Investigating Riboswitch-Containing Complexes via Molecular Dynamics Simulations

by

Yichen Liu

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
2024

Doctoral Committee:

Professor Nils G. Walter, Chair
Professor Charles L. Brooks III
Dr. Aaron T. Frank, Arrakis Therapeutics
Professor Markos Koutmos
Professor Wenjing Wang

Yichen Liu
twinsliu@umich.edu
ORCID iD: 0000-0001-9176-6121

© Yichen Liu 2024

DEDICATION

To my family and friends.

ACKNOWLEDGEMENTS

First and foremost I would like to thank my advisors, Dr. Aaron Frank and Dr. Nils Walter for all the help and support I have received during the past few years. I would like to thank Aaron for guiding me into the field of computational biophysics and for always being supportive. I would like to thank Nils for accepting me into the lab and for supporting me through the last two years of graduate school. I would like to thank both of you for being amazing mentors and providing insightful ideas on my research projects. I am truly honored and grateful to have both of you as my mentors.

Next I would like to thank all my committee members. I would like to thank Dr. Charles L. Brooks for providing me the opportunity to rotate in your lab and for all the valuable feedback on my research. I would like to thank Dr. Markos Koutmos for all the help, support and collaboration. I would like to thank Dr. Wenjing Wang for the support and the great feedback on my research. I would also like to thank Dr. Melanie Ohi for the help on our collaborative project.

To everyone in the Frank lab and Walter lab, I thank you for your support and help. I would like to thank Dr. Indrajit Deb for all the guidance and help I received during my time in graduate school. I would like to thank Dr. Kexin Zhang, Dr. Jingru Xie, Dr. Ciara Witt, Ziqiao Xu and Grace Arhin for the help and advice you gave me and for all the enjoyable time we had together. I would like to thank Dr. Adrien Chauvier and Emily Ellinger for your help in our collaborative project. I would also like to thank Rosa Romero, Dr. Elizabeth Franklin, Dr. Sujay Ray, Guoming Gao, Dr. Liuhan Dai, Jingxuan Tang, Emily Sumrall, Dr. Tanmay Chatterjee, Dr. Alexander Johnson-Buck, Dr. Pavel Banerjee, Dr. Laxmikanta

Khamari and Dr. Namra Siddiqui for your help inside and outside of lab.

Finally, I would like to express my gratitude to my family and friends. To my family, thank you for your unconditional love and support along the way. To my friends Yudan, Runxue, and Meichen, thank you for making my life enjoyable and for always supporting me.

TABLE OF CONTENTS

DEDICATION ii

ACKNOWLEDGEMENTS iii

LIST OF FIGURES viii

LIST OF TABLES xiii

LIST OF APPENDICES xiv

ABSTRACT xv

CHAPTER

1 Introduction 1

1.1 Central dogma and non-coding RNAs 1
1.1.1 Overview 1
1.1.2 Examples of non-coding RNAs 2
1.1.3 Riboswitches 3
1.2 Experimental methods for RNA structure determination 4
1.2.1 Overview of RNA structure 4
1.2.2 X-ray crystallography 6
1.2.3 Nuclear magnetic resonance spectroscopy (NMR) 6
1.2.4 Cryo-EM 7
1.2.5 Limitations 7
1.3 Molecular dynamics simulations 8
1.3.1 Overview of molecular dynamics simulations 8
1.3.2 Enhanced sampling methods 9
1.4 Application of Machine Learning in computational chemistry 11
1.5 Thesis outline 11

2 Using Selectively Scaled Molecular Dynamics Simulations to Assess Ligand Poses in RNA Aptamers 13

2.1 Introduction 14
2.2 Methods 16
2.2.1 Pose generation and selection 16

2.2.2 System preparation . . . 19
2.2.3 Molecular dynamics simulations . . . 19
2.2.4 Analysis of molecular dynamics simulations . . . 22
2.3 Results and Discussion . . . 24
2.3.1 Simulation-Derived Unbinding Profiles Resolve Native-Like Poses from Decoys . . . 24
2.3.2 Structure-based models can robustly predict simulation-derived unbinding profiles . . . 34
2.3.3 Structure-derived unbinding profiles help resolve native-like poses from non-native decoys . . . 36
2.4 Conclusion . . . 40

3 Pre-Equilibrium Kinetics May Underlie the Differential Activity of Selective Binders of the FMN-dependent Riboswitch . . . 42

3.1 Introduction . . . 43
3.2 Methods . . . 44
3.2.1 Selectively scaled MD simulations . . . 44
3.2.2 Contact Analysis . . . 46
3.2.3 Kinetic Models . . . 46
3.2.4 Selection of representative trajectories . . . 47
3.3 Results and Discussion . . . 48
3.3.1 WG-3 rapidly unbinds from the FNM aptamer compared to ribocil-B 48
3.3.2 WG-3 and ribocil-B exhibit distinct ssMD escape paths . . . 51
3.3.3 Unbinding of ribocil-B induces larger scale conformational changes compared to that of WG-3 . . . 54
3.4 Conclusions . . . 55

4 Structural Basis of Partitioning Between Termination and Antitermination of the Fluoride Riboswitch . . . 57

4.1 Introduction . . . 58
4.2 Methods . . . 60
4.2.1 RNAP expression and purification . . . 60
4.2.2 Preparation of elongation complex for cryo-EM . . . 61
4.2.3 Time pausing assays . . . 62
4.2.4 Time-pausing analysis . . . 63
4.2.5 Molecular dynamics flexible fitting simulations . . . 63
4.3 Results . . . 67
4.3.1 The in vitro assembled fluoride-PEC is functional . . . 67
4.3.2 Data collection . . . 69
4.3.3 Ligand binding induces riboswitch rotation and opens the RNAP exit channel . . . 71
4.3.4 Global conformational changes within the PEC . . . 77
4.3.5 Ligand binding alters RNA-protein interactions . . . 80
4.3.6 Time pausing assays validate switching behavior of the fluoride riboswitch . . . 83

4.4 Discussion 83
5 Summary and Future Outlook 87
5.1 Summary 87
5.2 Future Directions 89

APPENDICES 91

BIBLIOGRAPHY 108

LIST OF FIGURES

FIGURE

1.1 Total number of RNA structures deposited on the Protein Data Bank (PDB) through 2023 for structures of protein-nucleic acid complexes (top) and structures that contains only RNA (bottom). 5

1.2 The workflow of molecular dynamics simulations. 9

2.1 Structures of the five RNA-ligand complexes we studied. (a) PDB ID: 2B57. Purine riboswitch bound to 2,6-Diaminopurine. (b) PDB ID: 2XWN. Purine riboswitch bound to triazolo-triazole-diamine. (c) PDB ID: 3NPN. SAH riboswitch bound to S-adenosylhomocysteine. (d) PDB ID: 4XWF. ZTP riboswitch bound to ZMP. (e) PDB ID: 5KPY. 5-hydroxytryptophan aptamer bound to 5-hydroxy-L-tryptophan. 18

2.2 To observe the unbinding event in relatively short simulations, the interactions between the ligand atoms and water atoms are scaled up by the scaling factor $\alpha'_{ij} = 3.5$. Shown are the scaled potential (yellow) and the unscaled potential (orange). 21

