THEORETICAL AND COMPUTATIONAL MODELING OF RNA-LIGAND INTERACTIONS

A Dissertation presented to

the Faculty of the Graduate School

at the University of Missouri

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

by

YUANZHE ZHOU

Dr. Shi-Jie Chen, Dissertation Supervisor

DECEMBER 2021

The undersigned, appointed by the Dean of the Graduate School, have examined the dissertation entitled:

THEORETICAL AND COMPUTATIONAL MODELING OF
RNA-LIGAND INTERACTIONS

presented by Yuanzhe Zhou,

a candidate for the degree of Doctor of Philosophy and hereby certify that, in their opinion, it is worthy of acceptance.

Dr. Shi-Jie Chen

Dr. Xiao Heng

Dr. Gavin M. King

Dr. Ioan Kosztin

Dr. Xiaoqin Zou

ACKNOWLEDGMENTS

There are many who helped me along the way in my exploration of biological physics. I would like to thank all of them for their tremendous support.

I would first like to thank my advisor, Dr. Shi-Jie Chen. He continuously provided encouragement and was always enthusiastic to assist in any way he could throughout my academic journey. With immense knowledge and detailed insight in the field, his insightful feedback and guidance sharpened my thinking and brought my work to a higher level. This has truly made the research an inspiring experience for me. I am honored to have him as my advisor.

I am grateful to my other committee members, Drs. Xiao Heng, Gavin King, Ioan Kosztin, Xiaoqin Zou. I am so thankful to have them as my committee members, as well as their time and valuable advice, which greatly helped my research.

Thank you to all the members of Dr. Chen's group, they provided stimulating discussions as well as happy distractions to rest my mind outside of my research. Thanks to lab members: Sicheng Zhang, Drs. Jun Li and Lei Jin, and Drs. Yangwei Jiang, Travis Hurst, Chenhan Zhao, Dong Zhang, Yi Cheng, Xiaojun Xu, and Lizhen Sun. My accomplishments would not have been possible without their passionate participation and input.

Finally, I must express my very profound gratitude to my parents and friends for providing me with unfailing support and continuous encouragement throughout my years of study.

TABLE OF CONTENTS

CHAPTER

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

ABSTRACT

Ribonucleic acid (RNA) is a polymeric nucleic acid that plays a variety of critical roles in gene expression and regulation at the level of transcription and translation. Recently, there has been an enormous interest in the development of therapeutic strategies that target RNA molecules. Instead of modifying the product of gene expression, i.e., proteins, RNA-targeted therapeutics aims to modulate the relevant key RNA elements in the disease-related cellular pathways. Such approaches have two significant advantages. First, diseases with related proteins that are difficult or unable to be drugged become druggable by targeting the corresponding messenger RNAs (mRNAs) that encode the amino acid sequences. Second, besides coding mRNAs, the vast majority of the human genome sequences are transcribed to noncoding RNAs (ncRNAs), which serve as enzymatic, structural, and regulatory elements in cellular pathways of most human diseases. Targeting noncoding RNAs would open up remarkable new opportunities for disease treatment.

The first step in modeling the RNA-drug interaction is to understand the 3D structure of the given RNA target. With current theoretical models, accurate prediction of 3D structures for large RNAs from sequence remains computationally infeasible. One of the major challenges comes from the flexibility in the RNA molecule, especially in loop/junction regions, and the resulting rugged energy landscape. However, structure probing techniques, such as the "selective 2′-hydroxyl acylation analyzed by primer extension" (*SHAPE*) experiment, enable the quantitative detection of the relative flexibility and hence structure information of RNA structural elements. Therefore, one may incorporate the SHAPE data into RNA 3D structure prediction. In the first project, we investigate the feasibility of using a machine-learning-based approach to predict the SHAPE reactivity from the 3D RNA structure and compare the machine-learning result to that of a physics-based model. In the second project, in order to provide a user-friendly tool for RNA biologists, we developed a fully automated web interface, “SHAPE predictoR” (*SHAPER*) for predicting SHAPE profile from any given 3D RNA structure.

In a cellular environment, various factors, such as metal ions and small molecules, interact with an RNA molecule to modulate RNA cellular activity. RNA is a highly charged polymer with each backbone phosphate group carrying one unit of negative (electronic) charge. In order to fold into a compact functional tertiary structure, it requires metal ions to reduce Coulombic repulsive electrostatic forces by neutralizing the backbone charges. In particular, Mg^{2+} ion is essential for the folding and stability of RNA tertiary structures. In the third project, we introduce a machine-learning-based model, the “Magnesium convolutional neural network” (*MgNet*) model, to predict Mg^{2+} binding site for a given 3D RNA structure, and show the use of the model in investigating the important coordinating RNA atoms and identifying novel Mg^{2+} binding motifs.

Besides Mg^{2+} ions, small molecules, such as drug molecules, can also bind to an RNA to modulate its activities. Motivated by the tremendous potential of RNA-targeted drug discovery, in the fourth project, we develop a novel approach to predicting RNA-small molecule binding. Specifically, we develop a statistical potential-based scoring/ranking method (*SPRank*) to identify the native binding mode of the small molecule from a pool of decoys and estimate the binding affinity for the given RNA-small molecule complex. The results tested on a widely used data set suggest that SPRank can achieve (moderately) better performance than the current state-of-art models.

Chapter 1

Introduction

The backgrounds of my research projects are described in this chapter.

1.1 RNA and its biological significance — a brief overview

Ribonucleic acid (RNA) is a polymeric molecule transcribed from DNA in the cell nucleus and is essential in various biological roles in coding, decoding, regulation, and expression of genes. A nucleotide, which consists of a phosphate group, a ribose sugar, and a nucleobase. is the building blocks of an RNA molecule. There are four types of natural nucleotides: adenine (A), uracil (U), cytosine (C), or guanine (G). Like DNA, an RNA is assembled as a chain of nucleotides connected through phosphodiester bonds, but unlike DNA, an RNA is often found to be single-stranded and contain self-complementary sequences that allow parts of the RNA to fold and pair with itself to form highly structured tertiary conformation. RNA structure can often be viewed from three different structural levels (see Fig. 1.1): the primary (1D) level of an RNA structure is described by the linear

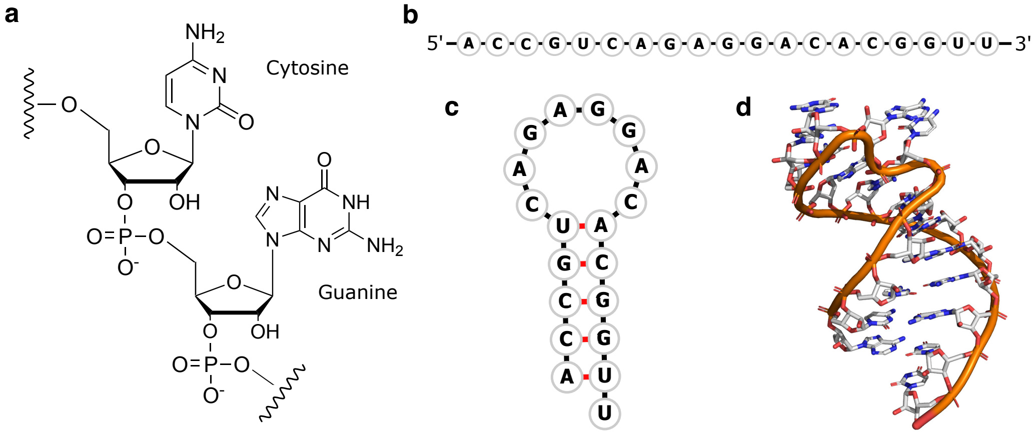


Figure 1.1: Example of nucleotides in RNA and RNA structure, viewed from different structural levels. (a) Two consecutive nucleotides (C and G) connected through a phosphodiester bond. (b) The sequence (primary (1D) structure) of the RNA molecule starts from 5′- to 3′-hydroxyl terminal functional groups. (c) The secondary (2D) structure shows the loop and helix. (d) The tertiary (3D) structure shows the three-dimensional geometry of the nucleotides and atoms.

sequence information of the nucleotides; the secondary structure is described at the two-dimensional (2D) level and is defined by the contact pairs (base pairs) of the nucleotides, and the tertiary structure is described at the three-dimensional (3D) level where all the interactions can be projected onto the structure in 3D space.

RNA molecules can be categorized into two types: coding RNAs that carry the genetic information and are translated into proteins, and the noncoding RNAs (ncRNAs), which constitute the vast majority of the RNA molecules and serve as enzymatic, structural, and regulatory elements for gene expression. Only approximately 1.5% of the human genome [1–10] encoding proteins. The vast majority of human genome encodes ncRNAs. Typical ncRNAs include those directly involved in protein synthesis, such as transfer RNAs (tRNAs) that deliver amino acids to the ribosome, and ribosomal RNAs (rRNAs) that link amino acids together to form coded proteins. In addition, there are many ncRNAs involved in post-transcriptional modification or gene regulation, such as various riboswitches, small nuclear RNA (snRNA), guide RNA (gRNA), long noncoding RNA (lncRNA), microRNA (miRNA), and small interfering RNA (siRNA), etc [11, 12].

With the ever increasing discoveries of new RNA structures and functions, RNA-based therapeutics is considered as a highly promising new strategy for disease treatment and has gained significant interest recently [2, 4, 13–21]. One of the common approaches in RNA-based therapeutics is modulating the activity of the target RNA molecule involved in the disease-related cellular process through small molecule binding [2, 4, 16–21]. However, correct identification of the potential drug molecule requires a comprehensive understanding of both RNA 3D structure and various interactions that affect RNA function, such as those between RNA and metal ions [22–40].

1.2 The significance of Mg^{2+} in RNA folding

Since RNA is a negatively charged molecule, metal ions, especially Mg^{2+}, are essential for shielding of charges in the polyanionic backbones and allowing RNA to adopt a diverse range of folded structures. Mg^{2+} is more effective in stabilizing RNA structure than other ions [22–25, 27, 28, 38–40]. First of all, Mg^{2+} has a higher charge than monovalent ions, which causes less entropic cost when becoming localized around the RNA [24–26]; In addition, it has a smaller radius than other divalent ions and can bind into well-defined narrow pockets and grooves in RNA [41–47]. For example, tRNA stability increases remarkably in the presence of monovalent (in particular Na^+ and K^+) and divalent (Mg^{2+}) cations [38]. And ribosome requires Mg^{2+} to stabilize due to its highly compact tertiary fold [42, 48, 49]. Researchers have shown that the growth of *Escherichia coli* cells under conditions of Mg^{2+} starvation results in ribosome depletion [50] and the *in vitro* association of the small and large ribosomal subunits to form intact ribosomes depends strongly on Mg^{2+}

concentration [51–53].

Mg^{2+} is usually considered as the natural cofactor to help recognize binding partners and mediate catalytic processes [31, 36, 37, 54–58]. Previous studies [30, 32–36] have shown that Mg^{2+} participate in the catalytic reactions of certain ribozymes. Hammerhead ribozymes are a well-known examples that require metal ions to be present both for forming the functional three-dimensional fold and performing the cleavage functions of a phosphodiester bond [31, 37, 56–58].

Metal ions are also relevant to drug discovery [59–61]. Metal ions can interfere with RNA-targeted antibiotic inhibition. For example, aminoglycoside bound to RNA can displace structurally important divalent metal ions and such a competitive binding can significantly influence drug efficacy [61]. Similar mechanisms of Mg^{2+} displacement have also been found in the inhibition of other ribozymes by neomycin. These findings suggested that aminoglycosides can compete with Mg^{2+} at functionally and structurally important ion binding sites [60].

However, experimental studies of RNA-Mg^{2+} interactions are challenging because it is difficult for X-ray crystallography to distinguish chemically distinct species with indistinguishable electron distributions. For example, Mg^{2+}, Na^{+}, and H_2O all have 10 electrons and are hard to be distinguished from the electron density maps alone [62]. Since partial and mixed occupancies of ions are also possible, other species, such as K^{+} with 55% occupation [63], although have different numbers of electrons, can have the same observed effective electron number as that of a Mg^{2+}, Table 1.1 shows a summary for the species with the same effective electron number. Due to the reason above, coordination distances and geometries are two additional criteria that are often used for assigning ionic species to electron density spots [64–69]. For example, the metal water coordination distance is what distinguishes Na^+ from Mg^{2+} and is only interpretable at a high resolution [70]. Misinterpreted locations of ions bound to macromolecules have been found in many macromolecule structures [71, 72] and Mg^{2+} is not an exception [73]. Because they all have 10 electrons and can be distinguished only in high-resolution structures, Mg^{2+} could be easily mistaken for water molecules or Na^+ or simply missing from the crystal structures [41]. A significant number of incorrectly identified Mg^{2+} sites can impose a strong (incorrect) bias on Mg^{2+} binding analysis. Thus, high resolution (1.5 A or better) and high occupancy are necessary to accurately identify Mg^{2+} in crystal structures. This demands the development of a computational model that can accurately predict Mg^{2+} binding sites. Such a model can be used to assist experimentalists to verify and validate their results and it can also further our understanding of the RNA-metal ion interaction.

Table 1.1: The table shows species with the same number of effective electrons at different occupancies observed in X-ray crystallography. The occupancy column lists the occupancy values, where full occupancy is indicated by 1.0 and mixed occupancy between two species is shown with a slash. Data collected from publication [63].

Species	Occupancy	Effective	Effective number
		radius Å	of electrons
H_2O	1.0	1.40	10
Na^+	1.0	0.95	10
Mg^{2+}	1.0	0.65	10
K^+	0.6	1.33	10
H_2O/Na^+	0.8/0.2	1.31	10
H_2O/K^+	0.8/0.1	1.39	10

1.3 Evaluating RNA-small molecule interaction

RNA molecule can serve as the target for drug (small molecule) binding [2, 4, 16–21], and this approach is analogous to protein-targeted drug discovery. For example, bacterial infection can be treated with antibiotics that target the active sites of the bacterial ribosomal RNAs (rRNAs) through the inhibition of the protein synthesis [74–77]. Another possible method is to regulate gene expression by modulating common riboswitches in bacterial cells through ligand-induced RNA conformational changes [78–92]. In addition to bacterial RNAs, another type of interesting RNA target is viral RNA with highly conserved structured motifs [16, 18, 93], such as the HIV transactivation response (TAR) element in the $5'$ untranslated region [94, 95], the internal ribosome entry site (IRES) element located in the hepatitis C virus (HCV) genome [96–100], and the influenza A virus RNA promoter [101, 102]. Previous studies have identified a small molecule compound against the atypical three-stemmed RNA pseudoknot that stimulated -1 programmed ribosomal frameshifting in SARS-CoV RNA genome, the compound shows inhibition of -1 ribosomal frameshifting with IC_{50} at 210 μM [103–106], which provides guidance for designing novel drugs for the treatment of SARS-CoV-2 disease.

Compared to protein, there are two unique challenges in modeling RNA-small molecule interactions. First, unlike a protein, RNA is highly charged, with each phosphate group carrying one electronic charge. Thus, RNA folding and ligand binding require the participation of metal ions such as Mg^{2+} and water molecules to stabilize the binding pocket structure of the RNA and to mediate ligand-RNA interactions [107–110]. Second, RNA molecules are often quite flexible, capable of folding into multiple stable conformations, and ligand binding often induces structural switches between different conformers of the RNA receptor. Compared with protein-ligand binding, ligand-binding sites on RNA can be less deep and more polar, solvated, and conformationally flexible [3, 18, 110], which adds further complexity to predicting RNA-small molecule interactions.

The prediction of RNA-ligand (small molecule) binding involves the generation of an ensemble of possible binding modes and the scoring/ranking of the different binding modes such that the best-scored (ranked) RNA-ligand binding mode is predicted. Current scoring functions generally fall into three types: physics-based approach [109, 111–120], knowledge-based approach [121–126], and machine-learning approach [127–129]. Compared to the physics-based approach, recently, there is much more active development in the knowledge-based and machine-learning scoring functions, with several newly published methods [130]. This trend reflects the increasing experimental data of RNA-ligand complexes which enables a more effective training process. The traditional physics-based approaches use either an atomistic based physical force field derived from thermodynamic data and *ab initio* calculations, or an empirical energy that contains the linear combination of terms for various physical interactions. The atomistic force-field approach typically uses a combination of molecular dynamic force field (e.g., AMBER and CHARMM force fields) and implicit solvent model (e.g., Poisson-Boltzmann surface area model [131–137] or Generalized-Born surface area model [138–146]) to model the energy changes due to RNA-small molecule interactions and molecule-solvent interactions. DOCK 6 [112] and MORDOR [111] are two such examples. However, for the consideration of computational efficiency, the majority of the physics-based approaches adopted empirical energy functions [109, 113–120]. By evaluating the total energy as a weighted sum of individual interactions, such as van der Waals, electrostatic, desolvation and hydrogen-bond interactions, these models are able to achieve a better balance between computational cost and accuracy. In fact, most RNA-small molecule docking software adopted this type of energy functions [109, 113–120].

In recent years, knowledge-based and machine-learning approaches are two types of emerging scoring functions for modeling RNA-small molecule interactions [121–129]. In the knowledge-based approach, the basic assumption is that the interaction potential between atom pairs can be derived from the statistics of the known RNA-small molecule complexes through inverse Boltzmann's law. In contrast, machine-learning approaches do not assume any pre-defined functional form of the interactions and can leverage the experimental data better with a much larger number of trainable parameters. A variety of machine-learning models such as support vector machine (SVM), random forest (RF), neural network (NN), and convolutional neural network (CNN) have been proposed and shown success to predict protein/RNA-small molecule interactions [127–129, 147–152]. However, although machine-learning models for protein folding [153–158] and protein-small molecule interactions [159–162] have shown significant success, the lack of a comprehensive and high-quality curated database of RNA-small molecule complexes imposes great challenges on both knowledge-based and machine-learning approaches.

1.4 Extracting 3D structural information from SHAPE data

Selective 2′-hydroxyl acylation analyzed by primer extension (SHAPE) is an effective chemical probing technique that provides insights into RNA local structure at single nucleotide resolution [163, 164]. SHAPE reagents are small-molecule electrophiles—1-methyl-7-nitroisatoic anhydride (1M7), 1-methyl-6-nitroisatoic anhydride (1M6), N-methylisotoic anhydride (NMIA), benzoyl cyanide (BzCN), 2-methyl-3-furoic acid imidazolide (FAI), and 2-methylnicotinic acid imidazolide (NAI)—that react preferentially with the 2′-hydroxyl group of the target RNA through acylation to form a 2′-O-adduct (see Fig. 1.2). Although the mechanism that governs the SHAPE reactivity is still not fully understood, several studies have suggested a variety of conformations that render a nucleotide reactivity of SHAPE [165, 166], and indicated that SHAPE reactivity can reflect local nucleotides flexibility [167–169]. Highly SHAPE-reactive nucleotides often come from the unconstrained region of the RNA, such as the flexible loop/junction region, which is capable of sampling multiple conformations and has greater probability to adopt the SHAPE-reactive conformations. On the other hand, nucleotides that are constrained by base-pairing and stacking interactions are less flexible and hence less SHAPE-reactive. This characteristic feature makes SHAPE a useful tool for the quantitative measurement of RNA local structural dynamics. In particular, SHAPE reactivity can be used to estimate whether a nucleotide is in a rigid base pair (either in a helix or in a structured loop) or remains unpaired in a flexible loop/junction. The local structural information provided by SHAPE can place effective constraints on the RNA conformational space, which results in a much more effective computational modeling of RNA structure. Studies [165, 166, 170–172] have shown that SHAPE reactivity can be used to guide 2D/3D structure predictions and exclude SHAPE-incompatible structures.

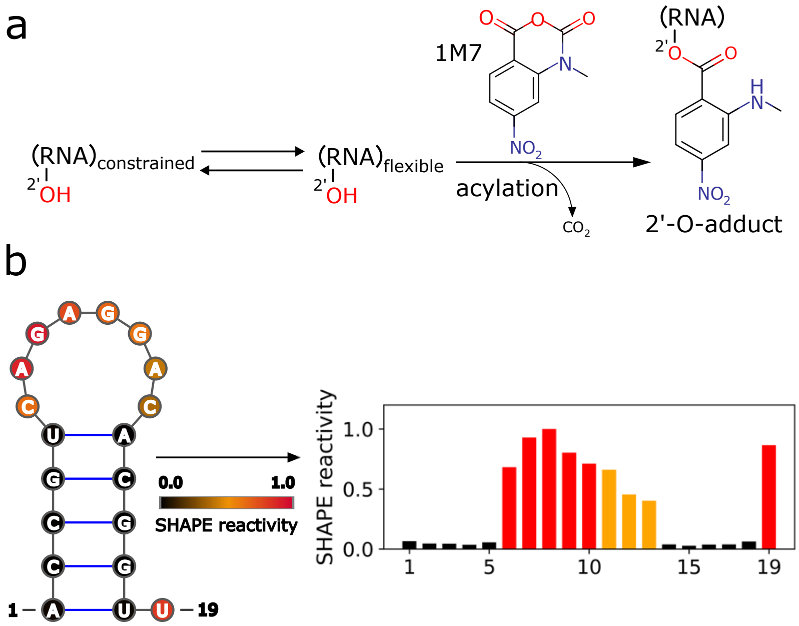


Figure 1.2: SHAPE chemical probing of RNA structure. (a) Illustration of the SHAPE acylation between a single 1M7 and an RNA molecule. (b) Normalized SHAPE reactivities (see scale) shown in both an RNA 2D structure plot and a bar plot.

1.5 Briefly summarizing the main results

1.5.1 Project 1: Developing a machine-learning approach to the prediction of SHAPE reactivity

Machine learning has shown unprecedented success in protein structure prediction [173–178], protein-ligand binding [179–182], regulatory genomics, and cellular imaging [183, 184]. In this project, we investigated the possibility of using a particular machine learning approach—convolutional neural network (CNN)—to see how the machine learning performance can be translated into the SHAPE prediction. We compared the performance of the CNN model to a previously developed analytical model, 3DSSR [171], for a set of RNA with 20 cases. Results have shown that indeed, the machine learning approach can give promising results if large amount of data is available. When only limited data is available, however, we found that the analytical model can provide better predictions. The result highlighted the importance of the size and quality of training data set for successful machine learning approaches.

The work led to a publication: Travis Hurst*, Yuanzhe Zhou*, and Shi-Jie Chen. "Analytical modeling and deep learning approaches to estimating RNA SHAPE reactivity from 3D structure" (* denotes equal first author). In: *Commun. Inf. Syst.* 19.3 (2019), pp. 299-319. DOI: https://dx.doi.org/10.4310/CIS.2019.v19.n3.a4. URL: https://www.intlpress.com/site/pub/pages/journals/items/cis/content/vols/0019/0003/a004/index.php.

1.5.2 Project 2: Developing a software pipeline and web server for predicting SHAPE reactivity

The SHAPE data serves as a convenient and efficient way to probe the RNA local flexibility. The information contained in the SHAPE reactivity for the target RNA can be used to guide the 2D/3D structure prediction [165, 166, 170–172]. To facilitate the research and provide a user-friendly interface for the end-users, we simplified the workflow of our most recent computational model, the reformulated 3DSSR [172], for predicting SHAPE reactivity, and compiled it into a software pipeline. We also provide a standalone web service, SHAPER [185], for our model. By predicting the SHAPE profile for any given RNA 3D structure and calculating the correlation between the predicted and experimental SHAPE profile, the SHAPER web server guides users to select the correct native structure based on experimental SHAPE data.

The work led to a publication: Yuanzhe Zhou, Jun Li, Travis Hurst, and Shi-Jie

Chen. “SHAPER: A Web Server for Fast and Accurate SHAPE Reactivity Prediction”. In: *Front. Mol. Biosci.* (2021), pp. 715. DOI: https://doi.org/10.3389/fmolb.2021.721955. URL: https://www.frontiersin.org/articles/10.3389/fmolb.2021.721955/full.

1.5.3 Project 3: Predicting Mg^{2+} binding sites for a given RNA structure using convolutional neural network (CNN) model

The ability to accurately predict the Mg^{2+} ion binding sites has a far-reaching impact to RNA structure prediction and RNA-targeted drug design. We recently developed a machine-learning-based computational model for predicting Mg^{2+} binding sites. By exploiting the local 3D shape (RNA volume) and electrostatic information associated with experimentally observed Mg^{2+} binding sites, our CNN model is able to achieve higher accuracy and efficiency in binding site prediction than traditional knowledge-based and physics-based models. Besides the comparison between various computational models, we also used the saliency analysis to reveal the most important coordinating atoms on the ion-binding sites. Further investigation on Mg^{2+} binding sites predicted by the CNN model led to the identification of two new Mg^{2+} binding motifs.

This work has been submitted for publication: Yuanzhe Zhou, and Shi-Jie Chen. “A method for decoding nucleic acid-magnesium ion interactions using deep learning convolutional neural network”, 2021.

1.5.4 Project 4: A critical review of computational models for RNA-small molecule interactions

The rapidly growing interest in RNA-targeted drug discovery causes leads to increasing demands for a fast and accurate computational tool that can facilitate the drug screening process. In this work, we presented a critical review for the currently available computational approaches for modeling RNA-small molecule interactions. Our critical assessment of the different models suggest that although recently developed models have led to encouraging improvements in the prediction of binding modes, the accuracy for the predictions of binding modes and binding affinities are generally quite low. We need more accurate models to achieve complete sampling and accurate scoring for RNA-ligand binding modes.

The work led to a publication: Yuanzhe Zhou, Yangwei Jiang, and Shi-Jie Chen. "RNA-ligand molecular docking: Advances and challenges". In: *Wiley Interdiscip. Rev. Comput. Mol. Sci.* (2021), pp. e1571. DOI: https://doi.org/10.1002/wcms.1571. URL: https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/wcms.1571.

1.5.5 Project 5: Developing a statistical potential approach to Identifying small molecule native binding mode

A successful identification of high-quality lead compounds in drug design requires an accurate scoring function for the interactions between RNA and small molecules. We developed a knowledge-based scoring function (SPRank) that uses statistical potential to estimate the binding affinity and to rank the small molecule. Specifically, based on a training data set with 130 experimentally determined RNA-small molecule complexes, SPRank employs an iterative process to derive the pairwise atomic potentials such that using the pairwise atomic potentials, the simulated pairwise distribution agrees with that of the experimental data set. Extensive tests indicate that SPRank outperforms other general scoring functions. A manuscript will be submitted soon to report this new model.

Chapter 2

Analytical modeling and deep learning approaches to estimating RNA SHAPE reactivity from 3D structure

This chapter was published[1].

The selective 2′-hydroxyl acylation analyzed by primer extension (SHAPE) chemical probing method provides information about RNA structure and dynamics at single nucleotide resolution. To facilitate understanding of the relationship between nucleotide flexibility, SHAPE reactivity, and RNA 3D structure, we developed an analytical 3D Structure-SHAPE Relationship (3DSSR) method and a predictive convolutional neural network (CNN) model that predict the SHAPE reactivity from RNA 3D structures. Starting from an RNA 3D structure, the analytical model combines key factors into a composite

[1] Travis Hurst, Yuanzhe Zhou, and Shi-Jie Chen. "Analytical modeling and deep learning approaches to estimating RNA SHAPE reactivity from 3D structure". In: *Communications in Information and Systems* 19.3 (2019), pp. 299–319. DOI: https://dx.doi.org/10.4310/CIS.2019.v19.n3.a4. URL: https://www.intlpress.com/site/pub/pages/journals/items/cis/content/vols/0019/0003/a004/.

function to predict conformational flexibility of each nucleotide and calculate the correlation between the prediction and experimental SHAPE reactivity. Here, we apply the 3DSSR and the deep learning SHAPE model to SHAPE data-assisted RNA 3D structure prediction. We show that the models provide an effective sieve to exclude 3D structures that are incompatible with experimental SHAPE data. Additionally, we compare the 3DSSR analytical model with the CNN deep learning model that recognizes structural and physical/chemical patterns to predict SHAPE data from RNA 3D structure. Depending on the training data set, the analytical model outperforms the deep learning approach for most test cases, indicating that insufficient data is available to adequately train the CNN at this juncture. For other test cases, the deep learning approach provides better predictions than the analytical model, suggesting that the deep learning approach may become increasingly promising as more SHAPE data becomes available.

2.1 Introduction

Galvanized by recent progress in RNA chemical probing technology, researchers developed efficient, data-driven experimental modeling approaches that place effective constraints on RNA structure to complement established template and physics-based methods [187–191]. Selective $2'$-hydroxyl acylation analyzed by primer extension (SHAPE) provides significant insights into local nucleotide structure and dynamics in RNA [163, 164]. SHAPE reagents are small ligands—such as 1-methyl-7-nitroisatoic anhydride (1M7) [192]—that covalently bind to the $2'$-hydroxyl group of a nucleotide (see Fig. 2.1) [193]. Previous studies [167–169] suggest that unconstrained nucleotides have a greater ability to sample more conformations and to adopt SHAPE-reactive postures, which causes them to have higher

SHAPE reactivity. In contrast, nucleotides that are constrained by base-pairing and stacking interactions have a lower propensity to sample a variety of poses and are much less reactive. By quantitatively measuring local nucleotide dynamics, SHAPE is an effective tool for probing whether a nucleotide is constrained by interactions with other nucleotides (in a helix or structured loop) or is located in a flexible loop/junction, without many interactions. In secondary structure modeling, use of SHAPE data substantially improves accuracy and efficiency [165, 166, 194–196], where SHAPE reactivity is used to provide additional structural constraints for free-energy based predictions [197]. Moreover, when used as the basis for advanced experimental approaches, such as differential SHAPE reactivity, mutate-and-map, and time-resolved SHAPE chemistry, SHAPE probing provides helpful information for the *in vitro* and *in vivo* determination of non-canonical tertiary interactions and RNA kinetics [198–204].

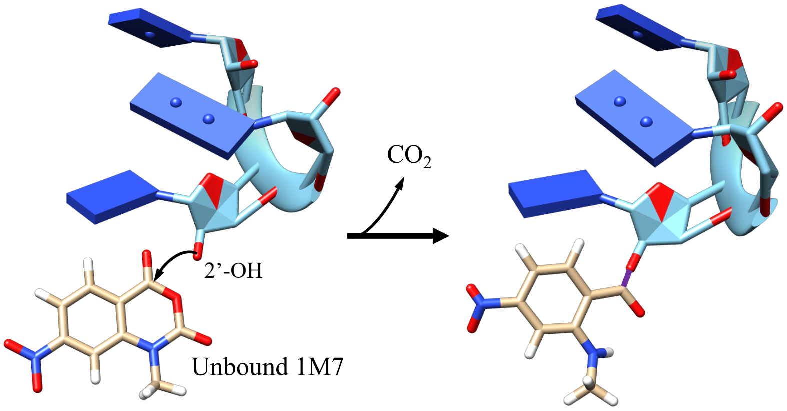


Figure 2.1: The SHAPE reaction. The RNA nucleotide 2′-OH group attacks the reactive carbon of 1M7, releases CO$_2$, and forms a covalent bond (purple) with the SHAPE reagent.

Machine learning is a general method of data analysis that automates analytical model building and is based on the idea that models can learn from data, extract patterns, and make decisions with minimal human intervention. Complex problems without clear underlying mathematical structures benefit from machine learning because manually constructed analytical models cannot easily capture all of the underlying mechanics. The appeal of machine learning methods is the ability to derive predictive models without a need for strong assumptions about underlying mechanisms, which are frequently unknown or insufficiently defined in computational biology. Machine learning has exhibited unprecedented performance in protein structure prediction [173–178], protein-ligand binding [179–182], regulatory genomics and cellular imaging [183, 184]. Deep learning is a subset of machine learning based on artificial neural networks, and "deep" refers to the presence of multiple hidden layers. The convolutional neural network (CNN) is one of the deep learning network models and has gained significant attention due to its success in computer visual recognition.

Previously, we developed an analytical function to quantitatively predict the SHAPE profile from individual RNA 3D structures [171]. We showed how our function can be applied to exclude SHAPE-incompatible structures. To establish the relationship between SHAPE reactivity and nucleotide dynamics, we generated conformational ensembles with MD simulations to measure the correlation between SHAPE reactivity and the conformational propensity of each nucleotide. Then, by combining key factors that account for physical properties implicated in the SHAPE mechanism—the nucleotide interaction strength, SHAPE ligand accessibility, and base-pairing pattern—we developed the analytical 3D Structure-SHAPE Relationship (3DSSR) function, which characterizes the local nucleotide flexibility and predicts SHAPE reactivity based on information about the nucleotide posture and local energetics. To test the discriminating ability of our tool, we used the 3DSSR function to show how SHAPE-incompatible decoy structures may be excluded based on the low correlation between their predicted SHAPE profile and experimental SHAPE data.

Here, we revisit the 3DSSR model and develop a novel convolutional neural network (CNN) model, which uses experimental structural data to predict the SHAPE reactivity for any given nucleotide. First, we briefly describe the formulation of the 3DSSR model on a molecule that was not originally used to test or train either the 3DSSR or CNN model. Then, we describe the methods used to obtain the CNN model. Finally, we compare the ability of the two models to make useful predictions of SHAPE reactivity on RNA molecules used in training and a molecule neither algorithm has seen, emphasizing that analytical formulations often provide more insight than pattern recognition methods when limited data is available.

Table 2.1: RNA structures used for validation. The Protein Database ID (PDBID), length of the RNA in nucleotides (nt), type of RNA, and organism of origin are displayed. The SHAPE profiles for these RNA molecules are from the published experimental data [167, 195, 196, 205, 206].

PDBID	Length (nt)	Type of RNA	Organism
2L8H	29	TAR RNA	*HIV-1*
1AUD	30	U1A protein binding site RNA	*H. sapiens*
2L1V	36	M-box riboswitch	*B. subtilis*
2K95*	48	Telomerase pseudoknot	*H. sapiens*
1Y26	71	Adenine riboswitch	*V. vulnificus*
1VTQ	75	PreQ1 riboswitch aptamer	*B. subtilis*
1EHZ	76	Aspartate tRNA	*Yeast*
1P5O*	77	IRES Domain II	*Hepatitis C*
2GDI	79	TPP riboswitch	*E. coli*
3IWN	93	Cyclic-di-GMP riboswitch	*V. cholera*
4KQY	117	SAM-I riboswitch	*B. subtilis*
1C2X*	120	5S rRNA	*E. coli*
3IVK*	128	Catalytic core of RNA polymerase ribozyme	*E. coli*
1NBS	154	Specificity domain of Ribonuclease P RNA	*B. subtilis*
3PDR	154	M-box riboswitch	*B. subtilis*
1GID*	158	Group 1 Ribozyme	*Synthetic*
3P49*	169	Glycine Riboswitch	*H. sapiens*
3DIG	174	Lysine riboswitch	*T. maritima*
4UE5*	299	SRP RNA	*C. lupus*
3G78*	421	Group II intron	*O. iheyensis*

* Denotes cases used to parameterize the CNN model, not the 3DSSR model.

2.2 Methods

2.2.1 Finding structures corresponding to SHAPE data

In order to find RNA structures that correspond to our SHAPE sequences, we used the sequence searching interface equipped with NCBI's BLAST (Basic Local Alignment Search Tool) program [207] provided by RCSB protein databank [208] to align the sequences. In the 3DSSR (CNN20) model 12 (20) RNA structures with an average length of $\sim$92 (120) nucleotides that have SHAPE reactivity data were used (see Table 2.1). For comparison, we also parameterized the CNN model using the same 12 structures as 3DSSR (CNN12). SHAPE reactivity data came from databases for sharing nucleic acid chemical probing data, the RNA mapping database (RMDB) [206] and the SNRNASM database [205]. To have comparable SHAPE reactivity values between different RNA structures, all of the negative values of SHAPE reactivity data are set to zero, in accordance with previous work [166]. Furthermore, the SHAPE profiles are scaled by the maximum reactivity value of each respective RNA structure, which confines SHAPE reactivity data to range from 0 to 1.

2.2.2 Reviewing 3DSSR methods and conclusions

Previously, we used simulations to show that SHAPE data corresponds with nucleotide flexibility, parameterize the 3DSSR model, and generate decoys to illustrate how our model can be used to exclude SHAPE-incompatible structures [171]. The ability of a nucleotide to react with SHAPE depends on the propensity of a nucleotide to sample SHAPE-reactive postures and the ability of the SHAPE ligand to access the reactive site. Capturing these

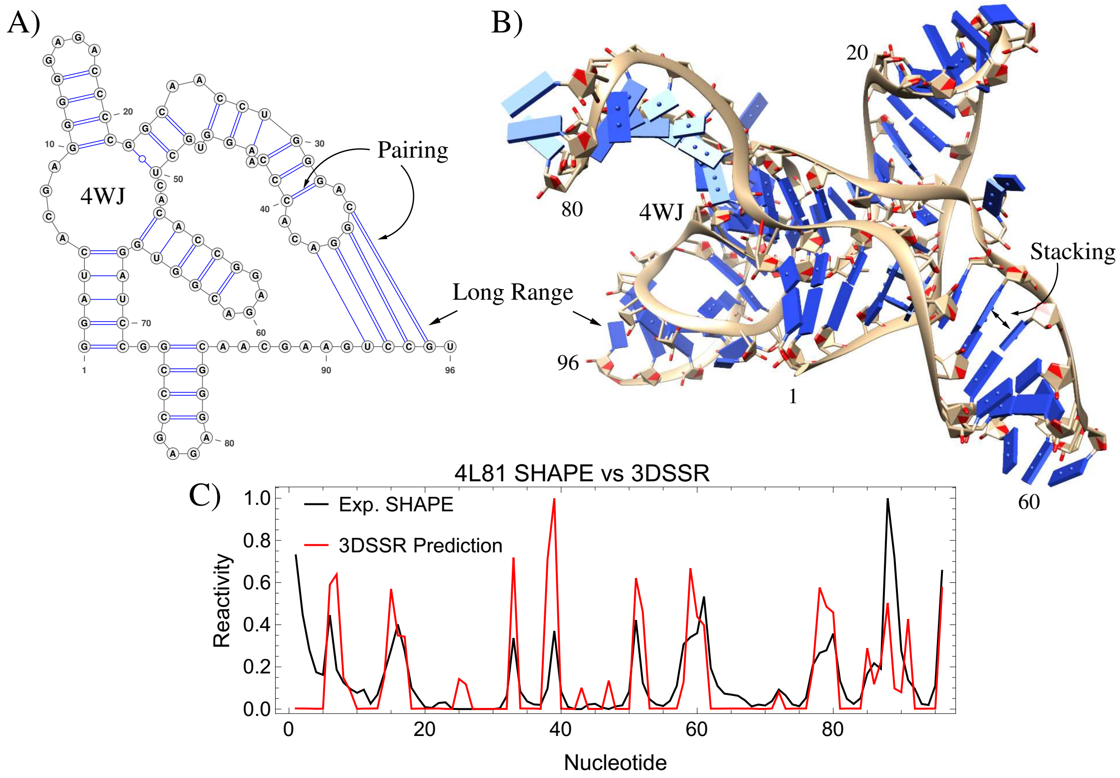


Figure 2.2: The 2D, 3D, and SHAPE reactivity of RNA-Puzzle 8 (PDBID: 4L81). A) The 2D structure [209] shows the four-way junction (4WJ), base-pairs, and long range interactions. B) The 3D structure shows the 4WJ and an example of a base stacking interaction. C) The experimental and predicted SHAPE profiles for the crystallized 4L81 structure show good agreement (Pearson correlation = 0.57).

concepts, we proposed the 3DSSR function

$$P(n) = BP(n) \cdot \frac{SAS(n) + S_0}{|II(n) - 1.0|} \tag{2.1}$$

to estimate the nucleotide stability and predict the SHAPE reactivity $P(n)$ for a nucleotide n. The base-pairing factor $BP(n)$ accounts for the 2D structure, which is characterized by the base-pairing pattern: a nucleotide n in a helix region is assigned $BP(n) = 0.01$ and a nucleotide in a loop or junction region is assigned $BP(n) = 1.0$. A 2D structure can always be extracted from a 3D structure (for example, using the RNApdbee 2.0 webserver [210]), and helix nucleotides are normally SHAPE-inert. The SHAPE ligand accessible 2′-OH surface area $SAS(n)$ describes the necessary requirement of a SHAPE ligand to access the nucleotide for a reaction to occur. If a nucleotide 2′-OH is buried inside the RNA structure, SHAPE reagents cannot react, which reduces the SHAPE reactivity. The unbound SHAPE ligand has an effective radius between 2.0 and 2.5 Å, and our results indicate that the 3DSSR function is not sensitive to different probe sizes within this range. The accessible surface of 2′-OH is calculated using VMD [211]. S_0 is a constant, accounting for the ability of a nucleotide to become accessible during experimental SHAPE probing. $II(n)$ is the interaction intensity for nucleotide n, which accounts for tertiary structure interactions. Through fitting, information from base-pairing and base-stacking interactions are combined to calculate the $II(n)$, a quasi-energy score for each nucleotide.

In the present study, we focus on a SHAPE data-assisted approach to RNA 3D structure prediction. For a given RNA sequence, we can generate an ensemble of possible conformations using, for example, the IsRNA coarse grained simulation model [212]. We then score each conformation by the correlation (similarity) between the (3DSSR-predicted) SHAPE profile of the conformation and the experimentally determined SHAPE data for the RNA molecule. Although due to the low-resolution energy model, the 3DSSR model might not be able to identify the native, crystal structure from SHAPE data alone, as shown below, the model can assist structure prediction by successfully excluding SHAPE-incompatible structures.

2.2.3 Applying the model to the SAM-I/IV riboswitch aptamer

For illustration, here we apply the 3DSSR model to the SAM-I/IV riboswitch aptamer (PDBID: 4L81) that was used in round 8 of RNA-Puzzles [213] (see Fig. 2.2), a community-wide, CASP-like blind test for RNA 3D structure prediction. This structure has not been previously used to train or test the 3DSSR model, and the structures submitted in the RNA-Puzzle competition by different labs give us objective decoys to show the ability of the 3DSSR model to exclude structures that are incompatible with SHAPE.

First, we access the submitted structures and assessment results from the RNA-Puzzles database (see Fig. 2.3A for a structure submitted to the competition). Next, we extract the 2D structures from the submitted 3D structures using the RNApdbee 2.0 webserver [210] (Fig. 2.3B). After that, we use RNAview software to identify the base pair types from the 3D structures [214] (Fig. 2.3C). Additionally, we directly calculate the stacking interaction information from the 3D structures: the angles and distances between different RNA bases (Fig. 2.3D). Then, we calculate the solvent accessible surface of each nucleotide $2'$-OH in the 3D structures with VMD [211] (see Fig. 2.3E for a visual representation). Finally, we use the 3DSSR function to combine all of the structural information and predict the SHAPE reactivity for each nucleotide (Fig. 2.3F). To evaluate the SHAPE-compatibility, we also calculate the Pearson correlation between the experimental and 3DSSR-predicted SHAPE profiles. Comparison of the 3DSSR-predicted and experimental SHAPE profiles on the native, crystal structure can be seen in Fig. 2.2C. Provided with candidate 3D structures and experimental SHAPE data, we can exclude SHAPE-incompatible structures on the basis of their 3DSSR-predicted SHAPE profile.

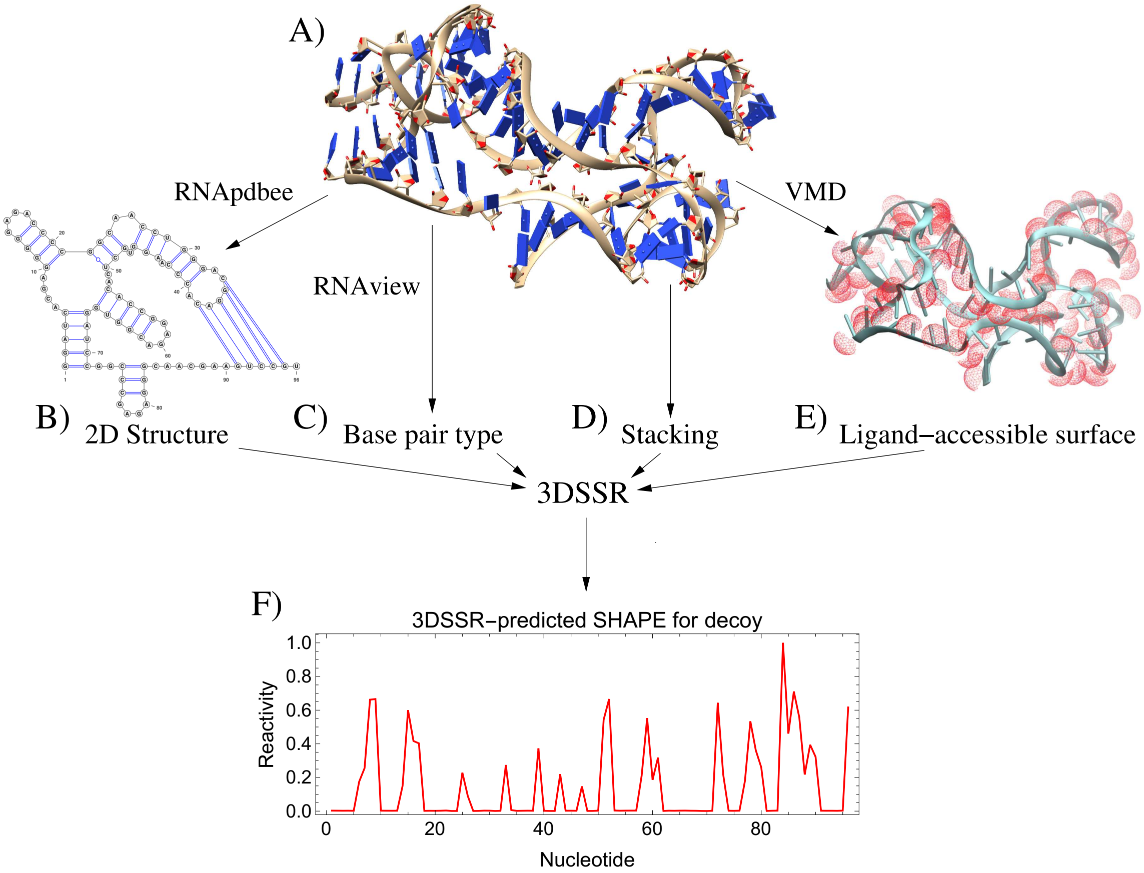


Figure 2.3: 3DSSR workflow on an RNA-Puzzle 8 decoy. A) A candidate 3D structure decoy is processed by RNApdbee 2.0 [210], RNAview [214], in-house software, and VMD [211] to B) produce a 2D structure, C) identify base pair types, D) extract stacking angle/distance information, and E) calculate the solvent accessible surface of the $2'$-OH, respectively. The information extracted from the structure is input into the 3DSSR function to produce F) the predicted SHAPE profile for each nucleotide.

The sensitivity of the model to structures with high RMSD and lower Interaction Network Fidelity (INF; a quantity to measure the similarity in interaction pattern) [215] can be seen in Fig. 2.5, where we apply the 3DSSR model on all of the submitted structures for

RNA-Puzzle 8 to show the ability of the 3DSSR function to exclude SHAPE-incompatible structures. Contributing to the objectivity of the test, the submitted 3D structures and assessment results (values of RMSD and INF for each structure) for RNA-Puzzle 8 were all taken from the RNA-Puzzles database. The results suggest that many of the 43 submitted structures could be discarded because they are incompatible with SHAPE. For example, the native crystal structure is ranked in the top ten, and we could comfortably discard the bottom 20 structures, which all have a correlation < 0.45. Only one structure ranked in the bottom 20 by the 3DSSR model has RMSD (INF) $<$ ($>$) 11.2 (0.80), and no structure in the bottom 20 has favorable assessment values for both RMSD and INF. As can be seen in Fig. 2.5A, the combination of assessment results indicate that a cutoff of 0.45 is quite conservative. We could discard the bottom 65 percent of structures (the bottom 28), which would keep all of the structures with favorable assessment results for both RMSD and INF. For RNA-Puzzle 8, discarding more than the bottom 70 percent would cause us to discard the native structure. However, the quality of the SHAPE data, the size of the RNA, and the quality of candidate structures all affect the number of structures that may be comfortably excluded on the basis of SHAPE data using 3DSSR. These factors should be known so that a reasonable sieving scheme can be found.

2.2.4 Using a CNN to predict SHAPE reactivity from structure

Describing the nucleotide environment

In agreement with SHAPE experiments, our CNN method probes RNA structure at single nucleotide resolution. For each nucleotide, the surrounding environment refers to neighboring atoms within a cubic volume of space around the nucleotide. Since we define the space surrounding a nucleotide as the space confined in a cube, the environment captured by this cube is not rotationally invariant. To remove the effects caused by the different choices of the cube orientation, we set a local Cartesian coordinate system for every given nucleotide. The coordinate system of a nucleotide is determined by the C1′, C4′, and O4′ atoms. Specifically, the origin of the local coordinate system is located at the atom O4′, and the local $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$ axes are defined as follows. First, we denote the $\mathbf{r_{C1'}}$, $\mathbf{r_{C4'}}$, and $\mathbf{r_{O4'}}$ as the coordinates of the selected atoms, C1′, C4′, and O4′. Second, we calculate three vectors $\mathbf{v_x}$, $\mathbf{v_y}$ and $\mathbf{v_z}$ with respect to the local origin as

$$\begin{aligned} \mathbf{v_x} &= \mathbf{r_{C4'}} - \mathbf{r_{O4'}} \\ \mathbf{v_y} &= \mathbf{r_{C1'}} - \mathbf{r_{O4'}} \\ \mathbf{v_z} &= \mathbf{v_x} \times \mathbf{v_y} \end{aligned} \tag{2.2}$$

where $\mathbf{v_x}$ represents the vector from atom O4′ to atom C4′, $\mathbf{v_y}$ represents the vector from atom O4′ to atom C1′ and $\mathbf{v_z}$ is just the cross product of $\mathbf{v_x}$ and $\mathbf{v_y}$. Then, the $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{z}$ axes are set according to the following Eq. 2.3,

$$\begin{aligned} \mathbf{x} &= \frac{\mathbf{v_x}}{\|\mathbf{v_x}\|} \\ \mathbf{z} &= \frac{\mathbf{v_z}}{\|\mathbf{v_z}\|} \\ \mathbf{y} &= \mathbf{z} \times \mathbf{x} \end{aligned} \tag{2.3}$$

The surrounding environment of each nucleotide is captured through a cube centered and oriented according to the local coordinate system. As shown in Fig. 2.4, the length of the cube is 24 Å and the atoms contained in the cube will be used to generate the image for CNN model.

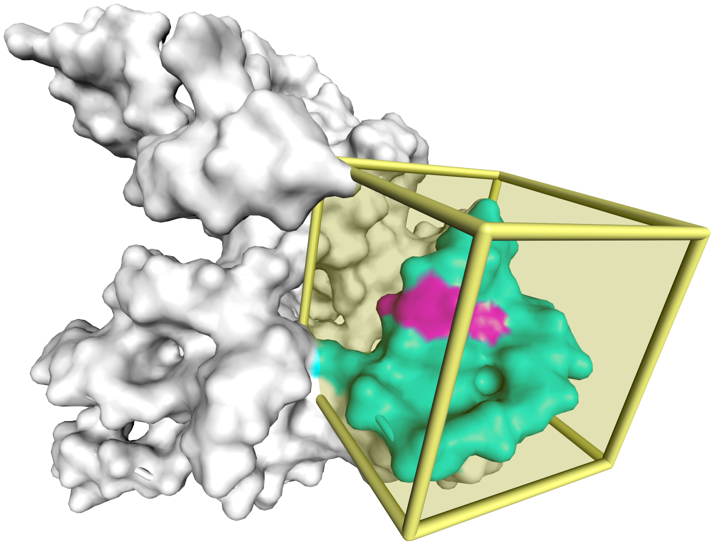


Figure 2.4: . Extracting the 3D image of an RNA nucleotide. The magenta color depicts the nucleotide under assessment and the surrounding environment is confined within the cube with length 24 Å. The surrounding atoms are drawn in cyan, and the cube boundaries are drawn with yellow solid lines.

Input: defining the 3D image as input into the CNN

As we described in the previous section, a 24 Å × 24 Å × 24 Å cube is used to provide the surrounding environment of each nucleotide. The corresponding image associated with this nucleotide is contained within this cube. As a normal 2D digital image has three color channels (RGB) with each channel represented by a 2D pixel matrix, the 3D image that we used to capture the surrounding environment is also composed of multiple channels. However, our 3D images do not simply use RGB color channels: the channels we selected represent certain physical or chemical features. In our CNN model, we defined 5 channels, which are fully described in Table 2.2.

Since we extract our 3D image from a cube, each channel of the 3D image is represented by a 3D matrix, and each position in this 3D matrix has a voxel (3D pixel) value. We set the length of our 3D image equal to the cube with an image resolution of 1 Å, so each

Table 2.2: Feature channels used for 3D images.

Feature	Description
Hydrophobic	Aliphatic or aromatic carbon atoms
Aromatic	Aromatic carbon atoms
Positive ionizable	Gasteiger positive charge
Negative ionizable	Gasteiger negative charge
Excluded volume	All atom types

voxel has a dimension of 1 Å × 1 Å × 1 Å. A step function fills the voxels of each channel. For example, the voxels of the excluded volume channel that are occupied by RNA atoms are filled with 1, and the rest are filled with 0, according to their Van der Waals radius. A similar procedure was used to generate other channels.

Describing the CNN architecture

Our CNN model takes the multi-channel images as input, and outputs a predicted SHAPE reactivity for each image. The network is a basic ResNet [216] architecture with only slight modifications and has 10 convolutional layers. The detailed architecture is shown in Table 2.3. The first layer accepts the 3D image in a convolutional layer and has 64 $7 \times 7 \times 7$ filters with a stride of 2. The next layer has 4 residual blocks, with each block containing two convolutional layers. Downsampling is directly performed in the first convolutional layer and by the beginning convolutional layers of blocks 2-4. Finally, the network ends with a global average pooling layer and a 512-way fully-connected layer with a sigmoid activation function. Except the first layer, all of the convolutional layers use $3 \times 3 \times 3$ sized filters. Batch normalization [217] was applied right after each convolutional layer and before 'Rectified Linear Unit' [218] activation, following [217]. In our network, two hidden layers inserted residual shortcut connections for every block. The shortcut takes an identical input from the previous block and maps this identity shortcut right before the activation of the second hidden layer within the block; the block is same as the original ResNet block [216]. We initialize the weights as in [216, 219] and train all residual nets from scratch. The only preprocessing we used is the subtraction of a mean value from each image. This mean value is calculated by averaging all the voxels of all images in the training set.

Table 2.3: Details of CNN Architectures. Each building block is shown with two convolutional layers. Downsampling is performed in every convolutional layer with a stride of 2.

	Layer name	Output size	Filter size	Filter num
first layer	conv1	$12 \times 12 \times 12$	$7 \times 7 \times 7$	64, stride 2
block1	conv2	$12 \times 12 \times 12$	$3 \times 3 \times 3$	64, stride 1
block1	conv3	$12 \times 12 \times 12$	$3 \times 3 \times 3$	64, stride 1
block2	conv4	$6 \times 6 \times 6$	$3 \times 3 \times 3$	128, stride 2
block2	conv5	$6 \times 6 \times 6$	$3 \times 3 \times 3$	128, stride 1
block3	conv6	$3 \times 3 \times 3$	$3 \times 3 \times 3$	256, stride 2
block3	conv7	$3 \times 3 \times 3$	$3 \times 3 \times 3$	256, stride 1
block4	conv8	$2 \times 2 \times 2$	$3 \times 3 \times 3$	512, stride 2
block4	conv9	$2 \times 2 \times 2$	$3 \times 3 \times 3$	512, stride 1
last layer	fc	$1 \times 1 \times 1$	average pool, 512-d fc, sigmoid	

For the network optimizer, we used Adam [220] with default parameters for momentum scheduling ($\beta_1 = 0.99, \beta_2 = 0.999$) provided by PyTorch [221], and a mini-batch size of 128 was used for training. The learning rate started from 0.01 and was divided by 10 when the training accuracy plateaued, and the models were trained for up to 100 epochs. For our loss function, we calculated the mean square error (MSE) loss between predicted SHAPE reactivities and experimental SHAPE reactivities as

$$\text{Loss} = \sum_{n=1}^{N} (P_n - G_n)^2 / N \tag{2.4}$$

where N is the number of images and $P_n(G_n)$ is the predicted(experimental) SHAPE reactivity for image n.

