Using Machine Learning to Better Predict the Structure of RNA and RNA Containing Complexes

by

Sahil Chhabra

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

Doctoral Committee:

Assistant Professor Aaron Terrence Frank, Chair
Assistant Professor Sarah Keane
Professor Ayyalusamy Ramamoorthy
Associate Professor Ambuj Tewari

Sahil Chhabra

itssahil@umich.edu

ORCID iD: 0000-0002-0602-3743

DEDICATION

This dissertation is dedicated to my mom and dad, who have been my source of inspiration and gave me strength, who continually provide me with their moral, spiritual, emotional and financial support.

ACKNOWLEDGEMENTS

Firstly, I would like to express my sincere gratitude to my advisor Dr. Aaron Frank for the continuous support of my doctoral study and related research, for his patience, motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis.

Besides my advisor, I would like to thank the rest of my thesis committee: Dr. Sarah Keane, Dr. Ayyalusamy Ramamoorthy and Dr. Ambuj Tewari for their insightful comments and encouragement, but also for the hard question which incented me to widen my research from various perspectives.

My sincere thanks also goes to Dr. Florence Tama and Dr. Osamu Miyashita, who provided me an opportunity to join their team as intern, and who gave access to the laboratory and research facilities. Their precious support and knowledge helped me grew as a research scientist in various ways.

TABLE OF CONTENTS

CHAPTER

LIST OF FIGURES

Figure

LIST OF TABLES

Table

ABSTRACT

Determining the structure of RNA in the presence of drug like molecules is a crucial step in any drug development campaign. Standard experimental approaches are expensive and time-consuming, and current state-of-the-art computational methods are too inaccurate to be useful. In principle, computer docking can be used to predict the 3D structure of RNA-ligand complexes. However the scoring functions which are accompanied by the available docking programs for pose ranking of RNA-ligand complexes miss-classify native like poses among a set of decoy poses. As such, there is a need for the development of fast, easy, and precise prediction methods for determining the 3D structure of RNAs. In theory, nuclear magnetic resonance (NMR) spectroscopy derived chemical shifts contain information about the local chemical environment at each site in a molecule and so can be a source of rich structural information. In this work, the goal is to predict the structure of RNA-ligand complexes using NMR chemical shifts. To that end, we explore the effect of different machine learning algorithms and ring current models to accurately predict the chemical shifts for standard RNA-ligand complexes. Extra-Randomized trees machine learning algorithms and Pople ring current model were found to be the most accurate ones at predicting the chemical shifts of RNA-ligand complexes.

Next we explored the use of chemical shifts to guide the 3D structure prediction of RNA-ligand complexes starting from RNA sequence. We applied CS-Fold, an in-house method which utilizes chemical shifts to guide the secondary structure pre-

diction of RNAs. From the best predicted secondary structures using CS-Fold, we generated de novo 3D models of RNAs using the Fragment Assembly of RNA with Full Atom Refinement (FARFAR) approach. We used chemical shifts predicted by LarmorD to refine those 3D structures. We found that CS-Fold (the CS-guided secondary structure prediction approach) combined with Rosetta de novo protocol for 3D motifs prediction significantly enhanced the recovery rates to 50% compared to 20% obtained by the RNAStructure and Rosetta combination. Next we used rDock to dock the ligand from the 10 best predicted 3D structures of the RNA and filter the poses based on the chemical shift errors. This study motivated us to build machine learning models based on a molecular fingerprinting approach that can recover native-like RNA-ligand structures from non-native ones in a decoy set as described below.

Next, we describe RNAPoser, a computational tool that estimate the relative "nativeness" of a set of RNA-ligand poses using machine learning pose classifiers. We trained our pose classifiers on molecular "fingerprints" that were a fusion of atomic fingerprints. These fingerprints encode the local "RNA environment" around ligand atoms. Using the classification scores from our RNAPoser classifiers and ranking the poses based on those scores, we found that the recovery of native like poses is significantly better than those obtained from just using the raw rdock docking scores. We also performed a leave-one-out validation approach and found that RNAPoser could recover ∼80% of the poses that were within 2.5 Å of the native poses, in 88 RNA-ligand complexes we explored. Likewise, on a validation set of 17 complexes, we could recover poses in ∼70% of the complexes. RNAPosers could be used as a tool to help in RNA-ligand pose prediction and hence we make it available to the academic community via https://github.com/atfrank/RNAPosers.

CHAPTER I

Introduction

1.1 Biological Context

1.1.1 Nucleic Acid Basics

Nucleic acids are the building blocks of life because of their role in the storage and transmission of genetic information as well as expression of that genetic information into proteins. There are two closely related types of nucleic acids, namely, deoxyribonucleic acid (DNA) and ribonucleic acid (RNA). While DNA is vital for storing the genetic instructions and its transfer to new offspring(s) for all living organisms and for the development of cells, RNA molecules act as a substrate for the ribosome mediated decoding of genetic instructions into proteins and regulation of gene expression.

The building blocks of RNA molecules are knows as nucleotides which consist of 3 components: a sugar, a base and a phosphate group. The sugar is a five-membered ring, where C1' connects to one of the bases and C5' connects to the phosphate group. The bases are adenine (A), cytosine (C), guanine (G), and uracil (U). The nucleic bases, purines (A and G) and pyrimidines (C and U) are aromatic heterocarbon rings and their sequence in a polynucleotide chain is the primary structure of RNAs. A nucleotide has six backbone torsion angles $(\alpha, \beta, \gamma, \delta, \epsilon, \zeta)$ within the sugar-phosphate backbone and one glycosidic torsion angle around the covalent bond connecting sugar and base moiety. All these structural features including the sugar and torsion angles constitute the structural parameters of a RNA three-dimensional(3D) structure.

RNA structure is very diverse due to the presence of wide range of conformational flexibility within the structural components. These structural components comprises of Primary component which is the nucleic base sequence, Secondary component which includes the base pairing information of the nucleic basis on top of their primary structure and Tertiary component which is the arrangement of the secondary structure in the three dimension space. The spatial arrangements of these secondary structural elements bring variations in the 3D structure of RNA molecules. Interestingly, the diversity in RNA 3D structure comes from the fact that RNA 3D structure is folded primarily from a single strand, rather than two complementary strands like DNA. Since the single stranded sequence is typically not self complementary unlike DNAs, only a few specific stretches of sequence can fold back on themselves to form double helical regions with Watson-Crick (WC) base pairs (A:T and G:C base-pairs) in the secondary structure. The possibility of formation of non-Watson-Crick G:U base pairing may also promote double helical structures. The bases which cannot form base pairing of any kind (unpaired bases) provide relatively unrestricted conformational contribution to the RNA 3D structure.

1.1.2 RNA Structure-Function Relationship

In the central dogma of molecular biology, depicted in Figure 1.1, ribonucleic acid (RNA) is the carrier of information from deoxyribonucleic acid (DNA) to proteins. From the structural point of view, RNA is very similar to DNA with three major differences: (1) RNA has nucleobase uracil (U) whereas DNA has thymine (T), (2) RNA has ribose sugar whereas DNA has deoxyribose sugar (which lacks 2'-hydroxyl group) and, (3) RNAs are single stranded where DNAs are double stranded. Apart

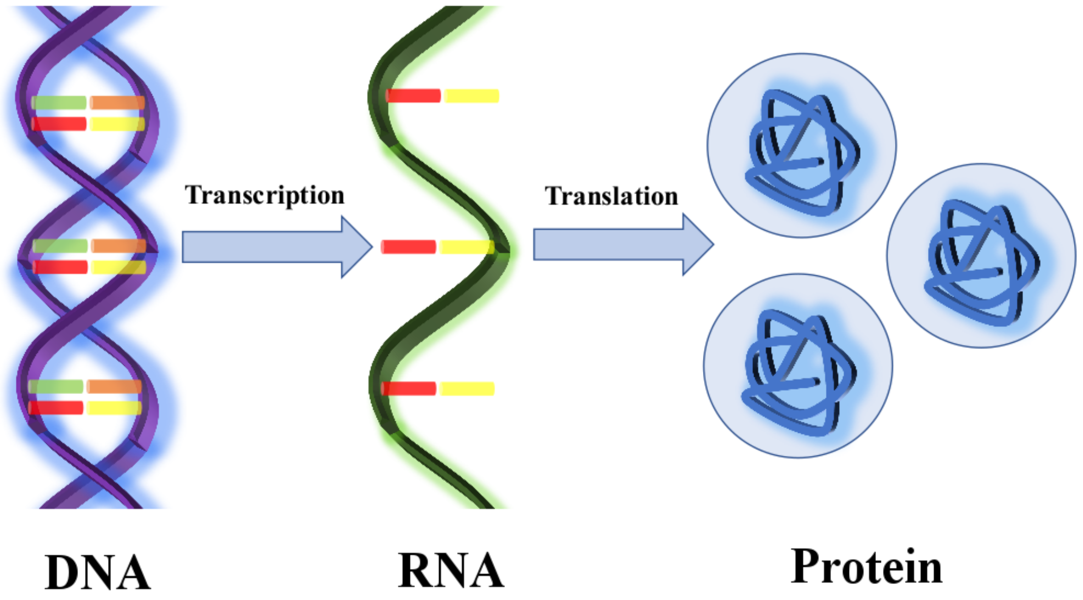


Figure 1.1: Central Dogma of Molecular Biology

from the primary process of translating information from DNA to proteins (a process commonly known as translation), RNAs are also involved in other gene regulatory processes such as transcription initiation and post-transcriptional modifications.[1–5] In general, RNAs are of two types: coding RNAs and non-coding RNAs. Coding RNA sequences (like mRNA) are transcribed from genes and translated into proteins. RNA molecules in the form of ribosomal RNA (rRNA), messenger RNA (mRNA) and transfer RNA (tRNA) work together within the ribosomal protein synthetic machinery to produce proteins. Other forms of RNA molecules, namely micro RNAs (miRNAs), small nuclear RNAs (snRNAs), small nucleolar RNAs (snoRNAs), long noncoding RNAs (lncRNAs), and chromosomal RNAs (known as non-coding RNAs, ncRNAs), in general, have been discovered and identified to be associated with the regulations of important cellular processes.[6–8] Even now, there still remain plenty of RNAs with functions and encoding genes that are not known.[9,10]

Recent discoveries highlight the important roles that RNAs play in modern cellular biology, and they have revealed important links between RNA structure and dynamics, and RNA functionality.[11–18] Initially it was though that RNAs are just the carrier of information from DNAs to proteins. However, besides the well known coding RNAs: messenger RNA (mRNA), ribosomal RNA (rRNA) and transfer RNA (tRNA), discovery of various forms of non-coding RNAs and post-transcriptional modifications in the past few decades have drastically changed our views towards the structure-function relationship of RNAs.[19–21] It has been shown that RNA molecules are capable of controlling the post-transcriptional expression levels of many genes[22] (RNA interference), regulate stem cell pluripotency and cell division[23] (long-non-coding RNAs) and up-regulate the transcription of the genes[24] (enhancer-RNAs). It is also evident from several experimental and theoretical studies that the structural dynamics of RNAs determine their functionality.[25,26] For example Wade and et. al showed a reduction in the bacterial gene expression of mRNA-effector complex which was attributed to the adoption of a distinct structure that sequesters the ribosome binding site.[27] On the structure-function relationship note, it would be interesting to develop methods that will enable RNA structure to be predicted with improved accuracy. Better accuracy in turn will ensure more reliably how the structural features of a certain RNA system are associated with its function, and then apply this knowledge to design drug agents that exploit these structural features to cure several diseases like dystrophy type 1 (DM1), prostate cancer, spinal muscular atrophy (SMA), Huntingtons disease-like 2 (HDL2) and autism.[28–30]

1.1.3 RNA as a drug target

There is increased awareness that RNA molecules influence every step of gene expression and regulation through activities that are attributable to its secondary

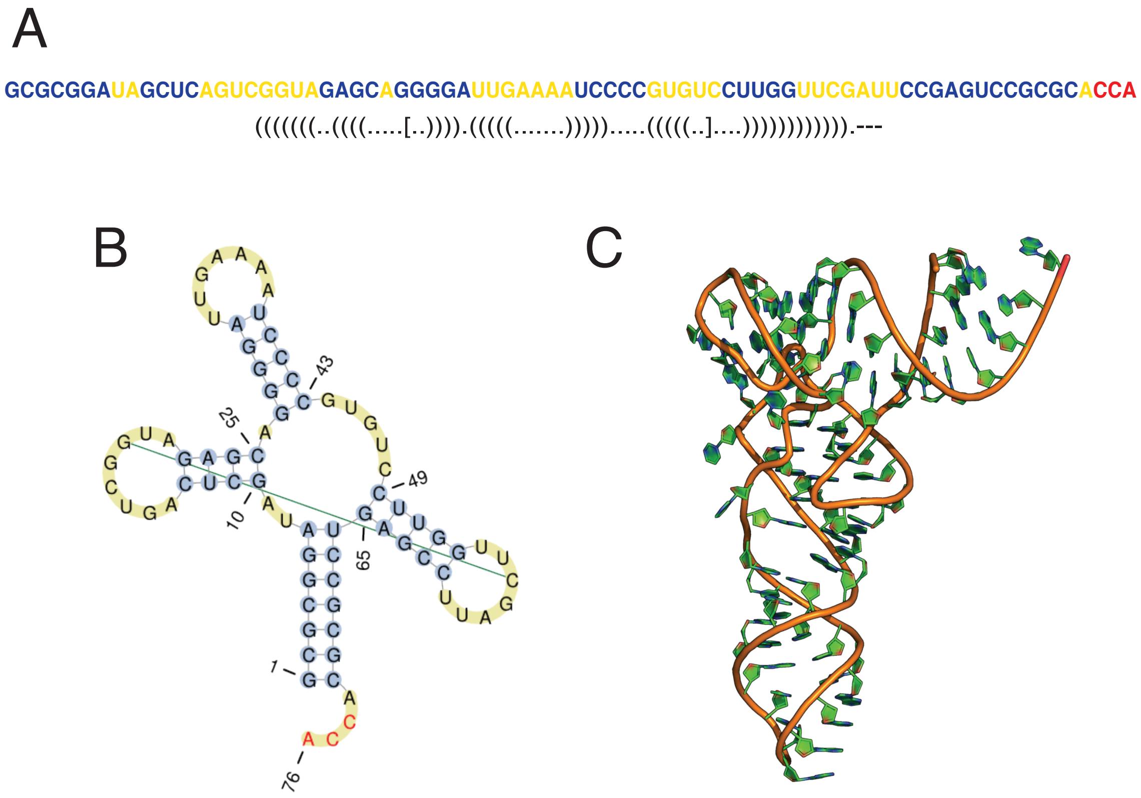


Figure 1.2: A)Primary B) Secondary and C) Tertiary Structure of RNA

and tertiary structures (Figure 1.2).[31–34] As such, targeting the RNA structure with a small molecule ligand that chemically binds in a structure specific manner can inhibit the functioning of mis-regulated RNAs. However, there is huge conformational flexibility associated with binding of a small molecule ligand to RNA, which comes from the ability of RNA molecules to undertake a variety of tertiary conformations. Extensive experiments by spectroscopic methods[35–40] have been undertaken to explore the various RNA tertiary structures. For example Bardaro and colleagues used (13)C NMR relaxation experiments to examine conformational changes in the motional landscape of HIV-1 TAR in the presence of ligands.[41] Those experiments have significantly improved our understanding of the RNA-ligand interactions and the structural changes in RNA after ligand binding.

RNAs also play a significant role in infectious diseases, especially in the case of RNA viruses that rely on a single stranded RNA genome, and hence have become attractive as potential drug targets.[42] The linking of single-stranded with double-stranded regions within RNAs leads to a well defined 3D structure that can be selectively targeted by small-molecule ligands. Small molecules can be very efficient inhibitors of a particular RNA target as they can recognize specific three-dimensional structures.[43,44] Examples of viral RNAs as drug targets include aminoglycosides which can act as inhibitors of the dimerization initiation site of Human immunodeficiency virus(HIV)-1 RNA[45] and Hepatitis C virus internal ribosome entry site (HCV IRES).[46]

RNA activity generally depends on how it interacts with the other molecules in the cell. For example, riboswitches regulate gene expression by interacting with small molecule ligands such as ions, amino acids, vitamin B12, thiamine pyrophosphate (TPP) or flavine mononucleotide (FMN).[47] Riboswitches are attractive antibacterial drug targets since they are very common in bacterial cells and rarely occur in eukaryotic cells.[48,49] Bacterial rRNAs are another (antibiotic) drug target since they constitute the active site of ribosomes. Antimicrobials that inhibit protein synthesis in bacteria act by binding to a particular site of the ribosomal RNA with various degrees of selectivity. Despite the lack of specificity, these antimicrobials have shown the viability of the small-molecule approach for RNA targeting and remain an important source of inspiration for the design of new compounds. Not only coding but also ncRNAs have been used for small-molecule targeting. An example of this is nucleotide repeat expansions like in fragile X syndrome, myotonic dystrophy or spinocerebellar ataxia and more recently oncogenic ncRNAs such as microRNAs. Apart from antibacterial targets such as bacterial ribosome and antiviral targets such as trans-activating response RNA (TAR) in HIV, there is a third class of RNA targets which is the human mRNA.[50]

RNA-ligand interactions are generally divided into 3 types: nonspecific electrostatic interactions (like between the +ve charged ligand and the -ve charged RNA phosphate backbone), specific interactions (like direct hydrogen bonding or van der Waals (VDW) interactions) and stacking interactions (between RNA bases and aromatic ligands).[51] The interactions between the ligands and the WC edge of the RNAs in combination with stacking interactions have been proposed to play a major role in ligand selectivity and recognition. All these RNA-ligand interactions fold into a particular 3D structure. Additionally, small synthetic compounds are structure specific, making the study of a given target RNA's structure of paramount importance to design ligands which are efficient and selective.[52,53]

1.2 Techniques and Methods

1.2.1 Structure Determination Techniques

The three most powerful experimental methods to determine RNA-ligand structures are X-ray crystallography,nuclear magnetic resonance (NMR) and cryogenic electron microscopy (cryo-EM). X-ray diffraction can solve very big complexes with high resolution, but requires a purifiable and crystallizable RNA, and provides limited information about the dynamics. NMR needs the purified sample to solve RNA structure in solution and provides solution state dynamics, but is limited to molecular weights below 50 kDa. With recent advances in direct electron detectors,[54,55] Cryo-EM is increasingly used for structure determination of macromolecules with molecular weights above 65 kDa. But the application to RNA-ligand and RNA-protein (RNP) complexes is limited due to extremely high conformational flexibility of large RNAs (e.g., viral RNA segments $>$ 200 kDa) and the inaccurate orientation assignment for smaller, more rigid RNAs (< 30 kDa).[36] In the past decade, much of the emphasis has been given to determine protein structures as opposed to RNA structures. The statistics from Protein Data Bank (PDB) (http://www.rcsb.org) clearly show this trend which has 145,836 (as of Nov 3, 2019) protein structures deposited, whereas the Nucleic Acid Database (http://ndbserver.rutgers.edu) has less than 3,300 RNA structures. About 20% of the RNA structures deposited in Nucleic Acid Database were determined by NMR with an average molecular weight of 8 kDa.[56] Despite the existence of numerous structural determination techniques, the current understanding of the RNA structure-function relationships is limited due to the lack of high-resolution structural information.

Each of the structural determination techniques mentioned above have some limitations. Developing computational methods that can leverage the experimental information provided by these techniques can lead to a better understanding of the RNA-world. X-ray crystallography cannot be applied to most RNAs as they are difficult to crystallize and even for the crystalline structure it only provides a partial picture of the relevant single averaged structure. NMR on the other hand can provide information about the solution state dynamics and there is abundant high-resolution experimental information readily available even in the presence of large scale flexibility, which makes for an excellent candidate in the development of structure determination methods using NMR observable. NMR spectroscopy provides dynamic structural information on pico-second to millisecond timescales meaning it is best suited to study RNA-bound structures that undergo structural changes on a variety of timescales. NMR chemical shift, the most accurate and abundantly available parameter provides details on the local structure and can be used to fingerprint the structure of RNA. We envisioned that combining computational modeling using

NMR observable would create breakthroughs in determining RNA structures in the presence of ligands and could help in expanding the RNA structure database.

1.2.2 NMR Chemical shift

In NMR spectroscopy, the energy levels of a spin active nuclei when placed in an external magnetic field are split in to half which gives rise to resonance frequency between those levels. The external applied magnetic field interacts with the internal local magnetic field around the nucleus and produces an effective resultant magnetic field which is responsible for the resonance frequency. The difference in energy levels corresponds to this resonance frequency, are directly proportional to the chemical shift, for the nuclei in the presence of the magnetic field.

$$B_{eff} = (1 - \sigma)B_0 \tag{1.1}$$

Where B_{eff} and B_0 are the effective and external magnetic field respectively, and σ is the chemical shielding.