2.3 Illustration of the simulation-based framework employed in this study. For individual poses, we generate a series of unbinding trajectories using the selectively scaled MD (ssMD) technique. We then apply contact analysis to the resulting trajectories and, from them, compute an unbinding profile, defined as the number of native contacts that persist as a function of time ($\langle M(t) \rangle$). We then fit the unbinding profile to a single-exponential function to extract a set of parameters (A, τ and B) that compactly describe the unbinding profile. Here, we assess whether we could use any of these parameters to correctly identify near-native poses. In an attempt to bypass costly validation simulations, we also explore whether we could learn from simulations and predict, directly from the structure of the initial pose, the parameters of the unbinding profile (green arrow). 23

2.4 (a) Overlay of poses for each of the five RNA-ligand complexes we examined in this study. (b) Corresponding ROC plots for discriminating native-like poses from non-native (decoy) poses based on the characteristic decay times (τ_{ssMD}) estimated from ssMD simulations of individual poses. (c) Comparison between the native (green) and best poses (pink) among the top 1, 3, and 5 poses, when we sort poses by τ_{ssMD}. 31

2.5 ROC plots when using the ssMD-derived exponential fit parameters (a) A, (b) τ, and (c) B of five RNA-ligand complexes, here identified using their PDB IDs. 32

2.6 ROC plots when ranking poses using the rDock interaction scores of five RNA-ligand complexes, here identified using their PBD IDs. 33
2.7 Results we obtained when using partial least squared (PLS) to predict parameters extracted from ssMD unbinding profiles (A, τ, and B) directly from the structure of individual poses. (a) Cross-validation RMSEP as function of the number of components in the PLS model used to predict $\log(A)$ (left), $\log(\tau)$ (middle), and $\log(C)$ (right). Gray dotted lines correspond to the absolute minimum RMSEP observed, and blue dotted lines indicate the optimal selected number of components. We chose the latter using the one-sigma rule as the smallest number of components that exhibited an RMSEP within 1σ of the minimum RMSEP. The error bars capture the variance in the estimated RMSEP over the leave-one-out cross-validation experiments. (b) Scatter plots comparing actual (ssMD-derived) $\log(A)$ (left), $\log(\tau)$ (middle), and $\log(C)$ (right) with those predicted using optimal number of components. Results correspond to those from leave-one-out cross-validation experiments. And we have annotated the scatter plots with the RMSEP and correlation coefficient (R). 35
2.8 ROC plots for discriminating native-like poses from non-native (decoy) poses based on the characteristic decay times predicted directly from the structure of individual poses using our structure-based PLS model (τ_{PLS}). Below each ROC plot, we show comparisons between the native (green) and best poses (pink) among the top 1, 3, and 5 poses, when poses are sorted by τ_{PLS}. 37
2.9 ROC plots for the 4NYA test case when using a 2.70 Å (a) and 2.85 Å (b) RMSD cutoff to distinguish native from non-native (decoy) poses. Shown in (c) is a closeup view of the native pose and a pose with 2.83 Å RMSD. With 2.70 Å RMSD cutoff, such poses would be considered non-native though their location overlaps with the native pose, albeit in a flipped conformation. The featurization approach we used is likely insensitive to such subtle differences thus yielding similar estimates for τ as poses with RMSD less than ∼2.80 Å. This, in part, explains the low AUC observed when using a 2.70 Å RMSD cutoff as such poses would be considered non-native. 39

3.1 (A) ssMD-derived relative estimates of the off-rate ($k_{\mathrm{off}}(\alpha)$) for WG-3 and ribocil-B as a function of scaling factor, α (where $\alpha = 1/\alpha'_{ij}$ and α'_{ij} set to 3.25, 3.50, and 3.75), used during the simulations. Results at unscaled potential (i.e., α=1 or equivalently $\alpha' = 1$) were obtained by extrapolating the $k_{\mathrm{off}}(\alpha)$ derived from short ssMD simulations. (B) Kinetic models for binding/unbinding of WG-3 (top panel) and ribocil-B (bottom panel) to and from the FMN-aptamer domain. For WG-3, k_{on} and k_{off} were derived from simulations; k_{on} was computed as $k_{\mathrm{on}} = k_{\mathrm{off}}/K_{\mathrm{D}}$ using experimental K_{D} and k_{off} estimated from our simulations via extrapolation shown in (A). For ribocil-B, k_{on} and k_{off} correspond to previously reported experimental values. (C) Pre-equilibrium dynamics of WG-3 (left) and ribocil-B (right), simulated using the kinetics models that are shown in (B). The Sensing Window is the time elapsed between the aptamer domain's transcription and the terminator helix's formation. 50

3.2 The maximum off-rate constants lead to specific time courses for WG-3, ribocil-B and FMN at the three scaling values. 51
3.3 Representative release pathways for (A) ribocil-B and (B) WG-3 from the aptamer of the FMN-dependent riboswitch. A48 and U61 are shown as red and blue space filling models, respectively. 53
3.4 Representative unbinding trajectories. Snapshots along the representative unbinding trajectories of (A) WG-3 and (B) ribocil-B. Highlighted in the dashed box is the interval during which unbinding occurs. Indicated below each snapshot is the RMSD relative to the initial structure along the unbinding trajectories. . 54

4.1 Comparison of the densities available for each cluster and the distribution of swivel state. (A) The densities of the RNA in each cluster and the description of the corresponding swiveled state and resolutions. The β subunit is colored in cyan and the β' subunit is colored in pink. (B) Zoomed view of densities available for riboswitch. (C) Illustration of the swiveled, partial swivel and unswiveled states of the RNAP. (D) Distribution of swivel states. The distribution shows that the ligand bound state favors the unswiveled state and the ligand-unbound state favors the swiveled state. 66
4.2 (A) Illustration of the structural alteration of the fluoride riboswitch in response to the binding of fluoride ligand. (B) Nucleic acid scaffold used for cryo-EM data acquisition. The secondary structure of the riboswitch and the sequence of the template and non-template DNA are shown. (C) Secondary structure of the ligand-bound and ligand-unbound structure. The regions fitted by MDFF and resolved by cryo-EM are boxed. (D) The model after attaching the aptamer domain of the fluoride riboswitch in the ligand unbound state. The region fitted through MDFF is rendered in gold, the attached aptamer is rendered in yellow. The cryo-EM density map rendered as transparent surface. (E) The model after modeling the aptamer domain of the fluoride riboswitch in the ligand bound state. The region fitted through MDFF is colored in gold, the attached aptamer is colored in yellow. The cryo-EM density map rendered as transparent surface. 68
4.3 RNAP in complex with the MDFF fitted region of the riboswitch. The fitted residues for the ligand bound state are colored in red. The fitted residues for the ligand unbound state are colored as blue. The RNAP and template/non-template DNA are rendered as transparent cartoons.The non-template DNA is colored in dark red; the template DNA is colored in black. The α subunit of the RNAP is colored in corn flower blue, the β subunit is colored in cyan, the β' subunit is colored in pink, and the ω subunit is colored in dark green. 70
4.4 (A). Plot of the global cross-correlation coefficients for the 100 structures generated for the ligand-unbound state. For the ligand-unbound state, trajectory 46 was used as the model as it has the highest GCCC among all of the fitted structures. (B) Plot of the global cross-correlation coefficients for the 100 structures generated for the ligand-bound state. For the ligand-bound state, trajectory 1 was used as the model as it has the highest GCCC among all of the fitted structures. 72