Output: predicting SHAPE reactivity with a CNN

For any given 3D image that describes the surrounding environment of the considered nucleotide, our CNN model will output a real number characterizing the predicted SHAPE reactivity. This output value is confined within range from 0 to 1.

Implementation and cross-validating

Based on the SHAPE data for 20 RNAs (totally 2455 nucleotides) collected by different experimental labs, we have 2455 SHAPE data along with the corresponding high-resolution atomic coordinates for all the nucleotides and their pertinent physical and chemical parameters. All the data together serve as the input for the CNN. To test and validate the deep learning approach, we used the leave-one-out cross-validation method to validate the performance of our model. Each time, our model was trained on 19 RNA cases with corresponding SHAPE reactivity data and tested on 1 RNA case. This process was carried out 20 times, leaving out each RNA in turn. The overall performance is evaluated by averaging the Pearson correlation coefficients of the 20 test cases over the leave-one-out process. We also carried out this procedure for the 12 cases used to parameterize the 3DSSR function. The results of the cross-validation process are summarized in Table 2.4. The Pearson correlation coefficient was used to measure the similarity between the predicted SHAPE profile and the experimentally derived SHAPE profile. For each training and validation set in the cross-validation, we chose the model that has the best performance on the validation set to avoid overfitting.

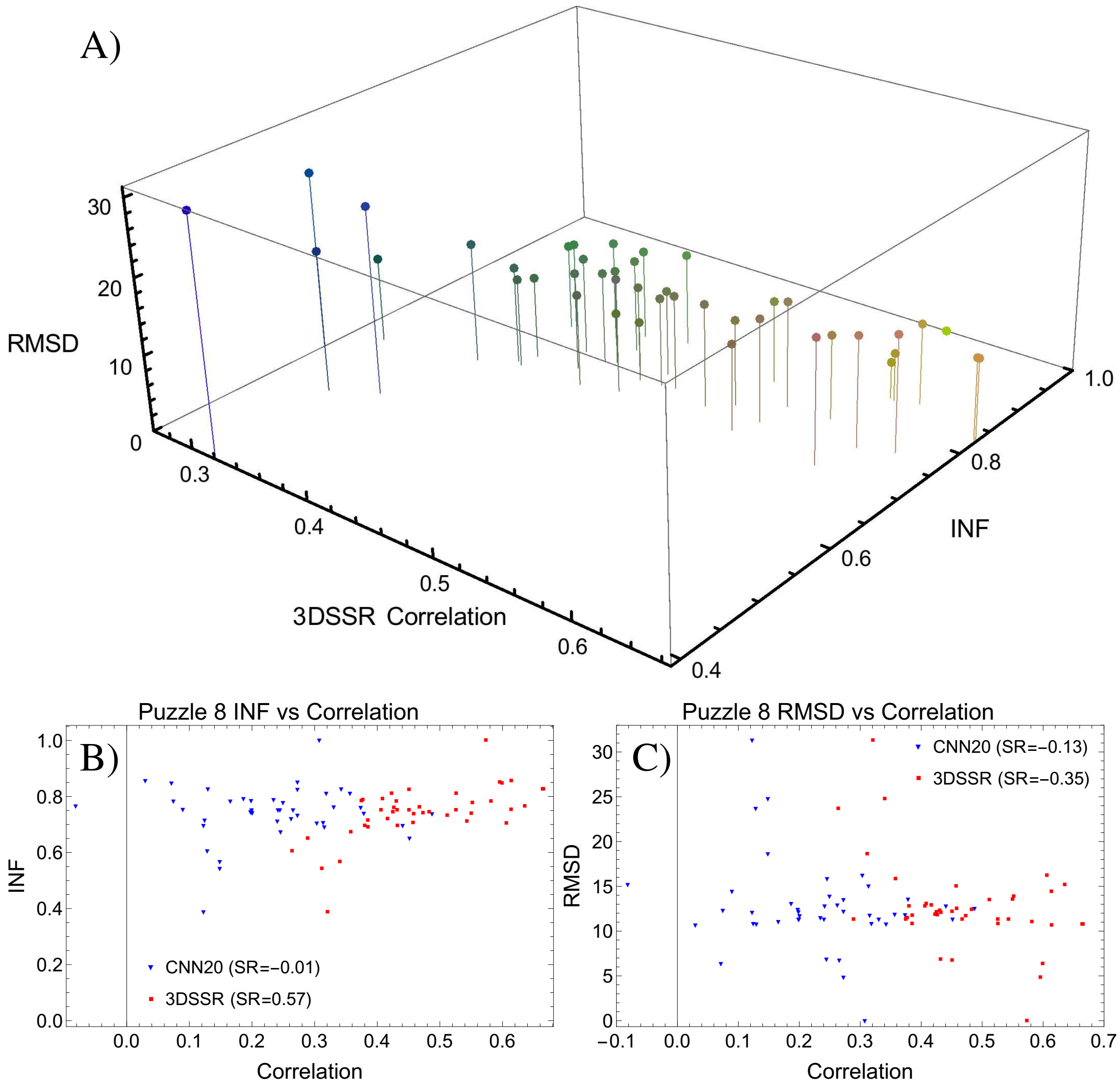


Figure 2.5: Sieving SHAPE-incompatible structures from RNA-Puzzle 8 submissions. A) The 3D representation shows the trend of the assessment results (RMSD and INF) with the correlation between 3DSSR-predicted SHAPE profiles and experimental SHAPE (3DSSR Correlation). Warmer colors indicate higher correlation, higher INF, and lower RMSD. The INF and RMSD values were taken from the RNA-Puzzles database. 2D plots of the B) INF and C) RMSD with respect to the 3DSSR and CNN20 correlations are also shown, along with their respective Spearman rank coefficients (SR).

2.2.5 Comparing 3DSSR to CNN models

As can be seen in Table 2.4, the 3DSSR model generally outperforms the CNN model, regardless of whether 20 or 12 structures are used to train the CNN. In contrast, the CNN model performs substantially better on 3PDR, which may indicate that information in the structure of 3PDR leading to its SHAPE reactivity profile is contained in the other cases. Because 3PDR has high performance in the CNN model in spite of its length, we may expect improvements in other cases once the amount of training data is increased. The relatively poor performance in other cases may indicate that factors that contribute to SHAPE reactivity in those RNA are not adequately represented by the structures provided in the training set. In addition, the small fluctuations captured in the 3DSSR model by using solvated, near-native representations to fit the unknown parameters may help boost its performance over the CNN.

However, the correlation alone does not show us the discerning ability of the 3DSSR and CNN models on decoy structures. For that, we turn to the results on RNA-Puzzle 8, where the Spearman rank correlation coefficient (SR) can tell us how well the models perform on ranking the structures in comparison to objective assessments (RMSD and INF). For INF(RMSD), the SR values were 0.57(-0.35) and -0.01(-0.13) for 3DSSR and CNN20, respectively, which shows that 3DSSR markedly outperforms the CNN20 model on both ranking assessments and can be used to exclude more SHAPE-incompatible structures (see Fig. 2.5BC).

Table 2.4: Pearson correlations between the experimental SHAPE data and the prediction algorithms: 3DSSR and the cross-validated CNN model trained on 11(19) cases and tested on the one left out, denoted as CNN12(CNN20).

PDB	Length (nt)	3DSSR	CNN20	CNN12
2L8H	29	0.96	0.85	0.87
1AUD	30	0.92	0.90	0.71
2L1V	36	0.83	0.81	0.79
1Y26	71	0.88	0.52	0.66
1VTQ	75	0.71	0.71	0.80
1EHZ	76	0.80	0.77	0.78
2GDI	79	0.89	0.81	0.66
3IWN	93	0.74	0.33	0.38
4KQY	117	0.75	0.58	0.64
1NBS	154	0.61	0.48	0.34
3PDR	154	0.61	0.81	0.83
3DIG	174	0.70	0.64	0.68
Average	92	0.78	0.68	0.68

2.3 Conclusion

Efficient chemical probing methods, like SHAPE, provide a wealth of information about RNA structure and dynamics. By formulating an analytical expression that captures the key factors determining SHAPE reactivity, we can predict SHAPE reactivity from individual RNA structures. After computing predictive SHAPE profiles for a set of candidate RNA 3D structures, we can sieve the structures based on the correlation between the predicted and experimental reactivities, and SHAPE-incompatible structures can be excluded. This general method of combining efficient experimental data with computational sieving may be transferred to other efficient probing methods, enabling more confident computational determination of RNA tertiary structure at lower cost.

Machine learning techniques are incapable of creating new concepts and require the training data to be a good representative of the test data. To put it simply, a dog classification model trained with only dog images can not be used to classify cats; the model cannot be generalized to predict information it has never seen during training. Because we only have 20 RNA structures with SHAPE reactivity profiles in our data set, there is a good chance that nucleotides in a test RNA are not well represented by the other 19 structures, which results in worse performance. Additionally, using features that are important for determining SHAPE reactivity of a given nucleotide can greatly facilitate the learning process. However, finding the right combination of image channel features is not easy since the underlying mechanism that governs SHAPE reactivity is still unclear.

Although the mechanism that governs SHAPE reactivity is not fully understood, our general understanding is enough to formulate a relatively simple analytical function—the 3DSSR model—to predict reactivity based on the sensitivity of SHAPE to local nucleotide dynamics and the accessibility of SHAPE-reactive nucleotides. Because there is not enough data to apply a trained, pattern recognizing CNN to new structures, our manually constructed, analytical 3DSSR function is better at ranking structures on the basis of experimental SHAPE data. Although machine learning and advanced data-processing methods are leading to rapid advances on many problems with ample data and unclear underlying mathematical structure, physics-based models can perform better in systems where limited data is available and underlying mechanisms are known well enough to mathematically express the mechanics, even if the mechanisms are incompletely understood. As more data becomes available, we expect performance of the CNN model to improve. In the meantime, we recommend using expressions of the underlying mechanics to predict SHAPE reactivity for guiding RNA structure prediction.

Chapter 3

SHAPER: A Web Server for Fast and Accurate SHAPE Reactivity Prediction

This chapter was published[1].

Selective $2'$-hydroxyl acylation analyzed by primer extension (SHAPE) chemical probing serves as a convenient and efficient experiment technique for providing information about RNA local flexibility. The local structural information contained in SHAPE reactivity data can be used as constraints in 2D/3D structure predictions. Here, we present SHAPE predictoR (SHAPER), a web server for fast and accurate SHAPE reactivity prediction. The main purpose of the SHAPER web server is to provide a portal that uses experimental SHAPE data to refine 2D/3D RNA structure selection. Input structures for the SHAPER server can be obtained through experimental or computational modeling. The SHAPER server can accept RNA structures with single or multiple conformations, and the predicted SHAPE profile and correlation with experimental SHAPE data (if provided) for each con-

[1] Yuanzhe Zhou et al. "SHAPER: A Web Server for Fast and Accurate SHAPE Reactivity Prediction". In: *Frontiers in Molecular Biosciences* 8 (2021), p. 715. ISSN: 2296-889X. DOI: 10.3389/fmolb.2021.721955. URL: https://www.frontiersin.org/article/10.3389/fmolb.2021.721955.

formation can be freely downloaded through the web portal. The SHAPER web server is available at http://rna.physics.missouri.edu/shaper/.

3.1 Introduction

With the development of novel ribonucleic acid (RNA) structure determination methods alongside discoveries of new RNA structures and cellular functions, RNA has become increasingly important, contributing new avenues in the development of therapeutic applications for human diseases. Computational modeling of RNA structures could greatly deepen our understanding of RNA folding mechanisms. However, computational prediction of RNA structures from the sequence remains a significant unsolved problem [222–224].

Although lacking complete structural information, some experimental methods can provide useful details for guiding structure prediction. The selective 2′-hydroxyl acylation analyzed by primer extension (SHAPE) method is a convenient and efficient RNA structure probing technology with single nucleotide resolution that can provide information about local nucleotide structural dynamics [163, 164]. The SHAPE reactivity of a nucleotide is reflected by the ability to bind SHAPE reagents—small ligands such as 1-methyl-7-nitroisatoic anhydride (1M7)—that preferentially bind to the oxygen of 2′-hydroxyl group of RNA nucleotides [193]. Previous studies [167–169] suggested that SHAPE reactivity is correlated with nucleotide flexibility, where unconstrained nucleotides tend to be more reactive while nucleotides constrained by base pairing, stacking, or other interactions are less reactive. The signals seen in SHAPE experiments intrinsically reflect interactions in the 3D structure, and can therefore be used to place effective constraints on the possible structures in a conformational pool generated by computational modeling software.

Since many RNA structure prediction studies would benefit from utilizing experimental SHAPE data, having a freely available, dedicated web server for rapidly predicting SHAPE profiles and filtering structural ensembles is essential. In this paper, we present our SHAPE predictoR (SHAPER) web server for predicting the SHAPE profile of any given RNA structure. The organization of the server is shown in Fig. 3.1. The SHAPER server only requires the 3D coordinates of the target RNA (in PDB format). These structures can come from experimental structures, simulation snapshots, or computational structure-prediction models, etc. The SHAPER server can accept either individual structures or a structural ensemble, and the output contains predicted SHAPE profiles with the correlations between predicted profiles and a provided experimental SHAPE profile (if available). The engine powering the SHAPER web server is the new SHAPE prediction model [172], which is an updated version of the original 3D Structure-SHAPE Relationship (3DSRR) model [171]. The SHAPER model incorporates RNA sequence-dependent bias into the prediction and is able to provide higher correlations between SHAPE data and the native RNA structure, which improves our ability to discern between SHAPE-compatible and -incompatible structures on decoys than the previous 3DSRR model.

3.2 Materials and Methods

3.2.1 Workflow

The following shows both the workflow and theoretical background of the SHAPER server. Detailed description and analysis of the SHAPER model can be found in the original pa-

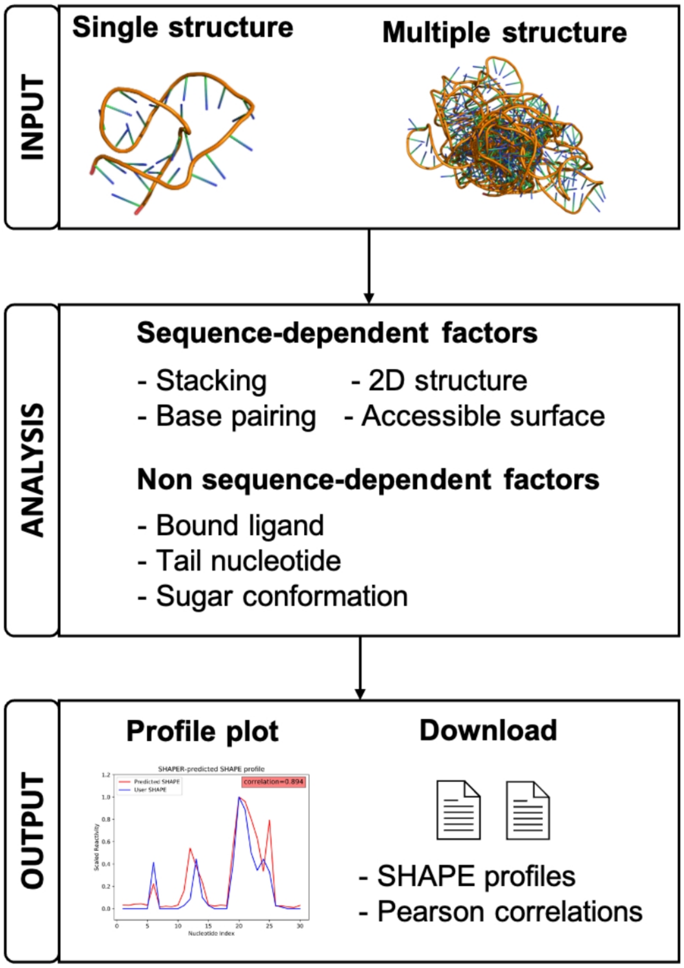


Figure 3.1: A schematic view of the organization and function of the SHAPER web server.

per [172].

Step 1: Uploading Input Data

As shown in Fig. 3.2, Step 1, the input parameters are the following: (1) the input RNA structure file in PDB format, (2) user provided SHAPE profile, (3) user provided MASK file for the target RNA (for masking nucleotides that interact with ligands), (4) an email address for delivery of the calculation results, and (5) a simple text verification to prevent robotic usage. Required parameters are labeled by red asterisks. After submitting the job, the user will be redirected to a waiting page (Fig. 3.2, Step 2), where they can view information about the running job. The information shown in the table in Fig. 3.2, Step 2 includes: JobID—an identification code used to look up the results— and the file names of the RNA, SHAPE, and MASK file uploaded by user, respectively.

Step 2: Calculating SHAPE on Server Side

After submitting the job, SHAPER will put the job in a queue and will run the job once the computational resources are available. Usually, it takes less than a minute for a single structure with around 100 nucleotides. The procedures taken by the SHAPER server are listed in the order of execution.

- *Validating input.*

  The input RNA file (in PDB format) is checked before any further processing. Entities other than RNA will be removed from the PDB file, only the backbone of modified residues and the first occurrence of atoms with multiple alternative locations will be kept for SHAPE reactivity calculations.

- *Identifying base pairing and stacking interactions.*

  Base pairs are identified by RNAView [214], while stacking nucleotides are identified by our in-house Perl script. Then, pairing and stacking energies are combined into the interaction energy score (IE) for a given nucleotide i as

$$E_{\text{IE}}(i) = \sum_{m}[A \cdot E_{\text{bp}}^{(t)}(i,m) + B] + \sum_{k} E_{\text{st}}^{(i)}(i,k) \tag{3.1}$$

where all the type-t base pairing energies $E_{bp}^{(t)}(i,m)$ and all the stacking energies $E_{st}^{(i)}(i,k)$ of nucleotide i are summed together. A and B are two extra parameters trained for the SHAPER model. The base pairing interactions $E_{bp}^{(t)}(i,m)$ were derived through a quasi-chemical statistical potential approach based on the statistical frequencies of the base pairing interactions extracted from the non-redundant RNA Basepair Catalog [225], and the stacking energies introduce $5' \rightarrow 3'$ polarity-dependence by using different weights and energy parameters for upstream $5'$ and downstream $3'$ nucleotides, respectively.

- *Extracting 2D structure.*

  Using the Dissecting the Spatial Structure of RNA (DSSR) tool [226], the 2D structure is extracted from the input 3D structure. A parameter $E_{2D}(i)$ is introduced to represent the energy contributed by the base pairing nucleotide i in the 2D structure.

- *Accounting for other structural features.*

  (1) Ligand Accessible Surface (A_{SAS}). The accessibility of the SHAPE reagent (1M7) to the $2'$-hydroxyl of each nucleotide is calculated using Visual Molecular Dynamics (VMD) [211] with a bead radius of 2.0 Å.

  (2) Ribose sugar conformations. Previous studies [227, 228] suggest that the conformation of the ribose sugar is important for SHAPE-reactivity. A correction F_{sug} determined by the pseudorotation angle of the ribose is employed to account for this effect.

  (2) Tail nucleotides. Simple parameter F_{term} on terminal nucleotide is used to account for the effect of the short nucleotide sequence added at the terminal regions during SHAPE experiments.

  (3) Bound ligands. Nucleotides interacting with a bound ligand need different treat-

ment. To account for these effects, a ligand binding energy penalty E_{lig} is introduced for the nucleotides that are interacting with bound ligands. This is achieved by masking the nucleotides that interact with the ligand. Users can supply their own mask file when submitting jobs on the web server, supplying 0 and 1 for non-interacting and interacting nucleotides, respectively. By default, the SHAPER server will treat all nucleotides as not interacting with ligand.

- *Accounting for the effects of neighboring nucleotides.*

  Due to observations that a free nucleotide next to rigid nucleotides will be less reactive than a free nucleotide that has flexible neighbors, we introduce a weighted averaging scheme to account for this type of correlative effect for E_{IE}, E_{2D}, and A_{SAS} terms as

$$\bar{E}_{\text{IE}}(i) = \frac{\sum_{j=0}^{3} w_j \times E_{\text{IE}}(i+j-1)}{\sum_{j=0}^{3} w_j} \tag{3.2}$$

$$\bar{E}_{\text{2D}}(i) = \frac{\sum_{j=0}^{3} d_j \times E_{2D}(i+j-1)}{\sum_{j=0}^{3} d_j} \tag{3.3}$$

$$\bar{A}_{\text{SAS}}(i) = \frac{\sum_{j=0}^{3} a_j \times A_{\text{SAS}}(i+j-1)}{\sum_{j=0}^{3} a_j} \tag{3.4}$$

  where $w_0 - w_3$, $d_0 - d_3$, and $a_0 - a_3$ are weights accounting for the influence of interactions involving the nucleotide of interest (NOI) and/or neighboring nucleotides.

- *Predicting the SHAPE profile.*

  The final SHAPE prediction is a combination of the interaction factors, written as

$$p_i = SF_i \times e^{SE_i} \tag{3.5}$$

where structural factors SF_i and energy-like scores SE_i are determined by

$$SF_i = (\bar{A}_{SAS}(i) + A^0_{SAS}) \times F_{sug}(i) \times F_{term}(i) \tag{3.6}$$

$$SE_i = \bar{E}_{2D}(i) + \bar{E}_{IE}(i) + E_{lig}(i) \tag{3.7}$$

and A^0_{SAS} is a parameter that accounts for the breathing of the RNA structure that may allow an apparently inaccessible nucleotide to become accessible to the SHAPE reagent. The model implies an effective ambient temperature when modeling SHAPE reactivity. Indeed, solution conditions including temperature can influence RNA conformational fluctuation and the reaction for SHAPE reagents (such as 1-methyl-7-nitroisatoic anhydride) to form $2'$-O-adducts with RNA nucleotides. Because SHAPE experimental data were collected under the folding conditions for the respective (folded) RNAs, the parameters in the model may be appropriate for the selection of folded RNA structures for the experimental conditions involved in the training data set. Considering that different SHAPE experimental data for different RNAs were often collected at different solution (such as temperature) conditions, the parameters in the model reflect an average effect of the different experimental conditions.

- *Calculating regular and noise-adjusted Pearson correlations.*

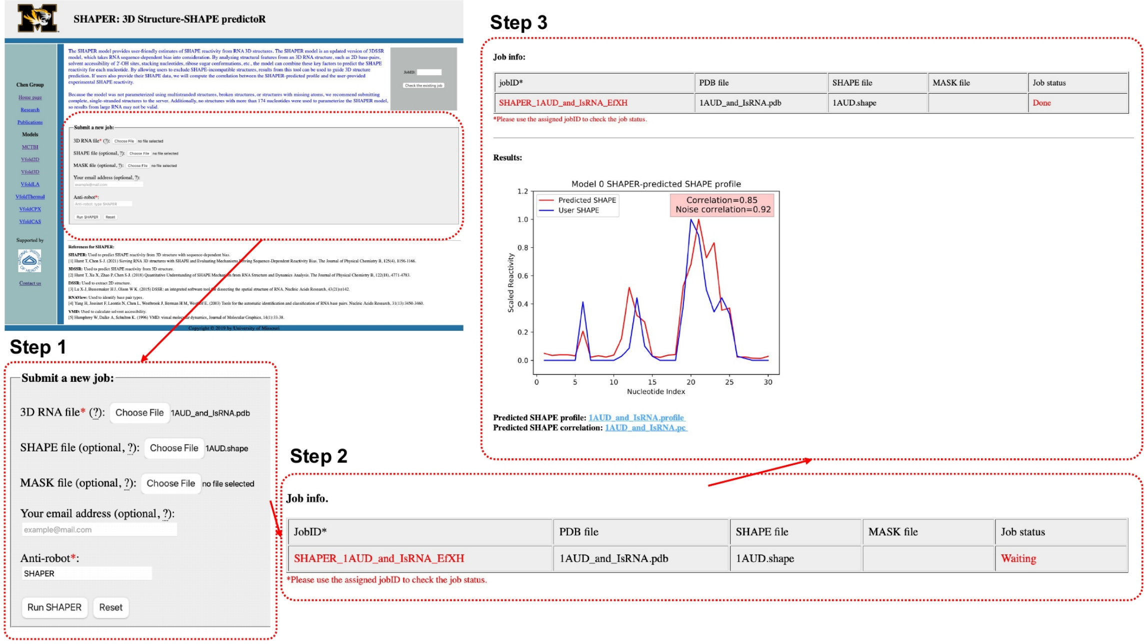
  In the original 3DSSR model, the relationship between the predicted SHAPE profile and experimental SHAPE data (if provided) was characterized by the Pearson correlation (PC). However, this regular PC does not account for the log-normality of SHAPE data and noise found by multiple previous studies [229, 230]. The newer SHAPER model uses a noise-adjusted normalization method to calculate the noise-adjusted PC between the predicted SHAPE profile and reweighted experimental SHAPE reactivities [172].

Step 3: Showing Output

After submitting a job, the user will be directed to a result page which shows the job status and information about the input files. This page will be refreshed every few seconds. Once the job is done, The result page will be updated and the Job status will change from "Waiting" to "Done". A plot of the predicted and user-provided (if any) SHAPE profile along with normal and noise-adjusted Pearson correlation coefficients will appear below the status table. Links to download corresponding SHAPE prediction and correlation files will appear at the bottom of the page (Fig. 3.2 Step 3). Existing results can be accessed by using the JobID, by bookmarking the address of the result page, or by checking email results (if provided).

3.2.2 Server Implementation

Several programming and scripting languages are used in the SHAPER server, including Bash, C++, Python, Perl, and Tcl. The SHAPE prediction module is implemented in C++ for performance. Third party software packages are used in other modules for preparing the necessary input files. Dissecting the Spatial Structure of RNA (DSSR) [226] is used to extract the 2D structure and torsional information of the ribose sugars from a 3D structure. RNAView [214] is used to identify base pair types shown in 3D structure, and the identification of stacking interactions is carried out by our in-house program written in Perl. The ligand accessible surface of the $2'$-hydroxyl for each nucleotide is calculated using Visual Molecular Dynamics (VMD) [211]. The above tools help automate the preparation process



SHAPER: 3D Structure-SHAPE predictoR

The SHAPER model provides user-friendly estimates of SHAPE reactivity from RNA 3D structures. The SHAPER model is an updated version of 3DSSR model, which takes RNA sequence-dependent bias into consideration. By analyzing structural features from an 3D RNA structure, such as 2D base-pairs, solvent accessibility of 2'-OH sites, stacking nucleotides, ribose sugar conformations, etc., the model can combine these key factors to predict the SHAPE reactivity for each nucleotide. By allowing users to exclude SHAPE-incompatible structures, results from this tool can be used to guide 3D structure prediction. If users also provide their SHAPE data, we will compute the correlation between the SHAPER-predicted profile and the user-provided experimental SHAPE reactivity.

Because the model was not parameterized using multistranded structures, broken structures, or structures with missing atoms, we recommend submitting complete, single-stranded structures to the server. Additionally, no structures with more than 174 nucleotides were used to parameterize the SHAPER model, so results from large RNA may not be valid.

JobID:

Check the existing job

Chen Group

Home page

Research

Publications

Models

MCTBI

Vfold2D

Vfold3D

VfoldLA

VfoldThermal

VfoldCPX

VfoldCAS

Supported by

Contact us

Submit a new job:

3D RNA file* (?): Choose File no file selected

SHAPE file (optional, ?): Choose File no file selected

MASK file (optional, ?): Choose File no file selected

Your email address (optional, ?): example@mail.com

Anti-robot*: Anti-robot: type SHAPER

Run SHAPER Reset

References for SHAPER:

SHAPER: Used to predict SHAPE reactivity from 3D structure with sequence-dependent bias.
[1] Hurst T, Chen S-J. (2021) Sieving RNA 3D structures with SHAPE and Evaluating Mechanisms Driving Sequence-Dependent Reactivity Bias. The Journal of Physical Chemistry B, 125(4), 1156-1166.

3DSSR: Used to predict SHAPE reactivity from 3D structure.
[2] Hurst T, Xu X, Zhao P, Chen S-J. (2018) Quantitative Understanding of SHAPE Mechanism from RNA Structure and Dynamics Analysis. The Journal of Physical Chemistry B, 122(18), 4771-4783.

DSSR: Used to extract 2D structure.
[3] Lu X-J, Bussemaker H J, Olson W K. (2015) DSSR: an integrated software tool for dissecting the spatial structure of RNA. Nucleic Acids Research, 43(21):e142.

RNAView: Used to identify base pair types.
[4] Yang H, Jossinet F, Leontis N, Chen L, Westbrook J, Berman H M, Westhof E. (2003) Tools for the automatic identification and classification of RNA base pairs. Nucleic Acids Research, 31(13):3450-3460.

VMD: Used to calculate solvent accessibility.
[5] Humphrey W, Dalke A, Schulten K. (1996) VMD: visual molecular dynamics, Journal of Molecular Graphics, 14(1):33-38.

Copyright © 2019 by University of Missouri

Step 1

Submit a new job:

3D RNA file* (?): Choose File 1AUD_and_IsRNA.pdb

SHAPE file (optional, ?): Choose File 1AUD.shape

MASK file (optional, ?): Choose File no file selected

Your email address (optional, ?): example@mail.com

Anti-robot*: SHAPER

Run SHAPER Reset

Step 2

Job info.

JobID*	PDB file	SHAPE file	MASK file	Job status
SHAPER_1AUD_and_IsRNA_EfXH	1AUD_and_IsRNA.pdb	1AUD.shape		Waiting

*Please use the assigned jobID to check the job status.

Step 3

Job info:

jobID*	PDB file	SHAPE file	MASK file	Job status
SHAPER_1AUD_and_IsRNA_EfXH	1AUD_and_IsRNA.pdb	1AUD.shape		Done

*Please use the assigned jobID to check the job status.

Results:

Predicted SHAPE profile: 1AUD_and_IsRNA.profile
Predicted SHAPE correlation: 1AUD_and_IsRNA.pc

Figure 3.2: Interface of the SHAPER web server and the steps involved in submitting a job. The overview of the interface of the starting page is shown in the top left, and the area within the dashed red box is updated in each step. There are three steps: uploading and choosing parameters (Step 1), waiting the job (Step 2), and checking results of the job (Step 3).

and greatly reduce the potential for human error. All modules were combined by Python and the web server is based on Apache 2.2.15.

3.3 Case Study

3.3.1 Sieving RNA 3D structures generated by 3D structure prediction software

To better illustrate the function of the SHAPER web server, we ran an example case with known experimental SHAPE data to show the ability of SHAPER to distinguish near-native conformations from a pool of decoys. The test RNA structure (PDB code: 2L8H) contains 29 nucleotides. We used our coarse grained (CG) simulation software (IsRNA) [212, 231] and an all-atom molecular dynamics (MD) simulation to generate 59 decoy conformations for the target RNA. We selected 20 near-native conformations and 39 non-native conformations generated with native and non-native 2D structures [172]. These decoys along with the native structure allow us to show the ability of the SHAPER server to distinguish native conformation from conformational pools. Then we put these 60 structures into the SHAPER web server, and the correlation coefficients (PC and noise-adjusted PC) between predicted SHAPE profiles and experimental SHAPE data were calculated. The root mean square deviations (RMSDs) between the native and decoy conformations were calculated for heavy atoms. As shown in Fig. 3.3C for the relationship between RMSD and SHAPE correlation coefficients, the native structure shows the highest correlation, and the near-native conformations around 2 Å of RMSD also have high correlations. However, similar correlations were also found for non-native conformations around 4 Å to 6 Å. This is because the 2D structural constraints (see, Fig. 3.3B) used to generate these decoys are similar to the native 2D structural constraints (see, Fig. 3.3A). As for the non-native conformations generated by using different 2D structural constraints (see, Fig. 3.3B), both correlation coefficients (PC and noise-adjusted PC) drop significantly relative to the values of the native conformation. The above results suggest that SHAPE correlation may serve as a useful measure to sieve structures and find the native and near-native 2D and 3D structures.

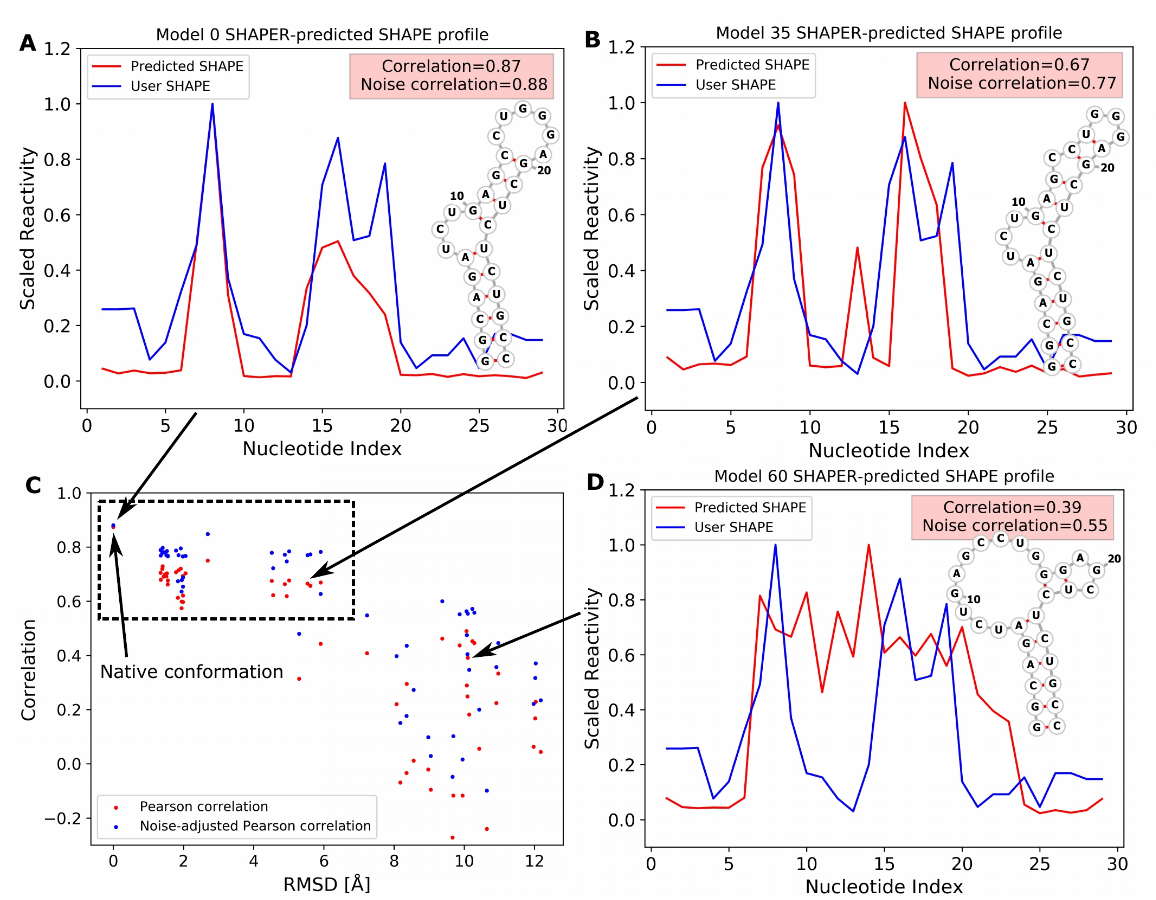


Figure 3.3: SHAPE profiles for native (A) and selected decoy (B,D) conformations at different RMSDs. Predicted and experimental (i.e., User SHAPE) profiles are shown in red and blue curves, respectively. 2D structure in (A) corresponds to the native structure, and 2D structures shown in (B,D) were used as constraints to run the simulations. (C) The relation between PC/noise-adjusted PC (red/blue) and RMSDs relative to the native structure for sixty tested conformations (include the native one, PDB code: 2L8H). The data point of native conformation is shown on the top left of (C) and pointed out by an arrow.

3.4 Conclusion

SHAPER is a fast and accurate web server to predict SHAPE profile for any given RNA structure. Compared to the original 3DSSR model, SHAPER greatly improves performance [171] by accounting for sequence-dependent bias, tail effects, and ligand binding. In addition, the SHAPER model better reflects that SHAPE reactivities are a direct reflection of the underlying system energetics and incorporates effects related to the log-normality of SHAPE data and noise. The server provides functionalities for predicting SHAPE profiles for RNA with either a single structure or a structural ensemble. Combined with the available experimental SHAPE data, SHAPER can provide a reliable measure of the nativeness of the target conformation and serves as a convenient tool to help researchers select the most probable RNA 3D structures from a pool of decoys.

Chapter 4

A method for decoding nucleic acid-magnesium ion interactions using a deep learning convolutional neural network

This chapter has been submitted[1].

Magnesium ions (Mg^{2+}) play a vital biological role as cofactors, interacting with RNA molecules to facilitate RNA folding and function. Previous attempts to accurately model RNA-Mg^{2+} binding have been plagued by difficulties arising from the challenge of accurately locating Mg^{2+} binding sites. Using experimental RNA structural data, we developed and applied MgNet, a machine-learning model, to predict Mg^{2+} binding sites in RNA molecules. This approach exploits local binding information associated with each nucleotide. In particular, electrostatic and 3D-shape (RNA volume) features are used to capture the key interaction patterns for the network to predict the density distribution of

[1] Yuanzhe Zhou and Shi-Jie Chen. "A method for decoding nucleic acid-magnesium ion interactions using a deep learning convolutional neural network". In: *submitted* (2021).

Mg^{2+} around the RNA molecules. Five-fold cross-validation on a dataset of 177 selected Mg^{2+}-containing structures and comparisons with three different types of methods validate the approach. Results show that this new approach predicts Mg^{2+} binding sites with higher accuracy and efficiency. We use saliency analysis for eight different Mg^{2+} binding motifs to reveal the coordinating atoms of Mg^{2+} ions. Furthermore, learning the relevant physical mechanism through in-depth training on the known ion-RNA complexes, MgNet also uncovers new Mg^{2+} binding motifs.

4.1 Introduction

The phosphodiester backbone of RNA carries an electronic charge per nucleotide, thus, metal ions, through binding to RNA, play a critical role in stabilizing an RNA structure. In particular, magnesium ions (Mg^{2+}) are essential for RNA tertiary structure folding, stability [22–25, 27–29], and function in biological processes [30–37]. However, experimental studies of RNA-Mg^{2+} interactions are challenging. The flexible nature of RNA can lead to an ensemble of low-energy conformations, and Mg^{2+} binding preferences may change in different conformations. Furthermore, using electron density maps to distinguish Mg^{2+} from water (H_2O) and sodium ion (Na^+) is challenging because they all have 10 electrons and can be distinguished only in high-resolution structures, so Mg^{2+} can be easily mistaken for H_2O or Na^+ [41, 62]. Alternatively, Mg^{2+} may be simply missing from crystal structures [41]. A significant number of misidentified Mg^{2+} binding sites can impose a strong and incorrect bias on Mg^{2+} binding analysis and prediction.

In addition to the obstacles created by RNA conformational multiplicity and misidentification of Mg^{2+} binding sites, a relative dearth of high-resolution data also imposes a barrier to understanding the relevant biological processes that depend on RNA-Mg^{2+} binding. As of April 16, 2021, 1558 structures that have RNA-Mg^{2+} interactions are available in the Nucleic Acid Database [225, 233]. Of these, 1555 are X-ray structures, and only 942 have high resolution (< 3.0 Å). Many of these structures come from the same molecule and organism with similar Mg^{2+} binding sites, making them redundant. High-resolution experimental studies are time-consuming and inefficient, which makes computational prediction of Mg^{2+} binding a highly desirable supplement to experimental benchmarks. The growing number of experimentally solved RNA structures motivates us to take advantage of the increasing amount of experimental information by developing a knowledge/data-based method to model the interactions between RNA and Mg^{2+}.

During the last few years, researchers have developed a number of novel approaches to predict RNA-metal ion binding sites. We can categorize these modeling efforts into physics-based approaches and knowledge-based approaches. Physics-based methods explicitly consider physical interactions. These models provide detailed information about the physical energetics and dynamics for RNA-ion binding. However, given their relatively complex functional forms, the physical approaches are often computationally intensive. All-atom MD [29, 234–236], Brownian dynamics [59, 237], Poisson-Boltzmann (PB)/generalized Born (GB) models [238, 239], and statistical mechanical models [240, 241] are all physics-based models, with varying levels of success. Knowledge-based methods rely on information extracted from experimentally determined structures. Such methods are usually much less computationally demanding than physics-based approaches, but the inability of these methods to take conformational dynamics into consideration also makes them ill-suited for ion binding prediction that involves conformational changes. FEATURE [242] and MetalionRNA [243] represent two important knowledge-based methods.

FEATURE [242] is a knowledge-based predictor that can predict the most typical metal ion-binding sites in RNA structures. By collecting a set of metal ion binding sites as “sites” and a set of control non-binding sites as “non-sites”, FEATURE [242] transforms the environments around these “sites” and “non-sites” into a set of feature vectors, where each feature vector describes a unique environment encoding either a binding site or a non-binding site. Collections of these feature vectors are called microenvironments. When given a query region in a new structure, the Wilcoxon rank-sum test is used to compare the feature vector of the query region with the collection of “sites” and “non-sites” microenvironments. This nonparametric test determines the group with which the query region is most similar and scores the likelihood that the query region is a binding site.

MetalionRNA [243] uses a representative set of 113 crystallographically determined structures to derive statistical potentials for Na^+, K^+, and Mg^{2+} ions. The model evaluates the three-body anisotropic contact frequencies between metal ions and a set of predefined covalently bonded RNA atom pairs that are known to make the strongest contributions to metal ion binding. The model then transforms the contact frequencies into statistical potentials through the inverse Boltzmann law. Given a new structure, MetalionRNA scores every grid point in the space according to statistical potentials derived from the observed contact frequencies in the training set. These scores are used to predict the final binding sites.

Here, we propose a convolutional neural network (CNN) model, MgNet, which uses experimental structural data to predict metal ion binding sites. CNNs have found success in various fields, especially in computer visual recognition. CNN models excel at pattern recognition by using convolutional operations to combine correlated data and identify underlying trends. Our CNN has a layered structure as shown in Fig. 4.1c. The more layers the CNN contains, the deeper the network is. The convolutional filters connect the different layers of the CNN. As the core components of the network, these convolutional filters contain trainable parameters. They perform convolutional operations on the output of each preceding layer. By extracting and processing the features from preceding layers, filters serve as feature extractors. Features extracted from an image can be combined, processed, and propagated through multiple layers, resulting in a high-level abstraction of the important features in the original image. This high-level abstraction of features allows the model to make knowledge-based predictions. Our approach transforms each RNA structure with Mg^{2+} into a collection of images for MgNet training, validation, or testing.

In this study, we apply a regression CNN with residual shortcuts similar to the ResNet [216] model. While normal CNNs read 2D images as input, our MgNet reads "3D images" that contain the local environment of the binding and non-binding sites as input. Just as each 2D image has three color channels that provide different information for a 2D CNN, our 3D images have multiple "color" channels, which contain partial charge information, volume occupancy information, or other chemical properties that contribute to the interactions between RNA and metal ions.

4.2 Results

4.2.1 Outline of the method

Our machine-learning model is a variant of the vanilla convolutional neural network. Compared to the traditional knowledge-based methods [242, 243] used for predicting RNA-

Mg^{2+} binding sites, our MgNet model has two distinct advantages: it does not assume any functional form of the interaction energy prior to training, and with the progression of the convolution through a layer-wise organization, it captures the long-range correlations between RNA atoms and Mg^{2+} ions that are missing from traditional knowledge-based approaches [242, 243].

In our method, electrostatic and 3D-shape (RNA volume) information taken from the Mg^{2+} binding environment were used as the input features. This information can usually be obtained rapidly, automatically, and reliably from macromolecular modeling software, such as UCSF Chimera [244]. Ion distributions around the RNA can then be identified by MgNet through 3D image analysis of the target RNA with computed electrostatic and 3D-shape (RNA volume) information.

We compare specific Mg^{2+} binding sites predicted by MgNet to experimental results and to those predicted from other methods: a knowledge-based approach [243], a molecular dynamics simulation-based approach [29] and a Brownian dynamics simulation-based approach [59]. In order to identify Mg^{2+} binding sites from the predicted ion probability distribution, we use the DBSCAN [245] method to cluster the ion binding sites of probability maxima. Within each high-probability region, k-means clustering was used to find the representative points of the region. These representative points were chosen as the predicted ion sites and were ranked based on the sum of the probabilities of the points within the corresponding cluster.

We use two criteria, true positive rate (TPR) and positive predictive value (PPV), to measure the predictive power of the model. Physically, TPR (PPV) is the ratio between the number of the correctly predicted ion binding sites and that of experimentally observed (theoretically predicted) bound ions. Generally speaking, although one may alter TPR and

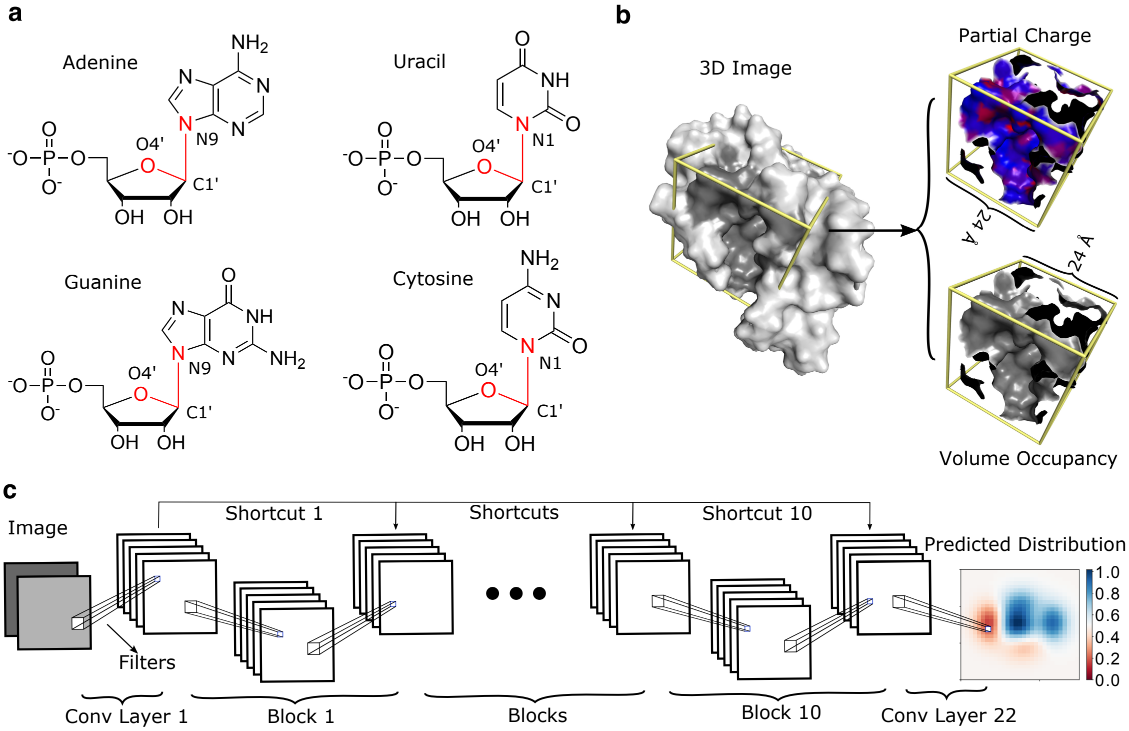


Figure 4.1: Overview of the MgNet. (a) Three key atoms (shown in red) in the sugar ring and bases are used to determine the local Cartesian coordinate system. The origin of the local coordinate system is set to the midpoint between the carbon atom (C1′) and nitrogen atom (N1 for pyrimidine or N9 for purine), where the vector formed by C1′ and oxygen atom (O4′), and the vector formed by C1′ and nitrogen atom (N1 or N9) are used to determine the x-y plane of the system. (b) Each 3D image is taken from a 24 Å x 24 Å x 24 Å cubic box centered at a given nucleotide, and is used to capture the information for binding and non-binding sites. The cubic box is shown with yellow frames. The local Cartesian coordinate system (i.e., orientation of the cubic box) is determined by the key atoms in the associated nucleotide. Two feature channels (partial charge and volume occupancy) are used to extract the relevant information from the image. (c) The MgNet is drawn in a 2D diagram for a better illustration, where all 3D images (3D cubic grids) are shown as 2D squares. From left to right, an image with two feature channels is fed into the MgNet, and information is then processed by different layers of filters and connected through various shortcuts. The final prediction is an ion density distribution map.

PPV by adjusting the definition of the "correctly" predicted sites, these two metrics are often antagonistic to each other except for a perfect model. In practice, increasing of the number of the predicted sites usually improves the TPR but in the meantime, causes the degradation of the PPV, and vice versa. Thus TPR and PPV together can provide an overall measure for the performance of the model.

To extract physical insights from the neural network, we performed saliency calculation and coordination classification. From the gradients of the predicted scores with respect to the input image pixels (saliency values), the saliency analysis identifies [246] the most sensitive pixels in the input image whose small variations cause substantial changes in the output result. The saliency technique allows us to uncover the RNA atoms that most sensitively determine Mg^{2+} binding. Furthermore, a thorough investigation on the configurations of RNA atoms around a bound Mg^{2+} ion reveals Mg^{2+} binding motifs.

4.2.2 Evaluating MgNet performance through cross-validation

We carried out a five-fold cross-validation on the selected dataset with 177 RNA-Mg^{2+} complex structures. Details of this dataset can be found in section "Methods" and Table 4.6-4.10. With two sets of 36 RNA structures and three sets of 35 RNA structures, we randomly split the 177 Mg^{2+}-containing structures into 5 subsets. For each cycle, we use one of the subsets for testing and the other four for training the MgNet model. The cross-validation approach ensures the complete sampling of the whole data sets while keeping test and training sets not overlapping in the same cycle. For each test set, we measure the accuracy using the root-mean-square-deviation (RMSD) between the predicted and the experimentally determined coordinate of the bound Mg^{2+} ion. The five test sets contain 234, 361, 283, 265 and 264 experimentally determined Mg^{2+} ions. As shown in Table 4.1, the MgNet model predicts 371, 490, 309, 346 and 347 Mg^{2+} binding sites for the 1st, 2nd, 3rd, 4th, and 5th test sets, respectively, and the predicted Mg^{2+} ion coordinate is within 3 Å from the experimentally determined position for 118, 179, 123, 133 and 108 predicted binding sites, respectively. In summary, for the 177 RNA-Mg^{2+} complex structures, there are 1407 experimentally determined Mg^{2+} binding sites, MgNet predicts 1863 Mg^{2+} binding sites, among which 661 Mg^{2+} binding sites (coordinates) were predicted within 3 Å from the experimentally results. Statistically speaking, the test result implies that the MgNet model is able to identify nearly half of true Mg^{2+} binding sites with high accuracy.

Table 4.1: TPR and PPV of the five-fold cross-validation test. Table shows values of TPR and PPV of MgNet model on the five-fold cross-validation test. The number after cv indicates the index of the validation. The column under the total is the averaged results of the five-fold cross-validation.

	cv1	cv2	cv3	cv4	cv5	total
TPR	50.43%	49.58%	43.46%	50.19%	40.91%	46.91%
PPV	31.81%	36.53%	39.81%	38.44%	31.12%	35.54%

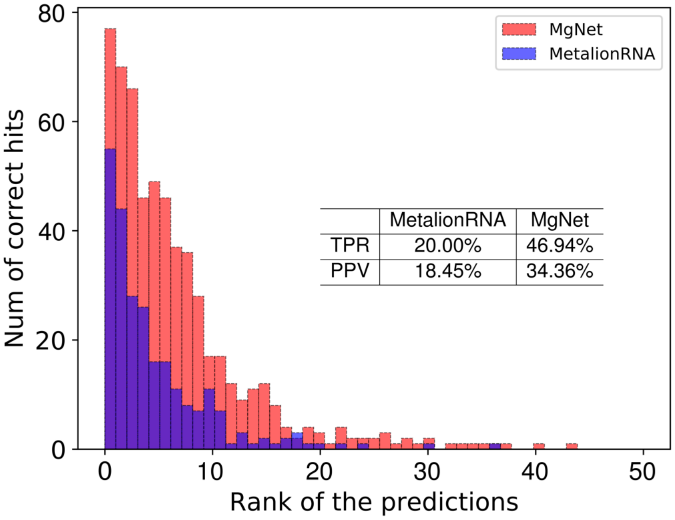
4.2.3 Comparing the performance between MgNet and MetalionRNA

By comparing MgNet to a knowledge-based method, MetalionRNA [243] model, we assess the performance of the CNN approach. Following the previous studies [242, 243], we first compare the performance of MgNet and MetalionRNA on a 58 nt fragment of *Escherichia coli* 23S rRNA which contains seven Mg^{2+} ions in the crystal structure (PDB code 1HC8). As shown in Table 4.2, MgNet and MetalionRNA can identify all the seven Mg^{2+} ions within the top-9 and top-29 ranked predictions with an accuracy of 0.5-3.6 Å and 0.6-3.8 Å, respectively. Lower ranked sites correspond to predicted sites with lower confidence. Among the top-9 predictions from MgNet, six out of the seven Mg^{2+} ions are predicted with an accuracy of 0.5-2.8 Å. The remaining experimental ion is found in between two experimentally determined ions (residue number 1161 and 1160) with a distance of 2.8 Å and 3.6 Å to the ions of residue number 1161 and 1160, respectively. The result suggests that these two Mg^{2+} sites could share a mutual binding area, see Fig. 4.3a. As shown in the *MgNet column of Table 4.2, the top-12 predicted Mg^{2+} ion coordinates give all the seven experimentally determined ions with accuracy of 0.5-2.3 Å. However, while the adjusted cluster setting increases the overall TPR values, it also reduces the PPV values of MgNet predictions on the five-fold cross-validation set. Thus, combining TPR and PPV, MgNet with the default clustering settings is used for the comparisons with other methods. With the default cluster setting, for the top-9 predicted sites, MetalionRNA and MgNet predict 4 and 6 correct Mg^{2+} binding sites with an accuracy of 0.6-1.9 Å and 0.5-2.8 Å, respectively.

Table 4.2: Comparison between the performance of MetalionRNA and of MgNet for the 58 nt fragment of 23S rRNA structure (PDB code: 1HC8). RMSD values and ranks of the predictions of MetalionRNA [243] and MgNet for Mg^{2+} ions in the 58 nt fragment of 23S rRNA structure (PDB code: 1HC8). The leftmost column lists the Mg^{2+} identifiers (residue number) as labeled in the PDB file. From the second column to rightmost column, we summarizes predictions made by MetalionRNA, MgNet with default cluster settings, and MgNet with an adjusted cluster settings, respectively. The 7 experimentally determined ion sites are successfully predicted by MetalionRNA, MgNet, and *MgNet within the top-29, 9 and 12 ranked hits, respectively. For each entry, the number in a parenthesis indicates the rank of the corresponding prediction. A dash line means there is no predicted ion for the corresponding experimental binding site.

Mg^{2+}	MetalionRNA(29)	MgNet(9)	*MgNet(12)
(res no.)	Å(rank)	Å(rank)	Å(rank)
1159	0.8 (1)	0.8 (8)	0.8 (7)
1160	1.9 (6)	-	0.6 (12)
1161	2.9 (29)	2.8 (7)	2.3 (8)
1163	0.6 (3)	1.8 (5)	1.3 (4)
1164	1.4 (2)	2.3 (3)	1.4 (5)
1167	3.8 (10)	0.5 (1)	0.5 (3)
1172	3.2 (13)	1.6 (9)	1.8 (9)

For a more comprehensive comparison, we use TPR and PPV to evaluate the performance of MetalionRNA web-server [243] on our cross-validation dataset. Table inside the Fig. 4.2 summaries the results. Fig. 4.2 also shows the distributions of the number of the correctly predicted sites from the MgNet model and the MetalionRNA web-server on the 176 RNA-containing structures, it can be easily seen that MgNet has a much better rate in giving the experimental ion binding sites for most of the top ranked predictions.



	MetalionRNA	MgNet
TPR	20.00%	46.94%
PPV	18.45%	34.36%

Figure 4.2: Comparison between MetalionRNA server [243] and MgNet on cross-validation set. Histogram shows the distribution of the number of correct hits over the top-prediction ranks. The horizontal axis represents the rank of the predictions, where n on the axis means the nth-ranked prediction for a given RNA, and the vertical axis represents the number of the experimentally determined ions in the cross-validation set that are correctly identified by the nth-ranked prediction of each RNA. The table in the figure shows values of TPR and PPV of the MgNet model and the MetalionRNA model, respectively, on the cross-validation set. The results shown here exclude the structure of PDB ID 3T1Y due to the failed retrieval of the prediction data from the MetalionRNA web-server. The cutoff RMSD for correct hits is 3 Å. MetalionRNA server results were obtained from the MetalionRNA server with default settings, and MgNet results were collected from our five-fold cross-validation models with default clustering settings. The details of the default clustering settings can be found in the **Methods** section.

