thirdly, since both X-ray and NMR determine structures *in vitro*, meaning that RNA molecules have to be taken out of the living cell and pre-processed before experiments

can be performed, the structures determined using these methods may not reflect the functional states of the RNA *in vivo* (in the cell). Chemical probing methods can overcome this limitation to some extent, but they lack the ability to determine the structural details and achieve the high accuracy that X-ray and NMR methods could achieve. Despite the limitations, experimental techniques provide insights into RNA function at atomic resolution and continue to be invaluable in structural biology. They are the foundation of other computational methods for structural and energetics determination that will be discussed in Section 1.3.

1.2. RNA intermolecular interactions

Most RNAs do not act in isolation but interact with other molecules to execute their functions, including solvents (water and ions), small-molecule ligands, other RNAs, and proteins. RNA-solvent interactions are considered to be the primary stabilizing factors for RNA secondary and tertiary structures, depending upon the nature of the surrounding liquid. A large number of RNAs co-exist with proteins and form RNA-Protein complexes (RNPs) [29]. Some RNAs interact with other RNAs to regulate translation and transcription [30, 31]. Moreover, RNA interacts with ligands (small molecules bound to RNAs), ranging from simple amino acids (the building block of proteins) to metabolites like flavine mononucleotide [32], to perform their biological function. In this section, I will discuss the interaction between RNA and other molecules, with more emphasis placed on solvents and small-molecule ligands.

1.2.1. RNA-solvent interaction

The presence of water molecules is crucial to the RNA tertiary structural stability and its function. Water molecules come in at every stage of RNA structure formation and serve as

lubricant when RNA interacts with other molecules [33]. Nucleic acids typically interact with water molecules via polar or ionic groups, which form hydrogen-bonds when coming into contact with hydrogen atoms in water molecules, and RNAs have an especially greater extent of hydration due to the extra oxygen atoms (O2') present on ribose (the five-carbon sugar).

RNAs also interact with ions, both cations (positively charged ions) and anions (negatively charged), through their highly negatively charged phosphodiester backbone. Both long-range and short-range forces come into play in RNA-ion interactions [34]. Among all ions, cations like potassium (K^+), magnesium (Mg^{2+}), and their compounds with water molecules have been shown to play important roles in stabilizing RNA tertiary structures and in modulating biological function in living organisms [35, 36]. Under-concentration of ions will result in the "collapse" of folded RNA tertiary structures [34]. The importance of positively-charged ions is mainly highlighted in two aspects. Firstly, since the RNA backbone carries negative charges, positively charged ions are essential in neutralizing the overall charge and maintaining structural stability [37, 38]. Many RNA structures solved in experiments include positively charged ions as stabilizers. For example, the first RNA structure solved experimentally, the yeast tRNA structure, is stabilized by Mg^{2+}ions [39]. Secondly, in some cases, RNAs bind metal ions selectively to perform their functions, and the presence of ions helps to facilitate the interaction between an RNA and other macromolecules. For example, Mg^{2+}participate in inducing conformational changes at the internal ribosome entry site of the hepatitis C virus [37].

1.2.2. RNA-ligand interaction

RNAs interact with ligands to induce a change in conformation as well as regulate biological activities. When characterizing the RNA-ligand interaction, several factors must be taken into account, including the type of ligands, the structure of the RNA, and the specific site of the RNA structure that the ligand binds. Specificity and affinity are the two dimensions used to assess the quality of interaction of RNA-ligand binding. Binding affinity is a direct measure of how tight a ligand binds to an RNA. Specificity, on the other hand, is a measure of how selective the binding site binds to a specific ligand. If a binding site on an RNA does not bind to any ligands but one, it is said to have high specificity to the one ligand it binds. Specificity, as well as affinity, is a useful measure in assessing RNA-ligand interaction, when designing drugs, for example, one would expect the drug binding site should be unique to the drug, but not a competitive site that the drug molecule has to compete with others. In the case when a binding site has high specificity, it could potentially achieve a high success rate even at a lower affinity.

Riboswitch

Riboswitches are a family of RNAs that bind to small-molecule ligands to regulate their own activity, which best exemplify the interaction between RNA with small-molecule ligands. Riboswitches are cis-regulatory RNA domains composed of the aptamer domain and gene expression platform. The aptamer domain is the part of the riboswitch that possesses binding pockets to bind small-molecule ligands. The gene expression platform, also called the actuator domain, is the functional part of the riboswitch responsible for transcription or translation. Riboswitches are known to bind to a large variety of ligands, and, most importantly, regulate gene expression without the need for protein factors [40]. The "switch"

mechanism works as follows: when the aptamer domain binds to the small molecules, conformational changes in the actuator domain are induced, which in turn block or mediate transcription or translation (see Figure 1.2). A riboswitch is just like a genetic switch, with the ability to flip genes "on" and "off" in response to small-molecule ligands. Structures of the riboswitch aptamer domain have been well studied in experiments, with over 100 structures solved under various conditions and deposited to the PDB [20].

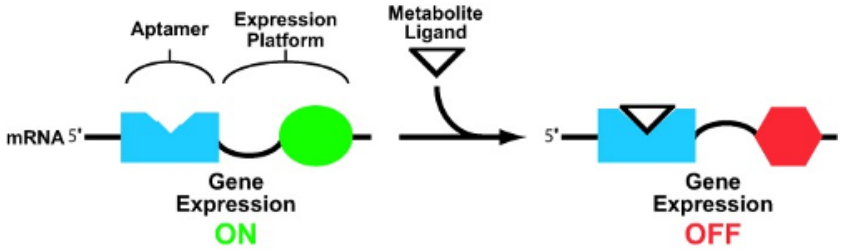


Figure 1.2: Illustration of riboswitch mechanism. Figure from [40].

1.3. Computational methods in RNA research

With the limitations of experimental methods and the increase in available computing power, computational methods have now been adapted and have become more and more popular to study RNA structure, dynamics, and energetics. Computational methods play an important role in determining of RNA secondary [41–43] and tertiary structures [44–48]. In this section, I will describe the two most commonly used computational tools in RNA research. One is Molecular Dynamics (MD) simulation, which utilize physics-based or statistical potential functions (force fields) to simulate dynamics of biomolecules [49]. The other one is Machine Learning (ML), the data-based computational strategy that constructs powerful models by directly learning the behavior and functional relationships from existing data.

1.3.1. Molecular dynamics simulation

MD simulations are one of the most common computational techniques in biophysical science and has been used in data preparation or verification in most chapters in this thesis. MD has been widely adopted as a tool to produce trajectories that describe the dynamical properties of each atom in a complex biological system. MD has assisted in determining structures, revealing biomolecular internal motions, uncovering RNA tetra-loop folding pathway[50] as well as assessing interactions between biomolecules [51].

MD simulations model the individual motions of each atom as a function of time using numerical methods, where the time dimension is discretized into small, finite steps to facilitate the numerical integration of equation of motions between time steps [52]. In MD simulations of biomolecules with N atoms, atoms are modeled as point particles with mass m_i, position x_i and velocity v_i. The system evolves following Newton's Law of motion:

$$m_i \frac{dv_i}{dt} = F_i = -\nabla_i V(x_1, ..., x_N) \tag{1.2}$$

$$\frac{dx_i}{dt} = v_i \tag{1.3}$$

With an initial set of positions and velocities for each atom in the system, the force experienced by each atom can be determined by the force field function V (though named as "force field", V is actually a potential function). The velocity and position for the next time step can be determined by integrating the corresponding differential equations. The force field, the function linking atomic coordinates to the time derivative of velocity, is therefore essential to produce the dynamics. In biological research, the force field typically consists of bonded and non-bonded interactions between atoms. Bonded interactions include stretch-

ing of bonds and angles between connected bonds. Non-bonded interactions are composed of longer-range interactions, including van der Waals and electrostatic interactions. The functional form and parameters involved in the force field function are either empirical or approximations of theoretical calculations. The simplicity of the framework has made MD the primary method for determining the conformational landscape and structural ensembles for RNAs and other biomolecules.

The large size of biomolecules leads to a natural tradeoff between accuracy and computational complexity in the MD simulation of biological systems. On the one hand, since simulations can be achieved in a way that atomic coordinates are updated by numerical integration at discrete timestep on the scale of femtoseconds, a reasonably small error in each time step can accumulate to become significant after thousands of steps. Although proper integration methods like back-Euler, Verlet integration have been used to reduce the numerical instability, the limitation of machine precision and inaccuracy resulting from the intrinsic error of the force field remains a problem that can lead to incorrect dynamics. On the other hand, biomolecules are usually hundreds of thousands of atoms in size. Biological reactions of interest, like folding and binding-unbinding, occur on the time scale of milliseconds up to minutes [53], which requires millions to trillions of steps of simulations. Thus, a highly efficient force field is required to perform relatively-long simulations.

Additionally, some limitations remain within MD simulations. Firstly, the approximations made in the force field and its empirical nature make it not accurate. For instance, polarizability is sacrificed in most classical force fields and replaced by fixed point charges. Over the course of long simulations, the error in the force field will accumulate and therefore drive the structures to unrealistic direction or result in incorrect conformational landscapes. Some solutions have been proposed for this problem. For example, one can incorporate experimental measurements to guide MD simulations [54], or derive high-accuracy force field

from quantum field theories. The other limitation in MD simulation is the speed sometimes not satisfactory. The simulation of a midsized 996-residue protein for 1 second can take over 171 years on a computer with 16 cores [55, 56]. High-parallel GPU calculations have been adopted, and specially designed hardware has been developed [57] to overcome this limitation.

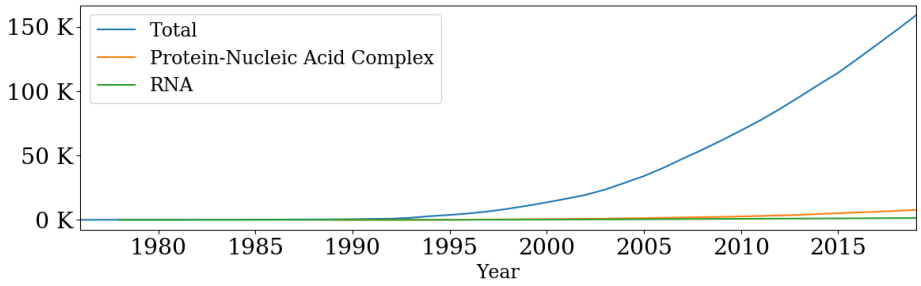
1.3.2. Machine learning

Machine learning is a data-driven method that utilizes computers to learn from experience. Machine learning has become more and more popular in recent years, with applications in computer vision, natural language processing (NLP), robotics, e.t.c. It has seen great success in commercial products like Siri and Alexa, self-driving cars, and the famous game AI AlphaGO[58] that, together with its successors [59], have beaten all human experts in the board game Go.

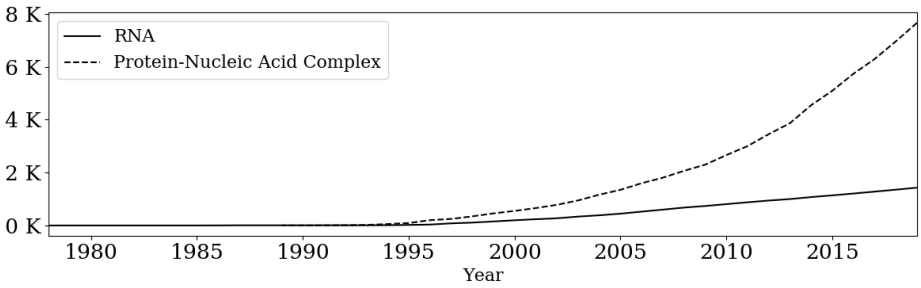
The growth of data available in biological research also made it suitable for the development of ML models. glsPDB [60], the website that holds all known tertiary structures of biomolecules has observed an exponential growth in the number of data entries since its announcement in 1971 [61] (Figure 1.3). Numerous ML models have been developed for proteins, protein-containing complexes, and biomolecules in general. For example, ML-based models utilized in protein research, such as for prediction of protein structures [62, 63], identification of protein targeting drugs [64–67] and many others, [68] have achieved comparable or even superior accuracies comparable to traditional rule-based models. Natural language processing (NLP) models have been developed to classify documents and extract information from biomedical literature [69–71]. ML models have

Though the slow progress of RNA structure determination in experiments (green line in

Figure 1.3a) compared to proteins has long been a limiting factor for the development of data-based methods of RNAs, there has been a significant improvement since the 1990s. There are now over 6000 structures for RNA-protein complexes and over 1000 RNA-alone structures in the Protein Data Bank (PDB) (Figure 1.3b), in combination resulting in millions of atomic coordinates available. This enormous, rich databank is an excellent resource that provides a lot of valuable information about local RNA structures and structural motifs. A number of ML-based models have been trained to study RNA-protein interactions and structural properties of RNAs [15, 72], to predict RNA secondary and tertiary structures from sequence [73, 74], and reversely, to predict RNA sequences from a tertiary structure in inverse RNA folding problems [75].



(a)



(b)

Figure 1.3: Growth of number of structure entries in the RCSB Protein Data Bank (RCSB PDB, http://www.rcsb.org) [60] of (a) all structures (including proteins) and (b) RNA and nucleic acid-protein complex. Data was taken from PDB website [76]

$$X = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \dots & x_{1n} \\ x_{21} & x_{22} & x_{23} & \dots & x_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{d1} & x_{d2} & x_{d3} & \dots & x_{dn} \end{bmatrix} \qquad y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \tag{1.4}$$

In general, ML is used to deal with problems consisting of a large set of multi-dimensional data samples, by building a "black-box" type of model to predict an observable without knowing the underlying relations. There are typically two types of machine learning problems, namely supervised learning and unsupervised learning. In supervised learning, a model is trained to map a set of features, X, to known targets, y, (Equation 1.4), where y can be either continuous or categorical values. When y is continuous, the problem is a "regression" problem, and the model is a regression model. For example, the prediction of the house pricing, for example, is a regression problem. When y is categorical, on the other hand, the problem (model) is named as "classification" problem (model). For example, predicting handwritten digits is a classification problem. Sometimes classification models can output probabilities for each category. In unsupervised learning, no targets y are known, and the problem is focused on clustering input data into groups or analyzing the underlying patterns within the features X.

The major aspects which affect the success rate of an ML model are data collection, feature engineering, and model selection. The training data is a set of data used to develop machine learning models, and it directly affects what the model is learning and how well it could optimally learn. Feature engineering is the step that requires the most domain knowledge as one has to handcraft features that are relevant to the problem, given the targets y and the available information. Feature engineering can either be done manually before the training or using some statistical tools like Principle Component Analysis (PCA). In some neural net-

work models, the feature engineering step can be bypassed, and the feature embeddings can be automatically learned in the training step. Model selection and hyperparameter tuning are all about choosing the right statistical model that fits the data. If a relationship to be learned is simple, a simple linear model would be sufficient and would likely work better than a complex model such as deep neural networks. In the instance that features are sparse or when feature dimensionality exceeds the number of training samples, regularized models and sparse learning can be used. The choice of models and hyperparameters largely depend on the nature of models. It is common practice to test a variety of models and parameters, and choose the one with the highest accuracy or lowest error on a separate validation dataset that does not overlap with the training data based on trial-and-error.

1.4. Outline

The work presented in this thesis aims to provide frameworks for predicting RNA structural ensembles as well as binding sites of ions and small-molecule ligands. First, a framework of constructing RNA structural ensemble based on SASA and Bayesian/maximum entropy (BME) is discussed in Chapter II and is applied to the SAM-I riboswitch. In Chapter III, a fingerprinting method capable of characterizing RNA local atomic environment is proposed and used to identify native-like RNA structures and bound-like RNA-ligand complexes. The fingerprint was used as features to predict Mg^{2+}ion binding sites and ligand binding poses. Finally, in Chapter VI, I developed a classifier to identify ligand-binding cavities in RNA structure and to extract bound-like conformations from an RNA conformational pool.

In Chapter II, I built a framework to construct RNA structural ensemble using atomic SASA values. To understand the functionally relevant structural ensembles of RNA, methods are needed that relate ensemble accessible measurements to the underlying conformational dis-

tributions comprising these ensembles. Several relatively rapid and ensemble-based experimental approaches, such as chemical probing methods, provide information about the accessibility of reactive groups to the solvent environment, mirroring the SASA of these groups. Relationships between the solvent accessibility of members of a computationally generated set of RNA conformations and the observed ensemble measurements of SASA may be useful in establishing a representative distribution of conformations to infer functionally relevant environmental responses, and to yield individual conformational members of such ensembles for the targeted design of therapeutics or the rationalization of a functionally relevant response. As such, I explored the development of an approach that yields an RNA ensemble that is consistent with the SASA-tied experimental observations. I developed a framework based on the Bayesian/maximum entropy (BME), which was then benchmarked on a set of RNAs and simulated ensemble-averaged SASA values. The benchmarking results suggested that using BME with ensemble-averaged SASA, it is possible to construct a structural ensemble with desired accuracy. Then I applied the framework to the SAM-I riboswitch, in which measured SASA derived from LASER experiments were applied, and I successfully captured the ensemble properties of SAM riboswitch ensemble in its free and bound states. In the free-state ensemble, I identified a possible ligand-binding pocket different from the experimental pocket. I anticipate that performing binding experiments on the new pocket in the future will further validate the results.

As mentioned above, feature engineering is an essential step in building machine learning models for RNA. In Chapter III, I developed a structure-based fingerprint using a simple Gaussian function of atomic pairwise distances. Time complexity analysis of the fingerprint shows that it is sufficient for small to medium-sized RNA molecules. I demonstrated the capabilities of the fingerprint in combination with machine learning methods to discover patterns in RNA structural information and develop predictive models to help predict the

structure and structure-related properties of RNAs. In particular, using the fingerprints as features, I trained ML models to predict Mg^{2+}binding sites in RNA structures and ligand binding poses (in Appendix) in RNA-ligand complexes. Improvements on these fronts, coupled with advancements in computer software and hardware, will enable the structure and dynamics of newly discovered RNAs to be rapidly characterized [77].

Inspired by the binding pocket identification form Chapter II, I realized there is a need for the identification of binding cavities in a structured RNA and developed a structural-based framework for the identification of binding sites in Chapter IV. In an RNA structure, especially larger ones, there are often a number of cavities that potentially bind to small-molecule ligands. To facilitate the virtual screening on those RNAs, the detailed mechanism and an efficient tool to identify the most probable binding pockets are needed. In Chapter IV, I focused on predicting drug-like cavities in an RNA structure, that is, without pre-knowledge of the ligand itself, I want to identify which binding pocket in an RNA structure is most probably bound to a small ligand. The model I built was able to recover druggable cavities among all cavities identified by cavity mapping methods within the top 3 positions in most cases. Furthermore, I also built a pipeline that could predict druggable conformations and identify the druggable cavities using the sequence information of an RNA. Combined with ligand and pose selection and cavity mapping methods, such a framework will drive faster and more accurate virtual screening.

Chapter 2.

Structural Ensembles of Ribonucleic Acids From Solvent Accessibility Data

This chapter was based on the unpublished manuscript https://www.biorxiv.org/content/10.1101/2020.05.21.108498v1.

In this chapter, I used simulated data to quantify the extent to which the RNA structural ensembles can be inferred from local solvent accessibility. To this end, I first constructed pairs of decoy and target ensembles (i.e., simulated ensembles) for a set of benchmark, single-stranded RNAs. Second, I reweighted the decoy ensembles using the ensemble-averaged solvent accessible surface area (SASA) data calculated from the target ensembles. Third, I quantified the extent to which the conformational distributions in the target ensembles resembled the target ensembles. In general, I found that for a specific set of ensemble-averaged SASA data, I could reweight the decoy ensembles such that they came "closer" to the target ensembles. However, I found that the ability to infer atomic ensembles from SASA data was sensitive to their errors, limiting the overall "restraining" power of SASA data. Based on this assessment of the scope, I constructed a pair of atomistic ensembles

of the S-adenosylmethionine (SAM)-responsive riboswitch using experimental SASA data derived from light-activated structural examination of RNA (LASER) reactivities measured in the absence and presence of SAM. The differences between the ensembles are consistent with a reshaping of the free-energy landscape of the riboswitch in the presence of SAM, and the results agree with the atomistic picture that emerged from ensembles previously generated using orthogonal approaches. Interestingly, within the ligand-free ensemble, I identified a conformer that potentially harbors a hidden binding pocket. Broadly, the results should pave the way for the direct utilization of experimentally-derived solvent accessibility data to construct atomistic ensembles of RNA.

2.1. Introduction

Changes in conformational equilibria – in response to changes in physiochemical conditions within the cell – underlie the biological function of many ribonucleic acids (RNAs). Such changes may include changes in temperature, pH, or the absence/presence of binding partner(s). This ability of RNA to respond to changes in local cellular conditions is best exemplified by riboswitches, which are cis-acting regulatory RNA elements located in the 5'-untranslated (UTR) region of mRNAs that change their conformations upon binding to specific ligands [78–80]. This conformational change either sequesters or releases sequence-motifs that, in turn, activate or deactivate transcription or translation. The aptamer domains of riboswitches bind to their cognate ligands with high specificity and confer the RNA with its sensing capabilities. As such, understanding the structure of the aptamer domain is critical to understanding relationships between the conformational equilibria of aptamers and the sensing capabilities of riboswitches. Mounting evidence suggests that the aptamer domain of riboswitches exhibits varying degrees of structural plasticity. Therefore, character-

izing the structural ensemble, comprised of the set of conformers that are accessible to a riboswitch under a specific set of conditions, is a critical step in describing and then rationalizing its response to cognate and non-cognate ligands. Analysis of such ensembles can reveal the existence of alternate binding pockets that may facilitate binding of non-cognate ligands, as well as novel allosteric sites that may facilitate ligand binding away from the site bound by cognate ligands.

Solution techniques can be used to probe the equilibrium conformations of RNAs by providing access to structure-dependent, ensemble-averaged measurements that can, in principle, be used to infer structural ensembles that capture the conformational distribution of an RNA under a specific set of conditions. For instance, chemical probing experiments can be used to identify reactive sites in RNA, both *in vitro* and *in vivo* [81]. Because the sites that are most "reactive" tend to be solvent-exposed, the reactivities obtained from these experiments provide an indirect "read-out" of the local solvent accessibility across the ensemble of structures populated by the RNA. The ensemble-averaged reactivities derived from light-activated structural examination of RNA (LASER) experiments, in particular, have been shown to correlate strongly with solvent accessible surface area (SASA) of the C8 atom in purine residues [23], suggesting that they might be useful for constructing such ensembles. Using SASA data derived from highly accessible measurements is advantageous as it provides a fast and efficient route to access structural ensembles that can be used to infer functionally relevant environmental responses in RNA, and to yield individual conformers for the structure-guided design of therapeutics.

To infer structural ensembles from ensemble-averaged experimental data like SASA, one of two strategies can be employed: restraining or reweighting [82]. Restraining involves carrying out molecular dynamics simulations with a force field augmented with restraint terms that ensure that the simulated ensemble-averaged observables match the experimen-

tal measurements [83–85]. Reweighting, on the other hand, is a post-processing approach that involves assigning weights to conformers within an ensemble, so that the ensemble-averaged observables computed using these weights match the measured observables [86–88]. As a post-processing approach, reweighting has the advantage of being computationally efficient and could be used to generate multiple ensembles from multiple sets of experiment data (possibly measured under differing conditions) from the same set of structures without requiring additional simulations.

In this chapter, I proposed a computational approach based on conformational sampling and Bayesian Maximum-Entropy (BME) inference to generate structural ensembles of RNA consistent with observed ensemble-averaged SASA. Clues to the potential of experimental data based on SASA to construct structural ensembles were recently provided by Madl and coworkers, who carried out solution NMR experiments in which they used solvent paramagnetic relaxation enhancements (sPRE) induced by the soluble, paramagnetic compound Gd(DTPA-BMA) to probe the structure of two benchmark RNAs [89]. They found that the inclusion of sPRE data during structural refinement significantly enhanced the quality of the resulting NMR ensemble [89]. Building off of this observation, I carried out a large-scale and systematic study to explore the robustness and error tolerance of SASA to construct structural ensembles. I then applied the approach to SAM riboswitches and successfully constructed the structural ensemble of the riboswitches both in the absence and presence of SAM. Furthermore, by a closer investigation into the highest weighted conformational members in the ensembles, I identified conformers of the SAM riboswitch that possess alternate binding pockets that are predicted to accommodate small-molecule ligands besides the cognate SAM ligand. The ability to construct such ensembles using SASA derived from highly accessible experimental data is advantageous, as it provides a fast and efficient route to access structural ensembles that can be used to infer functionally relevant environmen-

tal responses in RNA and to yield individual conformers for the structure-guided design of therapeutics or the rationalization of a functionally relevant response.

2.2. Methods

2.2.1. Generating decoy and target ensembles

I compiled a dataset of 45 RNA molecules, with length ranging from 14 to 53 residues (Table B.1). For each of the RNA, its sequence and native structure (crystal or NMR structure) were downloaded from Protein Data Bank [60], which were then used to generate a set of diverse conformations with FARFAR [44, 45] and KGSrna [46]. FARFAR (Fragment Assembly of RNA with Full Atom Refinement) is a Rosetta framework that predicts RNA tertiary structure from its sequence and, optionally, a secondary structure model. During this process, small RNA fragments are drawn from the crystallographic structure database, followed by Fragment Assembly of RNA (FARNA) to generate a set of low-resolution structures, then the low-resolution structures are refined by energy minimization with Rosetta full-atom energy function. KGSrna (Kino-geometric sampling for RNA), on the other hand, is a conformational sampling tool that utilizes experimental structure to model the dynamical ensemble of an RNA. KGSrna generates a set of structures by geometric perturbations to the experimental structure. The perturbations are carefully designed to preserve hydrogen bonding, and structures with clashes are excluded. I carried out KGSrna sampling using an RMSD radius of = 10 Å. Here I used FARFAR and KGSrna to generate 1000 structures respectively, resulting in a total of 2000 structures as the initial decoy. By combining structures generated by FARFAR and KGSrna, a large simulated decoy ensemble is formed that contains diverse yet realistic conformations.

To avoid bias over native and near-native structures inherent to conformational sampling tools, a subset of 200 structures were sampled from the initial decoy with a uniform distribution of RMSD with respect to the native structure. This subset was then used in all subsequent steps such as choosing target ensemble and reweighting. After subsampling I was left with an approximately equal number of native and non-native structures, such that I was not taking the majority of the dataset when the target ensemble was chosen at a certain RMSD radius to the native structure. After subsampling, target ensembles were sampled near the native structure at various RMSD radius, ranging from 2.0 to 11.0 Å.

2.2.2. Reweighting ensembles using SASA data

Once the decoy and target ensembles are generated, I seek to find a set of weights $\{w_i\}$ for each structure i in the decoy ensemble, that the reweighted ensemble best matches the target ensemble (Figure 2.1). To achieve this, I use C8-SASA to characterize the structure and apply Bayesian/Maximum-entropy Reweighting (BME) to obtain the most probable weights which make the reweighted C8-SASA consistent with target C8-SASA. I first computed atomwise SASA for each structure in both decoy and target ensemble. The calculation was done in the open-source tool FreeSASA [90] using Lee & Richards' algorithm [13] at 1000 slices per atom. Then the computed SASA was filtered by matching atom name with *C*8 to obtain C8-SASA. The target C8-SASA was assumed to follow a Gaussian distribution, with its mean and standard deviation computed from the target ensemble. Finally, BME [91, 92] was performed to obtain the most probable weights such that the reweighted average C8-SASA match target C8-SASA within error tolerance. I formulate this problem as a constrained optimization problem with the entropy of weights being the objective function, which is then solvable using the Lagrangian method. Here I use w_i to denote weights of each state in decoy ensemble,

$SASA_{i,k}$ the SASA of atom k in the state i and $SASA_k$ being the ensemble-averaged SASA of atom k in target ensemble. The problem can be formulated as follows:

$$\text{maximize} \left(- \sum_i w_i \log w_i \right) \tag{2.1}$$

subject to

$$| \sum_i w_i SASA_{i,k} - SASA_k | \leq \epsilon_k \tag{2.2}$$

where $SASA_{i,k}$ is the SASA of atom k in conformer i, $SASA_k$ is the ensemble-averaged SASA of atom k in the target ensemble, and ϵ_k is the error tolerance sampled from a Gaussian distribution $p(\epsilon_k)$:

$$p(\epsilon_k) = \exp \left(- \frac{\epsilon_k^2}{2\theta\sigma_k} \right). \tag{2.3}$$

where σ_k is the standard deviation calculated from the target SASA, and θ is a factor that scales σ_k. For each RNA in the benchmark data set, I utilized the BME approach to reweight the decoy ensemble using C8-SASA data from each of the 10 distinct target ensembles, which differ in terms of the RMSD cutoff to the native structure. Furthermore, for each decoy and target ensemble pair, I ran 500 trials with random initialization of ϵ_k. The reweighted weights were averaged over all trials to get stable and reliable results, which minimizes the effect of randomness introduced by the sampling of ϵ_k. Therefore, in total, I carried out 225000 ($45 \times 10 \times 500$) reweighting experiments; in each case, I used $\theta = 1.0$ when carrying out the reweighting. (Eq. 2.3).

2.2.3. Comparing ensembles

To compare the target, reweighted, and decoy ensembles, I calculated their atomic density maps using GROmaρs [94], a GROMACS-based density map analysis tool. Then, to quantify

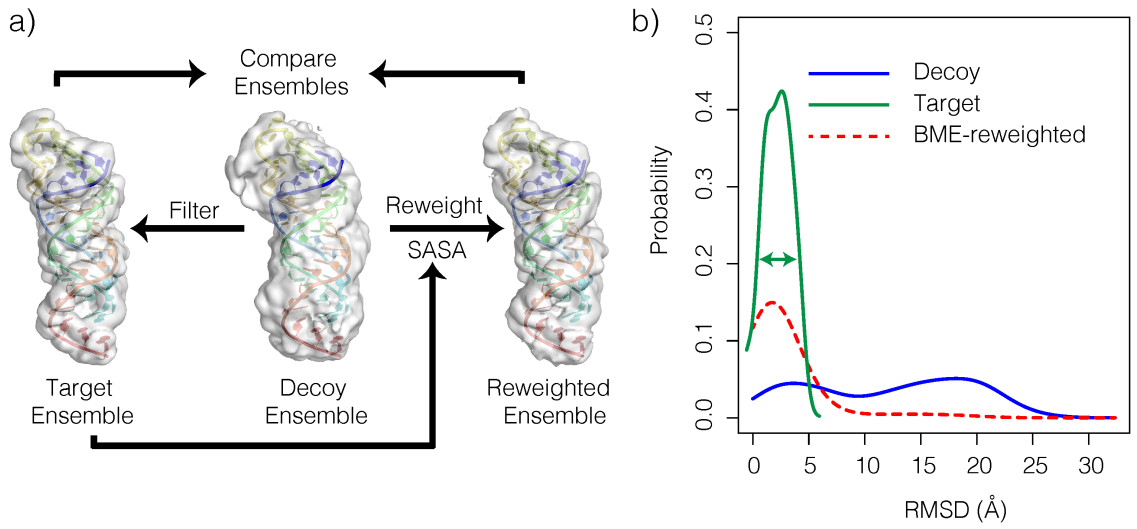


Figure 2.1: (a) Illustration of the workflow used to examine the ability of SASA data to recover representative structures. From a decoy ensemble, a target ensemble is constructed by filtering structures based on RMSD from the reference structure. SASA data calculated from the target ensemble is used to reweight the decoy ensemble, and then the reweighted ensemble is compared to the target ensemble. (b) Density plots comparing the RMSD distribution of decoy, target, and the BME-reweighted ensemble. These distributions correspond to those for the triple helix RNA (PDB ID: 2M8K [93]) with target ensemble comprised of structures within 3 Å of the native structure.

the extent to which decoy ensembles could be reweighted towards the target, I calculated κ (Equation 2.4), which I defined as the ratio between two cross-correlation values, CC_{RT} and CC_{DT}. The cross-correlation CC is the global correlation between density maps, as also implemented in GROmaρs [94]. CC_{RT} is defined as the cross-correlation between density maps of Reweighted(R) and Target (T) ensembles, while CC_{DT} is the cross-correlation between density maps of Decoy (D) and Target (T) ensembles.

$$\kappa = \frac{CC_{RT}}{CC_{DT}} = \frac{\text{reweighted to target ensemble correlation}}{\text{decoy to target ensemble correlation}} \tag{2.4}$$

Values of $\kappa > 1$ correspond to instances in which the atomic density maps of the reweighted ensemble more closely resembled the target ensemble than the decoy ensemble (i.e., $CC_{RT} >$

CC_{DT}). To visually compare the difference, I also computed the difference density maps between target and reweighted ensembles and between target and decoy ensembles. The difference maps were then rendered as volume maps using Pymol [6].