4.5 The plane generated in Chimera to quantify the twist upon ligand binding. The angle between the ligand-unbound aptamer (blue) and the ligand-bound (gold) aptamer was calculated. 74
4.6 (A) View from the outside of the RNAP exit channel with the attached aptamer domain of the ligand bound (gold) and unbound (blue) riboswitch. A twist of the aptamer domain is observed. (B) The exit channel of the ligand-unbound complex compared to the ligand-bound complex. The β-flap, β'-dock and β'-ZBD domains are highlighted for the ligand-bound (purple) and -unbound (green) complexes. The β'-ZBD domain of the ligand-bound complex shifts and opens the exit channel compared to the ligand-unbound complex. 75
4.7 Distance between residue A55 and U60 for the ligand bound (red) and ligand unbound (blue) system. Nucleotides A55 and U60 are shown. The residues A55 and U60 are closer to each other in the ligand free system. 76
4.8 Comparison of the ligand-unbound (green) and ligand-bound (purple) states of RNAP, which are predominantly swiveled and unswiveled, respectively. 78
4.9 Time pausing assays for the wild-type and mutants RNAP in the presence of NusG. The ligand-unbound state is colored in blue and the ligand-bound state is colored in red. 79
4.10 Interactions between the riboswitch and the RNAP. (A). Interactions between the RNA residues in the MDFF-fitted region and the RNA-DNA hybrid with the RNAP for the ligand-unbound state. (B). Interactions between the RNA residues in the MDFF-fitted region and the RNA-DNA hybrid with the RNAP for the ligand-bound state. 81
4.11 (A) Space filling model for the fluoride bound complex. (B) Space filling model for the ligand unbound complex. The RNA moves away from the β'-dock domain compared to the ligand-bound state. 82
4.12 Overlaid structure of E. coli RNAP (blue) in complex with NusG (yellow)123 and the fluoride-PEC structure resolved by cryo-EM (gray). (A) NusG binds to the ligand-bound state fluoride-PEC. (B) Zoomed view of NusG near the template/non-template DNA for the ligand-bound state fluoride-PEC. (C) NusG binds to the ligand unbound state fluoride-PEC. (D) Zoomed view of NusG near the template/non-template DNA for the ligand-unbound state fluoride-PEC which shows a steric clash. 85

A.1 Plot of the unbinding profile and the exponential curve for each pose of system 2B57. The x-axis corresponds to the timeframe of the trajectory, the y-axis corresponds to the proportion of native contacts that remains. The three parameters A, τ and B correspond to a, b, and c in each plot. 92
A.2 Plot of the unbinding profile and the exponential curve for each pose of system 2XNW. The x-axis corresponds to the timeframe of the trajectory, the y-axis corresponds to the proportion of native contacts that remains. The three parameters A, τ and B correspond to a, b, and c in each plot. 93

A.3 Plot of the unbinding profile and the exponential curve for each pose of system 3NPN. The x-axis corresponds to the timeframe of the trajectory, the y-axis corresponds to the proportion of native contacts that remains. The three parameters A, τ and B correspond to a, b, and c in each plot. 94
A.4 Plot of the unbinding profile and the exponential curve for each pose of system 4XNW. The x-axis corresponds to the timeframe of the trajectory, the y-axis corresponds to the proportion of native contacts that remains. The three parameters A, τ and B correspond to a, b, and c in each plot. 95
A.5 Plot of the unbinding profile and the exponential curve for each pose of system 4XNW. The x-axis corresponds to the timeframe of the trajectory, the y-axis corresponds to the proportion of native contacts that remains. The three parameters A, τ and B correspond to a, b, and c in each plot. 96

B.1 The secondary structure of the Mn-sensing riboswitch. This figure is generated by Adrien Chauvier. The P1 (blue)-P2 (pink) and P3 (green)-P4 (gray) form two superhelices connected by the L1 (purple) and L3 (orange) loops. The L1 and L3 loops form the binding pocket for metal ions. 99
B.2 Plot of the RMSDs of the nucleotides A22, A23, G24, A25, C26, A43, U44, and U45. The RMSDs of the ligand bound riboswitch are colored in red, and the RMSDs for the ligand unbound riboswitch are colored in blue. 104
B.3 (A). Secondary structure of the Mn-sensing riboswitch. Residues with larger RMSDs in the ligand unbound state are colored in black. (B). 3D structure of the Mn-sensing riboswitch. The residues with larger RMSDs in the ligand unbound state are colored in red. 105

LIST OF TABLES

TABLE

2.1 The PDB IDs, cutoff values used to cluster poses, and the number of poses used to carry out MD simulations for each system. The resolution of the structure is included in the parenthesis next to the PDB ID. 17
2.2 The simulation-derived parameters A, τ and B for system 2B57. Pose number 0 corresponds to the native pose, and the following poses correspond to the non-native poses generated by rDock. 26
2.3 The simulation-derived parameters A, τ and B for system 2XNW. Pose number 0 corresponds to the native pose, and the following poses correspond to the non-native poses generated by rDock. 27
2.4 The simulation-derived parameters A, τ and B for system 3NPN. Pose number 0 corresponds to the native pose, and the following poses correspond to the non-native poses generated by rDock. 28
2.5 The simulation-derived parameters A, τ and B for system 4XWF. Pose number 0 corresponds to the native pose, and the following poses correspond to the non-native poses generated by rDock. 29
2.6 The simulation-derived parameters A, τ and B for system 5KPY. Pose number 0 corresponds to the native pose, and the following poses correspond to the non-native poses generated by rDock. 30
2.7 Pose-prediction accuracy metrics when using τ estimated from ssMD-derived unbinding profiles to rank RNA-ligand poses. The resolution of each structure is included in the parenthesis next to the PDB ID. N is the number of poses, and the min RMSD listed is the lowest RMSD when considering the top 1/3/5 poses. The average value for N, AUC, and min RMSD are listed at the bottom row. . 33
2.8 Pose-prediction accuracy metrics when using structure-based τ_{PLS} to rank RNA-ligand poses. The resolution of each structure is included in the parenthesis next to the PDB ID. N is the number of poses, and the min RMSD listed is the lowest RMSD when considering the top 1/3/5 poses. 38
2.9 Pose-prediction accuracy metrics when using the rDock interaction scores to rank RNA-ligand poses. The resolution of each structure is included in the parenthesis next to the PDB ID. N is the number of poses, and the min RMSD listed is the lowest RMSD when considering the top 1/3/5 poses. 38

LIST OF APPENDICES

A Supporting Information for Chapter 2 . . . . . **91**

B Study of the Undocking Pathway of the Manganese Riboswitch **97**

ABSTRACT

Riboswitches are important structural elements that are predominantly found in bacteria, regulating the expression of downstream genes in response to the binding of ligands. Therefore, studying the structures of riboswitch-ligand systems and understanding their kinetics and structure-functional relationships are crucial steps toward the design of antibacterial drugs that target riboswitches.

Computational methods such as molecular dynamics simulations and docking have been widely applied to the study of biological systems. Despite the availability of docking software and machine learning-based platforms for structure prediction, accurately predicting the structures of RNA-ligand complexes remains a challenging problem. In this thesis, I have developed computational frameworks that utilize enhanced molecular dynamics simulations to improve the prediction of native-like poses for riboswitch-ligand systems, as well as to study the kinetics and structure-functional relationships of riboswitch containing complexes.

Inspired by the application of molecular dynamics simulations to predict the structures of protein-ligand and protein-protein complexes, I have developed a workflow that improves the prediction of the poses (conformations) of riboswitch-ligand complexes. First, I carried out selectively scaled molecular dynamics simulations to generate unbinding trajectories for several RNA-ligand systems. Then I conducted contact analyses to generate unbinding profiles where the unbinding parameters can be used to predict native-like poses. A machine learning model was trained using the simulation derived data and structural information, which then can be applied to predict native poses directly from structure.