4.2.4 Comparing performance between MgNet and a molecular dynamics (MD) simulation-based method

Although several physics-based methods have been developed to investigate the metal ion-RNA interactions, most of the methods focus on the dynamics or statistical properties instead of the ion binding sites. As suggested by Fischer et al. [29], a molecular dynamics (MD) method with explicit water can be applied to characterize Mg^{2+} distributions around folded RNA structures as well as predict Mg^{2+} positions. In the study [29], seven RNA structures consist of Mg^{2+} ions were selected as target system in MD simulation. Two of them fold into helical structure while the rest of the structures fold into more complex forms.

For Mg^{2+}-involved simulations, Fischer et al. [29] created two different systems for each RNA structure, with Mg^{2+} as the counterion (CI) (Mg^{2+}_{CI}) only and at the physiological salt (PS) concentration (Mg^{2+}_{PS}) with Mg^{2+} counterions and 0.15 M/l NaCl [29]. In order to investigate whether MD simulation can recover the experimental binding sites, ions were initially randomly placed in the simulation box. The predicted ion positions are determined by the occupancy of Mg^{2+} during the simulation using the software MobyWat [247, 248].

In order to compare the MgNet predictions with the MD simulations results for the seven RNA structures, we use a five-fold cross-validation procedure. First, we use the same five subsets of RNA structures generated from five-fold cross-validation. Second, for each subset, we remove any duplicate RNA structures that are the same as or similar to any of the seven test structures. The second step results in the removal of RNA structures with PDB codes 1D4R, 1Y95 and 4FRG, leaving 174 remaining RNA structures. We then performed the five-fold cross-validation for the five (modified) subsets. Finally, we used each trained model to predict the Mg^{2+} binding sites for the seven test RNA structures. Our results are shown as the MgNet column in Table 4.3.

Table 4.3: Comparison between the molecular dynamics (MD) simulation-based method and MgNet on seven test structures. RMSD values and the standard deviations between the predicted ion sites and the corresponding experimental ion sites, measured in angstrom. The PDB code and the corresponding experimental Mg^{2+} ions are listed in the first two columns. Column Mg^{2+}_{CI} and Mg^{2+}_{PS} show the average RMSD values and the standard deviations of MD simulation-based method. The top-50 predicted sites from the MD simulation-based method were used. Column MgNet shows the averaged RMSD values over the predictions of the five trained MgNet models. We note that MD simulation-based method did not provide the rank order for the predicted ions, thus we only listed average RMSD and the standard deviation for each Mg^{2+}. Ranks of the predictions of MgNet model can be found in Table 4.12.

PDB	Ion	Mg^{2+}_{CI}	Mg^{2+}_{PS}	MgNet
1D4R	MG-90	1.0 ± 0.5	1.1 ± 0.5	2.2 ± 0.6
	MG-91	5.0 ± 0.7	4.4 ± 0.7	3.7 ± 0.8
2MTK	MG-48	7.4 ± 3.2	5.8 ± 1.9	4.8 ± 0.5
	MG-49	3.9 ± 0.9	2.9 ± 1.6	3.6 ± 0.5
	MG-50	6.7 ± 3.0	5.7 ± 2.3	1.6 ± 0.2
	MG-51	3.2 ± 0.8	3.5 ± 0.4	7.1 ± 8.1
	MG-52	3.6 ± 0.5	3.8 ± 2.4	7.7 ± 0.4
	MG-53	2.1 ± 0.4	2.3 ± 1.1	2.1 ± 0.7
2QEK	MG-49	2.5 ± 0.2	2.5 ± 1.3	6.3 ± 4.4
4FRG	MG-179	2.4 ± 0.7	4.4 ± 0.8	1.2 ± 0.6
	MG-180	2.4 ± 0.8	5.3 ± 0.4	1.5 ± 0.4
	MG-181	2.8 ± 0.5	1.4 ± 0.5	18.8 ± 0.2
	MG-182	7.6 ± 0.5	7.0 ± 0.6	2.0 ± 0.1
	MG-183	3.7 ± 1.5	4.7 ± 2.9	1.6 ± 0.8
	MG-184	1.1 ± 0.3	2.0 ± 1.5	2.1 ± 0.9
	MG-185	3.7 ± 1.3	5.9 ± 1.4	4.8 ± 5.5
4JF2	MG-94	2.2 ± 1.1	2.7 ± 1.0	0.7 ± 0.2
	MG-95	3.2 ± 0.6	4.6 ± 0.7	4.0 ± 6.7
	MG-96	2.5 ± 0.8	2.9 ± 0.9	0.5 ± 0.1
	MG-97	18.3 ± 2.7	20.6 ± 0.8	0.6 ± 0.2
4KQY	MG-121	1.8 ± 0.4	3.4 ± 0.9	1.3 ± 0.4
	MG-122	1.8 ± 0.5	4.2 ± 2.0	6.4 ± 1.4
4P5J	MG-85	1.1 ± 0.7	1.8 ± 0.2	1.5 ± 0.4
	MG-86	2.5 ± 0.5	3.5 ± 2.1	1.3 ± 0.1

The MgNet model gives overall better predictions than the MD simulations for the locations of the bound ions, The different results between MgNet and MD simulations are due to several reasons. First, the RNA structures used in MgNet training are mainly crystal structures, thus the interaction patterns learned by MgNet may not be ideal for NMR solution structures, which causes slightly worse results for 2MTK (PDB ID), an NMR solution structure. Second, MD simulations for ions directly bound to RNA may suffer from the incomplete sampling problem due to the high barrier for Mg^{2+} dehydration.

4.2.5 Comparing the performance between MgNet and a Brownian dynamics (BD) simulation-based method

In Brownian dynamics (BD) simulations [59], diffuse cations move under the influence of random Brownian motion in the electrostatic field and the metal ion binding sites are identified by analyzing the trajectories of positively charged test spheres. Previous BD simulations were able to identify Mg^{2+} binding sites in the crystal structures of loop E of bacterial 5S rRNA (PDB code: 354D), $tRNA^{Phe}$ (PDB code: 4TRA) and $tRNA^{Asp}$ (PDB code: 3TRA). To compare MgNet with the BD simulations, we used the aforementioned five-fold cross-validation procedure with the test RNA structures removed from the training set. The resultant dataset contains 175 RNA structures. Table 4.4 shows the comparison between the BD simulations and our MgNet models.

Overall speaking, both BD simulations and MgNet show good performance for the tested RNA structures. However, there exist two notable differences between the predictions from the two approaches. Several trained models of MgNet failed to predict the binding sites within 10 Å from the experimental site for Mg^{2+} ion A-76 (354D) and ion

Table 4.4: Comparison between the Brownian dynamics (BD) simulation-based method and MgNet on three test structures. RMSD values between the predicted ion sites and the experimental ion sites, measured in angstrom. The PDB codes and the corresponding experimental Mg^{2+} ions are listed in the first two columns. By simulating many positively charged spheres under the influence of both random Brownian motion and the electrostatic field of the RNA, the predicted binding sites of BD simulations were identified as the regions where a significant number of the test charges are finally trapped. The results of the trained MgNet models are listed from columns cv1 to cv5 with the ranks shown in parentheses. Experimentally determined ion sites that theoretical models failed to predict within 10 Å are labeled with a dash.

PDB	Ion	BD	cv1	cv2	cv3	cv4	cv5
354D	A-203	1.8	0.9 (2)	1.1 (2)	1.6 (6)	1.2 (7)	1.0 (3)
	B-200	0.7	0.4 (1)	0.5 (3)	0.5 (2)	0.9 (1)	1.0 (2)
	B-201	1.3	1.2 (4)	1.6 (5)	1.3 (4)	1.0 (5)	0.9 (1)
	B-202	1.4	1.1 (5)	1.1 (1)	1.7 (3)	0.8 (4)	1.2 (6)
	B-204	2.7	1.4 (7)	1.6 (6)	5.8 (1)	1.6 (6)	6.4 (4)
3TRA	A-76	$\sim$5.0	-	-	-	4.3 (6)	-
4TRA	A-77	2.6	2.0 (4)	2.5 (6)	2.1 (7)	2.4 (4)	2.2 (4)
	A-78	2.1	1.0 (5)	0.7 (2)	0.9 (4)	0.5 (5)	1.4 (1)
	A-79	2.2	2.1 (7)	1.2 (3)	1.0 (5)	1.8 (6)	1.9 (5)
	A-80	0.3	6.2 (6)	-	-	6.1 (7)	6.0 (6)

A-80 (4TRA). One predicted site within 10 Å was captured for ion A-76, and the RMSDs of the MgNet-predicted ion A-80 are larger than that of BD simulation. For ion A-76 of 3TRA, the crystal structure of $tRNA^{Asp}$ contains a single Mg^{2+} located in the anticodon loop at the $C_{31} \cdot G_{39}$ base pair [59]. Both BD simulations and MgNet-predicted ion sites were found within a distance of $\sim$5 Å from the site in the crystal structure, and both were shifted upward in the anticodon stem towards the $G_{30} \cdot U_{40}$ wobble pair (Fig. 4.3b). This shifted ion binding pattern shares similarities with the experimentally found metal ion binding site at G·U pairs in the crystal structure of P4-P6 of group I intron [59]. The result might suggest a delocalized binding of metal ions in the anticodon loop of $tRNA^{Asp}$. As for ion A-80 of 4TRA, the predicted site deviates from the experimental site possibly because this particular ion is in close contact with a non-standard residues Wybutosine (yw). We note that Mg^{2+} binding to one or more non-standard residues is not common in our training set, the predictions of MgNet on such cases are less reliable.

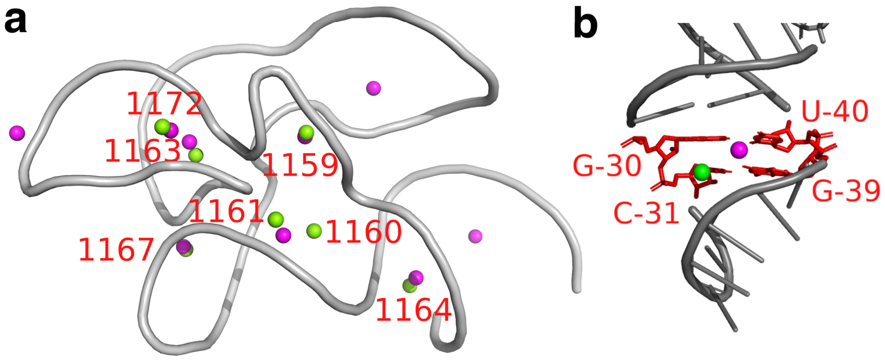


Figure 4.3: MgNet-predicted (magenta spheres) vs. experimentally determined (green spheres, labeled with residue identifiers) Mg^{2+} ion sites in (a) 58 nt fragment of *Escherichia coli* 23S rRNA (PDB ID: 1HC8) and (b) the anticodon loop in $tRNA^{Asp}$. The predicted site in (b) is shifted upward towards the $G_{30} \cdot U_{40}$ wobble pair. Four residues shown in red are labeled with the residue names and the residue sequence numbers.

4.2.6 MgNet-saliency analysis for metal ion binding sites

In machine-learning, a large saliency value means that a slight change in the corresponding input feature causes a markedly change in the output prediction score. Therefore, saliency analysis can identify the key physical features that most sensitively determine the predicted ion binding. In MgNet, from each input 3D image, the convolutional network predicts a 3D matrix where a matrix element p(i, j, k) is the probability of finding a bound ion at the grid site (i, j, k). From the gradients of the predicted ion distribution with respect to the images of the target binding site, the saliency analysis identifies the RNA atoms and the physical attributes that most critically determine the ion distribution; see the "Methods" section for a detailed description of the saliency calculation.

A previous survey on the Mg^{2+} sites in RNAs deposited in the Protein Data Bank [208] led to 13 Mg^{2+} binding motifs [41]. By examining the predictions for RNAs in cross-validation set and for several ribosomal RNAs which are not included in cross-validation set, we select eight example binding-sites with different motifs. These motifs differ by the type of the coordination (i.e., inner-sphere or outer-sphere coordination), the number and type of the coordinating atoms, and the geometry of the coordination. Six cases (Fig. 4.4a-f) have inner-sphere interactions with RNA atoms, while the rest (Fig. 4.4g-h) interact only with RNA atoms through outer-sphere hydrogen bonds (mediated by water molecules). Several motifs (Figs. 4.4a and 4.4b, 4.4c and 4.4d) share geometrical similarities. The "Magnesium clamp" [45, 47] and "Y-clamp" [41] use the bridging capability of phosphates to stabilize these close interactions through juxtaposition of two different strands or two distant segments of the same strand, very much similar to the disulfide bonds in proteins. The "U-phosphate" [41] and "G-phosphate" [42] both require the coordination of a phosphate oxygen and a nucleobase oxygen. The more complicated motifs, "Purine N7-seat" [41] , "G-G metal binding site" [46] , and "Triple G motif" [44] , contain complex water mediated coordination.

Saliency analysis for the above examples reveals important atoms that are critical for the stabilization of magnesium ions at the binding site. As shown in Fig. 4.4, atom saliency values for the two input channels (volume occupancy and partial charge) indicate specific coordinating atoms as the important factors in determining Mg^{2+} binding sites. Note that in Fig. 4.4a, two of the important phosphate oxygen atoms (OP1 of A34 and OP2 of G46) in the opposite direction have darker color, this is because there exists another Mg^{2+} that binds in the nearby location (shown as cyan sphere), causing these oxygen atoms to have a large saliency value. Note that saliency value of a particular atom reflects the sensitivity of the predicted ion density with respect to this particular atom, namely, small change in the pixel values (physical attributes) of the blue atoms shown in the figure would markedly alter the predicted ion (probability) density.

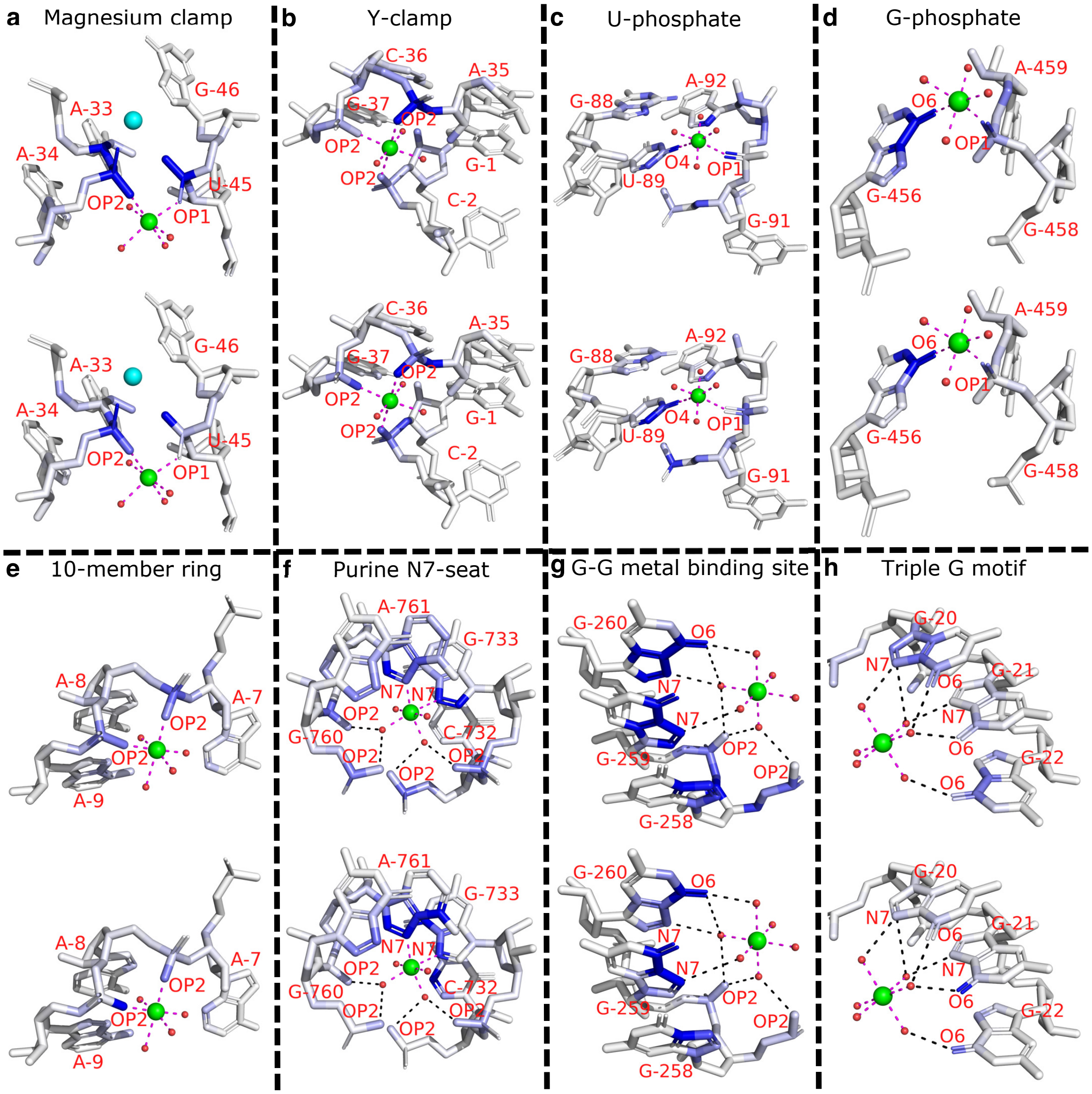


Figure 4.4: Example of saliency calculation for eight binding motifs. Saliency values were calculated for eight binding sites: (a) 3Q3Z-V85; (b) 2Z75-B301; (c) 2YIE-Z1116; (d) 1VQ8-08004; (e) 3DD2-B1000; (f) 2QBA-B3321; (g) 4TP8-A1601; (h) 3HAX-E200. And two input channels: volume occupancy (top) and partial charge (bottom). Experimentally determined positions of Mg^{2+} cation are indicated by green spheres, oxygen atoms in water molecules are shown in small red spheres. Direct coordinations (inner-sphere coordination) are shown as magenta dashes, and indirect coordinations (outer-sphere coordination, i.e., mediated by water molecules) are shown as black dashes. Residues and coordinating atoms other than oxygen of water molecules are labeled with red text. One extra Mg^{2+} in (a) is shown as a cyan sphere. The saliency values of each RNA atom are shown in blue scale, where the atom with larger saliency values are shown in a darker blue color.

As shown in Fig. 4.4, the coordinating atoms (connected through dashed lines) have relative large saliency values, indicating their importance in determining the Mg^{2+} binding sites. Indeed, as shown in Table 4.5, for the motifs shown in Fig. 4.4, As shown in Table 4.5, all of the binding sites can be successfully predicted by the MgNet model for the original RNA structures, however, after removing the coordinating atoms, MgNet fails to find the correct binding sites for six cases.

Table 4.5: Comparison between the performance of MgNet model for the original RNAs and for RNAs with the coordinating atoms removed. Number of successful MgNet predictions for each Mg^{2+} binding case. Predictions were made by five previously trained models obtained through five-fold cross-validation. However, for a binding case included in the cross-validation dataset (top-four cases), only the model trained without the case was used to make predictions. The purpose of this test is to show the importance of the coordinating atoms – the removal or change of the coordinating atoms would result in incorrect binding sites. The first column shows the PDB code and the Mg^{2+} identifier for the position of the bound ion. The column labeled with RNA and RNA^R shows the result for the original RNA and the RNA with coordinating atoms removed, respectively. The results are shown in the n/N format, where n and N represent the numbers of the MgNet model with successful predictions and of all the trained MgNet models, respectively. A prediction is successful if the the RMSD between the predicted ion site and the experimentally observed site is within 3 Å.

Mg^{2+}	RNA	RNA^R
2YIE-Z1116	1/1	0/1
3HAX-E200	1/1	1/1
3Q3Z-V85	1/1	0/1
2Z75-B301	1/1	0/1
3DD2-B1000	5/5	2/5
1VQ8-08004	5/5	0/5
4TP8-A1601	3/5	0/5
2QBA-B3321	5/5	0/5

To further investigate the spatial distribution of the RNA atoms around the bound ions, we classified four types of RNA atoms [41]: (i) O$_{\mathrm{ph}}$, phosphate oxygen (OP1/OP2); (ii) O$_{\mathrm{r}}$, oxygen in ribose (O2'/O4') or oxygen bridging phosphate and ribose (O3'/O5'); (iii) O$_{\mathrm{b}}$, nucleobase oxygen and (iv) $\mathrm{N_b}$, nucleobase nitrogen, where the last two types ($\mathrm{O_b}$ and $\mathrm{N_b}$) were further divided into subtypes according to the nucleotide type (purine or pyrimidine), resulting in overall six types.

We use radial distribution function (see **Methods**) to quantify the spatial distribution of the different types of atoms around a bound ion (see Fig. 4.5a). To further differentiate the effects of the different types of atoms, we define the radial distribution of the saliency value $h_{\mathrm{t}}(i)$ (for the correctly predicted ion binding sites):

$$h_{\mathrm{t}}(i) = \frac{s_{\mathrm{t}}(i)}{n_{\mathrm{t}}(i)} \tag{4.1}$$

where $s_{\mathrm{t}}(i)$ is the average of all the saliency values for the type-t RNA atom in the ith shell around the ion, and the denominator $n_{\mathrm{t}}(i)$ is the number of the type-t RNA atoms appearing in the ith shell. Physically, the saliency values $h_{\mathrm{t}}(i)$ indicates relative sensitivity of ion binding site to various types of RNA atoms.

The contact frequency distribution, as shown in Fig. 4.5a, shows two characteristic peaks at ∼2.3 Å and ∼4.3 Å, corresponding to inner-sphere and outer-sphere coordinations, respectively. The peak at ∼2.3 Å for $\mathrm{O_{ph}}$ indicates that $\mathrm{O_{ph}}$ is the most abundant inner-sphere coordinating atom, and the peak at ∼4.3 Å comes from the coordinations mediated by water molecules. For purine-$\mathrm{N_b}$, we found multiple nitrogen atoms in guanine/adenine residue that are spatially correlated, which explains the peaks around ∼4.3 Å and ∼6.3 Å. We note the the distribution curves become flat as distance increases, reflecting the relative abundance of these atom types in our cross-validation set.

The radial distributions of saliency values for volume occupancy and partial charge channels, as shown in Figs. 4.5b & c, are peaked at smaller radial distances than the contact frequency distribution shown in Fig. 4.5a. The shift in the peak positions is because $\mathrm{Mg^{2+}}$ is more sensitive to the coordinating atoms that have closer contacts with it. Furthermore, the saliency peaks of the different atom types in partial charge channel are higher than those in volume occupancy channel, except for $\mathrm{O_r}$. The result suggests that $\mathrm{Mg^{2+}}$ binding sites are more sensitive to the partial charges of the coordinating atoms than the occupancy of RNA atoms. The abnormal behavior of $\mathrm{O_r}$ may be caused by its spatial correlations with $\mathrm{O_{ph}}$. In the volume occupancy channel, $\mathrm{O_{ph}}$ and $\mathrm{O_r}$ often appear together as coordinating atoms thus show the similar peaks in the saliency distribution. In contrast, in the partial charge channel, the partial charge of an $\mathrm{O_r}$ is less than that of an $\mathrm{O_{ph}}$ thus shows a smaller peak (weaker sensitivity). Figs. 4.5c shows that purine-$\mathrm{O_b}$ atoms show the highest saliency peak. However, it is important to note that a saliency distribution in Figs. 4.5b & c represents only the average over different atoms for each RNA atom type. For example, the results for purine-$\mathrm{N_b}$ are averaged over ten different nitrogen atoms in a purine base, while the results for purine-$\mathrm{O_b}$ are only averaged over one oxygen in a purine base.

To identify the critical atoms for each case, we investigate the radial frequency distribution and the relative saliency distribution of each individual atom. Distributions of representative atoms within 3 Å (Fig. 4.5d) are similar to the radial frequency distributions (Fig. 4.5a), where the normalized distributions (Fig. 4.5d) are roughly twice as large due to the fact that O_{ph} contains two phosphate oxygen atoms (OP1 and OP2). The similar distributions suggest that these representative atoms are indeed the dominant inner-sphere coordinating atoms for each RNA atom type. Thus, the saliency distributions (Fig. 4.5e & f), which are dominated by RNA atoms with close contacts to $\mathrm{Mg^{2+}}$, also show similar trends as in Fig. 4.5b & c.

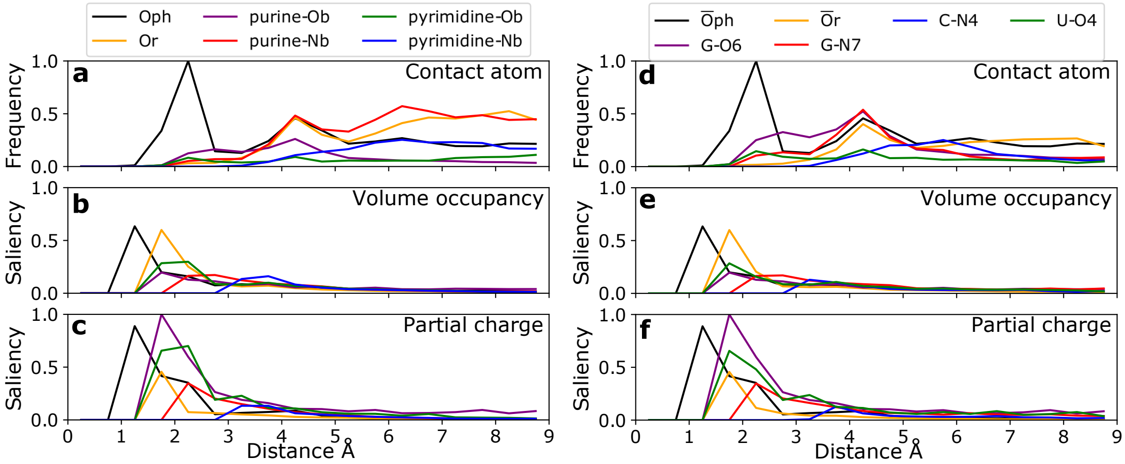


Figure 4.5: Radial frequency distributions and relative saliency distributions of different atom types and representative atoms around the correctly predicted Mg^{2+} ion sites. The figure shows the contact radial frequency distributions (a, d), the relative saliency distributions for the volume occupancies (b, e) and the partial charges (c, f) for the different RNA atom types, respectively. The frequencies and saliency values are normalized to the [0, 1] range. In (d-f), only the representative atom of each atom type is shown (with the same color as the corresponding atom type in (a-c)). $\overline{\mathrm{O}}_r$ is the average of two sugar oxygen atoms (O3′ and O5′) due to the similar radial frequencies and relative saliency distributions, and $\overline{\mathrm{O}}_{ph}$ is the average of the two phosphate oxygen atoms OP1 and OP2.

4.2.7 Identification of novel Mg^{2+} binding motifs

The MgNet approach led to two novel Mg^{2+} binding motifs [41]. Mg^{2+} binding motifs are defined as the recurring patterns of coordinating RNA atoms (i.e., geometric arrangement and atom type of the coordinating atoms). Typical Mg^{2+} can coordinate with 6 atoms forming octahedral geometry, these coordinating atoms are usually electronegative oxygen/nitrogen atoms from either water molecules or RNA molecules. In this study, we focused on motifs involving inner-sphere coordination with RNA atoms as MgNet does not treat outer-sphere coordinations (i.e., interactions mediated by water molecule).

For the MgNet prediction set with 373 representative sequences/structures (see **Meth-**

ods), MgNet predicts 1137 binding sites with inner-sphere coordinations, among which 313 are previously reported binding motifs and 654 are inner-sphere coordination binding sites with a single coordinating RNA atoms. For single atom-coordinated these sites, the bound Mg^{2+} ions could be partially dehydrated and it is possible that some of them belong to certain outer-sphere Mg^{2+} binding motifs if water-mediated outer-sphere interactions are considered. However, our current MgNet model is unable to identify the position of the coordinating water molecules thus Mg^{2+} coordinated by single RNA atom is not considered as a robust motif in this study.

The remaining 170 sites with inner-sphere coordination were examined and recurring specific patterns were identified. We found two new binding motifs, namely, the "16-member ring" and "Phosphate pyramid" (see Fig. 4.6). The 16-member ring motif involves two inner-sphere coordinating oxygen atoms from two phosphate groups, respectively, separated by one residue (not consecutive phosphate groups). The two coordinating oxygen atoms, the RNA backbone atoms in between, and the Mg^{2+} form a ring with 16 atoms (see Fig. 4.6a). The "Phosphate pyramid" motif contains either a "10-member ring" or a "16-member ring" with another inner-sphere ion coordinating the phosphate oxygen atoms, which makes motif look like a pyramid (see Fig. 4.6b).

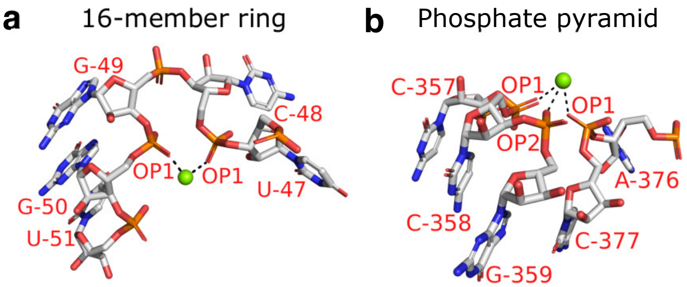


Figure 4.6: Representative sites for newly discovered Mg^{2+} binding motifs. Magnesium ions and inner-sphere interactions are shown in green spheres and black dashed lines, respectively. The coordinating RNA atoms and nearby nucleotides are labeled with red text. These representative sites are defined by PDB codes, chain id, and the predicted Mg^{2+} residue number as follows: (a) “16-member ring” (1QU2-T-9) and (b) “Phosphate pyramid” (4FAR-A-30).

We also calculated the relative abundance of the previously reported binding motifs and the newly found ones for both the MgRNA benchmark set [41] and the MgNet prediction set; see Fig. 4.7. The MgRNA benchmark set contains comprehensive high-quality Mg^{2+} binding sites selected from Mg^{2+}-containing RNA structures in Protein Data Bank [208]. This set was previously used in the study [41] to identify Mg^{2+} binding motifs. The percentage of each motif is calculated by dividing the number of the sites of the corresponding motif by the total number of sites with inner-sphere coordinating RNA atoms. For previously reported inner-sphere motifs, only top-5 abundant motifs are plotted. Histogram shows that the “Magnesium clamp” and “10-member ring” motifs are the top-2 abundant motifs in both the MgNet prediction set and the MgRNA benchmark set, and “G-phosphate”, “U-phosphate”, and “Y-clamp” motifs occur at similar levels. The newly discovered motifs are shown in the inset of the figure. The similar abundance of the “Phosphate pyramid” motif for both the MgNet prediction set and the MgRNA benchmark set indicates that this new motif was already presented in the MgRNA benchmark set and was probably overlooked in the previous study [41]. Interestingly, the abundance of the “16-

member ring" motif in MgRNA benchmark set is significantly lower than that in MgNet prediction set. By investigating the sites that are identified as a "16-member ring" motif in MgNet prediction set, we found that 65% of the sites belong to structures not included in the MgRNA benchmark set. Although these motifs are discovered by our machine-learning model, further computational and experimental studies are needed to validate these two newly identified motifs in RNA-Mg^{2+} interactions.

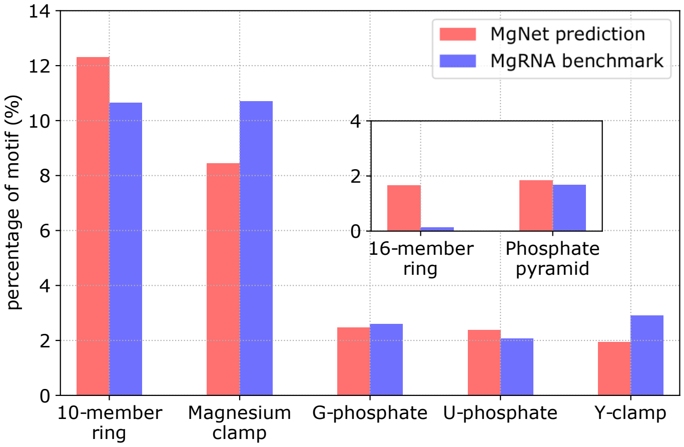


Figure 4.7: Relative abundance of top-5 previously reported and newly discovered inner-sphere Mg^{2+} binding motifs in MgNet prediction set (red) and MgRNA benchmark set [41] (blue). The two newly discovered motifs are shown in the inset. The percentage of each motif is calculated by dividing the number of the sites belonging to the corresponding motif by the total number of sites with inner-sphere coordinating RNA atoms.

4.3 Discussion

MgNet is a machine-learning method that uses a novel neural network (CNN) approach to predict Mg^{2+} binding sites for a given RNA structure. Currently, the model is trained to predict Mg^{2+} binding sites. With the increasing number of known RNA structures with bound ions, we can realistically expect the continuous improvement in the accuracy of

MgNet predictions and its applicability for other metal ions. Comparisons with other existing approaches such as MetalionRNA [243], MD simulations [29], and Brownian dynamics simulations [59] indicate that MgNet leads to notable improvements in the prediction of Mg^{2+} binding sites. Furthermore, saliency map analysis identifies and visualizes the RNA atoms that are most critical for Mg^{2+} binding. The information about the critical RNA atoms can facilitate our understanding of metal ion-RNA interactions. In contrast to physics-based models are usually excessively demanding in computational and human resources, with 3D RNA structures as the input and the predicted metal ion binding sites as the output, MgNet here can be conveniently implemented as a computationally efficient module that can be readily integrated into any automated processes.

4.4 Methods

4.4.1 Curating data sets

In order to generate a suitable collection of images, we used a set of 177 crystallographically determined structures containing both RNA and Mg^{2+} ions from the Protein Data Bank [208], including protein-RNA and DNA-RNA complexes (see Table 4.6-4.10). These 177 structures were selected according to the following criteria. Based on the sequence/structure equivalence classes ([249] version 3.54), for RNA PDB structures, we remove redundant structures of the same RNA molecule with similar Mg^{2+} binding sites. Due to computational limitations, for large RNAs, we select only 16S rRNA (∼1500 nucleotides). Compared with other large 18S, 23S and 28S rRNAs, 16S rRNA contains more binding sites for Mg^{2+} and less sites for other metal ions. Because the resolution of crystallographic structures is a key factor for accurate determination of the identity and position of Mg^{2+}, we kept only structures with resolution 3 Å or better. While allowing curation of a training set with sufficient data, this resolution cutoff serves to exclude structures that may misidentify Mg^{2+} binding sites. For structures with multiple models, we used the first model, and for residues with more than one alternative conformation, we used the first variant. In order to apply a five-fold cross-validation evaluation, the 177 RNA-containing structures were randomly divided into five subsets (Table 4.6-4.10).

Table 4.6: CV1 validation set.

1b23	1hq1	2a43	2g91	2oiu	301d	3cul	3l3c	3q51	3tzr
4l81	4qlm	4yco	5ew7	5ns4	5vjb	6b14	6cu1	1drz	1zz5
2cv1	2nok	2qus	354d	3ftm	3mei	3ssf	437d	4m30	4rge
5bjo	5ktj	5tpy	5wti	6c8d	6dta				

Table 4.7: CV2 validation set.

1duh	1ik5	1kxk	1nuj	2ann	2hw8	2yie	2zzn	3egz	3jxq
3loa	3ski	3v7e	4bwm	4nya	4p95	5dh6	5m0i	1feu	1j1u
1mms	1y26	2b8s	2oe5	2zzm	3cr1	3eph	3knc	3mxh	3t1y
430d	4g6r	4oji	4yb0	5lqt	5xus				

Table 4.8: CV3 validation set.

1jid	2nug	2qbz	3f4h	3hhn	3nd4	3oin	3u56	4frg	4m4o
4pdq	4xco	5d8h	5e54	5kpy	5ndh	5v0k	6dme	2fmt	2pjp
2val	3fs0	3ivn	3nkb	3td0	4en5	4ghl	4pcj	4pqv	4xw7
5ddp	5fj0	5mga	5u3g	5xtm					

Table 4.9: CV4 validation set.

1dfu	1f27	1hc8	1lnt	1mzp	1pjo	1yls	2ply	3cgs	3gvn
3la5	4oog	4znp	5btp	5lyv	5une	5y85	6dnr	1evv	1ffy
1hr2	1mji	1ntb	1y95	2g3s	364d	3d2x	3hax	3q3z	4z4f
5aox	5c9h	5t3k	5voe	6cc3					

Table 4.10: CV5 validation set.

1d4r	1jzv	1tra	2ao5	2q1r	2tra	3f2q	3oxd	4cs1	4jrc
4lx6	4tzx	5btm	5dar	5fj1	5u0q	6aso	6db9	1dk1	1l9a
1xjr	2fqn	2quw	2z75	3gx3	3zgz	4e8n	4k27	4rwn	4wkj
5ckk	5dhc	5nzd	5v2h	6c8o					

4.4.2 Defining 3D image

We used 24 Å $\times$ 24 Å $\times$ 24 Å cubic boxes to capture the information from binding and non-binding sites. The information contained in these boxes serve as the input “images” for learning. Similar to a 2D image having three color channels (red, green, blue), our 3D images contain two feature channels, volume occupancy and partial charge (see Table 4.11). For each channel in an image, there are $48 \times 48 \times 48$ voxels (pixels for 3D images), and each voxel has a volume of 0.5 Å $\times$ 0.5 Å $\times$ 0.5 Å. As a result, a 3D image is generated by two $48 \times 48 \times 48$ sized boxes stacked together.

Table 4.11: Feature channels used for the 3D descriptor.

Feature	Rule
volume occupancy	all the RNA atom types (not including Mg^{2+})
partial charge	partial charge values for all the RNA atom types (not including Mg^{2+})

4.4.3 Generating 3D images for RNA

For a given structure, each nucleotide is associated with an image. The midpoint between the backbone carbon atom C1$'$ and the base nitrogen atom connected to the C1$'$ atom is used as the origin for the corresponding image box. A local Cartesian coordinate system associated with each residue is set to avoid the need of image augmentation (i.e., 3D rotation transformation for each image). The space around the residue (within the image box) is discretized and filled with voxel values. The local coordinate system is set up according to the following steps. First, three key atoms are selected in the residue: O4$'$ and C1$'$ from the sugar ring and one nitrogen atom from base (N1 from uracil and cytosine or N9 from adenine and guanine, see Fig. 4.1a). Second, we calculated the vectors from C1$'$ to O4$'$ ($\mathbf{CO}$) and from C1$'$ to the base nitrogen atom ($\mathbf{CN}$). We selected vectors $\mathbf{CN}$ and $\mathbf{CN} \times \mathbf{CO}$ as the x- and the z-axis, respectively. The cross product of the z- and x-axes gives the y-axis.

We filled images with voxel values according to the Van der Waals radius r_{vdw} of each atom type. For each voxel in a property channel, we went through all RNA atoms to calculate the voxel occupancy. For example, we first calculate the distance r_{ij} between the RNA atom j and a given a voxel i. Then, we used a step-like function

$$n_i = f_j \times (1 - e^{-(\frac{r_{vdw}}{r_{ij}})^{12}}) \tag{4.2}$$

to evaluate the contribution of RNA atom j to the voxel value, where f_j represents the feature value associated with atom j. For the volume occupancy channel, f_j is 1, whereas for the partial charge channel, f_j is the partial charge of atom j. If more than one RNA atom contributes to the same voxel, we assign the average value from the contributors [250].

Only standard RNA residues were considered in this study.

In total, 15912 images were generated for the 177 structures. In the training process, we removed images with less than 300 non-zero voxels from the training set.

4.4.4 Labeling targets

Because training MgNet is a supervised learning task, we need to label each image with its true ion distribution and use image-label pairs to guide the learning process. In reality, the precision of ion positions in RNA structures is limited due to various factors. For example, X-ray diffraction can only resolve ion positions up to a certain resolution. In order to take these factors into consideration, we employed the distribution function in Eq. 4.2 (with $r_{vdw} = 2.5\text{Å}$ for Mg^{2+}) to account for the diffusiveness of the experimentally observed Mg^{2+} ions. The distribution of Mg^{2+} within each image box is used as the target label in MgNet training to compute the mean squared error (MSE) loss per voxel between the true and the predicted distributions. The minimization of the MSE loss guides the parameter training process in MgNet.

4.4.5 Choosing hyperparameters for MgNet

MgNet uses the two-channel 3D images of the RNA as the input and outputs a predicted Mg^{2+} distribution for each image. The network has 22 convolutional layers. Each of the first 21 layers contain 16 $3\times3\times3$ filters, and the last layer has only one $3\times3\times3$ filter. Here, we use $3 \times 3 \times 3$ filter to start with smaller filters. We used 16 filters each layer to optimize the usage of the GPU memory and the computer time spent on the training. Following a previous study [217], the batch normalization in each layer was applied immediately after the convolutional operation and before the "Rectified Linear Unit" [218] activation. The batch normalization [217] was also applied for the last layer before the final activation, and we replaced the "Rectified Linear Unit" [218] activation function with a sigmoidal activation function to keep the predicted voxel value in the range from 0 to 1. Based on the plain network, residual shortcut connections were inserted for every block with two hidden layers. The shortcut takes an identical input from a previous block and maps this identity shortcut right before the activation of the second hidden layer within the block (see Fig. 4.8 and Fig. 4.9). We initialize the weights [216, 219] and train all residual nets from scratch. To keep the input and output image sizes identical, we did not use any downsampling methods during the training.

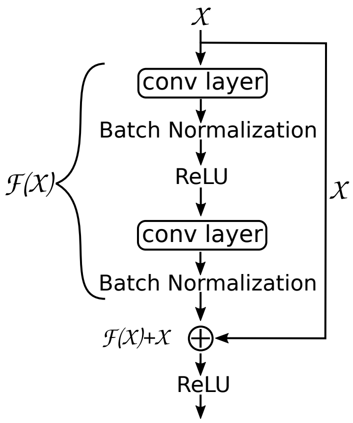


Figure 4.8: Block structure with a residual shortcut. Figure shows the entire block structure, where X is the input of this block (i.e., X is the output from the previous layer) and ReLU is the Rectified Linear Unit. Within this block, input X passes through two convolutional layers. The whole transformation in this block can be viewed as a function F, which maps input X to output $F(X)$, and an identity-mapping shortcut on the right-hand side adds X directly to the processed output $F(X)$.

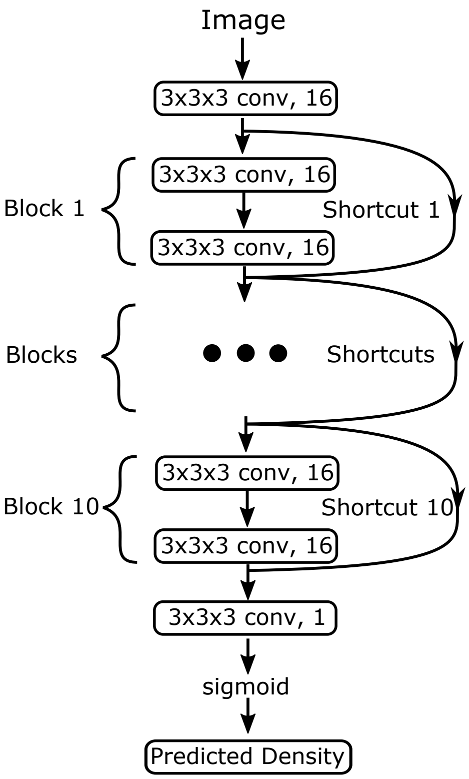


Figure 4.9: MgNet model. Ten blocks are stacked sequentially to make a 22-layer CNN. All convolutional layers have the same number of filters except for the last layer, which only has one filter. A sigmoidal activation function is applied to confine the predicted ion density within the $0 \sim 1$ range.

The only data preprocessing we used is the subtraction of the voxel mean from each image. For a given channel, the voxel mean was calculated by averaging the training set voxel values for all possible voxel positions in the corresponding channel. To center the data, we subtracted the voxel mean from each voxel value. This preprocessing was performed on the training, validation, and test sets.

For the network optimizer [220], we used default parameters provided by PyTorch [221] for momentum scheduling ($\beta_1 = 0.99, \beta_2 = 0.999$). A mini-batch size of 32 was used for training. The learning rate was initialized at 0.01 and divided by 10 at each plateau in training accuracy. The models were trained for up to 250 epochs. Our goal during the training is to minimize the weighted MSE loss function, L_w, which is calculated from the following equation

$$L_w = \sum_{n=1}^{N} \sum_{i,j,k=1}^{48} w_{ijk} \frac{(P_n(i,j,k) - G_n(i,j,k))^2}{48^3 N} \tag{4.3}$$

where N is the number of images, i, j, k is the voxel index, and $P_n(i,j,k)$ and $G_n(i,j,k)$ are the predicted and ground-truth ion distributions for the nth image, respectively. Further, the weights are defined as

$$w_{ijk} = \begin{cases} 1 & G_n(i,j,k) = 0 \\ 30 \cdot G_n(i,j,k) & G_n(i,j,k) \neq 0 \end{cases}$$

The above loss function gives the MSE between the predicted distributions and ground-truth distributions for all the voxels. Because the space sparsely occupied by Mg^{2+}, the data is highly imbalanced. The weighted loss function balances the learning process by increasing the penalty of a false negative prediction for positions that are truthfully occupied by Mg^{2+}.

4.4.6 Training and evaluating MgNet

To perform an unbiased evaluation for MgNet, we adopted a five-fold cross-validation procedure. For each fold, we trained MgNet for a total of 250 epochs with each epoch of training taking around 5 minutes. The training was conducted on 2 GTX 1080 Ti NVIDIA GPUs and one AMD Ryzen Threadripper 1950X 3.4 GHz 16-Core Processor. Because the loss quickly reached a plateau, we chose the model at epoch 40 as the final model.

4.4.7 Clustering to predict Mg^{2+} binding sites

Regions of the highest ion binding probability around the RNA were identified using the DBSCAN [245] clustering method. Within each high-probability region, we then used k-means clustering to generate N clusters, where N was determined from the ration between the volume (v) of the high density area and a preset cluster size (m). The representative points of the N clusters were chosen as predicted ion sites. These sites were combined and ranked based on the sum of the probabilities of all the voxels within the corresponding cluster. By changing the preset cluster size (m), we can adjust the number of clusters (N) within each high-probability region. The default clustering settings used in MgNet were obtained by optimizing the performance on the five-fold cross-validation through refining the preset cluster size (m).

4.4.8 Defining the evaluation metric

We used RMSD to evaluate the performance of MgNet for an individual test structure. The overall performance of the model for a large number of test structures was evaluated by the true positive rate (TPR) and the positive predictive value (PPV). TPR and PPV are calculated from the following equations:

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{P}} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{4.4}$$

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{4.5}$$

Here, P is the number of positive (experimentally observed) cases, TP (true positive) is the number of predicted Mg^{2+} that resides within a 3 Å sphere around an experimentally observed Mg^{2+}, and FP (false positive) is the number of predicted Mg^{2+} that falls outside the 3 Å range from experimentally observed Mg^{2+} ions.

4.4.9 Calculating radial frequency distribution

The radial frequency distribution in Fig. 4.5 is generated from the following steps. First, we find all the bound Mg^{2+} ions in the training set. Second, for each Mg^{2+} in the training set, the space within 9 Å around the ion is discretized into 18 spherical shells, each having a shell thickness of 0.5 Å. For each Mg^{2+}, we locate all the RNA atoms within the 9 Å sphere and bin them in the shells. Then, according to the different types of coordinating atoms, we counted the frequency of each coordinating atom type in the spherical shells for all the Mg^{2+} ions and computed the radial frequency distribution for every coordinating RNA atom type. The radial frequency in each spherical shell (or the distance bin) is normalized by the volume of the corresponding shell:

$$f_{\mathrm{t}}(i) = \frac{n_{\mathrm{t}}(i)}{v(i)} \tag{4.6}$$

where $n_{\mathrm{t}}(i)$ is the number of the type-t RNA atoms appearing in the ith shell for the bound Mg^{2+}, and $f_{\mathrm{t}}(i)$ is the frequency normalized by the corresponding shell volume $v(i)$.

4.4.10 Collecting dataset for motif identification

RNA structures used in motif identification were collected from nucleic-acid database (NDB). Initially, 980 crystallographically determined Mg^{2+}-containing structures with resolution better than 3 Å were downloaded. To avoid the redundancy in the dataset, we reduced the 980 Mg^{2+}-containing structures to 350 crystal structures with 373 representative sequence/structure equivalence classes according to the representative set of RNA 3D structures [249].

Table 4.12: RMSD table of the MD simulation-based method and MgNet on seven test structures. Column ${Mg^{2+}}_{CI}$ and ${Mg^{2+}}_{PS}$ are the average RMSD values and standard deviations of MD method between experimental and predicted binding sites during the production phase. Top 50 predicted sites were used in MD method. Columns cv1 to cv5 are the predictions made by MgNet with default clustering settings, shown in RMSD values with ranks in parentheses. In MgNet model, only predicted sites with RMSD less than 20 Å are listed in the table, experimental ions with no predicted sites within 20 Å are labeled with dash. (Contents of the table are shown in next page.)

PDB	Ion	$\mathrm{Mg}^{2+}_{\mathrm{CI}}$	$\mathrm{Mg}^{2+}_{\mathrm{PS}}$	cv1	cv2	cv3	cv4	cv5
1D4R	MG-90	1.0 ± 0.5	1.1 ± 0.5	2.9 (1)	2.4 (1)	1.6 (1)	1.6 (2)	2.3 (1)
	MG-91	5.0 ± 0.7	4.4 ± 0.7	2.8 (4)	5.0 (2)	3.4 (2)	3.6 (1)	3.5 (2)
2MTK	MG-48	7.4 ± 3.2	5.8 ± 1.9	4.9 (1)	5.2 (7)	4.7 (7)	3.9 (2)	5.2 (5)
	MG-49	3.9 ± 0.9	2.9 ± 1.6	3.1 (6)	4.0 (5)	4.1 (8)	3.7 (7)	3.0 (8)
	MG-50	6.7 ± 3.0	5.7 ± 2.3	1.3 (2)	1.5 (1)	1.8 (4)	1.8 (3)	1.5 (2)
	MG-51	3.2 ± 0.8	3.5 ± 0.4	19.5 (4)	1.5 (8)	11.5 (3)	1.2 (1)	2.0 (9)
	MG-52	3.6 ± 0.5	3.8 ± 2.4	7.6 (3)	8.0 (3)	7.6 (1)	8.2 (5)	7.2 (6)
	MG-53	2.1 ± 0.4	2.3 ± 1.1	2.0 (7)	1.6 (2)	3.2 (9)	1.5 (4)	2.0 (1)
2QEK	MG-49	2.5 ± 0.2	2.5 ± 1.3	1.5 (4)	1.6 (6)	9.8 (1)	8.9 (4)	9.9 (1)
4FRG	MG-179	2.4 ± 0.7	4.4 ± 0.8	1.5 (2)	0.9 (2)	1.8 (1)	1.5 (1)	0.4 (6)
	MG-180	2.4 ± 0.8	5.3 ± 0.4	0.9 (4)	1.7 (8)	1.4 (8)	1.6 (5)	1.9 (5)
	MG-181	2.8 ± 0.5	1.4 ± 0.5	18.6 (5)	18.9 (11)	-	-	18.8 (10)
	MG-182	7.6 ± 0.5	7.0 ± 0.6	1.8 (10)	2.0 (7)	1.9 (11)	2.1 (4)	-
	MG-183	3.7 ± 1.5	4.7 ± 2.9	2.2 (1)	2.5 (1)	0.7 (5)	2.0 (2)	0.8 (2)
	MG-184	1.1 ± 0.3	2.0 ± 1.5	1.2 (3)	3.4 (3)	2.1 (10)	1.3 (6)	2.6 (1)
	MG-185	3.7 ± 1.3	5.9 ± 1.4	1.1 (9)	0.8 (10)	7.5 (6)	13.3 (7)	1.4 (3)
4JF2	MG-94	2.2 ± 1.1	2.7 ± 1.0	0.4 (3)	0.7 (1)	0.8 (4)	0.9 (6)	0.5 (1)
	MG-95	3.2 ± 0.6	4.6 ± 0.7	16.0 (5)	1.1 (8)	1.1 (2)	1.0 (4)	1.0 (4)
	MG-96	2.5 ± 0.8	2.9 ± 0.9	0.4 (7)	0.6 (4)	0.6 (3)	0.5 (7)	0.5 (5)
	MG-97	18.3 ± 2.7	20.6 ± 0.8	0.7 (2)	0.7 (2)	0.9 (1)	0.6 (1)	0.3 (3)
4KQY	MG-121	1.8 ± 0.4	3.4 ± 0.9	1.9 (1)	1.3 (3)	1.2 (5)	0.7 (1)	1.4 (1)
	MG-122	1.8 ± 0.5	4.2 ± 2.0	7.1 (2)	5.4 (8)	5.7 (8)	5.3 (6)	8.5 (7)
4P5J	MG-85	1.1 ± 0.7	1.8 ± 0.2	1.6 (5)	1.1 (1)	1.4 (6)	1.2 (2)	2.0 (1)
	MG-86	2.5 ± 0.5	3.5 ± 2.1	1.3 (6)	1.3 (2)	1.4 (2)	1.3 (1)	1.2 (3)

Table 4.13: TPR and PPV of MD simulation-based method and MgNet on seven test structures. Table shows the TPR and PPV for both MD simulations and MgNet. Column $\mathrm{Mg}^{2+}_{\mathrm{CI}}$ and $\mathrm{Mg}^{2+}_{\mathrm{PS}}$ are the results of MD simulations with the different ion conditions. Columns cv1 to cv5 are the predictions made by the corresponding trained MgNet models. Only predicted sites with RMSD less than 3 Å are considered as true positive ones. For MD simulations, predictions were made by using the top 50 predicted sites. Since the default clustering settings of MgNet tend to give fewer predictions than the MD simulations (i.e., 50 sites), we provide MgNet results with two different settings. One uses default clustering settings and the other one uses the same number of predicted binding sites for each structure as the MD simulations (i.e., 350 predicted sites for seven structures). The results of these two settings are shown as TPR and PPV without and with asterisk, respectively).

	$\mathrm{Mg}^{2+}_{\mathrm{CI}}$	$\mathrm{Mg}^{2+}_{\mathrm{PS}}$	cv1	cv2	cv3	cv4	cv5
TPR	-	-	70.83%	70.83%	58.33%	66.67%	70.83%
PPV	-	-	31.48%	26.98%	25.93%	29.63%	32.08%
TPR*	54.17%	37.50%	87.50%	87.50%	87.50%	91.67%	91.67%
PPV*	3.71%	2.57%	6.00%	6.00%	6.00%	6.29%	6.29%

Table 4.14: Details of MgNet Architecture. Each building block is shown with two convolutional layers together without line separation. No downsampling is performed in this network, so the stride has size equal to 1 for all layers.

Block	Layer	Output Size	Filter size	Filter number	Padding
first conv	conv1	$48 \times 48 \times 48$	$3 \times 3 \times 3$	16	1
block1	conv2	$48 \times 48 \times 48$	$3 \times 3 \times 3$	16	1
	conv3	$48 \times 48 \times 48$	$3 \times 3 \times 3$	16	1
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
block10	conv20	$48 \times 48 \times 48$	$3 \times 3 \times 3$	16	1
	conv21	$48 \times 48 \times 48$	$3 \times 3 \times 3$	16	1
last conv	conv22	$48 \times 48 \times 48$	$3 \times 3 \times 3$	1	0

Chapter 5

RNA-ligand molecular docking: Advances and challenges

This chapter was published[1].