2.2.4. Constructing conformations of ligand-free state of the SAM riboswitch

To construct a conformational ensemble of the ligand-free (-SAM) state of the SAM riboswitch, I first generated a conformational pool including both free and bound states SAM riboswitch. A total of 32000 conformers were generated using KGSrna (as discussed in Section 2.2.1), with experimental structure for ligand-free (PDB ID: 3IQN, 3IQP [95]) and the ligand-bound (PDB ID: 2GIS [96], 3IQR [95]) states as reference. Next, from each conformer in the conformational pool, C8-SASA of all purine residues were computed using FreeSASA.[90] Then I used LASER-derived C8-SASA as targets to reweight the conformational pool. To estimate C8-SASA from LASER reactivity data, I fitted LASER reactivity data [23] to freeSASA-computed C8-SASAs for +SAM crystal structure (PDB ID: 2GIS) to obtain a linear function that maps LASER reactivity to C8-SASA [96]. The fit was then used to estimate C8-SASA in the ligand-free (-SAM) state from available -SAM LASER reactivity data.[23] LASER-estimated C8-SASA was then used as the mean of target distribution to reweight the conformational ensemble using the same BME reweighting technique that I employed for the previous computational experiments (see Section 2.2.2). BME was carried out with an error tolerance scale $\theta = 1$, and the standard deviation of target distribution was set to 2 $Å^2$.

2.2.5. Cavity mapping and docking experiments

For each of the four highest weighted conformers in the -SAM ensemble, I carried out cavity mapping to identify sites on the surface of the conformer that might facilitate interactions with small-molecule ligands. To achieve this, I utilized the two-sphere cavity mapping method implemented in the rbcavity program within the rDock modeling suite [17]. For cavity mapping, I set the maximum number of cavities to 10 and the minimum cavity volume to 50 $Å^2$. The resulting cavities were visualized in PyMOL (version 2.3.4) [6]. Next, I carried out a small-scale *in silico* screening by docking 500 small, drug-like molecules into the cavities identified using rbcavity. These 500 small molecules corresponded to a small library of drug-like compounds obtained from the ZINC library [97]. To identify the most conformationally selective compounds in the library, I first computed the selectivity index, $\gamma_{i,j}$, defined as

$$\gamma_{i,j} = \frac{\Delta G_{i,j}}{\langle \Delta G_{i,j} \rangle} \tag{2.5}$$

$$\gamma_i = \max_j \ \gamma_{i,j} \tag{2.6}$$

Here i runs over the compounds in the library, j runs over the conformers (docking receptors), $\Delta G_{i,j}$ is the docking score for compound i docked onto conformer j, and $\langle \Delta G_{i,j} \rangle$ is the average docking score for compound i across all j conformers. For each compound, I assign a selectivity (γ_i) as the maximum of the set of selectivity indices $\{\gamma_{i,j}\}$. As defined, compounds with high γ_i correspond to those that have a docking score ($\Delta G_{i,j}$) on a given conformer that is significantly more favorable (negative) than the average docking score ($\langle \Delta G_{i,j} \rangle$); such a compound was identified as being a "conformationally selective" compound.

2.3. Results

2.3.1. Reconstructing ensembles using SASA data.

I carried out a set of computational tests to quantify the degree to which SASAs could infer atomic ensembles of RNA (Figure 2.1). Specifically, SASA of C8 atoms in purine residues were used, as those atoms correspond to the sites probed in LASER experiments. LASER experiments were known to produce the reactivities profiles that have been shown to exhibit a strong, positive correlation with solvent accessibility [23]. To gauge the utility of C8-SASA in reconstructing atomistic ensembles of RNA, I generated pairs of decoy and target ensembles for a set of 45 RNAs, computed ensemble-averaged C8-SASAs over the target ensemble, and used the set of ensemble-averaged SASA data to reweight the decoy ensemble. To assess the performance of this SASA reweighting scheme, I calculated atomic density maps for target, decoy, and reweighted ensembles and calculated the cross-correlations between each pair. For each RNA in the benchmark set, I carried out simulations in which the width of the RMSD distribution within the target ensemble ranged between 2 and 11 Å.

Shown in Figure 2.2c are distributions of κ (Equation 2.4), the ratio of the cross-correlation between the atomic density maps of the reweighted ensemble (CC_{RT}) and the decoy ensemble (CC_{DT}) relative to the target ensemble, which I defined earlier in Section 2.2.3. $\kappa > 1$ corresponds to the case where density maps of the reweighted ensembles exhibited a higher correlation with the target than did the decoys ensembles. The median value of κ was > 1 for all the widths of target ensemble (Figure 2.2a). However, at target widths $\gtrsim$5 Å, for a small fraction of the benchmark set, κ was < 1. Note that for the examples in the benchmark set which exhibited $\kappa \leq 1$, the decoy ensemble is already very similar to the target ensemble, with relatively high cross-correlations ($CC > 0.90$). It explains why κ for these reweighted ensembles, especially in the noisy case, found it difficult to get over 1. Overall, these results

suggest that reweighting the decoy ensembles with SASA-derived conformational weights yielded atomic density maps that generally exhibit higher correlations to the target than did the corresponding decoy ensembles.

Next, I explored the sensitivity of the SASA reweighting scheme to both errors in the SASA data and the width of the target ensemble. To carry out this benchmarking, I first generated pairs of decoy and simulated target ensembles. Then, using ensemble-averaged C8-SASA from the target ensemble, I reweighted the decoy ensemble. Visual inspection of the difference atomic density maps of the reweighted ensembles relative to their target ensembles revealed that, in general, the C8-SASA reweighted ensembles exhibited smaller residual densities than the initial decoy ensembles (Figure 2.2c). This observation suggests that the reweighted ensembles tend to more closely resemble the target than did the decoy ensembles. Shown in Figure 2.2b are plots of κ (Equation 2.4) with respect to noise level in the target SASA and the width of the target ensembles. As defined, when the correlation between the reweighted and target ensemble is higher than that of the decoy and target ensembles, $\kappa > 1$. As might be expected, the lower the level of noise added to the target data and the narrower the width of the target ensemble, the higher the values of κ. Accordingly, κ is > 1 when the width of the target ensemble is $<= 4.2$ Å and the noise-level is $<= 1.3$ $Å^2$, which suggests that under such conditions, the C8-SASA data can be used to bring the decoy ensembles into better correspondence to the target ensembles.

2.3.2. SASA-based ensembles of the SAM riboswitch are consistent with reshaping the conformational pool in the presence of SAM

Next, using SASA data derived from LASER experiments, I constructed ensembles for the -SAM and +SAM states of the aptamer domain of the SAM riboswitch. To achieve this, I

first generated a diverse pool of conformers and computed the C8-SASA for purine residues, which correspond to the site in RNA probed by LASER experiments. Then I reweighted the pool of structures using the Bayesian maximum entropy (BME) reweighting method [92], using the C8-SASA predicted from LASER reactivity measured in -SAM and +SAM states as the target data, respectively [23]. Briefly, to generate the target data, I converted LASER reactivities to C8-SASA by fitting LASER reactivities of the +SAM to C8-SASA computed for the crystal structure of the +SAM. The mean error in the fit was 1.27 $Å^2$ (Figure B.1 in the Supporting Information). Based on the benchmarking results presented above (Figure 2.2c), I expect that ensembles reweighted using C8-SASA data with an inherent error ~1.27 $Å^2$ should be closer to the "true" ensembles than the initial conformational pool. Shown in Figure 2.3 is the comparison between the average structure in LASER/SASA-derived ensembles for the -SAM (Figure 2.3a) and +SAM (Figure 2.3b) states. The RMSD of the average ensemble structures were only 1.30 Å , consistent with X-ray crystallography result that the -SAM and +SAM structures were almost identical (RMSD = 0.52 Å). Despite the similarity of the average -SAM and +SAM structures, I did observe some subtle differences in the -SAM and +SAM ensembles along the distributions of the distance between residue 47 and residue 90, which I used as the reaction coordinate to describe the openness of P1 relative to P3 (Figure 2.3c). Comparison of the -SAM and +SAM distributions revealed that the mode of distribution shifts from 16.8 to 13.0 Å when going from the -SAM to the +SAM state, consistent with the -SAM state having a preference, relative to the +SAM state, for the open P1-P3 state.

Despite the global structure of the -SAM and +SAM being almost identical, an inspection of the crystal structures reveal that in the -SAM, A46 and U57 are base-paired (closed), whereas in the +SAM state, they are not base-paired (open); U57 instead forms a hydrogen bond with SAM in the +SAM state. To test whether the ensembles captured this subtle difference, I computed the distribution of the distance between A46-U57, which I used as a

reaction coordinate to describe the extent to which A46-U57 were base-paired (Figure 2.3d). In contrast to the distribution of the P1-P3 distance (Figure 2.3c), I observed a more dramatic difference between the -SAM and +SAM (Figure 2.3d). The distribution of the A46-U57 distance in the LASER/SASA ensembles are consistent with -SAM sampling both the closed and open A46/U57 states, whereas +SAM predominantly existing in the open A46/U57 state (Figure 2.3d). The LASER/SASA-derived ensembles, therefore, support a mechanism in which SAM shifts the population of states toward the closed A46/U57 state (Figure 2.3d).

2.3.3. The -SAM ensemble contains a conformer that is predicted to bind to small-molecules via a hidden pocket.

Since their discovery, riboswitches have garnered interest as under-explored drug targets [99]. Indeed, the recent discovery of the antibacterial small molecule, ribocil-B, which targets the flavin mononucleotide (FMN) riboswitch, supports the notion that riboswitches are druggable RNA targets [100]. As a result of this and related discoveries [101], the identification and design of small molecules that target riboswitches has become an active area of research. Because individual riboswitches have evolved to bind to a specific ligand, attempts to design compounds that target riboswitches have focused on identifying compounds that are analogs of the cognate ligand or compounds that can recapitulate its interaction pattern. Alternatively, one could envision identifying small molecules that bind to a riboswitch at a site other than that occupied by the cognate ligand. Once identified, these alternative sites can be targeted using structure-based methods like molecular docking [102].

To illustrate how one might attempt to detect such site computationally, I applied ensemble docking to the highest weighted conformers in the -SAM ensemble. First, I applied the two-sphere cavity mapping method to the four highest weighted conformers in the -SAM

ensemble; these four conformers have a cumulative weight of >0.70 (Figure 2.3e-h). Across the four conformers, I observed vast variations in both the location and size of the cavities. Next, I docked a set of 500 small, drug-like molecules (chosen from the ZINC library [97]) onto each of the four conformers. Because we were particularly interested in identifying dockable binding pockets, into which small molecules can fit, the focus of the analysis was on the short-range van der Waals (VDW) contribution to the binding free energy. Shown in Figure 2.4a are the distributions of the non-polar contribution to the binding free energy (ΔG) estimated using the rDock function. The mean ΔG across all four conformers was -24.0 kcal/mol, indicating that in general, the small molecules in the library were capable of forming favorable interaction with -SAM conformers 1-4. Overall, however, conformer 2 exhibited lower ΔG values than the other conformers (Figure 2.4a). Finally, of the screened molecules, I computed γ_i, the selective index (Equation 2.6), to identify those that exhibit conformational selectivity across the four conformers. Interestingly, the six most selective compounds in the library all exhibited a preference for conformer 2 (Table 2.1). Moreover, in this conformation, all six compounds are predicted to bind to the SAM riboswitch at the same binding site, which is located away from the binding pocket occupied by SAM in the +SAM crystal structure. Intriguingly, the pocket occupied by these molecules is near a set of conserved residues that participate in a base-triple that, though far away from the SAM binding site, have been shown to exhibit the strongest SAM-dependent stabilization [95]. In conformer 2, however, the base-triple is absent (Figure 2.4b), which results in the formation of the binding cavity that these compounds occupy. The pocket that these molecules occupy, therefore, represents a "hidden" pocket that is absent from the -SAM and +SAM crystal structures, and which only emerged after conformational sampling. Within the pocket, the six selective compounds are predicted to form stacking interactions with a pair of conserved residues, A62, and C65 (Figure 2.4b). Collectively, these results suggest that the simulated

-SAM ensemble contains conformers that harbor dockable pockets other than the pocket occupied by the cognate ligand, SAM.

Compound	ΔG_1 (kcal/mol)	ΔG_2 (kcal/mol)	ΔG_3 (kcal/mol)	ΔG_4 (kcal/mol)	$\gamma_i \times \min(\{\Delta G\})$ (kcal/mol)
1	-18.5	**-26.6**	-16.1	-14.9	-37.2
2	-10.2	**-24.0**	-15.3	-13.5	-36.4
3	-17.0	**-26.3**	-15.7	-17.3	-36.3
4	-17.3	**-26.1**	-16.6	-15.6	-35.9
5	-22.8	**-26.6**	-18.6	-10.7	-35.9
6	-15.7	**-27.5**	-20.2	-21.7	-35.4

Table 2.1: Docking scores of conformationally selective binders. For each, listed are the predictd binding free energy with conformer 1 (ΔG_1), 2 (ΔG_2), 3 (ΔG_3), and 4 (ΔG_4). Also listed for each compound is $\gamma_i \times \min(\{\Delta G\})$, the product of selectivity index, and the lowest docking score across the four conformers. Here the the binding free energy correspond to the non-polar (Van der Waals) contribution estimated using the rDock scoring function.

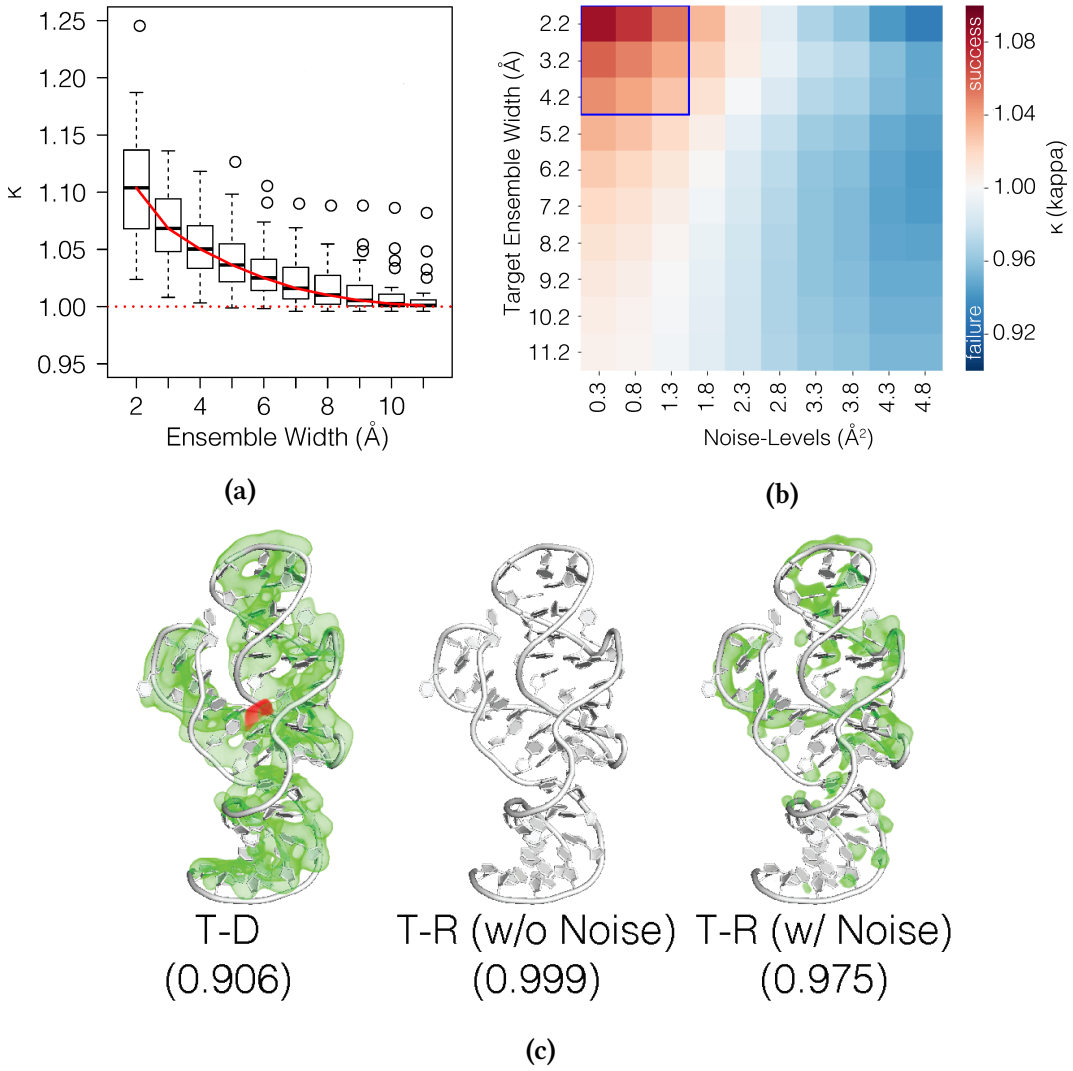


Figure 2.2: (a) Plots of κ versus the width of the target ensemble. (b) The heatmap of κ as a function of the width of the target ensembles and the noise-level in the corresponding target C8-SASA data. Ensemble width is defined as the maximum RMSD between a structure in the target ensemble and the native structure. The noise-level in target C8-SASA is simulated by adding random noise to the target ensemble-averaged C8-SASA. The absolute value of the noise was sampled from an exponential distribution with noise level as the scale parameter (or average noise). The map shown here is the average of κ values over all the benchmark data set. Note that for some RNA ensembles, the BME algorithm failed to converge. Accordingly, the averaging is performed on successful reweighting only. (c) Example of difference-atomic maps for the CR4/5 domain of medaka telomerase RNA (PDBID: 2MHI) [98].

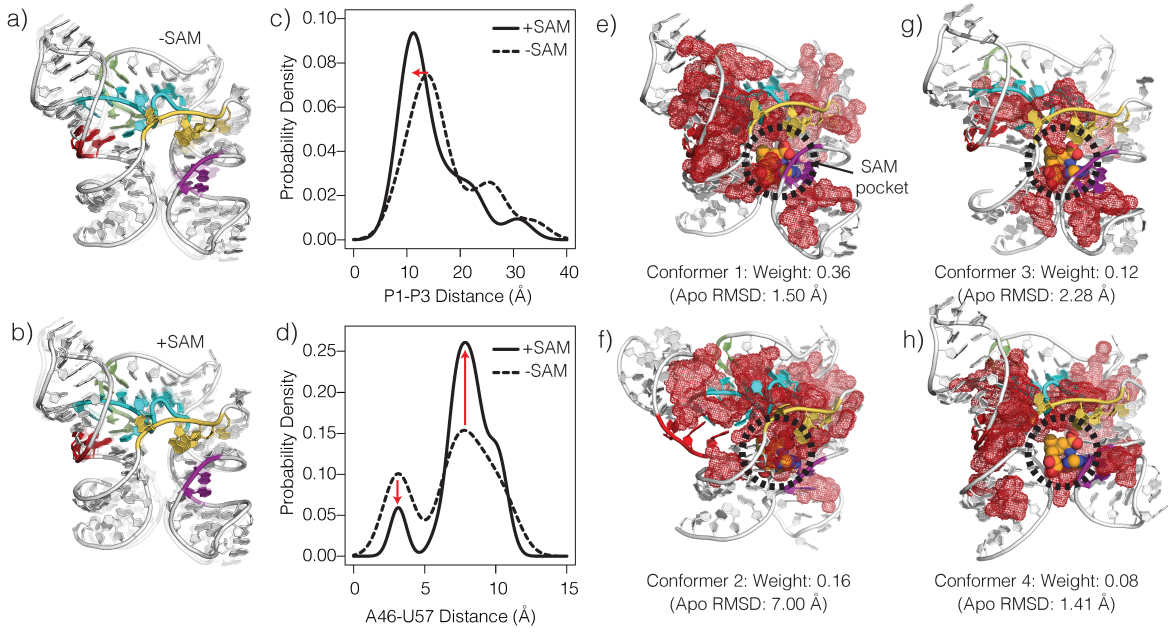


Figure 2.3: The LASER/SASA-derived SAM ensemble. Shown are the average structures in the LASER/SASA-derived ensemble of the -SAM (a) and +SAM (b) states of the SAM riboswitch. Shown in (c) is the distribution of the distance between P1 and P3 helices for both the -SAM and +SAM states. Similarly, shown in (d) are the distribution of the distances between residues A46 and U57, which are paired in the -SAM state (A46/U57 closed) and unpaired (A46/U57 open) in the +SAM state. (e-h) The four highest weighted conformers in the -SAM ensemble. For reference, the SAM is overlaid onto the images. The red mesh highlights the cavities identified in each conformer.

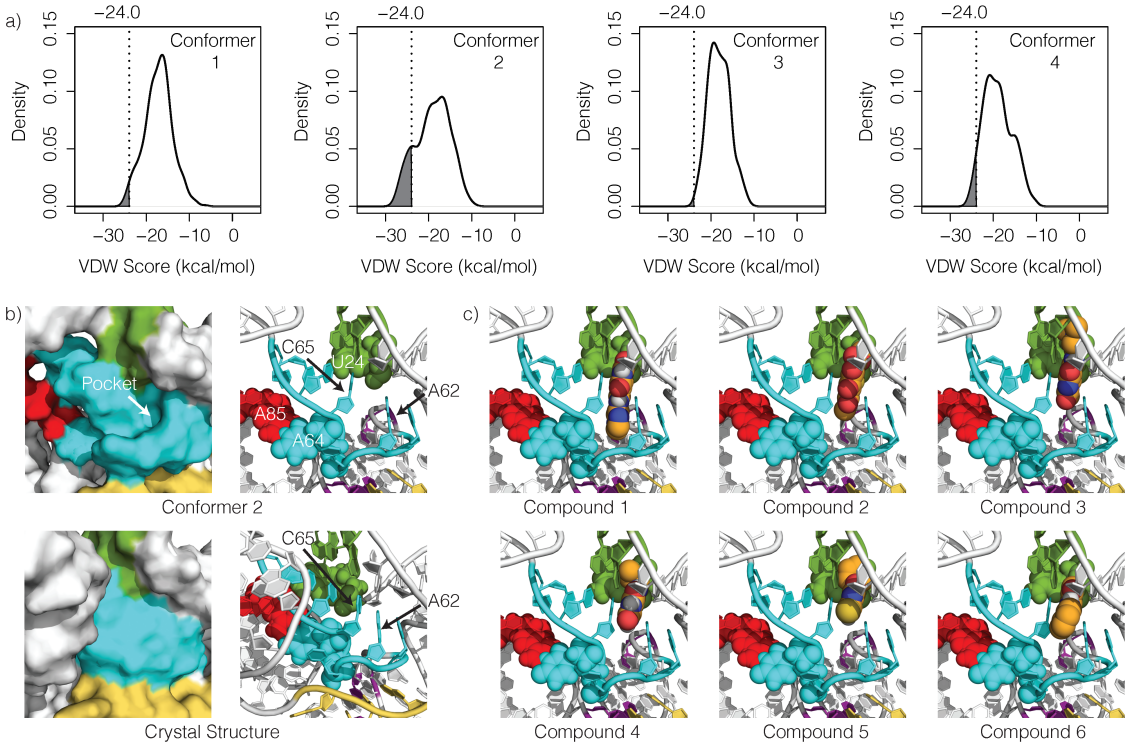


Figure 2.4: Ensemble docking. Distribution of docking scores across the conformer 1-4 in the -SAM ensemble. (b) A comparison between the binding site of the six most selective compounds in conformer 2 (top) and the corresponding site in +SAM crystal structure (bottom). The binding site is a hidden pocket, present in conformer 2 but absent in the +SAM crystal structure (bottom). Notably, the pocket features increased nucleobases A62-C65 distance and the absence of the nearby U24-A64-A85 base-triple. (c) Poses of the six most selective small molecules docked onto conformer 2. All six compounds form stacking interactions with C65 and A62.

2.4. Discussion

In this study, I examined the utility of conformational sampling and SASA Bayesian maximum entropy (BME) reweighting in constructing RNA structural ensembles and applied the framework to infer atomistic ensembles for the -SAM and +SAM states of the SAM riboswitch. I carried out tests using simulated ensembles to precisely quantify the inherent potential of SASA data for inferring ensembles of RNA. I concluded that the typical experimental error presented in chemical probing experimental methods like LASER is within the error tolerance of the reweighting framework. Results on SAM riboswitch suggests that the ensembles generated using SASA-BME framework were consistent with that -SAM state sampling a wider range of conformations relative to the +SAM state (Figure 2.3c-d), which was suggested by the existing biochemical and biophysical data [95]. Interestingly, by docking a small library of compounds against the four highest weighted conformers in the -SAM ensemble and identifying selective binders in the library, I was able to locate what appears to be a "hidden" binding pocket in the -SAM riboswitch. Because the residues that line this pocket are highly conserved, it represents an ideal pocket for small molecule targeting. Future work will center around executing a more exhaustive search for compounds that target this hidden site.

The use of the SASA to infer the atomistic ensemble of the SAM-responsive riboswitch was predicated on the assumption that SASA inherently contains conformational restraining power. Recently, Madl and coworkers carried out solution NMR experiments in which they used solvent paramagnetic relaxation enhancements (sPRE) induced by the soluble, paramagnetic compound Gd(DTPA-BMA) to probe the structure of two benchmark RNAs [89]. Like the reactivities derived from chemical probing, the sPRE data provided an indirect "read-out" of local solvent accessibility across the equilibrium ensemble. They found that the inclusion

of sPRE data during structural refinement significantly enhanced the quality of the resulting NMR ensemble [89]. Their findings, along with the results of this benchmarking study, strongly suggest that experimentally-derived SASA contains sufficient restraining power to infer structural ensembles of RNAs. Therefore, I envision that the SASA-based reweighting approach I utilize in this study will emerge as a robust yet straightforward strategy for using experimentally-derived SASA data to infer atomistic ensembles. Such ensembles can then be used to generate or test structure-function hypotheses and provide useful structural data to guide the discovery and design of RNA-targeting therapeutics.

Chapter 3.

Local Atomic Environment Characterization and Prediction of Magnesium Binding Sites in RNAs

In chapter II, I demonstrated how the chemical reactivity can be used as local atomic environment "fingerprint" to determine RNA structure ensembles. In this chapter, I will describe a set of numerical "fingerprints" that are capable of characterizing RNA global and local 3D structures using known atomic coordinates, and as descriptors for prediction of magnesium ion binding sites in RNA. The latter part in this chapter was done in close collaboration with Lichirui Zhang, a former visiting undergraduate student and now a Ph.D. candidate in Chemistry at Columbia University.

3.1. Introduction

3.1.1. The importance of Mg^{2+}ions in RNA

Metal ions are critical in stabilizing RNA structures and mediate its dynamics. Not only are the positive charges of metal ions necessary to compensate for the negative charges of the highly acidic phosphate backbone of RNA, but the presence of metal ions is critical to ensure proper folding of RNAs [37, 38]. Furthermore, metal ions could also mediate catalytic processes in some ribozymes [36, 103]. Magnesium ions (Mg^{2+}), in particular, are divalent with a small radius (0.72 Å) and bind tightly to RNA structures, and are found to stabilize the tertiary structures in many experimentally determined RNA structures. However, the determination of Mg^{2+}-binding sites in RNA structures remains a challenge.

3.1.2. Locating Mg^{2+}-binding sites

The preferred binding sites of metal ions in RNA can be determined experimentally by high-resolution X-ray crystallography. However, since Mg^{2+}, Na^{+} and H_2O all have 10 electrons each, many bound Mg^{2+}can be easily mistaken for Na^{+} or water molecules in the electron density profile, which X-ray crystallography relies on to determine structures. Also, the high ligand exchange rate of the metal ions commonly associated with nucleic acids [104] makes it difficult to investigate these ions in solution, resulting in the absence of Mg^{2+}in solution-state NMR structures. The difficulty of experimentally determining the location of Mg^{2+}ions motivates the development of methods to predict their positions based on the structure of an RNA.

Several computational models have been developed to gain insights into the RNA-ion interactions, which in turn provide information for binding site prediction [105]. For exam-

ple, the classical counterion condensation theory has been established to describe the short-range RNA-ion interactions, and Poisson Boltzmann's theory has been adapted to compute the long-range interactions. Based on these theories, models integrating MD, Monte Carlo sampling, energy minimization docking [106–115] and other statistical models [116, 117] have been developed as faster alternatives to treat RNA-ion interactions as well as to predict the binding sites of ions [118]. However, the source code for most of the methods is not publicly available, while methods that provide access are often have limited functionality. For example, the webserver for MCTBI simply cannot take any structures larger than 1MB (approximately 140-nts). To fill this gap, I implemented a machine learning-based that predicts Mg^{2+}-binding sites and made it freely available to the academic community via https://smaltr.org.

3.1.3. RNA 3D structure characterization

The local RNA structure determines where Mg^{2+}ions will bind. Mg^{2+}ions, or ions in general, bind to RNA molecules by interacting with RNA atoms through electrostatic interaction. The strength of these interactions depends on the RNA local structure at the binding site, e.g., the configuration of atoms in the surrounding and the atomic distances between an RNA atom and the binding site. A recent study also showed that the local chemical environment at each site in an RNA molecule could be used to identify and classify Mg^{2+}-binding sites [119].

Measurements based on atomic distances have been a common way to characterize RNA 3D structures. In this chapter, I present a fingerprinting tool that uniformly characterizes the local environment of selected atoms based on pairwise atomic distances. Then I carried out a set of clustering analyses to assess the structural sensitivity of the atomic fingerprint. I then

used the fingerprint as features to train machine learning models to predict Mg^{2+}-binding sites in RNA structures. An additional study that utilizes the molecular version of this fingerprint to obtain the prioritize ligand poses in RNA-ligand complexes was also carried out in a collaborative work with my former labmate, Dr. Sahil Chhabra, and is included in the Appendix.

3.2. Methods

3.2.1. Mathematical formulation of the fingerprinting methods

The structural environment of a reference atom is defined by the positions and properties of all neighboring atoms inside some sphere of radius R_c. Several factors need to be taken into account to describe this environment, including the type of atoms, the number of atoms, and distance to the reference atom. To construct a numerical descriptor of these properties, I developed a fingerprint (Figure 3.1) inspired by the directionally resolved atomic fingerprint by Botu and Ramprasad [120–122]. For each atom i in the system of interest, its atomic fingerprint is represented by the vector $\left[V_i(\eta, \nu), \nu \in \mathscr{V}\right]$, where $\mathscr{V}$ represents the set of atom types in the system of interest. The atomic fingerprint captures the atomic environment of atom i, described by the equation below:

$$V_i(\eta, \nu) = \sum_{\substack{j \neq i, \\ j \in \nu}} e^{-(r_{ij}/\eta)^2} \cdot f_d(r_{ij}) \tag{3.1}$$

Here, r_{ij} is the distance between atom i and j. η is a Gaussian parameter and has the unit of length, which can be tuned to put emphasis on nearby (smaller η) or distant (larger η) environment. ν indexes "atom types" (known as *atom name* in the nomenclature of chemistry) of

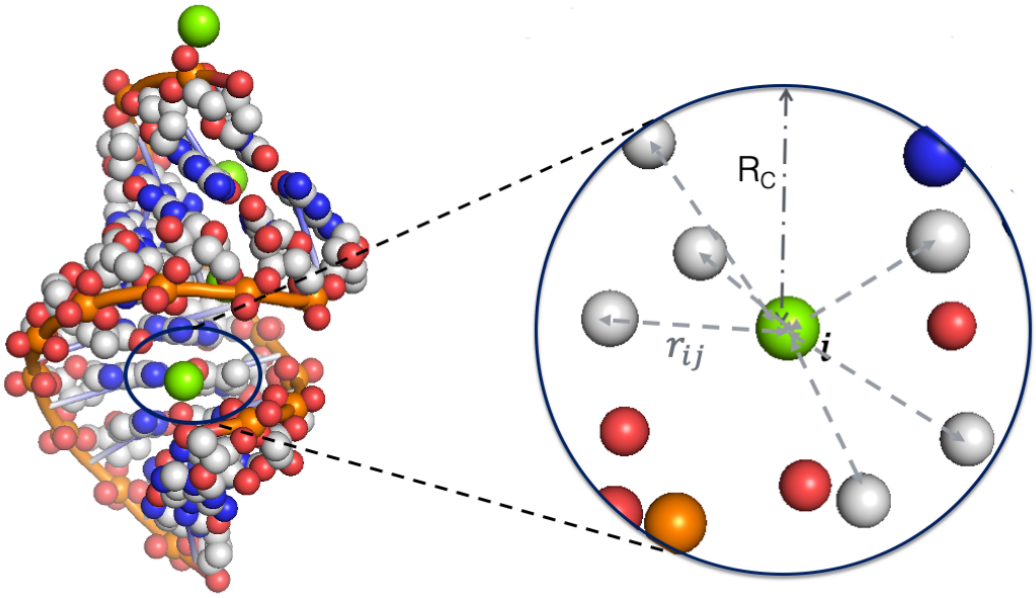


Figure 3.1: Illustration of the atomic fingerprint for a reference atom (colored in green) and one type of its neighbors (colored in gray). To generate the atomic fingerprint, a summation over all atoms of the same type based on atomic distances (indicated by dashed arrows) within cutoff distance R_C (indicated by the solid circle) is considered. The figure is rendered using the sphere representation of an RNA Dodecamer (PDB ID: 1DNO) [123]. Inspired by reference [124].

the neighboring atom j , for example, C1, N4, O6'. The atomic fingerprint is therefore a vector of length n, in which n is the number of *atom names* in the system of interest. One could also construct a longer fingerprint as a concatenation of multiple fingerprints corresponding to various η values, to include both nearby and distant environment. The summation in Equation 3.3 goes over all atom j in the neighborhood of atom i if the *atom name* of j is ν. f_d is a damping function that gradually approaches zero when r_{ij} increased to R_c and is expressed as:

$$f_d(r_{ij}) = \begin{cases} 0.5\left[\cos\left(\frac{\pi r_{ij}}{R_c}\right) + 1\right], & \text{if } r_{ij} < R_c \\ 0, & \text{otherwise} \end{cases} \tag{3.2}$$

The atomic fingerprint described in Equation 3.1 can be viewed as a sum of Gaussians multiplied by cutoff functions. The Gaussian term describes a spherical shell around the reference atom, with η defining the extent of the shell. As the atomic distance, r_{ij}, is always a positive number, this term is inversely proportional to the atomic distance r_{ij}. When r_{ij} is equal to zero, though not physically possible due to the inherent radius of atoms, this Gaussian term has its maximum of 1. The cutoff function, as described in Equation 3.2, decays smoothly with increased atomic distances. By including this term, the atomic fingerprint diminishes to zero at cutoff R_c. Finally, the sum over j includes the contribution of each neighbor with the same *atom name*. The atomic fingerprint resembles the radial symmetry function [125] proposed earlier, and was shown to be invariant to basic atomic transformation operations of translation, rotation, and permutation [124].