This effective resonance frequency is also referred to as the larmor precession, which varies because the actual magnetic field, B, at the nucleus is always less than the external field B_0 (Figure 1.3). Since the extent of shielding is proportional to the external magnetic field B_0, field independent units for the chemical shifts, values, which has units of parts per million (ppm).

$$\delta = \frac{\nu - \nu_{ref}}{\nu_{ref}} \times 10^6 \tag{1.2}$$

here:

$$\nu = \frac{\gamma B_0}{2\pi}(1 - \sigma) \tag{1.3}$$

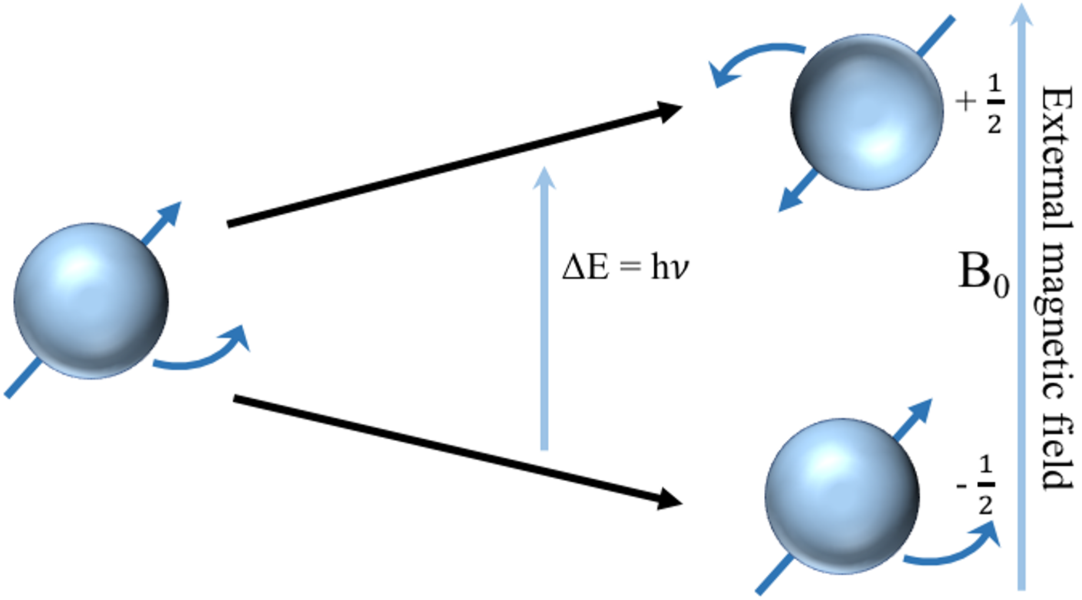


Figure 1.3: Nucleic level splitting of a spin active nuclei in the presence of external magnetic field

The magnetic field experienced by a nucleus in a molecular system is influenced by various factors, including the local electron distribution induced by the rotation of electrons around each of the nucleus. Further, the electron distribution depends on molecular geometry, such as bond lengths, bond angles, torsion angles, presence and absence of other molecules, etc. Differences in these molecular geometries account for the variation in the spin energy levels and resonance frequencies, which lead to the variations in observed NMR frequencies for the same kind of nucleus. It is for this reason that not all resonances occur at the same position for the same kind of nucleus. As such, the nucleus is shielded from the external magnetic field and the chemical shift is a measure of the extent by which this shielding is influenced by many structural features within the molecule.

Much effort has been spent working to understand the relationship between structure, dynamics and chemical shifts. Chemical shifts have been used as restraints in some computational structure determination workflows.[57] For example chemical shift guided Molecular Dynamic (MD) simulations were used to extract structural insights despite the affect which various conformational degrees of freedom pose on shielding.[58, 59] To completely understand chemical shift-structure relationship, reliable means of calculating chemical shifts needs to be developed. In theory, chemical shift can be determined using this expression

$$\sigma - 1 = \left(\frac{\partial^2 E}{\partial \mu \partial B_0} \right) \tag{1.4}$$

given that the magnetic interaction hamiltonian (or the wavefunction) is known. Where E is the interaction energy between a probe nucleus with magnetic dipole moment μ in the presence of effective magnetic field B_{eff} given by:

$$E = -\mu B_{eff} \tag{1.5}$$

Various methods within the quantum domain using ab-initio calculations and density functional theory (DFT) have been developed to calculate shielding,[60–62] but they require large basis sets, are computationally expensive, and are inaccurate at predicting shielding for large systems. Hybrid QM/MM approaches have been developed to reduce the computational expense. These hybrid methods operate by treating the nucleus and it's neighbors quantum mechanically and the rest of the system classically.[63, 64] Alternatives to quantum approaches are the empirical methods which replaces the electronic Hamiltonian and electronic interactions by parametric formulas. Most of the empirical methods assume the additive and local nature of chemical shift and are thus divided into short range, electrostatic and magnetic components.[65, 66] A lot of models have been developed to use the atom coordinates parametrized against a database of chemical shifts and solved structures.[67–69]

Chemical shifts have been extensively used to study proteins, but the use of chemical shifts to study RNA structure is rather limited. Some studies have tried to establish a connection between chemical shifts and torsion angles, ring currents in the aromatic rings, stacking interactions, electrostatic and magnetic interactions but only few studies have been carried out that infer structures from predicted chemical shifts.[70–73] These studies have shown that 1H and 13C chemical shifts are useful for structure validation and refinement. Hence accurately predicting the chemical shifts is vital for the structure determination of RNA and RNA containing complexes. In this dissertation, I will describe my modest approach to study the effects of ring currents on accurately predicting the chemical shifts of RNAs using machine learning approaches (Chapter 2). Following the successful prediction of RNA chemical shifts, I will explore the use of chemical shifts in predicting the structure of RNA, starting from sequence (primary structure), to secondary and tertiary structure in the absence and presence of ligands (Chapter 3). To aid the prediction of chemical shifts, I use machine learning techniques (described below) which are increasingly becoming popular and were found to be robust at predicting the chemical shifts.

1.2.3 Machine Learning (ML) Basics

According to Mitchell (1997), "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T , as measured by P, improves with experience E." In other words, machine learning is a field in which algorithms have an ability to learn by itself and make a prediction for unseen data.

More precisely, machine learning (ML) is a sub-field of computer science in which a computer (machine) can learn by itself (learning) given some initial knowledge. This initial knowledge is provided in terms of 'training' data and tested on a 'testing' data. An algorithm is a method or a function which are used on the training dataset to generate an ML model e.g. linear regression, decision trees, and ensemble methods. A model is a data structure which stores a representation of a dataset, which are known as its weights and biases. Models are taught by training an algorithm on a dataset. The attributes and values in those datasets are referred to as the features. If one is predicting a categorical output (discrete set of values) then its called Classification whereas if one is predicting a continuous output (range of possible values on number scale) then its called regression. Loss is the difference between the true value and the predicted value. Each model tries to minimize the loss on the training data which gives an indication of how the model is doing (the lower the loss the better the model). Any ML model is then assessed on a testing set by a performance metrics e.g. accuracy, sensitivity or confusion matrix.

An important property of ML predictive models is the bias- variance tradeoff. Bias comes from erroneous assumptions in the learning algorithm whereas variance stems from sensitivity to small fluctuations in the training set. High bias can cause the model to miss relevant information between the target and features (underfitting) whereas high variance can lead the model to incorporate random noise in the training data (overfitting). If the model is too simple and has very few parameters then it is characteristic of high bias and low variance. Conversely, a characteristic of a low bias model is if it has a large number of parameters. Finding the right balance is crucial to prevent overfitting and underfitting the data. This tradeoff exists because the algorithm cannot be more complex and less complex at the same time, and analyzing the bias-variance decomposition can help in inferring the learning algorithm's expected generalization error.

There are several algorithms within the ML field, like linear methods, decision trees, ensemble of trees, support vector machines, neural networks, etc. We will focus on two of these algorithmic domains: linear and ensemble methods. Linear methods are pretty straightforward to understand, explain, and implement. They can also be regularized to avoid under- or overfitting, and can also be updated when new data comes in. But linear models are harder at dissecting non-linear and complex relationships within the data. e.g. linear regression, ridge regression, lasso lars regression. Ensemble methods are several decision trees combined together with the final prediction being an average of the trees. Decision trees learn in hierarchical fashion by splitting the dataset into separate branches, which allows them to learn non-linear relationships. Ensemble methods are robust to outliers and perform very well in practice. Some exmaples of ensemble methods are Random Forest and Extra Randomized trees.

Linear Methods

The following are a set of linear methods used in this thesis and can be used for regression or classification. The target value is expected (assumed) to be a linear combination of the features given by

$$y = w_0 + w_1x_1 + ... + w_nx_n \tag{1.6}$$

and $w's$ are the weights of the models to be computed by the algorithm.

Linear Regression

Linear Regression fits a model by minimizing the residual sum of squares between the observed values and the predicted values. Mathematically:

$$\underset{w}{min} \|Xw - y\|_2^2 \tag{1.7}$$

The linear methods works best if the features are uncorrelated. If the features vectors are correlated, the design matrix becomes close to singular and the estimate becomes highly sensitive to random errors in the observed target.

Ridge Regression

Ridge regression tries to address some of the shortcomings of the Ordinary Least Squares by imposing a penalty of second order (residual sum of squares a.k.a. L2 prior) on the coefficients.

$$\underset{w}{min} \left\|Xw - y\right\|_2^2 + \alpha \left\|w\right\|_2^2 \tag{1.8}$$

where alpha controls the amount of shrinkage: The larger the value, the greater the shrinkage and which maintains the robustness to collinearity.

Lasso Regression

Lasso is a linear model that imposes a penalty (regularization) of first order (L1 prior)

$$\underset{w}{min} \left\|Xw - y\right\|_2^2 + \alpha \left\|w\right\|_1 \tag{1.9}$$

Lasso is very useful in estimating sparse coefficients due to its tendency to prefer solutions with lesser non-zero coefficients (and thereby reducing the number of features upon which the given solution is dependent). There are two implementations of Lasso one that uses coordinate descent to fit the coefficients and the other uses Least Angle Regression (LARS algorithm).

LassoLars

LassoLars is a lasso ML model that uses the LARS algorithm implementation and yields the exact solution, which is piece-wise linear as a function of the norm of its coefficients. Least-angle regression (LARS) in itself is a regression algorithm for high-dimensional data, is very similar to forward step-wise regression, and can be incorporated with the Lasso method. It works by finding the features that are most correlated with the target at each step.

Bayesian Ridge

In the Bayesian ridge, regression is formulated using probability distributions rather than point estimates, as in linear regression (above). The response, y, is drawn from a probability distribution, sampled from a normal distribution:

$$y \sim N(B^T X, \sigma^2 I) \tag{1.10}$$

The model parameters are estimated by maximizing the log marginal likelihood.

Ensemble Methods

Ensemble methods combines the predictions of several base estimators built with any given algorithm (decision trees) to improve generalizability and robustness over a single base estimator. The combined estimator is better than any single base estimator (on average). All the base estimators are build independently (in averaging methods) and then the predictions are averaged over them resulting in reduction of variance. There are mainly two types of ensemble methods: Bagging and Boosting. Bagging considers homogeneous weak learners, learns them independently from each other at the same time and combines those weak learners using a deterministic averaging strategy. Boosting considers homogeneous weak learners, learns them sequentially in a adjustable way and combines those weak learners using a deterministic process. Bagging reduces variance of the combined ensemble models compared to its component parts whereas boosting reduces bias of the combined ensemble models compared to its component parts (reducing variance at the same time if possible). Random forest and Extra Randomized trees are examples of bagging and gradient boosting is an example of boosting ensemble methods.

Bagging

In ensemble methods, bagging belongs to a class of algorithms that build several instances of a black-box estimator on random subsets of the train data, and average the individual predictions to form a final prediction. This helps to reduce the variance of a base estimator (a decision tree in our case), by introducing randomization into the procedure. Then these base estimators are combined to make an ensemble out of it which reduce over-fitting. Bagging work best with strong and complex models and in scikit-learn, bagging methods are offered as a unified Bagging Classifier or Regressor, which take a base estimator as an input along with parameters specifying the strategy to draw random subsets. We used decision trees as our base estimators for the Bagging method.

Random Forest

Each tree in the random forests (Classifier or Regressor) ensemble is built from a sample drawn with replacement (bootstrap) from the training dataset and the best split is computed from all the input features (or a random subset of them) during splitting each node while tree construction. Independent decision trees typically exhibit high variance (over-fitting) but these two sources of randomness in forests helps in decoupling the prediction errors (by averaging those predictions) thereby decreases the variance of the random forest estimator. Random forests tends to have a reduced variance by combining a lot of trees, but at the cost of a slight increase in bias.

Extra Randomized Trees

In Extra Randomized Trees, thresholds are drawn at random for each feature (instead of looking for the most distinctive thresholds for splitting as in RF). The best of these randomly generated thresholds is chosen as the splitting. This reduces the variance of the model even more, at the expense of a slight increase in bias.

Gradient Boosting

Gradient Boosting trains an estimator in a gradual, additive, and sequential manner.

It improves the shortcoming of the estimators by using gradients in the loss function (y=ax+b+e, where e is the error term). The loss function is a measure of how good the ensemble model's coefficients are at fitting the training data. By using gradient descent algorithm and updating our predictions based on a learning rate, we can find the values of the parameters where the error is minimized. In other words, we are updating the predictions at every step such that the sum of our residuals is close to 0 (or minimum) and predicted values are more or less close to actual values.

ML has become a popular and powerful tool for analyzing data and capturing insights from features embedded in the data. A number of studies have shown that ML can achieve state-of-the-art prediction performance in various learning tasks, from image and speech recognition to natural language processing. ML has also been successfully applied to solve many prediction problems in computational biophysics, such as protein structure prediction, RNA splicing prediction, RNA-protein binding prediction, and protein and RNA structure prediction.[74–78] There are several studies which use high and low resolution experimental information combined with ML approaches to accurately predict the structure of biomolecules.[79,80] These methods have been prevalent in protein structure prediction and are becoming increasingly popular in RNA structure prediction. The reason ML is becoming increasingly popular as opposed to the quantum approaches to calculate chemical shifts is because it is easier and faster to implement and does not require a lot of computational resources, yet still provides reasonably accurate results.

1.3 Aims and Objectives

In this thesis, I use linear and ensemble methods to study the effects of different ring currents models on the prediction of chemical shifts. Next I use those chemical shifts to refine the structure of RNA complexes from sequence up to 3D structure. I explore the possibility of chemical shifts to refine the RNA-ligand structure. Then I use a molecular fingerprinting approach and random forest ML classifier to build RNAPoser, which was able to recover native-like pose from a set of decoy pools.

In theory, NMR spectroscopy derived chemical shifts contain information about the local chemical environment at each site in a molecule, and so can be a source of rich structural information. Significant efforts has been been expended to better understand the relationship between structure and chemical shifts. As alluded earlier, standard experimental approaches are expensive and time-consuming, and current state-of-the-art computational methods are too inaccurate to be useful. As such, to enhance our understanding of chemical shifts and their relationship to the structure, precise methods needs to be developed to predict chemical shifts. Chemical shifts are local phenomena, and depend on their local chemical environment for a particular nuclei. Chemical shifts can be influenced by phenomena including hydrogen bonding, stacking interactions, electrostatic interactions, ring current, and magnetic anisotropy. There are three different ring current models known in current literature, namely, Pople, Johnson-Bovey and Haigh-Mallion, yet a side-by-side comparison of these ring current models in predicting chemical shifts is yet to be performed.

Knowledge of 3D structures of RNA-ligand complexes is necessary to explore molecular details and gain insights into the structure-function relationships. Secondary structural elements of RNAs (which include the WC and non-WC base pairs) contribute to the overall 3D structure of RNAs. Numerous studies have been reported that focus on the prediction of RNA secondary structure from the sequence, such as the RNAStructure software tool developed by David H. Mathews lab. Yet accurately predicting the structure of RNAs is still a challenge. As such, there is a need for the development of fast, easy, and precise prediction methods for determining the 3D structure of RNAs.

1.4 Thesis Plan

In chapter 2, I explore how different ML methods perform in terms of predicting the chemical shifts. In this chapter, I explore the accuracy of linear and ensemble methods in predicting chemical shifts. Then the effect of different ring current models is explored on the accuracy of chemical shift prediction.

In chapter 3, I explore the use of chemical shifts to guide the 3D structure prediction of RNA-ligand complexes. Here, I use chemical shifts to guide the secondary structure prediction of RNAs.[81] Then I use chemical shifts to refine those generated models. This entire approach set the groundwork for predicting RNA-ligand structure starting from sequence all the way to 3D structure using experimental information.

In chapter 4, I trained a set of ML classifiers to recover native-like poses of RNA-ligand complexes from non-native poses. The classifiers utilized a atomic fingerprinting approach, designed by my colleague Jingru, and encodes the local atomic environments as simple distance-based features. The classifier I trained using these fingerprint could more accurately recover native-like poses than current state-of-the-art methods.

BIBLIOGRAPHY

[1] Tina Glisovic, Jennifer L Bachorik, Jeongsik Yong, and Gideon Dreyfuss. RNA - binding proteins and post-transcriptional gene regulation. *FEBS letters*, 582(14):1977–1986, 2008.

[2] Cole Trapnell, David G Hendrickson, Martin Sauvageau, Loyal Goff, John L Rinn, and Lior Pachter. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nature biotechnology*, 31(1):46, 2013.

[3] Nina Sesto, Omri Wurtzel, Cristel Archambaud, Rotem Sorek, and Pascale Cossart. The excludon: A new concept in bacterial antisense RNA-mediated gene regulation. *Nature Reviews Microbiology*, 11(2):75, 2013.

[4] Pamela J Green, Ophry Pines, and Masayori Inouye. The role of antisense RNA in gene regulation. *Annual review of biochemistry*, 55(1):569–597, 1986.

[5] Kevin M Esvelt, Prashant Mali, Jonathan L Braff, Mark Moosburner, Stephanie J Yaung, and George M Church. Orthogonal Cas9 proteins for RNA-guided gene regulation and editing. *Nature methods*, 10(11):1116, 2013.

[6] T Phillips. Small non-coding RNA and gene expression. *Nature Education*, 1(1):115, 2008.

[7] Juliane CR Fernandes, Stephanie M Acuña, Juliana I Aoki, Lucile M Floeter-Winter, and Sandra M Muxel. Long non-coding RNAs in the regulation of gene expression: physiology and disease. *Non-coding RNA*, 5(1):17, 2019.

[8] Marios A Diamantopoulos, Panagiotis Tsiakanikas, and Andreas Scorilas. Non-coding RNAs: the riddle of the transcriptome and their perspectives in cancer. *Annals of Translational Medicine*, 6(12), 2018.

[9] Kimberly A Harris and Ronald R Breaker. Large noncoding RNAs in bacteria. *Microbiology spectrum*, 6(4), 2018.

[10] Jeremy E Wilusz, Hongjae Sunwoo, and David L Spector. Long noncoding RNAs: functional surprises from the RNA world. *Genes & development*, 23(13):1494–1504, 2009.

[11] Stefanie A Mortimer, Mary Anne Kidwell, and Jennifer A Doudna. Insights into RNA structure and function from genome-wide studies. *Nature reviews Genetics*, 15(7):469–479, 2014.

[12] José Almeida Cruz and Eric Westhof. The dynamic landscapes of RNA architecture. *Cell*, 136(4):604–609, 2009.

[13] M Bryan Warf and J Andrew Berglund. Role of RNA structure in regulating pre-mRNA splicing. *Trends in biochemical sciences*, 35(3):169–178, 2010.

[14] David M Mauger, Nathan A Siegfried, and Kevin M Weeks. The genetic code as expressed through relationships between mRNA structure and protein function. *FEBS letters*, 587(8):1180–1188, 2013.

[15] Eric J Strobel, Kyle E Watters, David Loughrey, and Julius B Lucks. RNA systems biology: uniting functional discoveries and structural tools to understand global roles of RNAs. *Current opinion in biotechnology*, 39:182–191, 2016.

[16] C Joel McManus and Brenton R Graveley. RNA structure and the mechanisms of alternative splicing. *Current opinion in genetics & development*, 21(4):373–379, 2011.

[17] Kelsey C Martin and Anne Ephrussi. mRNA localization: gene expression in the spatial dimension. *Cell*, 136(4):719–730, 2009.

[18] Alina Selega and Guido Sanguinetti. Trends and challenges in computational RNA biology, 2016.

[19] Amelia E Aranega and Diego Franco. Post-transcriptional regulation by proteins and non-coding RNAs. In *Congenital Heart Diseases: The Broken Heart*, pages 153–171. Springer, 2016.

[20] Rong-Zhang He, Di-Xian Luo, and Yin-Yuan Mo. Emerging roles of lncRNAs in the post-transcriptional regulation in cancer. *Genes & diseases*, 2019.

[21] Iain M Dykes and Costanza Emanueli. Transcriptional and post-transcriptional gene regulation by long non-coding RNA. *Genomics, proteomics & bioinformatics*, 15(3):177–186, 2017.

[22] Andrew Fire, SiQun Xu, Mary K Montgomery, Steven A Kostas, Samuel E Driver, and Craig C Mello. Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *nature*, 391(6669):806, 1998.

[23] John L Rinn and Howard Y Chang. Genome regulation by long noncoding RNAs. *Annual review of biochemistry*, 81:145–166, 2012.

[24] Ryan J Taft, Craig D Kaplan, Cas Simons, and John S Mattick. Evolution, biogenesis and function of promoter-associated RNAs. *Cell Cycle*, 8(15):2332–2338, 2009.

[25] Melanie A O'Neil and Jacqueline K Barton. 2-Aminopurine: a probe of structural dynamics and charge transfer in DNA and DNA: RNA hybrids. *Journal of the American Chemical Society*, 124(44):13053–13066, 2002.

[26] Jiri Sponer, Giovanni Bussi, Miroslav Krepl, Pavel Banas, Sandro Bottaro, Richard A Cunha, Alejandro Gil-Ley, Giovanni Pinamonti, Simón Poblete, Petr Jurecka, et al. RNA structural dynamics as captured by molecular simulations: A comprehensive overview. *Chemical reviews*, 118(8):4177–4338, 2018.

[27] Wade Winkler, Ali Nahvi, and Ronald R Breaker. Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression. *Nature*, 419(6910):952, 2002.

[28] Joseph D Puglisi, Ruoying Tan, Barbara J Calnan, Alan D Frankel, et al. Conformation of the TAR RNA-arginine complex by NMR spectroscopy. *Science*, 257(5066):76–80, 1992.