With the rapid evolution of computer algorithms and hardware, computational modeling of RNA-small molecule interactions has become an indispensable tool for novel drug discovery. Fast and accurate virtual screening has led to a drastic acceleration in the selection of effective, RNA-targeted small molecules as drug candidates. The docking-scoring method is the main approach for the virtual screening of RNA-targeted drugs. Accurate docking-scoring modeling needs to tackle four crucial problems: (1) conformational flexibility of the ligand, (2) conformational flexibility of the RNA, (3) complete sampling of binding sites and binding poses, and (4) accurate scoring of different binding modes. In addition to the problems associated with conformational flexibility, RNA molecules are neg-

[1] Yuanzhe Zhou, Yangwei Jiang, and Shi-Jie Chen. "RNA-ligand molecular docking: Advances and challenges". In: *WIREs Computational Molecular Science* (2021), e1571. DOI: https://doi.org/10.1002/wcms.1571. eprint: https://wires.onlinelibrary.wiley.com/doi/pdf/10.1002/wcms.1571. URL: https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/wcms.1571.

atively charged polymers, further complicating scoring functions for RNA-ligand binding. Advances in physics-based and knowledge-based scoring functions have shown highly encouraging success in predicting ligand binding modes and binding affinities. Furthermore, recent reports suggest that including dissociation kinetics (ligand residence time) in predictive models can improve performance in estimating in vivo drug efficacy. Moreover, the rise of deep-learning computational approaches has led to new tools for predicting RNA-small molecule binding. This review focuses on the recently developed computational methods for predicting RNA-ligand binding and their respective pros and cons.

Graphical Abstract

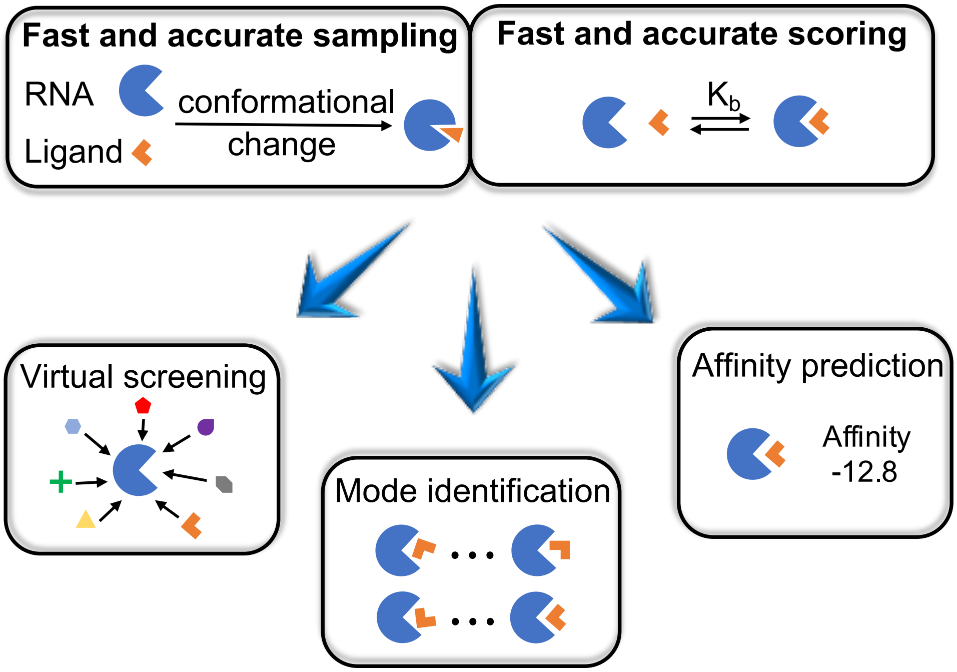


Figure 5.1: RNA-targeted drug discovery requires the synergy of enhanced sampling and accurate scoring with fast computational speed. The distinct aspects of RNA-ligand docking compared to protein-ligand docking pose unique challenges, which demand a new generation of molecular docking models. This review presents an overview of recently developed RNA-ligand docking methods for RNA-targeted drug discovery.

5.1 Introduction: targeting RNA with small molecules is a highly promising strategy for drug discovery

Ribonucleic acid (RNA) molecules are transcribed from DNA in the cell nucleus and play a variety of critical roles in gene expression and regulation at the level of transcription and translation. According to their cellular functions, RNA molecules can be categorized into two types: messenger (coding) RNAs (mRNAs) that encode the amino acid sequences and are translated into proteins, and noncoding RNAs (ncRNAs), which, instead of encoding amino acid sequence, serve as enzymatic, structural, and regulatory components for gene expression. With the coding RNAs occupying only <3% of the human genome [1–

3], the vast majority of the human genome sequences are transcribed to ncRNAs, such as ribosomal RNAs (rRNAs), microRNAs (miRNAs), small interfering RNAs (siRNAs), small nuclear RNAs (snRNAs), and various riboswitches [11, 12]. With the ever increasing discoveries of new RNA structures and cellular functions and the continuous developments of powerful RNA structure determination methods, RNA-based therapeutics are becoming new promising methods to treat human disease. In general, RNA-based therapeutics can be classified into two types. In the first type, therapeutic RNAs—including RNA aptamers, antisense oligonucleotides (ASO), small interfering RNAs (siRNA), and guide RNAs (gRNA)—bind to the target (e.g., RNA transcripts, DNA targets, and protein targets) to inhibit or induce targeted biochemical reactions. This approach has attracted tremendous interest in the field of gene therapy and has been under very active development [13–16]. In the second type of RNA-based therapeutics, an RNA molecule serves as the target for drug (small molecule) binding [2, 4, 16–21]. This second approach is analogous to protein-targeted drug discovery. However, in comparison to RNA targets, only $\sim$1.5% of the human genome encodes protein [2–7], and of these protein-encoding genes, only 10-15% are disease-related [2–4, 8–10]. The availability of druggable protein targets is further restricted by the structural and energetic fitness required for high-affinity drug binding. In contrast, genes that are undruggable or difficult to drug by targeting their associated proteins may be inhibited by drugs targeting the corresponding protein-encoding mRNA sequence. Therefore, compared to proteins, RNAs show much broader druggability. Additionally, noncoding RNAs play important roles in most human diseases from cancer to viral infection such as COVID-19. Targeting the large number of noncoding RNAs would open up remarkable new opportunities for drug discovery. For example, antibiotics targeting ribosomal RNA (rRNA), which forms the active site of a bacterial ribosome, effectively inhibit bacterial protein synthesis [74–77]. Specific small molecules (ligands) bound to common riboswitches in bacterial cells regulate gene expression through ligand-induced conformational changes of the RNA [78–92]. To inhibit viral replication, a potentially effective strategy is to use small molecule as a drug to target viral RNA motifs which are often often highly structured [16, 18, 93], such as the HIV transactivation response (TAR) element in the 5′ untranslated region [94, 95], the internal ribosome entry site (IRES) element located in the hepatitis C virus (HCV) genome [96–100], and the influenza A virus RNA promoter [101, 102]. Screening small molecule compounds selected from the compound library against an atypical three-stemmed RNA pseudoknot that stimulates -1 programmed ribosomal frameshifting [103, 104] in SARS-Cov RNA genome shows inhibition of the -1 ribosomal frameshifting of SARS-CoV with an IC_{50} of 210 μM [105, 106]. In addition to the above examples, many precursor messenger RNAs and microRNAs have shown great promise as therapeutic targets [16, 20, 21].

Compared with predicting protein-ligand interactions, which remains a challenging problem, modeling binding interactions between RNA and small ligand molecules presents three unique challenges. First, unlike a protein, RNA is highly charged, with each phosphate group carrying one electronic charge. Thus, RNA folding and ligand binding require the participation of metal ions such as Mg^{2+} and water molecules to stabilize the binding pocket structure of the RNA and to mediate ligand-RNA interactions [107–110]. Second, RNA molecules are often quite flexible, capable of folding into multiple stable conformations, and ligand binding often induces structural switches between different conformers or change the structure of an RNA receptor. Compared with protein-ligand binding, ligand binding sites on RNA can be less deep and more polar, solvated, and conformationally flexible [3, 18, 110], which adds further complexity to predicting RNA-small molecule interactions. Third, the fact that we have a limited number of experimentally determined structures for RNA molecules and RNA-ligand complexes makes pure knowledge-based approaches less effective for RNA-ligand predictions. In this regard, a physics-based approach or a hybrid knowledge-based and physics-based approach can yield unique advantages [109, 111–120, 251–253].

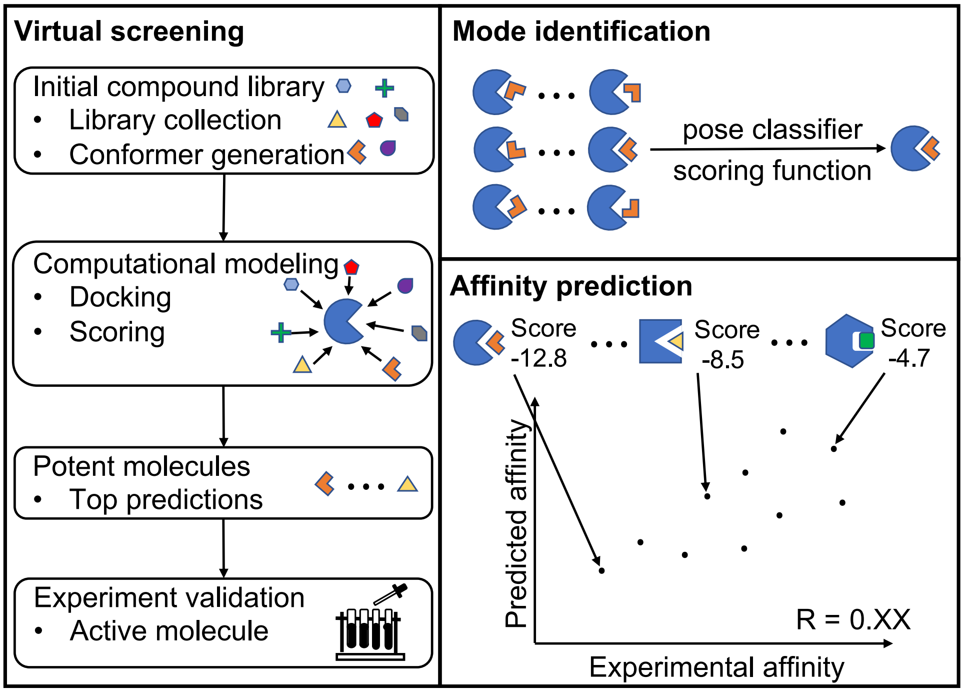


Figure 5.2: Three major applications of an RNA-ligand interaction model. Virtual screening involves docking against small molecules in a large library and scoring every docked pose. Top-scored selections are treated as the most promising candidates for putative binders. For a given RNA-ligand pair, computational models for ligand binding pose identification and RNA-ligand binding affinity prediction rely on scoring the possible RNA-ligand complex structures. An ideal scoring function for ligand binding pose identification should have the ability to distinguish the native pose from a large pool of docked decoy poses, while achieving the maximum correlation between the predicted scores and the experimental affinities for different RNA-ligand pairs.

Although successful computational tools have been developed for protein-ligand binding [159–162], the difference in chemical structure and energetics between RNA and proteins demands new methods for RNA-ligand interactions. These new methods for small molecule selection, shown in Fig. 5.2, are necessary for virtual screening, binding mode identification, and binding affinity prediction of specific RNA targets. A successful computational drug discovery requires the integration of three key components: (i) a method to identify the druggable RNA target, (ii) a computationally efficient sampling algorithm for RNA conformations, ligand conformers, and ligand poses, (iii) accurate scoring functions to assess the RNA-ligand complex structures and evaluate the binding affinity. In this review, we focus on computational challenges in predicting RNA-ligand interactions, with specific emphasis on the recent advances.

5.2 Methods for identifying druggable RNA targets and binding sites

5.2.1 Identifying druggable target RNAs

The druggability of a particular RNA target depends on the answers to three questions. First, does the inhibition/enhancement of the target RNA function lead to effective control of the disease? Second, is the RNA target accessible for the small molecule binding in cellular environment? Multiple factors can affect the accessibility of the target RNA, such as the abundance and lifetime of the target RNA in disease-related cellular processes [2, 4, 16]. Third, does the target RNA adopt binding site that enables small molecule binding with high affinity and high specificity? Small molecule targeting the particular RNA with high specificity can reduce the off-target side effects. An effective way to achieve high specificity is to target RNA that is unique in the diseased cells, pathogenic viruses or bacteria, such as riboswitches which are common in bacteria but rarely occur in eukaryotes. Another way is to computationally identify RNA motifs that is able to form unique and high-affinity pockets capable of small molecule binding [2, 4, 254].

Inforna [255, 256] is a template-based method capable of selecting RNA targets according to RNA secondary structure motifs such as hairpins, symmetric and asymmetric internal loops, and bulges. The current Inforna 2.0 template database [256] contains 1936 pairs of known RNA secondary structure motif-ligand bound complexes [256]. For a given RNA target, Inforna 2.0 [256] identifies RNA secondary structure motifs and from the template database, for a given motif, finds the corresponding ligand partners with fitness scores [257, 258]. The fitness score [257, 258] provides a measure of RNA-ligand binding affinity as well as the selectivity of the RNA motif against many other small molecules [256]. RNAs of high selectivity and affinity fitness scores are more likely to be druggable. With top scored small molecules as lead compounds, chemical similarity screening of compound library gives potential potent binders. Inforna [255, 256] has been proved to be successful in various studies [4], such as identifying small molecules that target oncogenic miRNA precursors [259, 260] and an A bulge in the (iron responsive element) IRE [261] of the SNCA mRNA related to Parkinson's disease [262].

On the basis of RNA secondary structures, Warner et al. [2] showed that information content [263, 264] can be used to identify druggable RNA motifs for potentially high-specificity and high-potency binding [265]. Information content measures the amount of information (in bits) required to specify the sequence and structure complexity of an RNA motif, where motifs with high bits ($\sim$30 bits) are more complex and more likely to be unique in the transcriptome [2]. In experiments, RNA structural information content is attainable through chemical probing techniques such as selective 2′-hydroxyl acylation analyzed by primer extension (SHAPE) [163, 266–268]. Focusing on RNA motifs with sufficient complexity (high information content) can lead us to RNAs with high binding specificity and affinity thus higher druggability. As an example, the binding affinities of GTP [264, 269] and targaprimir-96 [270] both show a strong correlation with the information content of the RNA motifs, where a 10 bits increase in information content results in a 10-fold increase in binding affinity [2].

5.2.2 Identifying binding sites for a given target RNA

An overall assessment of binding pockets

A recent statistical analysis demonstrates that many RNAs indeed fold into structures that form pockets amenable to selective small molecule binding [254]. To identify potential RNA suitable for ligand binding, Hewitt et al. [254] have evaluated RNA binding pockets using PocketFinder [265] for 1552 structured RNAs and all the proteins in the Protein Data Bank (PDB) [208], where a binding pocket is described by the volume and the solvent exposure of the pocket (buriedness) and the fraction of the pocket considered to be hydrophobic (hydrophobicity). The results suggest that although ligand-bound pockets on RNAs and proteins show overall similar physical properties, RNA pockets are on average less hydrophobic than their protein counterparts [254]. Moreover, compared to the unbound pockets of RNA, the ligand-bound pockets are generally larger in volume, more buried, and more hydrophobic [254].

Geometry-based methods

In search for binding sites based on RNA-ligand shape complementarity, DOCK 6 [112] selects the binding pockets from a negative image of the receptor surface, where each cavity is characterized by a set of overlapping spheres [271]. Similarly, rDock [118] applies a two-sphere mapping algorithm to identify the binding sites. Within the defined docking space, large spherical probes are placed on each grid point to rule out superficial and shallow sites. Then, small spherical probes are placed on the remaining unallocated grid points to map the cavities that serve as the possible binding sites [118]. Wide and shallow minor grooves of RNA, which can geometrically accommodate a wide range of ligand shapes and serve as non-specific binding sites, are often identified as putative binding pockets and cause false-positive predictions.

Energy-based methods

Other programs find binding sites by estimating the overall probe-pocket interaction energies, where the probes are virtual atoms and traverse the surface of the receptor. PocketFinder [265] and AutoLigand [272] are such cases and are equipped in ICM [273] and AutoDock [274], respectively. PocketFinder uses a Lennard-Jones (LJ) potential to describe the interactions between probe atoms and receptor atoms, and grid maps generated from the calculated interactions are used to identify the binding sites. AutoLigand uses a similar approach but involves an extra iterative step to identify the optimal binding sites from the grid maps, and it accounts for connections between the neighboring possible pockets.

Network and machine-learning approaches

By treating RNA-ligand interaction as a network of contacting atoms, network-based approaches have shown great promise in the prediction of the functional sites in RNA-ligand interactions. For example, using inter-nucleotide Euclidean (hamming) distance network for a 3D or 2D structure Rsite [275] and Rsite2 [276] predict the functional sites for RNA-ligand binding from the maximally closely clustered nucleotides. However, since the inter-nucleotide networks in Rsite and Rsite2 do not distinguish the different connection types between the nucleotides, these models often lead to false positive predictions. To distinguish the different connection types, RBind [277, 278] transforms an RNA structure into a graph, where a node and a edge denote a nucleotide and a noncovalent contact between the nucleotides, respectively, and predict the functional sites as regions formed by nucleotides of the maximum closeness. On a test set with 19 RNA-ligand complexes, RBind (average positive predictive value PPV = 0.67) outperforms Rsite (average PPV = 0.42) and Rsite2 (average PPV = 0.40) [277]. The result suggests that the different types of inter-nucleotide interactions encoded in the RNA structure provide important information for the prediction of the functional sites. RNAsite [279], a random forest-based model, uses sequence-based and/or structure-based descriptors to predict whether a given nucleotide belongs to the functional sites. In the model, a nucleotide is defined as part of a functional site if it contains an atom within 4 Å to the target ligand. In RNAsite, four different sets of features for each nucleotide are extracted: geometrical features of local surface convexity/concavity (Laplacian norm), topological features of the RNA nucleotide interaction network similar to the one used in RBind [277, 278], nucleotide-specific accessible surface areas, and position-specific evolutionary conservation of the nucleotide calculated from multiple sequence alignment. The model is trained on 60 RNAs with five-fold cross validation and tested on two separate sets with 19 (RB19) and 18 (TE18) RNAs, respectively. By using Mathews correlation coefficient (MCC) and area under the receiver operating characteristic curve (AUC), RNAsite [279] shows better performance compared to Rsite [275], Rsite2 [276] and RBind [277, 278], with 0.253 and 0.567 for MCC, and 0.776 and 0.877 for AUC on TE18 and RB19 sets, respectively. The promising results indicate the necessity to include more independent features as descriptors. Although these models have been trained specifically for RNA-small molecule complexes, further improvements are possible, for example, through a combination with other machine learning methods [280–283].

5.3 Methods for efficient sampling of ligand binding modes with flexible conformations — a major challenge in RNA-ligand docking

Exhaustive sampling of possible RNA-ligand complex structures is challenging due to the flexible nature of RNA and small molecules. Additionally, following the induced-fit effect or the conformational selection mechanism, RNA targets often undergo conformational changes in response to ligand binding [92, 284–286] (see Fig. 5.3a). This leads to the coupling between RNA folding, including cotranscriptional folding, and ligand binding when virtual screening is performed [92, 284–287]. A widely used approach to tackle this problem is ensemble docking [94, 95, 107, 121], where a ligand docks into an ensemble of RNA structures. An alternative approach is to sample the conformational changes on the fly in the docking process. Various methods [111, 112, 115, 117–119] have been developed to treat flexible docking. However, in part due to the required computational time for large-scale virtual screening, predicting large conformational changes remains a challenge.

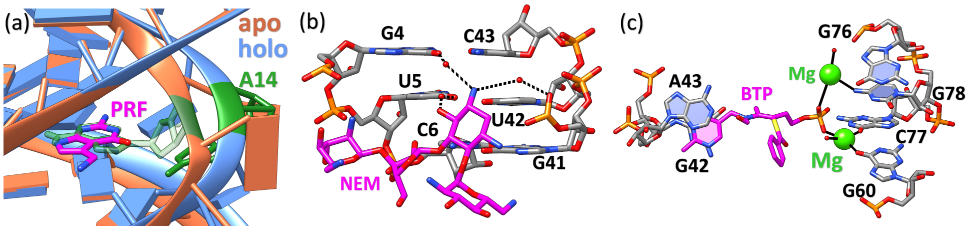


Figure 5.3: RNA conformational changes and binding interactions mediated by water molecules and ions. (a) The local structure difference of preQ1 riboswitches between apo (ligand-free) and holo (ligand-bound) states. The structure in orange denotes the apo state (PDB code: 6VUH [288]) and the structure in blue denotes the holo state (PDB code: 3Q50 [289]) with its bound small molecule (PRF) colored in magenta. Upon binding, the small molecule displaces residue A14 (colored in green for both apo and holo states) and causes the local structural transition. (b) Water molecules mediated RNA-ligand interactions. Water molecules form a bridge between small molecule Neomycin B (NEM, magenta) and 16S-rRNA A-site (PDB code: 2ET4 [290]). The isolated red dots denote the oxygen atoms in water molecules. The black dashed lines show the water-mediated hydrogen bonding contacts that promote NEM binding to the RNA receptor. (c) Metal ions in RNA-small molecule interactions. The ligand benfotiamine (BTP, magenta) interacts with residues G60, C77, and G78 of the *Thi*-box riboswitch through two magnesium ions (green) and the G42-A43 base stack (PDB code: 2HOO [291]). The black solid lines represent the inner sphere metal ion coordination. The polyanionic RNA recognizes the positively charged metal ion complex made up of the monophosphorylated compound and cations.

For ligand docking to RNA targets, such as ribosomal RNAs and riboswitches with reliable binding site information [292, 293], local sampling with a rigorous energy scoring function can often provide accurate predictions for the ligand binding pose. However, for a broad range of therapeutic RNAs including viral genomic RNAs [294, 295], the lack of the binding site information poses a great challenge to drug screening. For RNAs, unlike proteins, we have limited examples of RNA-ligand bound structures and scarce knowledge about the binding sites. This fact highlights the importance of blind docking (vs. local docking), where a small molecule is docked to the entire surface of the receptor without any prior knowledge of the binding site (see Fig. 5.4). Although computational models—including AutoDock Vina [117], GOLD [113], and Glide [114]—or models originally developed for protein-ligand docking, but optimized for RNA targets—such as AutoDock [274], ICM [273], DOCK 6 [112], and FITTED [251]—can be adopted for RNA-ligand docking, methods developed specifically for RNA targets, such as RiboDock [115], rDock [118], MORDOR [111], and RLDOCK [119, 120], have demonstrated advantages for RNA systems. See Table 5.1 for a list of docking software.

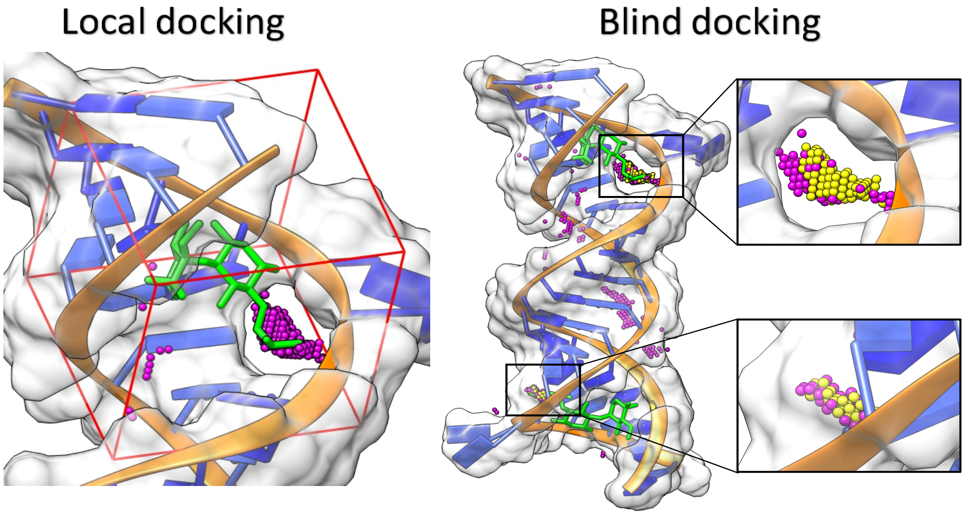


Figure 5.4: The difference between local and blind docking. A complex of an aminoglycoside antibiotic, gentamicin (green) and the 16S-rRNA A site of bacterial ribosome is used for illustration (PDB code: 2ET3 [290]). In this example, both docking (local & blind) processes are carried out using the RLDOCK model [119, 120]. In local docking, the binding pocket is predefined and the sampling is contained within the red dashed box. The small magenta spheres denote candidate binding sites predicted by RLDOCK. In blind docking, the binding site detection is performed across the whole surface of the RNA. The small yellow and magenta spheres denote the predicted high- and low-probability binding sites, respectively. Two cavities identified by RLDOCK (anchored by yellow spheres) are zoomed out separately.

Table 5.1: Docking programs available for RNA. Docking programs without dedicated binding site detection module are shown with a dash.

Docking program	Target	Conformational search algorithm	Binding site prediction
AutoDock Vina [117]	protein	Monte Carlo & quasi-Newton	-
GOLD [113]	protein	genetic algorithm	-
Glide [114]	protein	Monte Carlo	SiteMap [296]
AutoDock [274]	protein/RNA	genetic algorithm	AutoLigand [272]
ICM [273]	protein/RNA	Monte Carlo	PocketFinder [265]
DOCK 6 [112]	protein/RNA	incremental construction	sphgen module [271, 297]
FITTED [251]	protein/RNA	genetic algorithm	-
RiboDock [115]	RNA	genetic algorithm	two-sphere filter
rDock [118]	protein/RNA	genetic algorithm Monte Carlo simplex minimization	two-sphere filter
MORDOR [111]	RNA	molecular dynamics	grid-based systematic search
RLDOCK [119, 120]	RNA	multi-conformer docking	grid-based systematic search

5.3.1 Sampling of ligand conformations

There are three general ways to treat flexible ligand conformations in docking [298–300]: multi-conformer docking, incremental construction, and stochastic optimization. These three strategies differ in their computational speed and the conformational adaptability of the docked ligand to the geometric features of the binding-site.

Multi-conformer docking

The multi-conformer docking algorithm prepares a conformational ensemble for a ligand (small molecule) and performs rigid docking for each the ligand conformers against the same target [301]. For a given binding pocket, this method can be computationally fast if a limited number of conformers are docked. A key determinant for the success of this approach is that the near native conformations of the ligand must be included in the ligand conformer ensemble. Currently, there exist a number of ligand conformer generators, such as OMEGA [302, 303], RDKit [304], and Open Babel [305]. These models have been shown in benchmark studies to reproduce reliable conformational ensembles of small molecules within seconds [305, 306]. Combining a new molecular dynamics approach and a quantum mechanically-refined ligand-RNA interaction force field, a recently released conformer generator has led to improved accuracy for *in silico* drug design [307, 308]. Its web-based server and the database of bioactive conformational ensembles not only speed up the process of finding experimentally favorable ligand conformations through massive docking but also provide proper initial structures for further optimization [309]. In addition, in the docking process, a ligand conformer ensemble is constructed prior to conformational sampling, thus, the conformer ensemble can be appropriately built for the small molecules in question. For example, a ligand conformer ensemble can be generated with bias toward the low-energy states [119, 120] or with maximum diversity in the conformational space.

RNA–Ligand DOCKing (RLDOCK) [119, 120] is a recently developed docking model for flexible ligands using a multi-conformer approach. In RLDOCK, the ligand-binding mode is described by four variables (R, L, A, O), where L denotes the ligand conformer, A denotes the ligand atom placed at (anchor site) R, and O is the 3D rotation angle of L about A (at position R). For each RNA-ligand pair, the RLDOCK algorithm generates an ensemble of flexible ligand conformers and binding poses through the following steps.

1. The algorithm generates an ensemble of viable anchor sites R based on the following two criteria: (a) there should be no steric clash between ligand and RNA atoms and

(b) the RNA structural environment around R should form a pocket geometry.

2. Based on the viable anchor sites R generated above, by exhaustively enumerating all the different combinations of (R, L, A, O), RLDOCK samples all the possible ligand binding sites and binding poses. The results are stored for subsequent refinement. Before applying the scoring function to rank order all the binding modes, to accelerate computational speed, RLDOCK first sieved the exhaustive ensemble by removing those with high LJ potential between RNA and ligand.

3. All R sites with low LJ potentials U_{LJ}(R) (below the threshold) are selected as preferred R sites. Here U_{LJ}(R) is the minimum LJ potential over all possible (L, A, O) values for a given R.

4. For each preferred R, preferred ligand conformers L are selected from low LJ potentials U_{LJ}(R, L), the minimum LJ potential over all possible (A, O) values for a given set (R, L).

5. Similarly, for each preferred R and L, preferred ligand atoms A are selected from the low LJ potentials U_{LJ}(R, L, A), the minimum LJ potential over all possible O values for a given set (R, L, A).

After the above procedure, a preferred (R, L, A, O) ensemble with all the possible orientations (O) of the ligand is generated and subsequently ranked by the scoring function. To speed up the LJ potential calculation, RLDOCK employs a grid-based energy calculation, where each grid stores the LJ energy between RNA and a probe atom on the grid for fast computation of LJ energy for a given binding mode. Through the above procedure, RLDOCK generates millions of possible ligand configurations through exhaustive rotation and translation transformations at each putative binding site for each preconfigured conformer (see Fig. 5.5). Compared with other models, RLDOCK has the unique merit of using complete sampling for ligand conformers and binding modes.

Incremental construction

By anchoring rigid fragments through geometric matching and then incrementally building the ligand structure, on-the-fly ligand conformer sampling allows the local environment of the binding pocket to guide the growth of the small molecule. An inherent drawback of this approach is that small errors in the early steps can be amplified throughout the process, especially for large ligands. DOCK 6 [112] adopts this incremental construction strategy for ligand conformational sampling and search [105]. Unlike the original greedy algorithm (cluster-based pruning) to sieve the sampled ligand structures followed by clustering and ranking at each step, an improved algorithm, which skips the conformational clustering step in order to retain the original, diverse conformations of the flexible bonds, has led to an increase in the success rate by 10% for the prediction of binding poses [112]. The results have demonstrated the importance of maintaining the diversity of ligand conformers.

Stochastic optimization

The stochastic optimization method searches for binding modes on-the-fly by optimizing flexible torsional angles, orientation, and position of the small molecule. The Monte Carlo (MC) [114, 117] and Genetic algorithms (GA) are the most widely used stochastic optimization algorithms. A combination of different stochastic methods often lead to improved sampling and optimization results. For example, AutoDock Vina [117] adopts a hybrid approach with MC for global optimization and Broyden-Fletcher-Goldfarb-Shanno (BFGS)

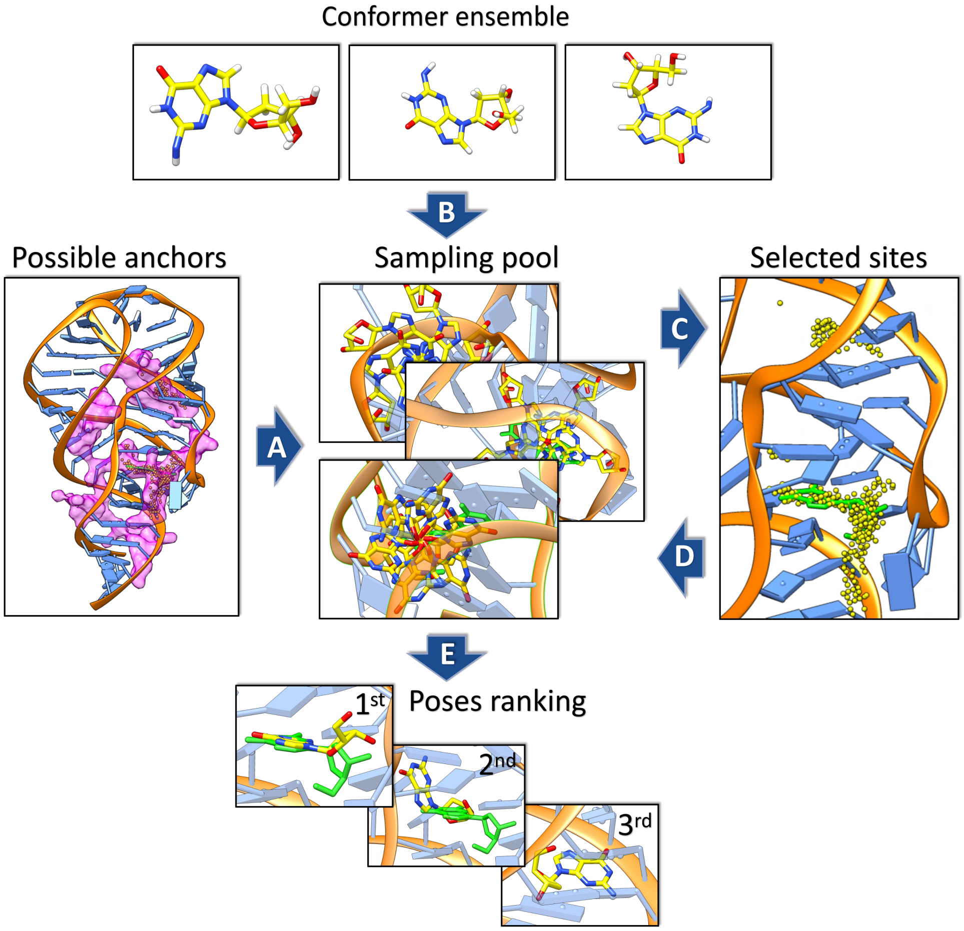


Figure 5.5: Illustration of conformational sampling methods used in RLDOCK, using the docking of 2'-deoxyguanosine to 2'-deoxyguanosine riboswitch (PDB code: 3SKL [310]) as an example. An ensemble of different conformers of the 2'-deoxyguanosine (dG) is constructed for flexible docking. The sampling and scoring procedures are shown in order and labeled through A to E. (A) First, the regions of possible anchor sites within the riboswitch, colored in magenta, are determined by the geometric features of the target RNA. (B) Second, with exhaustive sampling of these prepared conformers through translation and rotation around the anchor sites, (C) binding sites (yellow dots) are selected according to Lennard-Jones potential between RNA and ligand atoms. (D) Finally, the sampled ligand conformations associated with the selected binding sites are ranked (E) by a physics-based scoring function.

for local optimization [117]. Other examples include ICM [273] and RiboDock [115], which employ MC coupled with simulated annealing. Several protein docking programs, such as GOLD [113], AutoDock [274], and FITTED [251], use GA to sample and search for ligand conformations. Modifications to some of these methods for RNA targets have led to highly promising results. For example, through parameterizing the scoring function [311] or adding a new solvation term to the original scoring function [312], AutoDock can treat RNA-ligand interactions. Similarly, through proper optimization [110], FITTED can be used to predict RNA-ligand docking. The accuracy may be further improved once all possible RNA hydrogen bond donors and acceptors are considered, and after metal ions such as Mg^{2+} and Mn^{2+} are included as part of the receptor [110]. Like other stochastic optimization algorithms, a shortcoming of MC and GA is that the optimization process may become trapped in the local minima. This may pose a severe challenge due to the rugged energy landscape of RNA-ligand complexes. The problem can be alleviated through repetitive docking with random placement of the small molecule (ligand) and implementation of algorithms such as tabu search [313, 314] and stochastic tunneling [315] to accelerate the detrapping from the local minima. By minimizing the likelihood of poses being trapped in a local minimum in the early stages of the conformational search, rDock [118], a model for both nucleic acid and protein docking, employs a GA/MC hybrid method to enhance efficient sampling of ligand binding poses. The GA/MC hybrid method involves three rounds of GA, low-temperature MC, and Simplex minimization, each of which adopts an independent scoring function. An optimized set of scoring functions has been shown to significantly enhance the efficiency of sampling even with an unfavorable initial pose by minimizing the possibility of being trapped in a local minimum during the conformational search.

In summary, while multi-conformer docking provides a fast way to consider ligand flexibility prior to the docking calculation, its performance depends on the quality of the generated conformer ensemble. In contrast, stochastic and incremental sampling approaches can treat ligand flexibility during the docking process. However, such on-the-fly sampling approaches suffer from the problem that a small error in the early steps can be amplified in the later steps, and stochastic approaches suffer from the problem that poses can be potentially trapped in a local minima while docking. In addition, both approaches require additional energy terms to account for intra-ligand interactions. Although the conformational sampling modules in RNA-ligand docking software [109, 118–120, 126–128] have shown promising results for recovering native or near-native ligand poses, for a flexible RNA that undergoes conformational changes upon ligand binding, a search for a fast and accurate sampling method by combining folding and binding algorithms for both ligand and RNA continues.

5.3.2 Incorporation of RNA flexibility

It has been shown that for protein-ligand docking, ignoring the protein (receptor) flexibility can cause incorrectly predicted binding modes [316]. The problem can be more severe for ligand binding to an RNA, whose structure can be more flexible than a protein. To address this important issue in RNA-ligand docking, several successful approaches have been developed to incorporate RNA conformational changes in the docking algorithm. These approaches can be classified into three types: soft docking, ensemble docking, and fully flexible docking (See Fig. 5.6).

Soft docking

A mathematically convenient way for soft-docking is to decrease the energy penalties for steric clashes, thus tolerating some degrees of overlap between RNA and ligand [318].

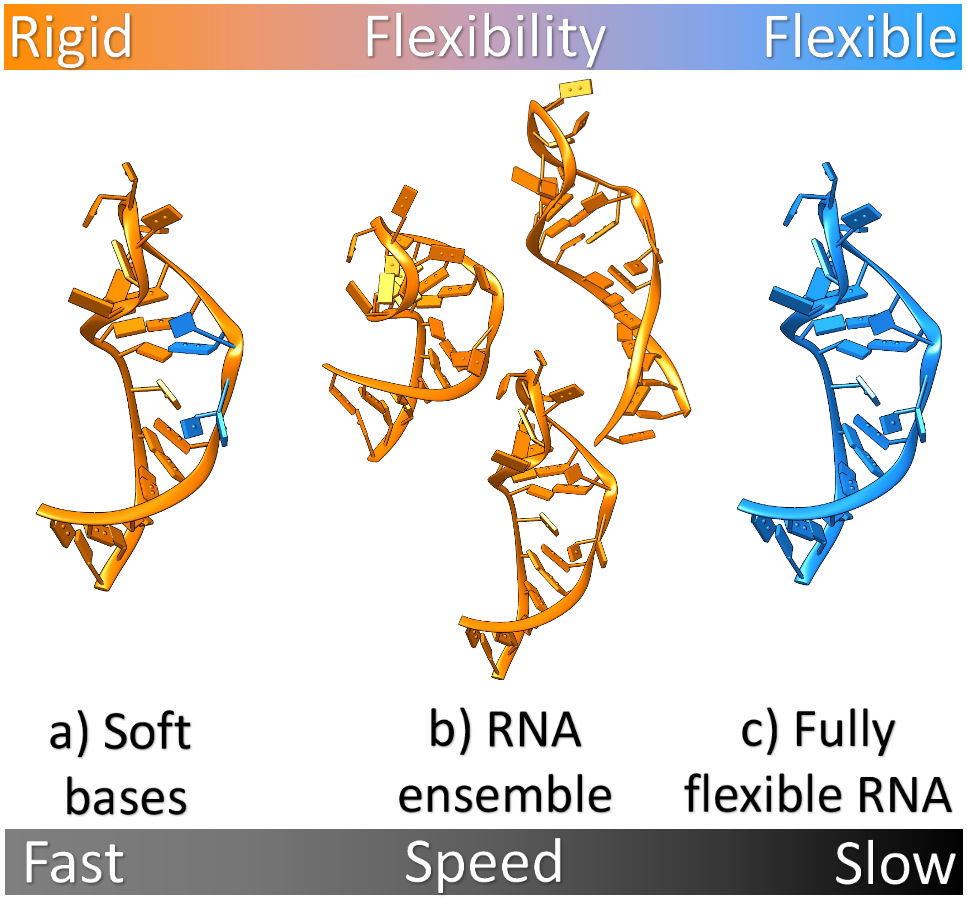


Figure 5.6: Different approaches to modeling RNA flexibility in RNA-ligand interactions illustrated using HIV-TAR RNA (PDB: 1ANR [317]) as an example. The orange and blue regions correspond to rigid and flexible portions of RNA, respectively. From left to right, a) bases from the active site are allowed to partially overlap with atoms from ligand through soft potential, b) an ensemble of various RNA conformations is used to perform docking, and c) RNA with full flexibility. Computational efficiency decreases from left to the right.

Glide [114] and GOLD [113] offer such options for users. In earlier work, Moitessier et al. [107] have employed this strategy to dock aminoglycosides in ribosomal A-site RNA, where RNA flexibility was considered using a set of soft van der Waals potentials, and the approach led to increased average accuracy. Soft docking is attractive for its convenience of implementation. However, the limited sampling space without the adjustment of backbone prevents its application to large conformational changes.

Ensemble docking

In ensemble docking, a given ligand docks into an ensemble of RNA conformations or an ensemble-averaged RNA conformation. Ensemble docking is found to be useful in several RNA-targeted studies [94, 95, 107], and has also been proved successful in protein-small molecule docking [319, 320]. In an attempt to reproduce the experimentally determined RNA-aminoglycoside complexes, soft docking to an ensemble-averaged RNA structure gives the best performance with an average RMSD of 2.49Å between the predicted and the experimentally measured binding mode [107]. Ensemble docking-based virtual screening with the ICM docking model [94] for HIV-TAR [94, 95] has predicted a TAR-targeting compound with high specificity. Furthermore, virtual screening for an experimentally derived TAR conformational ensemble against a ligand library composed of $\sim$100,000 drug-like organic compounds [95] provided an enriched family of TAR-targeting binders. From a practical perspective, the number of receptor conformations used in the ensemble docking is usually limited due to computational feasibility, thus receptor conformation selection can influence the accuracy of the prediction. Using only the conformational ensemble in the lowest-energy basin may not be the optimal strategy as ligand binding can stabilize and selectively enrich the population of conformations in other basins on the energy landscape [321–323]. Therefore, ensemble docking should not ignore low-populated ligand-free RNA conformations.

Molecular dynamics

Molecular dynamics (MD) simulation [324–331] can not only refine conformations of RNA-ligand complexes and generate ligand and RNA structures, but also shed light on the trajectory and the folding/unfolding of possible metastable states for both RNA and ligand in the docking process [95, 98, 105]. However, in practice, ligand binding events can occur in the timescale up to seconds [284, 286, 332], and an all-atom MD simulation for the process goes beyond the capacity of available computing power, especially when virtual screening for drug molecules is considered. Powered by advanced sampling techniques, several computational methods have enabled the characterization of RNA conformational changes upon ligand binding [333–337]. A non-equilibrium MD simulation [338] and an umbrella-sampling-based MD simulation [339] both have revealed the competitive relationship between the formation of the kissing-loop and the binding of the small molecule. A recent explicit-solvent MD simulation has shown a small molecule-induced stabilization effect in an adenine riboswitch and the ability of the riboswitch in the near-native states to attract small molecules through hydrogen bonding and base-stacking interaction [340]. These results demonstrated the unique advantage of MD simulations for the investigation of physical mechanisms in RNA-ligand binding.

Fully flexible RNA

Molecular dynamics simulation with proper force field can provide reliable sampling of RNA-ligand complex conformations. For example, by applying an RMSD penalty term to the conventional potential energy, MORDOR [111] (MOlecular Recognition with a Driven dynamics OptimizeR), by simulating ligand docking trajectories, can give conformational sampling and show ligand-induced conformational changes for RNA [111]. As an application, a MORDOR-based virtual screening has found a small family of binders targeting human telomerase RNAs (hTR) [341]. However, further applications of MORDOR are limited by the high computational cost, which can take up to hours for a docking run. To accelerate simulation, Supervised Molecular Dynamics (SuMD) has been proposed to sample the conformations of RNA-drug complexes [342]. SuMD accelerates the simulation by applying a tabu-like algorithm to guide the docking when the ligand is far away from the binding site and a conventional MD simulation when the ligand is close to the binding site. The hybrid method enables efficient simulation of the binding process within an affordable timescale. Although the simulated trajectory does not necessarily represent the physical binding process, SuMD may capture possible conformational changes. The reliability of SuMD for RNA-ligand docking is supported by success in predicting binding modes for several pharmaceutically important RNAs [342], where SuMD predicts RNA-ligand docking mode with a minimum RMSD of 0.34Å for the best case.

Similarly, another method based on elastic potential grids was initially proposed for modeling protein flexibility during docking [343] and later extended to RNA targets [121]. In this type of method, a 3D grid of the potential field of the initial RNA conformation is calculated in advance using DrugScore$^{\text{RNA}}$ [344]. After determining the potential grids, AutoDock [274] is used as a docking engine with precalculated elastic potential grids for docking. Due to its ability to account for RNA flexibility, docking to the deformable potential grids generated from unbound RNA has a much better performance than simply docking to unbound RNA alone. However, one of the limitations of the approach is that it requires *a priori* knowledge of the available end states of deformation. Moreover, the model cannot treat conformational changes caused by rotational flip motion and 2D structural rearrangements.

In summary, Molecular dynamics-based methods are time-consuming and thus not suitable for large-scale virtual screening. Rigid docking is fast but lacks accuracy. Soft docking and ensemble docking are in the middle ground between the two extremes as they sacrifice the ability of a more complete sampling of conformations in order to reduce the computational time. At the current stage, a versatile approach to accurately treat receptor flexibility awaits to be developed.

5.4 Accurate scoring functions for RNA-ligand docking: challenges and promises

Selecting a native ligand binding pose from an ensemble of candidates requires a reliable scoring function. There are three different approaches to the development of a scoring function: physics-based approach, knowledge-based approach, and machine-learning approach; see Table 5.2 and Table 5.3 for a list of reviewed scoring functions and the summary of the benchmark results, respectively.

5.4.1 Physics-based methods: physical principles of RNA-ligand binding lead to accurate scoring functions

Force-field approach

Atom-based physical force fields, originally derived from thermodynamic data and *ab initio* calculations, have enabled molecular dynamics simulations for nucleic acids-targeted drug discovery [346–350]. One of the key issues in physical force field-based computations for molecular docking is the solvent effect. Since the virtual screening of ligands against

Table 5.2: Summary of the reviewed scoring functions used in different models for predicting small molecule binding. ^{a} Some scoring functions optimized for protein may also be used for RNA. ^{b} Some models contain more than one scoring function, only the default one is listed. ^{c} The year that the model was first published, although some software is still under active development.

Category	Model	Target[a]	Score type[b]	Year[c]
Physics-based	MORDOR [111]	RNA	force fields	2008
	DOCK 6 [112]	RNA	force fields	2009
	GOLD [113]	protein	empirical terms	1997
	Glide [114]	protein	empirical terms	2004
	RiboDock [115]	RNA	empirical terms	2004
	AutoDock 4 [116]	protein	empirical terms	2007
	AutoDock Vina [117]	protein	empirical terms	2010
	iMDLScore1 [109] iMDLScore2 [109]	RNA	empirical terms	2012
	rDock [118]	protein nucleic acid	empirical terms	2014
	RLDOCK [119, 120]	RNA	empirical terms	2020
Knowledge-based	DrugScore$^{\text{RNA}}$ [121, 122]	RNA	statistical potentials	2000
	KScore [123]	protein nucleic acid	statistical potentials	2008
	LigandRNA [124]	RNA	statistical potentials	2013
	SPA-LN [125]	nucleic acid	iterative statistical potentials	2017
	ITScore-NL [126]	nucleic acid	iterative statistical potentials	2020
Machine-learning	T-Bind [345]	protein	gradient boosting trees	2018
	RNAPosers [128]	RNA	random forest	2020
	RNAmigos [129]	RNA	graph neural network	2020

an RNA target demands high computational efficiency for the docking calculation, implicit solvent models, such as Poisson-Boltzmann surface area (PB/SA) model [131–137] and the Generalized-Born surface area (GB/SA) model [138–146], would be highly promising due to the optimal balance between speed and accuracy. A hybrid force field that combines an implicit solvent model and an all-atom force field can often lead to accurate and efficient simulation of an RNA-ligand binding process. As shown by the success of DOCK 6 [112], generalized Born and Poisson Boltzmann implicit solvent models combined with the AMBER force fields can provide an effective energy model for an RNA-ligand docking system [112]. MORDOR [111], which combines an implicit solvent model GBSW (Generalized Born with Simple sWitching) [351] with the CHARMM-27 [352] force fields for the receptor and AMBER force fields [353] for ligand molecules, demonstrated how the hybrid energy function can lead to successful modeling of receptor-ligand binding. By using root-mean-square-distance constraints in energy minimization, the model allows local flexibility of the receptor to accommodate possible conformational changes induced by ligand binding and in the meantime, to guide the ligand to probe the surface of the target RNA.

In summary, physical force field-based scoring functions have the advantage of providing insights into the underlying physical mechanism of RNA-small molecule interaction, however, computational costs and the need for expert knowledge in simulating a specific system hinders the application of these models in large-scale virtual screening for drug discovery.

Empirical energy approach

Physically, different interactions in an RNA-ligand complex are correlated thus nonadditive. A simplified approach is to evaluate the total energy as a weighted sum of the component interactions such as van der Waals, electrostatic, desolvation and hydrogen-bond interactions:

$$\Delta G = \sum_i w_i \cdot \Delta G_i \tag{5.1}$$

where the weight coefficients w_i can be fitted by optimizing the success rate of the computational prediction for the training set. The above empirical scoring function has the advantage of high computational efficiency and adaptivity, which makes accurate prediction possible for specific types of RNA targets and ligands. Compared to the more rigorous force-field approach, empirical scoring functions, which often use "softer" energy forms, are more tolerant for minor clashes and suboptimal interactions during docking, thus partially alleviate the problem of incomplete sampling for receptors and ligand conformations. It is important to note that due to the different physical interactions and correlations between the different interactions in protein-ligand and RNA-ligand systems, parameters fitted from protein-ligand docking may not be transferable to RNA-ligand docking.

The semiempirical free energy function in AutoDock 4 [116], the fully empirical scoring function of AutoDock Vina [117], and other models such as GoldScore in GOLD [113] and GlideScore in Glide [114] have demonstrated success in predicting protein-ligand docking. These docking software packages, not specifically designed for nucleic acids, may not give optimal results for RNA-ligand docking. RNA-specific scoring functions such as iMDLScore1 and iMDLScore2 [109] have optimized the weight coefficients of the scoring terms [116] using multilinear regression (MLR) methods. In a comprehensive evaluation and comparison for eleven other scoring functions, iMDLScore1 and iMDLScore2 [109] have shown better performance in both binding mode and affinity predictions.

Several RNA-ligand docking software packages have incorporated their respective built-in empirical scoring functions [115, 118–120] specifically for RNA/DNA-small molecule interactions. The scoring function of rDock [118], the successor of the original model RiboDock [115], contains a weighted sum of various intermolecular and intramolecular interaction energies, including the van der Waals potential (vdW), an empirical energy term for attractive and repulsive polar interactions and the desolvation energy. Because virtual drug screening can benefit from our knowledge, such as pharmacophoric points and shape similarity, derived from known RNA-ligand complexes, rDock has added pseudo-energy restraint terms as an empirical bias, such as pharmacophoric restraints. Pharmacophoric restraints used in rDock ensures the generated ligand poses to satisfy the pharmacophores derived from the known RNA-ligand complexes or the hot-spot mapping methods. Applications to the virtual screening against Hsp90, both rDock and Glide show significant improvements with the inclusion of the bias [118]. Since rDock has been optimized for RNA docking, it outperforms Vina and Glide for RNA-ligand docking: For a set of 56 RNA-ligand complexes, the top-ranked poses predicted by rDock show a 54 ± 3 % success rate with a 2.5Å RMSD cut-off compared to 29± 2% for Vina and around 17.8% for Glide [118].

RLDOCK [119, 120] trained the weight coefficients for the different interaction terms such as van der Waals (vdW), electrostatic, polar and nonpolar hydration energies, and hydrogen-bond interactions for a set of 30 RNA-small molecule complexes. To enhance computational efficiency. RLDOCK adopts a two-step screening algorithm: In the first step, using a computationally efficient, crude estimation for the Born radii in the electrostatic energy calculation and the solvent-accessible surface area in the hydration energy, the model selects an initial pool of potential binding poses; in the second step, a rigorous scoring function is used to re-rank the binding poses to identify the top-ranked poses. Test on a separate set of 200 RNA-small molecule complexes indicates that the success rate of identifying the native/near-native binding modes increases significantly from the crude scoring function to the more refined scoring function, with 8.3%, 22.2%, and 29.6% for the crude scoring function and 17%, 40.4%, and 49.1% for the more rigorous scoring function within RMSD thresholds 1Å, 2Å, and 3Å, respectively. Considering the fluctuations in the ligand pose, RLDOCK groups similar ligand poses (according to the mutual RMSD) into clusters and rank the clustered poses. With the RLDOCK-ranked ligand poses (clusters), 44.3%, 74.3%, and 82.2% of the top-10 are within 1, 2, and 3Å (RMSD), respectively, to the native pose [119, 120], Furthermore, tested on a previously proposed set of 38 RNA-small molecule complexes [124], RLDOCK has demonstrated a higher success rate compared to other models for recovering the native ligand binding poses within RMSD of 2Å to the native pose. Specifically, RLDOCK shows a success rate of 55.3% (60.5%) for the top-1 (top-3) predicted poses as compared to 28.9% (39.5%), 36.8% (44.7%), 39.5% (47.4%), and 50.0% (57.9%) for DrugScore$^{\text{RNA}}$ [121, 344], DOCK 6 [112], LigandRNA [124, 354], and a combined LigandRNA [124, 354] and DOCK 6 [112] approach, respectively. Since RLDOCK distinguishes itself from other models by a global, complete sampling of all the possible binding sites and poses, the results above demonstrate the importance of high-quality sampling for ligand poses.

In summary, compared to atomistic force field-based approaches, the empirical energy function methods manage to reduce the computational burden using simple functional forms for RNA-small molecule interactions. However, this approach is subject to two main limitations: (a) neglecting the correlation between different interactions and (b) transferability of the weight coefficients between different RNA-ligand systems. The success of the model depends on the quality of the curated training set, thus the accuracy of the predictions is limited by the lack of available high-quality data for RNA-ligand complexes.

5.4.2 Knowledge-based scoring functions: statistical potential provides efficient scoring of binding modes

Statistical potential approach based on reference states

A statistical potential approach uses the inverse Boltzmann law to extract energy-like potential for user-defined interacting pairs from the experimental data:

$$E \propto -\sum_{i \in R} \sum_{j \in L} \ln \rho_{ij} \tag{5.2}$$

where and ρ_{ij} is the relative frequency of the user-defined interacting pair (i, j) between the receptor R and ligand L. Before being applied to RNA-ligand docking models [121, 123–126, 344], the statistical potential approach has been demonstrated to be effective for predicting protein-ligand docking [122, 159, 160, 162, 355–367]. Different variants of the statistical potential approach have been proposed for RNA-ligand systems since the development of DrugScore$^{\text{RNA}}$ [121, 344]. These statistical potential approaches mainly differ in two aspects: the choice of reference state and functional forms of potential energy terms. As early attempts, DrugScore$^{\text{RNA}}$ [121, 344], Kscore [123], and LigandRNA [124] have constructed the reference state by treating all the relevant atoms in the RNA-ligand complex as non-interacting particles, with different atom types differentiated or undifferentiated. In addition to the distance-dependent pairwise potential used in Kscore and DrugScore$^{\text{RNA}}$, LigandRNA, by taking into consideration the relative orientations between different atom pairs, has added a three-body anisotropic potential. Combined with DOCK 6 [112], this orientation-dependent potential shows a higher success rate than DrugScore$^{\text{RNA}}$ in predicting the native binding modes. Specifically, for a test set consisting of 42 RNA-small molecule complexes, with the 2Å RMSD criteria for a correctly predicted ligand pose, DrugScore$^{\mathrm{RNA}}$, DOCK 6, and LigandRNA show a success rate of 31.0%, 35.7% and 35.7%, respectively. A DOCK6 and LigandRNA hybrid score scheme further gives the success rate of 47.6%. The results show that the knowledge-based approach can benefit from a more accurate potential that accounts for more detailed information such as distance and angular correlations between the different interactions.

Iterative statistical potential approach

A major limitation of the above traditional approach is that the reference state ignores the many-body correlations between the different interactions. One way to circumvent this problem is to iteratively refine the energy function until the simulated probability distribution of the different atom pairs agrees with that observed from the experimental data [125, 126, 212, 231, 368–371]. Because the simulated distribution is based on sampling over the full energy landscape, an iterative approach can account for both native and nonnative interactions.