Vectorial Fingerprint

The atomic fingerprint in its vectorial form can be used to model atomic properties that have directional dependence (e.g., atomic forces). The vectorial fingerprint for the reference atom

is described as a 2D array of 3 rows and n columns. Each row in the fingerprint captures information of atomic environments of atom i in one of the three dimensions ($u \in \{x, y, z\}$). A row vector in direction u is given by the equation below:

$$V_i^u(\eta, \nu) = \sum_{\substack{j \neq i, \\ j \in \nu}} \frac{r_{ij}^u}{r_{ij}} \cdot e^{-(r_{ij}/\eta)^2} \cdot f_d(r_{ij}) \tag{3.3}$$

r_{ij}^u is the projection of r_{ij} in u direction. The first term in the product $\frac{r_{ij}^u}{r_{ij}}$ shows directional dependence of the fingerprint and is the only term that differs in different directions. The second and third terms are the same Gaussian functions and cutoff functions as in Equation 3.1.

This vectorial fingerprint is an extension of the basic atomic fingerprint (Equation 3.1) by including an additional directional dependence term. Each term of the basic atomic fingerprint could also be viewed as the root squared sum of the vectorial fingerprint in all directions:

$$V_{ij}(\eta, \nu) = \sqrt{V_{ij}^x(\eta, \nu)^2 + V_{ij}^y(\eta, \nu)^2 + V_{ij}^z(\eta, \nu)^2}. \tag{3.4}$$

With the additional directional dependency, the vectorial fingerprint, compared to the atomic fingerprint, is no longer invariant to rotational transformation. However, the vectorial fingerprint is advantageous as structural descriptors of atoms in prediction tasks that involve directional dependence. For example, the vectorial fingerprints have been used as features in a machine learning model, which predicts the atomic forces of simple substances to quantum accuracy [122].

Molecular Fingerprint

To characterize the global interaction between a reference molecule and other atom groups or molecules in the surrounding (e.g., solvents, ions, small-molecule ligands), a molecule-level descriptor is more desirable as "molecular fingerprint". Note that the term "molecular fingerprint" is used here to describe the surrounding environment of the reference molecule, particularly its interactions with other molecules, which is different from the cheminformatics definition as the "structure encoder" of a molecule.

One way to construct the molecular fingerprints is to simply aggregate the atomic fingerprints of all atoms in the reference molecule. However, the set of neighboring atoms has to be defined differently from the atomic fingerprints. When computing atomic fingerprints for a single reference atom, any atoms (within the cutoff distance R_c) are considered neighboring atoms and are included in the calculation of the fingerprint. While when computing atomic fingerprints to form molecular fingerprints, neighboring atoms should only be considered if they belong to atom groups or molecules different from the reference atom.

Below a type of molecular fingerprint, called "pose fingerprint", is described. The purpose of the pose fingerprint is to describe the orientation ("pose") of a small-molecule ligand to the RNA receptor when they bind to each other. The pose fingerprint was used as features to predict the preferred location and orientation of the small-molecule ligand to the RNA receptor to form stable RNA-ligand complex and is described in greater detail in Appendix A: RNAPosers [126].

In the pose fingerprint, the reference molecule is the small-molecule ligand. It is worth noting that the RNA molecule could also be used as reference, which results in the same fingerprint, but since the RNA molecule is typically two orders of magnitude larger than the ligand, it is more efficient to use ligand as reference. For a given ligand pose p, its fingerprint

vector F_p is the sum of atomic fingerprints with the same *atom-pair type s*. An *atom-pair* must contain one RNA atom and one ligand atom, and the *atom-pair type* is given by the atom types of both atoms. Unlike the *atom names* that describe the type of atoms in RNA, the atom types in small-molecule ligands are typically described by *SYBYL atom type*, which specifies the element and atomic hybridization (e.g., sp3 carbon). The *SYBYL atom type* of the ligand atom and the *atom name* of the RNA atom collectively define the atom-pair type *s*, and the set of all possible permutations of ligand atom types and RNA *atom names* are denoted as *S*. The pose fingerprint for a given pose p is a vector, and each element in the vector corresponds to an atom-pair type s given by

$$F_p(\eta, s) = \sum_{(i,j)\in s} e^{-(r_{ij}/\eta)^2} \cdot f_d(r_{ij}) \qquad \forall\, s \in S \tag{3.5}$$

For example, for a set of 20 ligand SYBYL atom types and 85 RNA *atom names*, the pose fingerprint of each pose $F_p = \{F_{p,s}, s \in S\}$ contains 1700 elements (20 SYBYL types × 85 RNA atom types). The η parameter and f_d function are the same Gaussian width parameter and cutoff function as in Equation 3.1 and 3.2. The pose fingerprint, together with a set of classifiers trained to predict the preferred orientation of ligand in a complex with an RNA molecule, is publicly available to the academic community via `https://github.com/atfrank/RNAPosers`.

3.2.2. Assessing the distinction power of fingerprints

One key factor in evaluating the quality of fingerprint is its ability to distinguish between different structures and groups of structures. To determine this distinction power of atomic fingerprint, I carried out three clustering analyses, aiming to cluster a set of RNA structures into different groups using their atomic fingerprints, such that the structures in the

same group share more similarity than to those in other groups. Below I describe the two algorithms used to form and visualize the clusters, the K-means clustering, and the t-SNE.

K-means clustering

The clustering analysis can be performed using unsupervised learning algorithms. K-means clustering is a widely used unsupervised learning method with a low cost and scales well with the number of samples. K-means cluster a set of data samples into k groups by minimizing the sum of squared distances between data points within each cluster and their corresponding cluster centroid. Given the data samples to be clustered {x}, the clusters $\{S : S_i\}$ and their cluster centroids μ_i, the objective function is as follows:

$$\text{minimize} \sum_{i=1}^{k} \sum_{x \in S_i} \|x - \mu_i\|^2 \tag{3.6}$$

Some of the major factors that affect the performance of K-means clustering include the number of clusters k, the initialization of cluster centroids, and the maximum iterations. As has been discussed in literature [127], the initialization of cluster centroids has an especially important influence on the performance. Here I chose the conformations that are furthest away from each other in terms of structural RMSD as initial cluster centroids. This method of cluster centroids initialization resembles the furthest point heuristic, or Maxmin heuristic, which generally leads to better clustering accuracy than random initialization. Once the initial cluster centroids were determined, the number of clusters and maximum iteration were determined by visual inspection of the formed clusters.

Visualizing clusters

To effectively visualize the clusters formed by high-dimensional data in 2D or 3D space, dimensionality reduction has to be applied. t-SNE [128] is a computational tool suitable for this purpose. t-SNE is an embedding method that generates a low-dimensional embedding of the high-dimensional data based on the pairwise similarity between data samples. The similarity is obtained by Gaussian function from pairwise Euclidean distances, as shown in Equation 3.7. t-SNE then maps high-dimensional data to 2D (or 3D) by minimizing the divergence between the distribution of pairwise similarity in high-dimensional space and its low-dimensional counterpart.

$$p_{j|i} = \frac{exp(-\|x_i - x_j\|/2\sigma^2)}{\sum_{k \neq i}(-\|x_i - x_k\|/2\sigma^2)} \tag{3.7}$$

3.2.3. Mg^{2+}-binding site predictor

I converted the prediction of Mg^{2+}-binding sites into a binary classification problem, which can be solved with machine learning techniques. I started by identifying a portion of the 3D space near the RNA surface, which are likely Mg^{2+}-binding sites, and discretizing the space into cartesian grids. Then, pseudo-atoms were placed in each grid, and the atomic fingerprint of each pseudo-atom was computed. Finally, a classification model was trained to predict whether a grid is a Mg^{2+}-binding site based on the pseudo-atom fingerprint. Next, I will describe the details in this workflow, including the dataset used for training, the featurization method, and the training and validation process.

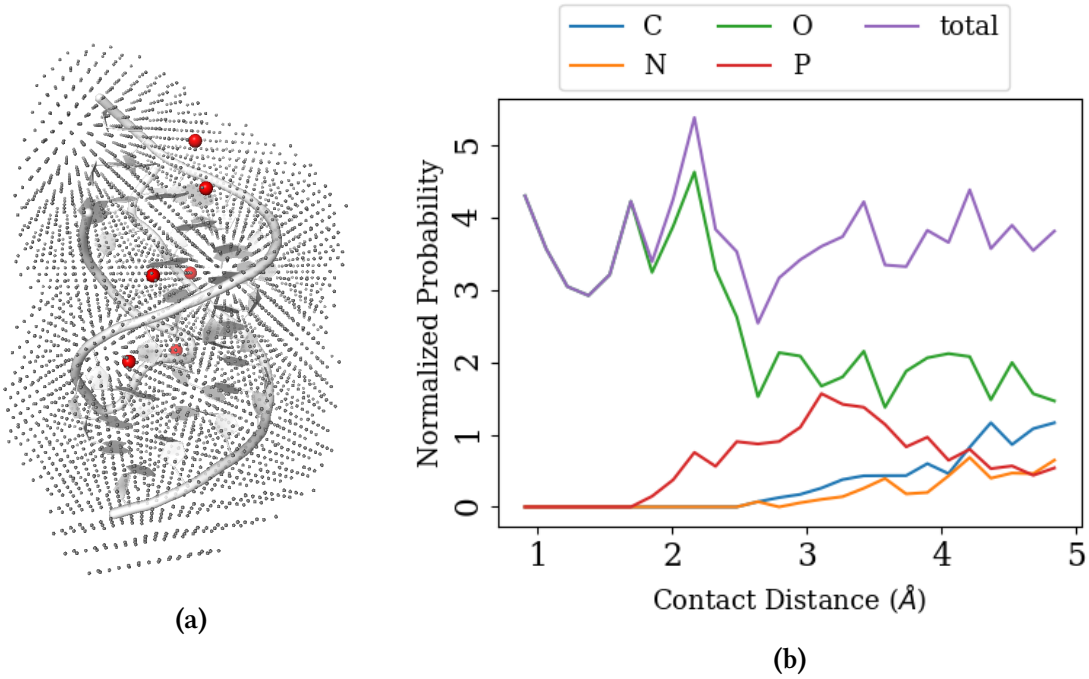
Dataset

There are over 400 X-ray structures of RNA so far in the PDB databank, which contain experimentally solved Mg^{2+}ions. However, not all of the Mg^{2+}positions in those X-ray structures were reliable. Several databases have been established to summarize the reliable Mg^{2+}-binding sites in experimental structures of RNA or nucleic acids, including MeRNA [129], MINAS [130], MetalionRNA [117] and MgRNA [131]. Here, I chose MgRNA [131] as our benchmarking dataset, as MgRNA is the most up-to-date and complete database which provides organized data with convenient access. I compiled a benchmarking dataset (Table B.2) containing a set of 156 RNAs with Mg^{2+}ions included in MgRNA [131]. This non-redundant list contains RNAs with various types of well-resolved Mg^{2+}-binding sites. Some of the PDB IDs were obsolete and were replaced with their successors. The entire dataset was used to train and validate the predictive model using a leave-one-out cross-validation approach.

Featurization

I placed pseudo-atoms in a grid-based manner around an RNA (Figure 3.2a). The lower and upper bound of the distances between a pseudo-atom and its nearest heavy atom in the RNA is set to 1.5 Å and 8 Å, which contains all the Mg^{2+}ions in the benchmark dataset. I do not consider binding pockets within 1.5 Å as it is too close to the RNA surface, given that the van der Waals radii of most heavy atoms were already larger than 1.5 Å[132]. I also excluded any sites that were further than 8 Å way from any RNA atoms because the interactions are minimal at that distance. The radial distribution of the distance between Mg^{2+}-binding sites and its nearest heavy atoms in the RNA is shown in Figure 3.2b. The heavy atoms that have the highest density in the nearby region of $Mg^{2+}$are oxygen atoms. This is expected because the highly positive electrostatic field of Mg^{2+}has a strong interaction with oxygen atoms in

water molecule and phosphate groups [37].



(a)

(b)

Figure 3.2: (a) Illustration of pseudo-atom placements in RNA. The grey spheres indicates a pseudo site for Mg^{2+}ion, and the actual Mg^{2+}were shown in red. The pseudo sites were placed in 3D-grids with separation 1.5 Å. (b) Distribution of atoms as a function of distances of $Mg^{2+}$to its nearest heavy atoms in RNA, normalized by $4\pi r^2 dr$ [133].

To further eliminate unlikely binding sites form the process of placing grids, a few more distance restrictions were placed. It has been shown that the preferred distances of the Mg^{2+}-binding site to its nearest heavy atoms are dependent on heavy-atom types, and the interaction frequencies between RNA and $Mg^{2+}$is almost exclusively determined by oxygen and nitrogens [131]. Therefore, I set a separate set of limitations on the cutoffs used for the distances of a pseudo-atom to the heavy atoms, as summarized in Table 3.1. A pseudo-atom will only be placed in grids where all the conditions were satisfied.

Table 3.1: Distance cutoffs used for Mg^{2+}-binding sites

	Mg^{2+}-N[1]	Mg^{2+}-OP1	Mg^{2+}-OP2	Mg^{2+}-others[2]
lower bound (Å)	1.70	1.60	1.50	1.65
upper bound (Å)	7.00	6.20	5.25	4.40

[1]N include all types of nitrogen atoms commonly seen in RNA: N1 N2 N3 N4 N6 N7 N9.

[2]Others include O O2 O4 O6 O2' O3' O4' O5'.

Training Algorithm

A random forest classifier was used to model whether a pseudo-atom could be a Mg^{2+}-binding site. Because the true binding sites only occupy a small number of the pseudo-atom grids, as shown in Figure 3.2a, the data set is highly imbalanced with the number of negative samples significantly exceeded the positive samples (the class ratio is around $30 : 1$). Therefore, when training the random forest classifier, class weight was set to "balanced" to assign a higher weight to each true sample. I found out that the model trained is less prone to hyper-parameters, and I chose to build a random forest model with 100 trees, and maximum depth for each tree is set to 5.

Leave-one-out cross-validation

I carried out a leave-one-out cross-validation on the benchmarking dataset. The leave-one-out approach works as follows:

1. First, I chose one Mg^{2+}-containing RNA structure as the validation structure and removed it from the benchmarking dataset.

2. Then I computed the sequence similarity of the picked structure to each of the rest RNAs and removed any RNA structures that have sequence similarity $> 80\%$.

3. Finally, a classification model was trained on the rest of the structures, and the model was assessed on the validation structure.

Due to the imbalance in the training data, I used the AUC to assess the model's performance. AUC is the area under the ROC curve, which is a plot of the true-positive rate versus the false-positive rate by varying the classification threshold (the prediction score threshold above which is considered a true binding site). A perfect model that completely separates the true and false samples by a classification threshold will have an AUC of 1. A model that makes random predictions will produce an AUC of approximately 0.5. The higher the AUC, the better the model can distinguish true from false samples.

3.3. Results

3.3.1. Time complexity of the atomic fingerprint scales squarely with number of atoms ($O(N^2)$)

Due to pairwise atomic distance calculations, the time complexity for computing atomic fingerprints of all atoms in a molecule is scaled as $O(n^2)$ (see Figure 3.3 for a plot of runtime in seconds versus the number of atoms in the molecule). For an RNA molecule with 29 nucleotides (about 940 atoms), it takes 0.41 seconds to compute its vectorial fingerprint and 0.17 seconds to compute the scalar fingerprint. This process could be accelerated, of course, by parallelizing the distance matrix calculation to multiple processors. However, as the main purpose of this tool is to analyze a single structure or short trajectories of small RNAs, the current speed is satisfactory.

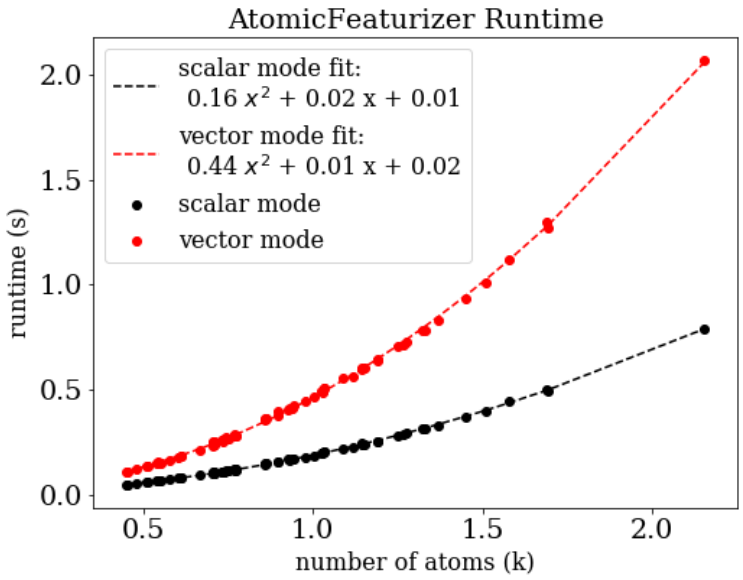


Figure 3.3: Atomic fingerprint and scalar fingerprint runtime illustration. The runtime was benchmarked on a dataset of 45 small RNAs, with length ranging from 14 to 53 nts. A trajectory of at least 1000 frames were generated by CHARMM for each RNA and used for this analysis. Runtime was based on average runtime per frame over the entire trajectory, using one core on an Intel(R) Xeon(R) CPU E5-2670 0 @ 2.60GHz.

3.3.2. RNA structures can be differentiated using atomic fingerprints

Free and bound states of miR-20b

To illustrate that the atomic fingerprints were sensitive to structural variations in RNA, I first used it to cluster structures of the 23-nt RNA, microRNA-20b (miR-20b) [134]. I created a 40-membered structural ensemble of miR-20b composed of structures from the 20-membered NMR bundle of the free state of miR-20b (PDB ID: 2N7X) and the structures from the 20-membered NMR bundle of the bound state of miR-20b (PDB ID: 2N82). In the presence of protein Rbfox RRM, miR-20b undergoes a structural change of 4.33 Å in RMSD that involves the disruption of several base-pairs in the apical loop region of miR-20b pre-element. Thus the structures of the free and bound forms of miR-20b are structurally distinct.

Shown in Figure 3.4 are the clustering results I obtained using atomic fingerprints of atoms in the RNA molecule. For each structure in the pool, atomic fingerprints were calculated for each atom and summed over all neighboring atom types ν, resulting in a single number for each atom. η was set to 2 Å. Fingerprints of all atoms in a structure were then aggregated into one vector, which was used as features for the clustering analysis. From t-SNE visualization (Figure 3.4a), there is a well defined linear boundary in the center of the graph between the two clusters, indicating that free and bound states are quite distinct in the atomic fingerprint space. Figure 3.4b is a visualization of all structures in the ensemble in PyMOL. By clustering the structures into two groups, I was able to group all bound-state structures into one cluster (blue), and the free-state structures into another cluster (red). Instead of using all-atom fingerprints, the correct clustering could also be achieved using a minimum number of two atomic fingerprints with the highest variance (*O2P* and *H5"* on residue 32 (the yellow-colored residue on the upper left corner of each structure in Figure 3.4b)). This simple analysis done for the 40-structure ensemble of miR-20b shows that using the atomic fingerprints as

clustering features and K-means clustering algorithm, I was able to identify the structural differences between the free and bound states RNA.

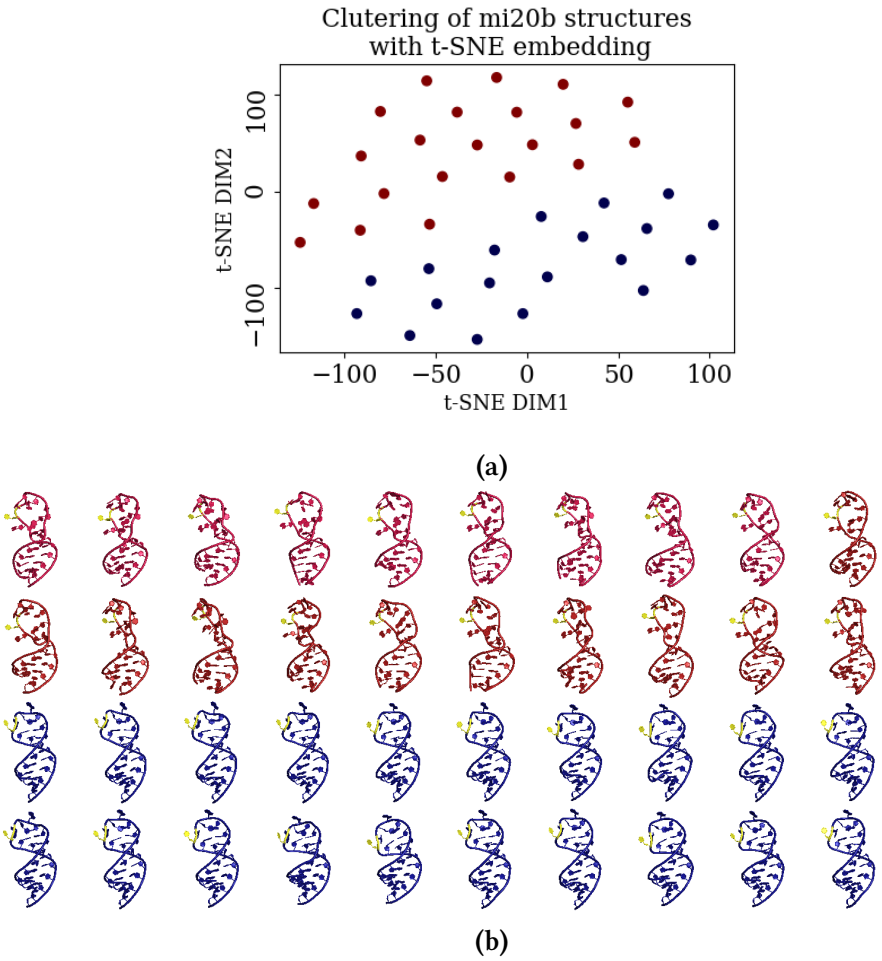
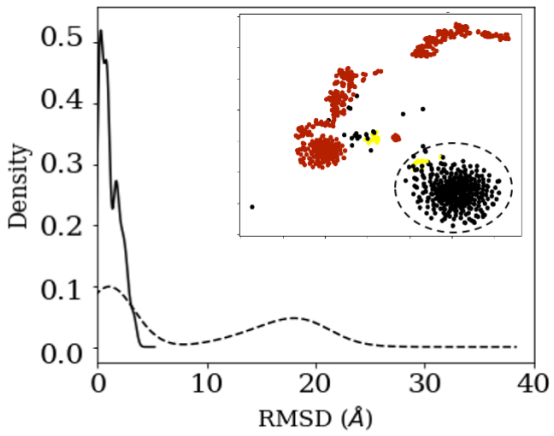


Figure 3.4: Results obtained by clustering the free-state and bound-state structures of miR-20b using their atomic fingerprints as features in (a) feature t-SNE space and (b) 3D space. The data points (structures) are colored differently based on their cluster IDs. (b) The first 20 structures correspond to bound-state structures and the latter 20 correspond to free-state structures.

Native and non-native structures of RNA pseudoknot

Next, I move on to study a more complicated system, a 927-membered pool of a 48-nt pseudoknot RNA (PDB ID: 2M8K). Compared with the miR-20b system, this pool exhibits higher complexity in conformational space and is a more stringent test of the atomic fingerprints in its distinction power. Some of the representative structures in the ensemble are shown in Figure 3.5a. The structures colored in black correspond to native structures, while structures colored in red and yellow correspond to non-native structures. This analysis aims to assess the ability of the atomic fingerprint to group them into a single native cluster. A clustering analysis similar to that applied to the miR-20b was carried out; the results are shown in Figure 3.5b. Three clusters were identified, and the native structure is assigned to the cluster colored in black. The distribution of structural RMSD of all structures in the ensemble relative to the reference native structure is shown in the dashed line. As can be seen, the structures exhibit high diversity in the pool with RMSD up to 40.0 ÅİIn comparison, the distribution of structural RMSD in the native cluster after clustering is shown as a solid line. All the clustered structures are within 5.0 Å in RMSD from the cluster center structure. This cluster also corresponds to the native cluster, which includes the native structure as a member structure. Visualizing the atomic fingerprints (Figure 3.5b Inset) in 2D space also reveals that structures in the native cluster are closer to each other in the fingerprint space. The results suggest that using atomic fingerprints as features, native and near-native structures in this complex ensemble of pseudoknot RNA could be separated from the non-native structures.

(a)



(b)

Figure 3.5: (a) Representative structures of the RNA pseudoknot (PDB ID: 2M8K) ensemble colored based on the RMSD to the native structures. (b) RMSD distribution relative to the reference native structure for all structures in the ensemble (dashed line) and within the native cluster (solid line).

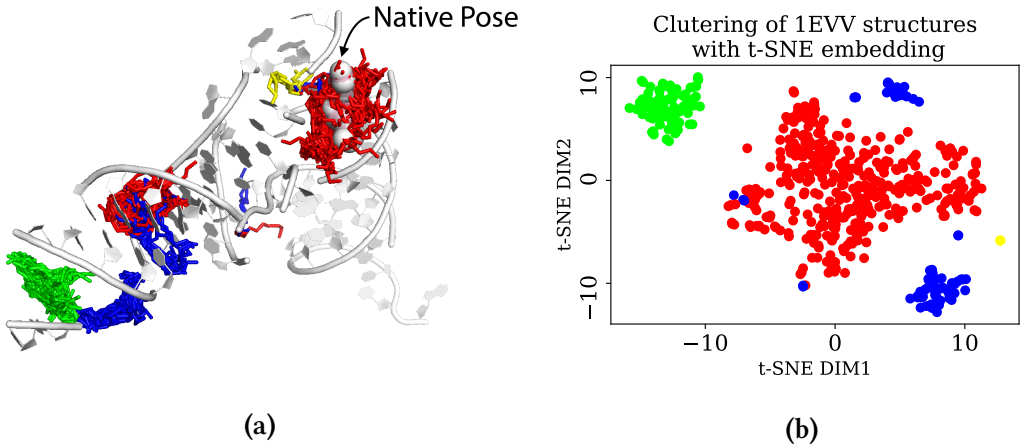
Native and non-native ligand poses in RNA-ligand systems

Finally, clustering analysis is carried out to extract native-like ligand poses from a set of 550 poses of an RNA-ligand complex using pose fingerprint. The data corresponded to a yeast phenylalanine tRNA in complex with spermine (PDB ID: 1EVV) and was taken from the dataset `https://doi.org/10.5281/zenodo.3711071` as part of the training set for RNAPosers [126]. Among all the 550 poses, only the ligand poses are varied (e.g., the ligand conformations or relative position to the RNA molecule), while the RNA structure is fixed.

The results for 4-cluster clustering analysis of the 550 poses are shown in Figure 3.6. The native cluster (cluster containing native poses) is colored red in Figure 3.6 a). Since the RNA structure is identical across all structures, the fingerprint only depends on (1) the ligand pose and (2) the RNA pocket, which ligand binds. As a result of the two factors, the native cluster captures the set of ligand poses that are either located in the native pocket (the RNA pocket the native ligand binds to) or have a high structural similarity to the native ligand pose. It suggests that the pose fingerprint could identify native-like poses or poses in the native pocket in RNA-ligand systems.

3.3.3. Classifiers based on atomic fingerprints accurately predict Mg^{2+} ions

Magnesium ions (Mg^{2+}) are critical for proper folding and functioning of Ribonucleic acid (RNA)s. However, there is a dearth of tools for identifying Mg^{2+}-binding sites in RNA. As such, I developed a classification tool that predicts Mg^{2+}-binding sites from an RNA's atomic coordinates. On average, the tool has a classification AUC of 0.86, and it out-performs a previous prediction tool in identifying Mg^{2+}-binding sites near kink turns in RNA structures.



(a) (b)

Figure 3.6: Clusters of ligand poses of Spermine in complex with yeast phenylalanine tRNA (PDB ID: 1EVV [135]) (a) in 3D space and (b) in feature t-SNE space. The native pose is shown in white spheres, and poses were rendered with colors corresponding to their cluster ID, with red being the native cluster (the cluster containing the native pose).

The classifier predicts Mg^{2+}ion with high accuracy

To evaluate the prediction performance of the Mg^{2+}-binding site classifier, I computed the area under the receiver operating characteristic curve (ROC), abbreviated as AUC, of each validation structure in the leave-one-out analysis. The AUC value describes the cumulative ability of this binding site prediction to recognize true positives and negatives while avoiding false positives and false negatives.

As can be seen from the AUC value for the 156 benchmarking RNAs (Figure 3.7 and Table B.2), the majority of individual structures demonstrate high AUC values, while a small minority show poor performance. The mean AUC is 0.86(±0.17), indicating that most of the Mg^{2+}-binding sites were ranked top among all potential sites in each structure. In particular, 2% of the dataset had AUC values below random (0.5), 75% > 0.83, 50% > 0.92, and 25% > 0.96. If AUC values above 0.85 are considered as a good indicator of recognition, 99 structures (61%) are above such a threshold.

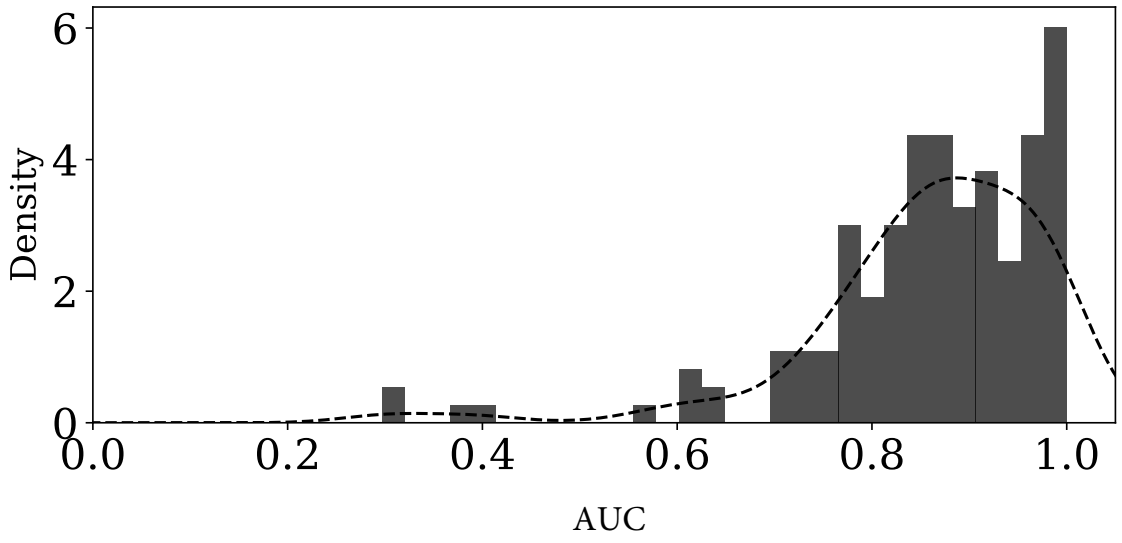


Figure 3.7: The summary results of the AUC on validation set in leave-one-out analysis.

The comparison to existing tools reveal the classifier better identifies site-binding Mg^{2+}ions in kink turns

Kink turns, often abbreviated as K-turns or kinks, are common structural motifs in RNA. K-turn is a type of junction of helical regions, where the RNA backbone has a "kink" that causes a sharp turn in the RNA helix (see Figure 3.8a for example of K-turns) [136]. K-turns are functionally important structural motifs because they are generally compact, tight regions of RNA structures that could serve as binding sites for other molecules [137], which interact with RNAs and alter RNA functions. The folding and stabilization of K-turns highly depend on the presence and binding of metal ions (Figure 3.8b), for example, one study on the prototypical K-turn Kt-7 found that the K-turn structure is only preserved in the presence of Mg^{2+}ions [138]. As such, the identification of Mg^{2+}binding sites in K-turns containing structures is essential for RNA structure prediction.