[29] Thomas A Cooper, Lili Wan, and Gideon Dreyfuss. RNA and disease. *Cell*, 136(4):777–793, 2009.

[30] Manel Esteller. Non-coding RNAs in human disease. *Nature reviews genetics*, 12(12):861, 2011.

[31] Christine S Chow and Felicia M Bogdan. A structural basis for RNA- ligand interactions. *Chemical reviews*, 97(5):1489–1514, 1997.

[32] Audrey Di Giorgio and Maria Duca. Synthetic small-molecule RNA ligands: future prospects as therapeutic agents. *MedChemComm*, 2019.

[33] Masayuki Matsui and David R Corey. Non-coding RNAs as drug targets. *Nature reviews Drug discovery*, 16(3):167, 2017.

[34] Kevin V Morris and John S Mattick. The rise of regulatory RNA. *Nature Reviews Genetics*, 15(6):423–437, 2014.

[35] VV Krishnan and B Rupp. Macromolecular structure determination: comparison of X-ray crystallography and NMR spectroscopy. *e LS*, 2001.

[36] Kaiming Zhang, Sarah C Keane, Zhaoming Su, Rossitza N Irobalieva, Muyuan Chen, Verna Van, Carly A Sciandra, Jan Marchant, Xiao Heng, Michael F Schmid, et al. Structure of the 30 kDa HIV-1 RNA dimerization signal by a hybrid cryo-EM, NMR, and molecular dynamics approach. *Structure*, 26(3):490–498, 2018.

[37] Michael P Latham, Darin J Brown, Scott A McCallum, and Arthur Pardi. NMR methods for studying the structure and dynamics of RNA. *Chembiochem*, 6(9):1492–1505, 2005.

[38] Seiki Baba, Ken-ichi Takahashi, Satoko Noguchi, Hiroshi Takaku, Yoshio Koyanagi, Naoki Yamamoto, and Gota Kawai. Solution RNA structures of the HIV-1 dimerization initiation site in the kissing-loop and extended-duplex dimers. *Journal of biochemistry*, 138(5):583–592, 2005.

[39] Nathan J Baird, Steven J Ludtke, Htet Khant, Wah Chiu, Tao Pan, and Tobin R Sosnick. Discrete structure of an RNA folding intermediate revealed by cryo-electron microscopy. *Journal of the American Chemical Society*, 132(46):16352–16353, 2010.

[40] Sarah C Keane and Michael F Summers. NMR studies of the structure and function of the HIV-1 5-leader. *Viruses*, 8(12):338, 2016.

[41] Michael F Bardaro Jr, Zahra Shajani, Krystyna Patora-Komisarska, John A Robinson, and Gabriele Varani. How binding of small molecule and peptide ligands to HIV-1 TAR alters the RNA motional landscape. *Nucleic acids research*, 37(5):1529–1540, 2009.

[42] Mark EJ Woolhouse, Kyle Adair, and Liam Brierley. RNA viruses: A case study of the biology of emerging infectious diseases. *Microbiology spectrum*, 1(1), 2013.

[43] Fareed Aboul-ela. Strategies for the design of RNA - binding small molecules. *Future medicinal chemistry*, 2(1):93–119, 2010.

[44] Anton A Komar and Maria Hatzoglou. Exploring internal ribosome entry sites as therapeutic targets. *Frontiers in oncology*, 5:233, 2015.

[45] Jean-Marc Jacque, Karine Triques, and Mario Stevenson. Modulation of HIV-1 replication by RNA interference. *Nature*, 418(6896):435, 2002.

[46] Peter J Lukavsky. Structure and function of HCV IRES domains. *Virus research*, 139(2):166–171, 2009.

[47] Jesse C Cochrane and Scott A Strobel. Riboswitch effectors as protein enzyme cofactors. *Rna*, 14(6):993–1002, 2008.

[48] Kenneth F Blount and Ronald R Breaker. Riboswitches as antibacterial drug targets. *Nature biotechnology*, 24(12):1558, 2006.

[49] Ronald R Breaker. Prospects for riboswitch discovery and analysis. *Molecular cell*, 43(6):867–879, 2011.

[50] Colleen M Connelly, Michelle H Moon, and John S Schneekloth Jr. The emerging role of RNA as a therapeutic target for small molecules. *Cell chemical biology*, 23(9):1077–1090, 2016.

[51] Efrat Kligun and Yael Mandel-Gutfreund. Conformational readout of RNA by small ligands. *RNA biology*, 10(6):981–989, 2013.

[52] Arturo López Castel, John D Cleary, and Christopher E Pearson. Repeat instability as the basis for human diseases and as a potential target for therapy. *Nature reviews Molecular cell biology*, 11(3):165, 2010.

[53] Hui Ling, Muller Fabbri, and George A Calin. MicroRNAs and other non-coding RNAs as targets for anticancer drug development. *Nature reviews Drug discovery*, 12(11):847–865, 2013.

[54] AR Faruqi and R Henderson. Electronic detectors for electron microscopy. *Current opinion in structural biology*, 17(5):549–555, 2007.

[55] G McMullan, AR Faruqi, D Clare, and R Henderson. Comparison of optimal performance at 300 keV of three direct electron detectors for use in low dose electron microscopy. *Ultramicroscopy*, 147:156–163, 2014.

[56] Ravi P Barnwal, Fan Yang, and Gabriele Varani. Applications of NMR to structure determination of RNAs large and small. *Archives of biochemistry and biophysics*, 628:42–56, 2017.

[57] Santrupti Nerli, Andrew C McShan, and Nikolaos G Sgourakis. Chemical shift-based methods in NMR structure determination. *Progress in nuclear magnetic resonance spectroscopy*, 106:1–25, 2018.

[58] Paul Robustelli, Kai Kohlhoff, Andrea Cavalli, and Michele Vendruscolo. Using NMR chemical shifts as structural restraints in molecular dynamics simulations of proteins. *Structure*, 18(8):923–933, 2010.

[59] Jiri Sponer, Giovanni Bussi, Miroslav Krepl, Pavel Banas, Sandro Bottaro, Richard A Cunha, Alejandro Gil-Ley, Giovanni Pinamonti, Simón Poblete, Petr Jurecka, et al. RNA structural dynamics as captured by molecular simulations: A comprehensive overview. *Chemical reviews*, 118(8):4177–4338, 2018.

[60] Vladimir G Malkin, Olga L Malkina, Mark E Casida, and Dennis R Salahub. Nuclear magnetic resonance shielding tensors calculated with a sum-over-states density functional perturbation theory. *Journal of the American Chemical Society*, 116(13):5898–5908, 1994.

[61] Trygve Helgaker, Philip J Wilson, Roger D Amos, and Nicholas C Handy. Nuclear shielding constants by density functional theory with gauge including atomic orbitals. *The Journal of Chemical Physics*, 113(8):2983–2989, 2000.

[62] Martin Kaupp, Vladimir G Malkin, Olga L Malkina, and Dennis R Salahub. Calculation of ligand NMR chemical shifts in transition-metal complexes using ab initio effective-core potentials and density functional theory. *Chemical physics letters*, 235(3-4):382–388, 1995.

[63] Daniel Sebastiani and Ursula Rothlisberger. Nuclear magnetic resonance chemical shifts from hybrid DFT QM/MM calculations. *The Journal of Physical Chemistry B*, 108(9):2807–2815, 2004.

[64] Xinsheng Jin, Tong Zhu, John ZH Zhang, and Xiao He. Automated fragmentation QM/MM calculation of NMR chemical shifts for protein-ligand complexes. *Frontiers in chemistry*, 6:150, 2018.

[65] Doree Sitkoff and David A Case. Theories of chemical shift anisotropies in proteins and nucleic acids. *Progress in nuclear magnetic resonance spectroscopy*, 32(2):165–190, 1998.

[66] Julio C Facelli. Chemical shift tensors: Theory and application to molecular structural problems. *Progress in nuclear magnetic resonance spectroscopy*, 58(3-4):176, 2011.

[67] Wolfgang Rieping and Wim F Vranken. Validation of archived chemical shifts through atomic coordinates. *Proteins: Structure, Function, and Bioinformatics*, 78(11):2482–2489, 2010.

[68] Stephen Neal, Alex M Nip, Haiyan Zhang, and David S Wishart. Rapid and accurate calculation of protein 1H, 13C and 15N chemical shifts. *Journal of biomolecular NMR*, 26(3):215–240, 2003.

[69] Wim F Vranken and Wolfgang Rieping. Relationship between chemical shift value and accessible surface area for all amino acid atoms. *BMC structural biology*, 9(1):20, 2009.

[70] Yang Shen and Ad Bax. Protein backbone and sidechain torsion angles predicted from NMR chemical shifts using artificial neural networks. *Journal of biomolecular NMR*, 56(3):227–241, 2013.

[71] Miha Plevnik, Zofia Gdaniec, and Janez Plavec. Solution structure of a modified 2, 5-linked RNA hairpin involved in an equilibrium with duplex. *Nucleic acids research*, 33(6):1749–1759, 2005.

[72] Ashok S Shetty, Jinshan Zhang, and Jeffrey S Moore. Aromatic π-stacking in solution as revealed through the aggregation of phenylacetylene macrocycles. *Journal of the American Chemical Society*, 118(5):1019–1027, 1996.

[73] Andrea Cavalli, Xavier Salvatella, Christopher M Dobson, and Michele Vendruscolo. Protein structure determination from NMR chemical shifts. *Proceedings of the National Academy of Sciences*, 104(23):9615–9620, 2007.

[74] Sai Zhang, Jingtian Zhou, Hailin Hu, Haipeng Gong, Ligong Chen, Chao Cheng, and Jianyang Zeng. A deep learning framework for modeling structural features of RNA-binding protein targets. *Nucleic acids research*, 44(4):e32–e32, 2015.

[75] Yifeng Cui, Qiwen Dong, Daocheng Hong, and Xikun Wang. Predicting protein-ligand binding residues with deep convolutional neural networks. *BMC bioinformatics*, 20(1):93, 2019.

[76] John A Capra and Mona Singh. Predicting functionally important residues from sequence conservation. *Bioinformatics*, 23(15):1875–1882, 2007.

[77] JD Fischer, Christian E Mayer, and Johannes Söding. Prediction of protein functional residues from sequence by probability density estimation. *Bioinformatics*, 24(5):613–620, 2008.

[78] Manish Kumar, M Michael Gromiha, and Gajendra Pal Singh Raghava. Prediction of RNA binding sites in a protein using SVM and PSSM profile. *Proteins: Structure, Function, and Bioinformatics*, 71(1):189–194, 2008.

[79] Xiaoyong Pan and Hong-Bin Shen. Predicting RNA–protein binding sites and motifs through combining local and global deep convolutional neural networks. *Bioinformatics*, 34(20):3427–3436, 2018.

[80] Sheng Wang, Jian Peng, Jianzhu Ma, and Jinbo Xu. Protein secondary structure prediction using deep convolutional neural fields. *Scientific reports*, 6:18962, 2016.

[81] Kexin Zhang and Aaron Terrence Frank. Conditional prediction of RNA secondary structure using NMR chemical shifts. *bioRxiv*, page 554931, 2019.

CHAPTER II

Effect of Different Ring Current Models on NMR Chemical Shift Prediction

2.1 Statement of Contribution

1. **Aaron T. Frank, PhD:** Conceived the project described in this chapter.
2. **Sahil Chhabra, PhD Candidate (Chemistry and Scientific Computing):** Generated the data to study the effect of ring current models on NMR chemical shift prediction in this chapter; Independently wrote this chapter.

2.2 Introduction

Determining the structure of the RNA in the presence of drug like molecules is a crucial step in any drug development campaign. NMR spectroscopy is an ideal tool for studying the RNA-ligand interactions as it provide a range of structural information about the RNA and RNA-binding properties of small molecule ligands.[1] For instance, NMR-derived chemical shifts of an RNA, which are site-specific probes of the local electronic environment within the RNA, change in the presence of ligands such as small drug-like molecules. As such, NMR spectroscopy can be used to monitor the binding of small molecules to RNA.

The data from various NMR techniques like ligand observed NMR techniques have proven to be highly useful to study the nucleic acid-ligand interactions.[1] The most significant usage of NMR derived chemical shifts in the presence of ligands is in the validation of hits from high throughput screening campaigns. Restraints from NMR spectroscopy combined with other experimental techniques like SAXS has been shown to be used to predict the RNA 3D structure using a coarse grained model.[2] In the "current affairs" Yu et.al.[3] (April 2017) developed the ^{1}H empirical chemical shift perturbation (HECSP) method and a scoring function NMRScoreP to refine the structure of the protein-ligand complex by comparing experimental and calculated chemical shift perturbation (CSP). NMR-CSP assisted docking has emerged as a powerful tool for locating the binding site of ligands.[4–17] However it is still difficult to extract the CSPs that are induced by only ligand binding because of conformational changes in RNA-ligand complexes. Chemical shifts which adjust themselves in presence and absence of ligands are ideal candidate to study the structure of RNA-ligand complexes.

As mentioned, standard experimental approaches to calculate chemical shifts are expensive and time consuming, and current state-of-the-art computational methods are too inaccurate to be useful. As such, there is a need for the development of fast, easy, and "precise prediction" methods for determining the 3D structure of RNAs. Since NMR spectroscopy derived CS contain information about the local chemical environment at each site in a molecule, so it can be a source of rich structural information. Developing models using this "structural information" with better accuracy, in turn, could elucidate how the key structural features of RNA systems are associated with its function, and then apply this knowledge to understand how systems interact, that is, what they bind to, and design drug agents that exploit these structural features to solve medical problems.

As eluded earlier, chemical shifts a.k.a. 'chemical fingerprints" of any molecule, which are readily accessible, precisely measured and sensitive to the structure can resolve the structure of RNA-ligand complex. However accurately predicting these chemical shifts is still a challenge. To address these challenge, I attempt to explore the effect of different ring current models and machine learning algorithms at predicting the chemical shifts of RNAs.

Ring current effects significantly influence the NMR chemical shifts in RNA and RNA-ligand complexes. Understanding these effects is particularly important for the development of accurate prediction methods of chemical shifts and the structures of RNAs. In literature there are 3 different ring current models as follows in order of their increasing complexity: Pople, Johnson-Bovey and Haigh-Mallion. In this work, we first analyzed the the effect of different machine learning algorithms: ensemble and linear, to accurately predict the chemical shifts for RNAs. Then we compare the effect of different ring current models in terms of their ability to accurately predict the chemical shifts.

The present study is aimed at elucidating the specific ring current effects of RNA conjugated rings on the chemical shifts of RNAs. We elucidate this description using the geometric factors and parameters that model the ring current effects of the conjugated rings on the chemical shifts of neighboring RNA atoms. Since most of the results presented here are transferable to any 5 and 6 membered rings, this study will find applicability in a lot of problems involving the analysis of ring current effects in NMR shielding. Here we combine the results of the best machine learning algorithms to predict the chemical shifts, with the results of exploring the major factors that determine the chemical shifts in RNAs to build a high-quality structure-based predictor of RNA chemical shifts. We use these types of chemical shifts predictions in chapter 3, to enable the refinement of RNA structures based on predicted chemical shift errors.

2.3 MATERIALS AND METHODS

In NMR spectroscopy, the nucleus is shielded from the external magnetic field, and the CS is the measure of the extent to which this shielding is influenced by the local electronic environment within the molecule. To predict the CS we encode the local environment as features derived from the coordinates of RNA, and build a machine learning model based on those features. In particular, for given spin-active nuclei on a given residue in an RNA, their chemical shifts are predicted based on hydrogen bonding status of the residue, the stacking interactions of the residue, the torsions of the residues as well as the local magnetic anisotropy, ring-current effects, and polarization effects. Below I give a brief description of each of the structural features used in the prediction of chemical shifts.

2.3.1 Hydrogen Bonding

Hydrogen bonding exists between the base in both the RNA and DNA. In RNA the canonical base pairs exist between adenine uracil pair (2 hydrogen bonds) and between cytosine - guanine pair (3 hydrogen bonds). Since the chemical shift is intrinsically related to the local electronic environment, the changes due to hydrogen bonding lead to a redistribution of the electron density thereby changing the CS of the nuclei involved in the bonding. There is always a downfield shift (higher frequency) for the electro-negative hydrogen bonded nucleus (O or N). The formation of hydrogen bonds in base pairs results in downfield shift for the amino and imino protons due to a decrease in the electron density around the hydrogen nucleus and de-shielding effects from the acceptor atom.

For the purpose of predicting chemical shifts, we calculate the number of hydrogen bonding interactions present between the base-base, base-backbone, and backbone-backbone inside each RNA residue. A distance cutoff of 3.5 Åand an angle cutoff of 135° between the hydrogen donor and acceptor was applied to determine the presence of a hydrogen bond. Both the conditions must be simultaneously fulfilled i.e. donor and acceptor atom must be within the distance cutoff and the angle must be grater the the angle cutoff to be accounted as a valid hydrogen bonding interaction.

2.3.2 Magnetic Anisotropy

In the presence of an external magnetic field the local circulation of electrons is stronger in some orientations of the molecule than in others. This anisotropic electron circulation causes shielding (right shift in NMR) and de-shielding (left shift in NMR) effects, which are called magnetic anisotropy effects. Aromatic rings like benzene cause very large shielding for H placed above the ring, and smaller deshielding for H to the side of the ring. This large up-field (above the ring) and down-filed (side of the ring) shifts, is caused due to the stronger electron circulation in case when the plane of the benzene ring is perpendicular to the magnetic field as opposed to when it is parallel to it.

The shielding effects due to the magnetic anisotropy on a query atom k due to the anisotropy of atom l were accounted for by Prado and Giessner-Prettre equation[18]

$$\delta_{ma,k} = \frac{1}{3r^5}\sum_{\alpha\beta}(3r_\alpha r_\beta - r^2)(1.967R_{\alpha\beta} - 5.368Q_{\alpha\beta}) \tag{2.1}$$

where r is the distance between the query atom k and atom l, α, β loop over all x,y,z and $R_{\alpha\beta}$ and $Q_{\alpha\beta}$ are the diamagnetic and paramagnetic components of the $\alpha\beta$

magnetic susceptibility tensor of atom l.

2.3.3 Ring Current

Another feature incorporated in the predictive modeling of chemical shifts is the ring current evaluated in the proximity of a simple conjugated system. The ring current is simply the product of the ring current intensity and a geometrical parameter, which captures the chemical environment around the ring and is calculated by the distance of the query point from the ring and the angle it forms with the center of the ring. The contribution to the CS by ring current (in ppm) is given by

$$\delta_R \times 10^{-6} = iBG(\rho, z, \phi) \tag{2.2}$$

Here i is the ring current intensity and G the geometric coefficient calculated by different ring current models and B is a constant. There are theoretical models of ring current effects, which emerged from both classical as well as quantum mechanical approaches. Of the numerous theoretical models, three have received a considerable attention owing to the ease of their implementation and the availability of the derived empirical tables - PO (Pople), JB (Johnson- Bovey) and HM (Haigh-Mallion).

Pople

As per the Pople[19] point dipole moment model, the expression for the change in the isotropic nuclear shielding constant in ppm originated by the ring current effect is given by

$$\Delta\sigma_{ring}^{PO} = 10^6 \times \frac{ne^2a^2}{4\pi mc^2} \times \frac{3\cos^2\Theta - 1}{r^3} \tag{2.3}$$

where the angle θ is between the query point and the ring normal, r is the distance from the center of the ring, n is the number of circulating electrons, a is the radius of the ring (taken to be benzene ring radius 1.39 Å) and e, m, c have their usual meaning.