SPA-LN [125] is an iterative statistical potential model for predicting nucleic acid-small molecule interactions. Using intrinsic specificity ratio (ISR), a measure of the native vs. nonnative binding modes discriminative power, the energy-like scoring function can account for both affinity and specificity. For binding affinity prediction, using Pearson correlation coefficient between the predicted and the experimentally measured affinity as a measure, for a set of 77 complexes from version 2014 of PDBbind database [372] and a separate set of 34 nucleic acid-small molecule complexes, SPA-LN gives Pearson correlation coefficients 0.58 and 0.60, respectively. For the binding pose prediction, for a test set of 56 nucleic acid-small molecule complexes [118], for the top-scored poses with 2.5Å RMSD cutoff for a pose considered to be native or near-native, SPA-LN [125], rDock [118], AutoDock Vina [117] and Glide [114] give a success rate of 54(±3)%, 54(±3)%, 29(±2)%, and 17.8%, respectively. The performance of SPA-LN suggests the importance of considering not only affinity but also the specificity of RNA-ligand binding.

To capture the stacking and electrostatic interactions in nucleic acids, ITScore-NL [126], an iterative statistical potential approach [370], adds an extra distance-dependent stacking potential term and an electrostatic potential energy term to the scoring function. Stacking potentials were calculated for all carbon-carbon atom pairs involved in stacking interactions between nucleobases and planar aromatic groups of a small molecule. Electrostatic potential was calculated for the polar atom pairs using the Debye-Hückel approximation. The model was compared with other methods in two datasets to validate the performance on native pose recovery and binding affinity prediction. With the same set of 77 nucleic acid-small molecule complexes used in the test of SPA-LN, ITScore-NL achieved a higher Pearson correlation coefficient (R = 0.64) than that shown in SPA-LN (R = 0.58). As for the success rate of native pose recovery, ITScore-NL [126] was able to correctly identify native binding mode of 71.43% (50.64% for SPA-LN [125]) complexes with RMSD cutoff 1.5Å if the top-3 predictions are selected and 90.90% (76.62% for SPA-LN [125]) complexes with RMSD cutoff 3.0Å if the top-5 predictions are selected. Compared to LigandRNA [124] on a 42 RNA-small molecule complexes with only top-scored poses being selected and poses with RMSD cutoff 2.0Å being used, the success rates of LigandRNA [124] and ITScore-NL [126] are 35.7% and 50.0%, respectively. The results indicate the importance of including stacking and electrostatic interactions in RNA-ligand docking.

In summary, compared to atomistic force field methods, the statistical potential approach is associated with a much higher computational efficiency. However, the choice of reference state places obstacles for accurate modeling of RNA-small molecule interactions. Even though iterative approaches have been developed to circumvent the reference state problem, constructing diverse and complete decoy sets for training remains a challenge for RNA-ligand complexes. Furthermore, because data-driven approaches rely on experimentally determined structural data, the success of the models suffers from limited structure data for RNAs and RNA-ligand complexes.

5.4.3 Machine-learning based scoring method for RNA-ligand docking: an emerging scoring approach with high promise

Machine-learning approach

With the success of machine-learning methods in various fields, a variety of machine-learning models such as support vector machine (SVM), random forest (RF), neural network (NN), and convolutional neural network (CNN) have been proposed and shown success to predict protein-small molecule interactions [147–152]. Machine learning approaches not only have the advantage of utilizing the experimental data and making fast predictions but also with a large number of trainable parameters, can leverage experimental data better than traditional knowledge-based methods. Furthermore, the relation between input features and output results is learned through the training process. Therefore, a machine-learning method can be readily adopted across different types of tasks by simply changing the input features and output format, which can be engineered for the corresponding task. Fig. 5.7 shows a typical workflow of training, validating, and testing a machine-learning model.

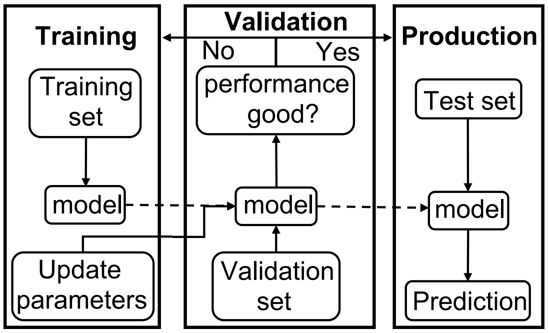


Figure 5.7: The typical workflow of a machine-learning approach. Training and validation cycle usually needs to be performed many times before the performance on the validation set reaches an acceptable level. After the training-validation cycle, the trained model is used to make predictions on the test set.

Importance of feature engineering

Although the quality and amount of the training data is vital to the performance of machine-learning methods, input feature engineering, which is often overlooked, is also critical to the success of the model. Generally, input features extracted from structure or geometry-based models for RNA-ligand binding often contain a large amount of detailed structural information, resulting in noise and excessively high dimensions in the parameter space. For example, there are many CNN-based approaches [181, 373–376] that simply extend the 2D image in the original CNN model by treating the binding site as a 3D image. However, this type of 3D image is not rotational invariant hence requires rotational augmentation when used in training and prediction. The extra dimensions added would significantly increase computer time for making a single prediction and for performing large-scale virtual screening for drug discovery. Additionally, many atoms which are shown as pixels in the image may not even contribute to the binding, thus further complicate the learning process. An optimal engineered feature should maximally simplify the input information while capturing the key features that determine RNA-ligand docking results.

T-Bind [345] is a method for protein-small molecule binding affinity prediction. What makes it interesting for its possible application to RNA-ligand binding is not only the machine-learning model but also, more importantly, its feature extraction method. In T-Bind [345], Cang et al. [345] introduces a novel mathematical concept, element specific persistent homology (ESPH) or multicomponent persistent homology, to capture the crucial topological information around the binding site. This feature extraction method offers a new way to embed geometric information into topological invariants and simplify the input features while still capturing the key information. Benchmark tests using PDBbind database [372, 377–381], a comprehensive collection of protein-small molecule binding affinity data together with 3D structures, have yielded Pearson correlation coefficients 0.818 and 0.767 for PDBbind v2007 core set [379] and PDBbind v2013 core set [380], respectively, and the T-Bind outperforms other scoring functions [345]. The result shows the merit of this feature extraction method and the promise of applying the same approach to the prediction of RNA-small molecule interactions.

RNAmigos [129] is a machine-learning model designed for the prediction of RNA-small molecule interactions. In RNAmigos [129], the base-pairing network around the binding site is simplified as a connected 2D graph with vertices and edges, where a nucleotide is represented by a vertex and backbone connectivity and base-pairing are represented by the different types of edges. The base-pairing interactions encoded in the 2D graph provide a signature for predicting the fingerprint of the small molecule that will most likely bind to the site. Furthermore, RNAmigos [129] shows the versatility of the machine-learning method. The model combines RNA base-pairing information at the binding site (in a 2D graph format) and graph neural network [382] (designed for data with 2D graph structure) to directly predict the fingerprint for the small molecule. RNAmigos [129] encodes the predicted fingerprint as a 166 bit MDL Molecular Access keys (MACCS) [383]. The small molecule can be found in a compound library simply by a similarity search against the predicted fingerprint. This procedure circumvents the traditional virtual screening route (docking then scoring) and substantially reduces the time for the search of a putative drug. RNAmigos [129] was trained with a set of 773 RNA-small molecule binding sites associated to 270 unique small molecules. Test results on an enrichment dataset with 176 unique RNA chains against 82 unique ligands have shown that RNAmigos outperforms a template-based method (Inforna 2.0 [255, 256]). Although RNAmigos shows promising results, it has two major limitations. First, RNAmigos requires prior knowledge of the binding site in order to generate a base-pairing network, and a misplaced binding site could led to a degradation in predictive power. However, accurate determination of the binding site itself can be a challenging task. Second, RNAmigos considers different RNA-small molecule complexes to occur equally likely. Such treatment can cause an effective bias in the training process because the binding affinities of the different RNA-small molecule complexes can span across a large range of values.

As another machine-learning model for RNA-small molecule binding, RNAPosers [128] contains a set of trained pose classifiers that can estimate the "nativeness" of a ligand for a given structure of the RNA and the ligand. The classifiers are based on the random forest method [384] with an ensemble of 1000 decision trees. For a given ligand, RNAPosers takes a pose fingerprint as its input, where the pose fingerprint is calculated as the sum of a Gaussian function multiplied by a cosine damping factor over the RNA-ligand interacting atom pairs. The comparison between RNAPosers and two knowledge-based methods, DrugScore$^{\text{RNA}}$ [121, 122] and SPA-LN [125], shows that RNAPosers [128] is able to yield higher success rates for the prediction of the native binding pose. Specifically, for a set of 31 RNA-small molecule complexes used in DrugScore$^{\mathrm{RNA}}$ [121, 122], RNAPosers gives a success rate of 61.9% as compared to 57.1% for DrugScore$^{\mathrm{RNA}}$. For another set of 56 RNA-small molecule complexes used as a validation set in SPA-LN [125], RNAPosers gives a success rate of 62.5% compared to 54.0% for SPA-LN. These results show the advantage of the machine learning method over traditional knowledge-based approaches.

Coarse-grained conformational representation, traditionally implemented in RNA folding models, can lead to a unique method for feature engineering. Recently Stefaniak and Bujnicki developed a new machine-learning model, AnnapuRNA [127], specifically for RNA-ligand interactions. The feature engineering algorithm in AnnapuRNA employs a coarse-grained representation of both RNA and ligand to derive RNA-ligand contact statistics. Specifically, RNA structure is coarse-grained with each nucleotide replaced with five beads [385], and a ligand is represented with the concept of pharmacophores [386]. Contact statistics collected from the coarse-grained representation with the assumption that the coarse-grained contact statistics can represent the core RNA-ligand interaction data. Five different machine-learning algorithms (Random Forest-RF, k Nearest Neighbors-kNN, Gaussian Naïve Bayes-GNB, Support Vector Machines with RBF kernel-SVM, and Deep-Learning-multi-layer feedforward artificial neural network-DL) have been trained on the coarse-grained statistics and each RNA-ligand complex is evaluated using a scoring function for the nativeness probability of the contacts and the ligand internal energy [353]. Benchmark test with a set of 33 RNA-ligand complexes has shown that kNN and DL algorithms give the best results and more extensive tests with 4 docking methods and 9 scoring functions have demonstrated that AnnapuRNA outperformed other programs tested. The results has indicated that the coarse-grained representation combined with the concept of pharmacophores can indeed provide an effective, simplified way of feature engineering for RNA-ligand binding.

In summary, although machine-learning models for protein folding [153–158] and protein-small molecule interactions [159–162] have shown significant success, modeling RNA-small molecule interactions using machine learning is a relatively new adventure. The machine learning approaches have several intrinsic limitations. Because the model involves a large number of trainable parameters, the training process is prone to overfitting, especially for cases with only limited training data. The problem is more important for RNA-small molecule binding due to the lack of a comprehensive and high-quality curated database. Although protein-focused libraries, such as PDB [208], PDBbind [372, 377–381], can contain data for RNA-small molecule complexes, a more comprehensive, dedicated NDB-like database [225, 233] for RNA-ligand complexes is needed.

5.4.4 Accounting for solvent-mediated interactions

The sugar-phosphate backbone is negatively charged and polar, resulting in an accumulation of water molecules, cations, and water/ion-mediated RNA-ligand interactions. However, most molecular docking models do not explicitly consider the bridging effects of water molecules (Fig. 5.3b) and metal ions (Fig. 5.3c). The neglect of such solvent effects is a notable drawback that can cause inaccurate predictions for RNA-ligand interactions. A viable approach is to use simulations with explicit waters and/or ions to refine RNA structures [105, 107]. Then, the positions of important water molecules can be retained for further docking against ligands. In this approach, the results are sensitive to the selection of the important water molecules. Because ligand-RNA interactions are sensitive to the positions and orientations of the water molecules and ions within the cavity space, achieving robust accuracy can be challenging with this approach.

An alternative approach is to predict the binding of water molecules and bound metal ions to the RNA prior to docking [240, 241, 387–392] then treat the predicted bound water molecules and/or ions as part of the receptor for RNA-small molecule docking. The Tightly Binding Ion (TBI) [240, 393, 394] model and the Monte Carlo TBI (MCTBI) model [241, 395] predict the ion distribution around an RNA structure. Through explicit sampling of the discrete ion distributions, TBI and MCTBI go beyond the mean-field Poisson-Boltzmann theory by accounting for the correlation between the different ions. 3D-RISM is another promising model for predicting the distribution of both solvent and ions around a given macromolecule [391]. In combination with a force field calibrated by prototype ionophores, 3D-RISM is able to recapitulate the water distribution around guanine quadruplexes by considering the correlation between solvent particles and treating the system as macromolecules in equilibrium with a bulk solvent at constant composition (and chemical potential) [396]. For the Oxytricha nova telomeric G-quadruplex structure as a test case, 35% (80%) of the 3D-RISM-predicted water binding modes are within 1Å (2Å) from the crystallographic modes, indicating that the model may be reliable for treating solvent effects [396]. SPLASH'EM (Solvation Potential Laid around Statistical Hydration on Entire Macromolecules) is a model for predicting bridging water molecules in nucleic acid-ligand complexes. Using the statistical information of water molecules around nucleotides in the PDB [208] and a scoring function containing a hydrogen-bonding potential with both directionality and polarization, SPLASH'EM has identified 62% of water molecules in nucleic acid-ligand complexes within 1Å [397].

5.5 Current achievements of RNA-ligand docking models: a performance comparison

Table 5.3 summarizes the benchmark test results of various computational methods for RNA-ligand docking. The data is adopted from the original publications of the models and the results are grouped according to the test sets. Cautions should be taken when comparing the performance of the different models. First of all, RNAs and ligands in different test systems can have different structural and physical features, thus a direct performance comparison of the different models based on the results from different test systems may not be appropriate. Second, even for methods evaluated based on the same test dataset, the interpretation of the performance comparison can be complicated. For example, for the same benchmark dataset, the decoy poses generated for pose identification can be quite different for the different tests. A robust and objective comparison between the different scoring functions demands a consistent and systematic benchmark test protocol for the generation of the decoy binding poses [109].

Nevertheless, the benchmark test results in Table 5.3 provides useful insights for the selection of computational methods. First, for pose identification, we find a clear trend that RNA-specific methods consistently outperform those developed for proteins or generic macromolecules. The result shows the importance of considering RNA-specific interactions and structural features for RNA-ligand docking. Second, recent advances in RNA-ligand scoring function have been mainly focused on knowledge-based/machine learning-based approaches [125–128]. The knowledge-based/machine learning-based approaches provide equal or better performance (especially for affinity prediction) than the traditional physics-based/empirical approaches [112, 113, 116–118], except for MORDOR [111] and RLDOCK [119, 120]. Third, even with the knowledge-based/machine learning-based approaches, the current success rate for affinity prediction is quite low. To improve the prediction accuracy, with the currently limited available data for RNA-ligand binding affinities and complex structures, new physics-based models that can accurately capture RNA-ligand interactions and conformational ensembles would be highly needed.

Table 5.3: Summary of the benchmark results in the literature. Results of different methods tested on the same test set are grouped together for comparison. The first column "Test set" shows the number of test cases and the original references reporting the test results. [a] Performance of affinity prediction is reported in terms of the square of the Pearson correlation coefficient, R^2. Correlation coefficient is calculated between experimental binding affinities and predicted binding affinities. Benchmarks without affinity prediction are shown as dashes. [b] Performance of pose identification is reported in success rate. The criteria for a correct prediction is shown in the (rank, RMSD) format. For example, (1, 2.5Å) means the top-1 prediction that has RMSD <2.5 Å to the native pose. Benchmarks without pose identification are shown as dashes. [c] RLDOCK, rDock, rDock_solv, AutoDock Vina use 38 instead of 42 complexes. MORDOR uses 32 instead of 42 complexes. [d] Only several top performing models evaluated in literature [109] are listed for this benchmark dataset. [e] Three outliers 3GX3, 2ESI and 1F1T are excluded in the binding affinity calculation. [f] Near native poses are sampled through rDock reference ligand method [118]. [g] Average and standard deviation from 100 sets of 100 random docking poses out of a pool of 1000 decoy conformations. [h] Native pose is included in pose identification. [i] RNA-adapted AutoDock scoring function [312]. [j] Scoring function is used to guide the docking instead of using default Vina scoring function.

Test set	Scoring function	Docking engine	Affinity[a] prediction(R^2)	Pose[b] identification(%)	
42 [119, 120, 124, 126] complexes			Correlation	(1, 2.0Å)	(3, 2.0Å)
	RLDOCK	RLDOCK	-	55.3[c]	60.5[c]
	ITScore-NL	DOCK6	-	50.0	54.7
	LigandRNA+DOCK6	DOCK6	-	47.6	54.8
	rDock_solv	rDock 2014	-	39.5[c]	55.3[c]
	DOCK6	DOCK6	-	35.7	45.2

	LigandRNA	DOCK6	-	35.7	42.9
	AutoDock Vina	AutoDock Vina	-	31.6[c]	44.7[c]
	DrugScore$^{\text{RNA}}$	DOCK6	-	31.0	42.9
	rDock	rDock 2014	-	28.9[c]	47.4[c]
	MORDOR	MORDOR	-	-	62.5[c]
34 [109, 125, 126] complexes[d]			Correlation[e]	(3, 1.5Å)	(5, 3.0Å)
	SPA-LN	rDock 2014	0.36	50.6	76.6
	Gold Fitness	GOLD5.0.1	0.25	42.9	73.2
	ASP	GOLD5.0.1	0.29	42.9	66.1
	rDock_solv	rDock 2006.2	0.18	41.1	73.2
	rDock	rDock 2006.2	0.15	33.9	60.7
	ITScore-NL	-	0.41	-	-
56 [118, 125, 128] complexes			Correlation	-	(1, 2.5Å)
	RNAPosers	rDock 2014	-	-	62.5[f]
	rDock_solv	rDock 2014	-	-	54±3[g]
	SPA-LN	rDock 2014	-	-	54±3[g,h]
	AutoDock Vina	AutoDock Vina	-	-	29±2[g]
	GlideScore	Glide (v.57111)	-	-	17.8
31 [122, 128, 312] complexes			Correlation	(1, 2.0Å)	(1, 2.5Å)
	RNAPosers	rDock 2014	-	57.1[f]	61.9[f]
	DrugScore$^{\text{RNA}}$	AutoDock 3.0.5[j]	-	41.9	45.2
	AutoDock[i]	AutoDock 3.0.5[j]	-	25.8	35.5
77 [125, 126, 377] complexes			Correlation	(3, 1.5Å)	(5, 3.0Å)
	ITScore-NL	AutoDock 4.2	0.41	71.4[h]	90.9[h]

	SPA-LN	rDock 2014	0.33	50.6[h]	76.6[h]

5.6 Nondocking based methods for modeling receptor-ligand binding

In addition to the scoring functions discussed above, other physics-based methods can also achieve high accuracy for determining the binding modes and affinities. Some of these methods, however, are not suitable for docking software due to either the expensive computational cost or the technical difficulty of incorporation of the method into a software. Because extensive reviews have been reported for free-energy methods [398], such as Molecular Mechanics/Poisson Boltzmann Surface Area (MM/PBSA) and Molecular Mechanics/Generalized Born Surface Area (MM/GBSA) [399–406], Linear Interaction Energy method (LIE) [407–409] Free-Energy Perturbation (FEP) [409, 410] and Thermodynamic Integration (TI) [404, 411], we here focus on several more recently developed physics-based methods for modeling small molecule binding problems, with the purpose of applying the methods to predict RNA-ligand binding.

5.6.1 Quantum mechanical methods

Quantum mechanical approaches, which can treat polarization, charge transfer, and many-body effects, are considered to be more accurate than force-fields in molecular mechanics. Methods to model small molecule binding range from less accurate semiempirical methods such as density functional theory (DFT) to more sophisticated methods, such as second-order Møller-Plesset perturbation theory (MP2), configuration interaction (CT) and strict coupled-cluster calculations (CC) [412]. Improvements in computer hardware have substantially reduced the computational time for quantum mechanical calculations. However, the cost is still relatively high, and systems under investigation usually need to be significantly simplified or divided into smaller fragments. Although fragmentation calculation enables the use of quantum mechanics-based methods for computing the energies of large biomolecules, such as proteins [413–415], the method has not been extensively applied to RNA-small molecule interacting systems. The difficulties may stem from long-range electrostatic interactions and many-body quantum mechanical effects in RNAs [415, 416].

Chen et al. [417] have extended molecular fractionation with conjugate caps (MFCC) scheme using quantum mechanical calculations to study DNA/RNA-small molecule interactions. Through the division of the system at each phosphate group, three oligo-nucleic acid interaction systems are decomposed into smaller subsystems, and the calculated interaction energy is found to be in excellent agreement with the results obtained from *ab initio* calculation for the original full system. In another study, Mlýnský et al. [418] have compared the abilities of various hybrid QM/MM methods, including *ab initio*, DFT and semiempirical approaches, to investigate the possible catalytic mechanism for the hairpin ribozyme. By using various hybrid QM/MM methods, Mlýnský et al. have computationally reconstructed potential and free energy surfaces for the catalysis reaction system. Among the tested methods, the activation barriers calculated from spin-component scaled Møller-Plesset (SCS-MP2) method and hybrid MPW1K functional show the best agreement with those derived from the experimental rate constant data. Recently, Bezerra et al. [252] have applied MFCC fragmentation scheme [419–421] within a density functional theory framework to calculate the interaction energy between ribosomal RNA and aminoglycoside hygromycin B. The calculation has revealed the regions where the drug molecule interacts strongly with the ribosome, and the result provides guidance for the improvement of drug-receptor affinity.

Compared to MM/MD based approaches, the quantum-mechanics approaches above are able to carry out more accurate calculations at the expense of expensive computational cost and hence the methods may not be suitable for large-scale virtual screening in drug discovery.

5.6.2 3D-RISM theory

By treating the distribution of bound ligands around the receptor as a receptor-ligand two-body correlation problem, three-dimensional reference interaction site model (3D-RISM) with Kovalenko-Hirata (KH) closure relation predicts the spatial distribution function of the ligand in the field created by the receptor [422–424]. The probable binding sites are identified from the peaks of the ligand spatial distribution function and the binding mode is determined based on the superposition approximation [253]. The 3D-RISM/KH approach has three unique advantages. First, it identifies the ligand-binding site from the distribution function for the mixture of solvent and ligand. Because the calculation circumvents the sampling and scoring in the traditional docking process, the 3D-RISM/KH approach avoids the limitations of sampling methods. Second, the theory explicitly accounts for the solvent effects. Therefore, the theory can treat ion and water-mediated interactions in receptor-ligand binding. Third, unlike the traditional continuum models, 3D-RISM/KH calculation can take the hydrogen bond between the solute and solvents into consideration. Several studies [425–430] have employed 3D-RISM/KH theory to predict the binding sites and binding modes of small molecules in proteins. In a recent study, Sugita and coworkers [253] have tested their 3D-RISM/KH approach on 18 different types of proteins and successfully predicted the native binding modes for half of the systems.

However, the 3D-RISM/KH theory is not without limitations. First, there is an upper limit of the query small molecules for a computationally feasible calculation. In practice, a large ligand needs to be divided into fragments which are later re-connected based on the calculated distributions. Second, 3D-RISM/KH takes longer than traditional docking calculation to evaluate tens of thousands of ligands as drug candidates for a given target. Third, the theory is based on correlation functions and may not be able to account for certain discrete configurations and interactions that are important for an accurate prediction.

5.6.3 Kinetic effects

For many drug molecules, binding equilibrium might not even be reached or maintained in the *in vivo* system. As a result, the thermodynamic equilibrium binding affinity may not be a proper indicator for the *in vivo* efficacy of the drug. In contrast, residence time, or the lifetime of the binary drug-target complex, measured by the inverse of the unbinding constant k_{off}, may be a better indicator for drug efficacy *in vivo* [431–435]. Indeed, a growing amount of evidence points to the direct correlation between the residence time of a drug molecule and its *in vivo* efficacy [431–441]; See Fig. 5.8 for a schematic visualization of the binding process for systems with simple and complex transition and intermediate states.

Currently, kinetic models such as unbiased MD, Markov state model, and weighted ensemble and metadynamics (MTD) have demonstrated the success of kinetic modeling in protein-targeted drug discovery [350, 434, 435, 442]. Similar kinetic studies for RNA-ligand binding is expected to provide a novel strategy for RNA-targeted drug design [443].

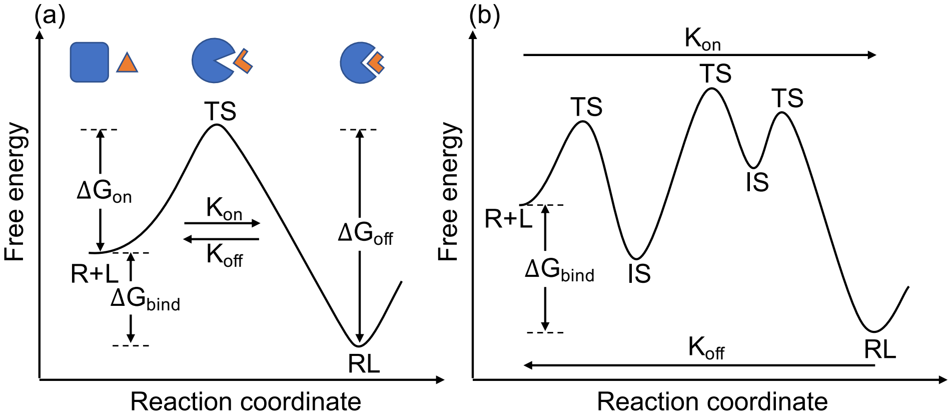


Figure 5.8: A simplified representation of the binding kinetics between the unbound receptor (R), unbound ligand (L) and the bound receptor-ligand complex (RL). (a) The binding kinetics of a system with only one transient state (TS) along the binding reaction coordinate. The figure shows a binding scenario where both receptor and ligand undergo conformational changes in the binding process. The kinetic residence time (i.e., the inverse of the RNA-ligand dissociation constant k_{off}) depends on the free energy difference (ΔG_{off}) between the bound state and transient state, while the thermodynamic binding energy (ΔG_{bind}) is determined by the free energy difference between the unbound state (R+L) and bound state (RL). (b) In practice, often the binding kinetic profile of a system contains multiple transient states (TS) and intermediate states (IS) with a much more complicated kinetic mechanism.

5.7 Conclusions and future perspectives

The rapidly growing therapeutic interest in RNA-targeted drug discovery causes an increasing demand for computational tools for predicting RNA-ligand interactions. Virtual screening remains an important first step in novel drug design when only the targeted RNA information is available. Various docking (sampling) methods and scoring functions have been developed to accelerate this process and in the meantime, have deepened our understanding of RNA-ligand binding mechanisms. Physics-based and knowledge-based approaches have shown promising success in predicting ligand binding poses and binding affinities. However, powered by advanced algorithms, machine-learning methods, although still in their infancy for RNA-ligand docking, have begun to show highly encouraging improved performance compared to traditional approaches.

Computer-aided drug discovery has come a long way. Past efforts have been mainly protein-centered. New RNA-based therapeutic design and computational methods have emerged. With the growth of the database of known structures and kinetic and thermodynamic measurements (e.g., binding affinity), data-driven methods, especially machine-learning methods, are expected to play a more and more important role. Furthermore, with the appreciation of RNA-ligand binding kinetic effects on *in vivo* efficacy of the drug, kinetics-based models, although currently have not been fully explored for RNA-drug binding, would be developed at an accelerated pace. With the development of various computational tools developed for RNA-targeted drug discovery, a CASP- and D3R-like [444, 445] community-wide events with blind tests and well-curated benchmark datasets, similar to the benchmarks widely used in the protein-ligand modeling community [367, 446–449], would be much needed.

Chapter 6

SPRank: an improved knowledge-based scoring function for modeling RNA-ligand interaction

Successful identification of high-quality lead compounds in the design of drug-like small molecules requires a scoring function that can give an accurate quantification for the interactions between RNA and small molecules. Here, we developed a knowledge-based statistical potential scoring function, SPRank, for predicting RNA-ligand interactions. A SYBYL-modified atom classification scheme is used to capture the intermolecular interactions between different chemical species. The parameters of SPRank are optimized through an iterative algorithm and the performance is evaluated by ten-fold cross-validation. On a widely used test set, SPRank outperforms other scoring functions with 66.7% success rate in identifying native binding modes for 42 RNA-ligand complexes, which is on par with other RNA-focused scoring functions. And the Pearson correlation coefficient between experimental affinities and SPRank predicted scores is above 0.66 for a test set consisting of

77 nucleic acid-ligand complexes.

6.1 Introduction

RNA molecules play critical roles in gene regulation [450–452] and protein synthesis [453–455]. Small molecules specifically targeting these functional RNAs can significantly affect the biological processes [456, 457]. Recently, RNA-based therapeutics has gained increasing interest in the field [2]. Diseases with related proteins that are "undruggable" or difficult to drug, i.e., proteins with a large flat surface without deep binding pockets, may be treated by designing drug-like small molecules to target the corresponding RNA elements that encode the proteins or regulate the biological processes. As an example, ligand-induced conformational switch is crucial to the regulatory mechanism of riboswitch, which serves as the critical regulatory element of its host messenger RNA [458]. One of the most extensively studied family of riboswitches is purine-sensing riboswitch. A small metabolite such as hypoxanthine or guanine bound to *B. subtilis xpt* guanine riboswitch can lead to the formation of a transcription terminator that "turns off" gene expression, In an adenine riboswitch, adenine is recognized by the *B. subtilis ydhL* riboswitch and gene expression can be activated by the disruption of the transcription terminator [459–461]. Due to their high selectivity for small molecules and ubiquity in bacteria, riboswitches can serve as potential antibacterial drug targets [462]. Another well-known example of small molecules affecting gene expression is the inhibition of protein synthesis induced by amino-modified glycoside, aminoglycoside [463, 464]. Aminoglycoside has the ability to interfere with eukaryotic translation mechanism by binding to the ribosomal decoding region. The mechanism has been used clinically as Gram-negative antibacterial drug [463]. The family of aminoglycoside antibiotics continues to grow with the addition of newly developed lead compounds, which was found to improve read-through activity and reduce toxicity [465, 466].

Computer-aided drug screening is a cost-effective way for identifying the lead for a large compound library. One of the core components to computationally model RNA-ligand interactions is the scoring function, from which we can identify the native binding mode and estimate binding affinity [467]. In the last decades, computational models have led to the discovery of several potent RNA-binding small molecules. These successful examples include the discovery of the lead compound (by the DOCK4 program) that disrupts the ribosomal frameshifting of SARS-CoV (severe acute respiratory syndrome coronavirus) [105, 106] and novel small molecules that target the HIV type 1 (HIV-1) TAR element (through the ICM program-aided virtual screening) [95]. These findings reveal the power of computer-aided RNA-targeted drug design.

One of the main challenges for developing the scoring function is the limited knowledge of the experimentally solved RNA-ligand complexes. Previous studies have included protein-ligand complexes into the training set [112, 118]. Through training the parameters for both RNA and protein systems, the approach could improve the overall performance of the scoring function and avoid overfitting in the training process development of RNA-targeted docking software. Examples are the scoring functions used in rDock [118] and DOCK6 [112]. In recent years, advancement in the X-ray crystallography [468], Nuclear Magnetic Resonance (NMR) Spectroscopy [469], and cryo-Electron Microscopy [470] leads to an rapid increase in in high-resolution three-dimensional structures of RNA-ligand complexes. Knowledge-based/machine learning scoring functions, such as SPA-LN [125], ITScore-NL [126], RNAposer [128], AnnapuRNA [127], whose parameters are derived from training sets solely consisting of RNA-ligand complexes has shown noticeable improvements.

However, compared to the development of the protein-specific scoring functions, the development of the RNA counterpart still lags. Much less effort is paid to predict the binding affinity, current scoring functions designed for RNA targets shows moderate success rate in native binding mode identification but limited performance in affinity estimation [130]. The current best Pearson correlation coefficients (R) between the predicted scores and the experimental binding affinities on the largest benchmark set (77 nucleic acid-small molecule complexes) is achieved by an iterative statistical potential approach, ITScore-NL [126], with R=0.64. The limited performance of current scoring functions demand a scoring function specifically designed for RNA system with better predictive ability in both binding mode identification and binding affinity estimation.

Knowledge-based/Statistical potential approach has been extended to RNA system [124–126] due to its success in modeling protein-ligand interactions [359, 369]. The performance of the derived scoring function highly depends on the chemical classification of the atoms and definition of the interacting pairwise potentials. DrugScore^{RNA} [344] uses SYBYL (mol2) [471] atom types to differentiate the interacting atom pairs through a distance-based potential. Kscore, a scoring function that employs a base-sensitive atom-typing scheme and distance-based potential, has achieved a Pearson correlation coefficient R=0.81 in binding affinity estimation for 15 RNA-ligand complexes [123]. In addition to the distance-dependent potential, LigandRNA [124] introduces an angle-dependent three-body potential, which improves the performance in binding mode identification. In this paper, we adopt a modified SYBYL-based atom classification scheme and an iterative procedure to derive an improved scoring function, namely, SPRank. The statistical potential derived from a training set with 130 non-redundant RNA-ligand complexes can improve the success rate of the native binding mode identification.

6.2 Materials and Methods

6.2.1 Curating the dataset

We downloaded all the RNA-ligand complexes from Protein Data Bank (PDB) [208]. Since the accuracy of the knowledge-based model depends on the quality of the training set [472], the following steps were used to compile our training set. First, we clustered redundant RNA-ligand complexes with different PDB identifiers (IDs) into the same group and selected the complex with the highest resolution as the representative case. Second, RNA structures with modified nucleotides were also discarded and ligands with a number of heavy atoms less than 5 were excluded. Finally, we removed RNA-ligand complexes included in the pose-identification set, which contains 42 RNA-ligand complexes and was used as a benchmark set in literature [124]. After the above steps, the training set was constructed with 130 RNA-ligand complexes. Besides the pose-identification set, a second test set (affinity-estimation set) was also prepared for model validation. Affinity-estimation set contains 77 nucleic acid-ligand complexes collected from the PDBBind database (version 2014) [377] with experimental binding affinity data. The PDB IDs of the training set and test sets can be found in Table 6.1.

To generate an ensemble of diverse decoys for the training set, we performed self docking and blind docking for each RNA-ligand complex in training set via rDock with $dock_{solv}$ score [118]. Reference-ligand method was used in self docking with the radius of the sphere set to 2, 4, 6, 8, and 10Åfor each docking run. In the blind docking, the two-sphere method was employed with the radius of the outer sphere set to 20, 30, and 50Åfor each docking run. 50 decoys were generated for each radius, and we were able to compile a conformational ensemble with 400 decoys for each training case. Decoys for pose-identification set and affinity-estimation set were generated in the same way as those generated in the training set.

Table 6.1: PDB IDs of training set and test sets

Training set									
1RAW	1ARJ	1ET4	1I9V	1LC4	1LVJ	1NTA	1O15	1O9M	1QD3
2GCV	1Y27	1YLS	1YRJ	2AU4	2B57	2ESJ	2ESI	2F4T	2G9C
2LWK	2HOO	2JUK	2KGP	2KTZ	2KU0	2KX8	2L1V	2L8H	2L94
2YIE	2M4Q	2MIY	2MXS	2O3V	2QWY	2XNW	2XNZ	2XO0	2XO1
3FO6	2Z74	3B4A	3DIL	3DIR	3DJ0	3GX3	3DVV	3E5C	3F4H
3MXH	3FU2	3G4M	3GAO	3GCA	3OWI	3GOT	3DS7	3IQN	3LA5
3SLM	3NPQ	3GER	3Q3Z	3Q50	3RKF	3S4P	3SD3	3SKI	3SKZ
4LVW	3SUH	3TZR	4ERJ	4FE5	4FRG	4JF2	4KQY	4L81	4LVV
4NYD	4LVX	4LVY	4LVZ	4LW0	4LX5	4LX6	4NYA	4NYB	4NYC
4YAZ	4NYG	4P95	4PDQ	4QK8	4QLN	4RZD	4TS2	4TZX	4XWF
5O62	4YB0	4ZNP	5BJO	5BTP	5BWS	5C45	5KPY	5KVJ	5KX9
6DLT	5OB3	5Z1H	5V3F	5XI1	5V0O	6AZ4	6BFB	6C63	6CK5
6QN3	6DMC	6E8S	6E1S	6E1W	6DN2	6FZ0	6HAG	6HBT	6HC5
Pose-identification set									
1AJU	1AM0	1BYJ	1EHT	1EI2	1F1T	1F27	1FMN	1FYP	1J7T
1KOC	1KOD	1MWL	1NBK	1NEM	1PBR	1Q8N	1TOB	1UTS	1UUD
1UUI	1XPF	1Y26	2BE0	2BEE	2ET8	2F4U	2FCZ	2FD0	2GDI
2O3X	2OE5	2PWT	2TOB	4P20	3D2X	3GX2	3SUX	1HNW	1FJG
1XBP	2OGN								
Affinity-estimation set									
1ARJ	1BYJ	1F1T	1F27	1FYP	1I9V	1Q8N	1QD3	1UTS	1UUD
1YKV	1YRJ	2AU4	2B57	2BE0	2BEE	2F4S	2F4T	2F4U	2G5K
2G9C	2KGP	2KTZ	2KU0	2KX8	2L94	2O3W	2O3X	2XNW	2XNZ
2XO0	2XO1	2YDH	2YGH	3DS7	3E5C	3FO4	3FO6	3FU2	3G4M
3GAO	3GER	3GES	3GOG	3GOT	3LA5	3MUM	3MUR	3MXH	3NPN
3OWZ	3Q3Z	3Q50	3S4P	3SD3	3SLM	4AOB	4ERJ	4FE5	4JF2
4KQY	4LVV	1CVX	1CVY	1DB6	1NZM	1P96	1QV4	1QV8	1R4E
2D55	2JWQ	2LOA	2MB3	316D	407D	408D			

6.2.2 Deriving the statistical potential with an iterative algorithm

The pairwise atomic interaction potential is derived from the inverse Boltzmann's law:

$$\Delta u_{i,j}(r) = -k_B T ln \frac{f_{i,j}^{OBS}(r)}{f_{i,j}^{REF}(r)} \tag{6.1}$$

Here, (i, j) denote the atom type i and j for an interacting atom pair. $f_{i,j}^{OBS}(r)$ and $f_{i,j}^{REF}(r)$ are the probability densities of the atom pair (i, j) at the distance r for the experimentally observed state and reference state, respectively. k_B is the Boltzmann constant and T is the absolute temperature.

The goal of the iterative algorithm is to derive a set of pairwise atomic potentials which can be used to distinguish the native/near-native binding pose from a pool of non-native poses (decoys). We followed the iterative algorithm that was first proposed by Thomas and Dill [369]. Previous studies have validated this iterative algorithm in both protein-ligand interaction modeling [359] and RNA structure prediction [473].

To determine the pairwise potential, we categorize atoms into 19 types based on their SYBYL types, see Table 6.2. For oxygen and nitrogen atoms, the classification depends on their ability to be hydrogen-bond acceptor/donor. And the classification of carbon atoms is adopted from ITScore [359]. Through the statistical analysis, we only keep the pairwise contacts with more than 300 occurrences in the training set which left us with 361 pairwise contacts.

Table 6.2: Atom definition for RNA and ligand.

SYBYL type	Atom type	Definition	SYBYL type	Atom type	Definition
C.2, C.ar, C.cat	C2NO	sp^2 carbon bonded to 1 negatively charged oxygen	N.2, N.am, N.ar, N.pl3	N21	sp^2 nitrogen bonded to 1 non-hydrogen atom
	C2PN	sp^2 carbon bonded to 1 positively charged nitrogen		N22	sp^2 nitrogen bonded to 2 non-hydrogen atoms
	C2N	amide carbon		N2	other sp^2 nitrogen
	C2O	carbonyl carbon except the above	N.4, N.3	NP	positively charged sp^3 nitrogen
	C2	other sp^2 carbon		N3	other sp^3 nitrogen
C.3	C3C	sp^3 carbon bonded to carbon or hydrogen only	O.co2	O2C	carbonyl oxygen
	C3	other sp^3 carbon	O.2	O2	sp^2 oxygen
S.2, S.3 S.o, S.o2	S	sulfur	O.3	O31	sp^3 oxygen bonded to 1 non-hydrogen atom
F, Cl, Br	Ha	halogen		O32	sp^3 oxygen bonded to 2 non-hydrogen atoms
P.3	P	phosphorus			

6.2.3 Training with ten-fold cross-validation

To maximize the training efficiency and avoid overfitting, we adopted ten-fold cross-validation to optimize the parameters. The training set was randomly divided into ten subsets. Each time the model was trained on nine subsets and the remaining subset was used to validate the model. The parameters with the best performance on the remaining subset were kept, and the average of the parameters obtained from ten trained models is used in the final scoring function.

6.3 Results and discussion

6.3.1 Examining the performance for native binding mode identification

Our SPRank model is compared with other state-of-art scoring functions, namely, ITScore-NL [126] and RLDOCK [119, 120]. As shown in Fig. 6.1, success rate is used to evaluate the performance of the model with both top-1 and top-3 predictions on the pose-identification set. The criteria for calculating the success rate is reported in the C(rank, RMSD) format. For example, C(1, 2.0) represents the success rate for the top-1 prediction that has RMSD <2.0 Å to the native pose. As shown in Fig. 6.1a, if C(1,2.0) (top-1 prediction, red) is used, RLDOCK correctly identify the binding modes of 21 complexes out of 38 non-ribosomal complexes with a success rate of 55.26%. Both SPRank and ITScore-NL can identify the native binding modes for over half of the 42 complexes, with success rates of 45.24% and 45.24%, respectively. For a loose criteria C(3,2.0) (top-3 prediction, blue), SPRank outperforms other scoring functions with a success rate of 61.90%.

A detailed analysis of the RMSD distribution for the correct identifications with C(1,2.0) is shown in Fig. 6.1b. The number of successful identifications within different RMSD cutoff intervals are shown with different colors. Among the 21 binding modes correctly predicted by SPRank with C(1,2.0), more than half of the cases have RMSDs below 1.0Å and shows a slightly better performance than RLDOCK. In general, both SPRank and RLDOCK have comparable performance and outperform other scoring functions in native binding mode identification.

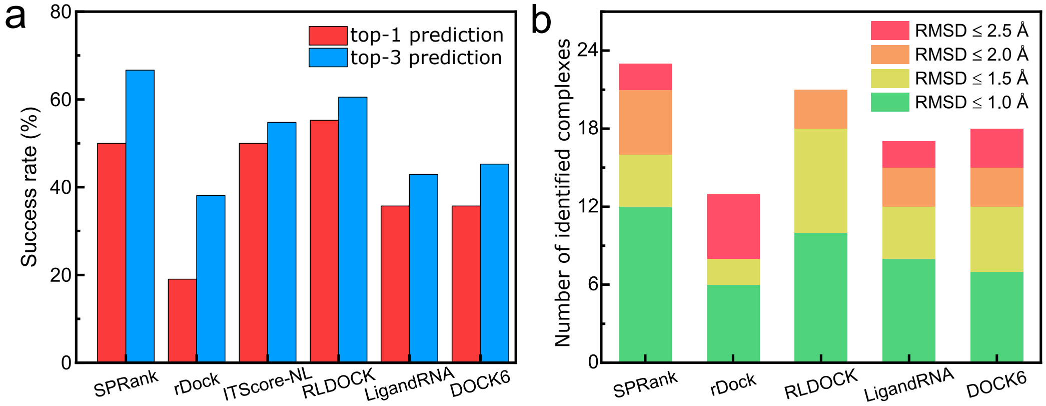


Figure 6.1: The comparison between SPRank and other scoring functions [119, 124, 126] on the pose-identification set with 42 RNA-ligand complexes. (a) The success rate of different scoring functions for the top-1 (red) and top-3 (blue) predictions. For both top-1 and top-3 predictions, correct prediction requires that at least one of the top-ranked poses is within RMSD 2.0Å relative to the native pose. (b) The number of the native binding modes correctly identified by various scoring functions with top-1 prediction and RMSD cutoff 2.0 Å. The successful cases are shown in different colors, where green/yellow/orange/red denotes the cases with the RMSD within (0.0Å,1.0Å)/(1.0Å,1.5Å)/(1.5Å,2.0Å)/(2.0Å,2.5Å) RMSD intervals, respectively. The data of ITScore-NL, RLDOCK, LigandRNA, and DOCK 6 were collected from previous publications [119, 124, 126].

6.3.2 Examining the performance for binding affinity estimation

Accurate prediction of the binding affinity for any given RNA-ligand complex is a much more challenging task than the binding mode identification. Because binding affinity is sensitive to small deviations of the relative positions between RNA and bound ligand. Optimal binding affinity estimation can only be obtained when the bound ligand is close enough to the native binding mode. However, with current docking software, it is difficult to position and orient the ligand to precisely reproduce the native binding mode, even when the pocket is already known. Moreover, the inherent flexibility of RNA molecules implies that the observed binding affinity should be the average of all the RNA-ligand complexes in a conformational ensemble. In practice, using a generated conformational ensemble for a given bound ligand and keeping the RNA molecule rigid is a much more reasonable approach to estimate the binding affinity due to the consideration of computational cost.

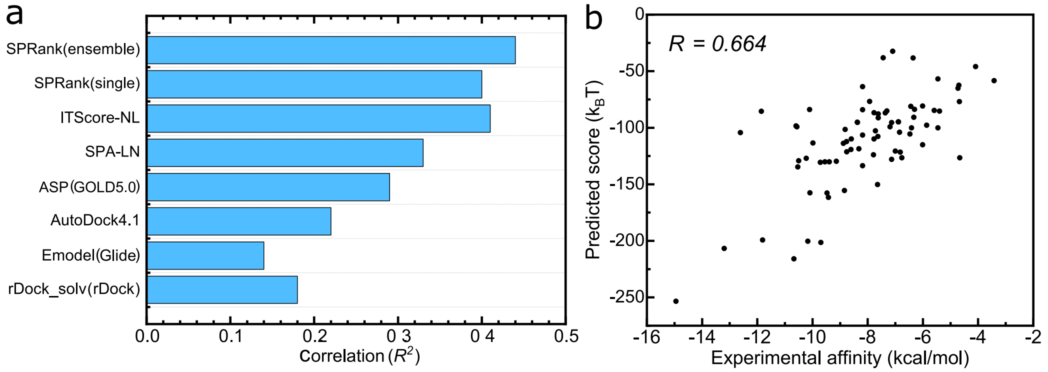


Figure 6.2: The Pearson correlation coefficients between the experimental affinities and the predicted scores on the affinity-estimation set (77 nucleic acid-ligand complexes) for various scoring functions. (a) The comparison of the correlation values between SPRank and other scoring functions. The SPRank(ensemble) and SPRank(single) represent predictions with only experimental solved structures and predictions with the generated conformational ensembles, respectively. The scoring function associated with a specific docking engine is shown in score(engine) format. (b) The plot of both experimental affinities and the scores predicted by SPRank (ensemble) on the affinity-estimation set.

In this study, we used SPRank with two different approaches to estimate the binding affinity for any given RNA-ligand complex, namely, SPRank (single) and SPRank (ensemble). In the first case, SPRank (single), only the experimentally solved complex structure is used for predicting binding affinity. The Pearson correlation coefficients (R) between experimental binding affinities and predicted scores of various scoring functions on the affinity-estimation set are shown in Fig. 6.2a. Clearly SPRank (single) and ITScore-NL both outperform other scoring functions and are able to achieve similar correlations with R=0.63 and R=0.64, respectively. In the second case, SPRank (ensemble), the predicted score is a Boltzmann weighted average across the entire ligand conformational ensemble for the given RNA-ligand complex. The ligand conformational ensemble is constructed by including all the sampled poses with RMSD less than 5.0Å relative to the native pose (i.e., the sampled poses within the binding pocket). This way, we are able to slightly increase the correlation from R=0.63 for SPRank (single) to R=0.664 for SPRank (ensemble).

6.4 Conclusion

We have developed a knowledge-based scoring function, SPRank, to identify the native ligand-binding mode and estimate the ligand-binding affinity. The pairwise potentials are defined by the modified SYBYL atom types and functions of the atom-pair distances. The parameters are optimized iteratively through ten-fold cross-validation. On both pose-identification set and affinity-estimation set, SPRank exhibits comparable performance as the state-of-art models. Results on the affinity-estimation set show that the correlation can be further improved through the ensemble-based scoring scheme.

Bibliography

[1] Sarah Djebali et al. “Landscape of transcription in human cells”. In: *Nature* 489.7414 (2012), pp. 101–108. ISSN: 1476-4687. DOI: 10.1038/nature11233. URL: https://doi.org/10.1038/nature11233.

[2] Katherine Deigan Warner, Christine E. Hajdin, and Kevin M. Weeks. “Principles for targeting RNA with drug-like small molecules”. In: *Nature Reviews Drug Discovery* 17.8 (2018), pp. 547–558. ISSN: 1474-1784. DOI: 10.1038/nrd.2018.93. URL: https://doi.org/10.1038/nrd.2018.93.

[3] Francesca Tessaro and Leonardo Scapozza. “How ‘Protein-Docking’ Translates into the New Emerging Field of Docking Small Molecules to Nucleic Acids?” In: *Molecules* 25.12 (2020). ISSN: 1420-3049. DOI: 10.3390/molecules25122749. URL: https://www.mdpi.com/1420-3049/25/12/2749.

[4] Matthew G. Costales et al. “How We Think about Targeting RNA with Small Molecules”. In: *Journal of Medicinal Chemistry* 63.17 (2020). PMID: 32212706, pp. 8880–8900. DOI: 10.1021/acs.jmedchem.9b01927. eprint: https:

//doi.org/10.1021/acs.jmedchem.9b01927. URL: https://doi.org/10.1021/acs.jmedchem.9b01927.

[5] Michele Clamp et al. "Distinguishing protein-coding and noncoding genes in the human genome". In: *Proceedings of the National Academy of Sciences* 104.49 (2007), pp. 19428–19433. ISSN: 0027-8424. DOI: 10.1073/pnas.0709013104. eprint: https://www.pnas.org/content/104/49/19428.full.pdf. URL: https://www.pnas.org/content/104/49/19428.

[6] Iakes Ezkurdia et al. "Multiple evidence strands suggest that there may be as few as 19 000 human protein-coding genes". In: *Human Molecular Genetics* 23.22 (June 2014), pp. 5866–5878. ISSN: 0964-6906. DOI: 10.1093/hmg/ddu309. eprint: https://academic.oup.com/hmg/article-pdf/23/22/5866/17261056/ddu309.pdf. URL: https://doi.org/10.1093/hmg/ddu309.

[7] Joseph C. Somody, Stephen S. MacKinnon, and Andreas Windemuth. "Structural coverage of the proteome for pharmaceutical applications". In: *Drug Discovery Today* 22.12 (2017), pp. 1792–1799. ISSN: 1359-6446. DOI: https://doi.org/10.1016/j.drudis.2017.08.004. URL: https://www.sciencedirect.com/science/article/pii/S1359644617301642.

[8] Andrew L. Hopkins and Colin R. Groom. “The druggable genome”. In: *Nature Reviews Drug Discovery* 1.9 (2002), pp. 727–730. ISSN: 1474-1784. DOI: 10.1038/nrd892. URL: https://doi.org/10.1038/nrd892.

[9] John P. Overington, Bissan Al-Lazikani, and Andrew L. Hopkins. “How many drug targets are there?” In: *Nature Reviews Drug Discovery* 5.12 (2006), pp. 993–996. ISSN: 1474-1784. DOI: 10.1038/nrd2199. URL: https://doi.org/10.1038/nrd2199.

[10] Scott J Dixon and Brent R Stockwell. “Identifying druggable disease-modifying gene products”. In: *Current Opinion in Chemical Biology* 13.5 (2009). Omics/Biopolymers/Model Systems, pp. 549 –555. ISSN: 1367-5931. DOI: https://doi.org/10.1016/j.cbpa.2009.08.003. URL: http://www.sciencedirect.com/science/article/pii/S1367593109001070.

[11] Phillip A. Sharp. “The Centrality of RNA”. In: *Cell* 136.4 (2009), pp. 577–580. ISSN: 0092-8674. DOI: https://doi.org/10.1016/j.cell.2009.02.007. URL: https://www.sciencedirect.com/science/article/pii/S0092867409001433.

[12] James Chappell et al. “The centrality of RNA for engineering gene expression”. In: *Biotechnology Journal* 8.12 (2013), pp. 1379–1395. DOI: https://doi.org/10.1002/biot.201300018. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/biot.201300018. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/biot.201300018.

[13] Stanley T. Crooke et al. “RNA-Targeted Therapeutics”. In: *Cell Metabolism* 27.4 (2018), pp. 714–739. ISSN: 1550-4131. DOI: https://doi.org/10.1016/j.cmet.2018.03.004. URL: http://www.sciencedirect.com/science/article/pii/S1550413118301827.

[14] Wei Yin and Mark Rogge. “Targeting RNA: A Transformative Therapeutic Strategy”. In: *Clinical and Translational Science* 12.2 (2019), pp. 98–112. DOI: https://doi.org/10.1111/cts.12624. eprint: https://ascpt.onlinelibrary.wiley.com/doi/pdf/10.1111/cts.12624. URL: https://ascpt.onlinelibrary.wiley.com/doi/abs/10.1111/cts.12624.

[15] Ai-Ming Yu et al. “RNA therapy: Are we using the right molecules?” In: *Pharmacology & Therapeutics* 196 (2019), pp. 91 –104. ISSN: 0163-7258. DOI: https://doi.org/10.1016/j.pharmthera.2018.11.011. URL: http://www.sciencedirect.com/science/article/pii/S016372581830216X.

[16] Ai-Ming Yu, Young Hee Choi, and Mei-Juan Tu. “RNA Drugs and RNA Targets for Small Molecules: Principles, Progress, and Challenges”. In: *Pharmacological Reviews* 72.4 (2020). Ed. by RHIAN M. TOUYZ, pp. 862–898. ISSN: 0031-6997. DOI: 10.1124/pr.120.019554. eprint: https://pharmrev.aspetjournals.org/content/72/4/862.full.pdf. URL: https://pharmrev.aspetjournals.org/content/72/4/862.

[17] Colleen M. Connelly, Michelle H. Moon, and John S. Schneekloth. "The Emerging Role of RNA as a Therapeutic Target for Small Molecules". In: *Cell Chemical Biology* 23.9 (2016), pp. 1077 –1090. ISSN: 2451-9456. DOI: https://doi.org/10.1016/j.chembiol.2016.05.021. URL: http://www.sciencedirect.com/science/article/pii/S2451945616302525.

[18] Thomas Hermann. "Small molecules targeting viral RNA". In: *WIREs RNA* 7.6 (2016), pp. 726–743. DOI: https://doi.org/10.1002/wrna.1373. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/wrna.1373. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/wrna.1373.

[19] Anita Donlic and Amanda E. Hargrove. "Targeting RNA in mammalian systems with small molecules". In: *WIREs RNA* 9.4 (2018), e1477. DOI: https://doi.org/10.1002/wrna.1477. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/wrna.1477. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/wrna.1477.

[20] Samantha M. Meyer et al. "Small molecule recognition of disease-relevant RNA structures". In: *Chem. Soc. Rev.* 49 (19 2020), pp. 7167–7199. DOI: 10.1039/D0CS00560F. URL: http://dx.doi.org/10.1039/D0CS00560F.

[21] Yanqiu Shao and Qiangfeng Cliff Zhang. "Targeting RNA structures in diseases with small molecules". In: *Essays in Biochemistry* 64.6 (2020), pp. 955–966. ISSN:

0071-1365. DOI: 10.1042/EBC20200011. URL: https://doi.org/10.1042/EBC20200011.

[22] Vinod K. Misra and David E. Draper. “On the role of magnesium ions in RNA stability”. In: *Biopolymers* 48.2-3 (1998), pp. 113–135. DOI: https://doi.org/10.1002/(SICI)1097-0282(1998)48:2<113::AID-BIP3>3.0.CO;2-Y. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/%28SICI%291097-0282%281998%2948%3A2%3C113%3A%3AAID-BIP3%3E3.0.CO%3B2-Y. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291097-0282%281998%2948%3A2%3C113%3A%3AAID-BIP3%3E3.0.CO%3B2-Y.