MetalionRNA [117] is another method for predicting metal ions binding sites in RNA

structures. It utilized a grid-based function in the polar coordinate system to create a statistical potential to describe the interaction between metal ions and RNA atom pairs. The statistical potential was parameterized using 113 RNA-metal ion systems with experimental structures of resolution < 2.0 Å and was tested with a 5-fold cross-validation approach. To compare the performance of the predictor trained in this work with that of MetalionRNA, I made a comparison of the predicted Mg^{2+}-binding sites in a fluoride riboswitch (PDB ID: 4ENC) and a group-I intron (PDB ID: 1HR2) using our leave-one-out results and the output given by MetalionRNA webserver.

Figure 3.8c summarizes the predictions for the Mg^{2+}ions present in the 4ENC and 1HR2 crystal structures. MetalionRNA results were obtained using all default parameter settings, with 5 and 26 Mg^{2+}-binding sites identified in 4ENC and 1HR2 based on their molecular size. When using our predictor, the same number of top-scored predicted binding sites were chosen to make a fair comparison. While MetalionRNA missed almost all Mg^{2+}-binding sites near the K-turns in the two RNAs, our predictor was able to identify most of them.

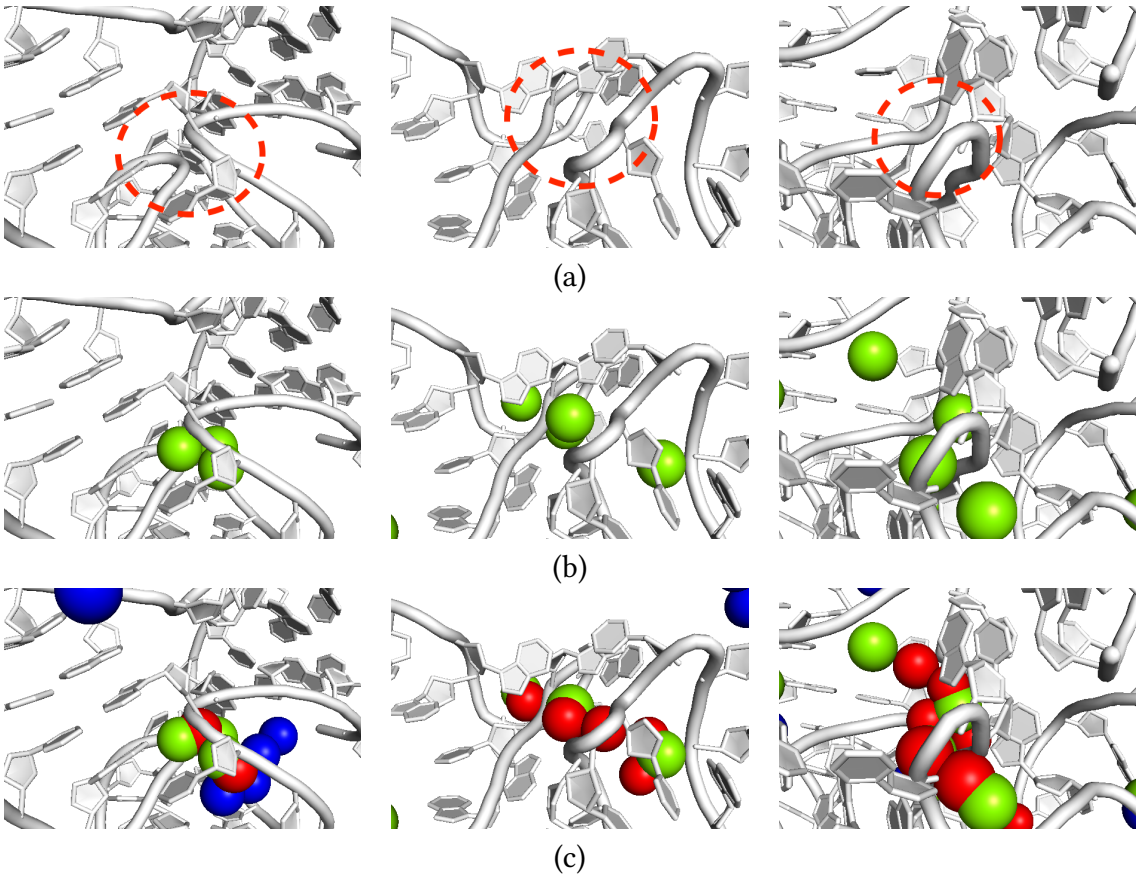


Figure 3.8: (a) K-turns in (left) fluoride riboswitch (PDB ID: 4ENC [139]) and (middle) chain B and (right) chain A of group-I intron (PDB ID: 1HR2 [140]). (b) Experimentally determined positions of Mg^{2+}cations in the K-turns indicated by green balls. (c) Top-scoring Mg^{2+}cations predicted by the classifier trained in this work (red) and MetalionRNA (blue).

3.4. Discussion

In this chapter, I described a computational structural fingerprint to represent the atomic environment around a reference atom. The fingerprinting method can be aggregated over all atoms in a molecule to describe the molecular environment. Clustering analyses to identify bound-like RNA conformations, native-like RNA structures as well as native-like poses of RNA-ligand systems were carried out. The clustering results suggest that the fingerprints

can resolve structural differences in conformations in RNA-only and RNA-ligand systems, especially in identifying native and native-like structures and poses. The atomic fingerprint runs in squared time, which may be a limiting factor in its utility. The next step in enhancing the atomic fingerprint will be to optimize its performance by accelerating the most time-consuming calculation of pairwise distances.

I next trained a machine learning classifier that takes atomic fingerprints as features to predict Mg^{2+}-binding sites in an RNA 3D structure. The leave-one-out analysis proved that Mg^{2+}positions predicted by the classifier successfully reproduce the crystallographically determined Mg^{2+}-binding sites. Comparison to another statistical-based prediction tool MetalionRNA demonstrated that our classifier could better identify Mg^{2+}-binding sites in the functionally critical K-turn regions. I anticipate this classifier could serve as a computational tool to predict Mg^{2+}-binding sites in computationally modeled structures. Since NMR spectra typically could not provide structural information of ion locations, this predictor could also be used to predict missing Mg^{2+}ions in NMR experimental structures.

There are other computational tools available for predicting Mg^{2+}-binding sites in RNA structures with high accuracy. However, most other models predict binding sites by simulating the dynamics or solving the non-linear Poisson-Boltzmann equations [107, 112, 115], which required much higher computational costs and specialized expertise to set up the system. Thus, they cannot be used to predict Mg^{2+}-binding sites in a large set of RNAs, as I did in this work. As such, I do not report a direct, full comparison between our classifier and these previously developed predictors.

The Mg^{2+}-binding sites predictor described in this chapter requires RNA 3D structures as an input. Here, I only analyzed its performance on high-accuracy, experimentally determined RNA structures. A possible next step would be to assess its ability to predict Mg^{2+}-binding sites in relatively low-resolution structures and explore the possibility of integrating

the ion prediction with RNA 3D structure prediction techniques, to facilitate RNA structure prediction in the absence of experimentally determined Mg^{2+} ions.

Chapter 4.

Mining For Bound-Like Conformations of RNA Using a Binding Cavity Screening Approach

There is now a keen interest in targeting functional, non-coding ribonucleic acids (RNA) with small-molecules ligands to modulate cellular processes. In principle, virtual screening could help identify small molecules that interact with the RNA by fitting virtual small molecules into binding cavities on its surface. However, for an RNA, especially large RNAs that bear pharmacological significance, there are usually multiple such cavities. Several cavity detection techniques have been developed, lacking, however, are methods for discriminating "druggable" cavities (cavities that bind to small molecules) from the so-called "decoy" cavities (geometrically feasible cavities that do not bind to small molecules). To identify the "druggable" cavities, I developed a binding cavity classifier, known as CavityPoser, using machine learning methods and the distance-based fingerprinting technique described in the previous chapter. In most instances, the CavityPoser was able to recover native-like "druggable" bind-

ing cavities, suggesting that it would be useful as a tool for "druggable" mining cavities in RNA targets. Moreover, I tested whether bound-like RNA conformations could be extracted from a conformational pool by recognizing those that harbored "druggable" cavities. Over a set of 6 RNAs, I found that by searching for conformations containing "druggable" pockets, we could recover conformations with structures similar to known bound-like conformations of the RNAs.

4.1. Introduction

Functional RNAs play essential roles in the cell, and abnormal RNA function is now known to cause many diseases. Moreover, RNA structures regulate many fundamental processes in pathogens. Therefore, targeting functional RNAs with small molecules offers opportunities to modulate RNA-mediated cellular processes, and in pathological cases, reverse the effects of RNA associated diseases [100, 141]. There are now many examples of small molecules that bind non-coding RNAs (ncRNAs) and modulate their function [100, 141]. *De novo* identification of such small molecules remains an unsolved challenge.

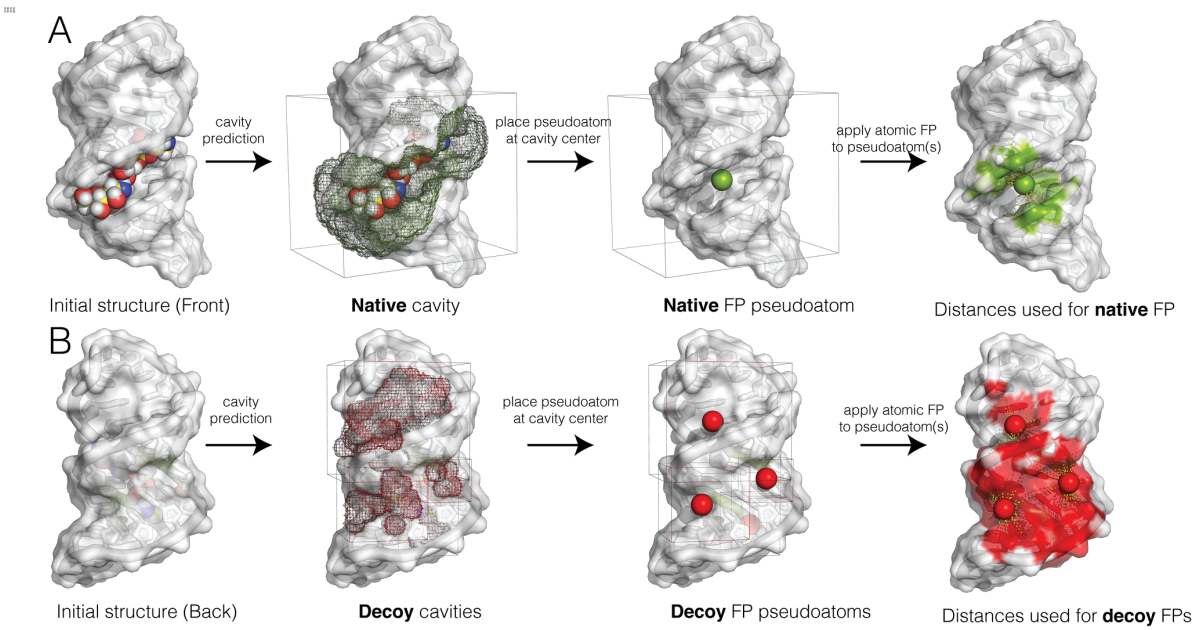
In theory, structure-based virtual screening holds great promise in identifying RNA-targeting small molecules [102, 142]. By definition, virtual screening is the process of searching small-molecule ligands that are likely to bind to a known target (here, an RNA), in so doing, narrow down the putative small molecule candidates from a large library to a few most promising ones before carrying out the experimental high throughput screening (HTS). Several fundamental components were required when considering RNA as drug targets in virtual screening, including the selection of RNA structure and the identification of "druggable" cavities (binding sites that could accommodate small molecules).

Selecting the RNA structure to use for virtual screening is non-trivial. First, like proteins,

RNAs fold into complex 3D structures, and a single RNA structure can possess multiple binding cavities. These cavities can be detected using various computational tools (cavity mapping methods) [143–146], such as reference ligand method and two probe sphere method [147]. However, since an RNA can possess multiple geometrically feasible cavities, docking against all possible cavities is time-consuming and may increase false-positive rates. Therefore, the foreknowledge of which of the detected cavities is "druggable" can enhance structure-based virtual screening [148]. Second, RNA structures are flexible and undergo rapid conformational changes, resulting in a highly flexible ensemble, which has been a challenge for choosing the right receptor structures in virtual screening. When the RNA comes into contact with a small molecule ligand, the conformational state that dominates the bound-state ensemble, also known as the *holo* state, is generally different from the conformational state that dominates free-state ensemble (*apo* state). As such, the traditional rigid "lock-key" model, which assumes that the RNA receptor adopts one fixed conformation, is unlikely to achieve good performance. One way to account for target-flexibility is ensemble docking. Ensemble docking involves applying rigid docking to each member of an ensemble of the receptor, instead of a single conformation. Several strategies can be used to construct RNA ensembles that are suitable for ensemble-docking. For instance, the ensemble can be constructed as a collection of experimentally determined [149] or computationally predicted structures of the target. Ensemble docking exhibited superior performance over most single receptor conformation [149] but was often limited by the computational costs, which scale linearly with the size of the ensemble [150]. One solution to this limitation is to identify a smaller subset of bound-like conformations from the ensemble containing "druggable" cavities to perform ensemble docking. By restricting the conformations used in ensemble docking to a few most important ones, a good compromise between computational speed and prediction accuracy could be achieved. Experimental data, such as residual dipo-

lar couplings (RDC) from NMR experiments [151] or ensemble-averaged reactivities from chemical probing experiments as discussed in Chapter II, could be used as constraints to select the bound-like conformations. In the absence of experimental data, methods to select ensembles enriched with bound-like conformations are currently lacking.

Here, I explored a data-driven framework to identify bound-like conformations by searching for cavities with similar structural properties to known "druggable" cavities. An increasing number of RNA-ligand complexes have been solved experimentally and could facilitate the development of data-driven predictive methods to discriminate between "druggable" cavities from non-druggable ones (decoys). In this work, I developed CavityPoser, a set of machine learning models that can classify "druggable" cavities from decoys. CavityPoser was able to rank the known "druggable" cavities within the first 1.65 positions on the test sets, and over 70% of the "druggable" cavities care ranked **1st** among all cavities. An importance analysis using one of the models suggests that the "druggable" cavities of an RNA have a preference for pyrimidine nucleotides and oxygen atoms in RNA major grooves. Then, I applied CavityPoser to identify bound-like conformations from a pool of lowest energy modeled structures. Results suggest that the conformers with the highest scored cavities have a similar structure to the experimental *holo* structures and possess cavities close to the known "druggable" cavities.



(a)

Figure 4.1: (a) Illustration of the cavity fingerprinting employed in this work.

4.2. Methods

4.2.1. Dataset

I compiled an extensive set of 131 RNA structures containing ligand-binding cavities for which crystal or NMR structures are available. For each RNA structure, I first used the cavity mapping program *rbcavity* from the rDock modeling suite [17] to identify all cavities in the RNA structure. A cavity is labeled as *decoy* if the center of geometry of the cavity's bounding box is more than 6.0 Å from the center of geometry of the known ligand, and labeled as *native* if within 6.0 Å. Next, I placed a pseudo-atom at the center of geometry of each cavity and computed the atomic fingerprint based on atomic distances between the pseudo-atom and nearby RNA atoms (as described in Chapter III, Equation 3.1), which was then used as features to train the CavityPoser. The training and test set were split among

structures, where two test sets, a set of 25 high-resolution X-ray structures, and a set of 24 NMR structures, were excluded from the training.

4.2.2. Training algorithm

To be able to discriminate "druggable" cavities from decoys, I trained a set of machine learning classification models (classifiers), including Extreme Gradient Boosting (XGB), Random Forest (RF), and a simple neural network classifier Multi-Layer Perceptron (MLP). All the classifiers take the cavity fingerprint as input and output the binding score. The binding score is a float number ranging from 0 to 1, and is an estimate of the probability of whether the associated cavity was "druggable". If the score is closer to 1, it indicates the associated cavity has a high probability of being "druggable". On the other hand, if the score is close to 0, it indicates the associated cavity is likely a decoy.

4.2.3. *De Novo* modeling of bound-like RNA conformations

De novo modeling is an approach for structure prediction, which produce a set of candidate structures and choose amongst them based on their properties. Here, I carried out a *de novo* modeling of bound-like RNA structures based on the "druggability" of possessed cavities. As model systems for examining the utility of binding cavity analysis in *de novo* modeling of RNA structures, I assembled an additional set of 5 RNAs containing binding cavities for small-molecule ligands (Table 4.1). Each of the RNAs has one experimentally identified cavity except for the synthetic neomycin-sensing riboswitch, which binds to two different ligands, composing a total of 6 RNA-ligand systems. The modeling protocol is described below.

1. Generate a set of RNA structures using the modeling program SimRNA [47] with se-

quence information and the corresponding secondary structures. The details of the SimRNA modeling was in Appendix B.1.

2. Identify cavities on each structure in the ensemble using *rbcavity* [17]. A pseudo-atom was placed at the center of the cavity's bounding box.

3. The cavity fingerprint and the corresponding binding score was obtained using the atomic fingerprint and the trained binding cavity classifier, CavityPoser.

4. Cavities among all structures in the ensemble were ranked based on their binding scores, and the structures with the top-scored cavities were identified and compared to the known *holo* structure.

Table 4.1: RNA-ligand Systems Used in *de novo* modeling

holo PDB	Description	ligand	apo PDB	reference
2L1V	preQ1 riboswitch (Class I) aptamer	PRQ	3Q51	[152]
2L94	HIV-1 frameshift site	L94	1Z2J	[153]
2LWK	influenza A virus RNA promoter	0EC	1JO7	[154]
2M4Q	E. coli ribosomela decoding site	AM2	3Q51	[155]
2MXS	synthetic neomycin-sensing riboswitch	PAR	-[u]	[156]
2N0J	synthetic neomycin-sensing riboswitch	RIO	-	[156]

[u] Unpublished data

4.3. Results

In this study, I developed CavityPoser, a binding cavity classifier for the prediction of ligand-binding cavities in RNA 3D structures, using experimental structures and machine learning methods. I sought to assess the extent to which CavityPoser could distinguish"druggable"

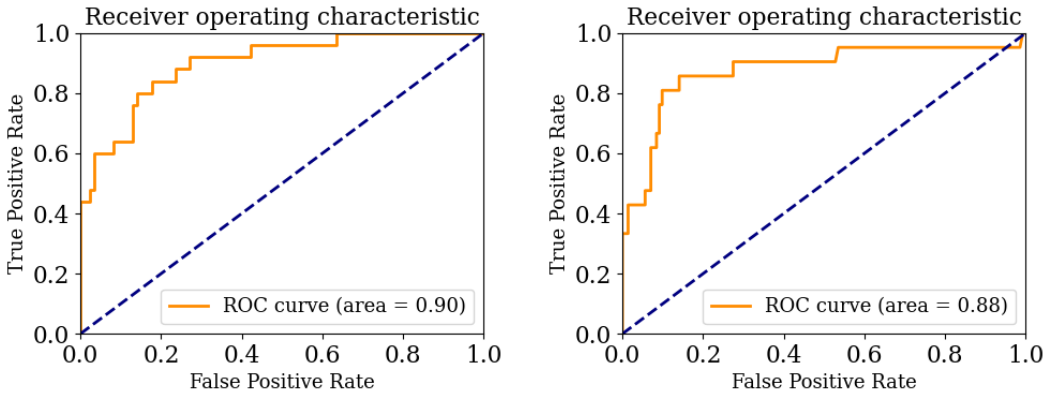
RNA binding cavities from "decoy" binding cavities, and analyze the chemical implication. I also explored whether bound-like RNA conformations could be identified from a conformational pool by screening for structures containing "druggable" binding cavities.

4.3.1. CavityPoser can distinguish "druggable" cavities from decoys.

To generate and test the CavityPoser, I began the study by developing a framework for estimating the "druggability" of RNA binding cavities. Cavities were modeled using rDock, and a cavity was defined as "druggable" if it is near an experimentally-identified ligand-binding site in holo RNA structures. The binding cavity fingerprint was computed as the atomic fingerprint for the pseudo-atom probe placed at the center of a cubic box bounding the cavity as described in Methods. With both known ligand-binding cavities and modeled decoy cavities, machine learning classifiers were trained that took the binding cavity fingerprint as input and returned a binding score ranging from 0 to 1, an estimation of the probability that the corresponding cavity is "druggable". The CavityPoser is the collection of three different classifiers, XGB, RF and MLP, and the output of the CavityPoser represents the average binding scores predicted with different classifiers.

Figure 4.2 and Table 4.2 summarize the performance of the CavityPoser on the 2 test sets. The first test set corresponds to a set of 25 RNA-ligand systems for which X-ray crystal structures were available (X-ray test set), and the second one corresponds to a set of 21 RNA-ligand systems for which NMR structures were available (NMR test set). Shown in Figure 4.2 are the ROC curves I obtained when the CavityPoser was applied on to the two test sets. The overall resulting AUC are 0.9 and 0.88, respectively, for the X-ray and NMR test sets. The cavities were ordered by decreasing binding score, and the ranking of the native cavity with respect to the total number of cavities identified for each RNA is shown in Table 4.2.

Only RNA structures with more than one cavities (at least one decoy cavities) were included in the table, excluding 11 X-ray structures and 2 NMR structures in which no cavities were identified as decoys by cavity mapping methods. The classifiers were able to identify most of the native cavities for X-ray and NMR structures within the first three ranking positions, with an average ranking of 1.36 and 1.65. The only exceptions are 2XNW (X-ray) and 1EHT (NMR). Furthermore, for 12 out of 14 X-ray structures and 12 out of 19 NMR structures, the native cavities were ranked first among all cavities of the same structure.



(a) (b)

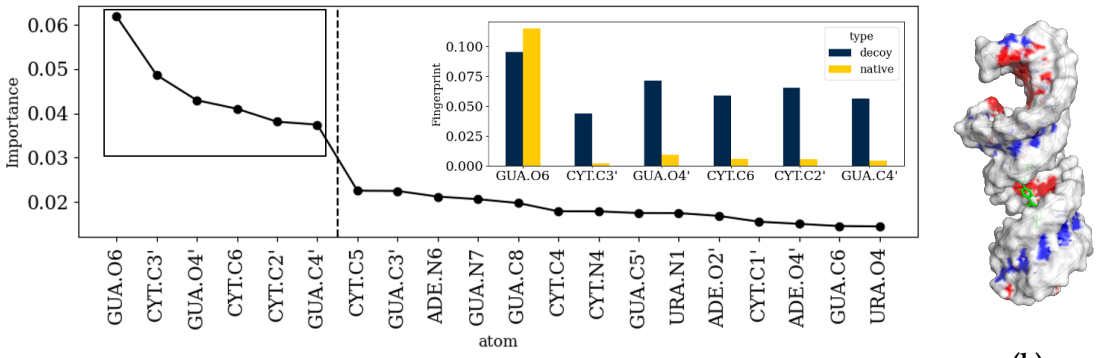
Figure 4.2: ROC curve of the cavity prediction for the systems in (a) test set 1 (X-ray structures) and (b) test set 2 (NMR structures).

4.3.2. Small-molecule ligands have a preference for RNA major groove.

Next, I explored the chemical implication of CavityPoser by examining which residues and atoms in the fingerprint are most important in the classification of "druggable" cavities. To accomplish this, I carried out feature importance analysis using the feature importance scores of one of the member classifiers in the trained CavityPoser, the random forest classifier. In theory, feature importance analysis could be carried out for any predictive models, but the random forest model provides the most straightforward method to obtain feature

importance scores. Shown in Figure 4.3a is the plot of importance scores associated with the few most important RNA atoms, defined by RNA residue types (ADE, GUA, CYT, and URA) and atom names. The majority of the most important atoms are in GUA and CYT residues, which is likely a result of the enrichment of GC base pairs in the training data.

A comparison of the average fingerprint of native and decoy cavities is shown as an inset in Figure 4.3a. As can be seen from the definition of the atomic fingerprint (Equation 3.1), the fingerprint for an RNA atom type is larger if more RNA atoms of the specified type are found within the cutoff distance from the pseudo-atom placed at the center of the cavity, or if the distance is smaller between the pseudo-atom and the center of the cavity. As such, based on the training data I used here, GUA:O6 atoms are favored by "druggable" cavities, while CYT:C3', GUA:O4', CYT:C6, CYT:C2' and GUA:C4' atoms are disfavored by the "druggable" cavities. In other words, if a cavity is found to be in the vicinity in the GUA:O6 atom, then it is more likely to be a "druggable" cavity. Interestingly, the GUA:O6 atoms, preferred by the native cavities, is the only one out of the six most important atoms located in the major groove, while all the other five atoms are located in the minor groove. Shown in Figure 4.3b is an illustration of major groove, minor groove and the ligand-binding site in an RNA molecule. The major groove is deep, and presents a richer ensemble of hydrogen bond acceptors and donors. The minor groove is shallow but is more accessible to the surrounding. Therefore, it is possible that the small-molecule ligands in the RNA-ligand complex have a preference for major grooves. Recent studies on RNA-ligand interactions also revealed that the deep major groove was the most preferred location for some small-molecule ligands [157]. However, I can not rule out that this preference is due to artifacts of the training data or classifier.



(a) (b)

Figure 4.3: The average fingerprint value of decoy and native cavities for the atoms contributing most to the cavity binding classification. (a) The x-axis label is in the format of "RES.Atom". The atoms were shown from left to right in importance descending order and the 6 atoms with highest importance scores were shown in this figure. The importances are identified using features importance scores yielded by the random forest classifier. (b) The major groove (red) and minor groove (blue) of an RNA molecule in complex with a small-molecule ligand (green).

4.3.3. Classifiers extract bound-like structures from ensembles.

Finally, I explored whether bound-like conformations could be distinguished from a conformational pool using only binding scores of possessed binding cavities. To test this, I generated ensembles of diverse structures using the modeling program SimRNA for 5 RNA structures, whose *holo* structures in complex with various small-molecule ligands are available. Each of the structures in the ensemble was fed into *rbcavity* to identify cavities and binding scores were predicted using CavityPoser. Three structures with the highest scored cavities in each RNA ensemble were identified and compared to the known experimental *holo* structures.

The results are summarized in Table 4.3 and Figure 4.4. In Table 4.3, structural RMSD between SimRNA structures and the corresponding *holo* structures and distances between the predicted cavity center and *holo* ligand center were shown. On average, the structures with top-3 scored cavities have a smaller RMSD and distance compared to the ensemble

average. Shown in Figure 4.4 are comparisons between the known holo conformations and the top 3 structures for the RNA structures, which exhibited the highest binding scores. Visual inspection of the modeled structures reveals that the top-scored cavities identified by the classifier fall within the opening of pockets surrounded by RNA residues. Overall, the comparison between the *holo* structure and modeled structures revealed that CavityPoser trained in this work was able to extract several bound-like structures based on binding scores of their cavities.

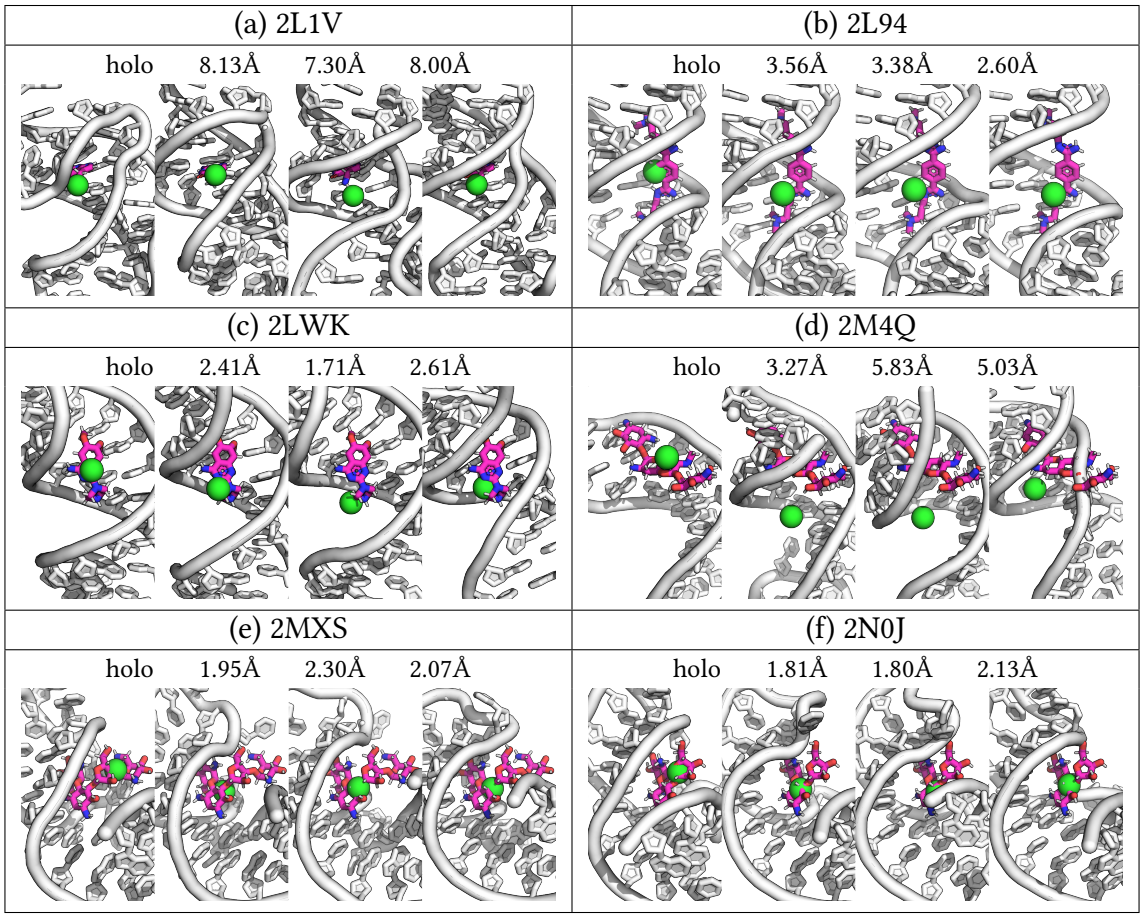


Figure 4.4: The holo structure and 3 modeled bound-like structures with highest-scored cavities for the additional test systems in complex with the holo ligand. Labeled are the heavy-atom RMSDs to the holo structure.

	PDB ID	Rank
X-ray	2B57	1/6
	2O3W	1/5
	2XNW	4/6
	2YDH	1/8
	3LA5	1/6
	3MUM	1/9
	3SD3	1/10
	3SLM	3/5
	4AOB	1/7
	4FE5	1/8
	4KQY	1/8
	4LX5	1/8
	4XWF	1/5
	5C7W	1/8
	average	1.36
NMR	1AJU	1/10
	1AKX	1/11
	1AM0	1/9
	1BYJ	1/10
	1EHT	6/6
	1EI2	1/8
	1FMN	2/10
	1KOC	2/3
	1KOD	1/4
	1LVJ	2/4
	1NEM	2/8
	1O9M	1/11
	1Q8N	2/11
	1QD3	1/11
	1TOB	1/11
	1UTS	1/8
	1UUD	1/9
	1UUI	1/11
	2TOB	2/6
	average	1.65

Table 4.2: Results of the cavity prediction for each of the systems in the X-ray and NMR test sets. Values in the *Rank* column represent the rank position of the native cavity over the total number of cavities identified by rbcavity.

RNA	RMSD(Å)	distance(Å)	Top-3 Scored Structures		
			score	RMSD(Å)	distance(Å)
2L1V	6.24 ± 1.46	3.05 ± 0.97	0.942	8.13	1.80
			0.917	7.30	2.72
			0.876	8.00	2.50
2L94	4.14 ± 1.12	5.88 ± 1.68	0.877	3.56	4.62
			0.875	3.38	4.89
			0.851	2.60	4.91
2LWK	2.43 ± 0.52	6.17 ± 1.06	0.997	2.41	4.95
			0.995	1.71	5.76
			0.991	2.61	5.85
2M4Q	5.76 ± 1.23	8.24 ± 1.25	0.965	3.27	5.82
			0.878	5.83	8.24
			0.842	5.03	7.34
2MXS	2.82 ± 1.00	5.86 ± 0.94	0.997	1.95	4.88
			0.996	2.30	5.50
			0.996	2.07	5.20
2N0J	2.72 ± 1.22	4.03 ± 1.04	0.996	1.81	3.74
			0.996	1.80	3.88
			0.993	2.13	3.04

Table 4.3: "RMSD": mean and standard deviation (shown in parenthesis) of the RMSD between aligned holo and modeled ensemble of structures. "distance": mean and standard deviation of the distances between holo cavity center and cavity centers in the modeled ensemble of structures. "Top-3 Scored Structures": and binding score, RMSD and distance of the 3 structures with highest scored cavities.