Johnson-Bovey

In the Johnson-Bovey model,[20] the complete classical description of the electric current circulating in a loop of radius "a" is considered. The ring current model was also extended to account the nature of the π orbitals by assigning two loops, above and below the ring plane and both the loops possesses $n/2$ circulating electrons, and the form of the equation for $\Delta\sigma_{ring}$ in ppm is given by

$$\Delta\sigma_{ring}^{JB} = 10^6 \times \frac{ne^2}{12\pi mc^2 a} \times \sum_{p=1}^{2} \left\{ \frac{1}{\sqrt{(1+\rho)^2 + z_p^2}} \times C \right\} \tag{2.4}$$

where C is given by:

$$C = \left\{ K(k)) + \frac{1 - \rho^2 - z_p^2}{(1-\rho)^2 + z_p^2} E(k)) \right\} \tag{2.5}$$

where $\Delta\sigma_{ring}$ is expressed in a cylindrical coordinate system centered at the ring center with z and rho given in the unit defined by the loop radius a. K and E are the complete elliptic integrals given by following equations.

$$K(m) = \int_0^1 \left[\left(1 - t^2\right) \left(1 - mt^2\right) \right]^{-\frac{1}{2}} dt \tag{2.6}$$

$$E(m) = \int_0^1 \left(1 - t^2\right)^{-\frac{1}{2}} \left(1 - mt^2\right)^{\frac{1}{2}} dt \tag{2.7}$$

$$where, k = \left(4\rho / \left[(1+\rho)^2 + z_p^2 \right]\right)^{1/2} \tag{2.8}$$

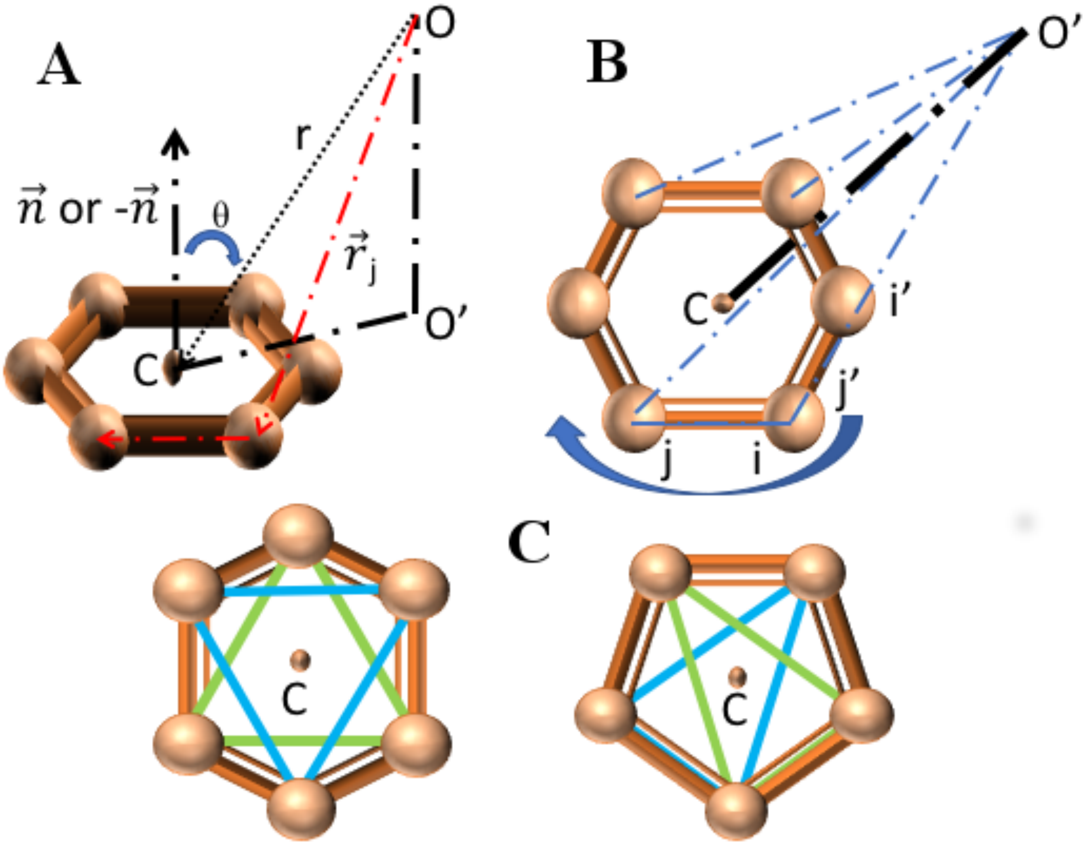


Figure 2.1: Geometric concepts used in the Ring Current Models

A separation of 1.28 Å(0.918 a, with radius a taken to be equal to the benzene ring radius) between the loops was found to be optimal to represent the hydrogen shielding in benzene.

Haigh-Mallion

$$\Delta\sigma_{ring}^{HM} = 10^6 \times KJ_{ring} \times \sum_{ij} S_{ij}\left\{\frac{1}{r_i^3} + \frac{1}{r_j^3}\right\} \tag{2.9}$$

In Haigh-Mallion[21] Ring current expression, J_{ring} is the algebraic (signed) triangle area formed by the O projection of the query point O onto the ring plane and the ring atoms i and j. r_i and r_j are the distances between O and atoms i and j, respectively.

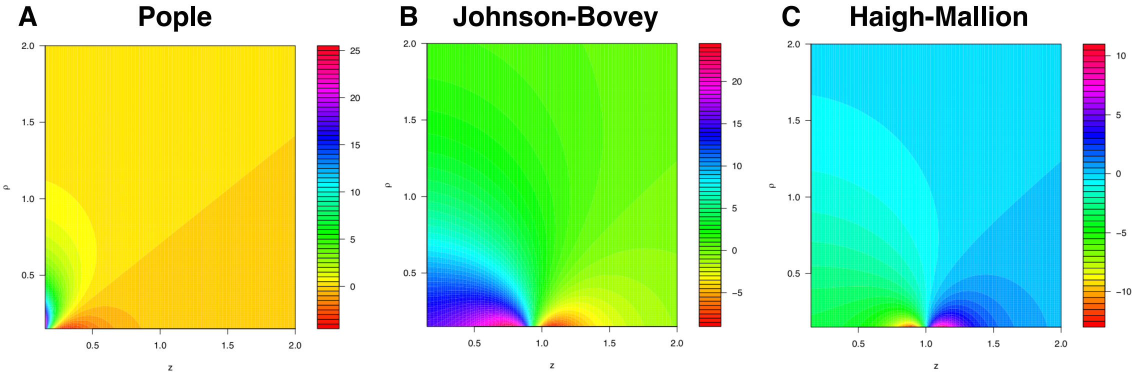


Figure 2.2: Maps of the ring current geometric factors in the proximity of a benzene ring for A)Pople B)Johnson-Bovey and C)Haigh-Mallion ring current models.

2.3.4 Stacking Interaction

The π stacking interactions are noncovalent interactions between aromatic rings as they consist of π bonds. These interactions are important in nucleo-base stacking within DNA and RNA molecules, which influence the electronic cloud around a nucleus and subsequently the chemical shift. Stacking interaction numbers were calculated similar to hydrogen bonding with a distance cutoff of 5 $\mathring{A}$ and angle cutoff of 30°. The distance between the two rings should be less than the distance cutoff and the angle between two normals of the ring plane should also be within the angle cutoff, with both conditions simultaneously true.

2.3.5 Torsion angles

Torsion angles in RNAs are dihedral angles, which are defined by 4 points in space. The six main chain torsion angles $(\alpha, \beta, \gamma, \delta, \epsilon, \zeta)$ around the covalent bonds, χ around the glycosidic bond, and five around the sugar pucker $(\upsilon_1, \upsilon_2, \upsilon_3, \upsilon_4)$ were calculated.

2.3.6 Polarization Effects

The electric field from a polar group or heavy atom can polarize an atom, thereby changing the local electron density by increasing or decreasing the shielding around that particular atom. The polarization effects were evaluated by the Buckingham equation[22] within the $10\mathring{A}$ distance cutoff as the electric filed effects decays with increasing distance:

$$\delta_{po} = -2 \times 10^{-12} E_Z - 10^{-18} E^2 \tag{2.10}$$

Here E is the electric field effect and E_z is the component of the filed in the direction of the bond. For computing the electric field, the partial charges were taken from the Amber topology files which were generated for all the RNAs.

2.3.7 Machine Learning

"A computer program is said to learn if its performance at a task T, as measured by a performance P, improves with experience E". ML is a sub-field of computer science in which computer (machine) can learn by itself (learning) given some initial knowledge. This initial knowledge is provided regarding training data and tested on a testing data. In our case, all of the structural features are calculated (serve as features for the various models), which are trained on a training set of 19 RNAs. The RNA structures were taken from the Protein Data Bank (PDB) and chemical shifts from the BMRB database.The RNAs were checked for referencing errors in ^{13}C and ^{1}H. chemical shifts assignments and were correctly referenced. The model is then tested on a set of 36 testing RNAs. Various ML models are tested, including linear and ensemble models, and the accuracy of each model is assessed.

Out of various linear ML methods Lasso Lars CV (LLCV) and out of Ensemble methods Extra-Randomized trees (ET) have been observed to have the best accuracy (Table 2.2). Lasso-Lars CV is a linear machine learning model trained with L1 prior as regularizer (Equation 2.11), which is numerically efficient when the number of dimensions is significantly greater than the number of points and takes care of the case in which two variables are correlated with the response.

$$LLCV_{CF} = min_w \frac{1}{2n_{samples}} \left\| Xw - y \right\|_2^2 + \alpha \left\| w \right\|_1 \tag{2.11}$$

Ensemble methods have emerged as the state-of-the-art ML methods which average the predictions of the independent decision trees generated by random subset of training database. For example in random forest machine learning algorithm, a random subset of the features are selected, and the best split is chosen at each node based on the subset of data. In extra randomized trees, randomness goes one step further in the way splits are computed (randomized) as opposed to the random forest. Extra randomized trees reduces the variance of the model a bit more, at the expense of a slightly greater increase in the bias.

2.3.8 Assessing Model Accuracy

We investigate the accuracy of various ML algorithms and ring current models, to predict the structure of RNAs in the presence of ligands. We use the predictors trained here on apo-chemical shift data to predict the holo-chemical shifts of 4 RNA-ligand complexes. We use the Normalized Sum of Logarithmic Ranks (NSLR), an error estimate between measured and predicted holo chemical shifts which quantifies the ability to discriminate the native-like molecules among a set of decoy pool of structures given by:

$$NSLR = \frac{SLR}{SLR_{max}} \tag{2.12}$$

Here SLR is defined as below and r_i is the rank achieved by the i_{th} structure by ordering in ascending order based on error type, and n is the total number of native-like structures and N is the total number of structure (native and non-native).

$$SLR = -\sum_{i=1}^{n} log(\frac{r_i}{N}) \tag{2.13}$$

Train Data Set: PDB (BMRB)				
1NC0 (5655)	2GM0 (7098)	2JXQ (15571)	2K3Z (15780)	2K41 (15781)
2JXS (15572)	2KYD (16980)	1KKA (5256)	1LC6 (5371)	1LDZ (4226)
1OW9 (5852)	1PJY (5834)	1R7W (6076)	1R7Z (6077)	1UUU (1UUU)
1YSV (6485)	2FDT (10018)	2KOC (5705)	2LBJ (17563)	2LBL (17565)
2LDL (17671)	2LDT (17682)	2LHP (17860)	2LI4 (17877)	2LK3 (17972)
2LP9 (18239)	2LPA (18240)	2LU0 (18503)	2LUB (18515)	2LV0 (18549)
2RN1 (11014)	2Y95 (16714)	4A4S (18036)	4A4T (18034)	4A4U (18035)
Test Data Set: PDB (BMRB)				
1SCL (1SCL)	1XHP (6320)	1Z2J (6543)	1ZC5 (6633)	2JWV (15538)
2K63 (15856)	2K64 (15857)	2K65 (15858)	2K66 (15859)	2LPS (5962)
2LQZ (18336)	2LUN (18532)	2LX1 (18656)	2M12 (18838)	2M21 (18891)
2M22 (18892)	2M23 (18893)	2M24 (18894)	2M8K (19260)	2MEQ (18975)
2MHI (19634)	2MI0 (19662)	2MIS (19692)	2QH2 (7403)	2QH3 (7404)
2QH4 (7405)				

Table 2.1: List of PDB ids and BMRB ids for train and test data sets

2.4 Results and Discussion

A collection of a training and test database (Table 2.1) of RNA and their chemical shifts are taken from the PDB and BMRB respectively. The extracted features from a given RNA are used to build predictors using the sci-kit learn python module.[23] These predictors were trained on a training database and tested on a testing set. The training database is carefully designed to ensure that the RNAs in the database provide sufficient coverage of known RNA chemical space.

2.4.1 Ensemble methods are more accurate than linear ones at predicting the chemical shift of RNAs

We applied different ML algorithms ranging from linear to the ensemble are summarized in (Table 2.2). From the results, we can observe that the Lasso Lars CV (LLCV) method perform best out of all the linear methods (MAE of 0.819; Reference Table 2.2) and extra randomized trees (ET) method perform best out of all the ensemble methods (MAE of 0.748; Reference Table 2.2). Both of the algorithms, LLCV and ET, help in reducing the variance at the cost of increase in bias to curb model over-fitting.

Linear	MAE(^{13}C)	MAE(^{1}H)	Ensemble	MAE(^{13}C)	MAE(^{1}H)
Linear Regression	0.893	0.159	Random Forest	0.757	0.145
LassoLarsCV	0.819	0.153	Extra Randomized	0.748	0.140
Ridge Regression	0.863	0.167	Gradient Boosting	0.854	0.150
Bayseian Ridge	0.866	0.169	Bagging	0.756	0.144

Table 2.2: Represents the Mean Absolute Errors (MAE) in ppm for different machine learning algorithms for ^{13}C and ^{1}H.

2.4.2 Pople model shows similar performance in terms of predicting CS as the other two ring current models

Next we examined the impact of various ring current models on chemical shift prediction accuracy. In particular, we explored three popular models have been developed: 1) JB, 2) PO, and 3) HM.[24,25] Extensive studies have been carried out on parameterizing these models, yet a side-by-side comparison of the performance of these models has yet to be carried out. In the present work, we build a set of empirical chemical shift predictors using these three ring-current models (Table 2.4) and then compared that based on their ability to (1) recapitulate experimental chemical shifts and (2) discriminate native from decoy models of the same RNA. As shown by Table 2.4 Pople (PO) model shows similar performance in terms of predicting CS as the other two models namely Johnson-Bovey (JB) and Haigh-Mallion (HM). The carbon chemical shifts MAE errors of PO, JB and HM are 0.754, 0.751 and 0.753 respectively. Chemical shift MAE error of 0.140 (Table 2.4) for hydrogen atom is exactly the same for PO, JB and HM ring current models.

RC Model	No RC	Pople	Johnson-Bovey	Haigh-Mallion
ExtraRandomized	0.773/0.158	0.754/0.140	0.751/0.140	0.753/0.140
LassoLarsCV	0.824/0.166	0.815/0.152	0.819/0.153	0.818/0.152

Table 2.3: Represents the Mean Absolute Errors (MAE) for different RC models for $^{13}C/^{1}H$.

RC Model	Pople	Johnson-Bovey	Haigh-Mallion
Computation time	2.59 sec	4.01 sec	33.17 sec

Table 2.4: Represents the computational time for each ring current model on a model benzene ring system.

Based on similar accuracy (as shown by Table 2.3) of all the ring current models in predicting chemical shifts, we next wanted to assess which model is the fastest in terms of computational power. We did a benchmark study on the benzene ring molecule and found that Pople model is the fastest in terms of computing power. It took only 2.59 seconds for Pople model to compute the ring current on a benzene ring model system with 1.8 GHz Dual-Core Intel Core i5 processor as opposed to 4.01 seconds for Johnson-Bovey and 33.17 seconds for Haigh-Mallion (Table 2.4). Based on similar accuracy of all three different ring current models, we decided to use Pople model for CS prediction based on it's computational efficiency.

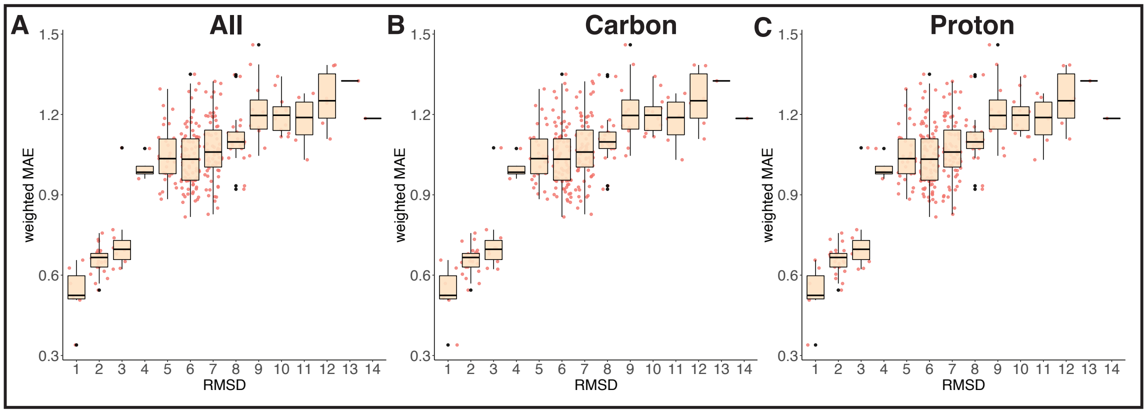


Figure 2.3: Represents the sensitivity analysis (weighted MAE v/s RMSD) for RNA 1LC6 with Pople ring current model and extra randomized tree machine learning algorithm for A)all, B)carbon and C)proton atom types.

2.4.3 Chemical shift errors are positively correlated with structural dissimilarity

The plot in Figure 2.3 shows the weighted mean absolute error (which is calculated on the predicted chemical shifts) v/s the Root Mean Square Deviation (RMSD) relative to the solved structure. The plot represents the sensitivity analysis for a structurally diverse conformational pool for the RNA PDB id of 1LC6. A structure with an RMSD of less than $2\AA$ is considered "native-like", and greater than $2\AA$ is considered "non-native". As the RMSD increases (non-native), the error between predicted and measured chemical shifts (weighted MAE) should also increase for a particular structure in a diverse conformational pool which is shown in Figure 2.3 as the structure that have low error, have low RMSD, and which have high error, have high RMSD for all, carbon and proton atom types.

RNAs	ET	LLCV	Combine	RNAs	ET	LLCV	Combine
2L1V	0.96	0.85	0.59	2L94	0.48	0.66	0.51
2LWK	0.39	0.57	0.49	2M4Q	0.48	0.44	0.28

Table 2.5: Given are the Normalized Sum of logarithmic ranks (NSLR) values for different Machine Learning methods with ET: Extra Randomized Trees, LLCV: Lasso-LARS linear method, combine: combination of ET and LLCV methods.

The NSLR ranges from 0 and 1, where a value of 1 corresponds to the perfect resolving power. Table 2.4 represents the results of the 4 RNA-ligand complexes studied and we found that predictors trained on just the apo-chemical shift data were able to resolve RNA-ligand complexes for at least 3 out of the 4 structures. Building on the use of chemical shift to predict the structure of RNA-ligand complexes, we next explore (in Chapter 3) the use of predicted chemical shifts to resolve the RNA-ligand structure starting from sequence of the RNAs.

2.4.4 Conclusion

In this study we have shown that, ET and LLCV ML algorithms were the most accurate ones in the ensemble and linear methods tested respectively, at predicting the chemical shifts of RNAs. We also observed the effect of different atom types on chemical shift prediction and found that for carbon, LLCV method is the most sensitive whereas for protons, ET is the most sensitive. The weighted MAE between predicted and measured CS increases as the structure becomes more non-native in a pool of decoy structures.

BIBLIOGRAPHY

[1] David R Calabrese, Colleen M Connelly, and John S Schneekloth Jr. Ligand-observed NMR techniques to probe RNA-small molecule interactions. *Methods in enzymology*, 623:131–149, 2019.

[2] Sebastian Doniach and Jan Lipfert. Use of small angle x-ray scattering (SAXS) to characterize conformational states of functional RNAs. In *Methods in enzymology*, volume 469, pages 237–251. Elsevier, 2009.

[3] Zhuoqin Yu, Pengfei Li, and Kenneth M Merz Jr. Using ligand-induced protein chemical shift perturbations to determine protein–ligand structures. *Biochemistry*, 56(18):2349–2362, 2017.

[4] Mattia Sturlese, Massimo Bellanda, and Stefano Moro. NMR-assisted molecular docking methodologies. *Molecular informatics*, 34(8):513–525, 2015.

[5] Mike P Williamson. Using chemical shift perturbation to characterise ligand binding. *Progress in nuclear magnetic resonance spectroscopy*, 73:1–16, 2013.

[6] Ulrich Schieborr, Martin Vogtherr, Bettina Elshorst, Marco Betz, Susanne Grimme, Barbara Pescatore, Thomas Langer, Krishna Saxena, and Harald Schwalbe. How much NMR data is required to determine a protein–ligand complex structure? *ChemBioChem*, 6(10):1891–1898, 2005.

[7] Marina Cioffi, Christopher A Hunter, Martin J Packer, and Andrea Spitaleri. Determination of protein–ligand binding modes using complexation-induced changes in 1H NMR chemical shift. *Journal of medicinal chemistry*, 51(8):2512–2517, 2008.

[8] Domingo González-Ruiz and Holger Gohlke. Steering protein- ligand docking with quantitative NMR chemical shift perturbations. *Journal of chemical information and modeling*, 49(10):2260–2271, 2009.

[9] Jaime Stark and Robert Powers. Rapid protein- ligand costructures using chemical shift perturbations. *Journal of the American Chemical Society*, 130(2):535–545, 2008.

[10] Clémentine Aguirre, Tim ten Brink, Olivier Cala, Jean-François Guichou, and Isabelle Krimm. Protein–ligand structure guided by backbone and side-chain proton chemical shift perturbations. *Journal of biomolecular NMR*, 60(2-3):147–156, 2014.

[11] Suzanne B Shuker, Philip J Hajduk, Robert P Meadows, and Stephen W Fesik. Discovering high-affinity ligands for proteins: SAR by NMR. *Science*, 274(5292):1531–1534, 1996.

[12] Xavier Morelli, Alain Dolla, Myrjam Czjzek, P Nuno Palma, Francis Blasco, Ludwig Krippahl, Jose JG Moura, and Françoise Guerlesquin. Heteronuclear NMR and soft docking: An experimental approach for a structural model of the cytochrome c 553- ferredoxin complex. *Biochemistry*, 39(10):2530–2537, 2000.

[13] Mark A McCoy and Daniel F Wyss. Alignment of weakly interacting molecules to protein surfaces using simulations of chemical shift perturbations. *Journal of biomolecular NMR*, 18(3):189–198, 2000.

[14] Mark A McCoy and Daniel F Wyss. Spatial localization of ligand binding sites from electron current density surfaces calculated from NMR chemical shift perturbations. *Journal of the American Chemical Society*, 124(39):11758–11763, 2002.

[15] Bing Wang, Kaushik Raha, and Kenneth M Merz. Pose scoring by NMR. *Journal of the American Chemical Society*, 126(37):11430–11431, 2004.

[16] Ivano Bertini, Marco Fragai, Andrea Giachetti, Claudio Luchinat, Massimiliano Maletta, Giacomo Parigi, and Kwon Joo Yeo. Combining in silico tools and NMR data to validate protein-ligand structural models: Application to matrix metalloproteinases. *Journal of medicinal chemistry*, 48(24):7544–7559, 2005.