[23] Ignacio Tinoco and Carlos Bustamante. “How RNA folds”. In: *Journal of Molecular Biology* 293.2 (1999), pp. 271–281. ISSN: 0022-2836. DOI: https://doi.org/10.1006/jmbi.1999.3001. URL: https://www.sciencedirect.com/science/article/pii/S0022283699930012.

[24] Vinod K. Misra and David E. Draper. “The linkage between magnesium binding and RNA folding11Edited by B. Honig”. In: *Journal of Molecular Biology* 317.4 (2002), pp. 507–521. ISSN: 0022-2836. DOI: https://doi.org/10.1006/jmbi.2002.5422. URL: https://www.sciencedirect.com/science/article/pii/S0022283602954227.

[25] DAVID E. DRAPER. “A guide to ions and RNA structure”. In: *RNA* 10.3 (2004), pp. 335–343. DOI: 10.1261/rna.5205404. eprint: http://rnajournal.

cshlp.org/content/10/3/335.full.pdf+html. URL: http://rnajournal.cshlp.org/content/10/3/335.abstract.

[26] David E. Draper, Dan Grilley, and Ana Maria Soto. “Ions and RNA Folding”. In: *Annual Review of Biophysics and Biomolecular Structure* 34.1 (2005). PMID: 15869389, pp. 221–243. DOI: 10.1146/annurev.biophys.34.040204.144511. eprint: https://doi.org/10.1146/annurev.biophys.34.040204.144511. URL: https://doi.org/10.1146/annurev.biophys.34.040204.144511.

[27] David E. Draper. “RNA Folding: Thermodynamic and Molecular Descriptions of the Roles of Ions”. In: *Biophysical Journal* 95.12 (2008), pp. 5489–5495. ISSN: 0006-3495. DOI: https://doi.org/10.1529/biophysj.108.131813. URL: https://www.sciencedirect.com/science/article/pii/S0006349508819716.

[28] David E. Draper. “Folding of RNA tertiary structure: Linkages between backbone phosphates, ions, and water”. In: *Biopolymers* 99.12 (2013), pp. 1105–1113. DOI: https://doi.org/10.1002/bip.22249. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/bip.22249. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/bip.22249.

[29] Nina M Fischer et al. “Influence of Na+ and Mg2+ ions on RNA structures studied with molecular dynamics simulations”. In: *Nucleic Acids Research* 46.10 (Apr. 2018), pp. 4872–4882. ISSN: 0305-1048. DOI: 10.1093/nar/gky221. eprint:

https://academic.oup.com/nar/article-pdf/46/10/4872/24962537/gky221.pdf. URL: https://doi.org/10.1093/nar/gky221.

[30] Anna Marie Pyle. "Ribozymes: A Distinct Class of Metalloenzymes". In: *Science* 261.5122 (1993), pp. 709–714. DOI: 10.1126/science.7688142. eprint: https://www.science.org/doi/pdf/10.1126/science.7688142. URL: https://www.science.org/doi/abs/10.1126/science.7688142.

[31] Snorri Th. Sigurdsson and Fritz Eckstein. "Structure-function relationships of hammerhead ribozymes: from understanding to applications". In: *Trends in Biotechnology* 13.8 (1995), pp. 286–289. ISSN: 0167-7799. DOI: https://doi.org/10.1016/S0167-7799(00)88966-0. URL: https://www.sciencedirect.com/science/article/pii/S0167779900889660.

[32] Jamie H. Cate, Raven L. Hanna, and Jennifer A. Doudna. "A magnesium ion core at the heart of a ribozyme domain". In: *Nature Structural Biology* 4.7 (1997), pp. 553–558. ISSN: 1545-9985. DOI: 10.1038/nsb0797-553. URL: https://doi.org/10.1038/nsb0797-553.

[33] Thomas Hermann et al. "Evidence for a hydroxide ion bridging two magnesium ions at the active site of the hammerhead ribozyme". In: *Nucleic Acids Research* 25.17 (Sept. 1997), pp. 3421–3427. ISSN: 0305-1048. DOI: 10.1093/nar/25.17.3421. eprint: https://academic.oup.com/nar/article-

pdf/25/17/3421/3646952/25-17-3421.pdf. URL: https://doi.org/10.1093/nar/25.17.3421.

[34] Shu-ou Shan et al. "Three metal ions at the active site of the Tetrahymena group I ribozyme". In: *Proceedings of the National Academy of Sciences* 96.22 (1999), pp. 12299–12304. ISSN: 0027-8424. DOI: 10.1073/pnas.96.22.12299. eprint: https://www.pnas.org/content/96/22/12299.full.pdf. URL: https://www.pnas.org/content/96/22/12299.

[35] Raven Hanna and Jennifer A Doudna. "Metal ions in ribozyme folding and catalysis". In: *Current Opinion in Chemical Biology* 4.2 (2000), pp. 166–170. ISSN: 1367-5931. DOI: https://doi.org/10.1016/S1367-5931(99)00071-X. URL: https://www.sciencedirect.com/science/article/pii/S136759319900071X.

[36] Mathias Brännvall and Leif A. Kirsebom. "Metal ion cooperativity in ribozyme cleavage of RNA". In: *Proceedings of the National Academy of Sciences* 98.23 (2001), pp. 12943–12947. ISSN: 0027-8424. DOI: 10.1073/pnas.221456598. eprint: https://www.pnas.org/content/98/23/12943.full.pdf. URL: https://www.pnas.org/content/98/23/12943.

[37] Joachim Schnabl and Roland KO Sigel. "Controlling ribozyme activity by metal ions". In: *Current Opinion in Chemical Biology* 14.2 (2010). Biocatalysis and Biotransformation/Bioinorganic Chemistry, pp. 269–275. ISSN: 1367-5931. DOI: https://doi.org/10.1016/j.cbpa.2009.11.024. URL:

https://www.sciencedirect.com/science/article/pii/S1367593109001951.

[38] Claus URBANKE, Roland RÖMER, and Günter MAASS. “Tertiary Structure of tRNAPhe (Yeast): Kinetics and Electrostatic Repulsion”. In: *European Journal of Biochemistry* 55.2 (1975), pp. 439–444. DOI: https://doi.org/10.1111/j.1432-1033.1975.tb02180.x. eprint: https://febs.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1432-1033.1975.tb02180.x. URL: https://febs.onlinelibrary.wiley.com/doi/abs/10.1111/j.1432-1033.1975.tb02180.x.

[39] Roland RÖMER and Renate HACH. “tRNA Conformation and Magnesium Binding”. In: *European Journal of Biochemistry* 55.1 (1975), pp. 271–284. DOI: https://doi.org/10.1111/j.1432-1033.1975.tb02160.x. eprint: https://febs.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1432-1033.1975.tb02160.x. URL: https://febs.onlinelibrary.wiley.com/doi/abs/10.1111/j.1432-1033.1975.tb02160.x.

[40] A Stein and DM Crothers. “Equilibrium binding of magnesium(II) by Escherichia coli tRNAfMet”. In: *Biochemistry* 15.1 (1976), 157—160. ISSN: 0006-2960. DOI: 10.1021/bi00646a024. URL: https://doi.org/10.1021/bi00646a024.

[41] Heping Zheng et al. “Magnesium-binding architectures in RNA crystal structures: validation, binding preferences, classification and motif detection”. In: *Nucleic*

Acids Research 43.7 (Mar. 2015), pp. 3789–3801. ISSN: 0305-1048. DOI: 10.1093/nar/gkv225. eprint: https://academic.oup.com/nar/article-pdf/43/7/3789/17436309/gkv225.pdf. URL: https://doi.org/10.1093/nar/gkv225.

[42] Daniel J. Klein, Peter B. Moore, and Thomas A. Steitz. "The contribution of metal ions to the structural stability of the large ribosomal subunit". In: *RNA* 10.9 (2004), pp. 1366–1379. DOI: 10.1261/rna.7390804. eprint: http://rnajournal.cshlp.org/content/10/9/1366.full.pdf+html. URL: http://rnajournal.cshlp.org/content/10/9/1366.abstract.

[43] Bernhard Lippert. "Multiplicity of metal ion binding patterns to nucleobases". In: *Coordination Chemistry Reviews* 200-202 (2000), pp. 487–516. ISSN: 0010-8545. DOI: https://doi.org/10.1016/S0010-8545(00)00260-5. URL: https://www.sciencedirect.com/science/article/pii/S0010854500002605.

[44] Ignacio Tinoco and Jeffrey S. Kieft. "The ion core in RNA folding". In: *Nature Structural Biology* 4.7 (1997), pp. 509–512. ISSN: 1545-9985. DOI: 10.1038/nsb0797-509. URL: https://doi.org/10.1038/nsb0797-509.

[45] E Ennifar et al. "The crystal structure of the dimerization initiation site of genomic HIV-1 RNA reveals an extended duplex with two adenine bulges". In: *Structure* 7.11 (1999), pp. 1439–1449. ISSN: 0969-2126. DOI: https://doi.org/10.1016/S0969-2126(00)80033-7. URL:

https://www.sciencedirect.com/science/article/pii/S0969212600800337.

[46] Carl C. Correll et al. “Metals, Motifs, and Recognition in the Crystal Structure of a 5S rRNA Domain”. In: *Cell* 91.5 (1997), pp. 705–712. ISSN: 0092-8674. DOI: https://doi.org/10.1016/S0092-8674(00)80457-2. URL: https://www.sciencedirect.com/science/article/pii/S0092867400804572.

[47] Anton S. Petrov et al. “Bidentate RNA–magnesium clamps: On the origin of the special role of magnesium in RNA folding”. In: *RNA* 17.2 (2011), pp. 291–297. DOI: 10.1261/rna.2390311. eprint: http://rnajournal.cshlp.org/content/17/2/291.full.pdf+html. URL: http://rnajournal.cshlp.org/content/17/2/291.abstract.

[48] Herbert L. Ennis and Michael Artman. “Ribosome size distribution in extracts of potassium-depleted Escherichia coli”. In: *Biochemical and Biophysical Research Communications* 48.1 (1972), pp. 161–168. ISSN: 0006-291X. DOI: https://doi.org/10.1016/0006-291X(72)90357-9. URL: https://www.sciencedirect.com/science/article/pii/0006291X72903579.

[49] Chiaolong Hsiao and Loren Dean Williams. “A recurrent magnesium-binding motif provides a framework for the ribosomal peptidyl transferase center”. In: *Nucleic Acids Research* 37.10 (Mar. 2009), pp. 3134–3142. ISSN: 0305-1048. DOI: 10.1093/nar/gkp119. eprint: https://academic.oup.com/nar/

article-pdf/37/10/3134/16751843/gkp119.pdf. URL: https://doi.org/10.1093/nar/gkp119.

[50] B.J. McCarthy. “The effects of magnesium starvation on the ribosome content of Escherichia coli”. In: *Biochimica et Biophysica Acta (BBA) - Specialized Section on Nucleic Acids and Related Subjects* 55.6 (1962), pp. 880–889. ISSN: 0926-6550. DOI: https://doi.org/10.1016/0926-6550(62)90345-6. URL: https://www.sciencedirect.com/science/article/pii/0926655062903456.

[51] Fu-Chuan Chao. “Dissociation of macromolecular ribonucleoprotein of yeast”. In: *Archives of Biochemistry and Biophysics* 70.2 (1957), pp. 426–431. ISSN: 0003-9861. DOI: https://doi.org/10.1016/0003-9861(57)90130-3. URL: https://www.sciencedirect.com/science/article/pii/0003986157901303.

[52] Fu-Chuan Chao and H.K. Schachman. “The isolation and characterization of a macromolecular ribonucleoprotein from yeast”. In: *Archives of Biochemistry and Biophysics* 61.1 (1956), pp. 220–230. ISSN: 0003-9861. DOI: https://doi.org/10.1016/0003-9861(56)90334-4. URL: https://www.sciencedirect.com/science/article/pii/0003986156903344.

[53] A. TISSIÈRES and J. D. WATSON. “Ribonucleoprotein Particles from Escherichia Coli”. In: *Nature* 182.4638 (1958), pp. 778–780. ISSN: 1476-4687. DOI: 10.1038/182778b0. URL: https://doi.org/10.1038/182778b0.

[54] Lasse Jenner et al. “Structural rearrangements of the ribosome at the tRNA proofreading step”. In: *Nature Structural & Molecular Biology* 17.9 (2010), pp. 1072–1078. ISSN: 1545-9985. DOI: 10.1038/nsmb.1880. URL: https://doi.org/10.1038/nsmb.1880.

[55] Jessica C Bowman et al. “Cations in charge: magnesium ions in RNA folding and catalysis”. In: *Current Opinion in Structural Biology* 22.3 (2012). Nucleic acids/Sequences and topology, pp. 262–272. ISSN: 0959-440X. DOI: https://doi.org/10.1016/j.sbi.2012.04.006. URL: https://www.sciencedirect.com/science/article/pii/S0959440X12000723.

[56] William G. Scott, John T. Finch, and Aaron Klug. “The crystal structure of an AII-RNAhammerhead ribozyme: A proposed mechanism for RNA catalytic cleavage”. In: *Cell* 81.7 (1995), pp. 991–1002. ISSN: 0092-8674. DOI: https://doi.org/10.1016/S0092-8674(05)80004-2. URL: https://www.sciencedirect.com/science/article/pii/S0092867405800042.

[57] Timothy J. Wilson and David M.J. Lilley. “RNA catalysis—is that it?” In: *RNA* 21.4 (2015), pp. 534–537. DOI: 10.1261/rna.049874.115. eprint: http://rnajournal.cshlp.org/content/21/4/534.full.pdf+html. URL: http://rnajournal.cshlp.org/content/21/4/534.short.

[58] David M.J. Lilley. “How RNA acts as a nuclease: some mechanistic comparisons in the nucleolytic ribozymes”. In: *Biochemical Society Transactions* 45.3 (June 2017),

pp. 683–691. ISSN: 0300-5127. DOI: 10.1042/BST20160158. eprint: https://portlandpress.com/biochemsoctrans/article-pdf/45/3/683/723681/bst-2016-0158c.pdf. URL: https://doi.org/10.1042/BST20160158.

[59] Thomas Hermann and Eric Westhof. “Exploration of metal ion binding sites in RNA folds by Brownian-dynamics simulations”. In: *Structure* 6.10 (1998), pp. 1303–1314. ISSN: 0969-2126. DOI: https://doi.org/10.1016/S0969-2126(98)00130-0. URL: https://www.sciencedirect.com/science/article/pii/S0969212698001300.

[60] Nils E. Mikkelsen et al. “Inhibition of RNase P RNA cleavage by aminoglycosides”. In: *Proceedings of the National Academy of Sciences* 96.11 (1999), pp. 6155–6160. ISSN: 0027-8424. DOI: 10.1073/pnas.96.11.6155. eprint: https://www.pnas.org/content/96/11/6155.full.pdf. URL: https://www.pnas.org/content/96/11/6155.

[61] Nils E. Mikkelsen et al. “Aminoglycoside binding displaces a divalent metal ion in a tRNA–neomycin B complex”. In: *Nature Structural Biology* 8.6 (2001), pp. 510–514. ISSN: 1545-9985. DOI: 10.1038/88569. URL: https://doi.org/10.1038/88569.

[62] Murad Nayal and Enrico Di Cera. “Valence Screening of Water in Protein Crystals Reveals Potential Na+Binding Sites”. In: *Journal of Molecular Biology* 256.2 (1996), pp. 228–234. ISSN: 0022-2836. DOI: https://doi.org/10.

1006/jmbi.1996.0081. URL: https://www.sciencedirect.com/science/article/pii/S0022283696900819.

[63] Loren Dean Williams. “Between Objectivity and Whim: Nucleic Acid Structural Biology”. In: *DNA Binders and Related Subjects: -/-*. Ed. by Michael J. Waring and Jonathan B. Chaires. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 77–88. ISBN: 978-3-540-31463-9. DOI: 10.1007/b100443. URL: https://doi.org/10.1007/b100443.

[64] Marjorie M. Harding. “The geometry of metal-ligand interactions relevant to proteins”. In: *Acta Crystallographica Section D* 55.8 (1999), pp. 1432–1443. DOI: 10.1107/S0907444999007374. URL: https://doi.org/10.1107/S0907444999007374.

[65] Marjorie M. Harding. “Metal-ligand geometry relevant to proteins and in proteins: sodium and potassium”. In: *Acta Crystallographica Section D* 58.5 (2002), pp. 872–874. DOI: 10.1107/S0907444902003712. URL: https://doi.org/10.1107/S0907444902003712.

[66] Marjorie M. Harding. “Small revisions to predicted distances around metal sites in proteins”. In: *Acta Crystallographica Section D* 62.6 (2006), pp. 678–682. DOI: 10.1107/S0907444906014594. URL: https://doi.org/10.1107/S0907444906014594.

[67] Charles W. Bock et al. “The Arrangement of First- and Second-shell Water Molecules Around Metal Ions: Effects of Charge and Size”. In: *Theoretical Chemistry Accounts* 115.2 (2006), pp. 100–112. ISSN: 1432-2234. DOI:

10.1007/s00214-005-0056-2. URL: https://doi.org/10.1007/s00214-005-0056-2.

[68] Yizhak Marcus. “Ionic radii in aqueous solutions”. In: *Chemical Reviews* 88.8 (1988), pp. 1475–1498.

[69] Yizhak Marcus. “Effect of Ions on the Structure of Water: Structure Making and Breaking”. In: *Chemical Reviews* 109.3 (2009). PMID: 19236019, pp. 1346–1370. DOI: 10.1021/cr8003828. eprint: https://doi.org/10.1021/cr8003828. URL: https://doi.org/10.1021/cr8003828.

[70] Pascal Auffinger, Neena Grover, and Eric Westhof. “Metal Ion Binding to RNA”. In: *Structural and Catalytic Roles of Metal Ions in RNA*. Ed. by Helmut Sigel, Astrid Sigel, and Roland K.O. Sigel. De Gruyter, 2015, pp. 1–36. DOI: doi:10.1515/9783110436648-006. URL: https://doi.org/10.1515/9783110436648-006.

[71] David R Cooper et al. “X-ray crystallography: assessment and validation of protein–small molecule complexes for drug discovery”. In: *Expert Opinion on Drug Discovery* 6.8 (2011). PMID: 21779303, pp. 771–782. DOI: 10.1517/17460441.2011.585154. eprint: https://doi.org/10.1517/17460441.2011.585154. URL: https://doi.org/10.1517/17460441.2011.585154.

[72] Edwin Pozharski, Christian X. Weichenberger, and Bernhard Rupp. “Techniques, tools and best practices for ligand electron-density analysis and results from their application to deposited crystal structures”. In: *Acta Crystallographica Section*

D 69.2 (2013), pp. 150–167. DOI: 10.1107/S0907444912044423. URL: https://doi.org/10.1107/S0907444912044423.

[73] Heping Zheng et al. "Validation of metal-binding sites in macromolecular structures with the CheckMyMetal web server". In: *Nature Protocols* 9.1 (2014), pp. 156–170. ISSN: 1750-2799. DOI: 10.1038/nprot.2013.172. URL: https://doi.org/10.1038/nprot.2013.172.

[74] Jacob Poehlsgaard and Stephen Douthwaite. "The bacterial ribosome as a target for antibiotics". In: *Nature Reviews Microbiology* 3.11 (2005), pp. 870–881. ISSN: 1740-1534. DOI: 10.1038/nrmicro1265. URL: https://doi.org/10.1038/nrmicro1265.

[75] David Bulkley et al. "Revisiting the structures of several antibiotics bound to the bacterial ribosome". In: *Proceedings of the National Academy of Sciences* 107.40 (2010), pp. 17158–17163. ISSN: 0027-8424. DOI: 10.1073/pnas.1008685107. eprint: https://www.pnas.org/content/107/40/17158.full.pdf. URL: https://www.pnas.org/content/107/40/17158.

[76] Dalia Deak et al. "Progress in the Fight Against Multidrug-Resistant Bacteria? A Review of U.S. Food and Drug Administration–Approved Antibiotics, 2010–2015". In: *Annals of Internal Medicine* 165.5 (2016), pp. 363–372. ISSN: 0003-4819. DOI: 10.7326/M16-0291. URL: https://www.acpjournals.org/doi/abs/10.7326/M16-0291.

[77] Jinzhong Lin et al. “Ribosome-Targeting Antibiotics: Modes of Action, Mechanisms of Resistance, and Implications for Drug Design”. In: *Annual Review of Biochemistry* 87.1 (2018). PMID: 29570352, pp. 451–478. DOI: 10.1146/annurev-biochem-062917-011942. eprint: https://doi.org/10.1146/annurev-biochem-062917-011942. URL: https://doi.org/10.1146/annurev-biochem-062917-011942.

[78] Maumita Mandal and Ronald R. Breaker. “Gene regulation by riboswitches”. In: *Nature Reviews Molecular Cell Biology* 5.6 (2004), pp. 451–463. ISSN: 1471-0080. DOI: 10.1038/nrm1403. URL: https://doi.org/10.1038/nrm1403.

[79] Brian J Tucker and Ronald R Breaker. “Riboswitches as versatile gene control elements”. In: *Current Opinion in Structural Biology* 15.3 (2005). Sequences and topology/Nucleic acids, pp. 342 –348. ISSN: 0959-440X. DOI: https://doi.org/10.1016/j.sbi.2005.05.003. URL: http://www.sciencedirect.com/science/article/pii/S0959440X05000874.

[80] Kenneth F. Blount and Ronald R. Breaker. “Riboswitches as antibacterial drug targets”. In: *Nature Biotechnology* 24.12 (2006), pp. 1558–1564. ISSN: 1546-1696. DOI: 10.1038/nbt1268. URL: https://doi.org/10.1038/nbt1268.

[81] Rebecca K. Montange and Robert T. Batey. “Riboswitches: Emerging Themes in RNA Structure and Function”. In: *Annual Review of Biophysics* 37.1 (2008). PMID: 18573075, pp. 117–133. DOI: 10.1146/annurev.biophys.37.032807.130000. eprint: https://doi.org/10.1146/annurev.biophys.

37.032807.130000. URL: https://doi.org/10.1146/annurev.biophys.37.032807.130000.

[82] Adam Roth and Ronald R. Breaker. “The Structural and Functional Diversity of Metabolite-Binding Riboswitches”. In: *Annual Review of Biochemistry* 78.1 (2009). PMID: 19298181, pp. 305–334. DOI: 10.1146/annurev.biochem.78.070507.135656. eprint: https://doi.org/10.1146/annurev.biochem.78.070507.135656. URL: https://doi.org/10.1146/annurev.biochem.78.070507.135656.

[83] Andrew D. Garst, Andrea L. Edwards, and Robert T. Batey. “Riboswitches: Structures and Mechanisms”. In: *Cold Spring Harbor Perspectives in Biology* 3.6 (2011). DOI: 10.1101/cshperspect.a003533. eprint: http://cshperspectives.cshlp.org/content/3/6/a003533.full.pdf+html. URL: http://cshperspectives.cshlp.org/content/3/6/a003533.abstract.

[84] Ronald R. Breaker. “Riboswitches and the RNA World”. In: *Cold Spring Harbor Perspectives in Biology* 4.2 (2012). DOI: 10.1101/cshperspect.a003566. eprint: http://cshperspectives.cshlp.org/content/4/2/a003566.full.pdf+html. URL: http://cshperspectives.cshlp.org/content/4/2/a003566.abstract.

[85] Alexander Serganov and Evgeny Nudler. “A Decade of Riboswitches”. In: *Cell* 152.1 (2013), pp. 17 –24. ISSN: 0092-8674. DOI: https://doi.org/10.

1016/j.cell.2012.12.024. URL: http://www.sciencedirect.com/science/article/pii/S0092867412015462.

[86] Alla Peselis and Alexander Serganov. “Themes and variations in riboswitch structure and function”. In: *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* 1839.10 (2014). Riboswitches, pp. 908 –918. ISSN: 1874-9399. DOI: https://doi.org/10.1016/j.bbagrm.2014.02.012. URL: http://www.sciencedirect.com/science/article/pii/S1874939914000339.

[87] Christian Berens and Beatrix Suess. “Riboswitch engineering — making the all-important second and third steps”. In: *Current Opinion in Biotechnology* 31 (2015). Analytical Biotechnology, pp. 10 –15. ISSN: 0958-1669. DOI: https://doi.org/10.1016/j.copbio.2014.07.014. URL: http://www.sciencedirect.com/science/article/pii/S0958166914001426.

[88] Phillip J. McCown et al. “Riboswitch diversity and distribution”. In: *RNA* 23.7 (2017), pp. 995–1011. DOI: 10.1261/rna.061234.117. eprint: http://rnajournal.cshlp.org/content/23/7/995.full.pdf+html. URL: http://rnajournal.cshlp.org/content/23/7/995.abstract.

[89] Zachary F. Hallberg et al. “Engineering and In Vivo Applications of Riboswitches”. In: *Annual Review of Biochemistry* 86.1 (2017). PMID: 28375743, pp. 515–539. DOI: 10.1146/annurev-biochem-060815-014628. eprint: https:

//doi.org/10.1146/annurev-biochem-060815-014628. URL: https://doi.org/10.1146/annurev-biochem-060815-014628.

[90] Ronald R. Breaker. “Riboswitches and Translation Control”. In: *Cold Spring Harbor Perspectives in Biology* 10.11 (2018). DOI: 10.1101/cshperspect.a032797. eprint: http://cshperspectives.cshlp.org/content/10/11/a032797.full.pdf+html. URL: http://cshperspectives.cshlp.org/content/10/11/a032797.abstract.

[91] Nikolet Pavlova, Dimitrios Kaloudas, and Robert Penchovsky. “Riboswitch distribution, structure, and function in bacteria”. In: *Gene* 708 (2019), pp. 38–48. ISSN: 0378-1119. DOI: https://doi.org/10.1016/j.gene.2019.05.036. URL: https://www.sciencedirect.com/science/article/pii/S0378111919304998.

[92] Anne-Sophie Vézina Bédard, Elsa D.M. Hien, and Daniel A. Lafontaine. “Riboswitch regulation mechanisms: RNA, metabolites and regulatory proteins”. In: *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* 1863.3 (2020), p. 194501. ISSN: 1874-9399. DOI: https://doi.org/10.1016/j.bbagrm.2020.194501. URL: https://www.sciencedirect.com/science/article/pii/S1874939919302834.

[93] A. Di Giorgio and M. Duca. “Synthetic small-molecule RNA ligands: future prospects as therapeutic agents”. In: *Med. Chem. Commun.* 10 (8 2019), pp. 1242–1255. DOI: 10.1039/C9MD00195F. URL: http://dx.doi.org/10.1039/C9MD00195F.

[94] Andrew C. Stelzer et al. "Discovery of selective bioactive small molecules by targeting an RNA dynamic ensemble". In: *Nature Chemical Biology* 7.8 (2011), pp. 553–559. ISSN: 1552-4469. DOI: 10.1038/nchembio.596. URL: https://doi.org/10.1038/nchembio.596.

[95] Laura R. Ganser et al. "High-performance virtual screening by targeting a high-resolution RNA dynamic ensemble". In: *Nature Structural & Molecular Biology* 25.5 (2018), pp. 425–434. ISSN: 1545-9985. DOI: 10.1038/s41594-018-0062-4. URL: https://doi.org/10.1038/s41594-018-0062-4.

[96] Punit P. Seth et al. "SAR by MS: Discovery of a New Class of RNA-Binding Small Molecules for the Hepatitis C Virus: Internal Ribosome Entry Site IIA Subdomain". In: *Journal of Medicinal Chemistry* 48.23 (2005). PMID: 16279767, pp. 7099–7102. DOI: 10.1021/jm050815o. eprint: https://doi.org/10.1021/jm050815o. URL: https://doi.org/10.1021/jm050815o.

[97] Maia Carnevali et al. "A Modular Approach to Synthetic RNA Binders of the Hepatitis C Virus Internal Ribosome Entry Site". In: *ChemBioChem* 11.10 (2010), pp. 1364–1367. DOI: https://doi.org/10.1002/cbic.201000177. eprint: https://chemistry-europe.onlinelibrary.wiley.com/doi/pdf/10.1002/cbic.201000177. URL: https://chemistry-europe.onlinelibrary.wiley.com/doi/abs/10.1002/cbic.201000177.

[98] Ryan B. Paulsen et al. "Inhibitor-induced structural change in the HCV IRES domain IIa RNA". In: *Proceedings of the National Academy of Sciences* 107.16

(2010), pp. 7263–7268. ISSN: 0027-8424. DOI: 10.1073/pnas.0911896107. eprint: https://www.pnas.org/content/107/16/7263.full.pdf. URL: https://www.pnas.org/content/107/16/7263.

[99] Sergey M. Dibrov et al. “Structure of a hepatitis C virus RNA domain in complex with a translation inhibitor reveals a binding mode reminiscent of riboswitches”. In: *Proceedings of the National Academy of Sciences* 109.14 (2012), pp. 5223–5228. ISSN: 0027-8424. DOI: 10.1073/pnas.1118699109. eprint: https://www.pnas.org/content/109/14/5223.full.pdf. URL: https://www.pnas.org/content/109/14/5223.

[100] Sergey M. Dibrov et al. “Hepatitis C Virus Translation Inhibitors Targeting the Internal Ribosomal Entry Site”. In: *Journal of Medicinal Chemistry* 57.5 (2014). PMID: 24138284, pp. 1694–1707. DOI: 10.1021/jm401312n. eprint: https://doi.org/10.1021/jm401312n. URL: https://doi.org/10.1021/jm401312n.

[101] Mi-Kyung Lee et al. “A novel small-molecule binds to the influenza A virus RNA promoter and inhibits viral replication”. In: *Chem. Commun.* 50 (3 2014), pp. 368–370. DOI: 10.1039/C3CC46973E. URL: http://dx.doi.org/10.1039/C3CC46973E.

[102] Angel Bottini et al. “Targeting Influenza A Virus RNA Promoter”. In: *Chemical Biology & Drug Design* 86.4 (2015), pp. 663–673. DOI: https://doi.org/10.1111/cbdd.12534. eprint: https://onlinelibrary.wiley.com/

doi/pdf/10.1111/cbdd.12534. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/cbdd.12534.

[103] Ewan P Plant et al. “A Three-Stemmed mRNA Pseudoknot in the SARS Coronavirus Frameshift Signal”. In: *PLOS Biology* 3.6 (May 2005). DOI: 10.1371/journal.pbio.0030172. URL: https://doi.org/10.1371/journal.pbio.0030172.

[104] Mei-Chi Su et al. “An atypical RNA pseudoknot stimulator and an upstream attenuation signal for -1 ribosomal frameshifting of SARS coronavirus”. In: *Nucleic Acids Research* 33.13 (Jan. 2005), pp. 4265–4275. ISSN: 0305-1048. DOI: 10.1093/nar/gki731. eprint: https://academic.oup.com/nar/article-pdf/33/13/4265/3824138/gki731.pdf. URL: https://doi.org/10.1093/nar/gki731.

[105] So-Jung Park, Yang-Gyun Kim, and Hyun-Ju Park. “Identification of RNA Pseudoknot-Binding Ligand That Inhibits the -1 Ribosomal Frameshifting of SARS-Coronavirus by Structure-Based Virtual Screening”. In: *Journal of the American Chemical Society* 133.26 (2011). PMID: 21591761, pp. 10094–10100. DOI: 10.1021/ja1098325. eprint: https://doi.org/10.1021/ja1098325. URL: https://doi.org/10.1021/ja1098325.

[106] Dustin B. Ritchie et al. “Anti-frameshifting Ligand Reduces the Conformational Plasticity of the SARS Virus Pseudoknot”. In: *Journal of the American Chemical Society* 136.6 (2014). PMID: 24446874, pp. 2196–2199. DOI: 10.1021/

ja410344b. eprint: https://doi.org/10.1021/ja410344b. URL: https://doi.org/10.1021/ja410344b.

[107] Nicolas Moitessier, Eric Westhof, and Stephen Hanessian. “Docking of Aminoglycosides to Hydrated and Flexible RNA”. In: *Journal of Medicinal Chemistry* 49.3 (2006). PMID: 16451068, pp. 1023–1033. DOI: 10.1021/jm0508437. eprint: https://doi.org/10.1021/jm0508437. URL: https://doi.org/10.1021/jm0508437.

[108] Yaozong Li et al. “Accuracy Assessment of Protein-Based Docking Programs against RNA Targets”. In: *Journal of Chemical Information and Modeling* 50.6 (2010), pp. 1134–1146. DOI: 10.1021/ci9004157. eprint: https://doi.org/10.1021/ci9004157. URL: https://doi.org/10.1021/ci9004157.

[109] Lu Chen, George A. Calin, and Shuxing Zhang. “Novel Insights of Structure-Based Modeling for RNA-Targeted Drug Discovery”. In: *Journal of Chemical Information and Modeling* 52.10 (2012). PMID: 22947071, pp. 2741–2753. DOI: 10.1021/ci300320t. eprint: https://doi.org/10.1021/ci300320t. URL: https://doi.org/10.1021/ci300320t.

[110] Jiaying Luo et al. “Challenges and current status of computational methods for docking small molecules to nucleic acids”. In: *European Journal of Medicinal Chemistry* 168 (2019), pp. 414 –425. ISSN: 0223-5234. DOI: https://doi.org/10.1016/j.ejmech.2019.02.046. URL:

http://www.sciencedirect.com/science/article/pii/S0223523419301606.

[111] Christophe Guilbert and Thomas L. James. “Docking to RNA via Root-Mean-Square-Deviation-Driven Energy Minimization with Flexible Ligands and Flexible Targets”. In: *Journal of Chemical Information and Modeling* 48.6 (2008). PMID: 18510306, pp. 1257–1268. DOI: 10.1021/ci8000327. eprint: https://doi.org/10.1021/ci8000327. URL: https://doi.org/10.1021/ci8000327.

[112] P. Therese Lang et al. “DOCK 6: Combining techniques to model RNA–small molecule complexes”. In: *RNA* 15.6 (2009), pp. 1219–1230. DOI: 10.1261/rna.1563609. eprint: http://rnajournal.cshlp.org/content/15/6/1219.full.pdf+html. URL: http://rnajournal.cshlp.org/content/15/6/1219.abstract.

[113] Gareth Jones et al. “Development and validation of a genetic algorithm for flexible docking”. In: *Journal of Molecular Biology* 267.3 (1997), pp. 727 –748. ISSN: 0022-2836. DOI: https://doi.org/10.1006/jmbi.1996.0897. URL: http://www.sciencedirect.com/science/article/pii/S0022283696908979.

[114] Richard A. Friesner et al. “Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy”. In: *Journal of Medicinal Chemistry* 47.7 (2004). PMID: 15027865, pp. 1739–1749. DOI: 10.

1021/jm0306430. eprint: https://doi.org/10.1021/jm0306430. URL: https://doi.org/10.1021/jm0306430.

[115] S. David Morley and Mohammad Afshar. "Validation of an empirical RNA-ligand scoring function for fast flexible docking using RiboDock®". In: *Journal of Computer-Aided Molecular Design* 18.3 (2004), pp. 189–208. ISSN: 1573-4951. DOI: 10.1023/B:JCAM.0000035199.48747.1e. URL: https://doi.org/10.1023/B:JCAM.0000035199.48747.1e.

[116] Ruth Huey et al. "A semiempirical free energy force field with charge-based desolvation". In: *Journal of Computational Chemistry* 28.6 (2007), pp. 1145–1152. DOI: https://doi.org/10.1002/jcc.20634. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jcc.20634. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.20634.

[117] Oleg Trott and Arthur J. Olson. "AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading". In: *Journal of Computational Chemistry* 31.2 (2010), pp. 455–461. DOI: https://doi.org/10.1002/jcc.21334. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jcc.21334. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.21334.

[118] Sergio Ruiz-Carmona et al. "rDock: A Fast, Versatile and Open Source Program for Docking Ligands to Proteins and Nucleic Acids". In: *PLOS Computational Biology*

10.4 (Apr. 2014), pp. 1–7. DOI: 10.1371/journal.pcbi.1003571. URL: https://doi.org/10.1371/journal.pcbi.1003571.

[119] Li-Zhen Sun et al. “RLDOCK: A New Method for Predicting RNA–Ligand Interactions”. In: *Journal of Chemical Theory and Computation* 16.11 (2020). PMID: 33095555, pp. 7173–7183. DOI: 10.1021/acs.jctc.0c00798. eprint: https://doi.org/10.1021/acs.jctc.0c00798. URL: https://doi.org/10.1021/acs.jctc.0c00798.

[120] Yangwei Jiang and Shi-Jie Chen. “RLDOCK method for predicting RNA-small molecule binding modes”. In: *Methods* (2021). ISSN: 1046-2023. DOI: https://doi.org/10.1016/j.ymeth.2021.01.009. URL: https://www.sciencedirect.com/science/article/pii/S1046202321000219.

[121] Dennis M. Krüger et al. “Target Flexibility in RNA-Ligand Docking Modeled by Elastic Potential Grids”. In: *ACS Medicinal Chemistry Letters* 2.7 (2011). PMID: 24900336, pp. 489–493. DOI: 10.1021/ml100217h. eprint: https://doi.org/10.1021/ml100217h. URL: https://doi.org/10.1021/ml100217h.

[122] Holger Gohlke, Manfred Hendlich, and Gerhard Klebe. “Knowledge-based scoring function to predict protein-ligand interactions”. In: *Journal of Molecular Biology* 295.2 (2000), pp. 337 –356. ISSN: 0022-2836. DOI: https://doi.org/10.1006/jmbi.1999.3371. URL: http://www.sciencedirect.com/science/article/pii/S0022283699933715.

[123] Xiaoyu Zhao et al. "An Improved PMF Scoring Function for Universally Predicting the Interactions of a Ligand with Protein, DNA, and RNA". In: *Journal of Chemical Information and Modeling* 48.7 (2008). PMID: 18553962, pp. 1438–1447. DOI: 10.1021/ci7004719. eprint: https://doi.org/10.1021/ci7004719. URL: https://doi.org/10.1021/ci7004719.

[124] Anna Philips et al. "LigandRNA: computational predictor of RNA–ligand interactions". In: *RNA* 19.12 (2013), pp. 1605–1616. DOI: 10.1261/rna.039834.113. eprint: http://rnajournal.cshlp.org/content/19/12/1605.full.pdf+html. URL: http://rnajournal.cshlp.org/content/19/12/1605.abstract.

[125] Zhiqiang Yan and Jin Wang. "SPA-LN: a scoring function of ligand–nucleic acid interactions via optimizing both specificity and affinity". In: *Nucleic Acids Research* 45.12 (Apr. 2017), e110–e110. ISSN: 0305-1048. DOI: 10.1093/nar/gkx255. eprint: https://academic.oup.com/nar/article-pdf/45/12/e110/25366972/gkx255.pdf. URL: https://doi.org/10.1093/nar/gkx255.

[126] Yuyu Feng and Sheng-You Huang. "ITScore-NL: An Iterative Knowledge-Based Scoring Function for Nucleic Acid–Ligand Interactions". In: *Journal of Chemical Information and Modeling* 60.12 (2020). PMID: 33291885, pp. 6698–6708. DOI: 10.1021/acs.jcim.0c00974. eprint: https://doi.org/10.1021/acs.jcim.0c00974. URL: https://doi.org/10.1021/acs.jcim.0c00974.

[127] Filip Stefaniak and Janusz M. Bujnicki. “AnnapuRNA: A scoring function for predicting RNA-small molecule binding poses”. In: *PLOS Computational Biology* 17.2 (Feb. 2021), pp. 1–31. DOI: 10.1371/journal.pcbi.1008309. URL: https://doi.org/10.1371/journal.pcbi.1008309.

[128] Sahil Chhabra, Jingru Xie, and Aaron T. Frank. “RNAPosers: Machine Learning Classifiers for Ribonucleic Acid–Ligand Poses”. In: *The Journal of Physical Chemistry B* 124.22 (2020). PMID: 32427491, pp. 4436–4445. DOI: 10.1021/acs.jpcb.0c02322. eprint: https://doi.org/10.1021/acs.jpcb.0c02322. URL: https://doi.org/10.1021/acs.jpcb.0c02322.

[129] Carlos Oliver et al. “Augmented base pairing networks encode RNA-small molecule binding preferences”. In: *Nucleic Acids Research* 48.14 (July 2020), pp. 7690–7699. ISSN: 0305-1048. DOI: 10.1093/nar/gkaa583. eprint: https://academic.oup.com/nar/article-pdf/48/14/7690/34131321/gkaa583.pdf. URL: https://doi.org/10.1093/nar/gkaa583.

[130] Yuanzhe Zhou, Yangwei Jiang, and Shi-Jie Chen. “RNA-ligand molecular docking: Advances and challenges”. In: *WIREs Computational Molecular Science* (2021), e1571. DOI: https://doi.org/10.1002/wcms.1571. eprint: https://wires.onlinelibrary.wiley.com/doi/pdf/10.1002/wcms.1571. URL: https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/wcms.1571.

[131] J. Warwicker and H.C. Watson. "Calculation of the electric potential in the active site cleft due to α-helix dipoles". In: *Journal of Molecular Biology* 157.4 (1982), pp. 671–679. ISSN: 0022-2836. DOI: https://doi.org/10.1016/0022-2836(82)90505-8. URL: https://www.sciencedirect.com/science/article/pii/0022283682905058.

[132] Nathan A. Baker et al. "Electrostatics of nanosystems: Application to microtubules and the ribosome". In: *Proceedings of the National Academy of Sciences* 98.18 (2001), pp. 10037–10041. ISSN: 0027-8424. DOI: 10.1073/pnas.181342398. eprint: https://www.pnas.org/content/98/18/10037.full.pdf. URL: https://www.pnas.org/content/98/18/10037.

[133] J. Andrew Grant, Barry T. Pickup, and Anthony Nicholls. "A smooth permittivity function for Poisson–Boltzmann solvation methods". In: *Journal of Computational Chemistry* 22.6 (2001), pp. 608–640. DOI: https://doi.org/10.1002/jcc.1032. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jcc.1032. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.1032.

[134] Walter Rocchia et al. "Rapid grid-based construction of the molecular surface and the use of induced surface charge to calculate reaction field energies: Applications to the molecular systems and geometric objects". In: *Journal of Computational Chemistry* 23.1 (2002), pp. 128–137. DOI: https://doi.org/10.1002/jcc.1161. eprint: https://onlinelibrary.wiley.com/doi/pdf/

10.1002/jcc.1161. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.1161.

[135] Antonio Morreale, Rubén Gil-Redondo, and Ángel R. Ortiz. "A new implicit solvent model for protein–ligand docking". In: *Proteins: Structure, Function, and Bioinformatics* 67.3 (2007), pp. 606–616. DOI: https://doi.org/10.1002/prot.21269. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/prot.21269. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.21269.

[136] Pengyu Ren et al. "Biomolecular electrostatics and solvation: a computational perspective". In: *Quarterly Reviews of Biophysics* 45.4 (2012), 427–491. DOI: 10.1017/S003358351200011X.

[137] Changhao Wang et al. "Recent Developments and Applications of the MMPBSA Method". In: *Frontiers in Molecular Biosciences* 4 (2018), p. 87. ISSN: 2296-889X. DOI: 10.3389/fmolb.2017.00087. URL: https://www.frontiersin.org/article/10.3389/fmolb.2017.00087.

[138] W Clark Still et al. "Semianalytical treatment of solvation for molecular mechanics and dynamics". In: *Journal of the American Chemical Society* 112.16 (1990), pp. 6127–6129.

[139] Gregory D. Hawkins, Christopher J. Cramer, and Donald G. Truhlar. "Pairwise solute descreening of solute charges from a dielectric medium". In: *Chemical Physics Letters* 246.1 (1995), pp. 122 –129. ISSN: 0009-2614. DOI: https:

//doi.org/10.1016/0009-2614(95)01082-K. URL: http://www.sciencedirect.com/science/article/pii/000926149501082K.

[140] Gregory D. Hawkins, Christopher J. Cramer, and Donald G. Truhlar. “Parametrized Models of Aqueous Free Energies of Solvation Based on Pairwise Descreening of Solute Atomic Charges from a Dielectric Medium”. In: *The Journal of Physical Chemistry* 100.51 (1996), pp. 19824–19839. DOI: 10.1021/jp961710n. eprint: https://doi.org/10.1021/jp961710n. URL: https://doi.org/10.1021/jp961710n.

[141] Di Qiu et al. “The GB/SA Continuum Model for Solvation. A Fast Analytical Method for the Calculation of Approximate Born Radii”. In: *The Journal of Physical Chemistry A* 101.16 (1997), pp. 3005–3014. DOI: 10.1021/jp961992r. eprint: https://doi.org/10.1021/jp961992r. URL: https://doi.org/10.1021/jp961992r.

[142] Alexey Onufriev, Donald Bashford, and David A. Case. “Modification of the Generalized Born Model Suitable for Macromolecules”. In: *The Journal of Physical Chemistry B* 104.15 (2000), pp. 3712–3720. DOI: 10.1021/jp994072s. eprint: https://doi.org/10.1021/jp994072s. URL: https://doi.org/10.1021/jp994072s.

[143] Vickie Tsui and David A. Case. “Molecular Dynamics Simulations of Nucleic Acids with a Generalized Born Solvation Model”. In: *Journal of the American Chemical Society* 122.11 (2000), pp. 2489–2498. DOI: 10.1021/ja9939385.

eprint: https://doi.org/10.1021/ja9939385. URL: https://doi.org/10.1021/ja9939385.

[144] Michael Feig, Wonpil Im, and Charles L. Brooks. “Implicit solvation based on generalized Born theory in different dielectric environments”. In: *The Journal of Chemical Physics* 120.2 (2004), pp. 903–911. DOI: 10.1063/1.1631258. eprint: https://doi.org/10.1063/1.1631258. URL: https://doi.org/10.1063/1.1631258.

[145] Hao-Yang Liu, Irwin D. Kuntz, and Xiaoqin Zou. “Pairwise GB/SA Scoring Function for Structure-based Drug Design”. In: *The Journal of Physical Chemistry B* 108.17 (2004), pp. 5453–5462. DOI: 10.1021/jp0312518. eprint: https://doi.org/10.1021/jp0312518. URL: https://doi.org/10.1021/jp0312518.

[146] Alexey Onufriev, Donald Bashford, and David A. Case. “Exploring protein native states and large-scale conformational changes with a modified generalized born model”. In: *Proteins: Structure, Function, and Bioinformatics* 55.2 (2004), pp. 383–394. DOI: https://doi.org/10.1002/prot.20033. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/prot.20033. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.20033.

[147] Maciej Wójcikowski, Pedro J. Ballester, and Pawel Siedlecki. “Performance of machine-learning scoring functions in structure-based virtual screen-

ing". In: *Scientific Reports* 7.1 (2017), p. 46710. ISSN: 2045-2322. DOI: 10.1038/srep46710. URL: https://doi.org/10.1038/srep46710.

[148] Xin Yang et al. "Concepts of Artificial Intelligence for Computer-Assisted Drug Discovery". In: *Chemical Reviews* 119.18 (2019). PMID: 31294972, pp. 10520–10594. DOI: 10.1021/acs.chemrev.8b00728. eprint: https://doi.org/10.1021/acs.chemrev.8b00728. URL: https://doi.org/10.1021/acs.chemrev.8b00728.

[149] Jessica Vamathevan et al. "Applications of machine learning in drug discovery and development". In: *Nature Reviews Drug Discovery* 18.6 (2019), pp. 463–477. ISSN: 1474-1784. DOI: 10.1038/s41573-019-0024-5. URL: https://doi.org/10.1038/s41573-019-0024-5.

[150] Duc Duy Nguyen et al. "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. ISSN: 1573-4951. DOI: 10.1007/s10822-018-0146-6. URL: https://doi.org/10.1007/s10822-018-0146-6.

[151] Hongjian Li et al. "Machine-learning scoring functions for structure-based virtual screening". In: *WIREs Computational Molecular Science* 11.1 (2020), e1478. DOI: https://doi.org/10.1002/wcms.1478. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/wcms.1478. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/wcms.1478.

[152] Hongjian Li et al. “Machine-learning scoring functions for structure-based drug lead optimization”. In: *WIREs Computational Molecular Science* 10.5 (2020), e1465. DOI: https://doi.org/10.1002/wcms.1465. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/wcms.1465. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/wcms.1465.

[153] Leyi Wei and Quan Zou. “Recent Progress in Machine Learning-Based Methods for Protein Fold Recognition”. In: *International Journal of Molecular Sciences* 17.12 (2016). ISSN: 1422-0067. DOI: 10.3390/ijms17122118. URL: https://www.mdpi.com/1422-0067/17/12/2118.

[154] Andrew W. Senior et al. “Protein structure prediction using multiple deep neural networks in the 13th Critical Assessment of Protein Structure Prediction (CASP13)”. In: *Proteins: Structure, Function, and Bioinformatics* 87.12 (2019), pp. 1141–1148. DOI: https://doi.org/10.1002/prot.25834. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/prot.25834. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.25834.

[155] Shaun M. Kandathil, Joe G. Greener, and David T. Jones. “Recent developments in deep learning applied to protein structure prediction”. In: *Proteins: Structure, Function, and Bioinformatics* 87.12 (2019), pp. 1179–1189. DOI: https://doi.org/10.1002/prot.25824. eprint: https://onlinelibrary.

wiley.com/doi/pdf/10.1002/prot.25824. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.25824.

[156] Brian Kuhlman and Philip Bradley. “Advances in protein structure prediction and design”. In: *Nature Reviews Molecular Cell Biology* 20.11 (2019), pp. 681–697. ISSN: 1471-0080. DOI: 10.1038/s41580-019-0163-x. URL: https://doi.org/10.1038/s41580-019-0163-x.

[157] Wenhao Gao et al. “Deep Learning in Protein Structural Modeling and Design”. In: *Patterns* 1.9 (2020), p. 100142. ISSN: 2666-3899. DOI: https://doi.org/10.1016/j.patter.2020.100142. URL: http://www.sciencedirect.com/science/article/pii/S2666389920301902.

[158] Frank Noé, Gianni De Fabritiis, and Cecilia Clementi. “Machine learning for protein folding and dynamics”. In: *Current Opinion in Structural Biology* 60 (2020). Folding and Binding - Proteins, pp. 77 –84. ISSN: 0959-440X. DOI: https://doi.org/10.1016/j.sbi.2019.12.005. URL: http://www.sciencedirect.com/science/article/pii/S0959440X19301447.

[159] Sérgio Filipe Sousa, Pedro Alexandrino Fernandes, and Maria João Ramos. “Protein–ligand docking: Current status and future challenges”. In: *Proteins: Structure, Function, and Bioinformatics* 65.1 (2006), pp. 15–26. DOI: https://doi.org/10.1002/prot.21082. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/prot.21082. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.21082.

[160] Sheng-You Huang, Sam Z. Grinter, and Xiaoqin Zou. “Scoring functions and their evaluation methods for protein–ligand docking: recent advances and future directions”. In: *Phys. Chem. Chem. Phys.* 12 (40 2010), pp. 12899–12908. DOI: 10.1039/C0CP00151A. URL: http://dx.doi.org/10.1039/C0CP00151A.

[161] Lucy J Colwell. “Statistical and machine learning approaches to predicting protein–ligand interactions”. In: *Current Opinion in Structural Biology* 49 (2018). Theory and simulation Macromolecular assemblies, pp. 123 –128. ISSN: 0959-440X. DOI: https://doi.org/10.1016/j.sbi.2018.01.006. URL: http://www.sciencedirect.com/science/article/pii/S0959440X17301525.

[162] Jin Li, Ailing Fu, and Le Zhang. “An Overview of Scoring Functions Used for Protein–Ligand Interactions in Molecular Docking”. In: *Interdisciplinary Sciences: Computational Life Sciences* 11.2 (2019), pp. 320–328. ISSN: 1867-1462. DOI: 10.1007/s12539-019-00327-w. URL: https://doi.org/10.1007/s12539-019-00327-w.

[163] Edward J. Merino et al. “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). PMID: 15783204, pp. 4223–4231. DOI: 10.1021/ja043822v. eprint: https://doi.org/10.1021/ja043822v. URL: https://doi.org/10.1021/ja043822v.

[164] Kevin A. Wilkinson, Edward J. Merino, and Kevin M. Weeks. "Selective 2′-hydroxyl acylation analyzed by primer extension (SHAPE): quantitative RNA structure analysis at single nucleotide resolution". In: *Nature Protocols* 1.3 (2006), pp. 1610–1616. ISSN: 1750-2799. DOI: 10.1038/nprot.2006.249. URL: https://doi.org/10.1038/nprot.2006.249.

[165] Katherine E. Deigan et al. "Accurate SHAPE-directed RNA structure determination". In: *Proceedings of the National Academy of Sciences* 106.1 (2009), pp. 97–102. ISSN: 0027-8424. DOI: 10.1073/pnas.0806929106. eprint: https://www.pnas.org/content/106/1/97.full.pdf. URL: https://www.pnas.org/content/106/1/97.

[166] Justin T. Low and Kevin M. Weeks. "SHAPE-directed RNA secondary structure prediction". In: *Methods* 52.2 (2010). RNA: From Sequence to Structure and Dynamics, pp. 150–158. ISSN: 1046-2023. DOI: https://doi.org/10.1016/j.ymeth.2010.06.007. URL: https://www.sciencedirect.com/science/article/pii/S1046202310001611.

[167] Costin M. Gherghe et al. "Strong Correlation between SHAPE Chemistry and the Generalized NMR Order Parameter (S2) in RNA". In: *Journal of the American Chemical Society* 130.37 (2008). PMID: 18710236, pp. 12244–12245. DOI: 10.1021/ja804541s. eprint: https://doi.org/10.1021/ja804541s. URL: https://doi.org/10.1021/ja804541s.

[168] Kevin M Weeks. "Advances in RNA structure analysis by chemical probing". In: *Current Opinion in Structural Biology* 20.3 (2010). Nucleic acids / Sequences and

topology, pp. 295–304. ISSN: 0959-440X. DOI: https://doi.org/10.1016/j.sbi.2010.04.001. URL: https://www.sciencedirect.com/science/article/pii/S0959440X10000667.

[169] Jennifer L. McGinnis et al. “The Mechanisms of RNA SHAPE Chemistry”. In: *Journal of the American Chemical Society* 134.15 (2012). PMID: 22475022, pp. 6617–6624. DOI: 10.1021/ja2104075. eprint: https://doi.org/10.1021/ja2104075. URL: https://doi.org/10.1021/ja2104075.

[170] Ronny Lorenz et al. “SHAPE directed RNA folding”. In: *Bioinformatics* 32.1 (Sept. 2015), pp. 145–147. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btv523. eprint: https://academic.oup.com/bioinformatics/article-pdf/32/1/145/16919548/btv523.pdf. URL: https://doi.org/10.1093/bioinformatics/btv523.

[171] Travis Hurst et al. “Quantitative Understanding of SHAPE Mechanism from RNA Structure and Dynamics Analysis”. In: *The Journal of Physical Chemistry B* 122.18 (2018). PMID: 29659274, pp. 4771–4783. DOI: 10.1021/acs.jpcb.8b00575. eprint: https://doi.org/10.1021/acs.jpcb.8b00575. URL: https://doi.org/10.1021/acs.jpcb.8b00575.

[172] Travis Hurst and Shi-Jie Chen. “Sieving RNA 3D structures with SHAPE and evaluating mechanisms driving sequence-dependent reactivity bias”. In: *The Journal of Physical Chemistry B* 125.4 (2021). PMID: 33497570, pp. 1156–1166. DOI: 10.1021/acs.jpcb.0c11365. eprint: https://doi.org/10.1021/

acs.jpcb.0c11365. URL: https://doi.org/10.1021/acs.jpcb.0c11365.

[173] R Evans et al. “De novo structure prediction with deeplearning based scoring”. In: *Annu Rev Biochem* 77.363-382 (2018), p. 6.

[174] Matt Spencer, Jesse Eickholt, and Jianlin Cheng. “A Deep Learning Network Approach to ¡italic¿ab initio¡/italic¿ Protein Secondary Structure Prediction”. In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 12.1 (2015), pp. 103–112. DOI: 10.1109/TCBB.2014.2343960.

[175] Rhys Heffernan et al. “Improving prediction of secondary structure, local backbone angles and solvent accessible surface area of proteins by iterative deep learning”. In: *Scientific Reports* 5.1 (2015), p. 11476. ISSN: 2045-2322. DOI: 10.1038/srep11476. URL: https://doi.org/10.1038/srep11476.