4.4. Discussion

In this study, I developed a method for predicting "druggable" cavities in an RNA 3D structure. The prediction task was first converted to a classification problem, allowing us to apply machine learning methods. I then used a distance-based fingerprinting method, as described in Chapter III, and a combination of machine learning methods to unveil "druggable" cavities. The resulting binding cavity classifier, named CavityPoser, successfully ranked "druggable" binding cavities. For the X-ray and NMR RNA-ligand test sets, respectively, I was able to rank the "druggable" cavities within the first 1.35 and 1.65 positions among all cavities identified by cavity mapping methods. In addition, the native drug cavity was ranked first place in 85% and 63% of the X-ray and NMR test cases.

I then used CavityPoser to predict the binding probability of an ensemble of structures, aiming to extract bound-like structures with "druggable" cavities. Starting from sequence information, based on the ensemble of lowest-energy structures generated with modeling program simRNA, I was able to identify several bound-like structures that possess "druggable" cavities and appear similar in shape to the known *holo* structures. These bound-like structures, in combination with the cavities, could serve as a starting point for the modeling of RNA receptors in ensemble-based virtual screening.

There are a few alternative approaches for predicting small-molecule binding sites in RNA 3D structures. Zheng et al. [158] calculate the Euclidean distances between each nucleotide and all the other nucleotides in an RNA molecule and determines the functional sites of ncRNAs as nucleotides that are the extreme points in the distance curve. Wang et al. [159] developed a network-based model, in which the RNA tertiary structure is transformed into a network with nucleotides as nodes and non-covalent interactions as edges; it then measures degree values and closeness values to identify the binding sites. Previous works to identify

small molecule binding sites of RNA, however, were typically based on the locations of individual nucleotides and their connections. In comparison, the framework developed here is centered around cavities, which were characterized using atomic fingerprint. The two different points of view are complementary to each other. Nucleotide-based methods provide evidence for ligand-binding cavities on nucleotide-level, whereas CavityPoser provides quantification of the interactions between the RNA molecule and the cavity region on the atomic-level. Together, these methods can provide complementary evidence for predicting ligand-binding RNA cavities.

Currently, CaviyPoser is only designed to predict the probability of a cavity that binds small-molecule ligands. However, to identify a cavity with pharmaceutical significance, it is also necessary to quantitatively assess the *binding affinity*, the strength of the binding interaction between the cavity and small-molecule ligands. Despite the limitation, CavityPoser is still helpful as it presents candidate cavities for further virtual screening and experiments. When combining with molecular docking, it is feasible to further investigate the binding affinity of the "druggable" cavities identified by the CavityPoser. It is also possible, with available binding affinity data, to train machine learning regression models that are capable of predicting the binding affinity of a cavity. The *Multi-task learning* (MLT) framework, in which one machine learning model can be used to make predictions on multiple correlated properties, can be used to model the "druggability" and binding affinity at the same time.

Chapter 5.

Conclusion

5.1. Summary

The structure and energetics of RNA are important topics in biophysical and medical sciences. RNA is a common macromolecule observed across various living organisms, and perform a diverse array of functions in the cell, including catalyzing in-cell reactions, guiding gene expression, and facilitating the assembling of cellular complexes. For these reasons, there is a vested interest in understanding the RNA structures better and characterizing their functions. The existing effort has been devoted to accurate determination of the low energy conformation states of RNA structures. However, there is a lack of tools that characterize these structures and utilize the structural information to examine the structural flexibility and quantitatively predict fundamental biophysical properties.

In this work, structural fingerprints based on solvent accessible surface area (SASA), as well as computationally crafted numerical descriptors, were used to characterize local atomic environments and build predictive models for RNA structures and structure-related properties. For the first case, a framework to systematically predict RNA structural ensembles from

known sequence and ensemble-averaged chemical reactivity was developed. This framework allowed us to assign weights to individual conformations in an RNA ensemble using Bayesian/maximum entropy (BME), and the weighted ensemble agrees with experimental measurements. The framework was applied onto the SAM-I riboswitch ensemble and identified a cryptic binding pocket. In the second example, an atomic fingerprinting method based on pairwise atomic distances were proposed. Using the fingerprinting methods and its extension to molecular level descriptors, machine learning models to predict magnesium ion binding sites and ligand binding poses were developed. The fingerprinting method was also applied to study the RNA-targeting drug design problem. Hereby once again, machine learning models were used to identify druggable RNA cavities, regions on the RNA surface that bind to small molecules to modulate potentially disease-related cellular functions.

This work has focused on the computational predictions of RNA structural ensembles and structure-related properties using various fingerprinting methods. It fits in the ultimate goal to elucidate structure-function relationships that govern RNA-based cellular functions. We witnessed a brief expose of the promise of structural fingerprints in the prediction of RNA, and the works presented here provide optimistic expectations on the fidelity of atomic environment fingerprints and provide guidance for future experimental work in RNA structure modeling and RNA drug design pipeline.

5.2. Future Directions

In Chapter II, one binding pocket different from the known experimental binding pocket emerged in the analysis of the free-state SAM-I riboswitch ensemble as the most probable site to bind small-molecule ligands. The possible next step would be to perform experiments to validate that small molecules could bind to the pocket. Experiments could further validate

the binding potential of this pocket and promote the understanding of the binding properties.

In Chapters III and IV, the prediction of Mg^{2+}binding sites and the ligand-binding cavities were modeled as machine learning classification problems. Nonetheless, both problems aim to resolve the ranking position of a sample (the Mg^{2+}or "druggable" binding site) in a set of samples (all possible binding sites in the same RNA molecule). A possible future direction would be to reformulate these problems as ranking problems and solve using Learning To Rank (L2R) algorithms. L2R is a class of supervised learning techniques aiming to produce a permutation of samples in a list of samples. Compared to traditional classification and regression problems that make predictions on a single instance at a time, L2R solves the problem involving a list of samples with emphases on their partial orders, for example, the rank position of each query in a list of search results. In Chapter III, the prediction of Mg^{2+}binding sites in each RNA structure could be formulated as a ranking problem, which generates the partial ordering of the binding sites in each RNA rather than the absolute binding scores. In Chapter IV, several cavities were identified from cavity mapping methods, and the rank position of the "druggable" cavity among all cavities was the main focus of the problem rather than the exact classification scores of each cavity. Although the classification models trained in this work have achieved good accuracies, formulating those prediction tasks into L2R problems and solving them using the well-developed ranking algorithms, like RankNet and LambdaRank [160], provide a new, more direct solution that could potentially achieve even better performance.

Appendix A.

RNAPosers - Machine Learning Classifiers For RNA-Ligand Poses

This appendix included the details for RNAPosers (a set of machine learning classifiers to identify the best RNA-ligand poses), as an additional example of the application of the atomic fingerprints described in Chapter II. This work was done in collaboration with my lab mate, Dr. Sahil Chhabra. The contents were published in the following article:
Chhabra, Sahil, Jingru Xie, and Aaron T. Frank. "RNAPosers: Machine Learning Classifiers for Ribonucleic Acid–Ligand Poses." *The Journal of Physical Chemistry B* (2020).

A.1. Introduction

Beyond acting as an intermediary between deoxyribonucleic acid (DNA) and proteins, ribonucleic acids (RNAs) play key regulatory roles within the cell [161–163]. For instance: ribosomal RNAs (rRNAs) catalyze protein synthesis[164]; riboswitches turn on and off RNA transcription or translation[165]; and short interfering RNAs (siRNAs)[166] and microR-

NAs (miRNAs)[167] silence the expression of targeted mRNAs. Indeed, many classes of "functional" RNAs are implicated in diseases[168] and are now considered viable drug targets[169–172]. Moreover, targeting RNAs with small molecules has garnered keen interest over the last decade[141, 173–175]. Rational structure-based methods promise to be a viable approach for identifying small molecules that can bind to and modulate the activity of structured RNAs.[102] Crucial to the success of rational structure-based approaches in RNA drug discovery is the ability to accurately predict the 3-dimensional (3D) structure of the complex formed between an RNA and a small molecule ligand. In principle, computer docking algorithms can be used to predict the 3D orientation and conformation (referred to as the pose) of a ligand bound to an RNA receptor. Unfortunately, "redocking" tests reveal that state-of-the-art scoring functions typically fail to recover the correct poses [176–180]. Moreover, several of the state-of-the-art RNA-specific scoring functions that have been recently developed are, for a variety of reasons, inaccessible to the scientific community. In these respects, there is an urgent need for methods that can distinguish "native-like" RNA-ligand poses from non-native decoy poses and are accessible to the wider RNA community. In this study, we sought to fill this critical void.

Recently, machine learning has been used to address several challenges associated with computer docking and virtual screening. For protein-ligand complexes in particular, machine learning has been used to develop more robust scoring functions for both pose and binding affinity prediction.[66, 181–185] Here, we used machine learning to train a set of pose classifiers that quantify the "nativeness" of RNA-ligand complexes. In what follows, we summarize our comparison between the ability of docking scores and machine learning classifiers to rank and identify atomically correct RNA-ligand poses. Compared with docking scores, we found that machine learning pose-classifiers were better able to discriminate native-like RNA-ligand poses from decoy poses. Accordingly, we make our pose classifiers

freely available to the scientific community via https://github.com/atfrank/RNAPosers.

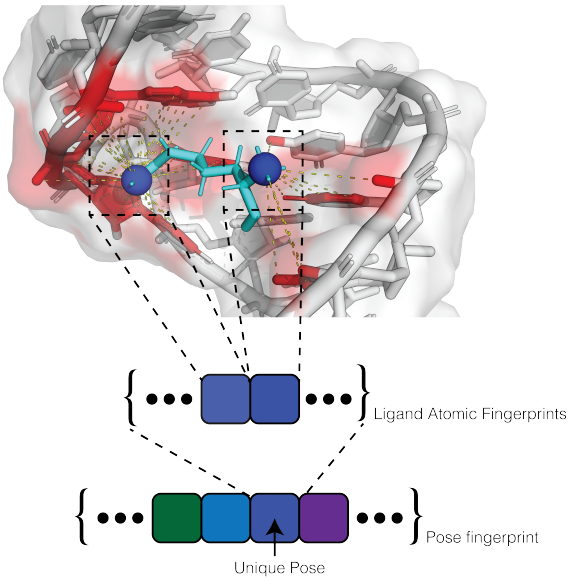


Figure A.1: Illustration of the fingerprinting approach we used to describe RNA-ligand interactions. From the structure of an RNA-ligand complex, atomic fingerprints were obtained through distances calculated between each ligand atom and its neighboring RNA atoms within 20 Å, and then all atomic fingerprints in the ligand are combined to construct the pose fingerprint. Each element in the final pose fingerprint is associated with a unique atom-pair type, as defined by the atom types of the ligand and the RNA.

A.2. Materials and Methods

A.2.1. Pose classifiers

Given the 3D coordinates of an RNA-ligand pose, we attempted to develop a method to estimate or classify whether the pose was native-like based solely from the atomic coordinates of the pose. In other words, we attempted to develop "pose classifiers". Here, machine learn-

ing was used to train a set of pose classifiers that estimate the "nativeness" of a RNA-ligand pose from a set of "pose features", which we define as any structural or structure-derived information that can be extracted or calculated from the atomic coordinates of the pose (Table A.1). First, we generated a set of classifiers for which the "pose features" correspond to individual scoring terms in the rDock scoring function.[17] Second, we generated a set of classifiers for which the "pose features" correspond to a novel pose fingerprints (FPs) that depends on the pairwise distance between heavy atoms in an RNA receptor and heavy atoms in a small molecule ligand (see below). For completeness, we also included pose predictions directly using rDock scoring terms, and classifiers trained on a combined feature using both rDock scoring terms and pose fingerprints.

Pose prediction method	Description
rDock	rDock scoring terms
Score Classifier	Classifiers trained on rDock score terms only
Pose FPs Classifier	Classifiers trained on pose fingerprints only
Score+Pose FPs Classifier	Classifiers trained on rDock score terms and pose fingerprints

Table A.1: Summary of the pose-prediction analyses carried out in this study. First, we carried out pose prediction using various score terms in the rDock RNA-ligand scoring function. Then, we explored using pose classifiers, which predict the nativeness of poses from rDock score terms, pose fingerprints (FPs), and a combination of rDock score terms and pose FPs, respectively.

A.2.2. Pose fingerprint

We utilized a pose fingerprint that is a composite of a set of atomic fingerprints (Figure A.1). For a given ligand atom i, its atomic fingerprint corresponds to the vector, $V_i = \{V_{i,j}\}$, whose elements $V_{i,j}(\eta)$ are given by

$$V_{i,j}(\eta) = e^{-(r_{ij}/\eta)^2} \cdot f_d(r_{ij}) \qquad \forall\, i \in ligand, j \in RNA \tag{A.1}$$

where r_{ij} is the distance between a heavy atom i in the ligand and heavy atom j in the RNA receptor. We only consider atom pairs within a cutoff distance $R_c = 20$ Å. η is the width of the Gaussian function (here we set $\eta = 2$). And $f_d(r_{ij})$ is the damping function given by

$$f_d(r_{ij}) = 0.5\left[\cos\left(\frac{\pi r_{ij}}{R_c}\right) + 1\right]. \tag{A.2}$$

We note that the atomic fingerprint based on Eq. A.1, which is a multi-element extension of the atomic fingerprint developed by Botu et.al. [122], is invariant to basic atomic transformation operations of translation, rotation and permutation.

For a given ligand pose p, its fingerprint vector F_p was composed of ligand atomic fingerprint by summing over all instances of a given atom-pair type s. Each atom-pair type is defined by two parts, namely SYBYL atom type of the ligand atom and residue and atom name of the RNA atom. The set of all possible permutations of ligand atom type and RNA atom types are denoted as S. As such, an element in the fingerprint for pose p and atom-pair type s is given by

$$F_{p,s} = \sum_{(i,j)\in s} V_{i,j}(\eta) \qquad \forall\, s \in S \tag{A.3}$$

We used a set of 20 SYBYL atom types for ligand, and 85 atom types for RNA (Table B.3). Thus, the pose fingerprint of each pose $F_p = \{F_{p,s}, s \in S\}$ contained 1700 elements (20 SYBYL types × 85 RNA atom types). Finally for each RNA-ligand system, each pose fingerprint was normalized by its ensemble median to ensure unity among various RNA-ligand systems. Coincidentally, our pose fingerprint closely resembles a recently described fingerprint that was successfully used to train machine learning pose and binding affinity predictors.[185]

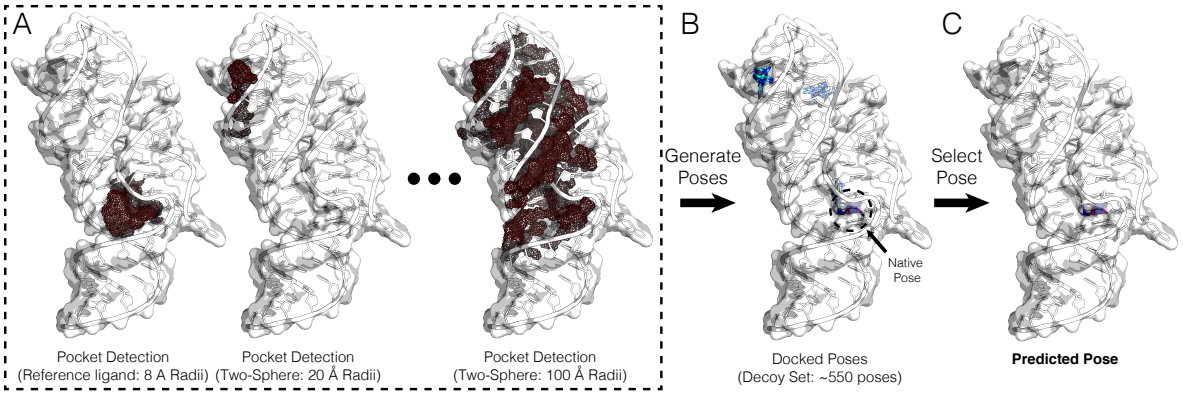


Figure A.2: Illustrated are the steps involved in generating the decoy sets used in this study. (A) Step 1 and 2, the actual binding pocket is mapped using the reference ligand method, and alternative pockets are mapped using two-sphere methods with increasingly large radii. (B) Step 3, poses were generated by docking the ligands into each of the mapped binding pockets and combined into a single decoy set. (C) The focus of this study is to develop and assess methods for selecting atomically-correct poses from these decoy poses (i.e., pose prediction).

A.2.3. Datasets

Leave-one-out training set

We compiled an initial dataset comprised of 80 RNA-ligand systems. For this dataset, the crystal structures of the RNA-ligand complexes were downloaded from the Protein Data Bank [60](see Table B.4 for a list of PDBIDs). To generate diverse decoy sets for each RNA-ligand system, computer docking was performed using the docking program rDock[17]. The following protocol was used to generate the poses with rDock (Figure A.2A and B). First, a set of poses were generated in the actual binding pocket, using the reference ligand method, with the sphere radii from the center of the known binding pocket set to 2, 3, 4, 5, 6, 7, and 8 Å, respectively. At each sphere radius, 50 poses were generated, for a total of 350 poses. Next, 250 additional poses were generated by docking into the binding pockets that were identified using the two-sphere method, with outer sphere radii set to 20, 40, 60, 80 and 100 Å, respectively. Docking was carried out with the default rDock scoring function

plus a solvation term from a desolvation potential defined as a weighted sum of solvent accessible surface area.[17, 186] Hence, in total, 600 poses were generated for each RNA-ligand complex. For some RNA-ligand complexes, the number of poses were less than 600 because the two-sphere method failed to identify binding pockets at one or more of the outer sphere radii we utilized for binding pocket detection. All pocket detection was carried out using the rDock utility program, rbcavity. The entire set of decoy poses can be accessed online at https://doi.org/10.5281/zenodo.3711071.

Validation set

We also compiled two additional, independent validation sets (Table A.2). The first was comprised of 17 RNA-ligand systems whose structures have been solved using X-ray crystallography. The second validation set was comprised of 21 RNA-ligand systems containing both X-ray and NMR structures. Excluded were systems in which the RNA shared high sequence similarity (> 80%) to RNAs in the leave-one-out training set.[187] For each RNA in these validation sets, a decoy set of ~600 poses was generated using the protocol identical to the one used to generate decoy poses in leave-one-out dataset. These dataset were then used to test the pose-recovery performance of the machine learning classifiers we trained on leave-one-out datset. The second dataset, the majority (20 out of 21) of which were NMR structures, was viewed as a particularly strong validation set because no NMR structures were included in the leave-one-out dataset that was used to train the classifiers. Moreover, all of the systems included in this validation set were also used to test the performance of the scoring function, DrugScoreRNA, thus facilitating a fair comparison between the performance of our predictors and a current state-of-the-art scoring function.

Dataset	Size	N	$f_{<2.5}$
Training set 1	80	43750	0.27
Validation set 1	17	8850	0.29
Validation set 2	21	12590	0.22

Table A.2: Summary of the primary datasets used in this study. Listed for each dataset are the size (i.e., the number of RNA-ligand complexes), N, the total number of poses, $f_{<2.5}$, and the fraction of poses with RMSD < 2.5 Å, respectively. See the Supporting Information for the exact composition of the datasets.

A.2.4. Training the pose classifiers

To train the pose classifiers, we employed the random forest method implemented in the sklearn Python module.[188] The classifiers comprised of an ensemble of 1000 decision trees with class weight set to balanced subsample. All other parameters were set to their default values. The classifiers were trained using a leave-one-out approach using the set of poses generated using rDock (see above). We trained separate classifiers with nativeness RMSD thresholds set to 1.0, 1.5, 2.0, and 2.5Å. Machine learning models can be susceptible to the so-called "twinning effect," which occurs when samples in the training set closely resemble samples in testing set. Here we have employed leave-one-out cross-validation in an attempt to mitigate the potential impact of "twinning" when assessing the performance of classifiers. In this leave-one-out approach, a single RNA-ligand system was removed from the training set and the classifiers were trained on the remaining 79 RNA-ligand complexes. The resulting classifier was then assessed on the excluded RNA-ligand system. *If the ligand in any of the other 79 RNA-ligand systems was identical to the ligand in the left-out system, they were removed prior to training the classifier used to assess the left-out system.*

A.2.5. Assessing classifiers

In order to quantify our ability to recover atomically correct poses using either docking scores from rDock scoring function or classification scores from our pose classifiers, we first sorted the poses to obtain the top-scored pose. When using docking scores, the pose with *lowest* (most negative) score was identified and RMSD relative to crystal pose was determined. When using classification scores, the pose with *highest* classification score was identified and RMSD relative to crystal pose was determined. We also calculated the success rates $S(X)$ (with X=1.0, 1.5, 2.0, and 2.5) as the percentage of RNA-ligand complexes for which the RMSD of the best pose (top-scored pose) was within X Å of the corresponding crystal pose.

A.3. Results and Discussion

For protein-ligand complexes, modern scoring functions have a reported success rate that exceeds ~75 %.[189] In contrast, for RNA-ligand complexes, state-of-the-art scoring functions have a success rate near 50 %.[17, 180] This discrepancy between the success rate of protein and RNA scoring functions motivated us to explore methods capable of enhancing our ability to discriminate native-like poses from non-native decoys.

A.3.1. Docking scores exhibit low success rates.

We began our study by assessing the ability of docking scores to recover the correct pose from decoy poses located in the experimental binding pocket as well as decoy poses located in alternate pockets on the surface of the RNA. To accomplish this, we initially generated decoys sets comprised of ~600 diverse poses for 80 RNA-ligand complexes (see Methods;

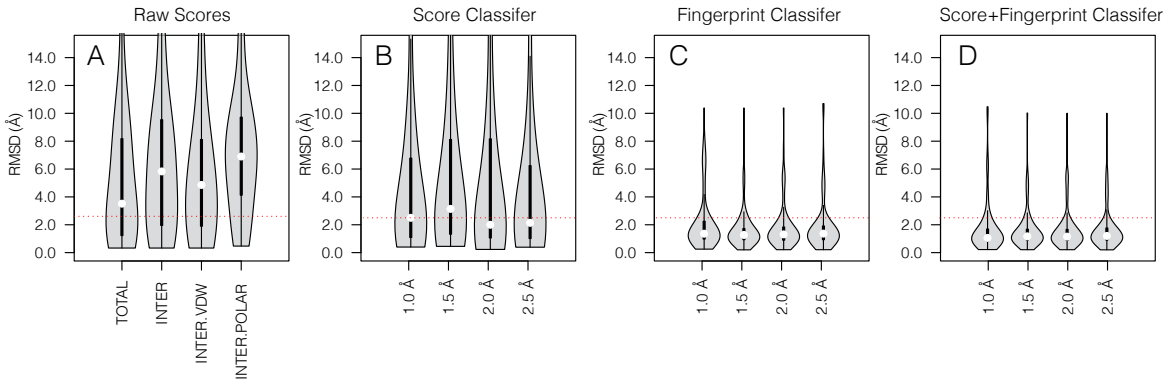


Figure A.3: RMSD distributions of predicted best poses over systems in the leave-one-out training set, when the best poses were predicted using (A) docking score terms, classification scores from classifiers trained using (B) docking score terms, (C) our pose fingerprint, (D) raw docking scores plus our pose fingerprint as features, respectively. Here, TOTAL, INTER, INTER.VDW, and INTER.POLAR refer to various terms in the rDock scoring function: TOTAL corresponds to total docking score, containing both RNA-ligand and intra-ligand contributions; INTER corresponds to the contribution of RNA-ligand interaction to the docking score; INTER.VDW corresponds to the non-polar, van der Waals contribution to the interaction docking score; and INTER.POLAR corresponds to the polar contribution to the interaction docking score. For the pose classifiers, results are shown for independent sets of classifiers that were trained with the nativeness threshold set to 1.0, 1.5, 2.0, and 2.5 Å. Here best poses correspond to the top-scoring pose. When using raw docking scores, the top-scoring pose is the pose with the lowest docking scores. When using the classifiers, the top-scoring pose is the one with the highest classification score. In each violin plot, the black bar in the center corresponds to quantile range, and the white squares are located at the corresponding median RMSD. For reference, the red dotted lines are placed at RMSD values of 2.50 Å.

Figure A.2A-C). In these decoys sets, the RNA receptors corresponded to the holo structures where only the ligand orientation and conformation varied.

Shown in Figure A.3A are distributions of RMSD (relative to the crystal pose) of the best poses selected from these decoy sets using individual score terms in the rDock scoring function.[17] When using the total docking score, the median RMSD of the predicted pose was 3.41 Å (Figure A.3A; Table B.6). We obtained similar results when using the total interaction, the van der Waals interaction, and the polar interaction score terms. In these cases,

the median RMSD were 5.72, 4.75, and 6.88 Å, respectively. To better quantify the ability of the score terms to select atomically correct poses, we also computed the success rate, $S(X)$, defined as the percentage of cases in which the predicted pose was within X Å of the native pose. Using total docking score, $S(1.0)$, $S(1.5)$, $S(2.0)$, and $S(2.5)$ were 22.7, 29.5, 37.5 and 42.0%, respectively. Similarly, $S(1.0)$, $S(1.5)$, $S(2.0)$, and $S(2.5)$ were 17.0, 21.6, 27.3, and 33.0%, respectively, when using total interaction, 18.2, 22.7, 28.4 and 36.4%, respectively, when using van der Waals interaction, and 8.0, 9.1, 12.5, and 21.6%, respectively, when using the polar interaction score terms. The docking score terms in the rDock scoring function, therefore, exhibited marginal ability to recover correct poses from diverse decoys poses.

A.3.2. Pose classifiers improve success rates on the leave-one-out dataset.

Next, we asked whether nonlinear machine learning classifiers could enhance our ability to recover correct poses from decoy poses. To test this, we cast the problem of recovering correct ligand poses as a classification problem and then trained machine learning models to discriminate correct poses from decoy poses. Briefly, we built random forest classification models that take a set of features as input and output "classification scores" that estimate the probability of a pose being native-like. To accomplish this, we first trained a series of random forest pose classifiers using a leave-one-out cross-validation approach in which we selected a single RNA-ligand from the dataset of 80 RNA-ligand systems (the leave-one-out dataset), and trained a classifier using decoy sets for the remaining 79 RNA-ligand systems. If the ligand in the left-out system was identical to any of the remaining 79 system, the data associated with these systems were removed prior to training the classifier (see Methods). After training, the performance of the resulting classifier was assessed on the left-out system. For the left-out system, classification scores for all poses were determined and then the pose

with the highest classification score was selected as the best (or predicted) pose for the left-out system. This procedure was repeated 80 times, i.e., one for each system in the leave-one-out dataset.

Shown in Figure A.3B are distributions of the RMSD of the best poses identified using classifiers trained with individual terms in rDock scoring function as learning features. Reported are results for the classifiers trained with nativeness RMSD threshold set to 1.0, 1.5, 2.0, and 2.5 Å, respectively. The corresponding success rates are also listed in Table B.6. In general, RMSD of the best poses identified using the score-based pose classifiers were lower than those selected using the terms in the rDock scoring function. For instance, for score-based pose classifiers trained with nativeness threshold set to 1.0, 1.5, 2.0, and 2.5 Å, median RMSD of the best poses were 2.50, 3.14, 2.08, and 2.14, respectively (Figure A.3B; Table B.6). The success rates, $S(1.0)$, $S(1.5)$, $S(2.0)$, and $S(2.5)$ were also higher for the score-based classifiers, with the best results obtained with nativeness threshold set to 2.0 and 2.5 Å. $S(1.0)$, $S(1.5)$, $S(2.0)$, and $S(2.5)$ were 21.6, 36.4, 50.0, and 54.5% respectively for classifiers trained with threshold set to 2.0 Å, and 25.0, 37.5, 48.9, and 54.5% for classifiers trained with threshold set to 2.5 Å. In comparison, the values obtained using total docking score were 22.7, 29.5 37.5 and 42.0%. These results suggest that pose classifiers trained using the scores terms as learning features could boost our ability to recover correct poses. The success rates, however, still pales in comparison to the success rates of protein-ligand pose prediction methods.

As such, we next asked whether we could further enhance the success rate of RNA-ligand pose prediction by training pose classifiers on features that more directly depend on RNA-ligand interactions. Specifically, we were interested in examining the utility of a distance-based atomic fingerprint that describes the local atomic environment near a given site which has shown promise in predicting properties like atomic forces[120] and resembles a pose fingerprint recently used for protein-ligand pose predictions.[185] To create a composite

fingerprint from atomic fingerprints, we summed and normalized all atomic fingerprints associated with specific ligand-RNA pair types (see Methods and Figure A.1). Using this composite RNA-ligand interaction fingerprint, we then trained another set of pose classifiers, again using the leave-one-out cross-validation approach. For comparison, we also trained classifiers that used the rDock score terms plus our pose fingerprint as features. Here again, separate classifiers were trained with nativeness threshold set to 1.0, 1.5, 2.0, and 2.5 Å. The results are summarize in Figure A.3C.

For the pose fingerprint classifiers trained with nativeness threshold set to 1.0, 1.5, 2.0, and 2.5 Å, median RMSD of the best poses were 1.36, 1.27, 1.31, and 1.42 Å, respectively. These fingerprint-based classifiers all exhibit similar success rates. For instance, $S(1.0)$, $S(1.5)$, $S(2.0)$, and $S(2.5)$ were 37.5, 63.6, 77.3, and 86.4%, respectively, for classifiers trained with nativeness threshold set to 1.5 Å which exhibited the lowest median RMSD of 1.27 Å. In comparison, $S(1.0)$, $S(1.5)$, $S(2.0)$, and $S(2.5)$ were 33.0, 58.0, 77.3, and 86.4%, respectively, for classifiers trained with nativeness threshold set to 2.5 Å which exhibited the highest median RMSD of 1.42 Å. We obtained comparable results for pose classifiers trained using the docking scores plus the fingerprint as features. Notable among these was the classifier trained with nativeness threshold 1.0 Å; for this set of classifiers, median RMSD of the best poses was 1.05 Å and $S(1.0)$, $S(1.5)$, $S(2.0)$, and $S(2.5)$ were 43.2, 70.5, 79.5, and 85.2%, respectively. Based on this leave-one-out analysis, the pose classifiers trained using pose fingerprint as well as classifiers trained using docking score terms plus pose fingerprint as features, both exhibited remarkable ability to recover atomically correct poses from the leave-one-out decoy sets.

A.3.3. Pose classifiers exhibit high success rates on two independent validation sets.

To further test our pose classifiers, we applied them to two additional validation sets, the first, consisting of a set of 17 RNA-ligand complexes for which X-ray structures were available and the second, consisting of 21 RNA-ligand complexes containing both X-ray and NMR structures. Figure A.4 summarizes the results. Globally, the results mirror those from our leave-one-out analysis: RMSD and success rates decreased and increased, respectively, when using raw docking scores (Figure A.4A, E), score classifier (Figure A.4B, F), fingerprint classifier (Figure A.4C, H) and score+fingerprint classifier (Figure A.4D, G) to classify poses in validation set 1 and 2. In general, however, RMSDs for validation set 1 and 2 were higher than the corresponding leave-one-out values, and the success rates were generally lower. For validation set 1, for instance, the lowest median RMSD was 1.25 Å (Table B.7), and the associated $S(1.0)$, $S(1.5)$, $S(2.0)$, and $S(2.5)$ were 38.9, 55.6, 55.6, and 61.1%, respectively (Table B.7). For validation set 2, the lowest median RMSD was 1.71 Å and $S(1.0)$, $S(1.5)$, $S(2.0)$, and $S(2.5)$ were 14.3, 23.8, 57.1, and 61.9% (Table B.8). Though median RMSD and success rates especially $S(2.0)$ and $S(2.5)$, for validation set 1 and 2 were similar, the distribution of RMSDs were narrower for validation set 1 (Figure A.4B-D) than set 2 (Figure A.4F-G), indicating that overall the classifiers performed better on validation set 1 than 2. The reason for this is that validation set 1 is composed of high resolution X-ray solved RNA-ligand complexes, whereas validation set 2 is mostly composed of low-resolution NMR solved RNA-ligand complex. The discrepancy between the results on validation set 1 and 2 could be a result of validation set 1 consisting of high-resolution structures. Furthermore, since the classifiers were trained on the leave-one-out training set which only consists of X-ray structures, it is not surprising that they performs better on X-ray structures.

In Figure A.5 and A.6 we show comparisons between actual and predicted poses for RNA-ligand complexes in validation set 1 and 2. This visual comparison revealed that even in cases where RMSD was high (in particular, for validation set 2), the predicted poses were typically within the known binding pocket. This is significant because when we constructed the decoys sets, we made a deliberate effort to ensure that included within the decoys sets were poses docking in pockets other than the known binding pocket. Collectively, these results suggest that the classifiers we trained using our pose fingerprint were capable of identifying native-like poses from diverse decoy poses.

One of the challenges in assessing machine learning models such as the classifiers we trained in this work is that, overlap between training and validation set can lead to overestimation of performance of the model. In our case, significant bias might be introduced into our assessment when ligand in a validation system is identical to a ligand in the training set. However, in both validation set 1 and 2, we observed several cases in which the ligand exhibit low similarity to ligand in the training set, yet the pose identified as best pose had a small RMSD relative to the corresponding crystal structure. Conversely, we also had examples of systems in which the ligand was identical to a ligand in the training set, yet the RMSD was high. These results suggest that chemical similarity to the ligand in the training set did not substantially bias our assessment our classifiers.