[17] Yuya Kodama, Koh Takeuchi, Nobuhisa Shimba, Kohki Ishikawa, Ei-ichiro Suzuki, Ichio Shimada, and Hideo Takahashi. Rapid identification of ligand-binding sites by using an assignment-free NMR approach. *Journal of medicinal chemistry*, 56(22):9342–9350, 2013.

[18] F Ribas Prado and C Giessner-Prettre. Parameters for the calculation of the ring current and atomic magnetic anisotropy contributions to magnetic shielding constants: Nucleic acid bases and intercalating agents. *Journal of Molecular Structure: THEOCHEM*, 76(1):81–92, 1981.

[19] JA Pople and KG Untch. Induced paramagnetic ring currents. *Journal of the American Chemical Society*, 88(21):4811–4815, 1966.

[20] CE Johnson Jr and FA Bovey. Calculation of nuclear magnetic resonance spectra of aromatic hydrocarbons. *The Journal of Chemical Physics*, 29(5):1012–1014, 1958.

[21] CW Haigh and RB Mallion. ring-currenteffects on 1H-NMR chemical shifts in linear acenes. *The Journal of Chemical Physics*, 76(8):4063–4066, 1982.

[22] AD Buckingham. Chemical shifts in the nuclear magnetic resonance spectra of molecules containing polar groups. *Canadian Journal of Chemistry*, 38(2):300–307, 1960.

[23] David Cournapeau. scikit-learn: machine learning in python.

[24] Aleksandr B Sahakyan and Michele Vendruscolo. Analysis of the contributions of ring current and electric field effects to the chemical shifts of RNA bases. *The Journal of Physical Chemistry B*, 117(7):1989–1998, 2013.

[25] GG Hall and A Hardisson. Ring currents and their effects in aromatic molecules. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 268(1334):328–338, 1962.

CHAPTER III

De Novo Prediction of RNA-Ligand Structure Guided by NMR Chemical Shifts

3.1 Statement of Contribution

1. **Aaron T. Frank, PhD:** Conceived of the project described in this chapter.
2. **Sahil Chhabra, PhD Candidate (Chemistry and Scientific Computing):** Generated the data and carried out the analysis described in this chapter; Independently wrote this chapter.
3. **Kexin Zhang, PhD Candidate (Chemistry and Scientific Computing):** Carried out the CS-Folding calculations; Prepared Figure 3.3.

3.2 Introduction

The recognition of small molecule compounds by RNA play important roles in regulating cellular function.[1] For instance, riboswitches which are a class of RNA that recognize small molecule ligands, bind to the cognate ligand, and turn on or off gene expression.[2,3] Also, several other classes of RNA such as microRNAs, group-II introns, and other structured mRNA elements can recognize small molecule ligands which in turn modulate their activity.[4–6] As such, structured RNAs have emerged as drug targets.[7–14]

Recent advances in understanding how these RNAs interact with ligands have revealed a lot of biological information.[15–18] However, approaches to correctly identify small molecule ligands in complexes with these biological regulatory elements have been limited. Determining the structure of RNA in complex with such small molecules is crucial to understanding the mechanism by which RNAs achieve specificity and selectivity for particular molecules.[19,20]

In principle, the structure of RNA in complex with small molecule ligand can be characterized using X-ray crystallography or NMR spectroscopy.[21–24] However, these methods can be time consuming and costly. As an alternative, computer docking can be used to predict the 3D structure of an RNA in complex with a small molecule.[25–27] However, current state-of-the-art methods fail at distinguishing native-like structures from incorrect, non-native structures.

The secondary structure prediction methods have been well developed by many researchers, with the most prominent among them being from the Mathews lab.[28–32] Most popular and widely used methods are the physics based free energy minimization approaches, and their performances to predict the secondary structure have been found to be 70%. The RNAstructure Fold program[33] devised by the Mathews lab predicts the most likely secondary structure that will occur at equilibrium on the basis of free energy minimization. The correct secondary structure is crucial to predict the 3D structure of the RNAs.

The 3D structure prediction of RNAs from secondary structure is mostly based on the fact that RNA folding is a hierarchical process.[34–36] Answers to the RNA folding problem, compared to the protein folding problem, are at an early stage, as current 3D RNA folding algorithms require manual manipulation or are generally limited to simple structures. Nonetheless many research groups have made strides in this direction and techniques like NAST,[37] BARNACLE,[38] iFoldRNA[39] and FARFAR[40] are available to the academic community for the prediction of RNA structure from secondary structure with varied accuracies. FARFAR can produce the best prediction model for small RNA sequences (20 nt). NAST and iFoldRNA have been shown to do better if SHAPE chemistry data is available. MC-SYM[41] has decent accuracy for medium sized RNAs but is limited by the difficulty of predicting long-range contacts. We used the FARFAR method incorporated in the Rosetta suite in this study to generate the de novo 3D RNA models. FARFAR was tested on a dataset of 43 structures of various lengths and motifs, and most predictions were found to have large RMSDs compared to the crystal structure (in the range of 6 or greater). That study also showed that the prediction accuracy improved with additional datapoints including the 2D structure and 3D contacts, but failure to detect long-range interactions remain a clear challenge.

Below, I describe my first attempt to enhance our ability to predict the structure of RNA-ligand complexes. Specifically, I describe the results of hybrid approach in which NMR derived CS are used to model the 3D structure of an RNA-small molecule complex, *starting from sequence information, only.* First, I will describe the CS guided model framework that we employed. Then the results of its application on the solution structure of a small molecule influenza RNA complex (2LWK), which we use here as a model system, will be discussed. Using the influenza A virus promoter RNA complex with a small-molecule (PDBID 2LWK, BMRBID 18633)[42] that inhibits viral replication as a model system, we have discovered that not only can computational methods sample native-like conformations of the RNA, but more importantly, the modeled RNA structure that best agree with CS data are within 3 of the NMR structure. Moreover, when the ligands are docked onto this structure, we can sample models of the full complex that are within 4 of the correct structure.

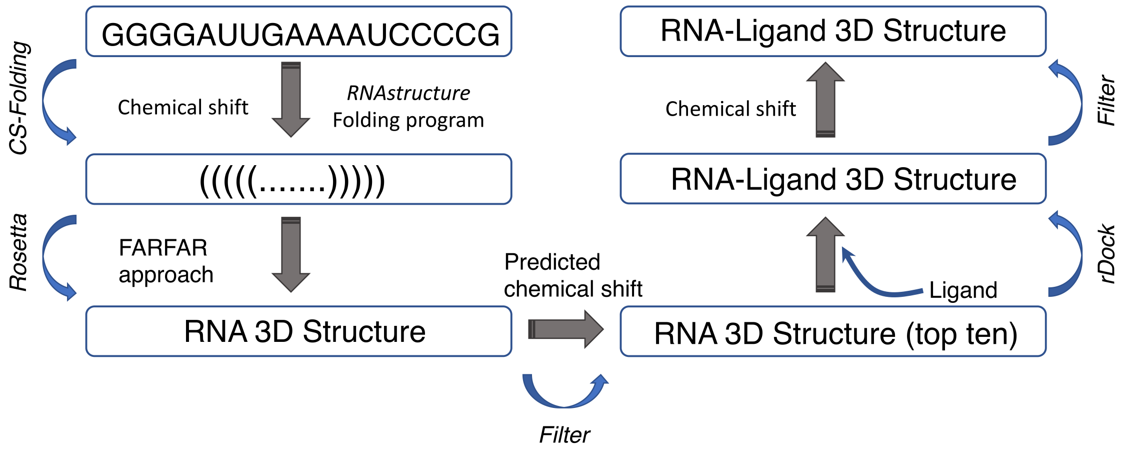


Figure 3.1: Schematic of the hybrid approach which uses NMR derived CS to model the 3D structure of an RNA-ligand complex starting from sequence information

3.3 MATERIALS AND METHODS

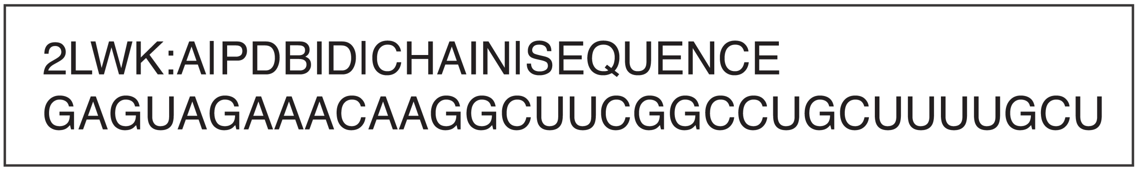
Secondary structures of RNAs and RNA-containing complexes are crucial for the understanding of their 3D structures and functions. RNAStructure[43] is one of the most widely used tool to predict RNA secondary structure. RNAStructure uses a free energy minimization algorithm that predicts the lowest free energy structure, that is, the most probable secondary structure. It also predicts other low free energy structures, called "suboptimal" structures, as possible alternative structures. The performance of physics-based free energy minimization approaches like RNAStructure is found to be 70% when comparing the predicted and native base pairs in an RNA. (Mathews et al. 2004; Xu et al. 2012). The accuracy of the predicted secondary structure impacts the tertiary structure prediction of a RNA. Hence, the accuracy of the former is critical to the performance of the latter. We used a CS guided approach devised by Kexin Zhang[44] (CS-Fold) in our lab to predict the secondary structure of RNA. From the best predicted secondary structure we used

Rosetta to generate de-novo 3D models and used CS predicted by LarmorD to assess the nativeness of the RNA structures.

3.3.1 Using RNAStructure to predict the secondary structure

We used the secondary structure prediction package named RNAStructure^{43} to predict the secondary structure of the RNA. We used the fold algorithm in RNAStructure which predicts the lowest free energy structure along with a set of low free energy structures. The fold algorithm incorporate four different prediction algorithms: partition function calculation, maximum free energy (MFE) structure prediction, finding structures with maximum expected accuracy (MEA), and pseudoknot prediction. RNAStructure Fold uses dynamic programming which ensures that the predicted lowest free energy structure is found. The entire secondary structure prediction problem is divided into smaller problems and then recursion is used to build the complete structure. The partition function is also calculated using a dynamic programming algorithm. Fold takes a sequence of RNA as the input and creates a group of secondary structures annotated with their corresponding prediction probabilities. It also includes other structures with varied probabilities of correctness as the minimum free energy structure may not be the correct one. We used the RNAStructure web server to predict the secondary structure of the RNA. http://rna.urmc.rochester.edu/RNAstructureWeb/ Servers/Predict1/Predict1.html. The input is a FASTA file/sequence, and for 2LWK the input looks like this:

We used the following parameters for secondary structure prediction in RNAStructure: T = 310.15 K, Max loop size =30, Maximum % Energy Difference (MFE, MEA)= 10, Maximum Number of Structures (MFE, MEA) = 10, Window Size (MFE, MEA) = 3, Gamma (MEA) =1, Iterations (Pseudoknot Prediction) = 1,



```
2LWK:A|PDBID|CHAIN|SEQUENCE
GAGUAGAAACAAGGCUUCGGCCUGCUUUUGCU
```

Figure 3.2: Fasta file example for 2LWK

Minimum Helix Length (Pseudoknot Prediction) =3. No constraints were provided for the secondary structure prediction process. We took the minimum free energy structure from Fold and used that to predicted the 3D structure.

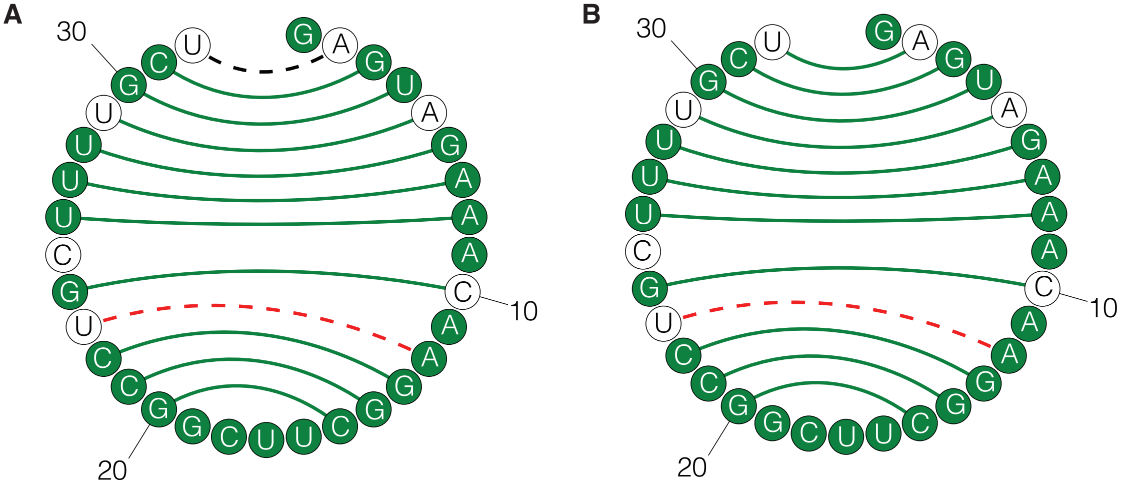


Figure 3.3: Secondary Structures generated by A) CS-Fold and B)ProbKnot for 2LWK

3.3.2 Chemical shift guided secondary structure prediction

We used another approach called CS-Fold developed by Kexin Zhang in the Frank lab to predict the secondary structure of RNA. CS-Folding framework uses assigned CS data to guide RNA secondary structure prediction. CS-Fold built an ML model that used CS to determine base-pairing status of individual RNAs. Then the predicted base-pairing status were used as restraints to guide the secondary structure prediction in the folding algorithm. It takes in input a sequence and CS file. We got the CS assignments for 2LWK from the BMRB database: http://www.bmrb.wisc.edu/. We also explored ProbKnot in RNAStructure which predicts a secondary structure of probable base pairs, which might include pseudoknots. Both the predicted structures for PDB id 2LWK are represented in Figure 3.1 for comparison.

3.3.3 Using Rosetta to generate RNA 3D Structures

We used Rosetta's Fragment Assembly of RNA with Full Atom Refinement (FARFAR) approach to produce de novo models of small RNA motifs. The FARFAR modeling algorithm for RNA structure in Rosetta is based on short fragments assembly from the existing RNA crystal structures with matching subsequences to the target RNA. The algorithm is comprised of 2 main steps. The first is Fragment Assembly of RNA i.e. FARNA, which is a Monte Carlo process guided by a low-resolution knowledge based energy function. The second step is refinement in an all-atom potential to yield more realistic structures with fewer clashes between the atoms. The output also provides an energy score, which is used to discriminate native-like conformations from non-native conformations.

The input for Rosetta is the secondary structure file in dot-bracket notation. Using secondary structure as the input, we create a set of low resolution models using FARNA. We compared our generated models to the reference model by using the RMSD values as the evaluation metric. We compared these RMSD values to the Rosetta energy scores to compare the native and non-native conformations and their energies. For FARNA step, we used the default mode that runs Monte Carlo fragment assembly optimized in a knowledge-based low-resolution potential. Next we did refinement (minimize rna) in the high-resolution Rosetta potential. The final models resulted in few steric clashes and improved energy scores. Those refined models also contain few chainbreaks and unrealistic atomic-level geometries which could be present due to the FARNA sampling method which uses rigid fragments of crystallographic RNA structures. This strategy of Full-Atom Refinement with FARNA (FARFAR) accounts for physical and chemical interactions like van der Waals, hydrogen bonding, backbone torsion angles and desolvation penalties for polar groups within RNA.

3.3.4 Generating RNA-ligand decoys

We generated the diverse decoy sets for each of the top 10 RNA-ligand systems obtained after the CS-Fold and Rosetta filtering using the rDock computer docking program. Blind docking was carried out in each of those structures using the 2 sphere method with outer spehere radii set to 50 Å. RbtSphereSiteMapper site mapper algorithm was used and the pocket detection was carried out using the rbcavity utility program. Maximum cavities was set to 4 with a minimum volume of 100 units. We generated 1000 docks for every top ten RNA structure, totalling 10k docks in total.

3.3.5 Assessing Structures

We used CS predicted by LarmorD to assess the quality of the structures generated by Rosetta. However to predict the CS for the RNA-ligand complex, we used the CS predictors described in chapter 2. LarmorD uses a interatomic distance based approach to build models that can predict the CS of the atoms in the RNA. It is trained on a set of 35 RNA structures taken from RCSB database and it is pretty fast and simply way to compute the shifts. The accuracy of LarmorD as reported in the paper is 0.19 and 1.09 ppm for protons and carbons, respectively. By comparison the Pople Ring current model described in Chapter 2 has prediction accuracy of

0.14 and 0.75 ppm for protons and carbons, respectively. We then compared the predicted and actual CS. We used the weighted Mean Absolute Error (rMAE) as our comparison metric as the weights scales the error such that nuclei with different ranges contribute proportionally to the MAE (for example 1H and 13C).

3.4 Results and Discussion

In this study, we explored the use of CS to guide the 3D structure prediction of RNA starting from it's sequence. We use the small molecule-influenza RNA complex (PDB id: 2LWK) as the test system to demonstrate this approach. We applied CS-Fold, a method developed by Zhang, which deploys CS to guide the secondary structure prediction of RNAs. From the best predicted secondary structure, we generated de novo 3D models of RNAs using the Fragment Assembly of RNA with Full Atom Refinement (FARFAR) approach. FARFAR uses a low-resolution knowledge based energy function and monte carlo sampling algorithm to produce motifs through Fragment Assembly of RNAs. The bottleneck for this approach is to achieve complete conformational sampling at the atomic level resolution. We use CS predicted by LarmorD to refine those generated structures in an attempt to address these shortcomings. We also explored the use of CS to identify native RNA-ligand structures from a set of decoy pools. We used rDock to dock the ligand in top 10 best predicted 3D structures of the RNA.

3.4.1 RNAStructure and Rosetta exhibits low recovery rates for RNA structure prediction

We began our study by assessing the ability of RNAStructure predicted secondary structures to generate the 3D models of RNAs. We used Rosetta and then CS errors to discriminate native like structures. To achieve this, we generated 10k Rosetta 3D

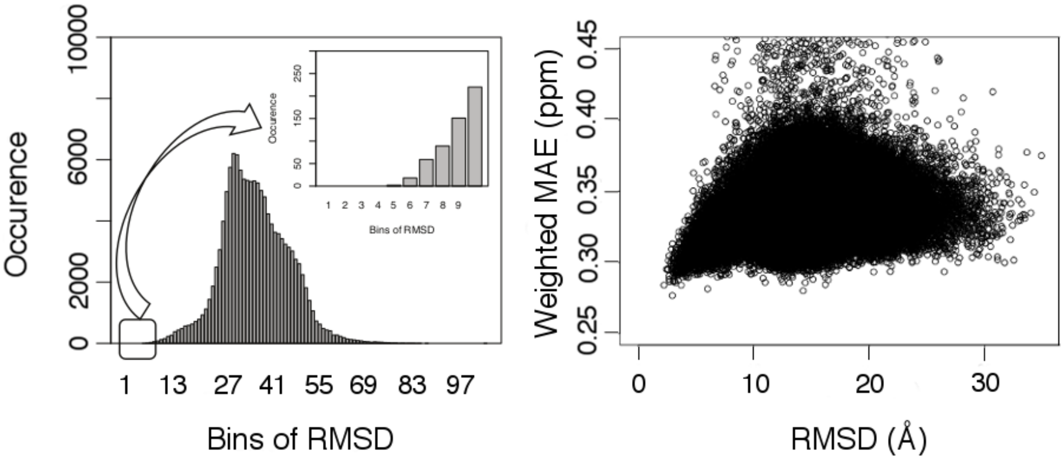


Figure 3.4: Left: Distribution of RMSDs for the generated 3D structures of the RNA by Rosetta after the minimization. Each bin in the plot comprises of 0.5 *Å* spacing. Right: weighted MAE of Chemical Shifts v/s RMSD for the predicted 3D structures of 2LWK using RNAStructure and Rosetta.

models for each of the 13 secondary structures predicted by RNAStructure. The RMSD distribution of all the 130k structures is shown in Figure 3.2. Part A in figure 3.2 represents the distribution after the FARNA step and before the minimization. Part B in figure 3.2 represents the distribution after the minimization step. Each bin in the plot contains 0.5 Åof spacing. The range of RMSDs before the minimization was 2.04 to 23.54 *Å* and after minimization was 2.22 to 41.79 *Å*. Refinement using FARFAR improves low-resolution models by relaxing them into more realistic conformations. We then took these 130k structures and predicted the CS for all of the structures using LarmorD. We then compared the ability of the Rosetta scores and CS-errors to predict the correct 3D structure of RNA. We first rank them based on Rosetta scores and top 10 structural RMSD, Rosetta scores and their corresponding CS errors are shown in Table 3.1 (left). We do not observe even a single structure within 3 Å(0% recovery rate) when ranked based on Rosetta scores. In fact, all the structures have RMSDs greater than 10 Å. The average CS error of top 10 structures was 2.30 ppm.

S.N.	RMSD(Å)	Chemshift Errors(ppm)	Rosetta Scores	RMSD(Å)	Chemshift Errors(ppm)	Rosetta Scores
1	11.79	2.33	-197.31	14.2	2.04	-133.69
2	11.41	2.25	-195.44	2.90	2.05	-149.69
3	13.11	2.32	-190.88	16.10	2.07	-140.38
4	11.85	2.36	-190.08	5.90	2.07	-117.80
5	12.65	2.26	-188.56	14.40	2.08	-159.69
6	12.11	2.23	-188.25	16.10	2.082	-145.27
7	11.53	2.22	-188.01	2.22	2.084	-166.62
8	12.04	2.39	-187.69	12.40	2.086	-153.28
9	12.37	2.28	-187.48	13.00	2.092	-170.49
10	13.00	2.26	-187.44	16.40	2.094	-97.60

Table 3.1: Left: Ability of the Rosetta scores and Right: CS errors to predict the correct 3D structure of RNA.