[176] Sheng Wang et al. “Protein Secondary Structure Prediction Using Deep Convolutional Neural Fields”. In: *Scientific Reports* 6.1 (2016), p. 18962. ISSN: 2045-2322. DOI: 10.1038/srep18962. URL: https://doi.org/10.1038/srep18962.

[177] Jian Zhou and Olga Troyanskaya. “Deep Supervised and Convolutional Generative Stochastic Network for Protein Secondary Structure Prediction”. In: *Proceedings of the 31st International Conference on Machine Learning*. Ed. by Eric P. Xing and Tony Jebara. Vol. 32. Proceedings of Machine Learning Research 1. Bejing, China: PMLR, 2014, pp. 745–753. URL: https://proceedings.mlr.press/v32/zhou14.html.

[178] Sheng Wang et al. "Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model". In: *PLOS Computational Biology* 13.1 (Jan. 2017), pp. 1–34. DOI: 10.1371/journal.pcbi.1005324. URL: https://doi.org/10.1371/journal.pcbi.1005324.

[179] Hongjian Li et al. "Improving AutoDock Vina Using Random Forest: The Growing Accuracy of Binding Affinity Prediction by the Effective Exploitation of Larger Data Sets". In: *Molecular Informatics* 34.2-3 (2015), pp. 115–126. DOI: https://doi.org/10.1002/minf.201400132. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/minf.201400132. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/minf.201400132.

[180] Zixuan Cang and Guo-Wei Wei. "TopologyNet: Topology based deep convolutional and multi-task neural networks for biomolecular property predictions". In: *PLOS Computational Biology* 13.7 (July 2017), pp. 1–27. DOI: 10.1371/journal.pcbi.1005690. URL: https://doi.org/10.1371/journal.pcbi.1005690.

[181] José Jiménez et al. "KDEEP: Protein–Ligand Absolute Binding Affinity Prediction via 3D-Convolutional Neural Networks". In: *Journal of Chemical Information and Modeling* 58.2 (2018). PMID: 29309725, pp. 287–296. DOI: 10.1021/acs.jcim.7b00650. eprint: https://doi.org/10.1021/acs.jcim.7b00650. URL: https://doi.org/10.1021/acs.jcim.7b00650.

[182] Izhar Wallach, Michael Dzamba, and Abraham Heifets. “AtomNet: A Deep Convolutional Neural Network for Bioactivity Prediction in Structure-based Drug Discovery”. In: *CoRR* abs/1510.02855 (2015). arXiv: 1510.02855. URL: http://arxiv.org/abs/1510.02855.

[183] Christof Angermueller et al. “Deep learning for computational biology”. In: *Molecular Systems Biology* 12.7 (2016), p. 878. DOI: https://doi.org/10.15252/msb.20156651. eprint: https://www.embopress.org/doi/pdf/10.15252/msb.20156651. URL: https://www.embopress.org/doi/abs/10.15252/msb.20156651.

[184] Jeffrey Skolnick et al. “Computational biology: deep learning”. In: *Emerging Topics in Life Sciences* 1.3 (Nov. 2017), pp. 257–274. ISSN: 2397-8554. DOI: 10.1042/ETLS20160025. eprint: https://portlandpress.com/emergtoplifesci/article-pdf/1/3/257/481511/etls-2016-0025c.pdf. URL: https://doi.org/10.1042/ETLS20160025.

[185] Yuanzhe Zhou et al. “SHAPER: A Web Server for Fast and Accurate SHAPE Reactivity Prediction”. In: *Frontiers in Molecular Biosciences* 8 (2021), p. 715. ISSN: 2296-889X. DOI: 10.3389/fmolb.2021.721955. URL: https://www.frontiersin.org/article/10.3389/fmolb.2021.721955.

[186] Travis Hurst, Yuanzhe Zhou, and Shi-Jie Chen. “Analytical modeling and deep learning approaches to estimating RNA SHAPE reactivity from 3D structure”. In: *Communications in Information and Systems* 19.3 (2019), pp. 299–319. DOI: https://dx.doi.org/10.4310/CIS.2019.v19.n3.a4. URL:

https://www.intlpress.com/site/pub/pages/journals/items/cis/content/vols/0019/0003/a004/.

[187] Clarence Y. Cheng et al. “RNA structure inference through chemical mapping after accidental or intentional mutations”. In: *Proceedings of the National Academy of Sciences* 114.37 (2017), pp. 9876–9881. ISSN: 0027-8424. DOI: 10.1073/pnas.1619897114. eprint: https://www.pnas.org/content/114/37/9876.full.pdf. URL: https://www.pnas.org/content/114/37/9876.

[188] Sichun Yang et al. “RNA Structure Determination Using SAXS Data”. In: *The Journal of Physical Chemistry B* 114.31 (2010). PMID: 20684627, pp. 10039–10048. DOI: 10.1021/jp1057308. eprint: https://doi.org/10.1021/jp1057308. URL: https://doi.org/10.1021/jp1057308.

[189] Marc Parisien and François Major. “Determining RNA three-dimensional structures using low-resolution data”. In: *Journal of Structural Biology* 179.3 (2012). Structural Bioinformatics, pp. 252–260. ISSN: 1047-8477. DOI: https://doi.org/10.1016/j.jsb.2011.12.024. URL: https://www.sciencedirect.com/science/article/pii/S1047847712000627.

[190] Feng Ding et al. “Three-dimensional RNA structure refinement by hydroxyl radical probing”. In: *Nature Methods* 9.6 (2012), pp. 603–608. ISSN: 1548-7105. DOI: 10.1038/nmeth.1976. URL: https://doi.org/10.1038/nmeth.1976.

[191] Zhen Xia et al. “RNA 3D Structure Prediction by Using a Coarse-Grained Model and Experimental Data”. In: *The Journal of Physical Chemistry B* 117.11 (2013). PMID: 23438338, pp. 3135–3144. DOI: 10.1021/jp400751w. eprint: https://doi.org/10.1021/jp400751w. URL: https://doi.org/10.1021/jp400751w.

[192] Stefanie A. Mortimer and Kevin M. Weeks. “A Fast-Acting Reagent for Accurate Analysis of RNA Secondary and Tertiary Structure by SHAPE Chemistry”. In: *Journal of the American Chemical Society* 129.14 (2007). PMID: 17367143, pp. 4144–4145. DOI: 10.1021/ja0704028. eprint: https://doi.org/10.1021/ja0704028. URL: https://doi.org/10.1021/ja0704028.

[193] Byron Lee et al. “Comparison of SHAPE reagents for mapping RNA structures inside living cells”. In: *RNA* 23.2 (2017), pp. 169–174. DOI: 10.1261/rna.058784.116. eprint: http://rnajournal.cshlp.org/content/23/2/169.full.pdf+html. URL: http://rnajournal.cshlp.org/content/23/2/169.abstract.

[194] Wipapat Kladwang et al. “Understanding the Errors of SHAPE-Directed RNA Structure Modeling”. In: *Biochemistry* 50.37 (2011). PMID: 21842868, pp. 8049–8056. DOI: 10.1021/bi200524n. eprint: https://doi.org/10.1021/bi200524n. URL: https://doi.org/10.1021/bi200524n.

[195] Christine E. Hajdin et al. “Accurate SHAPE-directed RNA secondary structure modeling, including pseudoknots”. In: *Proceedings of the National Academy of Sciences* 110.14 (2013), pp. 5498–5503. ISSN: 0027-8424. DOI: 10.1073/pnas.

1219988110. eprint: https://www.pnas.org/content/110/14/5498.full.pdf. URL: https://www.pnas.org/content/110/14/5498.

[196] Christopher W. Leonard et al. “Principles for Understanding the Accuracy of SHAPE-Directed RNA Structure Modeling”. In: *Biochemistry* 52.4 (2013). PMID: 23316814, pp. 588–595. DOI: 10.1021/bi300755u. eprint: https://doi.org/10.1021/bi300755u. URL: https://doi.org/10.1021/bi300755u.

[197] Douglas H. Turner and David H. Mathews. “NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure”. In: *Nucleic Acids Research* 38.suppl_1 (Oct. 2009), pp. D280–D282. ISSN: 0305-1048. DOI: 10.1093/nar/gkp892. eprint: https://academic.oup.com/nar/article-pdf/38/suppl_1/D280/11217894/gkp892.pdf. URL: https://doi.org/10.1093/nar/gkp892.

[198] Stefanie A. Mortimer and Kevin M. Weeks. “Time-resolved RNA SHAPE chemistry: quantitative RNA structure analysis in one-second snapshots and at single-nucleotide resolution”. In: *Nature Protocols* 4.10 (2009), pp. 1413–1421. ISSN: 1750-2799. DOI: 10.1038/nprot.2009.126. URL: https://doi.org/10.1038/nprot.2009.126.

[199] Wipapat Kladwang et al. “A two-dimensional mutate-and-map strategy for non-coding RNA structure”. In: *Nature Chemistry* 3.12 (2011), pp. 954–962. ISSN:

1755-4349. DOI: 10.1038/nchem.1176. URL: https://doi.org/10.1038/nchem.1176.

[200] Kady-Ann Steen, Greggory M. Rice, and Kevin M. Weeks. "Fingerprinting Noncanonical and Tertiary RNA Structures by Differential SHAPE Reactivity". In: *Journal of the American Chemical Society* 134.32 (2012). PMID: 22852530, pp. 13160–13163. DOI: 10.1021/ja304027m. eprint: https://doi.org/10.1021/ja304027m. URL: https://doi.org/10.1021/ja304027m.

[201] Matthew J. Smola et al. "SHAPE reveals transcript-wide interactions, complex structural domains, and protein interactions across the Xist lncRNA in living cells". In: *Proceedings of the National Academy of Sciences* 113.37 (2016), pp. 10322–10327. ISSN: 0027-8424. DOI: 10.1073/pnas.1600008113. eprint: https://www.pnas.org/content/113/37/10322.full.pdf. URL: https://www.pnas.org/content/113/37/10322.

[202] Kyle E. Watters et al. "Characterizing RNA structures in vitro and in vivo with selective 2′-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq)". In: *Methods* 103 (2016). Advances in RNA Structure Determination, pp. 34–48. ISSN: 1046-2023. DOI: https://doi.org/10.1016/j.ymeth.2016.04.002. URL: https://www.sciencedirect.com/science/article/pii/S104620231630072X.

[203] Rosa Diaz-Toledano, Gloria Lozano, and Encarnacion Martinez-Salas. "In-cell SHAPE uncovers dynamic interactions between the untranslated regions of the foot-and-mouth disease virus RNA". In: *Nucleic Acids Research* 45.3 (Sept.

2016), pp. 1416–1432. ISSN: 0305-1048. DOI: 10.1093/nar/gkw795. eprint: https://academic.oup.com/nar/article-pdf/45/3/1416/16665957/gkw795.pdf. URL: https://doi.org/10.1093/nar/gkw795.

[204] Meghan Zubradt et al. "DMS-MaPseq for genome-wide or targeted RNA structure probing in vivo". In: *Nature Methods* 14.1 (2017), pp. 75–82. ISSN: 1548-7105. DOI: 10.1038/nmeth.4057. URL: https://doi.org/10.1038/nmeth.4057.

[205] Philippe Rocca-Serra et al. "Sharing and archiving nucleic acid structure mapping data". In: *RNA* 17.7 (2011), pp. 1204–1212. DOI: 10.1261/rna.2753211. eprint: http://rnajournal.cshlp.org/content/17/7/1204.full.pdf+html. URL: http://rnajournal.cshlp.org/content/17/7/1204.abstract.

[206] Pablo Cordero, Julius B. Lucks, and Rhiju Das. "An RNA Mapping DataBase for curating RNA structure mapping experiments". In: *Bioinformatics* 28.22 (Sept. 2012), pp. 3006–3008. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bts554. eprint: https://academic.oup.com/bioinformatics/article-pdf/28/22/3006/16909397/bts554.pdf. URL: https://doi.org/10.1093/bioinformatics/bts554.

[207] Stephen F. Altschul et al. "Basic local alignment search tool". In: *Journal of Molecular Biology* 215.3 (1990), pp. 403–410. ISSN: 0022-2836. DOI: https://doi.org/10.1016/S0022-2836(05)80360-2. URL:

https://www.sciencedirect.com/science/article/pii/S0022283605803602.

[208] Helen M. Berman et al. “The Protein Data Bank”. In: *Nucleic Acids Research* 28.1 (Jan. 2000), pp. 235–242. ISSN: 0305-1048. DOI: 10.1093/nar/28.1.235. eprint: https://academic.oup.com/nar/article-pdf/28/1/235/9895144/280235.pdf. URL: https://doi.org/10.1093/nar/28.1.235.

[209] Kévin Darty, Alain Denise, and Yann Ponty. “VARNA: Interactive drawing and editing of the RNA secondary structure”. In: *Bioinformatics* 25.15 (Apr. 2009), pp. 1974–1975. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btp250. eprint: https://academic.oup.com/bioinformatics/article-pdf/25/15/1974/16887892/btp250.pdf. URL: https://doi.org/10.1093/bioinformatics/btp250.

[210] Tomasz Zok et al. “RNApdbee 2.0: multifunctional tool for RNA structure annotation”. In: *Nucleic Acids Research* 46.W1 (Apr. 2018), W30–W35. ISSN: 0305-1048. DOI: 10.1093/nar/gky314. eprint: https://academic.oup.com/nar/article-pdf/46/W1/W30/25110258/gky314.pdf. URL: https://doi.org/10.1093/nar/gky314.

[211] William Humphrey, Andrew Dalke, and Klaus Schulten. “VMD: Visual molecular dynamics”. In: *Journal of Molecular Graphics* 14.1 (1996), pp. 33–38. ISSN: 0263-7855. DOI: https://doi.org/10.1016/0263-7855(96)00018-5.

URL: https://www.sciencedirect.com/science/article/pii/0263785596000185.

[212] Dong Zhang and Shi-Jie Chen. "IsRNA: An Iterative Simulated Reference State Approach to Modeling Correlated Interactions in RNA Folding". In: *Journal of Chemical Theory and Computation* 14.4 (2018). PMID: 29499114, pp. 2230–2239. DOI: 10.1021/acs.jctc.7b01228. eprint: https://doi.org/10.1021/acs.jctc.7b01228. URL: https://doi.org/10.1021/acs.jctc.7b01228.

[213] Zhichao Miao et al. "RNA-Puzzles Round III: 3D RNA structure prediction of five riboswitches and one ribozyme". In: *RNA* 23.5 (2017), pp. 655–672. DOI: 10.1261/rna.060368.116. eprint: http://rnajournal.cshlp.org/content/23/5/655.full.pdf+html. URL: http://rnajournal.cshlp.org/content/23/5/655.abstract.

[214] Huanwang Yang et al. "Tools for the automatic identification and classification of RNA base pairs". In: *Nucleic Acids Research* 31.13 (July 2003), pp. 3450–3460. ISSN: 0305-1048. DOI: 10.1093/nar/gkg529. eprint: https://academic.oup.com/nar/article-pdf/31/13/3450/9487193/gkg529.pdf. URL: https://doi.org/10.1093/nar/gkg529.

[215] Marc Parisien et al. "New metrics for comparing and assessing discrepancies between RNA 3D structures and models". In: *RNA* 15.10 (2009), pp. 1875–1885. DOI: 10.1261/rna.1700409. eprint: http://rnajournal.cshlp.org/

content/15/10/1875.full.pdf+html. URL: http://rnajournal.cshlp.org/content/15/10/1875.abstract.

[216] Kaiming He et al. “Deep Residual Learning for Image Recognition”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016.

[217] Sergey Ioffe and Christian Szegedy. “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift”. In: *Proceedings of the 32nd International Conference on Machine Learning*. Ed. by Francis Bach and David Blei. Vol. 37. Proceedings of Machine Learning Research. Lille, France: PMLR, 2015, pp. 448–456. URL: https://proceedings.mlr.press/v37/ioffe15.html.

[218] Vinod Nair and Geoffrey E. Hinton. “Rectified Linear Units Improve Restricted Boltzmann Machines”. In: *ICML*. 2010, pp. 807–814. URL: https://icml.cc/Conferences/2010/papers/432.pdf.

[219] Kaiming He et al. “Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37.9 (2015), pp. 1904–1916. DOI: 10.1109/TPAMI.2015.2389824.

[220] Diederik P Kingma and Jimmy Ba. “Adam: A method for stochastic optimization”. In: *arXiv preprint arXiv:1412.6980* (2014).

[221] Adam Paszke et al. “Automatic differentiation in pytorch”. In: (2017).

[222] Bruce A. Shapiro et al. "Bridging the gap in RNA structure prediction". In: *Current Opinion in Structural Biology* 17.2 (2007). Theory and simulation / Macromolecular assemblages, pp. 157–165. ISSN: 0959-440X. DOI: https://doi.org/10.1016/j.sbi.2007.03.001. URL: https://www.sciencedirect.com/science/article/pii/S0959440X07000310.

[223] Christian Laing and Tamar Schlick. "Computational approaches to 3D modeling of RNA". In: *Journal of Physics: Condensed Matter* 22.28 (2010), p. 283101. DOI: 10.1088/0953-8984/22/28/283101. URL: https://doi.org/10.1088/0953-8984/22/28/283101.

[224] Zhichao Miao and Eric Westhof. "RNA structure: advances and assessment of 3D structure prediction". In: *Annual Review of Biophysics* 46.1 (2017). PMID: 28375730, pp. 483–503. DOI: 10.1146/annurev-biophys-070816-034125. eprint: https://doi.org/10.1146/annurev-biophys-070816-034125. URL: https://doi.org/10.1146/annurev-biophys-070816-034125.

[225] Buvaneswari Coimbatore Narayanan et al. "The Nucleic Acid Database: new features and capabilities". In: *Nucleic Acids Research* 42.D1 (Oct. 2013), pp. D114–D122. ISSN: 0305-1048. DOI: 10.1093/nar/gkt980. eprint: https://academic.oup.com/nar/article-pdf/42/D1/D114/3696991/gkt980.pdf. URL: https://doi.org/10.1093/nar/gkt980.

[226] Xiang-Jun Lu, Harmen J. Bussemaker, and Wilma K. Olson. "DSSR: an integrated software tool for dissecting the spatial structure of RNA". In: *Nucleic Acids Re-*

search 43.21 (July 2015), e142–e142. ISSN: 0305-1048. DOI: 10.1093/nar/gkv716. eprint: https://academic.oup.com/nar/article-pdf/43/21/e142/17435026/gkv716.pdf. URL: https://doi.org/10.1093/nar/gkv716.

[227] Quentin Vicens et al. “Local RNA structural changes induced by crystallization are revealed by SHAPE”. In: *RNA* 13.4 (2007), pp. 536–548. DOI: 10.1261/rna.400207. eprint: http://rnajournal.cshlp.org/content/13/4/536.full.pdf+html. URL: http://rnajournal.cshlp.org/content/13/4/536.abstract.

[228] Elisa Frezza et al. “The interplay between molecular flexibility and RNA chemical probing reactivities analyzed at the nucleotide level via an extensive molecular dynamics study”. In: *Methods* 162-163 (2019). Experimental and Computational Techniques for Studying Structural Dynamics and Function of RNA, pp. 108–127. ISSN: 1046-2023. DOI: https://doi.org/10.1016/j.ymeth.2019.05.021. URL: https://www.sciencedirect.com/science/article/pii/S104620231830392X.

[229] Fei Deng et al. “Data-directed RNA secondary structure prediction using probabilistic modeling”. In: *RNA* 22.8 (2016), pp. 1109–1119. DOI: 10.1261/rna.055756.115. eprint: http://rnajournal.cshlp.org/content/22/8/1109.full.pdf+html. URL: http://rnajournal.cshlp.org/content/22/8/1109.abstract.

[230] Sana Vaziri, Patrice Koehl, and Sharon Aviran. “Extracting information from RNA SHAPE data: Kalman filtering approach”. In: *PLOS ONE* 13.11 (Nov. 2018), pp. 1–29. DOI: 10.1371/journal.pone.0207029. URL: https://doi.org/10.1371/journal.pone.0207029.

[231] Dong Zhang, Jun Li, and Shi-Jie Chen. “IsRNA1: De Novo Prediction and Blind Screening of RNA 3D Structures”. In: *Journal of Chemical Theory and Computation* 17.3 (2021). PMID: 33560836, pp. 1842–1857. DOI: 10.1021/acs.jctc.0c01148. eprint: https://doi.org/10.1021/acs.jctc.0c01148. URL: https://doi.org/10.1021/acs.jctc.0c01148.

[232] Yuanzhe Zhou and Shi-Jie Chen. “A method for decoding nucleic acid-magnesium ion interactions using a deep learning convolutional neural network”. In: *submitted* (2021).

[233] H. M. Berman et al. “The nucleic acid database. A comprehensive relational database of three-dimensional structures of nucleic acids”. eng. In: *Biophysical journal* 63.3 (1992). 1384741[pmid], pp. 751–759. ISSN: 0006-3495. DOI: 10.1016/S0006-3495(92)81649-1. URL: https://pubmed.ncbi.nlm.nih.gov/1384741.

[234] Christian A. Hanke and Holger Gohlke. “Force Field Dependence of Riboswitch Dynamics”. In: *Computational Methods for Understanding Riboswitches*. Ed. by Shi-Jie Chen and Donald H. Burke-Aguero. Vol. 553. Methods in Enzymology. Academic Press, 2015, pp. 163–191. DOI: https://doi.org/10.1016/

bs.mie.2014.10.056. URL: https://www.sciencedirect.com/science/article/pii/S0076687914000573.

[235] Justin A. Lemkul, Sirish Kaushik Lakkaraju, and Alexander D. MacKerell. “Characterization of Mg2+ Distributions around RNA in Solution”. In: *ACS Omega* 1.4 (2016). PMID: 27819065, pp. 680–688. DOI: 10.1021/acsomega.6b00241. eprint: https://doi.org/10.1021/acsomega.6b00241. URL: https://doi.org/10.1021/acsomega.6b00241.

[236] Lorenzo Casalino et al. “Development of Site-Specific Mg2+–RNA Force Field Parameters: A Dream or Reality? Guidelines from Combined Molecular Dynamics and Quantum Mechanics Simulations”. In: *Journal of Chemical Theory and Computation* 13.1 (2017). PMID: 28001405, pp. 340–352. DOI: 10.1021/acs.jctc.6b00905. eprint: https://doi.org/10.1021/acs.jctc.6b00905. URL: https://doi.org/10.1021/acs.jctc.6b00905.

[237] Bernd N. M. van Buuren et al. “Brownian-dynamics simulations of metal-ion binding to four-way junctions”. In: *Nucleic Acids Research* 30.2 (Jan. 2002), pp. 507–514. ISSN: 0305-1048. DOI: 10.1093/nar/30.2.507. eprint: https://academic.oup.com/nar/article-pdf/30/2/507/9901127/300507.pdf. URL: https://doi.org/10.1093/nar/30.2.507.

[238] Vinod K Misra and David E Draper. “Mg2+ binding to tRNA revisited: the nonlinear poisson-boltzmann model11Edited by B. Honig”. In: *Journal of Molecular Biology* 299.3 (2000), pp. 813–825. ISSN: 0022-2836. DOI: https://doi.org/

10.1006/jmbi.2000.3769. URL: https://www.sciencedirect.com/science/article/pii/S0022283600937690.

[239] Carmen Burkhardt and Martin Zacharias. “Modelling ion binding to AA platform motifs in RNA: a continuum solvent study including conformational adaptation”. In: *Nucleic Acids Research* 29.19 (Oct. 2001), pp. 3910–3918. ISSN: 0305-1048. DOI: 10.1093/nar/29.19.3910. eprint: https://academic.oup.com/nar/article-pdf/29/19/3910/9906087/293910.pdf. URL: https://doi.org/10.1093/nar/29.19.3910.

[240] Zhi-Jie Tan and Shi-Jie Chen. “Electrostatic correlations and fluctuations for ion binding to a finite length polyelectrolyte”. In: *The Journal of Chemical Physics* 122.4 (2005), p. 044903. DOI: 10.1063/1.1842059. eprint: https://doi.org/10.1063/1.1842059. URL: https://doi.org/10.1063/1.1842059.

[241] Li-Zhen Sun and Shi-Jie Chen. “Monte Carlo Tightly Bound Ion Model: Predicting Ion-Binding Properties of RNA with Ion Correlations and Fluctuations”. In: *Journal of Chemical Theory and Computation* 12.7 (2016). PMID: 27311366, pp. 3370–3381. DOI: 10.1021/acs.jctc.6b00028. eprint: https://doi.org/10.1021/acs.jctc.6b00028. URL: https://doi.org/10.1021/acs.jctc.6b00028.

[242] D. Rey Banatao, Russ B. Altman, and Teri E. Klein. “Microenvironment analysis and identification of magnesium binding sites in RNA”. In: *Nucleic Acids Research* 31.15 (Aug. 2003), pp. 4450–4460. ISSN: 0305-1048. DOI: 10.1093/nar/

gkg471. eprint: https://academic.oup.com/nar/article-pdf/31/15/4450/3865113/gkg471.pdf. URL: https://doi.org/10.1093/nar/gkg471.

[243] Anna Philips et al. “MetalionRNA: computational predictor of metal-binding sites in RNA structures”. In: *Bioinformatics* 28.2 (Nov. 2011), pp. 198–205. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btr636. eprint: https://academic.oup.com/bioinformatics/article-pdf/28/2/198/16908676/btr636.pdf. URL: https://doi.org/10.1093/bioinformatics/btr636.

[244] Eric F. Pettersen et al. “UCSF Chimera-A visualization system for exploratory research and analysis”. In: *Journal of Computational Chemistry* 25.13 (2004), pp. 1605–1612. DOI: https://doi.org/10.1002/jcc.20084. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jcc.20084. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.20084.

[245] Martin Ester et al. “A density-based algorithm for discovering clusters in large spatial databases with noise.” In: *kdd*. Vol. 96. 34. 1996, pp. 226–231.

[246] Daniel Smilkov et al. “SmoothGrad: removing noise by adding noise”. In: *CoRR* abs/1706.03825 (2017). arXiv: 1706.03825. URL: http://arxiv.org/abs/1706.03825.

[247] Norbert Jeszenői et al. “Mobility-based prediction of hydration structures of protein surfaces”. In: *Bioinformatics* 31.12 (Feb. 2015), pp. 1959–1965.

ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btv093. eprint: https://academic.oup.com/bioinformatics/article-pdf/31/12/1959/17100695/btv093.pdf. URL: https://doi.org/10.1093/bioinformatics/btv093.

[248] Norbert Jeszenői et al. "Exploration of Interfacial Hydration Networks of Target–Ligand Complexes". In: *Journal of Chemical Information and Modeling* 56.1 (2016). PMID: 26704050, pp. 148–158. DOI: 10.1021/acs.jcim.5b00638. eprint: https://doi.org/10.1021/acs.jcim.5b00638. URL: https://doi.org/10.1021/acs.jcim.5b00638.

[249] Neocles B. Leontis and Craig L. Zirbel. "Nonredundant 3D Structure Datasets for RNA Knowledge Extraction and Benchmarking". In: *RNA 3D Structure Analysis and Prediction*. Ed. by Neocles Leontis and Eric Westhof. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 281–298. ISBN: 978-3-642-25740-7. DOI: 10.1007/978-3-642-25740-7_13. URL: https://doi.org/10.1007/978-3-642-25740-7_13.

[250] S. Doerr et al. "HTMD: High-Throughput Molecular Dynamics for Molecular Discovery". In: *Journal of Chemical Theory and Computation* 12.4 (2016). PMID: 26949976, pp. 1845–1852. DOI: 10.1021/acs.jctc.6b00049. eprint: https://doi.org/10.1021/acs.jctc.6b00049. URL: https://doi.org/10.1021/acs.jctc.6b00049.

[251] Christopher R. Corbeil, Pablo Englebienne, and Nicolas Moitessier. "Docking Ligands into Flexible and Solvated Macromolecules. 1. Development and Validation

of FITTED 1.0". In: *Journal of Chemical Information and Modeling* 47.2 (2007). PMID: 17305329, pp. 435–449. DOI: 10.1021/ci6002637. eprint: https://doi.org/10.1021/ci6002637. URL: https://doi.org/10.1021/ci6002637.

[252] Katyanna S. Bezerra et al. "Ribosomal RNA–Aminoglycoside Hygromycin B Interaction Energy Calculation within a Density Functional Theory Framework". In: *The Journal of Physical Chemistry B* 123.30 (2019). PMID: 31283875, pp. 6421–6429. DOI: 10.1021/acs.jpcb.9b04468. eprint: https://doi.org/10.1021/acs.jpcb.9b04468. URL: https://doi.org/10.1021/acs.jpcb.9b04468.

[253] Masatake Sugita et al. "New Protocol for Predicting the Ligand-Binding Site and Mode Based on the 3D-RISM/KH Theory". In: *Journal of Chemical Theory and Computation* 16.4 (2020). PMID: 32176492, pp. 2864–2876. DOI: 10.1021/acs.jctc.9b01069. eprint: https://doi.org/10.1021/acs.jctc.9b01069. URL: https://doi.org/10.1021/acs.jctc.9b01069.

[254] William M. Hewitt, David R. Calabrese, and John S. Schneekloth. "Evidence for ligandable sites in structured RNA throughout the Protein Data Bank". In: *Bioorganic & Medicinal Chemistry* 27.11 (2019), pp. 2253–2260. ISSN: 0968-0896. DOI: https://doi.org/10.1016/j.bmc.2019.04.010. URL: https://www.sciencedirect.com/science/article/pii/S0968089618321473.

[255] Sai Pradeep Velagapudi, Steven M. Gallo, and Matthew D. Disney. "Sequence-based design of bioactive small molecules that target precursor microRNAs". In: *Nature Chemical Biology* 10.4 (2014), pp. 291–297. ISSN: 1552-4469. DOI: 10.1038/nchembio.1452. URL: https://doi.org/10.1038/nchembio.1452.

[256] Matthew D. Disney et al. "Inforna 2.0: A Platform for the Sequence-Based Design of Small Molecules Targeting Structured RNAs". In: *ACS Chemical Biology* 11.6 (2016). PMID: 27097021, pp. 1720–1728. DOI: 10.1021/acschembio.6b00001. eprint: https://doi.org/10.1021/acschembio.6b00001. URL: https://doi.org/10.1021/acschembio.6b00001.

[257] Sai Pradeep Velagapudi, Steven J. Seedhouse, and Matthew D. Disney. "Structure-Activity Relationships through Sequencing (StARTS) Defines Optimal and Suboptimal RNA Motif Targets for Small Molecules". In: *Angewandte Chemie International Edition* 49.22 (2010), pp. 3816–3818. DOI: https://doi.org/10.1002/anie.200907257. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/anie.200907257. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/anie.200907257.

[258] Sai Pradeep Velagapudi et al. "Defining the RNA Internal Loops Preferred by Benzimidazole Derivatives via 2D Combinatorial Screening and Computational Analysis". In: *Journal of the American Chemical Society* 133.26 (2011). PMID: 21604752, pp. 10111–10118. DOI: 10.1021/ja200212b. eprint: https:

//doi.org/10.1021/ja200212b. URL: https://doi.org/10.1021/ja200212b.

[259] Sai Pradeep Velagapudi et al. "Defining RNA-Small Molecule Affinity Landscapes Enables Design of a Small Molecule Inhibitor of an Oncogenic Noncoding RNA". In: *ACS Central Science* 3.3 (2017). PMID: 28386598, pp. 205–216. DOI: 10.1021/acscentsci.7b00009. eprint: https://doi.org/10.1021/acscentsci.7b00009. URL: https://doi.org/10.1021/acscentsci.7b00009.

[260] Matthew D. Disney et al. "Chapter Three - Identifying and validating small molecules interacting with RNA (SMIRNAs)". In: *RNA Recognition*. Ed. by Amanda E. Hargrove. Vol. 623. Methods in Enzymology. Academic Press, 2019, pp. 45–66. DOI: https://doi.org/10.1016/bs.mie.2019.04.027. URL: https://www.sciencedirect.com/science/article/pii/S0076687919301429.

[261] Peiyuan Zhang et al. "Translation of the intrinsically disordered protein α-synuclein is inhibited by a small molecule targeting its structured mRNA". In: *Proceedings of the National Academy of Sciences* 117.3 (2020), pp. 1457–1467. ISSN: 0027-8424. DOI: 10.1073/pnas.1905057117. eprint: https://www.pnas.org/content/117/3/1457.full.pdf. URL: https://www.pnas.org/content/117/3/1457.

[262] Woojin Scott Kim, Katarina Kågedal, and Glenda M. Halliday. "Alpha-synuclein biology in Lewy body diseases". In: *Alzheimer's Research & Therapy* 6.5 (2014),

p. 73. ISSN: 1758-9193. DOI: 10.1186/s13195-014-0073-2. URL: https://doi.org/10.1186/s13195-014-0073-2.

[263] Thomas D. Schneider et al. "Information content of binding sites on nucleotide sequences". In: *Journal of Molecular Biology* 188.3 (1986), pp. 415–431. ISSN: 0022-2836. DOI: https://doi.org/10.1016/0022-2836(86)90165-8. URL: https://www.sciencedirect.com/science/article/pii/0022283686901658.

[264] James M. Carothers et al. "Informational Complexity and Functional Activity of RNA Structures". In: *Journal of the American Chemical Society* 126.16 (2004). PMID: 15099096, pp. 5130–5137. DOI: 10.1021/ja031504a. eprint: https://doi.org/10.1021/ja031504a. URL: https://doi.org/10.1021/ja031504a.

[265] Jianghong An, Maxim Totrov, and Ruben Abagyan. "Pocketome via Comprehensive Identification and Classification of Ligand Binding Envelopes*". In: *Molecular & Cellular Proteomics* 4.6 (2005), pp. 752–761. ISSN: 1535-9476. DOI: https://doi.org/10.1074/mcp.M400159-MCP200. URL: https://www.sciencedirect.com/science/article/pii/S1535947620314742.

[266] Stefanie A. Mortimer et al. "SHAPE-Seq: High-Throughput RNA Structure Analysis". In: *Current Protocols in Chemical Biology* 4.4 (2012), pp. 275–297. DOI: https://doi.org/10.1002/9780470559277.ch120019. eprint: https://currentprotocols.onlinelibrary.wiley.

com/doi/pdf/10.1002/9780470559277.ch120019. URL: https://currentprotocols.onlinelibrary.wiley.com/doi/abs/10.1002/9780470559277.ch120019.

[267] David Loughrey et al. “SHAPE-Seq 2.0: systematic optimization and extension of high-throughput chemical probing of RNA secondary structure with next generation sequencing”. In: *Nucleic Acids Research* 42.21 (Oct. 2014), e165–e165. ISSN: 0305-1048. DOI: 10.1093/nar/gku909. eprint: https://academic.oup.com/nar/article-pdf/42/21/e165/14122800/gku909.pdf. URL: https://doi.org/10.1093/nar/gku909.

[268] Nathan A. Siegfried et al. “RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP)”. In: *Nature Methods* 11.9 (2014), pp. 959–965. ISSN: 1548-7105. DOI: 10.1038/nmeth.3029. URL: https://doi.org/10.1038/nmeth.3029.

[269] James M. Carothers et al. “Solution structure of an informationally complex high-affinity RNA aptamer to GTP”. In: *RNA* 12.4 (2006), pp. 567–579. DOI: 10.1261/rna.2251306. eprint: http://rnajournal.cshlp.org/content/12/4/567.full.pdf+html. URL: http://rnajournal.cshlp.org/content/12/4/567.abstract.

[270] Sai Pradeep Velagapudi et al. “Design of a small molecule against an oncogenic noncoding RNA”. In: *Proceedings of the National Academy of Sciences* 113.21 (2016), pp. 5898–5903. ISSN: 0027-8424. DOI: 10.1073/pnas.1523975113.

eprint: https://www.pnas.org/content/113/21/5898.full.pdf. URL: https://www.pnas.org/content/113/21/5898.

[271] Irwin D. Kuntz et al. “A geometric approach to macromolecule-ligand interactions”. In: *Journal of Molecular Biology* 161.2 (1982), pp. 269–288. ISSN: 0022-2836. DOI: https://doi.org/10.1016/0022-2836(82)90153-X. URL: https://www.sciencedirect.com/science/article/pii/002228368290153X.

[272] Rodney Harris, Arthur J. Olson, and David S. Goodsell. “Automated prediction of ligand-binding sites in proteins”. In: *Proteins: Structure, Function, and Bioinformatics* 70.4 (2008), pp. 1506–1517. DOI: https://doi.org/10.1002/prot.21645. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/prot.21645. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.21645.

[273] Ruben Abagyan, Maxim Totrov, and Dmitry Kuznetsov. “ICM-A new method for protein modeling and design: Applications to docking and structure prediction from the distorted native conformation”. In: *Journal of Computational Chemistry* 15.5 (1994), pp. 488–506. DOI: https://doi.org/10.1002/jcc.540150503. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jcc.540150503. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.540150503.

[274] Garrett M. Morris et al. “Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function”. In: *Journal of Com-*

putational Chemistry 19.14 (1998), pp. 1639–1662. DOI: https://doi.org/10.1002/(SICI)1096-987X(19981115)19:14<1639::AID-JCC10>3.0.CO;2-B. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/%28SICI%291096-987X%2819981115%2919%3A14%3C1639%3A%3AAID-JCC10%3E3.0.CO%3B2-B. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291096-987X%2819981115%2919%3A14%3C1639%3A%3AAID-JCC10%3E3.0.CO%3B2-B.

[275] Pan Zeng et al. “Rsite: a computational method to identify the functional sites of noncoding RNAs”. In: *Scientific Reports* 5.1 (2015), p. 9179. ISSN: 2045-2322. DOI: 10.1038/srep09179. URL: https://doi.org/10.1038/srep09179.

[276] Pan Zeng and Qinghua Cui. “Rsite2: an efficient computational method to predict the functional sites of noncoding RNAs”. In: *Scientific Reports* 6.1 (2016), p. 19016. ISSN: 2045-2322. DOI: 10.1038/srep19016. URL: https://doi.org/10.1038/srep19016.

[277] Kaili Wang et al. “RBind: computational network method to predict RNA binding sites”. In: *Bioinformatics* 34.18 (Apr. 2018), pp. 3131–3136. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bty345. eprint: https://academic.oup.com/bioinformatics/article-pdf/34/18/3131/25732095/bty345.pdf. URL: https://doi.org/10.1093/bioinformatics/bty345.

[278] Huiwen Wang and Yunjie Zhao. "RBinds: A user-friendly server for RNA binding site prediction". In: *Computational and Structural Biotechnology Journal* 18 (2020), pp. 3762 –3765. ISSN: 2001-0370. DOI: https://doi.org/10.1016/j.csbj.2020.10.043. URL: http://www.sciencedirect.com/science/article/pii/S2001037020304657.

[279] Hong Su, Zhenling Peng, and Jianyi Yang. "Recognition of small molecule–RNA binding sites using RNA sequence and structure". In: *Bioinformatics* (Jan. 2021). btaa1092. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btaa1092. eprint: https://academic.oup.com/bioinformatics/advance-article-pdf/doi/10.1093/bioinformatics/btaa1092/35906225/btaa1092.pdf. URL: https://doi.org/10.1093/bioinformatics/btaa1092.

[280] J Jiménez et al. "DeepSite: protein-binding site predictor using 3D-convolutional neural networks". In: *Bioinformatics* 33.19 (May 2017), pp. 3036–3042. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btx350. eprint: https://academic.oup.com/bioinformatics/article-pdf/33/19/3036/25164841/btx350.pdf. URL: https://doi.org/10.1093/bioinformatics/btx350.

[281] Radoslav Kriválk and David Hoksza. "P2Rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure". In: *Journal of Cheminformatics* 10.1 (2018), p. 39. ISSN: 1758-2946. DOI: 10.1186/

s13321-018-0285-8. URL: https://doi.org/10.1186/s13321-018-0285-8.

[282] Igor Kozlovskii and Petr Popov. “Spatiotemporal identification of druggable binding sites using deep learning”. In: *Communications Biology* 3.1 (2020), p. 618. ISSN: 2399-3642. DOI: 10.1038/s42003-020-01350-0. URL: https://doi.org/10.1038/s42003-020-01350-0.

[283] Jingtian Zhao, Yang Cao, and Le Zhang. “Exploring the computational methods for protein-ligand binding site prediction”. In: *Computational and Structural Biotechnology Journal* 18 (2020), pp. 417–426. ISSN: 2001-0370. DOI: https://doi.org/10.1016/j.csbj.2020.02.008. URL: https://www.sciencedirect.com/science/article/pii/S2001037019304465.

[284] Hashim M Al-Hashimi and Nils G Walter. “RNA dynamics: it is about time”. In: *Current Opinion in Structural Biology* 18.3 (2008). Nucleic acids / Sequences and topology, pp. 321–329. ISSN: 0959-440X. DOI: https://doi.org/10.1016/j.sbi.2008.04.004. URL: https://www.sciencedirect.com/science/article/pii/S0959440X08000602.

[285] Laura R. Ganser et al. “The roles of structural dynamics in the cellular functions of RNAs”. In: *Nature Reviews Molecular Cell Biology* 20.8 (2019), pp. 474–489. ISSN: 1471-0080. DOI: 10.1038/s41580-019-0136-0. URL: https://doi.org/10.1038/s41580-019-0136-0.

[286] Catherine E. Scull et al. “Transcriptional Riboswitches Integrate Timescales for Bacterial Gene Expression Control”. In: *Frontiers in Molecular Biosciences* 7 (2021), p. 480. ISSN: 2296-889X. DOI: 10.3389/fmolb.2020.607158. URL: https://www.frontiersin.org/article/10.3389/fmolb.2020.607158.

[287] Peinan Zhao, Wenbing Zhang, and Shi-Jie Chen. “Cotranscriptional folding kinetics of ribonucleic acid secondary structures”. In: *The Journal of Chemical Physics* 135.24 (2011), p. 245101. DOI: 10.1063/1.3671644. eprint: https://doi.org/10.1063/1.3671644. URL: https://doi.org/10.1063/1.3671644.

[288] Griffin M Schroeder et al. “Analysis of a preQ1-I riboswitch in effector-free and bound states reveals a metabolite-programmed nucleobase-stacking spine that controls gene regulation”. In: *Nucleic Acids Research* 48.14 (June 2020), pp. 8146–8164. ISSN: 0305-1048. DOI: 10.1093/nar/gkaa546. eprint: https://academic.oup.com/nar/article-pdf/48/14/8146/34131194/gkaa546.pdf. URL: https://doi.org/10.1093/nar/gkaa546.

[289] Jermaine L. Jenkins et al. “Comparison of a PreQ1 Riboswitch Aptamer in Metabolite-bound and Free States with Implications for Gene Regulation*”. In: *Journal of Biological Chemistry* 286.28 (2011), pp. 24626–24637. ISSN: 0021-9258. DOI: https://doi.org/10.1074/jbc.M111.230375. URL: https://www.sciencedirect.com/science/article/pii/S0021925819486262.

[290] Boris François et al. “Crystal structures of complexes between aminoglycosides and decoding A site oligonucleotides: role of the number of rings and positive charges in the specific binding leading to miscoding”. In: *Nucleic Acids Research* 33.17 (Jan. 2005), pp. 5677–5690. ISSN: 0305-1048. DOI: 10.1093/nar/gki862. eprint: https://academic.oup.com/nar/article-pdf/33/17/5677/6319035/gki862.pdf. URL: https://doi.org/10.1093/nar/gki862.

[291] Thomas E. Edwards and Adrian R. Ferré-D'Amaré. “Crystal Structures of the Thi-Box Riboswitch Bound to Thiamine Pyrophosphate Analogs Reveal Adaptive RNA-Small Molecule Recognition”. In: *Structure* 14.9 (2006), pp. 1459–1468. ISSN: 0969-2126. DOI: https://doi.org/10.1016/j.str.2006.07.008. URL: https://www.sciencedirect.com/science/article/pii/S0969212606003303.

[292] Jeffrey B.-H. Tok and Lanrong Bi. “Aminoglycoside and its Derivatives as Ligands to Target the Ribosome”. In: *Current Topics in Medicinal Chemistry* 3.9 (2003), pp. 1001–1019. ISSN: 1568-0266/1873-4294. DOI: 10.2174/1568026033452131. URL: https://sci-hub.se/http://www.eurekaselect.com/node/81184/article.

[293] Peter Daldrop et al. “Novel Ligands for a Purine Riboswitch Discovered by RNA-Ligand Docking”. In: *Chemistry & Biology* 18.3 (2011), pp. 324–335. ISSN: 1074-5521. DOI: https://doi.org/10.1016/j.chembiol.2010.12.020.

URL: https://www.sciencedirect.com/science/article/pii/S1074552111000391.

[294] Giel P. Göertz et al. “Functional RNA during Zika virus infection”. In: *Virus Research* 254 (2018). Advances in ZIKA Research, pp. 41–53. ISSN: 0168-1702. DOI: https://doi.org/10.1016/j.virusres.2017.08.015. URL: https://www.sciencedirect.com/science/article/pii/S016817021730521X.

[295] Ramya Rangan et al. “RNA genome conservation and secondary structure in SARS-CoV-2 and SARS-related viruses: a first look”. In: *RNA* 26.8 (2020), pp. 937–959. DOI: 10.1261/rna.076141.120. eprint: http://rnajournal.cshlp.org/content/26/8/937.full.pdf+html. URL: http://rnajournal.cshlp.org/content/26/8/937.abstract.

[296] Tom Halgren. “New Method for Fast and Accurate Binding-site Identification and Analysis”. In: *Chemical Biology & Drug Design* 69.2 (2007), pp. 146–148. DOI: https://doi.org/10.1111/j.1747-0285.2007.00483.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1747-0285.2007.00483.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1747-0285.2007.00483.x.

[297] Renee L. DesJarlais et al. “Using shape complementarity as an initial screen in designing ligands for a receptor binding site of known three-dimensional structure”. In: *Journal of Medicinal Chemistry* 31.4 (1988). PMID: 3127588, pp. 722–729.

DOI: 10.1021/jm00399a006. eprint: https://doi.org/10.1021/jm00399a006. URL: https://doi.org/10.1021/jm00399a006.

[298] Didier Rognan. “Docking Methods for Virtual Screening: Principles and Recent Advances”. In: *Virtual Screening*. John Wiley & Sons, Ltd, 2011. Chap. 6, pp. 153–176. ISBN: 9783527633326. DOI: https://doi.org/10.1002/9783527633326.ch6. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9783527633326.ch6. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/9783527633326.ch6.

[299] Bohdan Waszkowycz, David E. Clark, and Emanuela Gancia. “Outstanding challenges in protein–ligand docking and structure-based virtual screening”. In: *WIREs Computational Molecular Science* 1.2 (2011), pp. 229–259. DOI: https://doi.org/10.1002/wcms.18. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/wcms.18. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/wcms.18.

[300] Thomas Wehler and Ruth Brenk. “Structure-Based Discovery of Small Molecules Binding to RNA”. In: *RNA Therapeutics*. Ed. by Amanda L. Garner. Cham: Springer International Publishing, 2018, pp. 47–77. ISBN: 978-3-319-68091-0. DOI: 10.1007/7355_2016_29. URL: https://doi.org/10.1007/7355_2016_29.

[301] Simon K. Kearsley et al. “Flexibases: A way to enhance the use of molecular docking methods”. In: *Journal of Computer-Aided Molecular Design* 8.5 (1994),

pp. 565–582. ISSN: 1573-4951. DOI: 10.1007/BF00123666. URL: https://doi.org/10.1007/BF00123666.

[302] *OMEGA 4.1.0.0: OpenEye Scientific Software*. Santa Fe, NM. URL: http://www.eyesopen.com.

[303] Paul C. D. Hawkins et al. “Conformer Generation with OMEGA: Algorithm and Validation Using High Quality Structures from the Protein Databank and Cambridge Structural Database”. In: *Journal of Chemical Information and Modeling* 50.4 (2010). PMID: 20235588, pp. 572–584. DOI: 10.1021/ci100031x. eprint: https://doi.org/10.1021/ci100031x. URL: https://doi.org/10.1021/ci100031x.

[304] *RDKit: Open-Source Cheminformatics Software*. URL: http://www.rdkit.org.

[305] Naruki Yoshikawa and Geoffrey R. Hutchison. “Fast, efficient fragment-based coordinate generation for Open Babel”. In: *Journal of Cheminformatics* 11.1 (2019), p. 49. ISSN: 1758-2946. DOI: 10.1186/s13321-019-0372-5. URL: https://doi.org/10.1186/s13321-019-0372-5.

[306] Paul C. D. Hawkins. “Conformation Generation: The State of the Art”. In: *Journal of Chemical Information and Modeling* 57.8 (2017). PMID: 28682617, pp. 1747–1756. DOI: 10.1021/acs.jcim.7b00221. eprint: https://doi.org/10.1021/acs.jcim.7b00221. URL: https://doi.org/10.1021/acs.jcim.7b00221.

[307] David Moreno et al. “DFFR: A New Method for High-Throughput Recalibration of Automatic Force-Fields for Drugs”. In: *Journal of Chemical Theory and Computation* 16.10 (2020). PMID: 32856910, pp. 6598–6608. DOI: 10.1021/acs.jctc.0c00306. eprint: https://doi.org/10.1021/acs.jctc.0c00306. URL: https://doi.org/10.1021/acs.jctc.0c00306.

[308] Sanja Zivanovic et al. “Exploring the Conformational Landscape of Bioactive Small Molecules”. In: *Journal of Chemical Theory and Computation* 16.10 (2020). PMID: 32786895, pp. 6575–6585. DOI: 10.1021/acs.jctc.0c00304. eprint: https://doi.org/10.1021/acs.jctc.0c00304. URL: https://doi.org/10.1021/acs.jctc.0c00304.

[309] Sanja Zivanovic et al. “Bioactive Conformational Ensemble Server and Database. A Public Framework to Speed Up In Silico Drug Discovery”. In: *Journal of Chemical Theory and Computation* 16.10 (2020). PMID: 32786900, pp. 6586–6597. DOI: 10.1021/acs.jctc.0c00305. eprint: https://doi.org/10.1021/acs.jctc.0c00305. URL: https://doi.org/10.1021/acs.jctc.0c00305.

[310] Olga Pikovskaya et al. “Structural principles of nucleoside selectivity in a 2’-deoxyguanosine riboswitch”. In: *Nature Chemical Biology* 7.10 (2011), pp. 748–755. ISSN: 1552-4469. DOI: 10.1038/nchembio.631. URL: https://doi.org/10.1038/nchembio.631.

[311] Florent Barbault et al. “Parametrization of a specific free energy function for automated docking against RNA targets using neural networks”. In: *Chemometrics*

and Intelligent Laboratory Systems 82.1 (2006). Selected Papers from the International Conference on Chemometrics and Bioinformatics in Asia, pp. 269–275. ISSN: 0169-7439. DOI: https://doi.org/10.1016/j.chemolab.2005.05.014. URL: https://www.sciencedirect.com/science/article/pii/S0169743905001280.

[312] Carsten Detering and Gabriele Varani. "Validation of Automated Docking Programs for Docking and Database Screening against RNA Drug Targets". In: *Journal of Medicinal Chemistry* 47.17 (2004). PMID: 15293991, pp. 4188–4201. DOI: 10.1021/jm030650o. eprint: https://doi.org/10.1021/jm030650o. URL: https://doi.org/10.1021/jm030650o.

[313] Cao Tongcheng and Li Tonghua. "A combination of numeric genetic algorithm and tabu search can be applied to molecular docking". In: *Computational Biology and Chemistry* 28.4 (2004), pp. 303–312. ISSN: 1476-9271. DOI: https://doi.org/10.1016/j.compbiolchem.2004.08.002. URL: https://www.sciencedirect.com/science/article/pii/S1476927104000623.

[314] Douglas B. Kitchen et al. "Docking and scoring in virtual screening for drug discovery: methods and applications". In: *Nature Reviews Drug Discovery* 3.11 (2004), pp. 935–949. ISSN: 1474-1784. DOI: 10.1038/nrd1549. URL: https://doi.org/10.1038/nrd1549.

[315] Johannes Flick, Frank Tristram, and Wolfgang Wenzel. "Modeling loop backbone flexibility in receptor-ligand docking simulations". In: *Journal of Computational*

Chemistry 33.31 (2012), pp. 2504–2515. DOI: https://doi.org/10.1002/jcc.23087. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jcc.23087. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.23087.

[316] Maxim Totrov and Ruben Abagyan. “Flexible ligand docking to multiple receptor conformations: a practical alternative”. In: *Current Opinion in Structural Biology* 18.2 (2008). Theory and simulation / Macromolecular assemblages, pp. 178–184. ISSN: 0959-440X. DOI: https://doi.org/10.1016/j.sbi.2008.01.004. URL: https://www.sciencedirect.com/science/article/pii/S0959440X08000080.

[317] Fareed Aboul-ela, Jonathan Karn, and Gabriele Varani. “Structure of HIV-1 TAR RNA in the Absence of Ligands Reveals a Novel Conformation of the Trinucleotide Bulge”. In: *Nucleic Acids Research* 24.20 (Oct. 1996), pp. 3974–3981. ISSN: 0305-1048. DOI: 10.1093/nar/24.20.3974. eprint: https://academic.oup.com/nar/article-pdf/24/20/3974/7064185/24-20-3974.pdf. URL: https://doi.org/10.1093/nar/24.20.3974.

[318] Anna Maria Ferrari et al. “Soft Docking and Multiple Receptor Conformations in Virtual Screening”. In: *Journal of Medicinal Chemistry* 47.21 (2004). PMID: 15456251, pp. 5076–5084. DOI: 10.1021/jm049756p. eprint: https://doi.org/10.1021/jm049756p. URL: https://doi.org/10.1021/jm049756p.

[319] Sheng-You Huang and Xiaoqin Zou. “Ensemble docking of multiple protein structures: Considering protein structural variations in molecular docking”. In: *Proteins: Structure, Function, and Bioinformatics* 66.2 (2007), pp. 399–421. DOI: https://doi.org/10.1002/prot.21214. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/prot.21214. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.21214.

[320] Rommie E. Amaro et al. “Ensemble Docking in Drug Discovery”. In: *Biophysical Journal* 114.10 (2018), pp. 2271–2278. ISSN: 0006-3495. DOI: https://doi.org/10.1016/j.bpj.2018.02.038. URL: https://www.sciencedirect.com/science/article/pii/S0006349518303242.

[321] Tobias Santner et al. “Pseudoknot Preorganization of the PreQ1 Class I Riboswitch”. In: *Journal of the American Chemical Society* 134.29 (2012). PMID: 22775200, pp. 11928–11931. DOI: 10.1021/ja3049964. eprint: https://doi.org/10.1021/ja3049964. URL: https://doi.org/10.1021/ja3049964.

[322] Anke Reining et al. “Three-state mechanism couples ligand and temperature sensing in riboswitches”. In: *Nature* 499.7458 (2013), pp. 355–359. ISSN: 1476-4687. DOI: 10.1038/nature12378. URL: https://doi.org/10.1038/nature12378.

[323] Bo Zhao, Alexandar L. Hansen, and Qi Zhang. "Characterizing Slow Chemical Exchange in Nucleic Acids by Carbon CEST and Low Spin-Lock Field $R_{1\rho}$ NMR Spectroscopy". In: *Journal of the American Chemical Society* 136.1 (2014). PMID: 24299272, pp. 20–23. DOI: 10.1021/ja409835y. eprint: https://doi.org/10.1021/ja409835y. URL: https://doi.org/10.1021/ja409835y.

[324] Wendy D. Cornell et al. "A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules". In: *Journal of the American Chemical Society* 117.19 (1995), pp. 5179–5197. DOI: 10.1021/ja00124a002. eprint: https://doi.org/10.1021/ja00124a002. URL: https://doi.org/10.1021/ja00124a002.

[325] David A. Case et al. "The Amber biomolecular simulation programs". In: *Journal of Computational Chemistry* 26.16 (2005), pp. 1668–1688. DOI: https://doi.org/10.1002/jcc.20290. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jcc.20290. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.20290.

[326] B. R. Brooks et al. "CHARMM: The biomolecular simulation program". In: *Journal of Computational Chemistry* 30.10 (2009), pp. 1545–1614. DOI: https://doi.org/10.1002/jcc.21287. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jcc.21287. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.21287.

[327] Elizabeth J. Denning et al. “Impact of 2’-hydroxyl sampling on the conformational properties of RNA: Update of the CHARMM all-atom additive force field for RNA”. In: *Journal of Computational Chemistry* 32.9 (2011), pp. 1929–1943. DOI: https://doi.org/10.1002/jcc.21777. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jcc.21777. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.21777.