Obtaining the structures of RNA-ligand complexes is only the first step for the study of their mechanisms. Understanding the kinetics of ligand unbinding is important to study

kinetically controlled riboswitches. In chapter 3, selectively scaled molecular dynamics simulations and kinetic modeling were applied to elucidate the differential activities of ligands of the FMN riboswitch. The workflow can be applied to find hits that target a specific riboswitch.

Finally, I applied molecular dynamics flexible fitting to predict the structure of ligand bound and ligand unbound states of the fluoride riboswitch using cryo-EM density maps of paused elongation complexes and previously determined crystal structures. Next, an analysis of the structures was conducted and, combined with biochemical assays, we proposed an RNA-dependent mechanism for transcription termination induced by the fluoride riboswitch.

The works presented here enhance the understanding of the mechanisms of riboswitch-mediated regulation of gene expression, forming a potential basis to develop novel antibacterial drugs targeting riboswitches. In the future, the workflows for pose prediction and kinetic and structural modeling can be tested on different biological systems to potentially extend the scope of their application.

CHAPTER 1

Introduction

1.1 Central dogma and non-coding RNAs

1.1.1 Overview

Ribonucleic acid (RNA) is one of the most important biological macromolecules in living organisms. For example, messenger (m)RNAs transfer genetic information from deoxyribonucleic acid (DNA) to protein. The central dogma of molecular biology states that genetic information is stored in DNA and transferred from DNA to RNA through a process called transcription. RNAs are then translated to proteins that perform many essential biological functions [1]. Therefore, RNAs are considered important intermediates for genetic information transfer. Although it has been assumed that most genetic information and biological functions are carried out by proteins, recent discoveries show that the majority of the eukaryotic genome is transcribed into non-coding RNAs (ncRNAs) [2].

In eukaryotes, up to 90% of genomic DNA is transcribed [3],whereas only 2-3% of the genome is translated into proteins [3, 4, 5], indicating that the majority of RNA transcripts are not translated into proteins and thus have to be considered ncRNAs. NcRNAs play essential roles in regulating gene expression at the transcriptional or translational level by producing functional RNAs, ensuring proper cellular functions and homeostasis [5, 6, 7]. Several examples of non-coding RNAs include microRNAs [6, 7], transfer RNAs (tRNAs) [8], long ncRNAs (lncRNAs) [9, 10, 11], and ribosomal RNAs (rRNAs) [12, 13, 14]. Despite the abundance and important biological functions of ncRNAs, structural information is limited

while new ncRNAs are continuously being discovered and studied [15].

1.1.2 Examples of non-coding RNAs

One category of ncRNAs are microRNAs, small, single-stranded ncRNAs containing 19 to 25 nucleotides that participate in post-transcriptional gene regulation [16, 17]. MicroRNAs affect gene silencing by binding to target protein-coding mRNAs, as the target mRNAs contain sequences that are complementary to the microRNA [6, 18, 19]. Several mechanisms by which microRNAs reduce gene expression have been reported. One mechanism is that the microRNA binds with full complementarity, including base-pairs in the central region, resulting in the cleavage and the degradation of mRNA[7]. With cleavage occurring at the phosphodiester bond connecting the target nucleotides that are opposite to microRNA bases 10-11, mismatches are not tolerated for cleavage [20, 21]. For a microRNA with target complementarity in the seed sequence of nucleotides 2-8 of the microRNA, mRNA degradation is unlikely, and it functions instead through translational repression [7]. It has also been shown that either the dysfunction or excess biosynthesis of microRNAs is related to diseases such as asthma, allergic rhinitis, and cancer [16, 22].

Another category of non-coding RNAs are ribozymes, the catalytic RNA that was discovered in 1982 [23]. Unlike proteins, which are made up of 20 different amino acids, ribozymes only contain four different nucleotides to perform catalytic functions [24]. Ribozymes can be classified as self-splicing ribozymes and self-cleaving ribozymes [24]. Self-splicing ribozymes require the recruitment of metal-ion cofactors to perform catalytic functions while self-cleaving ribozymes mostly use nucleotide bases for catalytic functions [25, 26]. Ribozymes can affect multiple RNA processing reactions such as those involved in RNA splicing and tRNA synthesis.

1.1.3 Riboswitches

Beside the ncRNAs mentioned above, riboswitches are of particular interest as they play critical roles in gene regulation. Riboswitches are structured elements that typically are located in the 5'-untranslated region (5'-UTR) of a messenger RNA (mRNA), regulating the expression of downstream genes through the binding of ligands [27, 28, 29]. Riboswitches consist of an aptamer domain and an expression platform. The aptamer domain is highly structured and contains a binding pocket that recognizes specific small molecules or ions. The binding of the ligand alters the conformation of the downstream expression platform, which then either promotes or suppresses the expression of downstream genes [29, 30].

Riboswitches can be categorized as transcriptional riboswitches or translational riboswitches based on their mechanism of action. Transcriptional riboswitches regulate gene expression at the transcriptional level, while translational riboswitches regulate gene expression at the translational level [27, 31]. Transcriptional riboswitches are predominantly found in Gram-positive bacteria, whereas translational riboswitches are commonly found in Gram-negative bacteria expressing the conserved Rho protein, which acts as a transcription terminator [30, 32].

Riboswitches can also be classified as ON-switches or OFF-switches based on whether the binding of ligand will result in the expression or suppression of downstream genes, respectively. ON-switches promote the expression of downstream genes upon the binding of ligand. One example of ON-switch is the fluoride riboswitch. It selectively binds to fluoride ions, and the binding of ligand promotes the expression of a protein that exports the fluoride ions, which can be toxic for the cell [33]. In contrast, OFF-switches terminate the transcription or inhibit translation of downstream genes upon the binding of ligand [34]. One example of OFF-switches is the FMN riboswitch, which selectively binds flavin mononucleotide (FMN). The binding of the ligand then triggers the termination of downstream expression of a gene that encodes for FMN biosynthesis [35, 36]. Since gene regulation modulated by riboswitches is essential for proper function of the cell and maintenance of cellular homeostasis, most OFF-

switches terminate the expression of genes related to the production of essential metabolites, whereas ON-switches typically regulate genes related to the production of export proteins for cytotoxic ligands [34].

Riboswitches appear predominantly in bacteria and are involved in the regulation of genes essential for cellular function. In addition to their important biological function, riboswitches contain a highly conserved aptamer domain with a structured binding pocket, making them a promising target for antibacterial drug design [30, 37].

1.2 Experimental methods for RNA structure determination

1.2.1 Overview of RNA structure

Elucidating the structures of RNAs is an important first step in the study of their biological functions. RNAs are polymeric biomolecules made up of four monomers termed nucleotides. Each nucleotide consists of a nucleobase, a ribose sugar ring, and a phosphate group. The four nucleotides include two purines and two pyrimidines, namely, adenine (A), guanine (G) and cytosine (C), uracil (U), respectively.