A.3.4. Comparisons to other knowledge-based scoring function.

DrugScoreRNA[177] and SPA-LN[180] are two RNA specific knowledge-based scoring functions that can be used for pose prediction. For these score functions, pair-potentials from pairwise atomic distances are fit to a pre-determined functional form (using the inverse Boltzmann relation), and their final score (the binding affinity), which is the sum over all

pair-potentials, can be used to rank poses. In contrast to knowledge-based potential, we do not assume functional relationships between pairwise distances and our target (the pose classification). For our classifiers, interactions between RNA atom and ligand atom pairs are analyzed and grouped based on different atom type combinations to form a set of pose fingerprints, and the final classification score is the majority vote from an ensemble of nonlinear machine learning models trained on a set of RNA-ligand complexes. For DrugScoreRNA and SPA-LN, all RNA-ligand interactions are described using the SYBYL atom types of the RNA and ligand atoms, respectively. Like DrugScoreRNA and SPA-LN, we use SYBYL atom types to describe the heavy atoms in the ligand. For RNA, however, we defined atom types based on their RNA residue name and atom names, not their SYBYL types. We note that in preliminary work, we found that the classifiers trained using SYBYL atom types to describe the RNA performed poorly relative to the classifiers trained using the residue and atom names to define atom types.

In Table A.3 is the comparison between our classifiers and DrugScoreRNA (see also Table B.30 for the RMSDs of individual RNAs). For DrugScoreRNA, the data is taken from Table B.5 in Supporting Information in DrugScoreRNA paper.[177] In general, the results obtained using our classifiers were superior to those obtained using DrugScoreRNA, despite DrugScoreRNA being parameterized using a much larger dataset (670 nucleic acid-ligand and -protein complexes versus our 80 RNA-ligand complexes). The median RMSD for our classifiers were 1.71 Å compared to 1.95 Å for DrugScoreRNA. The success rate at 2.5 Å (S(2.5)) of our classifier was higher than DrugScoreRNA (61.9% compared to 57.1%) and the S(1.0), S(1.5) and S(2.0) were identical to DrugScoreRNA (Table A.3).

Shown in Table A.4 are the results we obtained on the SPA-LN validation set, which was composed of 56 RNA-ligand complexes and were the same RNA-ligand complexes used to test the pose prediction accuracy of SPA-LN (testing dataset 3 in the SPA-LN article[180]).

Method	RMSD(Å)	$S(1.0)(\%)$	$S(1.5)(\%)$	$S(2.0)(\%)$	$S(2.5)(\%)$
RNAPosers	**1.71**	14.3	23.8	57.1	**61.9**
DrugScoreRNA	**1.95**	14.3	23.8	57.1	**57.1**

Table A.3: Comparing the recovery performance of RNAPosers to the recovery performance of DrugScoreRNA scoring function on a common dataset (validation set 2). The scores for DrugScoreRNA is taken from Table B.5 in Supporting Information of the literature.[177]

For this validation set, we trained a new set of classifiers on 130 RNA-ligand complexes, a subset of the 437 nucleic-acid complexes used to train the SPA-LN scoring function. Due to the lack of access to SPA-LN performances on individual RNAs, we were not able to carry out a structure-wise comparison but only the success rate at 2.5 Å; the success rate $S(2.5)$ was ~54% for SPA-LN, while for our classifiers it is ~62%. We note that although we made our comparison to other scoring functions based on the same set of RNA-ligand complexes, the coordinates for the poses differ from those original work of DrugScore^{RNA} and SPA-LN. Therefore, the comparisons presented above should not be regarded as direct comparisons. In order to facilitate future comparisons between various pose prediction methods, we have made all the coordinate data used in our study publicly available.

Method	RMSD(Å)	$S(1.0)(\%)$	$S(1.5)(\%)$	$S(2.0)(\%)$	$S(2.5)(\%)$
RNAPosers	1.92	26.8	37.5	50.0	**62.5**
SPA-LN	–	–	–	–	**54.0**

Table A.4: Median RMSD and success rates for systems in an additional validation set, which was comprised of 56 RNA-ligand complexes. These 56 RNA-ligand complexes correspond to a subset of RNA-ligand complexes that overlapped with testing dataset 3 in the SPA-LN publication[180]). The classifier used in this analysis was trained on a set of RNA-ligand complexes corresponding to a subset of SPA-LN training set. Listed are the results obtained when the best poses were selected using docking scores plus our pose fingerprint as learning features, with nativeness threshold set to 2.0Å.

A.3.5. Contacts with ribose atoms in adenine residues emerge as important pose prediction features.

Next, we explored the chemical implication of our machine learning model by examining which elements in our pose fingerprint were most important for pose prediction. To accomplish this, we carried out feature importance analysis on our pose fingerprints. Shown in Figure A.7 are the importance maps associated with specific RNA-ligand contact, which were defined by RNA atom names and residues types (ADE, GUA, CYT and URA) and SYBYL atom types in the ligand.

As might be expected, contacts with purine residues (Figure A.7A, B) exhibited larger relative importance scores than pyrimidines (Figure A.7C, D). We speculate that the importance scores are relatively higher for purine contacts because they provide more opportunities for stabilizing contacts via $\pi - \pi$ stacking and hydrogen bonding interactions. Consistent with our speculation, nucleobase atoms in the purines (N7 and C8 in ADE (Figure A.7A) and N1, N2, C2, N3 and C6 in GUA (Figure A.7B)) exhibit high importance scores. This is also shown in the atomwise importance scores for each residue (Figure A.8). Surprisingly, though, ligand contacts with the highest importance scores reside on the ribose of ADE, namely, O2' and C2' atoms (Figure A.8A). Intriguingly, previous analysis of RNA-ligand complexes in both ribosomal and non-ribosomal RNA identified several signature features of RNA-ligand interactions, among these was the presence of unusual pucker conformations in residues with the binding pockets.[157] It is possible that the apparent ribose hotspot on ADE (O2' and C2') is due to the presence of stabilizing ligand interactions with O2' on ADE that adopt unusual pucker conformations. We cannot, however, rule out that these apparent hotspots are artifacts of our classifiers or our training data.

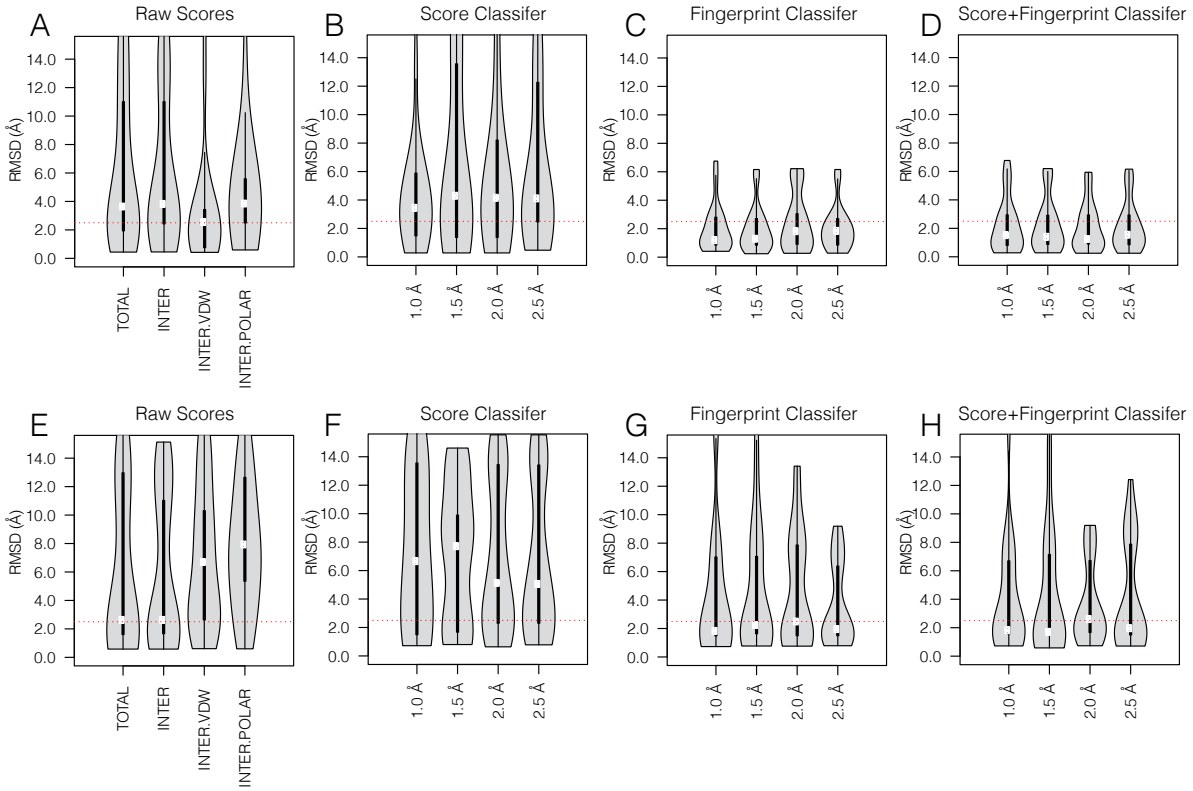


Figure A.4: Violin plots of the RMSD distributions of predicted best poses over systems in the validation set 1 (A-D) and 2 (E-H). Here best poses correspond to the top-scoring pose. When using raw docking scores, the top-scoring pose is the pose with the lowest docking scores. When using the classifiers, the top-scoring pose is the one with the highest classification score. Results shown here are RMSD distributions of best poses predicted using (A and E) docking score terms, classification scores from classifiers trained using (B and F) docking score terms, (C and G) our pose fingerprint, (D and H) raw docking scores plus our pose fingerprint as features, respectively. Here, TOTAL, INTER, INTER.VDW, and INTER.POLAR refer to various terms in the rDock scoring function: TOTAL, corresponds to total docking score, containing both RNA-ligand and intra-ligand contributions); INTER, corresponds to the contribution of RNA-ligand interaction to the docking score; INTER.VDW correspond to the non-polar, van der Waals contribution to the interaction docking score; and INTER.POLAR correspond to the polar contribution to the interaction docking score. For the pose classifiers (B-D and F-H), 4 independent sets of classifiers were trained with nativeness threshold set to 1.0, 1.5, 2.0, and 2.5 Å and results are shown in the plots for side-by-side comparison. In each violin plot, the black bar in the center corresponds to quantile range, and the white squares are located at the corresponding median RMSD. For reference, the red dotted lines are placed at RMSD values of 2.50 Å.

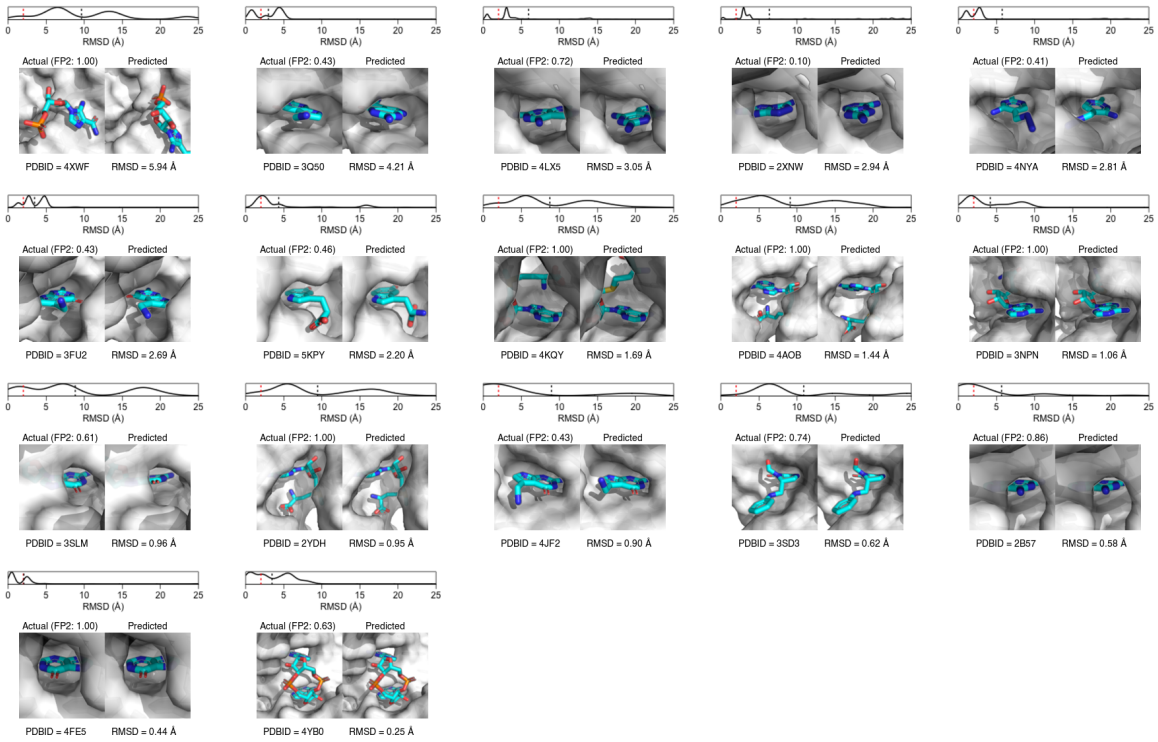


Figure A.5: Comparison between actual and predicted poses for validation set 1. Here, the poses were selected using the classifier trained using raw docking scores plus our pose fingerprint as features and the nativeness threshold was set to 2.0 Å. Shown for each case is the RMSD distribution over the decoy set, from which poses were selected. In each distribution plot, the black dotted line is placed at the mean RMSD value and, for reference, the red dotted line is placed at RMSD value 2.0 Å.

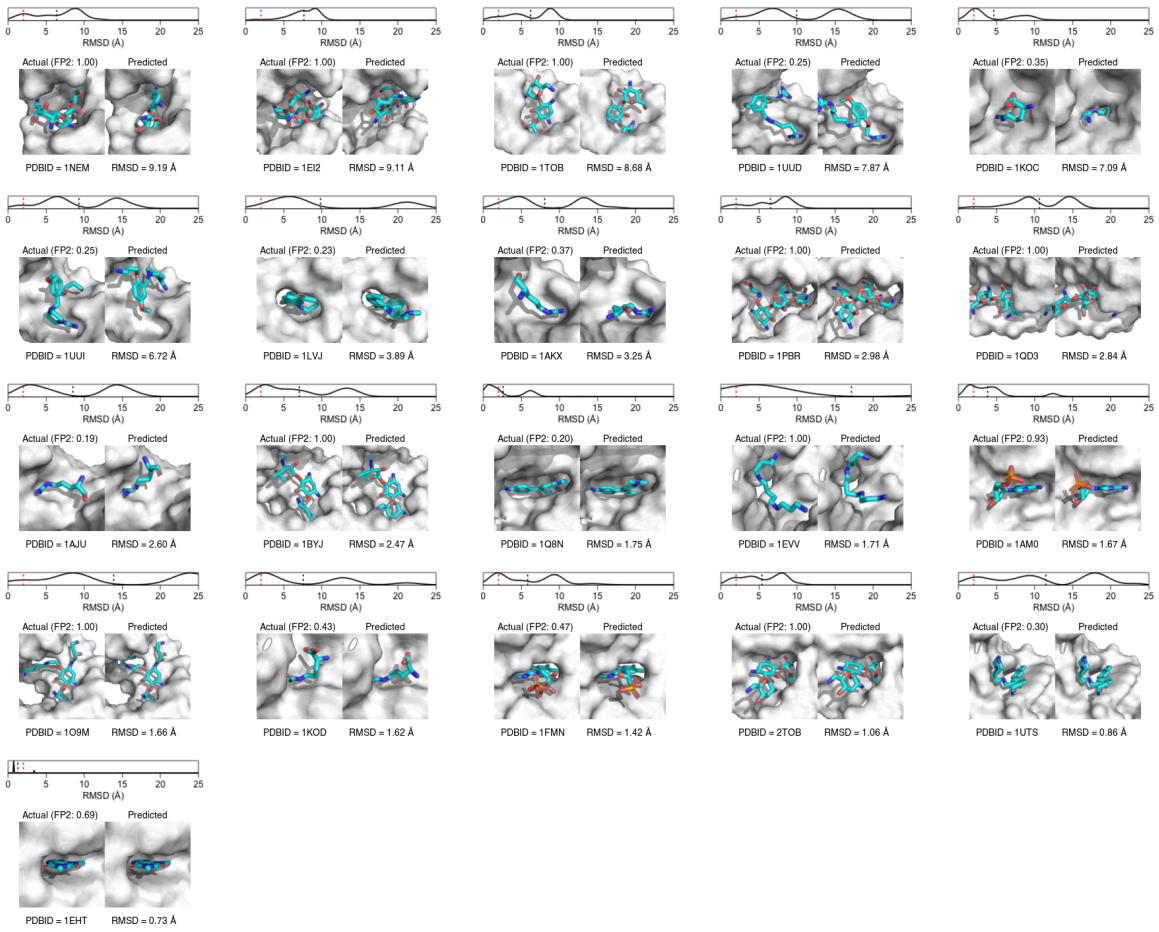


Figure A.6: Comparison between actual and predicted poses for validation set 2. Here, the poses were selected using the classifier trained using raw docking scores plus our pose fingerprint as features and the nativeness threshold was set to 2.0 Å. Shown for each case is the RMSD distribution over the decoy set, from which poses were selected. In each distribution plot, the black dotted line is placed at the mean RMSD value and, for reference, the red dotted line is placed at RMSD value 2.0 Å.

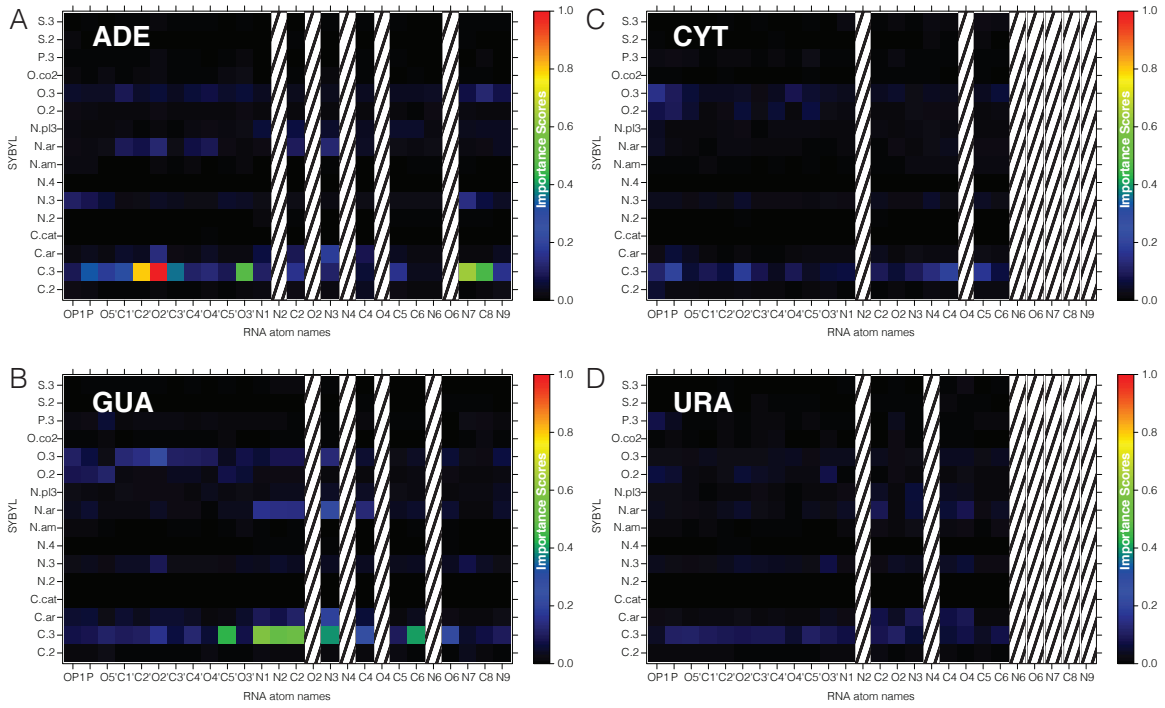


Figure A.7: Relative importance matrix for our pose fingerprint. Note that every element in the matrix correspond to unique RNA-ligand pairs, which are defined by the atoms names in ADE, GUA, CYT, and URA, respectively, and the SYBYL atom types for ligands. Results are shown for a random forest classifier trained on data for all 80 RNA-ligand in the original leave-one-out dataset and with nativeness threshold set to 2.5 Å. High values correspond to more important features.

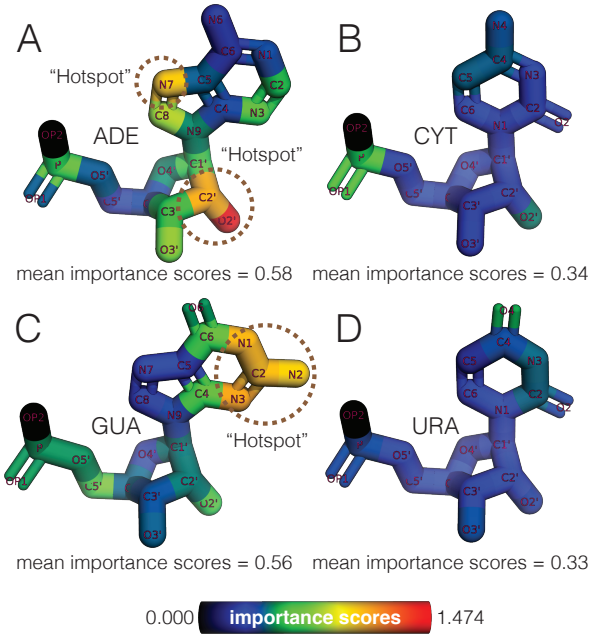


Figure A.8: Atomwise Importance Scores. Shown are cartoons of the importance scores projected onto the atoms in (A) ADE, (B) GUA, (C) CYT, and (D) URA residues. These scores correspond to the sum over the individual SYBYL atom types in the importance matrix (Figure A.7). The striped bars are placed at non-existent atoms in the residues.

A.4. Conclusion

In this study, we showed that machine learning classifiers, trained using a novel pose-fingerprint, was able to enhance RNA-ligand pose prediction over baseline docking scores. Due to the promising results we obtained using our pose classifiers, we have incorporated them into the software tool, RNAPosers (https://github.com/atfrank/RNAPosers), which consist of a commandline tool and an accompanying a PyMOL plugin. To facilitate the development and testing of other RNA-ligand pose prediction methods, we also made accessible coordinates of all the decoy sets used in this study (https://doi.org/10.5281/zenodo.3711071). In the context of RNA-ligand pose prediction, RNAPosers should find utility as a tool to assess the relative quality of a set of poses derived either from purely computational methods or from hybrid modeling methods that incorporate experimental data such as chemical shift perturbation data. Also, within the context of virtual screening, we envision that RNAPosers may find utility as a tool to identify high-confidence poses that can be brought forward for binding affinity prediction using physics-based free energy calculation methods like, MM-PBSA and FEP calculations as well as to facilitate structure-activity-relationship studies in the absence of experimental structural information.

A.5. Acknowledgements

We thank the members of the Frank lab for many useful discussion about this work. The authors were funded by University of Michigan research and computational startup funds. We also thank Dr. Sean M. Law who wrote the MoleTools (https://github.com/atfrank/MoleTools) library we used to implement the pose fingerprint used in this study.

Appendix B.

Supporting Information

B.1. SimRNA parameter setting and simulation details in Chapter IV.

SimRNA is an RNA 3D structure prediction tool, which simulates the RNA folding with a coarse-grained representation and generates a set of conformations using Monte Carlo sampling tools. The folding simulation was run for 64,000,000 steps, and structures were written out every 10,000 steps. Then for the simulated annealing, the initial temperature was set to 1.35 and the final temperature was set to 0.90. The weights for the bond, angles, η_θ terms in the coarse-grained force field employed in SimRNA was set to 1.00, 1.00, and 0.40, respectively. For our calculations, we ran SimRNA in replica exchange Monte Carlo mode, with the number of replicas set to 10. The resulting trajectories for each replica combined into a single trajectory, and the resulting collection of structures clustered with maximum RMSD threshold set to 5.0 Å. For each resulting cluster, the structure with the lowest energy was selected. After applying the protocol described above, a total of 300 low energy structures

were obtained and used as the final set of modeled structures.

B.2. Supporting tables

Table B.1: List of structures used for benchmarking the SASA-BME framework in ensemble reweighing.

PDB ID	Chain Length	Purine Length	Coverage
1KKA	17	10	0.58824
1L1W	29	17	0.58621
1LDZ	30	19	0.63333
1NC0	24	13	0.54167
1OW9	23	16	0.69565
1PJY	22	12	0.54545
1R7W	34	16	0.47059
1SCL	29	17	0.58621
1UUU	19	8	0.42105
1XHP	32	16	0.5
1YSV	27	14	0.51852
1ZC5	41	23	0.56098
2FDT	36	18	0.5
2JWV	29	16	0.55172
2K66	22	13	0.59091
2KOC	14	6	0.42857
2L3E	35	15	0.42857
2LBJ	17	8	0.47059
2LBL	17	7	0.41176
2LDT	31	18	0.58065
2LHP	37	19	0.51351
2LK3	24	14	0.58333
2LP9	16	9	0.5625
2LPA	15	8	0.53333
2LQZ	27	12	0.44444
2LU0	49	28	0.57143
2LUB	37	19	0.51351
2LUN	28	14	0.5
2LV0	24	14	0.58333
2M12	23	14	0.6087

PDB ID	Chain Length	Purine Length	Coverage
2M21	21	9	0.42857
2M22	23	13	0.56522
2M24	29	15	0.51724
2M4W	17	11	0.64706
2M5U	22	11	0.5
2M8K	48	18	0.375
2MEQ	19	10	0.52632
2MFD	19	9	0.47368
2MHI	53	24	0.45283
2MNC	29	14	0.48276
2N2O	23	11	0.47826
2N2P	23	11	0.47826
2QH2	24	12	0.5
2QH4	18	10	0.55556
2Y95	14	7	0.5

Table B.3 - B.16 listed the supporting tables for RNAPosers.

PDB ID	AUC	PDB ID	AUC	PDB ID	AUC	PDB ID	AUC
2GCS	0.886	3DIL	0.966	4DR6	0.789	4NXM	0.792
2GCV	0.964	3EGZ	0.834	4DR7	0.772	4NXN	0.737
2GDI	0.93	3G71	0.968	4DUY	0.785	4OJI	0.958
2H0S	1.0	3GX5	0.606	4DUZ	0.765	4PQV	1.0
2H0W	0.839	3HHN	0.86	4DV0	0.758	4Q9Q	0.701
2H0X	0.901	1K9M	0.841	4DV1	0.77	1QU2	0.715
2HO6	0.909	3IVK	0.921	4DV2	0.769	1QVF	0.886
1FFY	0.715	3L0U	0.842	4DV4	0.819	4QLM	0.945
2NZ4	0.967	3LA5	0.877	1NTB	0.306	1QVG	0.879
2OTJ	0.849	3MUM	0.297	4DV7	0.854	1S72	0.841
2OTL	0.835	1KD1	0.809	4E8M	0.769	4TRA	0.832
1FUF	0.989	3MUT	0.826	4E8N	1.0	4TZX	0.953
2QBZ	0.961	3MXH	0.983	4ENA	0.984	6TNA	0.912
2QEX	0.913	3NKB	0.972	4ENB	0.831	1TRA	0.959
1HR2	0.967	1KQS	0.868	4ENC	0.996	1VQ4	0.857
2R8S	0.978	1L2X	0.899	1NUJ	0.961	1VQ5	0.87
2YIE	0.783	3OWI	0.741	4FAW	0.557	1VQ6	0.866
2YIF	0.993	3OWW	0.826	1NUV	0.915	1VQ7	0.856
2Z75	0.979	3OWZ	0.851	4IFD	0.61	1VQ8	0.904
2ZXU	0.972	3OX0	0.825	4JF2	0.945	1VQ9	0.905
354D	0.973	3OXB	0.939	4JI0	0.781	1VQK	0.922
3B4A	0.998	3OXD	0.739	4JI1	0.834	1VQL	0.882
3B4B	0.896	3OXM	0.763	4JI2	0.789	1VQO	0.931
3B4C	1.0	3P49	0.874	4JI3	0.812	1VQP	0.888
3BO3	0.406	3RER	0.984	4JI4	0.803	1EHZ	0.981
3CC2	0.873	1LNG	0.983	4JI5	0.732	1X8W	0.984
1JJ2	0.897	3U2E	0.698	4JI6	0.788	1XJR	0.991
3CC7	0.862	3UCZ	0.965	4JI7	0.807	1YHQ	0.858
3CCJ	0.846	1M1K	0.821	4JI8	0.802	1EVV	0.981
3CCL	0.888	1M90	0.881	1PJO	1.0	1YI2	0.863
3CCM	0.919	3V7E	0.852	1Q7Y	0.865	1YIJ	0.857
3CCR	0.925	3ZGZ	0.958	4KZD	0.98	1YJ9	0.857
3CCU	0.869	462D	0.971	4LF7	0.646	1YJN	0.857
3CCV	0.919	1MMS	0.99	4LF8	0.646	1YLS	0.949
3CXC	0.902	1N78	0.919	4LF9	0.607	2A43	0.927
3D2G	0.913	4DR2	0.818	4LFB	0.793	1FEU	0.943
3D2V	0.961	4DR3	0.763	1Q81	0.885	1K73	0.915
3D2X	0.373	4DR4	0.788	4M30	0.933	4DR5	0.858
3DD2	0.87	1NJI	0.868	4M4O	0.995	1Q82	0.914

Table B.2: Training dataset and leave-one-out validation results for the Mg^{2+} binding site classifiers. Listed are the structure PDB IDs and the corresponding AUC when the structure is used as validation in the leave-one-out analysis.

Atom Types									
Ligand (SYBYL)		C.1	C.2	C.3	C.ar	C.cat	N.1	N.2	N.3	N.4
		N.ar	N.am	N.pl3	O.2	O.3	O.co2	S.2	S.3	S.o
		S.o2	P.3							
RNA	ADE	C1′	C2	C2′	C3′	C4	C4′	C5	C5′	C6
		C8	N1	N3	N6	N7	N9	O2′	O3′	O4′
		O5′	OP1	OP2	P					
	CYT	C1′	C2	C2′	C3′	C4	C4′	C5	C5′	C6
		N1	N3	N4	O2	O2′	O3′	O4′	O5′	OP1
		OP2	P							
	GUA	C1′	C2	C2′	C3′	C4	C4′	C5	C5′	C6
		C8	N1	N2	N3	N7	N9	O2′	O3′	O4′
		O5′	O6	OP1	OP2	P				
	URA	C1′	C2	C2′	C3′	C4	C4′	C5	C5′	C6
		N1	N3	O2	O2′	O3′	O4	O4′	O5′	OP1
		OP2	P							

Table B.3: Atom types considered in pose fingerprints (FPs). We used 20 SYBYL atom types for the ligand, and 85 RNA atom types (we consider unique combinations of residues and atom types) for the RNA, resulting in a total of 1700 pair of interactions.

Dataset	PDB IDs
Training (80)	1FUF 1LC4 1ZZ5 1F27 1J7T 1MWL 1NTA 1NTB 1O9M 1U8D 1YRJ 2EEU 2EEV 2EEW 2ET5 2HOJ 2HOM 2HOO 2O3V 2O3Y 2QWY 2BE0 2BEE 2ET3 2ET4 2ET8 2F4S 2F4T 2F4U 2FCX 2FCZ 2G5Q 2O3W 2O3X 2OE5 2OE8 3D2G 3D2V 3D2X 3E5E 3F2Q 3F2T 3F4H 3GX2 3GX3 3GX5 3GX6 3GX7 3IQN 3IQR 3NPQ 3SUH 3TZR 3WRU 3C44 4F8U 4F8V 4FAW 4GPX 4GPY 4LVX 4LVY 4LW0 4P20 4PDQ 4QLM 4QLN 4TS2 4TZX 4TZY 4B5R 4K32 4WCR 4ZNP 5BTP 5BWS 5BXK 5C45 5KX9 6BFB
Validation 1 (17)	2XNW 3FU2 3Q50 3SD3 3SLM 4FE5 4JF2 4LX5 4NYA 4XWF 4YB0 2B57 2YDH 3NPN 4AOB 4KQY 5KPY
Validation 2 (21)	1AJU 1AKX 1AM0 1BYJ 1EHT 1EI2 1EVV 1FMN 1KOC 1KOD 1LVJ 1NEM 1O9M 1PBR 1Q8N 1QD3 1TOB 1UTS 1UUD 1UUI 2TOB

Table B.4: PDB IDs for the leave-one-out training dataset and validation datasets 1 and 2.