Next we asked if the predicted CS could recover native like structures from the decoy of 130k structures. Ranked are all the structures based on the CS error as shown in Table 3.1 (Right). We observe two structures within 3 Ånative-ness cutoff (20% recovery rate in top 10). The two lowest CS error structures with RMSD of 2.90 and 2.22 had Rosetta scores of -149.69 and -166.62 respectively. CS error based filtering is successful in identifying native like structures from a set of decoy pool as shown in Figure 3.3. This CS error based scoring provides improvement over the Rosetta based scoring, but there still is a lot of scope for further improvement.

3.4.2 CS-Fold and Rosetta improves the recovery rates of RNA Structure Prediction

As eluded earlier the performance of secondary structure prediction is critical to the prediction of RNA 3D structure. So next we try to better predict the secondary structure using the CS-Folding approach. We took the best predicted secondary structure from the CS-Fold and followed the same protocol of generating the 3D structures using Rosetta. The RMSDs, CS errors, and Rosetta scores of top 10 predicted structures are given in Table 3.2 (structures ranked based on Rosetta scores). We observe 5 structures within 3 Ånativeness cutoff (50% recovery rate in top 10). Also 9 out of 10 structures are observed within 4 Å, giving a 90 % recovery rate under 4 Å. And all the structures are within 5 Å. Interestingly, the data frame is the same when ranked based on CS error. The 2 lowest CS error structures with CS error of 2.05 and 2.08 made in the 6th and 1st position respectively. These top 10 structures are represented in Figure 3.4 with native structure superimposed in gray color on each of the predicted structures. This approach based on the CS guided structure prediction, exhibited remarkable ability to predict the 3D structure of RNAs, starting with a sequence all the way to a 3D structure.

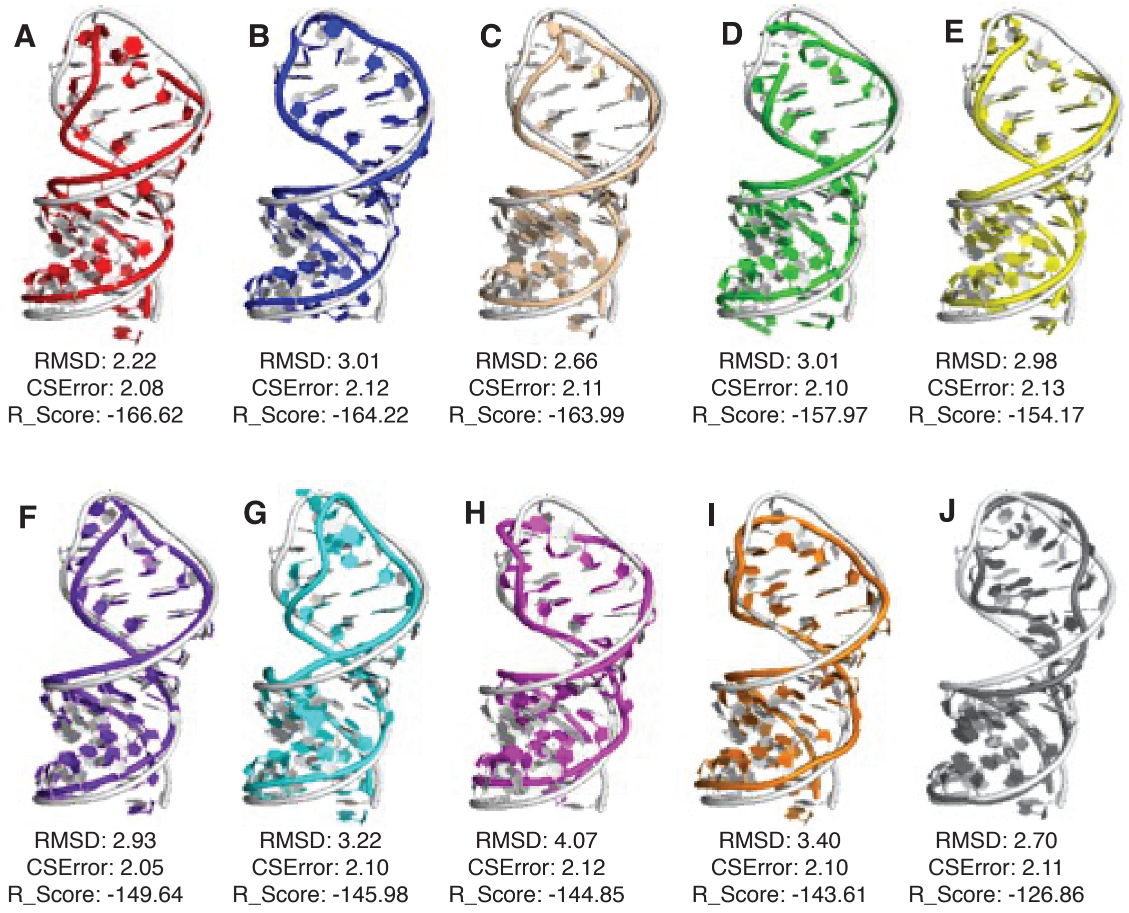


Figure 3.5: Top 10 RNA Structures for 2LWK predicted from sequence using CS-Fold and Rosetta. Also mentioned are the RMSD, Chemical Shift errors (CSErrors) and the Rosetta Scores (RScores). The Chemical Shifts were predicted using the LarmorD method. The structures are superimposed on native structure represented in gray on each image from A-J.

3.4.3 Chemical Shift Error v/s RMSD of the structure

Next we investigate the use of CS to recover native-like poses of RNA-ligand complexes. To achieve this, we took the 10 best predicted structures (as shown in Figure 3.4) from Rosetta and CS-Fold approach described above in this chapter. We used rDock to dock the ligand in all of the top 10 structures. We predicted CS in the presence of ligand using the chemical shift predictors described in Chapter 2, and used them to recover native like RNA-ligand poses. The results are shown in Figure 3.6 with the weighted MAE in predicting CS v/s the RMSD of the structures with respect to the native structure. The lowest CS error structure is the one in blue with RMSD >7 Å, and the second lowest is the one in orange with RMSD <3 Å. The remaining 8 structures had RMSDs of 3-4 Åbut had higher CS errors. There are many possible reasons for this low success rate in RNA-ligand complexes compared to the prediction of RNAs. First, the error is compiled over from the steps involved in predicting the RNA-ligand complex. Starting from the error in predicting sequence to secondary structure by RNAStructure or CS-Fold. Then the error in predicting the 3D structure from secondary structure using Rosetta, and lastly and more importantly the error occurred during the docking process using rDock. Apart with these 3 sources of error, it is also difficult to know the exact structure of the RNA-ligand complex in solution as most biologically relevant RNAs like the 2LWK are present as multiple conformations and isoforms thus rendering structural studies by NMR very difficult.

3.4.4 Conclusion

With our present knowledge of the RNA structure, from RNA sequences to RNA 3D structures, computational prediction of RNA tertiary structure with the help of experimental data has been developed over the years. Here in our study we wanted to improve on those approaches and test the use of CS to predict the RNA tertiary structure. We saw that the use of RNAStructure for secondary structure prediction combined with Rosetta de-novo approach for 3D motif prediction from secondary structures was unable to recover the native-like RNAs from a of decoy pool of 130k structures based on Rosetta scores. The success rate was sightly improved from 0% to 20 % when we used the predicted CS to score the structures based on the CS errors. Next we showed that CS-Fold CS guided secondary structure prediction combined with Rosetta de-novo for 3D motifs prediction protocol significantly enhance the recovery rates to 50%. Finally using rDock and CS error based filtering, we were able to recover native like RNA-ligand complexes starting from just the sequence of RNAs.

This approach established the foundation for using experimental data like CS to guide the prediction of RNA 3D structure. We have shown this approach work for an influenza A virus (PDB id 2WLK). In future work we will test this approach on other RNA-ligand systems. In this study we observed that both the rosetta scores and CS errors work very similarly in predicting the correct 3D structure of RNAs especially when using CS to guide the structure prediction. A follow-up to this study could combine Rosetta Scores and CS error to make a combine metric to filter native line RNA structure from the decoy pools. This approach can also be extended to RNA-containing complexes especially RNA-protein complexes in future.

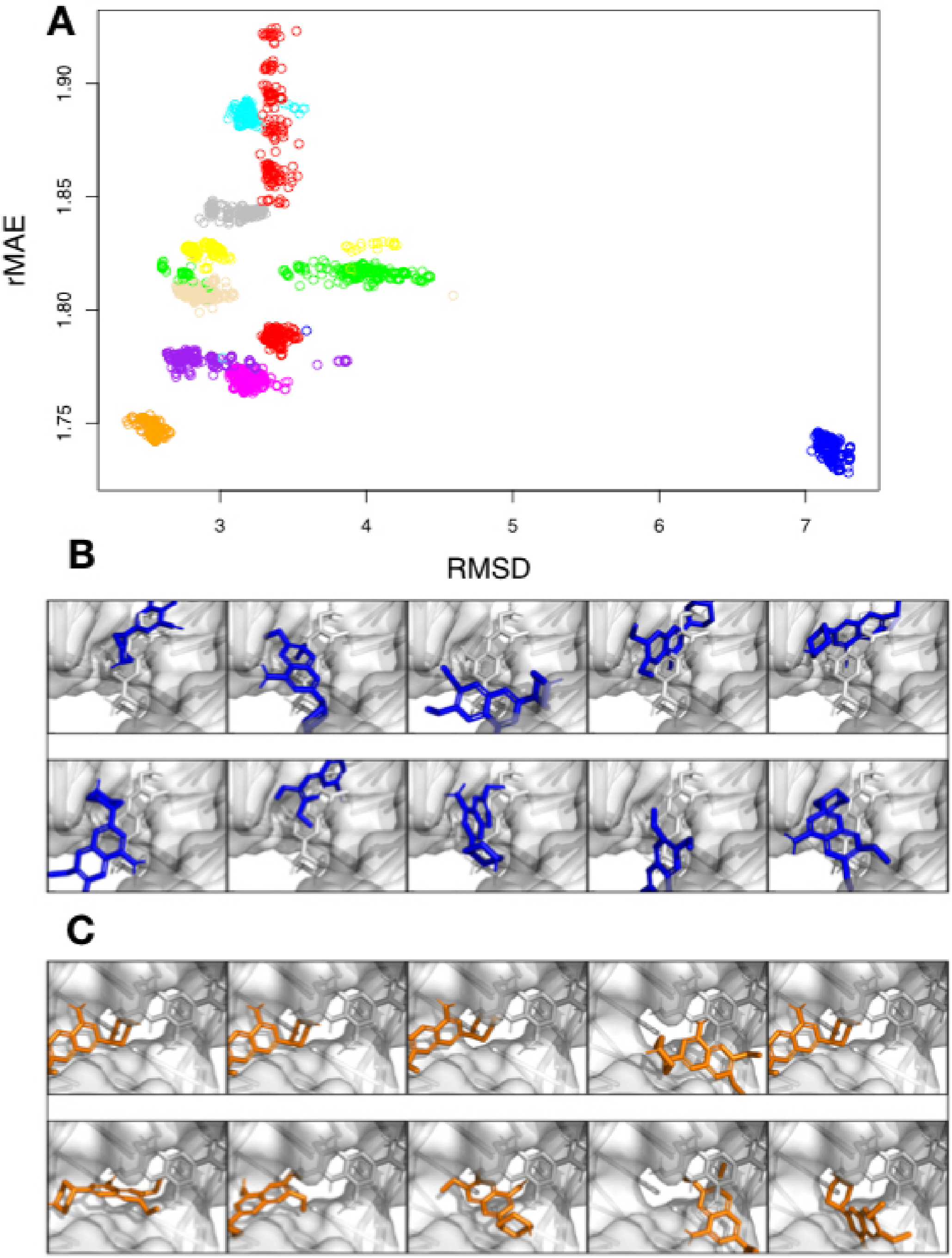


Figure 3.6: (A) rMAE v/s RMSD for the predicted RNA-ligand structures. The ligand was docked by rDock in top 10 structures shown in Figure 3.4. (B) Top 10 best predicted poses for the structure in blue in A. (C) Top 10 best predicted poses for the structure in orange in A.

BIBLIOGRAPHY

[1] Michelle H. Moon Colleen M. Connelly and Jr John S. Schneekloth. The emerging role of RNA as a therapeutic target for small molecules. *Cell Chem Biol.*, 23(9):1077–1090, 2016.

[2] Kenneth F Blount, Joy Xin Wang, Jinsoo Lim, Narasimhan Sudarsan, and Ronald R Breaker. Antibacterial lysine analogs that target lysine riboswitches. *Nature chemical biology*, 3(1):44, 2007.

[3] Sergey M Dibrov, Kejia Ding, Nicholas D Brunn, Matthew A Parker, B Mikael Bergdahl, David L Wyles, and Thomas Hermann. Structure of a hepatitis C virus RNA domain in complex with a translation inhibitor reveals a binding mode reminiscent of riboswitches. *Proceedings of the National Academy of Sciences*, 109(14):5223–5228, 2012.

[4] Kiranmai Gumireddy, Douglas D Young, Xin Xiong, John B Hogenesch, Qihong Huang, and Alexander Deiters. Small-molecule inhibitors of microRNA miR-21 function. *Angewandte Chemie International Edition*, 47(39):7482–7484, 2008.

[5] Yuta Naro, Meryl Thomas, Matthew D Stephens, Colleen M Connelly, and Alexander Deiters. Aryl amide small-molecule inhibitors of microRNA miR-21 function. *Bioorganic & medicinal chemistry letters*, 25(21):4793–4796, 2015.

[6] Sai Pradeep Velagapudi, Steven M Gallo, and Matthew D Disney. Sequence-based design of bioactive small molecules that target precursor microRNAs. *Nature chemical biology*, 10(4):291, 2014.

[7] Stefan Arenz and Daniel N Wilson. Blast from the past: reassessing forgotten translation inhibitors, antibiotic selectivity, and resistance mechanisms to aid drug development. *Molecular cell*, 61(1):3–14, 2016.

[8] Angel Bottini, Surya K De, Bainan Wu, Changyan Tang, Gabriele Varani, and Maurizio Pellecchia. Targeting influenza a virus RNA promoter. *Chemical biology & drug design*, 86(4):663–673, 2015.

[9] Amy Davidson, Darren W Begley, Carmen Lau, and Gabriele Varani. A small-molecule probe induces a conformation in HIV TAR RNA capable of binding drug-like fragments. *Journal of molecular biology*, 410(5):984–996, 2011.

[10] Thomas Hermann. Drugs targeting the ribosome. *Current opinion in structural biology*, 15(3):355–366, 2005.

[11] Stephan A Ohnmacht and Stephen Neidle. Small-molecule quadruplex-targeted drug discovery. *Bioorganic & medicinal chemistry letters*, 24(12):2602–2612, 2014.

[12] Suzanne G Rzuczek, Mark R Southern, and Matthew D Disney. Studying a drug-like, RNA-focused small molecule library identifies compounds that inhibit RNA toxicity in myotonic dystrophy. *ACS chemical biology*, 10(12):2706–2715, 2015.

[13] Katherine Deigan Warner, Philip Homan, Kevin M Weeks, Alison G Smith, Chris Abell, and Adrian R Ferre-DAmare. Validating fragment-based drug discovery for biological RNAs: lead fragments bind and remodel the TPP riboswitch specifically. *Chemistry & biology*, 21(5):591–595, 2014.

[14] Yu Zhou and Niu Huang. Binding site druggability assessment in fragment-based drug design. In *Fragment-Based Methods in Drug Discovery*, pages 13–21. Springer, 2015.

[15] Rajarshi Guha. On exploring structure–activity relationships. In *In Silico Models for Drug Discovery*, pages 81–94. Springer, 2013.

[16] Birte Seebeck, Markus Wagener, and Matthias Rarey. From activity cliffs to target-specific scoring models and pharmacophore hypotheses. *ChemMedChem*, 6(9):1630–1639, 2011.

[17] Ashutosh Banerjee, Dirk Schepmann, Jens Köhler, Ernst-Ulrich Würthwein, and Bernhard Wünsch. Synthesis and SAR studies of chiral non-racemic dexoxadrol analogues as uncompetitive NMDA receptor antagonists. *Bioorganic & medicinal chemistry*, 18(22):7855–7867, 2010.

[18] William WL Wong and Forbes J Burkowski. A constructive approach for discovering new drug leads: Using a kernel methodology for the inverse-QSAR problem. *Journal of cheminformatics*, 1(1):4, 2009.

[19] Jayakrishnan Nandakumar, Elaine R Podell, and Thomas R Cech. How telomeric protein POT1 avoids RNA to achieve specificity for single-stranded dna. *Proceedings of the National Academy of Sciences*, 107(2):651–656, 2010.

[20] Yicheng Long, Ben Bolanos, Lihu Gong, Wei Liu, Karen J Goodrich, Xin Yang, Siming Chen, Anne R Gooding, Karen A Maegley, Ketan S Gajiwala, et al. Conserved RNA-binding specificity of polycomb repressive complex 2 is achieved by dispersed amino acid patches in EZH2. *Elife*, 6:e31558, 2017.

[21] Tim R Blower, Xue Y Pei, Francesca L Short, Peter C Fineran, David P Humphreys, Ben F Luisi, and George PC Salmond. A processed noncoding RNA regulates an altruistic bacterial antiviral system. *Nature structural & molecular biology*, 18(2):185, 2011.

[22] Jin-Biao Ma, Yu-Ren Yuan, Gunter Meister, Yi Pei, Thomas Tuschl, and Dinshaw J Patel. Structural basis for 5-end-specific recognition of guide RNA by the A. fulgidus Piwi protein. *Nature*, 434(7033):666, 2005.

[23] Michael Petersen, Kent Bondensgaard, Jesper Wengel, and Jens Peter Jacobsen. Locked nucleic acid (LNA) recognition of RNA: NMR solution structures of LNA: RNA hybrids. *Journal of the American Chemical Society*, 124(21):5974–5982, 2002.

[24] Gabriele Varani, Fareed Aboul-ela, and Frédéric H-T Allain. NMR investigation of RNA structure. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 29(1-2):51–127, 1996.

[25] Nicolas Moitessier, Eric Westhof, and Stephen Hanessian. Docking of aminoglycosides to hydrated and flexible RNA. *Journal of medicinal chemistry*, 49(3):1023–1033, 2006.

[26] Carsten Detering and Gabriele Varani. Validation of automated docking programs for docking and database screening against RNA drug targets. *Journal of medicinal chemistry*, 47(17):4188–4201, 2004.

[27] Thomas Lengauer and Matthias Rarey. Computational methods for biomolecular docking. *Current opinion in structural biology*, 6(3):402–406, 1996.

[28] Katherine E Deigan, Tian W Li, David H Mathews, and Kevin M Weeks. Accurate SHAPE-directed RNA structure determination. *Proceedings of the National Academy of Sciences*, 106(1):97–102, 2009.

[29] David H Mathews, Matthew D Disney, Jessica L Childs, Susan J Schroeder, Michael Zuker, and Douglas H Turner. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proceedings of the National Academy of Sciences*, 101(19):7287–7292, 2004.

[30] David H Mathews. Revolutions in RNA secondary structure prediction. *Journal of molecular biology*, 359(3):526–532, 2006.

[31] David H Mathews and Douglas H Turner. Prediction of RNA secondary structure by free energy minimization. *Current opinion in structural biology*, 16(3):270–278, 2006.

[32] David H Mathews, Walter N Moss, and Douglas H Turner. Folding and finding RNA secondary structure. *Cold Spring Harbor perspectives in biology*, 2(12):a003665, 2010.

[33] Jessica S Reuter and David H Mathews. RNAstructure: software for RNA secondary structure prediction and analysis. *BMC bioinformatics*, 11(1):129, 2010.

[34] Michal J Boniecki, Grzegorz Lach, Wayne K Dawson, Konrad Tomala, Pawel Lukasz, Tomasz Soltysinski, Kristian M Rother, and Janusz M Bujnicki. SimRNA: a coarse-grained method for RNA folding simulations and 3D structure prediction. *Nucleic acids research*, 44(7):e63–e63, 2015.

[35] Bruce A Shapiro, Yaroslava G Yingling, Wojciech Kasprzak, and Eckart Bindewald. Bridging the gap in RNA structure prediction. *Current opinion in structural biology*, 17(2):157–165, 2007.

[36] Mariusz Popenda, Marta Szachniuk, Maciej Antczak, Katarzyna J Purzycka, Piotr Lukasiak, Natalia Bartol, Jacek Blazewicz, and Ryszard W Adamiak. Automated 3D structure composition for large RNAs. *Nucleic acids research*, 40(14):e112–e112, 2012.

[37] Magdalena A Jonikas, Randall J Radmer, Alain Laederach, Rhiju Das, Samuel Pearlman, Daniel Herschlag, and Russ B Altman. Coarse-grained modeling of large RNA molecules with knowledge-based potentials and structural filters. *Rna*, 15(2):189–199, 2009.

[38] Jes Frellsen, Ida Moltke, Martin Thiim, Kanti V Mardia, Jesper Ferkinghoff-Borg, and Thomas Hamelryck. A probabilistic model of RNA conformational space. *PLoS computational biology*, 5(6):e1000406, 2009.

[39] Shantanu Sharma, Feng Ding, and Nikolay V Dokholyan. iFoldRNA: three-dimensional RNA structure prediction and folding. *Bioinformatics*, 24(17):1951–1952, 2008.