The structure of RNA can be classified into three levels, namely, primary structure, secondary structure, and tertiary structure. Primary structure is also called 1D structure, which is the sequence of the RNA. Secondary structure or 2D structure contains information about the connectivity and base pairing of the RNA, including four Watson-Crick base pairs (A-U, U-A, C-G, G-C) and two wobble base pairs (G-U, U-G) which form stem regions, and single-stranded regions which form hairpins, junctions, and internal loops. Because of the structural interchangeability of these base pairs, many primary structures can give rise to a given secondary structure. The tertiary structure of RNA is the three-dimensional structure

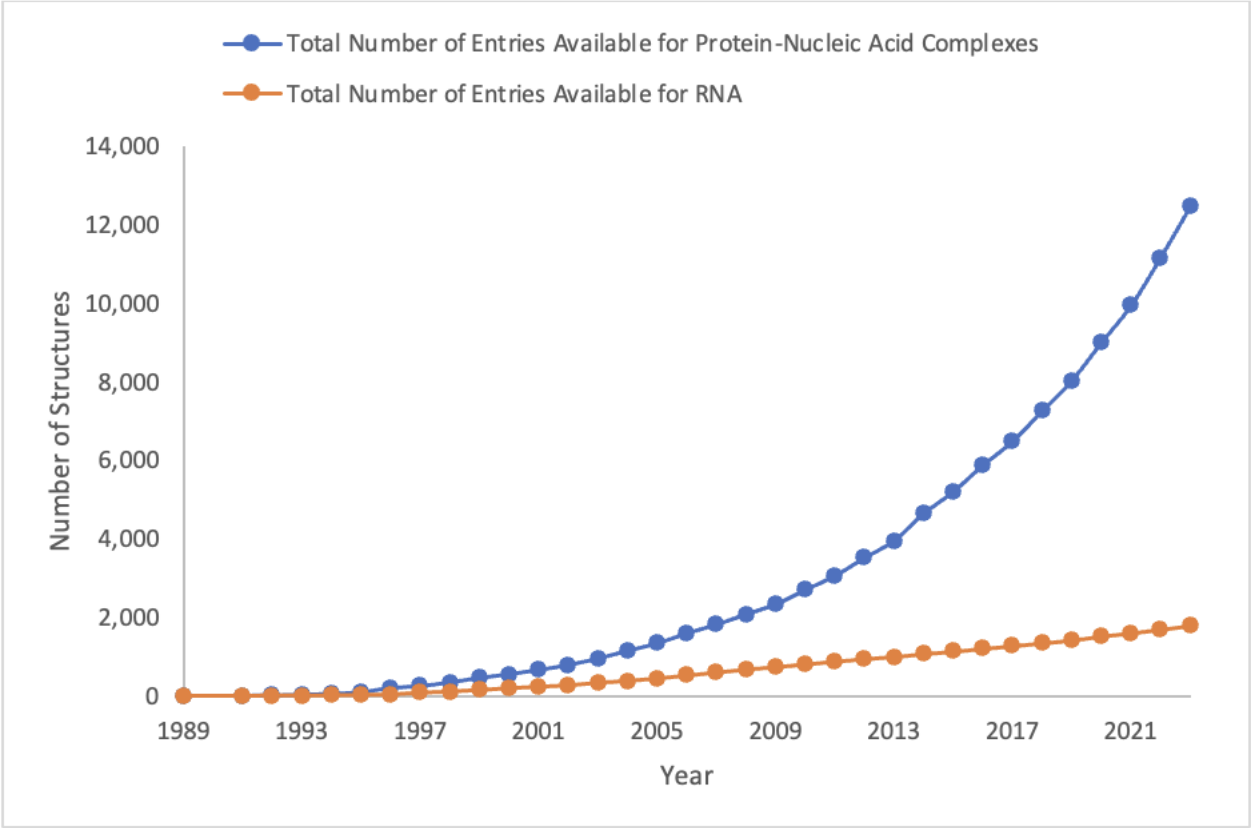


Figure 1.1: Total number of RNA structures deposited on the Protein Data Bank (PDB) through 2023 for structures of protein-nucleic acid complexes (top) and structures that contains only RNA (bottom).

of the RNA, which shows how the RNA arranges in 3-dimensional space. The function of RNAs typically depends on the tertiary structure, although secondary structures such as pseudoknots are shown to be important for RNA functions such as splicing and riboswitch conformational changes [38].

Experimental methods such as X-ray crystallography, nuclear magnetic resonance (NMR) and cryo-electron microscopy (cryo-EM) have been developed to elucidate the three-dimensional structures of RNAs. With the advancement of experimental techniques, the number of RNA structures available in the Protein Data Bank has increased significantly over time. The number of RNA-containing structures and the number of RNA-only structures by the end of 2023 is shown in Figure1.1.

1.2.2 X-ray crystallography

X-ray crystallography is an experimental technique that can accurately determine the three-dimensional structures of biological molecules such as proteins and nucleic acids. In X-ray crystallography, a sample is purified and then crystalized at high concentration [39]. Then the crystals are exposed to an X-ray beam where the diffraction patterns can be used to get structural information such as symmetry and the size of the repeating unit. Then a map of electron density is calculated from the diffraction spot [40]. Although the first X-ray crystallography structure obtained was yeast tRNA [40], most structures resolved using X-ray crystallography are proteins structures., RNA containing structures only account for a very small portion of the available crystal structures [41]. The sparsity of the RNA containing crystal structures is due to the difficulty in obtaining pure and properly folded RNA and the low sequence diversity, which makes it hard to design crystal contacts [41].

1.2.3 Nuclear magnetic resonance spectroscopy (NMR)

Nuclear magnetic resonance (NMR) spectroscopy is another widely used method to resolve the structures of RNAs in solution [42]. NMR spectroscopy takes advantage of the charged nature of nuclei of atoms. In NMR, an external magnetic field and the frequency of the local magnetic field is measured. Since the frequency depends on the nucleus and the local chemical environment, the chemical shifts calculated using the resonance frequency of the nucleus and the reference frequency are then used to provide information of the local structure and used to determine the structures of RNAs.

One advantage of solution NMR is that instead of determining the structure of a crystalized molecule, the measurement is carried out in solution so that it can capture the dynamics of the biomolecule at the molecular level, providing insights for the interactions between biomolecules. However, despite the fact that nearly half of all RNA structures were solved by NMR, the size of the RNA structures that can be resolved by NMR is typically

limited to 100 nucleotides for intermediate resolution and 50 nucleotides for high resolution [43]. This shortcoming is due to the fact that the increase in size will cause significant signal broadening [43], making it difficult to accurately resolve structural features.

1.2.4 Cryo-EM

Cryogenic electron microscopy (cryo-EM) is another experimental method for RNA structure determination. Like X-ray crystallography, cryo-EM relies on the principle of electron density, although here the scattering of electron beams is used to generate structures. One advantage of cryo-EM over X-ray crystallography is that it can capture the conformational dynamics of RNAs. For cryo-EM, an electron beam is passing through the sample embedded in a thin amorphous ice layer, and the scattered electrons are recorded and processed to produce the three-dimensional structure of the molecule [44, 45]. Although cryo-EM can capture the conformational dynamics of biomolecules, it can be challenging when using cryo-EM to resolve RNA structures due to their conformational flexibility [44, 46], which give rise to the lower resolution of the structures.

1.2.5 Limitations

Although the experimental methods mentioned above are powerful tools to resolve the structures of biomolecules, there are several limitations that require the use of computational methods to gain a detailed view of biological processes. For X-ray crystallography, the molecules of interest are crystalized, so this technique does not apply to molecules that do not form crystals, and it cannot capture the dynamic nature of the molecules. For NMR and cryo-EM, even though the dynamics can be captured at the molecular level, the transient interactions that are essential for understanding mechanisms typically cannot be resolved at

an atomic level.

1.3 Molecular dynamics simulations

1.3.1 Overview of molecular dynamics simulations

The most widely used physics-based computational method to study biological systems is molecular dynamics (MD) simulation, a computational method that simulates the motion of systems of molecules or particles [47, 48, 49]. MD simulations can be classified as coarse-grained MD simulations [50, 51] and all-atom simulations based on the representation of molecules.