Dataset	PDB IDs
SPA-LN Training (130)	1DDY 1ET4 1EVV 1FUF 1TN1 1TN2 1YLS 1ZZ5 2B57 2CKY 2EES 2EET 2EEU 2EEV 2EEW 2ET5 2GDI 2GIS 2HO6 2HO7 2HOJ 2HOL 2HOM 2HOO 2L1V 2L94 2LWK 2M4Q 2MIY 2MXS 2N0J 2NPY 2O3Y 2QWY 3B4A 3B4B 3B4C 3D0U 3D2G 3D2V 3D2X 3DIG 3DIL 3DIM 3DIX 3DIY 3DIZ 3DJ0 3DJ2 3DVV 3E5E 3E5F 3F2Q 3F2T 3F4H 3FAW 3GCA 3GS8 3GX2 3GX3 3GX5 3GX6 3GX7 3IQN 3IQR 3NPQ 3RKF 3SKI 3SKL 3SLQ 3SUH 3SUX 3TD1 3TZR 3WRU 4B5R 4E8N 4E8Q 4F8U 4F8V 4FAW 4FEJ 4FEL 4FEN 4FEO 4FEP 4GPW 4GPX 4GPY 4JIY 4K32 4L81 4LVV 4LVW 4LVX 4LVY 4LVZ 4LW0 4LX5 4LX6 4NYA 4NYD 4NYG 4P20 4P3S 4P5J 4P95 4PDQ 4QLM 4QLN 4RZD 4TS2 4TZX 4TZY 4WCR 4XNR 4XW7 4XWF 4Y1I 4YAZ 4YB0 4ZNP 5BTP 5BWS 5BXK 5C45 5KX9 5LWJ 5UZA 6BFB
SPA-LN Validation (56)	1AJU 1AKX 1AM0 1BYJ 1EHT 1EI2 1F1T 1F27 1FMN 1FYP 1J7T 1KOC 1KOD 1LC4 1LVJ 1MWL 1NEM 1NTA 1NTB 1NYI 1O15 1O9M 1PBR 1Q8N 1QD3 1TOB 1U8D 1UTS 1UUD 1UUI 1XPF 1Y26 1YRJ 2AU4 2BE0 2BEE 2ESJ 2ET3 2ET4 2ET8 2F4S 2F4T 2F4U 2FCX 2FCY 2FCZ 2FD0 2G5Q 2JUK 2O3V 2O3W 2O3X 2OE5 2OE8 2TOB 3C44

Table B.5: PDB IDs for SPA-LN training and validation set used for comparison to the SPA-LN scoring function.

In Tables B.6-B.9, median RMSD and success rates for systems in the training set and validation sets are listed. In each table we reported the results obtained when the best poses were selected using docking score terms and classifiers that trained using docking score terms, our pose fingerprint, and docking scores plus our pose fingerprint as learning features. For the pose classifiers, we include results for classifiers that we trained with the nativeness threshold set to 1.0, 1.5, 2.0, and 2.5 Å.

Selection Metric	*RMSD*(Å)	*S*(1.00)(%)	*S*(1.50)(%)	*S*(2.00)(%)	*S*(2.50)(%)
TOTAL	3.41	22.7	29.5	37.5	42.0
INTER	5.72	17.0	21.6	27.3	33.0
INTER.VDW	4.75	18.2	22.7	28.4	36.4
INTER.POLAR	6.88	8.0	9.1	12.5	21.6
Score CL (1.0 Å)	2.50	21.6	31.8	44.3	50.0
Score CL (1.5 Å)	3.14	18.2	28.4	40.9	47.7
Score CL (2.0 Å)	2.08	21.6	36.4	50.0	54.5
Score CL (2.5 Å)	2.14	25.0	37.5	48.9	54.5
Fingerpint CL (1.0 Å)	1.36	27.3	56.8	70.5	79.5
Fingerpint CL (1.5 Å)	1.27	37.5	63.6	77.3	86.4
Fingerpint CL (2.0 Å)	1.31	34.1	59.1	78.4	85.2
Fingerpint CL (2.5 Å)	1.42	33.0	58.0	77.3	86.4
Score+Fingerpint CL (1.0 Å)	**1.05**	**43.2**	**70.5**	**79.5**	**85.2**
Score+Fingerpint CL (1.5 Å)	1.17	40.9	67.0	80.7	88.6
Score+Fingerpint CL (2.0 Å)	1.15	42.0	65.9	78.4	88.6
Score+Fingerpint CL (2.5 Å)	1.20	36.4	64.8	80.7	86.4

Table B.6: Median RMSD and success rates for leave-one-out training set

Selection Metric	$RMSD$(Å)	S(1.00)(%)	S(1.50)(%)	S(2.00)(%)	S(2.50)(%)
TOTAL	3.07	17.6	17.6	29.4	35.3
INTER	3.07	17.6	23.5	23.5	29.4
INTER.VDW	2.42	29.4	41.2	47.1	52.9
INTER.POLAR	3.85	11.8	11.8	17.6	23.5
Score CL (1.0 Å)	3.09	23.5	29.4	29.4	29.4
Score CL (1.5 Å)	3.95	17.6	29.4	29.4	29.4
Score CL (2.0 Å)	3.95	11.8	29.4	29.4	29.4
Score CL (2.5 Å)	3.58	11.8	23.5	23.5	29.4
Fingerpint CL (1.0 Å)	1.11	47.1	58.8	64.7	70.6
Fingerpint CL (1.5 Å)	1.05	47.1	52.9	64.7	76.5
Fingerpint CL (2.0 Å)	1.44	35.3	52.9	52.9	58.8
Fingerpint CL (2.5 Å)	1.80	29.4	47.1	58.8	76.5
Score+Fingerpint CL (1.0 Å)	1.11	47.1	52.9	58.8	64.7
Score+Fingerpint CL (1.5 Å)	1.27	35.3	52.9	58.8	64.7
Score+Fingerpint CL (2.0 Å)	**1.44**	**41.2**	**52.9**	**58.8**	**64.7**
Score+Fingerpint CL (2.5 Å)	1.54	35.3	47.1	58.8	70.6

Table B.7: Median RMSD and success rates for systems in validation set 1.

Selection Metric	$RMSD$(Å)	S(1.00)(%)	S(1.50)(%)	S(2.00)(%)	S(2.50)(%)
TOTAL	2.63	19.0	23.8	33.3	38.1
INTER	2.63	14.3	23.8	28.6	42.9
INTER.VDW	6.70	4.8	4.8	4.8	19.0
INTER.POLAR	7.93	4.8	4.8	9.5	9.5
Score CL (1.0 Å)	6.66	9.5	28.6	33.3	33.3
Score CL (1.5 Å)	7.71	14.3	19.0	28.6	28.6
Score CL (2.0 Å)	5.13	4.8	14.3	23.8	28.6
Score CL (2.5 Å)	5.06	14.3	14.3	23.8	33.3
Fingerpint CL (1.0 Å)	1.83	9.5	28.6	52.4	57.1
Fingerpint CL (1.5 Å)	2.24	14.3	19.0	47.6	57.1
Fingerpint CL (2.0 Å)	2.49	9.5	28.6	47.6	52.4
Fingerpint CL (2.5 Å)	1.96	14.3	28.6	52.4	52.4
Score+Fingerpint CL (1.0 Å)	1.82	14.3	19.0	52.4	57.1
Score+Fingerpint CL (1.5 Å)	**1.71**	**14.3**	**23.8**	**57.1**	**61.9**
Score+Fingerpint CL (2.0 Å)	2.60	9.5	19.0	42.9	47.6
Score+Fingerpint CL (2.5 Å)	1.96	14.3	33.3	52.4	52.4

Table B.8: Median RMSD and success rates for systems in validation set 2.

Selection Metric	$RMSD$(Å)	S(1.00)(%)	S(1.50)(%)	S(2.00)(%)	S(2.50)(%)
TOTAL	3.29	19.6	26.8	32.1	33.9
INTER	5.64	14.3	19.6	23.2	30.4
INTER.VDW	6.77	7.1	8.9	8.9	19.6
INTER.POLAR	7.89	3.6	3.6	5.4	7.1
Score CL (1.0 Å)	4.21	10.7	21.4	25.0	30.4
Score CL (1.5 Å)	4.15	14.3	21.4	30.4	42.9
Score CL (2.0 Å)	4.86	16.1	21.4	28.6	37.5
Score CL (2.5 Å)	3.62	7.1	14.3	23.2	37.5
Fingerpint CL (1.0 Å)	1.84	30.4	37.5	55.4	55.4
Fingerpint CL (1.5 Å)	1.80	30.4	42.9	53.6	62.5
Fingerpint CL (2.0 Å)	1.92	23.2	39.3	50.0	66.1
Fingerpint CL (2.5 Å)	1.85	26.8	41.1	53.6	57.1
Score+Fingerpint CL (1.0 Å)	**1.56**	**37.5**	**48.2**	**55.4**	**62.5**
Score+Fingerpint CL (1.5 Å)	1.62	33.9	46.4	55.4	62.5
Score+Fingerpint CL (2.0 Å)	1.92	26.8	37.5	50.0	62.5
Score+Fingerpint CL (2.5 Å)	1.74	30.4	42.9	57.1	58.9
SPA-LN	–	–	–	–	~54.0

Table B.9: Median RMSD and success rates for systems in SPA-LN validation set, which was comprised of data for set 56 RNA-ligand complexes. These 56 RNA-ligand complexes correspond to RNA-ligand complexes in testing dataset 3 in the SLA-LN publication[180]). The classifiers used in this analysis were trained on a seperate training set, consistent with the training set of SPA-LN (Table B.5).

PDB IDs	RMSD (Å)	PDB IDs	RMSD (Å)	PDB IDs	RMSD (Å)
1F27	1.19	2HOM	1.10	3WRU	1.57
1FUF	1.44	2HOO	1.44	4B5R	0.67
1J7T	1.44	2O3V	2.66	4F8U	0.82
1LC4	0.75	2O3W	2.61	4F8V	0.63
1MWL	0.95	2O3X	1.62	4FAW	4.20
1NTA	4.28	2O3Y	3.88	4GPW	0.63
1NTB	0.79	2OE5	1.21	4GPX	0.66
1O9M	0.64	2OE8	0.91	4GPY	0.40
1U8D	0.22	2QWY	1.08	4K32	0.77
1YRJ	9.99	3B4B	1.41	4L81	0.73
1ZZ5	2.17	3B4C	0.64	4LVX	5.93
2BE0	10.47	3C44	0.96	4LVY	0.94
2BEE	2.72	3D2G	1.92	4LW0	1.01
2CKY	1.69	3D2V	2.12	4P20	0.97
2EEU	0.87	3D2X	1.03	4P3S	0.89
2EEV	0.42	3E5E	7.24	4PDQ	0.96
2EEW	1.00	3F2Q	1.07	4QLM	1.10
2ET3	1.01	3F2T	0.98	4QLN	2.10
2ET4	2.68	3F4H	0.99	4TS2	1.04
2ET5	0.68	3GX2	0.92	4TZX	0.38
2ET8	1.19	3GX3	0.74	4TZY	0.46
2F4S	0.63	3GX5	1.31	4WCR	2.23
2F4T	1.05	3GX6	1.41	4YAZ	1.70
2F4U	0.88	3GX7	0.76	4ZNP	6.16
2FCX	1.29	3IQN	0.84	5BTP	2.42
2FCY	1.69	3IQR	1.92	5BWS	1.18
2FCZ	1.62	3NPQ	1.17	5BXK	0.92
2G5Q	1.01	3SUH	1.26	5C45	0.52
2HOJ	2.52	3TZR	1.04	5KX9	0.49
6BFB	0.80				

Table B.10: Top-scored poses RMSD for each system in leave-one-out training set.

In Tables B.11-B.13, the maximum chemical similarity between the ligands in the validation sets and training set are listed, in addition to their top-scored poses RMSD using RNAPosers and/or DrugScore^{RNA} when available. For validation set 1 and 2, the relevant training set is that listed in Table B.4. For SPA-LN validation set, the relevant training set is that listed in Table B.5. For the ligand in each system in the validation set, we calculated the chemical similarity to all the ligands in the relevant training set as the maximum Tanimoto score between their chemical fingerprints. The Tanimoto scores were computed using FP2 fingerprint with OpenBabel.[190]

PDB IDs	Chem. Similarity	RMSD (Å)
2XNW	0.097	2.94
3FU2	0.434	2.69
3Q50	0.434	4.21
3SD3	0.744	0.61
3SLM	0.613	0.96
4FE5	1.000	0.49
4JF2	0.434	0.90
4LX5	0.723	3.05
4NYA	0.409	2.81
4XWF	1.000	5.94
4YB0	0.631	0.27
2B57	0.857	0.58
2YDH	1.000	0.95
3NPN	1.000	1.06
4AOB	1.000	1.44
4KQY	1.000	1.04
5KPY	0.464	2.20

Table B.11: Chemical, sequence similarity in training set for systems in validation set 1.

PDB IDs	Chem. Similarity	RNAPosers RMSD (Å)	DrugScore^{RNA} RMSD (Å)
1AJU	0.194	15.86	7.32
1AKX	0.371	13.14	7.04
1AM0	0.932	1.56	2.86
1BYJ	1.000	1.53	1.99
1EHT	0.691	0.79	1.95
1EI2	1.000	7.61	0.84
1EVV	1.000	1.71	10.29
1FMN	0.474	1.56	1.55
1KOC	0.352	7.15	1.61
1KOD	0.435	1.23	1.87
1LVJ	0.228	3.91	3.09
1NEM	1.000	9.19	0.66
1O9M	1.000	0.62	8.69
1PBR	1.000	1.02	1.05
1Q8N	0.205	1.80	3.65
1QD3	1.000	0.58	0.88
1TOB	1.000	2.29	1.52
1UTS	0.299	1.62	11.04
1UUD	0.246	7.87	1.59
1UUI	0.246	6.72	5.56
2TOB	1.000	1.67	1.45

Table B.12: Chemical, sequence similarity to training set, RMSDs of top-scored poses obtained using RNAPosers and DrugScore^{RNA} for systems in validation set 2.

IDs	Chem. Similarity	RMSD (Å)	IDs	Chem. Similarity	RMSD (Å)
1AJU	0.220	16.59	1UUD	0.333	2.5
1AKX	0.455	13.59	1UUI	0.323	4.6
1AM0	0.932	1.26	1XPF	1.000	1.59
1BYJ	1.000	1.52	1Y26	1.000	0.56
1EHT	0.691	0.74	1YRJ	1.000	10.34
1EI2	1.000	7.61	2AU4	0.286	5.1
1F1T	0.494	0.62	2BE0	0.623	8.44
1F27	0.361	3.18	2BEE	0.872	2.03
1FMN	0.534	1.42	2ESJ	1.000	3.24
1FYP	1.000	3.41	2ET3	1.000	0.87
1J7T	1.000	1.37	2ET4	1.000	0.61
1KOC	0.319	18.08	2ET8	0.889	0.56
1KOD	0.350	4.62	2F4S	0.889	0.54
1LC4	0.944	0.81	2F4T	0.800	0.9
1LVJ	0.249	3.48	2F4U	0.943	0.88
1MWL	1.000	0.53	2FCX	0.889	0.96
1NEM	1.000	0.88	2FCY	1.000	1.69
1NTA	0.821	8.44	2FCZ	1.000	1.08
1NTB	0.821	1.46	2FD0	1.000	11.6
1NYI	0.942	6.09	2G5Q	1.000	0.87
1O15	0.691	1.18	2JUK	0.874	18.69
1O9M	0.943	0.64	2O3V	0.889	5.72
1PBR	1.000	8.45	2O3W	1.000	8.58
1Q8N	0.375	0.7	2O3X	1.000	2.44
1QD3	1.000	8.86	2OE5	1.000	0.97
1TOB	0.944	2.19	2OE8	1.000	0.71
1U8D	1.000	0.2	2TOB	0.944	0.95
1UTS	0.305	1.68	3C44	1.000	0.96

Table B.13: Chemical, sequence similarity to SPA-LN training set and RMSD of top-scored poses obtained using RNAPosers for systems in SPA-LN validation set.

Results using less stringent criterion for success

The results presented in the main manuscript were obtained by selecting the best (top 1) poses. It is customary to present the results obtained when selecting the best pose among that top N, where N is typical ranged between 2-5. Below are the results we obtained for $N = 3$.

Selection Metric	$RMSD$(Å)	S(1.00)(%)	S(1.50)(%)	S(2.00)(%)	S(2.50)(%)
TOTAL	2.38	28.4	34.1	46.6	52.3
INTER	2.57	23.9	34.1	42.0	47.7
INTER.VDW	2.54	23.9	30.7	40.9	48.9
INTER.POLAR	3.54	18.2	26.1	35.2	39.8
Score CL (1.0 Å)	1.64	30.7	47.7	60.2	64.8
Score CL (1.5 Å)	1.76	26.1	45.5	61.4	64.8
Score CL (2.0 Å)	1.51	28.4	47.7	59.1	63.6
Score CL (2.5 Å)	1.63	29.5	47.7	62.5	67.0
Fingerpint CL (1.0 Å)	1.15	40.9	64.8	79.5	86.4
Fingerpint CL (1.5 Å)	1.00	52.3	73.9	87.5	94.3
Fingerpint CL (2.0 Å)	1.00	51.1	73.9	85.2	90.9
Fingerpint CL (2.5 Å)	0.98	54.5	72.7	86.4	92.0
Score+Fingerpint CL (1.0 Å)	0.96	61.4	77.3	84.1	90.9
Score+Fingerpint CL (1.5 Å)	0.98	54.5	77.3	86.4	90.9
Score+Fingerpint CL (2.0 Å)	0.98	54.5	72.7	84.1	90.9
Score+Fingerpint CL (2.5 Å)	0.99	52.3	73.9	86.4	92.0

Table B.14: Median RMSD and success rates for systems in the leave-one-out set when selecting the best among the top 3 poses. For the pose classifiers, we include results for classifiers that we trained with the nativeness threshold set to 1.0, 1.5, 2.0, and 2.5 Å.

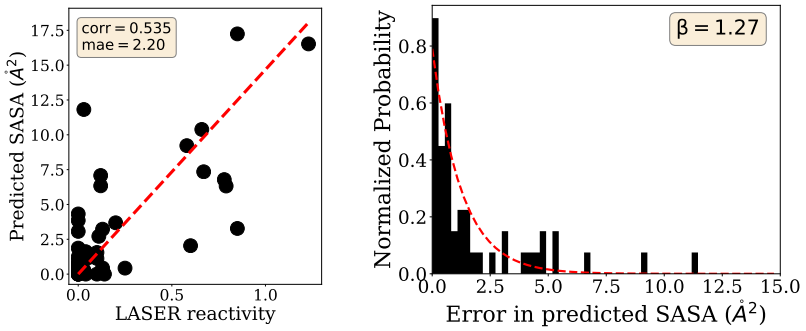
Selection Metric	*RMSD*(Å)	*S*(1.00)(%)	*S*(1.50)(%)	*S*(2.00)(%)	*S*(2.50)(%)
TOTAL	2.80	17.6	23.5	35.3	41.2
INTER	2.81	17.6	23.5	23.5	29.4
INTER.VDW	1.42	35.3	52.9	52.9	58.8
INTER.POLAR	1.92	29.4	41.2	52.9	58.8
Score CL (1.0 Å)	1.39	23.5	52.9	52.9	52.9
Score CL (1.5 Å)	2.82	17.6	35.3	35.3	35.3
Score CL (2.0 Å)	2.81	17.6	41.2	41.2	41.2
Score CL (2.5 Å)	3.05	17.6	29.4	35.3	41.2
Fingerpint CL (1.0 Å)	1.00	52.9	64.7	64.7	70.6
Fingerpint CL (1.5 Å)	1.05	47.1	64.7	64.7	76.5
Fingerpint CL (2.0 Å)	1.03	47.1	64.7	64.7	70.6
Fingerpint CL (2.5 Å)	1.49	35.3	52.9	64.7	76.5
Score+Fingerpint CL (1.0 Å)	1.06	47.1	58.8	58.8	64.7
Score+Fingerpint CL (1.5 Å)	1.05	47.1	58.8	58.8	64.7
Score+Fingerpint CL (2.0 Å)	1.00	52.9	64.7	64.7	70.6
Score+Fingerpint CL (2.5 Å)	1.39	35.3	58.8	64.7	76.5

Table B.15: Median RMSD and success rates for systems in validation set 1 when selecting the best among the top 3 poses. For the pose classifiers, we include results for classifiers that we trained with the nativeness threshold set to 1.0, 1.5, 2.0, and 2.5 Å.

Selection Metric	*RMSD*(Å)	*S*(1.00)(%)	*S*(1.50)(%)	*S*(2.00)(%)	*S*(2.50)(%)
TOTAL	2.45	19.0	33.3	42.9	52.4
INTER	2.50	19.0	33.3	42.9	52.4
INTER.VDW	4.69	14.3	14.3	33.3	38.1
INTER.POLAR	5.03	4.8	19.0	33.3	42.9
Score CL (1.0 Å)	2.23	19.0	38.1	42.9	52.4
Score CL (1.5 Å)	3.73	23.8	28.6	33.3	42.9
Score CL (2.0 Å)	3.53	9.5	23.8	33.3	42.9
Score CL (2.5 Å)	5.06	19.0	23.8	33.3	38.1
Fingerpint CL (1.0 Å)	1.72	9.5	42.9	57.1	61.9
Fingerpint CL (1.5 Å)	1.92	19.0	42.9	52.4	57.1
Fingerpint CL (2.0 Å)	2.49	14.3	38.1	47.6	52.4
Fingerpint CL (2.5 Å)	1.76	14.3	42.9	52.4	52.4
Score+Fingerpint CL (1.0 Å)	1.80	23.8	38.1	52.4	61.9
Score+Fingerpint CL (1.5 Å)	1.56	28.6	42.9	61.9	66.7
Score+Fingerpint CL (2.0 Å)	1.59	14.3	42.9	52.4	57.1
Score+Fingerpint CL (2.5 Å)	1.75	14.3	42.9	52.4	52.4

Table B.16: Median RMSD and success rates for systems in validation set 2 when selecting the best among the top 3 poses. For the pose classifiers, we include results for classifiers that we trained with the nativeness threshold set to 1.0, 1.5, 2.0, and 2.5 Å.

B.3. Supporting figures



(a) (b)

Figure B.1: Linear Fit of C8-SASA to LASER reactivity. (a) the linear fit of C8-SASA to LASER reactivity. (b) exponential distribution of absolute error of the linear fit

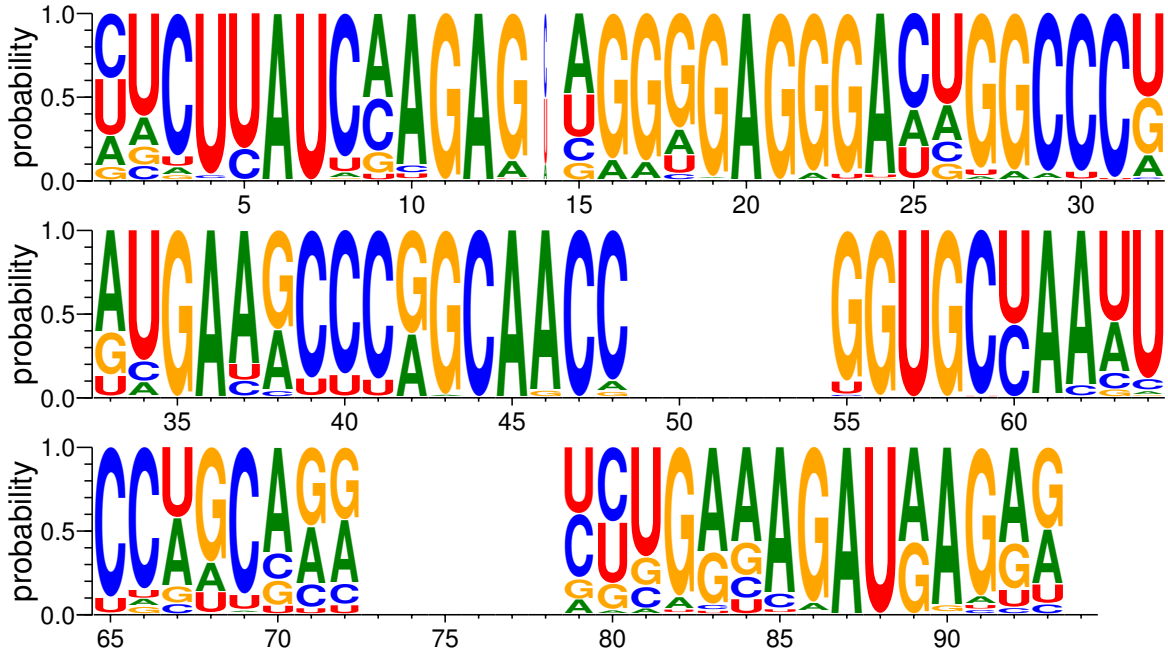


Figure B.2: Conservation map for SAM-I riboswitch. Credit to Dr. Indrajit Deb.

Bibliography

[1] S. E. Butcher and A. M. Pyle. “The molecular interactions that stabilize RNA tertiary structure: RNA motifs, patterns, and networks”. In: *Accounts of chemical research* 44.12 (2011), pp. 1302–1311.

[2] C. S. Chow and F. M. Bogdan. “A structural basis for RNA- ligand interactions”. In: *Chemical Reviews* 97.5 (1997), pp. 1489–1514.

[3] M. Chawla, S. Abdel-Azeim, R. Oliva, and L. Cavallo. “Higher order structural effects stabilizing the reverse Watson–Crick Guanine-Cytosine base pair in functional RNAs”. In: *Nucleic acids research* 42.2 (2014), pp. 714–726.

[4] J. Cabello-Villegas, M. E. Winkler, and E. P. Nikonowicz. “Solution conformations of unmodified and A37N6-dimethylallyl modified anticodon stem-loops of Escherichia coli tRNAPhe”. In: *Journal of molecular biology* 319.5 (2002), pp. 1015–1034.

[5] P. Kerpedjiev, S. Hammer, and I. L. Hofacker. “Forna (force-directed RNA): simple and effective online RNA secondary structure diagrams”. In: *Bioinformatics* 31.20 (2015), pp. 3377–3379.

[6] L. Schrodinger. “The PyMOL molecular graphics system, version 1.8”. In: *Schrodinger LLC, New York, NY* (2015).

[7] K. B. Hall. “RNA does the folding dance of twist, turn, stack”. In: *Proceedings of the National Academy of Sciences* 110.42 (2013), pp. 16706–16707.

[8] E. A. Dethoff, K. Petzold, J. Chugh, A. Casiano-Negroni, and H. M. Al-Hashimi. “Visualizing transient low-populated structures of RNA”. In: *Nature* 491.7426 (2012), pp. 724–728.

[9] C. E. Hajdin, F. Ding, N. V. Dokholyan, and K. M. Weeks. “On the significance of an RNA tertiary structure prediction”. In: *Rna* 16.7 (2010), pp. 1340–1349.

[10] D. R. Bell, S. Y. Cheng, H. Salazar, and P. Ren. “Capturing RNA folding free energy with coarse-grained molecular dynamics simulations”. In: *Scientific reports* 7 (2017), p. 45812.

[11] K. K. Q. Nguyen, Y. K. Gomez, M. Bakhom, A. Radcliffe, P. La, D. Rochelle, J. W. Lee, and E. J. Sorin. “Ensemble simulations: folding, unfolding and misfolding of a high-efficiency frameshifting RNA pseudoknot”. In: *Nucleic acids research* 45.8 (2017), pp. 4893–4904.

[12] C. Guilbert and T. 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.

[13] B. Lee and F. M. Richards. “The interpretation of protein structures: estimation of static accessibility”. In: *Journal of molecular biology* 55.3 (1971), 379–IN4.

[14] C. J. Alden and S.-H. Kim. “Solvent-accessible surfaces of nucleic acids”. In: *Journal of molecular biology* 132.3 (1979), pp. 411–434.

[15] C. M. Livi and E. Blanzieri. “Protein-specific prediction of mRNA binding using RNA sequences, binding motifs and predicted secondary structures”. In: *BMC bioinformatics* 15.1 (2014), p. 123.

[16] C. R. Munteanu, A. C. Pimenta, C. Fernandez-Lozano, A. Melo, M. N. Cordeiro, and I. S. Moreira. “Solvent accessible surface area-based hot-spot detection methods for protein–protein and protein–nucleic acid interfaces”. In: *Journal of chemical information and modeling* 55.5 (2015), pp. 1077–1086.

[17] S. Ruiz-Carmona, D. Alvarez-Garcia, N. Foloppe, A. B. Garmendia-Doval, S. Juhos, P. Schmidtke, X. Barril, R. E. Hubbard, and S. D. Morley. “rDock: a fast, versatile and open source program for docking ligands to proteins and nucleic acids”. In: *PLoS computational biology* 10.4 (2014), e1003571.

[18] J. Weiser, P. S. Shenkin, and W. C. Still. “Approximate atomic surfaces from linear combinations of pairwise overlaps (LCPO)”. In: *Journal of Computational Chemistry* 20.2 (1999), pp. 217–230.

[19] N. Hansen and W. F. Van Gunsteren. “Practical aspects of free-energy calculations: a review”. In: *Journal of chemical theory and computation* 10.7 (2014), pp. 2632–2647.

[20] E. Westhof. “Twenty years of RNA crystallography”. In: *Rna* 21.4 (2015), pp. 486–487.

[21] P. Tijerina, S. Mohr, and R. Russell. “DMS footprinting of structured RNAs and RNA–protein complexes”. In: *Nature protocols* 2.10 (2007), p. 2608.

[22] E. J. Merino, K. A. Wilkinson, J. L. Coughlan, and K. M. Weeks. “RNA structure analysis at single nucleotide resolution by selective 2 ‘-hydroxyl acylation and primer extension (SHAPE)”. In: *Journal of the American Chemical Society* 127.12 (2005), pp. 4223–4231.

[23] C. Feng, D. Chan, J. Joseph, M. Muuronen, W. H. Coldren, N. Dai, I. R. Corrêa Jr, F. Furche, C. M. Hadad, and R. C. Spitale. “Light-activated chemical probing of nucleobase solvent accessibility inside cells”. In: *Nature chemical biology* 14.3 (2018), p. 276.

[24] J. H. Cate and J. A. Doudna. "[12] Solving large RNA structures by X-ray crystallography". In: (2000).

[25] A. L. Edwards, A. D. Garst, and R. T. Batey. "Determining structures of RNA aptamers and riboswitches by X-ray crystallography". In: *Nucleic Acid and Peptide Aptamers*. Springer, 2009, pp. 135–163.

[26] A. Marchanka, B. Simon, G. Althoff-Ospelt, and T. Carlomagno. "RNA structure determination by solid-state NMR spectroscopy". In: *Nature communications* 6.1 (2015), pp. 1–7.

[27] T. Carlomagno. "Present and future of NMR for RNA–protein complexes: a perspective of integrated structural biology". In: *Journal of Magnetic Resonance* 241 (2014), pp. 126–136.

[28] B. Zhao and Q. Zhang. "Characterizing excited conformational states of RNA by NMR spectroscopy". In: *Current opinion in structural biology* 30 (2015), pp. 134–146.

[29] B. M. Lunde, C. Moore, and G. Varani. "RNA-binding proteins: modular design for efficient function". In: *Nature reviews Molecular cell biology* 8.6 (2007), pp. 479–490.

[30] D. Banerjee and F. Slack. "Control of developmental timing by small temporal RNAs: a paradigm for RNA-mediated regulation of gene expression". In: *Bioessays* 24.2 (2002), pp. 119–129.

[31] J. F. Kugel and J. A. Goodrich. "An RNA transcriptional regulator templates its own regulatory RNA". In: *Nature chemical biology* 3.2 (2007), pp. 89–90.

[32] W. C. Winkler, S. Cohen-Chalamish, and R. R. Breaker. "An mRNA structure that controls gene expression by binding FMN". In: *Proceedings of the National Academy of Sciences* 99.25 (2002), pp. 15908–15913.

[33] D. Laage, T. Elsaesser, and J. T. Hynes. "Water dynamics in the hydration shells of biomolecules". In: *Chemical Reviews* 117.16 (2017), pp. 10694–10725.

[34] D. E. Draper. "Folding of RNA tertiary structure: linkages between backbone phosphates, ions, and water". In: *Biopolymers* 99.12 (2013), pp. 1105–1113.

[35] S. R. Holbrook, J. L. Sussman, R. W. Warrant, G. M. Church, and S.-H. Kim. "RNA-ligand interactions:(I) magnesium binding sites in yeast tRNA Phe". In: *Nucleic acids research* 4.8 (1977), pp. 2811–2820.

[36] R. Sigel and H. Sigel. "Metal ion interactions with nucleic acids and their constituents". In: *Comprehensive inorganic chemistry II* 3 (2013), pp. 623–660.

[37] S. K. Kolev, P. S. Petkov, M. A. Rangelov, D. V. Trifonov, T. I. Milenov, and G. N. Vayssilov. "Interaction of Na+, K+, Mg 2+ and Ca 2+ counter cations with RNA". In: *Metallomics* 10.5 (2018), pp. 659–678.

[38] N. A. Denesyuk and D. Thirumalai. "How do metal ions direct ribozyme folding?" In: *Nature chemistry* 7.10 (2015), p. 793.

[39] J. L. Sussman, S. R. Holbrook, R. W. Warrant, G. M. Church, and S.-H. Kim. "Crystal structure of yeast phenylalanine transfer RNA: I. Crystallographic refinement". In: *Journal of molecular biology* 123.4 (1978), pp. 607–630.