[40] Rhiju Das, John Karanicolas, and David Baker. Atomic accuracy in predicting and designing noncanonical RNA structure. *Nature methods*, 7(4):291, 2010.

[41] Marc Parisien and Francois Major. The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature*, 452(7183):51, 2008.

[42] Mi-Kyung Lee, Angel Bottini, Meehyein Kim, Michael F Bardaro, Ziming Zhang, Maurizio Pellecchia, Byong-Seok Choi, and Gabriele Varani. A novel small-molecule binds to the influenza A virus RNA promoter and inhibits viral replication. *Chemical communications*, 50(3):368–370, 2013.

[43] Stanislav Bellaousov, Jessica S Reuter, Matthew G Seetin, and David H Mathews. RNAstructure: web servers for RNA secondary structure prediction and analysis. *Nucleic acids research*, 41(W1):W471–W474, 2013.

[44] Kexin Zhang and Aaron Terrence Frank. Conditional prediction of RNA secondary structure using NMR chemical shifts. *bioRxiv*, page 554931, 2019.

CHAPTER IV

RNAPoser: Tool to Recover Native Poses for RNA-Ligand Complexes

4.1 Statement of Contribution

1. **Aaron T. Frank, PhD:** Conceived of the project described in this chapter; wrote the manuscript on which this chapter is *heavily* based on; and generated all figures presented herein.

2. **Sahil Chhabra, PhD Candidate (Chemistry and Scientific Computing):** Generated the data used to train and test the machine learning classifiers developed in this chapter; assisted in the writing of the manuscript.

3. **Jingru Xie, PhD Candidate (Physics):** As part of her thesis work, designed and implemented the featurization approach used in this chapter; implemented the RNAPosers software that resulted from the work described in this chapter.

4.2 Introduction

Rational structure-based methods promise to be a viable approach for identifying small molecules that can bind to and modulate the activity of therapeutically relevant RNAs. Crucial to the success of rational structure-based approaches in RNA drug discovery is the ability to accurately predict the 3-dimensional (3D) structure of the complex formed between an RNA and a small molecule ligand. In principle, computer docking algorithms can be used to predict the 3D orientation and conformation (referred to as the pose) of a ligand bound to an RNA receptor. Unfortunately, "redocking" tests reveal that state-of-the-art scoring functions typically fail to recover the correct poses.[1–5] In this respect, there is an urgent need for methods that can accurately distinguish "native-like" RNA-ligand poses from non-native decoy poses.

Recently, machine learning has been used to address several challenges associated with computer docking and virtual screening. For protein-ligand complexes in particular, machine learning has been used to develop more robust scoring functions for both pose and binding affinity prediction.[6–10] The success of RNA-ligand pose prediction with the help of chemical shifts filtering in chapter 3 motivated us to explore the RNA-ligand pose prediction without the use of chemical shifts. Here, we used machine learning to train a set of pose classifiers that quantify the "nativeness" of RNA-ligand complexes. In what follows, we summarize our comparison between the ability of docking scores and machine learning classifiers to rank and identify atomically correct RNA-ligand poses. Compared with docking scores, we found that machine learning pose-classifiers were better able to discriminate native-like RNA-ligand poses from decoy poses.

4.3 MATERIALS AND METHODS

4.3.1 Decoy sets

We compiled an initial dataset comprised of 88 RNA-ligand systems. An additional set of 17 RNA-ligand system was compiled and used for final validation. For both datasets, the crystal structures of the RNA-ligand complexes were downloaded from the Protein Data Bank (PDB:http://www.pdb.org). To generate diverse decoy sets for each RNA-ligand system, computer docking was performed using the docking

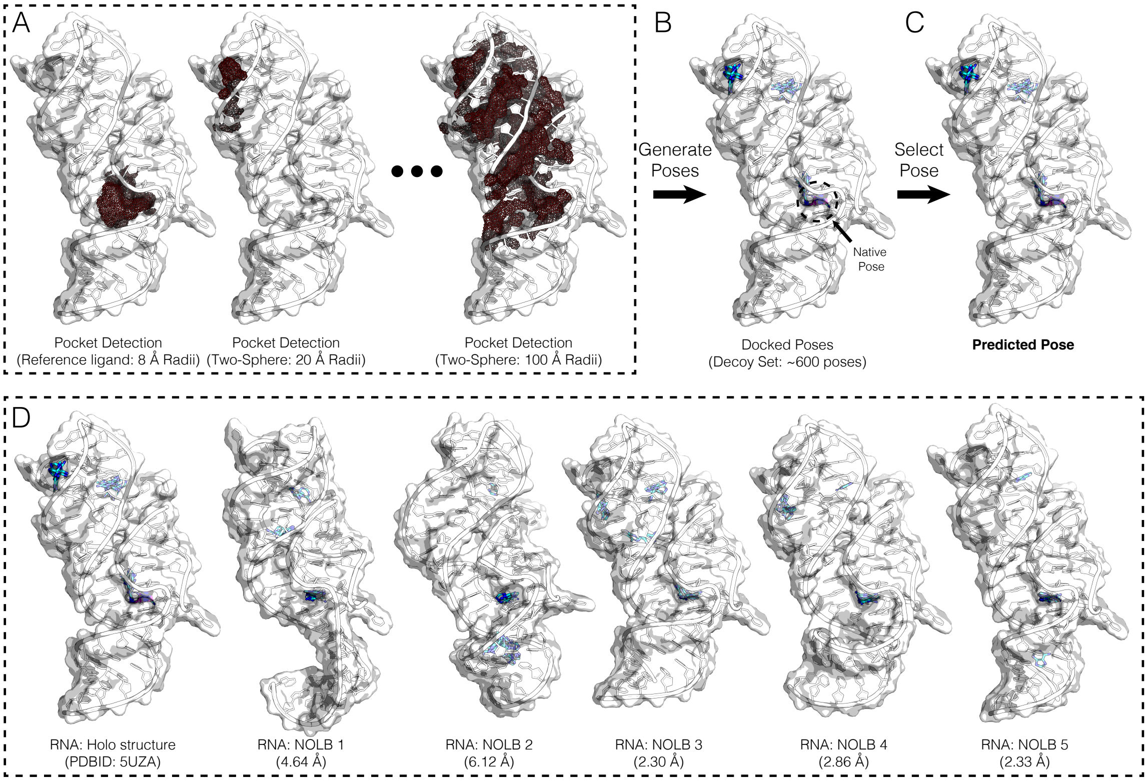


Figure 4.1: Illustrated are the steps involved in generating the decoy sets used in this study. (A) First, the actual binding pocket is mapped using the reference ligand method, and second, alternative pockets are mapped using the two-sphere methods, with increasingly large radii. (B) Third, poses were generated by docking the ligands into each of the mapped binding pockets and then combining all poses into a single decoy set. (C) The focus of this study was to develop and assess methods for selecting atomically-correct poses from these decoy poses. (D) Example of an augmented decoy set in which both the RNA and ligand are flexible. For a given RNA-ligand complex, we generated the augmented decoy sets by deforming the holo RNA structure along its non-linear normal modes. In this study, we generated five such structures for each RNA and then generated docked poses for each, and then all poses combined to form a final augmented decoy set. Shown here is that augmented decoy set for PDBID: 5UZA. Indicated under each of the deformed structures is the RMSD relative to the holo structure.

program rDock (citeruiz2014rdock). The following protocol was used to generate the poses with rDock (Figure 4.1A and B). First, a set of poses were generated in the actual binding pocket, using the reference ligand method, with the sphere radii from the center of the known binding pocket set to 2, 3, 4, 5, 6, 7, and 8 Å, respectively. At each sphere radius, 50 poses were generated, for a total of ∼350 poses. Next, 250 additional poses were generated by docking into the binding pockets that were identified using the two-sphere method, with outer sphere radii set to 20, 40, 60, 80 and 100 Å, respectively. Hence, in total, ∼600 poses were generated for each RNA-ligand complex. For some RNA-ligand complexes, the number of poses were less than 600 because the two-sphere method failed to identify binding pockets at one or more of the outer sphere radii we utilized for binding pocket detection. All pocket detection was carried out using the rDock utility program, rbcavity. The entire set of decoy poses can be accessed at https://github.com/atfrank/RNAPosers.

4.3.2 Pose classifiers

Machine learning was used to train a set of pose classifiers that take a set of "pose features" as input and output a measure of the "nativeness" of the pose. First, we generated a set of classifiers for which the "pose features" correspond to individual scoring terms in the rDock scoring function.[11] Second, we generated a set of classifiers for which the "pose features" correspond a pose novel fingerprint the depends on the pairwise distance between heavy atoms in the an RNA receptor and a the heavy atoms in a small molecule ligand (see below). To train the pose classifiers, we employed the random forest method implemented in the sklearn Python module.[12] The classifiers comprised of an ensemble of 1000 decision trees with class weight set to balanced subsample. All other parameters were set to their default values. The classifiers were trained using a leave-one-out approach using the set of poses generated using rDock (see above). We trained separate classifiers with the nativeness RMSD thresholds set to ≤1.0, 1.5, 2.0, and 2.5Å. Machine learning models can be susceptible to the so-called "twinning effect," which occurs when samples in the training set closely resemble samples in testing set. Here we have employed leave-one-out cross-validation in an attempt to mitigate the potential impact of "twinning" when assessing the performance of classifiers. In this leave-one-out approach, a single

RNA-ligand system was removed from the training set and the classifiers were trained on the remaining 87 RNA-ligand complex. The resulting classifier was then assessed on the excluded RNA-ligand system. *If the ligand in any of the other 87 RNA-ligand systems was identical to the ligand in the left-out system, they were removed prior to training the classifier used to assess the left-out system.*

4.3.3 Pose fingerprint

We utilized a pose fingerprint that is a composite of a set of atomic fingerprints. For a given ligand atom, the atomic fingerprint correspond to the vector, $\{V_i\}$, whose elements $V_i(\eta, v)$ are given by

$$V_i(\eta,\nu) = \sum_{\substack{j\neq i \\ j\in\nu}} e^{-(r_{ij}/\eta)^2} \cdot f_d(r_{ij}) \tag{4.1}$$

where r_{ij} is the distance between the heavy atom i in a ligand and the heavy atom j in the RNA receptor, η is the width of a Gaussian function (here we set $\eta = 2$), ν is a set of unique RNA atom types, and $f_d(r_{ij})$ is the damping function given by

$$f_d(r_{ij}) = 0.5\left[\cos\left(\frac{\pi r_{ij}}{R_c}\right) + 1\right]. \tag{4.2}$$

Here, R_c is a cutoff distance and in this study, it was set to 20 Å. We note that the atomic fingerprint based on Eq. 4.1, which is a multi-element extension of the atomic fingerprint developed by Botu and et.al.,[13] is invariant to the basic atomic transformation operations of translation, rotation and permutation.

For a given ligand pose, i, a fingerprint vector, F_i, was generated from the atomic fingerprint defined by Eq. 4.1 by summing over all instances of a given atom-pair type, which is defined by the SYBYL atom types in the ligand and atom types in the RNA. We denote each unique ligand-RNA pair as S. As such, an element in the fingerprint for pose i and atom pair type S, $F_{i,S}$, is given by

$$F_{i,S} = \sum_{s} V_s(\eta) \tag{4.3}$$

Here s runs over all instances of pair type S in pose i. If pair-type S is not present in i, $F_{i,S} = 0$. The set of 21 SYBYL atom types we used were: {C.1, C.2, C.3, C.ar, C.cat, N.1, N.2, N.3, N.4, N.ar, N.am, N.pl3, O.2, O.3, O.co2, S.2, S.3, S.o, S.o2, P.3}. The set of 85 RNA atom types we used were: ADE:{C1′, C2, C2′, C3′ , C4, C4′, C5, C5′, C6, C8, N1, N3, N6, N7, N9, O2′, O3′, O4′, O5′, OP1, OP2, P}; CYT:{C1′, C2, C2′, C3′, C4, C4′, C5, C5′, C6, N1, N3, N4, O2, O2′, O3′, O4′, O5′, OP1, OP2, P}; GUA:{C1′, C2, C2′, C3′, C4, C4′, C5, C5′, C6, C8, N1, N2, N3, N7, N9, O2′, O3′, O4′, O5′, O6, OP1, OP2, P}; URA:{C1′, C2, C2′, C3′, C4, C4′, C5, C5′, C6, N1, N3, O2, O2′, O3′, O4, O4′, O5′, OP1, OP2, P}. Thus, the final pose fingerprint $F_i = \{F_{i,S}\}$, which was normalized for each RNA-ligand system, contained 1785 elements (21 SBYL types × 85 RNA atom types). Coincidentally, our pose fingerprint closely resembles a recently described fingerprint that was successfully used to train machine learning pose and binding affinity predictors.[10]

4.3.4 Assessing classifiers

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

4.4 Results and Discussion

For protein-ligand complexes, modern scoring functions have a reported success rate that exceeds ∼75 %.[14] In contrast, for RNA-ligand complexes, state-of-the-art scoring functions have a success rate near 50 %.[5,11] This discrepancy between the success rate of protein and RNA scoring functions motivated us to explore methods capable of enhancing our ability to discriminate native-like poses from non-native decoys.

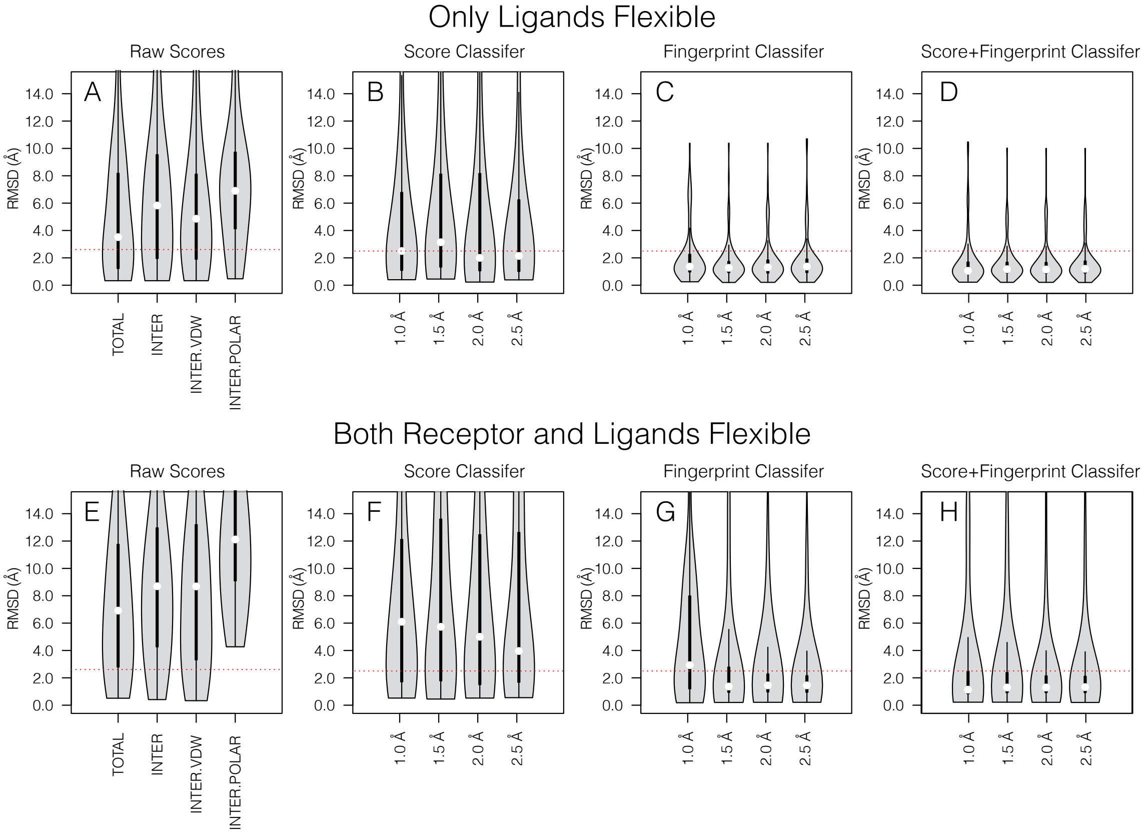


Figure 4.2: RMSD distributions of the best predicted poses over the systems in the leave-one-out training database when the best poses were predicted using (A) docking score terms and classifiers training using the (B) docking score terms, (C) our pose fingerprint, and (D) raw docking scores and our pose fingerprint as features, respectively. For the pose classifiers, results are shown for independent sets of classifiers that were trained with the nativeness threshold set to 1.0, 1.5, 2.0, and 2.5 Å.

4.4.1 Docking scores exhibit low success rates.

We began our study by assessing the ability of docking scores to recover the correct pose from decoy poses located in the experimental binding pocket as well as decoy poses located in alternate pockets on the surface of the RNA. To accomplish this, we initially generated decoys sets comprised of ∼600 diverse poses for 88 RNA-ligand complexes (see Methods). In these decoys sets, the RNA receptors corresponded to the holo structures where only the ligand orientation and conformation varied.

Shown in Figure 4.2A are the distributions of the RMSD (relative to the crystal pose) of the best poses selected from these decoy sets using individual score terms in the rDock scoring function.[11] When using the total docking score, the median RMSD of the predicted pose was 3.41 Å (Figure 4.2A; Table 4.1). We obtained similar results when using the total interaction, the van der Waals interaction, and the polar interaction score terms. In these cases, the median RMSD were 5.72, 4.75, and 6.88 Å, respectively. To better quantify the ability of the score terms to select atomically correct poses, we also computed the success rate, $S(X)$, defined as the percentage of cases in which the predicted pose was within X Å of the native pose. Using the total docking score, the $S(1.00)$, $S(1.50)$, $S(2.00)$, and $S(2.5)$ were 22.7, 29.5, 37.5, 42.0, and 44.3 %, respectively (Table 4.1). Similarly, the $S(1.00)$, $S(1.50)$, $S(2.00)$, and $S(2.5)$ were 17.0, 21.6, 27.3, and 33.0%, respectively, when using total interaction, 18.2, 22.7, 28.4 and 36.4%, respectively, when using the van der Waals interaction, and 8.0, 9.1, 12.5, and 21.6%, respectively, when using the polar interaction score terms (Table 4.1). The docking score terms in the rDock scoring function, therefore, exhibited marginal ability to recover correct poses from diverse decoys poses.

4.4.2 Pose classifiers improve success rates.

Next, we asked whether nonlinear classifiers could enhance our ability to recover the correct poses from decoy poses. To test this, we cast the problem of recovering correct ligand poses as a "classification" problem and then machine learning models were trained to discriminate correct poses from decoy poses. Briefly, we built a set of random forest classification models that take a set of features as input and output "classification scores" that estimate the probability of a pose being native-like. To accomplish this, we first trained a series of random forest pose classifiers using a leave-one-out cross-validation approach in which we selected a single RNA-ligand from the dataset of 88 RNA-ligand systems (the leave-one-out dataset), and trained a classifier using the decoy sets for the remaining 87 RNA-ligand systems. After training, the performance of the resulting classifier was assessed using the RNA-ligand system set of the left-out system. For that system, the classification scores for all decoy poses were determined and then the pose with the highest classification score was selected as the "best" (or predicted) pose for the left-out system. This procedure was repeated 88 times, i.e., one for each system in the leave-one-out dataset.

Shown in Figure 4.3 are the relationship between RMSD and the Classification Scores (CLS) of the predicted poses for classifiers trained on the molecular fingerprints using RMSD threshold of 2.5 Å. Reported are results for the 88 Leave One Out (LOO) and 17 validation set RNAs. In general, for both LOO and validation set the RNAs which have RMSD 2Åhave a CLS score of greater than 0.5.

Shown in Figure 4.2B are the distributions of the RMSD of the predicted poses that were identified using the classifiers that used the individual terms in the rDock scoring function as learning features. Reported are results for the classifiers trained with nativeness RMSD threshold set to 1.0, 1.5, 2.0, and 2.5 Å, respectively. Listed

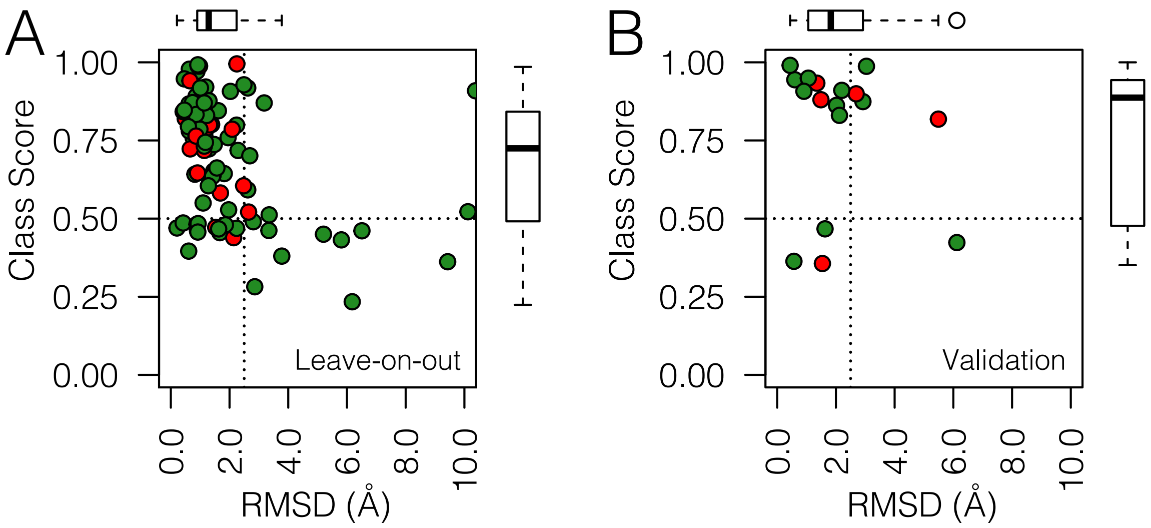


Figure 4.3: Illustrated are the CLS vs RMSD for LOO and validation sets

in Table 4.1 are the corresponding success rates. In general, the RMSD of the best poses that were identified using the score-based pose classifiers were lower than those of the best poses selected using the terms in the rDock scoring function. For instance, for the score-based pose classifiers trained with nativeness threshold set to 1.0, 1.5, 2.0, and 2.5 Å, the the median RMSD of the best poses were 2.50, 3.14, 2.08, and 2.14, respectively (Figure 4.2B; Table 4.1). The success rates, $S(1.00)$, $S(1.50)$, $S(2.00)$, and $S(2.5)$ were also generally larger for the score-based classifiers, with the best results obtained with the nativeness threshold set to 2.0 and 2.5 Å, respectively. $S(1.00)$, $S(1.50)$, $S(2.00)$, and $S(2.5)$ were 21.6, 36.4, 50.0, and 54.5%, respectively, for the classifiers trained with the threshold set to 2.0 Åand 25.0, 37.5, 48.9, and 54.5%, respectively, for the classifiers trained with the threshold set to 2.5 Å. In comparison, the values obtained when using the total docking score to identify the best pose were 37.5, 42.0, and 44.3 %, respectively (Table 4.1). These results suggest that the pose classifiers that were trained using the scores terms as learning features could boost our ability to recover correct poses. The success rates, however, still pales in comparison to the success rates of protein-ligand pose prediction methods, several of which achieve success rates near 75%.