Coarse-grained simulations use particles to represent groups of atoms rather than individual atoms [51]. The advantage of coarse-grained simulations is the ability to simulate larger systems for longer simulation times at lower compute cost. However, this advantage is traded for accuracy, since coarse-grained simulations can only provide approximate simulations for the system of interest. Coarse-grained simulations are typically used for simulating large systems such as antibody solutions [52] and viral capsids [53].

An all-atom MD simulation starts with initial positions and initial velocities of each atom. If the initial velocities are not specified, they will be randomly assigned based on the Maxwell-Boltzmann distribution at the specified temperature. Then the force experienced by each atom is calculated with a force field, which refers to the set of formula and parameters that are used to calculate the interaction energies of the system as a function of atomic coordinates [54]. For a given time interval (time step), the motions of the atoms are calculated by solving Newton's equations of motion and the energy of the system is computed over a time interval of typically 1 fs, then saved as a frame of the trajectory. The process continues until it reaches the last specified time frame. The workflow of MD simulations is shown in Figure 1.2.

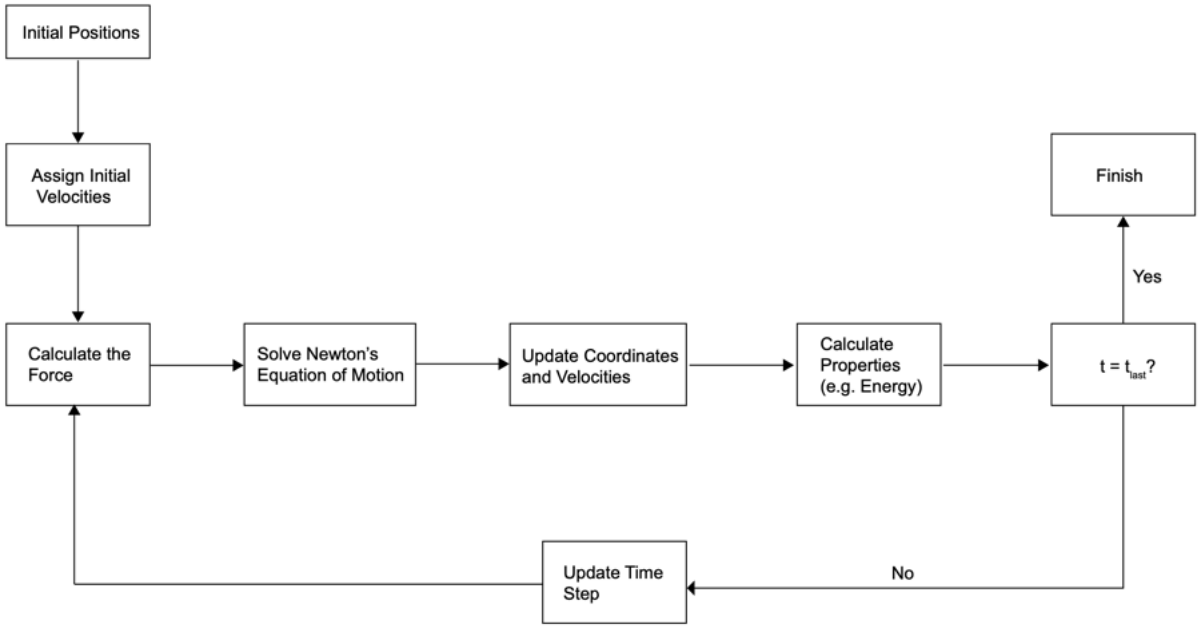


Figure 1.2: The workflow of molecular dynamics simulations.

All-atom MD simulations are more computationally expensive than coarse-grained simulations since the atoms are represented explicitly and there are more degrees of freedom. These features also allow MD simulations to show the atomistic details of the molecules and to provide insight into interactions that are hard to observe experimentally.

1.3.2 Enhanced sampling methods

Many biological processes have relatively long timescales, usually on the millisecond order or above. Despite the increase in availability of computational resources, the timescale of those biological processes makes it infeasible to observe the transitions of interest in a reasonable amount of time using conventional MD simulations. To better sample the conformational transitions, enhanced sampling methods have been developed. Commonly used enhanced sampling methods include scaled MD simulations, metadynamics, and steered

MD simulations [55, 56, 57, 58].

1.3.2.1 Selectively scaled MD simulations

Selectively scaled MD simulation is particularly useful in the study of ligand unbinding for biological systems. It has been used to study the ligand unbinding kinetics and drug residence times [59, 60]. In selectively scaled MD simulations, a specific energy term of the potential energy function is scaled to accelerate a process of interest. For example, the Lennard-Jones potential can be scaled to accelerate ligand unbinding [59]. Aside from the ability to simulate processes that are beyond the scope of conventional MD simulations, scaled MD simulation enables the recovery of kinetics without the use of a reaction coordinate, which improves the computational efficiency [56]. Selectively scaled MD is a major technique used in chapter 2 and chapter 3 of this dissertation for the study of RNA-ligand systems.

1.3.2.2 Steered MD (SMD) simulations

Another enhanced sampling method that is suitable for the study of biological systems with ligand is steered MD simulations. Steered MD simulations have been used in rational drug design and the investigation of the mechanical functions of proteins [61, 62].

In steered MD, one atom or a group of atoms are fixed, and a steered atom/atoms group (SMD atom) are picked where a force is applied to pull the SMD atom to induce the desired conformational changes. There are two ways of pulling the SMD atom, namely, constant force pulling and constant velocity pulling. For constant force pulling, the force applied to the SMD atom is a constant and is in the direction of the vector pointing to the SMD atom from the fixed atom. For constant velocity pulling, the SMD atom is pulled at constant velocity via a virtual spring, meaning the magnitude of the force increases with time.

1.4 Application of Machine Learning in computational chemistry

Machine learning describes data-driven methods that have been widely used in the study of biological systems. Machine learning can be classified as supervised learning, unsupervised learning, and reinforcement learning. Supervised learning is the most used method as it leverages the available experimental data to predict the properties for the system of interested. For example, supervised learning has been used to predict properties such as chemical shifts, native poses and solvent accessible area [63, 64, 65, 66].

Supervised learning models are supplied with labeled input data; the algorithm then maps the input data to the labels. To test the performance of the model, the data sets are typically divided into training sets and testing sets, where only the data from the training set are fed into the model. The data from the testing set is used to check the accuracy of the prediction. The results of the validation can be used to tune the parameters used for the model, in order to improve the accuracy of prediction. However, supervised learning models have the risk of overfitting, which gives poor accuracy for the testing set compared to the training set.

With more data being generated experimentally and computationally, the interest in applying machine learning to predict the properties of biological system continues to surge.

1.5 Thesis outline

Accurately determining the structure of RNAs and RNA-containing complexes is an important step in investigating their biological functions. Recent advanced in enhanced sampling methods of MD simulations enable the study of biological processes such as ligand unbinding.

Inspired by the application of MD simulations to predict the structures of protein-ligand systems, a method that leverages enhanced molecular dynamics simulations and machine

learning to predict the native-like poses for an RNA aptamer-ligand system was developed. First, selectively scaled MD simulations were conducted on five RNA-ligand systems, and we obtained the unbinding trajectories from the simulations. Then contact analysis was conducted on the trajectories to generate unbinding profiles. The characteristic decay time derived from the unbinding trajectory was shown to be able to resolve native poses. A partial least square model was then trained to predict the native-like poses directly from structures. The study of RNA-ligand pose prediction is presented in chapter 2.