[40] W. C. Winkler and R. R. Breaker. "Genetic control by metabolite-binding riboswitches". In: *Chembiochem* 4.10 (2003), pp. 1024–1032.

[41] J. A. Jaeger, D. H. Turner, and M. Zuker. "Improved predictions of secondary structures for RNA". In: *Proceedings of the National Academy of Sciences* 86.20 (1989), pp. 7706–7710.

[42] I. L. Hofacker, M. Fekete, and P. F. Stadler. "Secondary structure prediction for aligned RNA sequences". In: *Journal of molecular biology* 319.5 (2002), pp. 1059–1066.

[43] M. Zuker. "Mfold web server for nucleic acid folding and hybridization prediction". In: *Nucleic acids research* 31.13 (2003), pp. 3406–3415.

[44] C. Y. Cheng, F.-C. Chou, and R. Das. "Modeling complex RNA tertiary folds with Rosetta". In: *Methods in enzymology*. Vol. 553. Elsevier, 2015, pp. 35–64.

[45] R. Das, J. Karanicolas, and D. Baker. "Atomic accuracy in predicting and designing noncanonical RNA structure". In: *Nature methods* 7.4 (2010), p. 291.

[46] R. Fonseca, H. van den Bedem, and J. Bernauer. "KGSrna: efficient 3D kinematics-based sampling for nucleic acids". In: *International Conference on Research in Computational Molecular Biology*. Springer. 2015, pp. 80–95.

[47] M. J. Boniecki, G. Lach, W. K. Dawson, K. Tomala, P. Lukasz, T. Soltysinski, K. M. Rother, and J. M. Bujnicki. "SimRNA: a coarse-grained method for RNA folding simulations and 3D structure prediction". In: *Nucleic acids research* 44.7 (2015), e63–e63.

[48] M. Parisien and F. Major. "The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data". In: *Nature* 452.7183 (2008), pp. 51–55.

[49] L. G. Smith, J. Zhao, D. H. Mathews, and D. H. Turner. "Physics-based all-atom modeling of RNA energetics and structure". In: *Wiley Interdisciplinary Reviews: RNA* 8.5 (2017), e1422.

[50] A. A. Chen and A. E. García. "High-resolution reversible folding of hyperstable RNA tetraloops using molecular dynamics simulations". In: *Proceedings of the National Academy of Sciences* 110.42 (2013), pp. 16820–16825.

[51] M. Karplus and J. A. McCammon. "Molecular dynamics simulations of biomolecules". In: *Nature structural biology* 9.9 (2002), pp. 646–652.

[52] H. Gould, J. Tobochnik, and W. Christian. *An introduction to computer simulation methods*. Vol. 1. Addison-Wesley New York, 1988.

[53] J. C. Schlatterer, J. S. Martin, A. Laederach, and M. Brenowitz. "Mapping the kinetic barriers of a large RNA molecule's folding landscape". In: *PloS one* 9.2 (2014).

[54] S. Bottaro, G. Bussi, S. D. Kennedy, D. H. Turner, and K. Lindorff-Larsen. "Conformational ensembles of RNA oligonucleotides from integrating NMR and molecular simulations". In: *Science advances* 4.5 (2018), eaar8521.

[55] B. R. Brooks, C. L. Brooks III, A. D. Mackerell Jr, L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, et al. *CHARMM Parallel Performance (using DOMDEC)*. URL: https://www.charmm.org/charmm/program/performance/ (visited on 05/27/2020).

[56] B. R. Brooks, C. L. Brooks III, A. D. Mackerell Jr, L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, et al. "CHARMM: the biomolecular simulation program". In: *Journal of computational chemistry* 30.10 (2009), pp. 1545–1614.

[57] D. E. Shaw, R. O. Dror, J. K. Salmon, J. Grossman, K. M. Mackenzie, J. A. Bank, C. Young, M. M. Deneroff, B. Batson, K. J. Bowers, et al. "Millisecond-scale molecular dynamics simulations on Anton". In: *Proceedings of the conference on high performance computing networking, storage and analysis*. 2009, pp. 1–11.

[58] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al. "Mastering the game of Go with deep neural networks and tree search". In: *nature* 529.7587 (2016), p. 484.

[59] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, et al. "Mastering the game of go without human knowledge". In: *Nature* 550.7676 (2017), pp. 354–359.

[60] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. "The protein data bank". In: *Nucleic acids research* 28.1 (2000), pp. 235–242.

[61] P. D. Bank. "Protein data bank". In: *Nature New Biol* 233 (1971), p. 223.

[62] S. M. Law, A. T. Frank, and C. L. Brooks III. "PCASSO: A fast and efficient Cα-based method for accurately assigning protein secondary structure elements". In: *Journal of computational chemistry* 35.24 (2014), pp. 1757–1761.

[63] J. Feng and D. Shukla. "FingerprintContacts: Predicting Alternative Conformations of Proteins from Coevolution". In: *The Journal of Physical Chemistry B* (2020).

[64] T. Wang, M.-B. Wu, R.-H. Zhang, Z.-J. Chen, C. Hua, J.-P. Lin, and L.-R. Yang. "Advances in computational structure-based drug design and application in drug discovery". In: *Current topics in medicinal chemistry* 16.9 (2016), pp. 901–916.

[65] J. Iglesias, S. Saen-oon, R. Soliva, and V. Guallar. "Computational structure-based drug design: Predicting target flexibility". In: *Wiley Interdisciplinary Reviews: Computational Molecular Science* 8.5 (2018), e1367.

[66] J. A. Morrone, J. K. Weber, T. Huynh, H. Luo, and W. D. Cornell. "Combining Docking Pose Rank and Structure with Deep Learning Improves Protein–Ligand Binding Mode Prediction over a Baseline Docking Approach". In: *Journal of Chemical Information and Modeling* (2020).

[67] P. Gainza, F. Sverrisson, F. Monti, E. Rodola, D. Boscaini, M. Bronstein, and B. Correia. "Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning". In: *Nature Methods* 17.2 (2020), pp. 184–192.

[68] A. T. Frank, S.-H. Bae, and A. C. Stelzer. "Prediction of RNA 1H and 13C chemical shifts: a structure based approach". In: *The Journal of Physical Chemistry B* 117.43 (2013), pp. 13497–13506.

[69] C. Friedman, P. Kra, H. Yu, M. Krauthammer, and A. Rzhetsky. "GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles". In: *ISMB (supplement of bioinformatics)*. 2001, pp. 74–82.

[70] V. Hatzivassiloglou, P. A. Duboue, and A. Rzhetsky. "Disambiguating proteins, genes, and RNA in text: a machine learning approach". In: *Bioinformatics* 17.suppl_1 (2001), S97–S106.

[71] W. Hersh. *Information retrieval: a health and biomedical perspective*. Springer Science & Business Media, 2008.

[72] J. Xie, K. Zhang, and A. T. Frank. "PyShifts: A PyMOL Plugin for Chemical Shift-Based Analysis of Biomolecular Ensembles". In: *Journal of Chemical Information and Modeling* 60.3 (2020), pp. 1073–1078.

[73] R. Lorenz, S. H. Bernhart, C. H. Zu Siederdissen, H. Tafer, C. Flamm, P. F. Stadler, and I. L. Hofacker. "ViennaRNA Package 2.0". In: *Algorithms for molecular biology* 6.1 (2011), p. 26.

[74] K. Zhang and A. T. Frank. "Conditional Prediction of Ribonucleic Acid Secondary Structure Using Chemical Shifts". In: *The Journal of Physical Chemistry B* 124.3 (2019), pp. 470–478.

[75] P. Eastman, J. Shi, B. Ramsundar, and V. S. Pande. "Solving the RNA design problem with reinforcement learning". In: *PLoS computational biology* 14.6 (2018), e1006176.

[76] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. *PDB Statistics: Overall Growth of Released Structures Per Year*. URL: https://www.rcsb.org/stats (visited on 05/27/2020).

[77] R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. von Lilienfeld. "Big data meets quantum chemistry approximations: The Δ-machine learning approach". In: *Journal of chemical theory and computation* 11.5 (2015), pp. 2087–2096.

[78] M. Mandal and R. R. Breaker. "Gene regulation by riboswitches". In: *Nature Reviews Molecular Cell Biology* 5.6 (2004), pp. 451–463.

[79] E. Nudler. “Flipping riboswitches”. In: *Cell* 126.1 (2006), pp. 19–22.

[80] R. R. Breaker. “Complex riboswitches”. In: *Science* 319.5871 (2008), pp. 1795–1797.

[81] D. Mitchell III, S. M. Assmann, and P. C. Bevilacqua. “Probing RNA structure in vivo”. In: *Current Opinion in Structural Biology* 59 (2019), pp. 151–158.

[82] R. Rangan, M. Bonomi, G. T. Heller, A. Cesari, G. Bussi, and M. Vendruscolo. “Determination of structural ensembles of proteins: restraining vs reweighting”. In: *Journal of chemical theory and computation* 14.12 (2018), pp. 6632–6641.

[83] R. B. Best and M. Vendruscolo. “Determination of protein structures consistent with NMR order parameters”. In: *Journal of the American Chemical Society* 126.26 (2004), pp. 8090–8091.

[84] J. W. Pitera and J. D. Chodera. “On the use of experimental observations to bias simulated ensembles”. In: *Journal of chemical theory and computation* 8.10 (2012), pp. 3445–3451.

[85] A. Cavalli, C. Camilloni, and M. Vendruscolo. “Molecular dynamics simulations with replica-averaged structural restraints generate structural ensembles according to the maximum entropy principle”. In: *The Journal of chemical physics* 138.9 (2013), 03B603.

[86] K. A. Beauchamp, V. S. Pande, and R. Das. “Bayesian energy landscape tilting: towards concordant models of molecular ensembles”. In: *Biophysical journal* 106.6 (2014), pp. 1381–1390.

[87] G. Hummer and J. Köfinger. “Bayesian ensemble refinement by replica simulations and reweighting”. In: *The Journal of chemical physics* 143.24 (2015), 12B634_1.

[88] H. T. A. Leung, O. Bignucolo, R. Aregger, S. A. Dames, A. Mazur, S. Bernèche, and S. Grzesiek. “A rigorous and efficient method to reweight very large conformational ensembles using average experimental data and to determine their relative information content”. In: *Journal of chemical theory and computation* 12.1 (2016), pp. 383–394.

[89] C. Hartlmueller, J. C. Guenther, A. C. Wolter, J. Woehnert, M. Sattler, and T. Madl. “RNA structure refinement using NMR solvent accessibility data”. In: *Scientific reports* 7.1 (2017), p. 5393.

[90] S. Mitternacht. “FreeSASA: An open source C library for solvent accessible surface area calculations”. In: *F1000Research* 5 (2016).

[91] A. Cesari, S. Reißer, and G. Bussi. “Using the maximum entropy principle to combine simulations and solution experiments”. In: *Computation* 6.1 (2018), p. 15.

[92] S. Bottaro, T. Bengtsen, and K. Lindorff-Larsen. “Integrating molecular simulation and experimental data: a Bayesian/maximum entropy reweighting approach”. In: *Structural Bioinformatics*. Springer, 2020, pp. 219–240.

[93] D. D. Cash, O. Cohen-Zontag, N.-K. Kim, K. Shefer, Y. Brown, N. B. Ulyanov, Y. Tzfati, and J. Feigon. "Pyrimidine motif triple helix in the Kluyveromyces lactis telomerase RNA pseudoknot is essential for function in vivo". In: *Proceedings of the National Academy of Sciences* 110.27 (2013), pp. 10970–10975.

[94] R. Briones, C. Blau, C. Kutzner, B. L. de Groot, and C. Aponte-Santamaría. "GROmaρs: a GROMACS-based toolset to analyze density maps derived from molecular dynamics simulations". In: *Biophysical journal* 116.1 (2019), pp. 4–11.

[95] C. D. Stoddard, R. K. Montange, S. P. Hennelly, R. P. Rambo, K. Y. Sanbonmatsu, and R. T. Batey. "Free state conformational sampling of the SAM-I riboswitch aptamer domain". In: *Structure* 18.7 (2010), pp. 787–797.

[96] R. K. Montange and R. T. Batey. "Structure of the S-adenosylmethionine riboswitch regulatory mRNA element". In: *Nature* 441.7097 (2006), pp. 1172–1175.

[97] T. Sterling and J. J. Irwin. "ZINC 15–ligand discovery for everyone". In: *Journal of chemical information and modeling* 55.11 (2015), pp. 2324–2337.

[98] N.-K. Kim, Q. Zhang, and J. Feigon. "Structure and sequence elements of the CR4/5 domain of medaka telomerase RNA important for telomerase function". In: *Nucleic acids research* 42.5 (2014), pp. 3395–3408.

[99] K. F. Blount and R. R. Breaker. "Riboswitches as antibacterial drug targets". In: *Nature biotechnology* 24.12 (2006), pp. 1558–1564.

[100] J. A. Howe, H. Wang, T. O. Fischmann, C. J. Balibar, L. Xiao, A. M. Galgoci, J. C. Malinverni, T. Mayhood, A. Villafania, A. Nahvi, et al. "Selective small-molecule inhibition of an RNA structural element". In: *Nature* 526.7575 (2015), pp. 672–677.

[101] N. F. Rizvi, J. A. Howe, A. Nahvi, D. J. Klein, T. O. Fischmann, H.-Y. Kim, M. A. McCoy, S. S. Walker, A. Hruza, M. P. Richards, et al. "Discovery of selective RNA-binding small molecules by affinity-selection mass spectrometry". In: *ACS chemical biology* 13.3 (2018), pp. 820–831.

[102] P. Daldrop, F. E. Reyes, D. A. Robinson, C. M. Hammond, D. M. Lilley, R. T. Batey, and R. Brenk. "Novel ligands for a purine riboswitch discovered by RNA-ligand docking". In: *Chemistry & biology* 18.3 (2011), pp. 324–335.

[103] M. J. Fedor. "The role of metal ions in RNA catalysis". In: *Current opinion in structural biology* 12.3 (2002), pp. 289–295.

[104] S. Johannsen, M. M. Korth, J. Schnabl, and R. K. Sigel. "Exploring metal ion coordination to nucleic acids by NMR". In: *CHIMIA International Journal for Chemistry* 63.3 (2009), pp. 146–152.

[105] F. Dong, B. Olsen, and N. A. Baker. "Computational methods for biomolecular electrostatics". In: *Methods in cell biology* 84 (2008), pp. 843–870.

[106] A. A. Chen, D. E. Draper, and R. V. Pappu. "Molecular simulation studies of monovalent counterion-mediated interactions in a model RNA kissing loop". In: *Journal of molecular biology* 390.4 (2009), pp. 805–819.

[107] G. M. Giambaşu, T. Luchko, D. Herschlag, D. M. York, and D. A. Case. "Ion counting from explicit-solvent simulations and 3D-RISM". In: *Biophysical journal* 106.4 (2014), pp. 883–894.

[108] C. Burkhardt and M. Zacharias. "Modelling ion binding to AA platform motifs in RNA: a continuum solvent study including conformational adaptation". In: *Nucleic acids research* 29.19 (2001), pp. 3910–3918.

[109] H. Chen, S. P. Meisburger, S. A. Pabit, J. L. Sutton, W. W. Webb, and L. Pollack. "Ionic strength-dependent persistence lengths of single-stranded RNA and DNA". In: *Proceedings of the National Academy of Sciences* 109.3 (2012), pp. 799–804.

[110] S. Kirmizialtin and R. Elber. "Computational exploration of mobile ion distributions around RNA duplex". In: *The Journal of Physical Chemistry B* 114.24 (2010), pp. 8207–8220.

[111] R. L. Hayes, J. K. Noel, U. Mohanty, P. C. Whitford, S. P. Hennelly, J. N. Onuchic, and K. Y. Sanbonmatsu. "Magnesium fluctuations modulate RNA dynamics in the SAM-I riboswitch". In: *Journal of the American Chemical Society* 134.29 (2012), pp. 12043–12053.

[112] R. L. Hayes, J. K. Noel, A. Mandic, P. C. Whitford, K. Y. Sanbonmatsu, U. Mohanty, and J. N. Onuchic. "Generalized Manning condensation model captures the RNA ion atmosphere". In: *Physical review letters* 114.25 (2015), p. 258105.

[113] Y.-Y. Wu, Z.-L. Zhang, J.-S. Zhang, X.-L. Zhu, and Z.-J. Tan. "Multivalent ion-mediated nucleic acid helix-helix interactions: RNA versus DNA". In: *Nucleic acids research* 43.12 (2015), pp. 6156–6165.

[114] P. S. Henke and C. H. Mak. "Free energy of RNA-counterion interactions in a tight-binding model computed by a discrete space mapping". In: *The Journal of chemical physics* 141.6 (2014), 08B612_1.

[115] S. D. Fried, L.-P. Wang, S. G. Boxer, P. Ren, and V. S. Pande. "Calculations of the electric fields in liquid solutions". In: *The Journal of Physical Chemistry B* 117.50 (2013), pp. 16236–16248.

[116] C. Mak and P. S. Henke. "Ions and RNAs: free energies of counterion-mediated RNA fold stabilities". In: *Journal of chemical theory and computation* 9.1 (2013), pp. 621–639.

[117] A. Philips, K. Milanowska, G. Lach, M. Boniecki, K. Rother, and J. M. Bujnicki. "MetalionRNA: computational predictor of metal-binding sites in RNA structures". In: *Bioinformatics* 28.2 (2012), pp. 198–205.

[118] L.-Z. Sun, J.-X. Zhang, and S.-J. Chen. “MCTBI: a web server for predicting metal ion effects in RNA structures”. In: *RNA* 23.8 (2017), pp. 1155–1165.

[119] D. R. Banatao, R. B. Altman, and T. E. Klein. “Microenvironment analysis and identification of magnesium binding sites in RNA”. In: *Nucleic acids research* 31.15 (2003), pp. 4450–4460.

[120] V. Botu and R. Ramprasad. “Adaptive machine learning framework to accelerate ab initio molecular dynamics”. In: *International Journal of Quantum Chemistry* 115.16 (2015), pp. 1074–1083.

[121] V. Botu and R. Ramprasad. “Learning scheme to predict atomic forces and accelerate materials simulations”. In: *Physical Review B* 92.9 (2015), p. 094306.

[122] V. Botu, R. Batra, J. Chapman, and R. Ramprasad. “Machine learning force fields: construction, validation, and outlook”. In: *The Journal of Physical Chemistry C* 121.1 (2016), pp. 511–522.

[123] H. Robinson, Y.-G. Gao, R. Sanishvili, A. Joachimiak, and A. H.-J. Wang. “Hexahydrated magnesium ions bind in the deep major groove and at the outer mouth of A-form nucleic acid duplexes”. In: *Nucleic Acids Research* 28.8 (2000), pp. 1760–1766.

[124] V. Botu. “Surface Chemistry with Machine Learning and Quantum Mechanics”. In: (2016).

[125] J. Behler. “Atom-centered symmetry functions for constructing high-dimensional neural network potentials”. In: *The Journal of chemical physics* 134.7 (2011), p. 074106.

[126] S. Chhabra, J. Xie, and A. T. Frank. “RNAPosers: Machine Learning Classifiers for Ribonucleic Acid–Ligand Poses”. In: *The Journal of Physical Chemistry B* (2020).

[127] P. Fränti and S. Sieranoja. “How much can k-means be improved by using better initialization and repeats?” In: *Pattern Recognition* 93 (2019), pp. 95–112.

[128] L. v. d. Maaten and G. Hinton. “Visualizing data using t-SNE”. In: *Journal of machine learning research* 9.Nov (2008), pp. 2579–2605.

[129] L. R. Stefan, R. Zhang, A. G. Levitan, D. K. Hendrix, S. E. Brenner, and S. R. Holbrook. “MeRNA: a database of metal ion binding sites in RNA structures”. In: *Nucleic acids research* 34.suppl_1 (2006), pp. D131–D134.

[130] J. Schnabl, P. Suter, and R. K. Sigel. “MINAS—a database of Metal Ions in Nucleic Acids”. In: *Nucleic acids research* 40.D1 (2012), pp. D434–D438.

[131] H. Zheng, I. G. Shabalin, K. B. Handing, J. M. Bujnicki, and W. Minor. “Magnesium-binding architectures in RNA crystal structures: validation, binding preferences, classification and motif detection”. In: *Nucleic acids research* 43.7 (2015), pp. 3789–3801.

[132] S. S. Batsanov. “Van der Waals radii of elements”. In: *Inorganic materials* 37.9 (2001), pp. 871–885.

[133] M. Nayal and E. Di Cera. "Predicting Ca (2+)-binding sites in proteins." In: *Proceedings of the National Academy of Sciences* 91.2 (1994), pp. 817–821.

[134] Y. Chen, L. Zubovic, F. Yang, K. Godin, T. Pavelitz, J. Castellanos, P. Macchi, and G. Varani. "Rbfox Proteins Regulate Microrna Biogenesis by Sequence-specific Binding to Their Precursors and Target Downstream Dicer". In: *Nucleic Acids Res.* 44.9 (2016), pp. 4381–4395.

[135] L. Jovine, S. Djordjevic, and D. Rhodes. "The crystal structure of yeast phenylalanine tRNA at 2.0 Å resolution: cleavage by Mg2+ in 15-year old crystals". In: *Journal of molecular biology* 301.2 (2000), pp. 401–414.

[136] D. Klein, T. Schmeing, P. Moore, and T. Steitz. "The kink-turn: a new RNA secondary structure motif". In: *The EMBO journal* 20.15 (2001), pp. 4214–4221.

[137] L. Huang and D. M. Lilley. "The kink turn, a key architectural element in RNA structure". In: *Journal of molecular biology* 428.5 (2016), pp. 790–801.

[138] T. A. Goody, S. E. Melcher, D. G. Norman, and D. M. Lilley. "The kink-turn motif in RNA is dimorphic, and metal ion-dependent". In: *Rna* 10.2 (2004), pp. 254–264.

[139] A. Ren, K. R. Rajashankar, and D. J. Patel. "Fluoride ion encapsulation by Mg 2+ ions and phosphates in a fluoride riboswitch". In: *Nature* 486.7401 (2012), pp. 85–89.

[140] K. Juneau, E. Podell, D. J. Harrington, and T. R. Cech. "Structural basis of the enhanced stability of a mutant ribozyme domain and a detailed view of RNA–solvent interactions". In: *Structure* 9.3 (2001), pp. 221–231.

[141] O. Fedorova, G. E. Jagdmann, R. L. Adams, L. Yuan, M. C. Van Zandt, and A. M. Pyle. "Small molecules that target group II introns are potent antifungal agents". In: *Nature chemical biology* 14.12 (2018), p. 1073.

[142] A. C. Stelzer, A. T. Frank, J. D. Kratz, M. D. Swanson, M. J. Gonzalez-Hernandez, J. Lee, I. Andricioaei, D. M. Markovitz, and H. M. Al-Hashimi. "Discovery of selective bioactive small molecules by targeting an RNA dynamic ensemble". In: *Nature chemical biology* 7.8 (2011), p. 553.

[143] A. T. Laurie and R. M. Jackson. "Q-SiteFinder: an energy-based method for the prediction of protein–ligand binding sites". In: *Bioinformatics* 21.9 (2005), pp. 1908–1916.

[144] B. Huang and M. Schroeder. "LIGSITE csc: predicting ligand binding sites using the Connolly surface and degree of conservation". In: *BMC structural biology* 6.1 (2006), p. 19.

[145] J. Dundas, Z. Ouyang, J. Tseng, A. Binkowski, Y. Turpaz, and J. Liang. "CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues". In: *Nucleic acids research* 34.suppl_2 (2006), W116–W118.

[146] V. Le Guilloux, P. Schmidtke, and P. Tuffery. “Fpocket: an open source platform for ligand pocket detection”. In: *BMC bioinformatics* 10.1 (2009), p. 168.

[147] S. D. Morley and M. Afshar. “Validation of an empirical RNA-ligand scoring function for fast flexible docking using RiboDock®”. In: *Journal of computer-aided molecular design* 18.3 (2004), pp. 189–208.

[148] G. R. Bowman and P. L. Geissler. “Equilibrium fluctuations of a single folded protein reveal a multitude of potential cryptic allosteric sites”. In: *Proceedings of the National Academy of Sciences* 109.29 (2012), pp. 11681–11686.

[149] M. Rueda, G. Bottegoni, and R. Abagyan. “Recipes for the selection of experimental protein conformations for virtual screening”. In: *Journal of chemical information and modeling* 50.1 (2010), pp. 186–193.

[150] E. Lionta, G. Spyrou, D. K Vassilatis, and Z. Cournia. “Structure-based virtual screening for drug discovery: principles, applications and recent advances”. In: *Current topics in medicinal chemistry* 14.16 (2014), pp. 1923–1938.

[151] A. T. Frank, A. C. Stelzer, H. M. Al-Hashimi, and I. Andricioaei. “Constructing RNA dynamical ensembles by combining MD and motionally decoupled NMR RDCs: new insights into RNA dynamics and adaptive ligand recognition”. In: *Nucleic acids research* 37.11 (2009), pp. 3670–3679.

[152] M. Kang, R. Peterson, and J. Feigon. “Structural insights into riboswitch control of the biosynthesis of queuosine, a modified nucleotide found in the anticodon of tRNA”. In: *Molecular cell* 33.6 (2009), pp. 784–790.

[153] R. J. Marcheschi, M. Tonelli, A. Kumar, and S. E. Butcher. “Structure of the HIV-1 frameshift site RNA bound to a small molecule inhibitor of viral replication”. In: *ACS chemical biology* 6.8 (2011), pp. 857–864.

[154] M.-K. Lee, A. Bottini, M. Kim, M. F. Bardaro, Z. Zhang, M. Pellecchia, B.-S. Choi, and G. Varani. “A novel small-molecule binds to the influenza A virus RNA promoter and inhibits viral replication”. In: *Chemical communications* 50.3 (2013), pp. 368–370.

[155] A. Tsai, S. Uemura, M. Johansson, E. V. Puglisi, R. A. Marshall, C. E. Aitken, J. Korlach, M. Ehrenberg, and J. D. Puglisi. “The impact of aminoglycosides on the dynamics of translation elongation”. In: *Cell reports* 3.2 (2013), pp. 497–508.

[156] E. Duchardt-Ferner, S. R. Gottstein-Schmidtke, J. E. Weigand, O. Ohlenschläger, J.-P. Wurm, C. Hammann, B. Suess, and J. Wöhnert. “What a Difference an OH Makes: Conformational Dynamics as the Basis for the Ligand Specificity of the Neomycin-Sensing Riboswitch”. In: *Angewandte Chemie International Edition* 55.4 (2016), pp. 1527–1530.

[157] E. Kligun and Y. Mandel-Gutfreund. “Conformational readout of RNA by small ligands”. In: *RNA Biology* 10.6 (2013), pp. 981–989.

[158] P. Zeng, J. Li, W. Ma, and Q. Cui. "Rsite: a computational method to identify the functional sites of noncoding RNAs". In: *Scientific reports* 5 (2015), p. 9179.

[159] K. Wang, Y. Jian, H. Wang, C. Zeng, and Y. Zhao. "RBind: computational network method to predict RNA binding sites". In: *Bioinformatics* 34.18 (2018), pp. 3131–3136.

[160] C. J. Burges. "From ranknet to lambdarank to lambdamart: An overview". In: *Learning* 11.23-581 (2010), p. 81.

[161] O. Leslie E. "Prebiotic chemistry and the origin of the RNA world". In: *Critical Reviews in Biochemistry and Molecular Biology* 39.2 (2004), pp. 99–123.

[162] S. R. Eddy. "Non–coding RNA genes and the modern RNA world". In: *Nature Reviews Genetics* 2.12 (2001), p. 919.

[163] M. Egli, A. Flavell, A. M. Pyle, S. Allen, J. Fisher, S. I. Haq, J. W. Engels, J. A. Grasby, B. Luisi, C. Laughton, et al. *Nucleic acids in chemistry and biology*. Royal Society of Chemistry, 2006.

[164] P. Nissen, J. Hansen, N. Ban, P. B. Moore, and T. A. Steitz. "The structural basis of ribosome activity in peptide bond synthesis". In: *Science* 289.5481 (2000), pp. 920–930.

[165] B. J. Tucker and R. R. Breaker. "Riboswitches as versatile gene control elements". In: *Current Opinion in Structural Biology* 15.3 (2005), pp. 342–348.

[166] Y. Dorsett and T. Tuschl. "siRNAs: applications in functional genomics and potential as therapeutics". In: *Nature Reviews Drug Discovery* 3.4 (2004), p. 318.

[167] N. Bushati and S. M. Cohen. "microRNA functions". In: *Annu. Rev. Cell Dev. Biol.* 23 (2007), pp. 175–205.

[168] T. A. Cooper, L. Wan, and G. Dreyfuss. "RNA and disease". In: *Cell* 136.4 (2009), pp. 777–793.

[169] A. Serganov and D. J. Patel. "Ribozymes, riboswitches and beyond: regulation of gene expression without proteins". In: *Nature Reviews Genetics* 8.10 (2007), p. 776.

[170] H. Ling, M. Fabbri, and G. A. Calin. "MicroRNAs and other non-coding RNAs as targets for anticancer drug development". In: *Nature Reviews Drug Discovery* 12.11 (2013), p. 847.

[171] M. Matsui and D. R. Corey. "Non-coding RNAs as drug targets". In: *Nature Reviews Drug Discovery* 16.3 (2017), p. 167.

[172] G. Zhu, M. Ye, M. J. Donovan, E. Song, Z. Zhao, and W. Tan. "Nucleic acid aptamers: an emerging frontier in cancer therapy". In: *Chemical Communications* 48.85 (2012), pp. 10472–10480.

[173] J. B. Opalinska and A. M. Gewirtz. "Nucleic-acid therapeutics: basic principles and recent applications". In: *Nature Reviews Drug Discovery* 1.7 (2002), p. 503.

[174] R. R. Breaker. “Riboswitches and the RNA world”. In: *Cold Spring Harbor Perspectives in Biology* 4.2 (2012), a003566.

[175] P. Machtel, K. Bąkowska-Żywicka, and M. Żywicki. “Emerging applications of riboswitches–from antibacterial targets to molecular tools”. In: *Journal of Applied Genetics* 57.4 (2016), pp. 531–541.

[176] D. B. Kitchen, H. Decornez, J. R. Furr, and J. Bajorath. “Docking and scoring in virtual screening for drug discovery: methods and applications”. In: *Nature Reviews Drug Discovery* 3.11 (2004), p. 935.

[177] P. Pfeffer and H. Gohlke. “DrugScoreRNA Knowledge-Based Scoring Function To Predict RNA- Ligand Interactions”. In: *Journal of Chemical Information and Modeling* 47.5 (2007), pp. 1868–1876.

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

[179] A. Philips, K. Milanowska, G. Łach, and J. M. Bujnicki. “LigandRNA: computational predictor of RNA–ligand interactions”. In: *RNA* 19.12 (2013), pp. 1605–1616.

[180] Z. Yan and J. 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.

[181] J. D. Durrant and J. A. McCammon. “NNScore 2.0: a neural-network receptor–ligand scoring function”. In: *Journal of chemical information and modeling* 51.11 (2011), pp. 2897–2903.

[182] M. Ragoza, J. Hochuli, E. Idrobo, J. Sunseri, and D. R. Koes. “Protein–ligand scoring with convolutional neural networks”. In: *Journal of chemical information and modeling* 57.4 (2017), pp. 942–957.

[183] J. Jiménez, M. Skalic, G. Martinez-Rosell, and G. De Fabritiis. “K DEEP: Protein–Ligand Absolute Binding Affinity Prediction via 3D-Convolutional Neural Networks”. In: *Journal of chemical information and modeling* 58.2 (2018), pp. 287–296.

[184] J. Pei, Z. Zheng, H. Kim, L. F. Song, S. Walworth, M. R. Merz, and K. M. Merz. “Random Forest refinement of pairwise potentials for protein-ligand decoy detection”. In: *Journal of Chemical Information and Modeling* (2019).

[185] D. D. Nguyen, Z. Cang, K. Wu, M. Wang, Y. Cao, and G.-W. Wei. “Mathematical deep learning for pose and binding affinity prediction and ranking in D3R Grand Challenges”. In: *Journal of Computer-Aided Molecular Design* 33.1 (2019), pp. 71–82.

[186] J. Wang, W. Wang, S. Huo, M. Lee, and P. A. Kollman. “Solvation model based on weighted solvent accessible surface area”. In: *The Journal of Physical Chemistry B* 105.21 (2001), pp. 5055–5067.

[187] T. Rognes, T. Flouri, B. Nichols, C. Quince, and F. Mahé. “VSEARCH: a versatile open source tool for metagenomics”. In: *PeerJ* 4 (2016), e2584.

[188] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[189] J. Dittrich, D. Schmidt, C. Pfleger, and H. Gohlke. “Converging a Knowledge-Based Scoring Function: DrugScore2018”. In: *Journal of chemical information and modeling* 59.1 (2018), pp. 509–521.

[190] N. M. O’Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, and G. R. Hutchison. “Open Babel: An open chemical toolbox”. In: *Journal of Cheminformatics* 3.1 (2011), p. 33.