[328] Nathan Schmid et al. “Definition and testing of the GROMOS force-field versions 54A7 and 54B7”. In: *European Biophysics Journal* 40.7 (2011), p. 843. ISSN: 1432-1017. DOI: 10.1007/s00249-011-0700-9. URL: https://doi.org/10.1007/s00249-011-0700-9.

[329] Jiří Šponer et al. “RNA Structural Dynamics As Captured by Molecular Simulations: A Comprehensive Overview”. In: *Chemical Reviews* 118.8 (2018). PMID: 29297679, pp. 4177–4338. DOI: 10.1021/acs.chemrev.7b00427. eprint: https://doi.org/10.1021/acs.chemrev.7b00427. URL: https://doi.org/10.1021/acs.chemrev.7b00427.

[330] Pavel Banáš et al. “Molecular Mechanism of preQ1 Riboswitch Action: A Molecular Dynamics Study”. In: *The Journal of Physical Chemistry B* 116.42 (2012). PMID: 22998634, pp. 12721–12734. DOI: 10.1021/jp309230v. eprint: https://doi.org/10.1021/jp309230v. URL: https://doi.org/10.1021/jp309230v.

[331] Petra Kührová et al. “Improving the Performance of the Amber RNA Force Field by Tuning the Hydrogen-Bonding Interactions”. In: *Journal of Chemical Theory and Computation* 15.5 (2019). PMID: 30896943, pp. 3288–3305. DOI: 10.1021/acs.jctc.8b00955. eprint: https://doi.org/10.1021/acs.jctc.8b00955. URL: https://doi.org/10.1021/acs.jctc.8b00955.

[332] Andrea Haller, Marie F. Soulière, and Ronald Micura. “The Dynamic Nature of RNA as Key to Understanding Riboswitch Mechanisms”. In: *Accounts of Chemical Research* 44.12 (2011). PMID: 21678902, pp. 1339–1348. DOI: 10.1021/ar200035g. eprint: https://doi.org/10.1021/ar200035g. URL: https://doi.org/10.1021/ar200035g.

[333] Monika Sharma, Gopalakrishnan Bulusu, and Abhijit Mitra. “MD simulations of ligand-bound and ligand-free aptamer: Molecular level insights into the binding and switching mechanism of the add A-riboswitch”. In: *RNA* 15.9 (2009), pp. 1673–1692. DOI: 10.1261/rna.1675809. eprint: http://rnajournal.cshlp.org/content/15/9/1673.full.pdf+html. URL: http://rnajournal.cshlp.org/content/15/9/1673.abstract.

[334] Alessandra Villa, Jens Wöhnert, and Gerhard Stock. “Molecular dynamics simulation study of the binding of purine bases to the aptamer domain of the guanine sensing riboswitch”. In: *Nucleic Acids Research* 37.14 (June 2009), pp. 4774–4786. ISSN: 0305-1048. DOI: 10.1093/nar/gkp486. eprint: https://academic.oup.com/nar/article-pdf/37/14/4774/3807305/gkp486.pdf. URL: https://doi.org/10.1093/nar/gkp486.

[335] U. Deva Priyakumar and Alexander D. MacKerell. “Role of the Adenine Ligand on the Stabilization of the Secondary and Tertiary Interactions in the Adenine Riboswitch”. In: *Journal of Molecular Biology* 396.5 (2010), pp. 1422–1438. ISSN: 0022-2836. DOI: https://doi.org/10.1016/j.jmb.2009.12.024. URL: https://www.sciencedirect.com/science/article/pii/S0022283609015290.

[336] Zhou Gong et al. “Role of Ligand Binding in Structural Organization of Add A-riboswitch Aptamer: A Molecular Dynamics Simulation”. In: *Journal of Biomolecular Structure and Dynamics* 29.2 (2011). PMID: 21875158, pp. 403–416. DOI: 10.1080/07391102.2011.10507394. eprint: https://doi.org/10.1080/07391102.2011.10507394. URL: https://doi.org/10.1080/07391102.2011.10507394.

[337] Francesco Di Palma, Francesco Colizzi, and Giovanni Bussi. “Ligand-induced stabilization of the aptamer terminal helix in the add adenine riboswitch”. In: *RNA* 19.11 (2013), pp. 1517–1524. DOI: 10.1261/rna.040493.113. eprint: http://rnajournal.cshlp.org/content/19/11/1517.full.pdf+html. URL: http://rnajournal.cshlp.org/content/19/11/1517.abstract.

[338] Phuong H. Nguyen, Philippe Derreumaux, and Gerhard Stock. “Energy Flow and Long-Range Correlations in Guanine-Binding Riboswitch: A Nonequilibrium Molecular Dynamics Study”. In: *The Journal of Physical Chemistry B* 113.27 (2009). PMID: 19569726, pp. 9340–9347. DOI: 10.1021/jp902013s.

eprint: https://doi.org/10.1021/jp902013s. URL: https://doi.org/10.1021/jp902013s.

[339] Olof Allnér, Lennart Nilsson, and Alessandra Villa. "Loop-loop interaction in an adenine-sensing riboswitch: A molecular dynamics study". In: *RNA* 19.7 (2013), pp. 916–926. DOI: 10.1261/rna.037549.112. eprint: http://rnajournal.cshlp.org/content/19/7/916.full.pdf+html. URL: http://rnajournal.cshlp.org/content/19/7/916.abstract.

[340] Lei Bao, Jun Wang, and Yi Xiao. "Molecular dynamics simulation of the binding process of ligands to the add adenine riboswitch aptamer". In: *Phys. Rev. E* 100 (2 2019), p. 022412. DOI: 10.1103/PhysRevE.100.022412. URL: https://link.aps.org/doi/10.1103/PhysRevE.100.022412.

[341] Irene Gómez Pinto et al. "Discovery of Ligands for a Novel Target, the Human Telomerase RNA, Based on Flexible-Target Virtual Screening and NMR". In: *Journal of Medicinal Chemistry* 51.22 (2008), pp. 7205–7215. DOI: 10.1021/jm800825n. eprint: https://doi.org/10.1021/jm800825n. URL: https://doi.org/10.1021/jm800825n.

[342] Maicol Bissaro, Mattia Sturlese, and Stefano Moro. "Exploring the RNA-Recognition Mechanism Using Supervised Molecular Dynamics (SuMD) Simulations: Toward a Rational Design for Ribonucleic-Targeting Molecules?" In: *Frontiers in Chemistry* 8 (2020), p. 107. ISSN: 2296-2646. DOI: 10.3389/

fchem.2020.00107. URL: https://www.frontiersin.org/article/10.3389/fchem.2020.00107.

[343] Sina Kazemi et al. “Elastic Potential Grids: Accurate and Efficient Representation of Intermolecular Interactions for Fully Flexible Docking”. In: *ChemMedChem* 4.8 (2009), pp. 1264–1268. DOI: https://doi.org/10.1002/cmdc.200900146. eprint: https://chemistry-europe.onlinelibrary.wiley.com/doi/pdf/10.1002/cmdc.200900146. URL: https://chemistry-europe.onlinelibrary.wiley.com/doi/abs/10.1002/cmdc.200900146.

[344] Patrick Pfeffer and Holger Gohlke. “DrugScoreRNAKnowledge-Based Scoring Function To Predict RNA-Ligand Interactions”. In: *Journal of Chemical Information and Modeling* 47.5 (2007). PMID: 17705464, pp. 1868–1876. DOI: 10.1021/ci700134p. eprint: https://doi.org/10.1021/ci700134p. URL: https://doi.org/10.1021/ci700134p.

[345] Zixuan Cang and Guo-Wei Wei. “Integration of element specific persistent homology and machine learning for protein-ligand binding affinity prediction”. In: *International Journal for Numerical Methods in Biomedical Engineering* 34.2 (2018). e2914 cnm.2914, e2914. DOI: https://doi.org/10.1002/cnm.2914. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cnm.2914. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/cnm.2914.

[346] Marco De Vivo and Andrea Cavalli. "Recent advances in dynamic docking for drug discovery". In: *WIREs Computational Molecular Science* 7.6 (2017), e1320. DOI: https://doi.org/10.1002/wcms.1320. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/wcms.1320. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/wcms.1320.

[347] Aravindhan Ganesan, Michelle L. Coote, and Khaled Barakat. "Molecular dynamics-driven drug discovery: leaping forward with confidence". In: *Drug Discovery Today* 22.2 (2017), pp. 249 –269. ISSN: 1359-6446. DOI: https://doi.org/10.1016/j.drudis.2016.11.001. URL: http://www.sciencedirect.com/science/article/pii/S1359644616304147.

[348] Ratna S. Katiyar and Prateek K. Jha. "Molecular simulations in drug delivery: Opportunities and challenges". In: *WIREs Computational Molecular Science* 8.4 (2018), e1358. DOI: https://doi.org/10.1002/wcms.1358. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/wcms.1358. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/wcms.1358.

[349] Xuewei Liu et al. "Molecular dynamics simulations and novel drug discovery". In: *Expert Opinion on Drug Discovery* 13.1 (2018). PMID: 29139324, pp. 23–37. DOI: 10.1080/17460441.2018.1403419. eprint: https://doi.org/10.

1080/17460441.2018.1403419. URL: https://doi.org/10.1080/17460441.2018.1403419.

[350] Sergio Decherchi and Andrea Cavalli. “Thermodynamics and Kinetics of Drug-Target Binding by Molecular Simulation”. In: *Chemical Reviews* 120.23 (2020). PMID: 33006893, pp. 12788–12833. DOI: 10.1021/acs.chemrev.0c00534. eprint: https://doi.org/10.1021/acs.chemrev.0c00534. URL: https://doi.org/10.1021/acs.chemrev.0c00534.

[351] Wonpil Im, Michael S. Lee, and Charles L. Brooks III. “Generalized born model with a simple smoothing function”. In: *Journal of Computational Chemistry* 24.14 (2003), pp. 1691–1702. DOI: https://doi.org/10.1002/jcc.10321. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jcc.10321. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.10321.

[352] Alexander D. MacKerell Jr., Nilesh Banavali, and Nicolas Foloppe. “Development and current status of the CHARMM force field for nucleic acids”. In: *Biopolymers* 56.4 (2000), pp. 257–265. DOI: https://doi.org/10.1002/1097-0282(2000)56:4<257::AID-BIP10029>3.0.CO;2-W. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/1097-0282%282000%2956%3A4%3C257%3A%3AAID-BIP10029%3E3.0.CO%3B2-W. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/1097-0282%282000%2956%3A4%3C257%3A%3AAID-BIP10029%3E3.0.CO%3B2-W.

[353] Junmei Wang et al. “Development and testing of a general amber force field”. In: *Journal of Computational Chemistry* 25.9 (2004), pp. 1157–1174. DOI: https://doi.org/10.1002/jcc.20035. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jcc.20035. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.20035.

[354] Anna Philips, Grzegorz Łach, and Janusz M. Bujnicki. “Chapter Eleven - Computational Methods for Prediction of RNA Interactions with Metal Ions and Small Organic Ligands”. In: *Computational Methods for Understanding Riboswitches*. Ed. by Shi-Jie Chen and Donald H. Burke-Aguero. Vol. 553. Methods in Enzymology. Academic Press, 2015, pp. 261–285. DOI: https://doi.org/10.1016/bs.mie.2014.10.057. URL: https://www.sciencedirect.com/science/article/pii/S0076687914000585.

[355] Ingo Muegge and Yvonne C. Martin. “A General and Fast Scoring Function for Protein-Ligand Interactions: A Simplified Potential Approach”. In: *Journal of Medicinal Chemistry* 42.5 (1999). PMID: 10072678, pp. 791–804. DOI: 10.1021/jm980536j. eprint: https://doi.org/10.1021/jm980536j. URL: https://doi.org/10.1021/jm980536j.

[356] Wijnand T. M. Mooij and Marcel L. Verdonk. “General and targeted statistical potentials for protein–ligand interactions”. In: *Proteins: Structure, Function, and Bioinformatics* 61.2 (2005), pp. 272–287. DOI: https://doi.org/10.1002/prot.20588. eprint: https://onlinelibrary.wiley.com/

doi/pdf/10.1002/prot.20588. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.20588.

[357] Hans F. G. Velec, Holger Gohlke, and Gerhard Klebe. "DrugScore^{CSD} Knowledge-Based Scoring Function Derived from Small Molecule Crystal Data with Superior Recognition Rate of Near-Native Ligand Poses and Better Affinity Prediction". In: *Journal of Medicinal Chemistry* 48.20 (2005). PMID: 16190756, pp. 6296–6303. DOI: 10.1021/jm050436v. eprint: https://doi.org/10.1021/jm050436v. URL: https://doi.org/10.1021/jm050436v.

[358] Chi Zhang et al. "A Knowledge-Based Energy Function for Protein-Ligand, Protein-Protein, and Protein-DNA Complexes". In: *Journal of Medicinal Chemistry* 48.7 (2005). PMID: 15801826, pp. 2325–2335. DOI: 10.1021/jm049314d. eprint: https://doi.org/10.1021/jm049314d. URL: https://doi.org/10.1021/jm049314d.

[359] Sheng-You Huang and Xiaoqin Zou. "An iterative knowledge-based scoring function to predict protein–ligand interactions: I. Derivation of interaction potentials". In: *Journal of Computational Chemistry* 27.15 (2006), pp. 1866–1875. DOI: https://doi.org/10.1002/jcc.20504. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jcc.20504. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.20504.

[360] Sheng-You Huang and Xiaoqin Zou. "An iterative knowledge-based scoring function to predict protein–ligand interactions: II. Validation of the scoring

function". In: *Journal of Computational Chemistry* 27.15 (2006), pp. 1876–1882. DOI: https://doi.org/10.1002/jcc.20505. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jcc.20505. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.20505.

[361] Chao-Yie Yang, Renxiao Wang, and Shaomeng Wang. "M-Score: A Knowledge-Based Potential Scoring Function Accounting for Protein Atom Mobility". In: *Journal of Medicinal Chemistry* 49.20 (2006). PMID: 17004706, pp. 5903–5911. DOI: 10.1021/jm050043w. eprint: https://doi.org/10.1021/jm050043w. URL: https://doi.org/10.1021/jm050043w.

[362] Sheng-You Huang and Xiaoqin Zou. "Inclusion of Solvation and Entropy in the Knowledge-Based Scoring Function for Protein-Ligand Interactions". In: *Journal of Chemical Information and Modeling* 50.2 (2010). PMID: 20088605, pp. 262–273. DOI: 10.1021/ci9002987. eprint: https://doi.org/10.1021/ci9002987. URL: https://doi.org/10.1021/ci9002987.

[363] Gerd Neudert and Gerhard Klebe. "DSX: A Knowledge-Based Scoring Function for the Assessment of Protein–Ligand Complexes". In: *Journal of Chemical Information and Modeling* 51.10 (2011). PMID: 21863864, pp. 2731–2745. DOI: 10.1021/ci200274q. eprint: https://doi.org/10.1021/ci200274q. URL: https://doi.org/10.1021/ci200274q.

[364] Zheng Zheng and Kenneth M. Merz. "Development of the Knowledge-Based and Empirical Combined Scoring Algorithm (KECSA) To Score Protein–Ligand Inter-

actions". In: *Journal of Chemical Information and Modeling* 53.5 (2013). PMID: 23560465, pp. 1073–1083. DOI: 10.1021/ci300619x. eprint: https://doi.org/10.1021/ci300619x. URL: https://doi.org/10.1021/ci300619x.

[365] Sheng-You Huang and Xiaoqin Zou. "A knowledge-based scoring function for protein-RNA interactions derived from a statistical mechanics-based iterative method". In: *Nucleic Acids Research* 42.7 (Jan. 2014), e55–e55. ISSN: 0305-1048. DOI: 10.1093/nar/gku077. URL: https://doi.org/10.1093/nar/gku077.

[366] Pin Chen et al. "DLIGAND2: an improved knowledge-based energy function for protein–ligand interactions using the distance-scaled, finite, ideal-gas reference state". In: *Journal of Cheminformatics* 11.1 (2019), p. 52. ISSN: 1758-2946. DOI: 10.1186/s13321-019-0373-4. URL: https://doi.org/10.1186/s13321-019-0373-4.

[367] Minyi Su et al. "Comparative Assessment of Scoring Functions: The CASF-2016 Update". In: *Journal of Chemical Information and Modeling* 59.2 (2019), pp. 895–913. DOI: 10.1021/acs.jcim.8b00545. eprint: https://doi.org/10.1021/acs.jcim.8b00545. URL: https://doi.org/10.1021/acs.jcim.8b00545.

[368] Paul D. Thomas and Ken A. Dill. "Statistical Potentials Extracted From Protein Structures: How Accurate Are They?" In: *Journal of Molecular Biology* 257.2 (1996), pp. 457 –469. ISSN: 0022-2836. DOI: https://doi.org/10.

1006/jmbi.1996.0175. URL: http://www.sciencedirect.com/science/article/pii/S0022283696901758.

[369] P D Thomas and K A Dill. “An iterative method for extracting energy-like quantities from protein structures”. In: *Proceedings of the National Academy of Sciences* 93.21 (1996), pp. 11628–11633. ISSN: 0027-8424. DOI: 10.1073/pnas.93.21.11628. eprint: https://www.pnas.org/content/93/21/11628.full.pdf. URL: https://www.pnas.org/content/93/21/11628.

[370] Sheng-You Huang and Xiaoqin Zou. “An iterative knowledge-based scoring function for protein–protein recognition”. In: *Proteins: Structure, Function, and Bioinformatics* 72.2 (2008), pp. 557–579. DOI: https://doi.org/10.1002/prot.21949. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/prot.21949. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.21949.

[371] Travis Hurst et al. “A Bayes-inspired theory for optimally building an efficient coarse-grained folding force field”. In: *Communications in Information and Systems* 21.1 (2021), pp. 65–83.

[372] Zhihai Liu et al. “PDB-wide collection of binding data: current status of the PDBbind database”. In: *Bioinformatics* 31.3 (Oct. 2014), pp. 405–412. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btu626. eprint: https://academic.oup.com/bioinformatics/article-pdf/31/3/405/17126045/btu626.pdf. URL: https://doi.org/10.1093/bioinformatics/btu626.

[373] Matthew Ragoza et al. “Protein-Ligand Scoring with Convolutional Neural Networks”. In: *Journal of Chemical Information and Modeling* 57.4 (2017). PMID: 28368587, pp. 942–957. DOI: 10.1021/acs.jcim.6b00740. eprint: https://doi.org/10.1021/acs.jcim.6b00740. URL: https://doi.org/10.1021/acs.jcim.6b00740.

[374] Fergus Imrie et al. “Protein Family-Specific Models Using Deep Neural Networks and Transfer Learning Improve Virtual Screening and Highlight the Need for More Data”. In: *Journal of Chemical Information and Modeling* 58.11 (2018). PMID: 30273487, pp. 2319–2330. DOI: 10.1021/acs.jcim.8b00350. eprint: https://doi.org/10.1021/acs.jcim.8b00350. URL: https://doi.org/10.1021/acs.jcim.8b00350.

[375] Marta M Stepniewska-Dziubinska, Piotr Zielenkiewicz, and Pawel Siedlecki. “Development and evaluation of a deep learning model for protein–ligand binding affinity prediction”. In: *Bioinformatics* 34.21 (May 2018), pp. 3666–3674. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bty374. eprint: https://academic.oup.com/bioinformatics/article-pdf/34/21/3666/26147031/bty374.pdf. URL: https://doi.org/10.1093/bioinformatics/bty374.

[376] Jocelyn Sunseri et al. “Convolutional neural network scoring and minimization in the D3R 2017 community challenge”. In: *Journal of Computer-Aided Molecular Design* 33.1 (2019), pp. 19–34. ISSN: 1573-4951. DOI: 10.1007/s10822-

018-0133-y. URL: https://doi.org/10.1007/s10822-018-0133-y.

[377] Renxiao Wang et al. “The PDBbind Database: Collection of Binding Affinities for Protein-Ligand Complexes with Known Three-Dimensional Structures”. In: *Journal of Medicinal Chemistry* 47.12 (2004). PMID: 15163179, pp. 2977–2980. DOI: 10.1021/jm030580l. eprint: https://doi.org/10.1021/jm030580l. URL: https://doi.org/10.1021/jm030580l.

[378] Renxiao Wang et al. “The PDBbind Database: Methodologies and Updates”. In: *Journal of Medicinal Chemistry* 48.12 (2005). PMID: 15943484, pp. 4111–4119. DOI: 10.1021/jm048957q. eprint: https://doi.org/10.1021/jm048957q. URL: https://doi.org/10.1021/jm048957q.

[379] Tiejun Cheng et al. “Comparative Assessment of Scoring Functions on a Diverse Test Set”. In: *Journal of Chemical Information and Modeling* 49.4 (2009). PMID: 19358517, pp. 1079–1093. DOI: 10.1021/ci9000053. eprint: https://doi.org/10.1021/ci9000053. URL: https://doi.org/10.1021/ci9000053.

[380] Yan Li et al. “Comparative Assessment of Scoring Functions on an Updated Benchmark: 1. Compilation of the Test Set”. In: *Journal of Chemical Information and Modeling* 54.6 (2014). PMID: 24716849, pp. 1700–1716. DOI: 10.1021/ci500080q. eprint: https://doi.org/10.1021/ci500080q. URL: https://doi.org/10.1021/ci500080q.

[381] Zhihai Liu et al. "Forging the Basis for Developing Protein–Ligand Interaction Scoring Functions". In: *Accounts of Chemical Research* 50.2 (2017). PMID: 28182403, pp. 302–309. DOI: 10.1021/acs.accounts.6b00491. eprint: https://doi.org/10.1021/acs.accounts.6b00491. URL: https://doi.org/10.1021/acs.accounts.6b00491.

[382] Michael Schlichtkrull et al. "Modeling Relational Data with Graph Convolutional Networks". In: *The Semantic Web*. Ed. by Aldo Gangemi et al. Cham: Springer International Publishing, 2018, pp. 593–607. ISBN: 978-3-319-93417-4.

[383] Joseph L. Durant et al. "Reoptimization of MDL Keys for Use in Drug Discovery". In: *Journal of Chemical Information and Computer Sciences* 42.6 (2002). PMID: 12444722, pp. 1273–1280. DOI: 10.1021/ci010132r. eprint: https://doi.org/10.1021/ci010132r. URL: https://doi.org/10.1021/ci010132r.

[384] Leo Breiman. "Random forests". In: *Machine learning* 45.1 (2001), pp. 5–32.

[385] Michal J. Boniecki et al. "SimRNA: a coarse-grained method for RNA folding simulations and 3D structure prediction". In: *Nucleic Acids Research* 44.7 (Dec. 2015), e63–e63. ISSN: 0305-1048. DOI: 10.1093/nar/gkv1479. eprint: https://academic.oup.com/nar/article-pdf/44/7/e63/17438019/gkv1479.pdf. URL: https://doi.org/10.1093/nar/gkv1479.

[386] Jonatan Taminau, Gert Thijs, and Hans De Winter. "Pharao: Pharmacophore alignment and optimization". In: *Journal of Molecular Graphics and Modelling* 27.2 (2008), pp. 161–169. ISSN: 1093-3263. DOI: https://doi.org/10.1016/

j.jmgm.2008.04.003. URL: https://www.sciencedirect.com/science/article/pii/S109332630800048X.

[387] Magdalena Gebala et al. "Does Cation Size Affect Occupancy and Electrostatic Screening of the Nucleic Acid Ion Atmosphere?" In: *Journal of the American Chemical Society* 138.34 (2016). PMID: 27479701, pp. 10925–10934. DOI: 10.1021/jacs.6b04289. eprint: https://doi.org/10.1021/jacs.6b04289. URL: https://doi.org/10.1021/jacs.6b04289.

[388] Huan-Xiang Zhou. "Macromolecular electrostatic energy within the nonlinear Poisson-Boltzmann equation". In: *The Journal of Chemical Physics* 100.4 (1994), pp. 3152–3162. DOI: 10.1063/1.466406. eprint: https://doi.org/10.1063/1.466406. URL: https://doi.org/10.1063/1.466406.

[389] Vinod K Misra and David E Draper. "The interpretation of Mg2+ binding isotherms for nucleic acids using Poisson-Boltzmann theory". In: *Journal of Molecular Biology* 294.5 (1999), pp. 1135–1147. ISSN: 0022-2836. DOI: https://doi.org/10.1006/jmbi.1999.3334. URL: https://www.sciencedirect.com/science/article/pii/S002228369993334X.

[390] C. H. Mak and Paul S. Henke. "Ions and RNAs: Free Energies of Counterion-Mediated RNA Fold Stabilities". In: *Journal of Chemical Theory and Computation* 9.1 (2013). PMID: 26589060, pp. 621–639. DOI: 10.1021/ct300760y. eprint: https://doi.org/10.1021/ct300760y. URL: https://doi.org/10.1021/ct300760y.

[391] George M. Giambaşu et al. “Ion Counting from Explicit-Solvent Simulations and 3D-RISM”. In: *Biophysical Journal* 106.4 (2014), pp. 883–894. ISSN: 0006-3495. DOI: https://doi.org/10.1016/j.bpj.2014.01.021. URL: https://www.sciencedirect.com/science/article/pii/S0006349514000915.

[392] Ryan L. Hayes et al. “Generalized Manning Condensation Model Captures the RNA Ion Atmosphere”. In: *Phys. Rev. Lett.* 114 (25 2015), p. 258105. DOI: 10.1103/PhysRevLett.114.258105. URL: https://link.aps.org/doi/10.1103/PhysRevLett.114.258105.

[393] Zhi-Jie Tan and Shi-Jie Chen. “Predicting Electrostatic Forces in RNA Folding”. In: *Biophysical, Chemical, and Functional Probes of RNA Structure, Interactions and Folding: Part B*. Vol. 469. Methods in Enzymology. Academic Press, 2009, pp. 465–487. DOI: https://doi.org/10.1016/S0076-6879(09)69022-4. URL: https://www.sciencedirect.com/science/article/pii/S0076687909690224.

[394] Yuhong Zhu, Zhaojian He, and Shi-Jie Chen. “TBI Server: A Web Server for Predicting Ion Effects in RNA Folding”. In: *PLOS ONE* 10.3 (Mar. 2015), pp. 1–13. DOI: 10.1371/journal.pone.0119705. URL: https://doi.org/10.1371/journal.pone.0119705.

[395] Li-Zhen Sun, Jing-Xiang Zhang, and Shi-Jie Chen. “MCTBI: a web server for predicting metal ion effects in RNA structures”. In: *RNA* 23.8 (2017), pp. 1155–1165. DOI: 10.1261/rna.060947.117. eprint: http://rnajournal.

cshlp.org/content/23/8/1155.full.pdf+html. URL: http://rnajournal.cshlp.org/content/23/8/1155.abstract.

[396] George M. Giambaşu, David A. Case, and Darrin M. York. "Predicting Site-Binding Modes of Ions and Water to Nucleic Acids Using Molecular Solvation Theory". In: *Journal of the American Chemical Society* 141.6 (2019), pp. 2435–2445. DOI: 10.1021/jacs.8b11474. eprint: https://doi.org/10.1021/jacs.8b11474. URL: https://doi.org/10.1021/jacs.8b11474.

[397] Wanlei Wei et al. "Predicting Positions of Bridging Water Molecules in Nucleic Acid–Ligand Complexes". In: *Journal of Chemical Information and Modeling* 59.6 (2019). PMID: 30998377, pp. 2941–2951. DOI: 10.1021/acs.jcim.9b00163. eprint: https://doi.org/10.1021/acs.jcim.9b00163. URL: https://doi.org/10.1021/acs.jcim.9b00163.

[398] Vittorio Limongelli. "Ligand binding free energy and kinetics calculation in 2020". In: *WIREs Computational Molecular Science* 10.4 (2020), e1455. DOI: https://doi.org/10.1002/wcms.1455. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/wcms.1455. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/wcms.1455.

[399] Bernd Kuhn and Peter A. Kollman. "Binding of a Diverse Set of Ligands to Avidin and Streptavidin: An Accurate Quantitative Prediction of Their Relative Affinities by a Combination of Molecular Mechanics and Continuum Solvent Models". In:

Journal of Medicinal Chemistry 43.20 (2000). PMID: 11020294, pp. 3786–3791. DOI: 10.1021/jm000241h. eprint: https://doi.org/10.1021/jm000241h. URL: https://doi.org/10.1021/jm000241h.

[400] Hiroaki Gouda et al. "Free energy calculations for theophylline binding to an RNA aptamer: Comparison of MM-PBSA and thermodynamic integration methods". In: *Biopolymers* 68.1 (2003), pp. 16–34. DOI: https://doi.org/10.1002/bip.10270. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/bip.10270. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/bip.10270.

[401] James B. Murray et al. "Interactions of Designer Antibiotics and the Bacterial Ribosomal Aminoacyl-tRNA Site". In: *Chemistry & Biology* 13.2 (2006), pp. 129–138. ISSN: 1074-5521. DOI: https://doi.org/10.1016/j.chembiol.2005.11.004. URL: https://www.sciencedirect.com/science/article/pii/S1074552105003753.

[402] Samy O. Meroueh and Shahriar Mobashery. "Conformational Transition in the Aminoacyl t-RNA Site of the Bacterial Ribosome both in the Presence and Absence of an Aminoglycoside Antibiotic". In: *Chemical Biology & Drug Design* 69.5 (2007), pp. 291–297. DOI: https://doi.org/10.1111/j.1747-0285.2007.00505.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1747-0285.2007.00505.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1747-0285.2007.00505.x.

[403] Holly Freedman et al. “Explicitly Solvated Ligand Contribution to Continuum Solvation Models for Binding Free Energies: Selectivity of Theophylline Binding to an RNA Aptamer”. In: *The Journal of Physical Chemistry B* 114.6 (2010). PMID: 20099932, pp. 2227–2237. DOI: 10.1021/jp9059664. eprint: https://doi.org/10.1021/jp9059664. URL: https://doi.org/10.1021/jp9059664.

[404] Guodong Hu, Aijing Ma, and Jihua Wang. “Ligand Selectivity Mechanism and Conformational Changes in Guanine Riboswitch by Molecular Dynamics Simulations and Free Energy Calculations”. In: *Journal of Chemical Information and Modeling* 57.4 (2017). PMID: 28345904, pp. 918–928. DOI: 10.1021/acs.jcim.7b00139. eprint: https://doi.org/10.1021/acs.jcim.7b00139. URL: https://doi.org/10.1021/acs.jcim.7b00139.

[405] Saikiran Reddy Peddi, Sree Kanth Sivan, and Vijjulatha Manga. “Molecular dynamics and MM/GBSA-integrated protocol probing the correlation between biological activities and binding free energies of HIV-1 TAR RNA inhibitors”. In: *Journal of Biomolecular Structure and Dynamics* 36.2 (2018). PMID: 28081678, pp. 486–503. DOI: 10.1080/07391102.2017.1281762. eprint: https://doi.org/10.1080/07391102.2017.1281762. URL: https://doi.org/10.1080/07391102.2017.1281762.

[406] Fu Chen et al. “Assessing the performance of MM/PBSA and MM/GBSA methods. 8. Predicting binding free energies and poses of protein–RNA complexes”. In: *RNA* 24.9 (2018), pp. 1183–1194. DOI: 10.1261/rna.065896.118. eprint: http:

//rnajournal.cshlp.org/content/24/9/1183.full.pdf+html. URL: http://rnajournal.cshlp.org/content/24/9/1183.abstract.

[407] Johan Åqvist, Carmen Medina, and Jan-Erik Samuelsson. "A new method for predicting binding affinity in computer-aided drug design". In: *Protein Engineering, Design and Selection* 7.3 (Mar. 1994), pp. 385–391. ISSN: 1741-0126. DOI: 10.1093/protein/7.3.385. eprint: https://academic.oup.com/peds/article-pdf/7/3/385/4286250/7-3-385.pdf. URL: https://doi.org/10.1093/protein/7.3.385.

[408] Johan Åqvist, Victor B. Luzhkov, and Bjørn O. Brandsdal. "Ligand Binding Affinities from MD Simulations". In: *Accounts of Chemical Research* 35.6 (2002). PMID: 12069620, pp. 358–365. DOI: 10.1021/ar010014p. eprint: https://doi.org/10.1021/ar010014p. URL: https://doi.org/10.1021/ar010014p.

[409] Johan Sund, Christoffer Lind, and Johan Åqvist. "Binding Site Preorganization and Ligand Discrimination in the Purine Riboswitch". In: *The Journal of Physical Chemistry B* 119.3 (2015). PMID: 25014157, pp. 773–782. DOI: 10.1021/jp5052358. eprint: https://doi.org/10.1021/jp5052358. URL: https://doi.org/10.1021/jp5052358.

[410] Robert W. Zwanzig. "High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases". In: *The Journal of Chemical Physics* 22.8 (1954),

pp. 1420–1426. DOI: 10.1063/1.1740409. eprint: https://doi.org/10.1063/1.1740409. URL: https://doi.org/10.1063/1.1740409.

[411] John G. Kirkwood. “Statistical Mechanics of Fluid Mixtures”. In: *The Journal of Chemical Physics* 3.5 (1935), pp. 300–313. DOI: 10.1063/1.1749657. eprint: https://doi.org/10.1063/1.1749657. URL: https://doi.org/10.1063/1.1749657.

[412] Ulf Ryde and Pär Söderhjelm. “Ligand-Binding Affinity Estimates Supported by Quantum-Mechanical Methods”. In: *Chemical Reviews* 116.9 (2016). PMID: 27077817, pp. 5520–5566. DOI: 10.1021/acs.chemrev.5b00630. eprint: https://doi.org/10.1021/acs.chemrev.5b00630. URL: https://doi.org/10.1021/acs.chemrev.5b00630.

[413] Michael A. Collins and Ryan P. A. Bettens. “Energy-Based Molecular Fragmentation Methods”. In: *Chemical Reviews* 115.12 (2015). PMID: 25843427, pp. 5607–5642. DOI: 10.1021/cr500455b. eprint: https://doi.org/10.1021/cr500455b. URL: https://doi.org/10.1021/cr500455b.

[414] Krishnan Raghavachari and Arjun Saha. “Accurate Composite and Fragment-Based Quantum Chemical Models for Large Molecules”. In: *Chemical Reviews* 115.12 (2015). PMID: 25849163, pp. 5643–5677. DOI: 10.1021/cr500606e. eprint: https://doi.org/10.1021/cr500606e. URL: https://doi.org/10.1021/cr500606e.

[415] Xinsheng Jin, John Z. H. Zhang, and Xiao He. “Full QM Calculation of RNA Energy Using Electrostatically Embedded Generalized Molecular Fractionation with

Conjugate Caps Method”. In: *The Journal of Physical Chemistry A* 121.12 (2017). PMID: 28264557, pp. 2503–2514. DOI: 10.1021/acs.jpca.7b00859. eprint: https://doi.org/10.1021/acs.jpca.7b00859. URL: https://doi.org/10.1021/acs.jpca.7b00859.

[416] E.L. Albuquerque et al. “DNA-based nanobiostructured devices: The role of quasiperiodicity and correlation effects”. In: *Physics Reports* 535.4 (2014). DNA-based nanobiostructured devices: The role of quasiperiodicity and correlation effects, pp. 139 –209. ISSN: 0370-1573. DOI: https://doi.org/10.1016/j.physrep.2013.10.004. URL: http://www.sciencedirect.com/science/article/pii/S0370157313003797.

[417] Xi H. Chen and John Z. H. Zhang. “Theoretical method for full ab initio calculation of DNA/RNA–ligand interaction energy”. In: *The Journal of Chemical Physics* 120.24 (2004), pp. 11386–11391. DOI: 10.1063/1.1737295. eprint: https://doi.org/10.1063/1.1737295. URL: https://doi.org/10.1063/1.1737295.

[418] Vojtěch Mlýnský et al. “Comparison of ab Initio, DFT, and Semiempirical QM/MM Approaches for Description of Catalytic Mechanism of Hairpin Ribozyme”. In: *Journal of Chemical Theory and Computation* 10.4 (2014). PMID: 26580373, pp. 1608–1622. DOI: 10.1021/ct401015e. eprint: https://doi.org/10.1021/ct401015e. URL: https://doi.org/10.1021/ct401015e.

[419] Roner F. da Costa et al. “Explaining statin inhibition effectiveness of HMG-CoA reductase by quantum biochemistry computations”. In: *Phys. Chem. Chem. Phys.*

14 (4 2012), pp. 1389–1398. DOI: 10.1039/C1CP22824B. URL: http://dx.doi.org/10.1039/C1CP22824B.

[420] K.B. Mota et al. "A quantum biochemistry model of the interaction between the estrogen receptor and the two antagonists used in breast cancer treatment". In: *Computational and Theoretical Chemistry* 1089 (2016), pp. 21 –27. ISSN: 2210-271X. DOI: https://doi.org/10.1016/j.comptc.2016.05.006. URL: http://www.sciencedirect.com/science/article/pii/S2210271X16301773.

[421] B.G. de Sousa et al. "Molecular modelling and quantum biochemistry computations of a naturally occurring bioremediation enzyme: Alkane hydroxylase from Pseudomonas putida P1". In: *Journal of Molecular Graphics and Modelling* 77 (2017), pp. 232 –239. ISSN: 1093-3263. DOI: https://doi.org/10.1016/j.jmgm.2017.08.021. URL: http://www.sciencedirect.com/science/article/pii/S1093326317304588.

[422] Andriy Kovalenko and Fumio Hirata. "Three-dimensional density profiles of water in contact with a solute of arbitrary shape: a RISM approach". In: *Chemical Physics Letters* 290.1 (1998), pp. 237 –244. ISSN: 0009-2614. DOI: https://doi.org/10.1016/S0009-2614(98)00471-0. URL: http://www.sciencedirect.com/science/article/pii/S0009261498004710.

[423] Andriy Kovalenko and Fumio Hirata. "Self-consistent description of a metal–water interface by the Kohn–Sham density functional theory and the three-dimensional

reference interaction site model". In: *The Journal of Chemical Physics* 110.20 (1999), pp. 10095–10112. DOI: 10.1063/1.478883. eprint: https://doi.org/10.1063/1.478883. URL: https://doi.org/10.1063/1.478883.

[424] Andriy Kovalenko and Fumio Hirata. "Hydration free energy of hydrophobic solutes studied by a reference interaction site model with a repulsive bridge correction and a thermodynamic perturbation method". In: *The Journal of Chemical Physics* 113.7 (2000), pp. 2793–2805. DOI: 10.1063/1.1305885. eprint: https://doi.org/10.1063/1.1305885. URL: https://doi.org/10.1063/1.1305885.

[425] Norio Yoshida et al. "Selective Ion-Binding by Protein Probed with the 3D-RISM Theory". In: *Journal of the American Chemical Society* 128.37 (2006). PMID: 16967934, pp. 12042–12043. DOI: 10.1021/ja0633262. eprint: https://doi.org/10.1021/ja0633262. URL: https://doi.org/10.1021/ja0633262.

[426] Takashi Imai et al. "Three-Dimensional Distribution Function Theory for the Prediction of Protein-Ligand Binding Sites and Affinities: Application to the Binding of Noble Gases to Hen Egg-White Lysozyme in Aqueous Solution". In: *The Journal of Physical Chemistry B* 111.39 (2007). PMID: 17824692, pp. 11585–11591. DOI: 10.1021/jp074865b. eprint: https://doi.org/10.1021/jp074865b. URL: https://doi.org/10.1021/jp074865b.

[427] Takashi Imai et al. “Ligand Mapping on Protein Surfaces by the 3D-RISM Theory: Toward Computational Fragment-Based Drug Design”. In: *Journal of the American Chemical Society* 131.34 (2009). PMID: 19655800, pp. 12430–12440. DOI: 10.1021/ja905029t. eprint: https://doi.org/10.1021/ja905029t. URL: https://doi.org/10.1021/ja905029t.

[428] Takashi Imai et al. “Functionality Mapping on Internal Surfaces of Multidrug Transporter AcrB Based on Molecular Theory of Solvation: Implications for Drug Efflux Pathway”. In: *The Journal of Physical Chemistry B* 115.25 (2011). PMID: 21526784, pp. 8288–8295. DOI: 10.1021/jp2015758. eprint: https://doi.org/10.1021/jp2015758. URL: https://doi.org/10.1021/jp2015758.

[429] Yasuomi Kiyota, Norio Yoshida, and Fumio Hirata. “A New Approach for Investigating the Molecular Recognition of Protein: Toward Structure-Based Drug Design Based on the 3D-RISM Theory”. In: *Journal of Chemical Theory and Computation* 7.11 (2011). PMID: 26598271, pp. 3803–3815. DOI: 10.1021/ct200358h. eprint: https://doi.org/10.1021/ct200358h. URL: https://doi.org/10.1021/ct200358h.

[430] Dragan Nikolić et al. “3D-RISM-Dock: A New Fragment-Based Drug Design Protocol”. In: *Journal of Chemical Theory and Computation* 8.9 (2012). PMID: 26605742, pp. 3356–3372. DOI: 10.1021/ct300257v. eprint: https://doi.org/10.1021/ct300257v. URL: https://doi.org/10.1021/ct300257v.

[431] Peter J. Tummino and Robert A. Copeland. "Residence Time of Receptor-Ligand Complexes and Its Effect on Biological Function". In: *Biochemistry* 47.20 (2008). PMID: 18412369, pp. 5481–5492. DOI: 10.1021/bi8002023. eprint: https://doi.org/10.1021/bi8002023. URL: https://doi.org/10.1021/bi8002023.

[432] Hao Lu and Peter J Tonge. "Drug-target residence time: critical information for lead optimization". In: *Current Opinion in Chemical Biology* 14.4 (2010). Next Generation Therapeutics, pp. 467 –474. ISSN: 1367-5931. DOI: https://doi.org/10.1016/j.cbpa.2010.06.176. URL: http://www.sciencedirect.com/science/article/pii/S1367593110000785.

[433] Rumin Zhang and Frederick Monsma. "The importance of drug-target residence time." In: *Current opinion in drug discovery & development* 12.4 (2009), p. 488.

[434] Robert A. Copeland. "The drug-target residence time model: a 10-year retrospective". In: *Nature Reviews Drug Discovery* 15.2 (2016), pp. 87–95. ISSN: 1474-1784. DOI: 10.1038/nrd.2015.18. URL: https://doi.org/10.1038/nrd.2015.18.

[435] Doris A. Schuetz et al. "Kinetics for Drug Discovery: an industry-driven effort to target drug residence time". In: *Drug Discovery Today* 22.6 (2017), pp. 896 – 911. ISSN: 1359-6446. DOI: https://doi.org/10.1016/j.drudis.2017.02.002. URL: http://www.sciencedirect.com/science/article/pii/S1359644617300958.

[436] Robert A. Copeland, David L. Pompliano, and Thomas D. Meek. “Drug-target residence time and its implications for lead optimization”. In: *Nature Reviews Drug Discovery* 5.9 (2006), pp. 730–739. ISSN: 1474-1784. DOI: 10.1038/nrd2082. URL: https://doi.org/10.1038/nrd2082.

[437] David C. Swinney. “Applications of Binding Kinetics to Drug Discovery”. In: *Pharmaceutical Medicine* 22.1 (2008), pp. 23–34. ISSN: 1179-1993. DOI: 10.1007/BF03256679. URL: https://doi.org/10.1007/BF03256679.

[438] David C Swinney. “The role of binding kinetics in therapeutically useful drug action.” In: *Current opinion in drug discovery & development* 12.1 (2009), p. 31.

[439] Albert C. Pan et al. “Molecular determinants of drug–receptor binding kinetics”. In: *Drug Discovery Today* 18.13 (2013), pp. 667 –673. ISSN: 1359-6446. DOI: https://doi.org/10.1016/j.drudis.2013.02.007. URL: http://www.sciencedirect.com/science/article/pii/S1359644613000627.

[440] Ning Yin, Jianfeng Pei, and Luhua Lai. “A comprehensive analysis of the influence of drug binding kinetics on drug action at molecular and systems levels”. In: *Mol. BioSyst.* 9 (6 2013), pp. 1381–1389. DOI: 10.1039/C3MB25471B. URL: http://dx.doi.org/10.1039/C3MB25471B.

[441] M. Bernetti, A. Cavalli, and L. Mollica. “Protein-ligand (un)binding kinetics as a new paradigm for drug discovery at the crossroad between experiments and modelling”. In: *Med. Chem. Commun.* 8 (3 2017), pp. 534–550. DOI: 10.1039/C6MD00581K. URL: http://dx.doi.org/10.1039/C6MD00581K.

[442] Mattia Bernetti et al. "Kinetics of Drug Binding and Residence Time". In: *Annual Review of Physical Chemistry* 70.1 (2019). PMID: 30786217, pp. 143–171. DOI: 10.1146/annurev-physchem-042018-052340. eprint: https://doi.org/10.1146/annurev-physchem-042018-052340. URL: https://doi.org/10.1146/annurev-physchem-042018-052340.

[443] Aline Umuhire Juru, Neeraj N. Patwardhan, and Amanda E. Hargrove. "Understanding the Contributions of Conformational Changes, Thermodynamics, and Kinetics of RNA–Small Molecule Interactions". In: *ACS Chemical Biology* 14.5 (2019), pp. 824–838. DOI: 10.1021/acschembio.8b00945. eprint: https://doi.org/10.1021/acschembio.8b00945. URL: https://doi.org/10.1021/acschembio.8b00945.

[444] John Moult et al. "A large-scale experiment to assess protein structure prediction methods". In: *Proteins: Structure, Function, and Bioinformatics* 23.3 (1995), pp. ii–iv. DOI: https://doi.org/10.1002/prot.340230303. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/prot.340230303. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.340230303.

[445] Symon Gathiaka et al. "D3R grand challenge 2015: Evaluation of protein–ligand pose and affinity predictions". In: *Journal of Computer-Aided Molecular Design* 30.9 (2016), pp. 651–668. ISSN: 1573-4951. DOI: 10.1007/s10822-016-9946-8. URL: https://doi.org/10.1007/s10822-016-9946-8.

[446] Yan Li et al. “Comparative Assessment of Scoring Functions on an Updated Benchmark: 2. Evaluation Methods and General Results”. In: *Journal of Chemical Information and Modeling* 54.6 (2014). PMID: 24708446, pp. 1717–1736. DOI: 10.1021/ci500081m. eprint: https://doi.org/10.1021/ci500081m. URL: https://doi.org/10.1021/ci500081m.

[447] Yan Li et al. “Assessing protein–ligand interaction scoring functions with the CASF-2013 benchmark”. In: *Nature Protocols* 13.4 (2018), pp. 666–680. ISSN: 1750-2799. DOI: 10.1038/nprot.2017.114. URL: https://doi.org/10.1038/nprot.2017.114.

[448] Niu Huang, Brian K. Shoichet, and John J. Irwin. “Benchmarking Sets for Molecular Docking”. In: *Journal of Medicinal Chemistry* 49.23 (2006). PMID: 17154509, pp. 6789–6801. DOI: 10.1021/jm0608356. eprint: https://doi.org/10.1021/jm0608356. URL: https://doi.org/10.1021/jm0608356.

[449] Michael M. Mysinger et al. “Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking”. In: *Journal of Medicinal Chemistry* 55.14 (2012). PMID: 22716043, pp. 6582–6594. DOI: 10.1021/jm300687e. eprint: https://doi.org/10.1021/jm300687e. URL: https://doi.org/10.1021/jm300687e.

[450] John S. Mattick. “RNA regulation: a new genetics?” In: *Nature Reviews Genetics* 5.4 (2004), pp. 316–323. ISSN: 1471-0064. DOI: 10.1038/nrg1321. URL: https://doi.org/10.1038/nrg1321.

[451] Elizabeth A. Dethoff et al. "Functional complexity and regulation through RNA dynamics". In: *Nature* 482.7385 (2012), pp. 322–330. ISSN: 1476-4687. DOI: 10.1038/nature10885. URL: https://doi.org/10.1038/nature10885.

[452] Joshua L. Payne, Fahad Khalid, and Andreas Wagner. "RNA-mediated gene regulation is less evolvable than transcriptional regulation". In: *Proceedings of the National Academy of Sciences* 115.15 (2018), E3481–E3490. ISSN: 0027-8424. DOI: 10.1073/pnas.1719138115. eprint: https://www.pnas.org/content/115/15/E3481.full.pdf. URL: https://www.pnas.org/content/115/15/E3481.

[453] M. Nirenberg et al. "The RNA Code and Protein Synthesis". In: *Cold Spring Harbor Symposia on Quantitative Biology* 31 (1966), pp. 11–24. DOI: 10.1101/SQB.1966.031.01.008. eprint: http://symposium.cshlp.org/content/31/11.full.pdf+html. URL: http://symposium.cshlp.org/content/31/11.short.

[454] Gideon Dreyfuss, V. Narry Kim, and Naoyuki Kataoka. "Messenger-RNA-binding proteins and the messages they carry". In: *Nature Reviews Molecular Cell Biology* 3.3 (2002), pp. 195–205. ISSN: 1471-0080. DOI: 10.1038/nrm760. URL: https://doi.org/10.1038/nrm760.

[455] Kevin V. Morris and John S. Mattick. "The rise of regulatory RNA". In: *Nature Reviews Genetics* 15.6 (2014), pp. 423–437. ISSN: 1471-0064. DOI: 10.1038/nrg3722. URL: https://doi.org/10.1038/nrg3722.

[456] Masayuki Matsui and David R. Corey. “Non-coding RNAs as drug targets”. In: *Nature Reviews Drug Discovery* 16.3 (2017), pp. 167–179. ISSN: 1474-1784. DOI: 10.1038/nrd.2016.117. URL: https://doi.org/10.1038/nrd.2016.117.

[457] Emile N. Van Meter, Jackline A. Onyango, and Kelly A. Teske. “A review of currently identified small molecule modulators of microRNA function”. In: *European Journal of Medicinal Chemistry* 188 (2020), p. 112008. ISSN: 0223-5234. DOI: https://doi.org/10.1016/j.ejmech.2019.112008. URL: https://www.sciencedirect.com/science/article/pii/S0223523419311651.

[458] Katherine E. Deigan and Adrian R. FerrÉ-D’AmarÉ. “Riboswitches: Discovery of Drugs That Target Bacterial Gene-Regulatory RNAs”. In: *Accounts of Chemical Research* 44.12 (2011). PMID: 21615107, pp. 1329–1338. DOI: 10.1021/ar200039b. eprint: https://doi.org/10.1021/ar200039b. URL: https://doi.org/10.1021/ar200039b.

[459] Robert T. Batey. “Structure and mechanism of purine-binding riboswitches”. In: *Quarterly Reviews of Biophysics* 45.3 (2012), 345–381. DOI: 10.1017/S0033583512000078.

[460] Jane N. Kim and Ronald R. Breaker. “Purine sensing by riboswitches”. In: *Biology of the Cell* 100.1 (2008), pp. 1–11. DOI: https://doi.org/10.1042/BC20070088. eprint: https://onlinelibrary.wiley.com/doi/

pdf/10.1042/BC20070088. URL: https://onlinelibrary.wiley.com/doi/abs/10.1042/BC20070088.

[461] Maumita Mandal and Ronald R. Breaker. “Adenine riboswitches and gene activation by disruption of a transcription terminator”. In: *Nature Structural & Molecular Biology* 11.1 (2004), pp. 29–35. ISSN: 1545-9985. DOI: 10.1038/nsmb710. URL: https://doi.org/10.1038/nsmb710.

[462] Vipul Panchal and Ruth Brenk. “Riboswitches as Drug Targets for Antibiotics”. In: *Antibiotics* 10.1 (2021). ISSN: 2079-6382. DOI: 10.3390/antibiotics10010045. URL: https://www.mdpi.com/2079-6382/10/1/45.

[463] Frank Walter, Quentin Vicens, and Eric Westhof. “Aminoglycoside–RNA interactions”. In: *Current Opinion in Chemical Biology* 3.6 (1999), pp. 694–704. ISSN: 1367-5931. DOI: https://doi.org/10.1016/S1367-5931(99)00028-9. URL: https://www.sciencedirect.com/science/article/pii/S1367593199000289.

[464] Chi-Huey Wong et al. “Specificity of aminoglycoside antibiotics for the A-site of the decoding region of ribosomal RNA”. In: *Chemistry & Biology* 5.7 (1998), pp. 397–406. ISSN: 1074-5521. DOI: https://doi.org/10.1016/S1074-5521(98)90073-4. URL: https://www.sciencedirect.com/science/article/pii/S1074552198900734.

[465] Jeyakumar Kandasamy et al. “Increased Selectivity toward Cytoplasmic versus Mitochondrial Ribosome Confers Improved Efficiency of Synthetic Aminoglycosides

in Fixing Damaged Genes: A Strategy for Treatment of Genetic Diseases Caused by Nonsense Mutations". In: *Journal of Medicinal Chemistry* 55.23 (2012). PMID: 23148581, pp. 10630–10643. DOI: 10.1021/jm3012992. eprint: https://doi.org/10.1021/jm3012992. URL: https://doi.org/10.1021/jm3012992.

[466] Narayana Murthy Sabbavarapu et al. "Exploring eukaryotic versus prokaryotic ribosomal RNA recognition with aminoglycoside derivatives". In: *Med. Chem. Commun.* 9 (3 2018), pp. 503–508. DOI: 10.1039/C8MD00001H. URL: http://dx.doi.org/10.1039/C8MD00001H.

[467] Isabella A. Guedes, Felipe S. S. Pereira, and Laurent E. Dardenne. "Empirical Scoring Functions for Structure-Based Virtual Screening: Applications, Critical Aspects, and Challenges". In: *Frontiers in Pharmacology* 9 (2018), p. 1089. ISSN: 1663-9812. DOI: 10.3389/fphar.2018.01089. URL: https://www.frontiersin.org/article/10.3389/fphar.2018.01089.

[468] Andrea L. Edwards, Andrew D. Garst, and Robert T. Batey. "Determining Structures of RNA Aptamers and Riboswitches by X-Ray Crystallography". In: *Nucleic Acid and Peptide Aptamers: Methods and Protocols*. Ed. by Günter Mayer. Totowa, NJ: Humana Press, 2009, pp. 135–163. ISBN: 978-1-59745-557-2. DOI: 10.1007/978-1-59745-557-2_9. URL: https://doi.org/10.1007/978-1-59745-557-2_9.

[469] Huaqun Zhang and Sarah C. Keane. "Advances that facilitate the study of large RNA structure and dynamics by nuclear magnetic resonance spectroscopy". In:

WIREs RNA 10.5 (2019), e1541. DOI: https://doi.org/10.1002/wrna.1541. eprint: https://wires.onlinelibrary.wiley.com/doi/pdf/10.1002/wrna.1541. URL: https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/wrna.1541.

[470] Kaiming Zhang et al. "Cryo-EM structure of a 40 kDa SAM-IV riboswitch RNA at 3.7 Å resolution". In: *Nature Communications* 10.1 (2019), p. 5511. ISSN: 2041-1723. DOI: 10.1038/s41467-019-13494-7. URL: https://doi.org/10.1038/s41467-019-13494-7.

[471] Matthew Clark, Richard D. Cramer III, and Nicole Van Opdenbosch. "Validation of the general purpose tripos 5.2 force field". In: *Journal of Computational Chemistry* 10.8 (1989), pp. 982–1012. DOI: https://doi.org/10.1002/jcc.540100804. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jcc.540100804. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.540100804.

[472] Jacqueline Schmitt et al. "Predictive model-based quality inspection using Machine Learning and Edge Cloud Computing". In: *Advanced Engineering Informatics* 45 (2020), p. 101101. ISSN: 1474-0346. DOI: https://doi.org/10.1016/j.aei.2020.101101. URL: https://www.sciencedirect.com/science/article/pii/S1474034620300707.

[473] Liang Liu and Shi-Jie Chen. "Coarse-Grained Prediction of RNA Loop Structures". In: *PLOS ONE* 7.11 (Nov. 2012), pp. 1–15. DOI: 10.1371/journal.

pone.0048460. URL: https://doi.org/10.1371/journal.pone.0048460.

VITA

Yuanzhe Zhou was born on March 12, 1994, in Xinmi, Henan, People's Republic of China. He has earned B.S. degrees in Physics at Southern University of Science and Technology (SUSTech) in 2016. He then went on to pursue his PhD degree in Computational Biological Physics at the University of Missouri–Columbia and began research with Dr. Shi-Jie Chen. He earned his M.S. degree in Physics at University of Missouri–Columbia in 2019, advised by Dr. Shi-Jie Chen.