In chapter 3, selectively scaled MD simulations were applied to the activities of three small molecules that bind to the FMN riboswitch. In this work, we rationalized the different activity of the ligands in inhibiting bacterial growth by extracting kinetic information from the unbinding simulations. The on-rate and off-rate were estimated from the MD simulations, and representative paths of the unbinding of the ligands were identified.

In chapter 4, I employed another simulation method, molecular dynamics flexible fitting (MDFF) simulations to resolve the structures of the fluoride riboswitch in paused transcription elongation complexes with ligand bound and unbound. To this end, I used an available crystal structure of the fluoride riboswitch and the density maps obtained from the cryo-EM experiments. In this work, we combined computational method with biochemical techniques including biochemical assays and cryo-electron microscopy to study the coupling between the fluoride riboswitch and the termination of RNAP transcription.

The main goal of my thesis is to apply molecular dynamics simulations to study the pose prediction, kinetics, and mechanisms of riboswitch-containing complexes. A summary of the thesis and outlook will be presented in chapter 5.

CHAPTER 2

Using Selectively Scaled Molecular Dynamics Simulations to Assess Ligand Poses in RNA Aptamers

The contents of this chapter were published in the following reference:

Liu, Y.; Frank, A.T. Using Selectively Scaled Molecular Dynamics Simulations to Assess Ligand Poses in RNA Aptamers. *J. Chem. Theory Comput.* **2022**, 18(9), 5703-5709.

Aaron T. Frank conceived the project, Yichen Liu carried out simulations. Yichen Liu and Aaron T. Frank carried out the analysis and wrote and reviewed the manuscript.

Abstract

Predicting the structure of RNA-ligand complexes is an important problem in RNA structural biology. Although computational docking tools such as rDock and AutoDock can be used to generate poses for RNA-ligand complexes, accurately resolving the native poses from non-native poses is still a challenge for RNA-ligand systems. To improve the prediction of

native or near-native poses for RNA-ligand complexes, we developed a method that leverage enhanced molecular dynamics simulations and machine learning to predict the native poses. Based on the assumption that native-like poses are more stable than non-native poses and are thus less likely to dissociate during molecular dynamics simulations, we used enhanced simulations to promote ligand unbinding for poses of five RNA aptamer-ligand complexes. We then fitted unbinding profiles to a single exponential and identified the characteristic decay time (τ) as particularly effective at identifying native-like poses from decoys. It was also found that a simple partial least square regression model trained to predict the simulation-derived parameters directly from structure could also discriminate ligand poses for similar RNA aptamers. Characterizing the unbinding properties of poses may therefore be an effective way to enhance pose prediction accuracy for RNA aptamer-ligand complexes. A similar workflow might be applicable to other RNA-ligand systems; however, further analysis will be required to confirm this hypothesis.

2.1 Introduction

Inspired by the approval of the first RNA-targeting small molecule drug [67] and the discovery of small molecules that target structured RNA that found in microbes [68, 69, 70], interest in applying structure-centric methods to identify RNA-targeting small molecules continues to increase. For a well-validated target and a clear and demonstrable relationship between its conformational states and biological function [71], structure-based methods can be used to identify small molecules that bind to and perturb the conformational distribution of that RNA. By designing conformationally specific RNA binders, one could use them to alter RNA function in a therapeutically desirable manner.

Structure-based virtual screening (SBVS) is a commonly used method for identifying small molecules that target a specific conformation of RNA [72, 73]. SBVS involves docking

a library of small molecule ligands to a structure of the target and based on the docking scores (binding affinity estimates), sorting them to identify the subset of compounds that are likely to bind to the target. If hits were found within a small subset of compound selected based on the docking, the screening is considered successful. Although docking software such as rDock and AutoDock can generated poses, the quality of the pose prediction for RNA-ligand systems are not as good as that for protein-ligand systems, due to the flexible nature of RNA. The inaccurate pose generated for RNA-ligand complexes will then lead to inaccurate binding energy calculations, because the binding affinity calculation is based on the poses generated. Therefore, accurately predicting the native poses of RNA-ligand systems is crucial for RNA drug discovery.

Inspired by the application of molecular dynamics (MD) simulations to assist the prediction of protein-ligand [74, 75, 76, 77] and protein-protein [78, 79] poses, we attempted to explore whether we can use information derived from MD simulations to enhance pose prediction of RNA-ligand complexes. In this project, we focused on the aptamers of riboswitches, highly structured RNA elements with well-defined binding pockets, regulating the expression of downstream genes. We hypothesized that the native or near-native poses of the ligand are more stable and thus are less likely to dissociate during MD simulations. Since the timescale of the unbinding is beyond the scope of conventional molecular dynamics simulations, an enhanced sampling method, selectively scaled MD simulations (ssMD) [59], was employed to accelerate the dissociation of ligand. Then we generated the unbinding profiles from the resulting trajectories for individual poses. Several parameters were derived from the binding profiles, and we assessed the resolving power of these parameters in terms of identifying the native pose. Overall, the characteristic decay time (τ) generated from the unbinding profiles can be used to discriminating native poses from non-native RNA-ligand poses. Furthermore, inspired by previous studies that show that one can use machine learning to learn from the data generated by simulations [80], we showed that we could reproduce the simulation-derived unbinding parameters directly from the structure of individual poses.

We demonstrated that the characteristic decay time τ predicted from structure using the partial least squares regression model retained the ability to identify native poses. We envision that this work should pave the way for using a kinetic framework to assess docked poses of RNA-ligand complexes.

2.2 Methods

2.2.1 Pose generation and selection

We conducted selectively-scaled molecular dynamics (ssMD) simulations on 5 RNA-ligand complexes [59]. The systems correspond to five riboswitch aptamers and their cognate ligand. The structures of the 5 RNA aptamers were downloaded from the Protein Data Bank (PDB: http://www.pdb.org), corresponding to the following PDB IDs: 2B57, 2XNW, 3NPN, 4XWF, and 5KPY. The structures of the five RNA-ligand complexes are shown in Figure 2.1. For each system, 100 non-native poses were generated using rDock [81]. Since the native poses and the binding pockets are known, the cavity generation was conducted using the reference ligand method, where the binding site of the native ligand was used as reference for cavity mapping. The maximum number of cavities was set to 1 so that the poses was generated in the same binding pocket as the native pose. The mapping algorithm used was RbtLigandSiteMapper, which is the default mapping algorithm of rDock. The radius of the cavity mapping region was set to 7.0 Å and the radius of the small probe was 1.0 Å. The 100 poses generated for each system was ranked based on the estimated interaction energies calculated using the standard scoring function of rDock. The standard scoring function includes the intermolecular, intra-ligand, and intra-target terms. The poses generated for each system were clustered in order to reduce the computational cost for molecular dynamics simulations. The Multiscale Modeling Tools for Structural Biology (MMTSB) tool set was used to carry out the clustering [82]. Poses were clustered based on the Cartesian coordinate

root mean squared deviation (RMSD) between structures, where poses with similar coordinates were grouped into the same cluster. The calculation of RMSD only included heavy atoms. In order to keep the number of clusters to be less than or equal to 20, different cut-off values were used for the clustering of each system. The RMSD cutoff values used for clustering the 5 systems were 1.0 Å, 1.5 Å, 4.5 Å, 4.0 Å, and 2.0 Å, respectively. And the number of clusters generated for the non-native poses of the 5 system is 18, 19, 13, 18 and 13, respectively. The PDB ID, cutoff value, and the total number of clusters generated for each system was summarized in Table 2.1. Since the total number of clusters showed in Table 2.1 also included the native pose, the number of clusters showed in column N would be 19, 20, 14, 19 and 14, respectively. For each system, the native pose and the best pose from each cluster were selected as candidate poses for subsequent MD simulations, the best pose was selected based on the docking score provided by rDock.