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

Table 4.1: Median RMSD and success rates for systems in the leave-one-out training database. Listed are the results obtained when the best poses were selected using the docking score terms and classifiers that were trained using the docking score terms, our pose fingerprint, and docking scores plus our pose fingerprint as learning features. For the pose classifiers, we include results for classifiers that we trained with the nativeness threshold set to 1.0, 1.5, 2.0, and 2.5 Å.

As such, we next asked whether we could further enhance the success rate of RNA-ligand pose prediction by training pose classifiers on features that more directly depend on RNA-ligand interactions. Specifically, we were interested in examining the utility of a simple distance-based atomic fingerprint that describes the local atomic environment near a given site which has shown promise in predicting properties like atomic forces[15] and resembles a pose fingerprint recently used for protein-ligand pose predictions.[10] To create a composite fingerprint from atomic fingerprints, we summed and normalized all atomic fingerprints associated with specific ligand-RNA pair "types" (see Methods). Using this composite RNA-ligand interaction fingerprint, we then trained another set of pose classifiers, again using the leave-out-one cross-validation approach. For comparison, we also trained classifiers that used the rDock score terms plus our pose fingerprint as features. Here again, separate classifiers were trained with the nativeness threshold set to 1.0, 1.5, 2.0, and 2.5 Å.

For the pose fingerprint classifiers trained with nativeness threshold set to 1.0, 1.5, 2.0, and 2.5 Å, the median RMSD of the best poses were 1.36, 1.27, 1.31, and 1.42 Å, respectively (Figure 4.2C; Table 4.1). These fingerprint-based classifiers all exhibit similar success rates. For instance, $S(1.00)$, $S(1.50)$, $S(2.00)$, and $S(2.5)$ were 37.5, 63.6, 77.3, and 86.4 %, respectively, for the classifiers trained with the nativeness threshold set to 1.5 Å, and which had the lowest median RMSD of 1.27 Å(Table 4.1). In comparison, $S(1.00)$, $S(1.50)$, $S(2.00)$, and $S(2.5)$ were 33.0, 58.0, 77.3, and 86.4 %, respectively, for the classifiers trained with the nativeness threshold set to 2.5 Å, and which had the highest median RMSD of 1.42 Å(Table 4.1). We obtained comparable results for the pose classifiers that were trained using the docking scores plus the fingerprint as features. Notable among these was the classifier trained with the nativeness threshold set to 1.0 Å; for this set of classifiers, the median RMSD of the best poses was 1.05 Åand the $S(1.00)$, $S(1.50)$, $S(2.00)$, and $S(2.5)$ were 43.2, 70.5, 79.5, and 85.2, respectively. Based on this leave-one-out analysis, the pose classifiers trained using the pose fingerprint as well as the classifiers trained using docking score terms and pose fingerprint as features, both exhibited remarkable ability to recover atomically correct poses from the leave-one-out decoy sets.

4.4.3 Pose classifiers improve success rates on augment decoys sets in which both the RNA and the ligand are flexible.

As a more robust test of our classifiers, the original decoy set for each RNA-ligand system in our dataset was augmented with poses that were generated by docking against a set of five perturbed structures of the corresponding RNA receptor. As might be expected, the individual docking score terms failed to recover poses

Selection Metric	$\text{RMSD}_{\text{median}}$(Å)	S(1.00)(%)	S(1.50)(%)	S(2.00)(%)	S(2.50)(%)
TOTAL	6.81	6.2	11.2	18.8	23.8
INTER	8.59	2.5	6.2	11.2	16.2
INTER.VDW	8.58	6.2	10.0	15.0	21.2
INTER.POLAR	12.00	0.0	0.0	0.0	0.0
Score Classifier (1.0 Å)	6.10	15.0	21.2	30.0	33.8
Score Classifier (1.5 Å)	6.02	15.0	21.2	31.2	37.5
Score Classifier (2.0 Å)	5.00	15.0	25.0	37.5	41.2
Score Classifier (2.5 Å)	3.96	13.8	22.5	32.5	37.5
Fingerpint Classifier (1.0 Å)	2.98	20.0	30.0	38.8	46.2
Fingerpint Classifier (1.5 Å)	1.35	33.8	53.8	63.8	72.5
Fingerpint Classifier (2.0 Å)	1.42	31.2	53.8	72.5	80.0
Fingerpint Classifier (2.5 Å)	1.44	31.2	53.8	72.5	82.5
Score+Fingerpint Classifier (1.0 Å)	1.10	41.2	62.5	70.0	76.2
Score+Fingerpint Classifier (1.5 Å)	1.31	36.2	57.5	68.8	76.2
Score+Fingerpint Classifier (2.0 Å)	1.30	35.0	57.5	70.0	80.0
Score+Fingerpint Classifier (2.5 Å)	1.32	31.2	56.2	72.5	77.5

Table 4.2: Median RMSD and success rates for the augmented decoy sets (Figure 4.1D) of the systems in the leave-one-out training database. Listed are the results obtained when the best poses were selected using the docking score terms and classifiers that were trained using the docking score terms, our pose fingerprint, and docking scores plus our pose fingerprint as learning features. For the pose classifiers, we include results for classifiers that we trained with the nativeness threshold set to 1.0, 1.5, 2.0, and 2.5 Å.

in these augmented decoy sets, with the best results obtained when ranking and selecting poses based on their total docking score. In this case, the median of RMSD of the best pose was 6.81 Å (Figure 4.2E; Table 4.2). Marginally better results were obtained using the score classifiers (Figure 4.2F; Table 4.2), with the classifiers trained with a nativeness threshold of 2.5 Å exhibiting the lowest median value of 3.96 Å, and $S(1.00)$, $S(1.50)$, $S(2.00)$, and $S(2.5)$ values of 13.8, 22.5, 32.5, and 37.5 %, respectively (Table 4.2). By contrast, the pose classifiers trained using the pose fingerprint as features and the composite score and pose fingerprint features were typically able to recover correct poses (Figure 4.2G and H). Except for the fingerprint classifier trained with the nativeness threshold set to 1.00 Å, the median RMSD for the predicted poses, were <1.50 Å. A representative of these pose classifiers was the fingerprint classifier trained with nativeness threshold set to 2.50 Å. For this classifier, the median RMSD of the predicted poses was 1.44 Å and the success rates,

$S(1.00)$, $S(1.50)$, $S(2.00)$, and $S(2.5)$, were 31.2, 53.8, 72.5, and 82.5 %, respectively (Table 4.2).

Selection Metric	$\text{RMSD}_{\text{median}}$(Å)	$S(1.00)$(%)	$S(1.50)$(%)	$S(2.00)$(%)	$S(2.50)$(%)
TOTAL	4.96	5.9	5.9	17.6	17.6
INTER	7.16	5.9	11.8	11.8	11.8
INTER.VDW	3.29	5.9	17.6	17.6	23.5
INTER.POLAR	11.67	0.0	0.0	0.0	0.0
Score Classifier (1.0 Å)	4.19	11.8	17.6	17.6	17.6
Score Classifier (1.5 Å)	6.62	5.9	17.6	17.6	17.6
Score Classifier (2.0 Å)	5.89	5.9	17.6	17.6	17.6
Score Classifier (2.5 Å)	8.39	5.9	17.6	23.5	23.5
Fingerpint Classifier (1.0 Å)	1.89	35.3	47.1	52.9	58.8
Fingerpint Classifier (1.5 Å)	2.05	41.2	41.2	47.1	64.7
Fingerpint Classifier (2.0 Å)	2.69	29.4	41.2	41.2	47.1
Fingerpint Classifier (2.5 Å)	2.12	23.5	35.3	41.2	58.8
Score+Fingerpint Classifier (1.0 Å)	2.69	35.3	41.2	41.2	47.1
Score+Fingerpint Classifier (1.5 Å)	2.71	29.4	41.2	41.2	47.1
Score+Fingerpint Classifier (2.0 Å)	2.69	29.4	41.2	41.2	47.1
Score+Fingerpint Classifier (2.5 Å)	2.05	29.4	41.2	47.1	58.8

Table 4.3: Median RMSD and success rates for the augmented decoy sets of the systems in the independent validation set. Listed are the results obtained when the best poses were selected using the docking score terms and classifiers that were trained using the docking score terms, our pose fingerprint, and docking scores plus our pose fingerprint as learning features. For the pose classifiers, we include results for classifiers that we trained with the nativeness threshold set to 1.0, 1.5, 2.0, and 2.5 Å. For all pose classifiers, results are shown for independent sets of classifiers that were trained with nativeness threshold set to 1.0, 1.5, 2.0, and 2.5 Å.

Next, we repeated the analysis described above for an independent dataset comprised of 17 RNA-ligand systems that we did not include in the leave-one-out dataset used to train the pose classifiers. These tests were carried out using the more challenging augmented decoys sets. When using the docking scores to rank and select poses, the van der Waals interaction scores exhibited the lowest median RMSD poses (3.29 Å) and when using the score-based pose classifiers, the classifier trained with nativeness threshold of 1.0 Å exhibited the lowest median RMSD (4.19 Å). In contrast, the corresponding values for the fingerprint-based classifiers and scores plus fingerprint-based classifiers were 1.89 Å (nativeness threshold of 1.00 Å) and 2.05 Å (nativeness threshold of 2.50 Å), respectively (Table 4.3). In terms of the success rates, the best performing classifiers were the fingerprint-based pose classifiers that were trained with nativeness threshold of 1.00 and 1.50 Å, respectively. The success rates, $S(1.00)$, $S(1.50)$, $S(2.00)$, and $S(2.5)$ were 35.3, 47.1, 52.9, and 58.8 for the fingerprint-based pose classifier that were trained with nativeness threshold of 1.00 Å and 41.2, 41.2, 47.1, and 64.7 for the corresponding classifier that were trained with nativeness threshold of 1.50 ÅÅs such, though the success rates on the validation set were lower than rates estimated from our leave-one-out analysis, they are significantly higher than the results obtained using either the raw docking scores or the baseline score-based classifiers (Table 4.2).

When we challenged the classifiers with the augmented decoys sets, in which both the RNA conformation and ligand poses were varied, one possible reason why their overall performance degraded is that we trained the classifiers themselves on decoys sets in which only the ligand poses were varied (i.e., for each RNA-ligand complex, the RNA was fixed in the *holo* conformation). It seems reasonable that the performance of the classifiers could enhanced the by training them using the augmented decoy sets, in which both the RNA conformation and ligand poses were varied. Unfortunately, such augmented decoy sets are *severely* imbalanced, which hampers the training of robust random forest classifiers. In future work, we explore methods to train robust classifiers on these highly imbalanced decoy sets as well as developing similar classifiers to RNA-protein complexes.

Here, we focused on training classifiers that can be used to post-process docked RNA-ligand poses. The enhanced ability to recover atomically accurate poses indicate that the relationships between the pose fingerprint and the nativeness of individual poses captured by our classifiers might also be useful in guiding conformational sampling during docking. In principle, we could convert these classification scores into a pseudo-energy term of the form $-kT\ln(1+p)$, and can add it as an additional term to existing scoring functions. Alternatively, a new pair-wise scoring function could be developed using the random forest refinement strategy recently described by Merz and coworkers.[9] In either case, we could then assess the classifier-informed scoring functions by quantifying the extent to which the distribution of docked poses shift towards or away from native-like poses. Future work will explore this further.

4.4.4 Conclusion

In this study, we showed that machine learning classifiers significantly enhance RNA-ligand pose prediction accuracy, especially when applied to set of ligand poses that were docked against the *holo* conformations of RNAs. Due to the promising results we obtained using our pose classifiers, we have incorporated them into the software tool, RNAPosers (https://github.com/atfrank/RNAPosers). To facilitate the development and testing of other RNA-ligand pose prediction methods and scoring functions, we also make accessible all of the decoy sets used in this study. In the context of RNA-ligand pose prediction, RNAPosers should find utility as a tool to assess the relative quality of a set of poses derived either from purely computational methods or from hybrid modeling methods that incorporate experimental data such as chemical shift perturbation. Also, within the context of virtual screening, we envision that RNAPosers may find utility as a tool to identify high-confidence poses that can be brought forward for binding affinity prediction using physics-based free energy calculation methods like, MM-PBSA and FEP calculations.

BIBLIOGRAPHY

[1] Douglas B Kitchen, Hélène Decornez, John R Furr, and Jürgen Bajorath. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nature reviews Drug discovery*, 3(11):935, 2004.

[2] Patrick Pfeffer and Holger Gohlke. DrugScoreRNA knowledge-based scoring function to predict RNA- ligand interactions. *Journal of chemical information and modeling*, 47(5):1868–1876, 2007.

[3] Lu Chen, George A Calin, and Shuxing Zhang. Novel insights of structure-based modeling for RNA-targeted drug discovery. *Journal of chemical information and modeling*, 52(10):2741–2753, 2012.

[4] Anna Philips, Kaja Milanowska, Grzegorz Łach, and Janusz M Bujnicki. LigandRNA: computational predictor of RNA–ligand interactions. *RNA*, 2013.

[5] Zhiqiang Yan and Jin Wang. SPA-LN: a scoring function of ligand–nucleic acid interactions via optimizing both specificity and affinity. *Nucleic acids research*, 45(12):e110–e110, 2017.

[6] Jacob D Durrant and J Andrew McCammon. NNScore 2.0: a neural-network receptor–ligand scoring function. *Journal of chemical information and modeling*, 51(11):2897–2903, 2011.

[7] Matthew Ragoza, Joshua Hochuli, Elisa Idrobo, Jocelyn Sunseri, and David Ryan Koes. Protein–ligand scoring with convolutional neural networks. *Journal of chemical information and modeling*, 57(4):942–957, 2017.

[8] José Jiménez, Miha Skalic, Gerard Martinez-Rosell, and Gianni De Fabritiis. K DEEP: Protein–ligand absolute binding affinity prediction via 3D-convolutional neural networks. *Journal of chemical information and modeling*, 58(2):287–296, 2018.

[9] Jun Pei, Zheng Zheng, Hyunji Kim, Lin Frank Song, Sarah Walworth, Margaux R Merz, and Kenneth M Merz. Random forest refinement of pairwise potentials for protein-ligand decoy detection. *Journal of Chemical Information and Modeling*, 2019.

[10] Duc Duy Nguyen, Zixuan Cang, Kedi Wu, Menglun Wang, Yin Cao, and Guo-Wei Wei. Mathematical deep learning for pose and binding affinity prediction and ranking in D3R grand challenges. *Journal of computer-aided molecular design*, 33(1):71–82, 2019.

[11] Sergio Ruiz-Carmona, Daniel Alvarez-Garcia, Nicolas Foloppe, A Beatriz Garmendia-Doval, Szilveszter Juhos, Peter Schmidtke, Xavier Barril, Roderick E Hubbard, and S David Morley. rDock: a fast, versatile and open source program for docking ligands to proteins and nucleic acids. *PLoS computational biology*, 10(4):e1003571, 2014.

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

[13] Venkatesh Botu, Rohit Batra, James Chapman, and Rampi Ramprasad. Machine learning force fields: construction, validation, and outlook. *The Journal of Physical Chemistry C*, 121(1):511–522, 2016.

[14] Jonas Dittrich, Denis Schmidt, Christopher Pfleger, and Holger Gohlke. Converging a knowledge-based scoring function: Drugscore2018. *Journal of chemical information and modeling*, 59(1):509–521, 2018.

[15] Venkatesh Botu and Rampi Ramprasad. Adaptive machine learning framework to accelerate ab initio molecular dynamics. *International Journal of Quantum Chemistry*, 115(16):1074–1083, 2015.

CHAPTER V

Overall Summary

5.1 Major Findings

In this work we developed methods to predict the structure of RNAs and RNA-ligand complexes. Understanding of the structure is crucial to gain insights into the structure-function relationship of biomolecules. CS obtained from NMR studies encodes local chemical interactions between the atoms. We devised precise prediction methods to predict the CS by featurizing the local chemical environment. We studied the impact of different Ring Current Models, namely, Pople, Johnson-Boevy and Haigh-Mallion, on the accuracy of predicting CS. We found that the complexity of calculating ring current varies between these three models, with Pople being the simplest and Haigh-Mallion being the most complex. We also explored the accuracy of different ML algorithms on predicting the chemical shifts. Next we used predicted CS to filter and predict the RNA secondary and tertiary structures starting from sequence. CS were also able to recover native-like RNA-ligand poses from a decoy set. The ligand in that study was docked onto the top 10 RNA tertiary structures that were predicted using a CS-guided prediction approach, starting from the RNA sequence. This RNA-ligand pose prediction approach motivated us to design a molecular fingerprinting based classifier to rank different RNA-ligand poses. The molecular fingerprinting approach which combined the atomic fingerprints that were based on SYBYL atom types was able to provide unique featurization for a single RNA-ligand pose. We then fed these molecular fingerprints to the random forest ML classifier which were successfully able to resolve the RNA-ligand poses.

5.2 Innovation

A side by side comparison of all the ring current models ability to predict chemical shifts was lacking in the literature. In chapter 2, we compared all the ring current models and found that they have relatively similar performance at predicting the chemical shifts despite the difference in the time complexity of their implementations. Based on these observations, we decided to use the Pople ring current model because of it's lowest time complexity compared to the other two models namely Johnson-Bovey and Haigh-Mallion. Also in terms of machine learning algorithms, ensemble methods outperformed the linear methods in terms of accuracy in predicting the CS. Random forest and lasso lars cross validation algorithms were found to be the most accurate machine learning algorithms in ensemble and linear categories respectively. In chapter 3, we came up with novel de-novo approach to predict the RNA structure for PDB id 2LWK with better accuracy compared to existing methods using chemical shifts filtering. We were able to recover 50% of the RNA structures within $2.5\AA$ using CS guided secondary structure prediction combined with Rosetta de-novo 3D-structure prediction protocol. Chemical shifts error based filtering was also able to recover native like RNA-ligand complexes starting from just the sequence of RNAs. In chapter 4, RNAPoser which was build on a novel fingerprinting approach was able to significantly enhance RNA-ligand pose prediction accuracy to 83.3% when compared with the accuracy of 46.7% for LigandRNA and 40.0% for DrugscoreRNA on the 21 common RNA-ligand systems in all the three studies. We also compared RNAPoser's accuracy with the SPA-LN scoring function. On the testing dataset of 56 common RNA-ligand decoys with 1000 structures for each complex, the SPA-LN and rDock both had 54% accuracy. On our decoy sets of the same 56 RNA-ligand complexes with around 600 structures for each complex, rDock had an accuracy of 34% whereas RNAPoser had an accuracy of 58.9%. These results show that RNAPoser is able to better predict the poses for RNA-ligand complexes.

5.3 Limitations

Most of the structure prediction protocols presented in this thesis have some limitations, which are presented below. As in any machine learning study, the confidence in accuracy of the model depends on the confidence in the data used to build those models. For example, in predicting NMR chemical shifts, the accuracy of the predictors which were build depends on the accuracy of the reported chemical shifts used to train those predictors. The less the error is in those reported chemical shifts, the better our predictors are at predicting the chemical shifts. Similarly, the strength of the molecular fingerprints and the predictors used in RNAPoser depends on the accuracy of featurizer, which in turn depends on the accuracy of the structures used to calculate those features. The more accurate those structural coordinates are, the better the accuracy of the features and hence the predictors as well. Another limitation in using structure prediction/generation tools like RNAStructure, Rosetta and rDock are the inherent errors in the accuracy of the scoring functions which are mostly energy based or empirical based . Since most of the studies in this thesis uses the above mentioned software tools, the accuracy of our methods in turn depends on accuracy of those softwares.

5.4 Future Work

Despite the limitations mentioned above, we believe that the accuracy of RNA structure prediction can be further improved by integrating various methods presented in this thesis. For example, one could combine the predicted chemical shifts and RNAPoser classification scores, to build a hybrid scoring function to recover native like RNA-ligand poses. One could also explore combining the chemical shift filtering approach with the Rosetta scores to predict the de-novo RNA-ligand structure and test it on different RNA-ligand complexes. One could also explore the transferability of RNAPoser to other systems. For example the classifiers we trained on RNA-small molecule data can be extended to be used for recovering RNA-protein structures. We expect that RNAPoser can also be used to score the RNA residues to access the quality of RNA structures.