System	**Cutoff**	N
2B57 (2.15 Å)	1.0 Å	19
2XNW (1.50 Å)	1.5 Å	20
3NPN (2.79 Å)	4.5 Å	14
4XWF (1.80 Å)	4.0 Å	19
5KPY (2.00 Å)	2.0 Å	14

Table 2.1: The PDB IDs, cutoff values used to cluster poses, and the number of poses used to carry out MD simulations for each system. The resolution of the structure is included in the parenthesis next to the PDB ID.

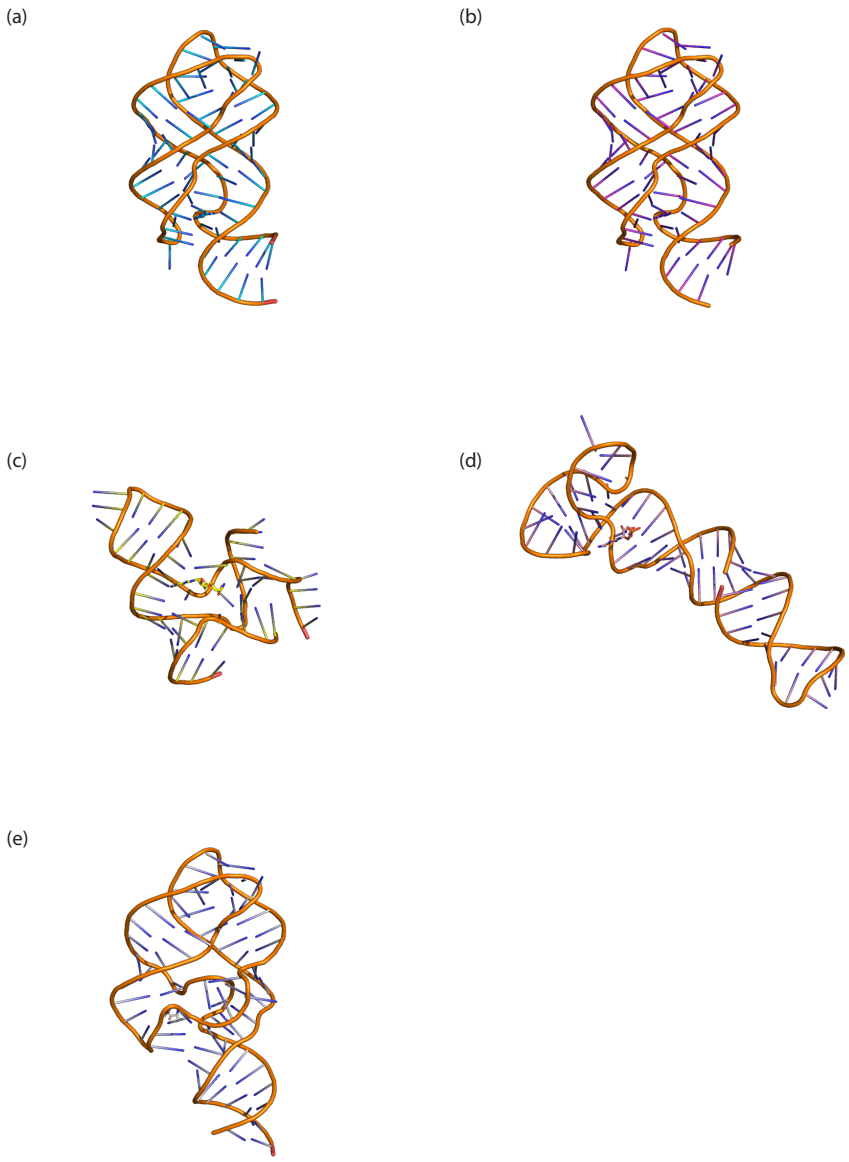


Figure 2.1: Structures of the five RNA-ligand complexes we studied. (a) PDB ID: 2B57. Purine riboswitch bound to 2,6-Diaminopurine. (b) PDB ID: 2XWN. Purine riboswitch bound to triazolo-triazole-diamine. (c) PDB ID: 3NPN. SAH riboswitch bound to S-adenosylhomocysteine. (d) PDB ID: 4XWF. ZTP riboswitch bound to ZMP. (e) PDB ID: 5KPY. 5-hydroxytryptophan aptamer bound to 5-hydroxy-L-tryptophan.

2.2.2 System preparation

The coordinate files of the 5 RNA-ligand complexes were submitted to the CHARMM-GUI webserver to generate the input files needed for MD simulations [83]. The CHARMM General Force Field (CGenFF) was used to simulate the ligand and the CHARMM36 nucleic acid force field was used to simulate RNA [84]. The complexes were solvated in a cubic TIP3P water box with the minimum distance between the solute and the boundary set to 10 Å. To neutralize the systems, potassium ions (K+) and chloride ions (Cl-) were added to the systems.

2.2.3 Molecular dynamics simulations

The simulations were carried out using the CHARMM/OpenMM interface [85, 86]. The simulation consists of an equilibration run and a production run. One graphic processing unit (GPU) and one core were used for the equilibration run, while one GPU and eight cores were used for the production run. The NVT condition was used for equilibration simulation and the SHAKE algorithm was used to constrain the hydrogen atoms [87]. Harmonic restrains were applied to the heavy atoms of RNA. For backbone heavy atoms, the force constant used was 1.0 kcal/mol/Å^2, while for nucleobase heavy atoms, the force constant used was 0.1 kcal/mol/Å^2. The time step of the equilibration simulation was set to 1 fs and each equilibration run consists of 2,000,000 steps, so the length of the equilibration simulation was 2 ns.

The equilibrated structures were used as the starting structures of the production simulations. For production simulations, the hydrogen atoms were constrained using the SHAKE algorithm. To accelerate the unbinding of ligand, selectively scaled MD simulations was used where the selected energy terms are scaled [59]. For the 5 RNA-ligand systems, the Lennard-Jones potential between water molecules and ligand was scaled up since the assumption is that by increasing the interactions between water molecules and the ligand, the ligand is

more likely to dissociate from the binding pocket, and thus reducing the simulation time required to observe unbinding events. The ϵ_{ij} term of the of the Lennard-Jones potential of the water and ligand atoms was scaled up by the scaling factor α'_{ij}.

$$V_{ij} = \epsilon_{ij} \times \alpha'_{ij} \left[\left(\frac{r_{ij,\min}}{r_{ij}} \right)^{12} - \left(\frac{r_{ij,\min}}{r_{ij}} \right)^{6} \right] \tag{2.1}$$

The scaling factor α'_{ij} was set to be 3.5. Figure 2.2 shows the scaling of the Lennard-Jones potential. The NPT conditions were used for the production simulations, and for each candidate poses from each of the RNA-ligand complex, 20 independent simulations were carried out. The time step for the production run was set to 2 fs and each simulation consists of 5,000,000 steps, meaning that the length of the simulation was 10 ns. Coordinates of the complexes were written out to a trajectory file every 5000 steps so that each trajectory has 1000 frames. The trajectories generated from the production simulations were used for subsequent analysis.

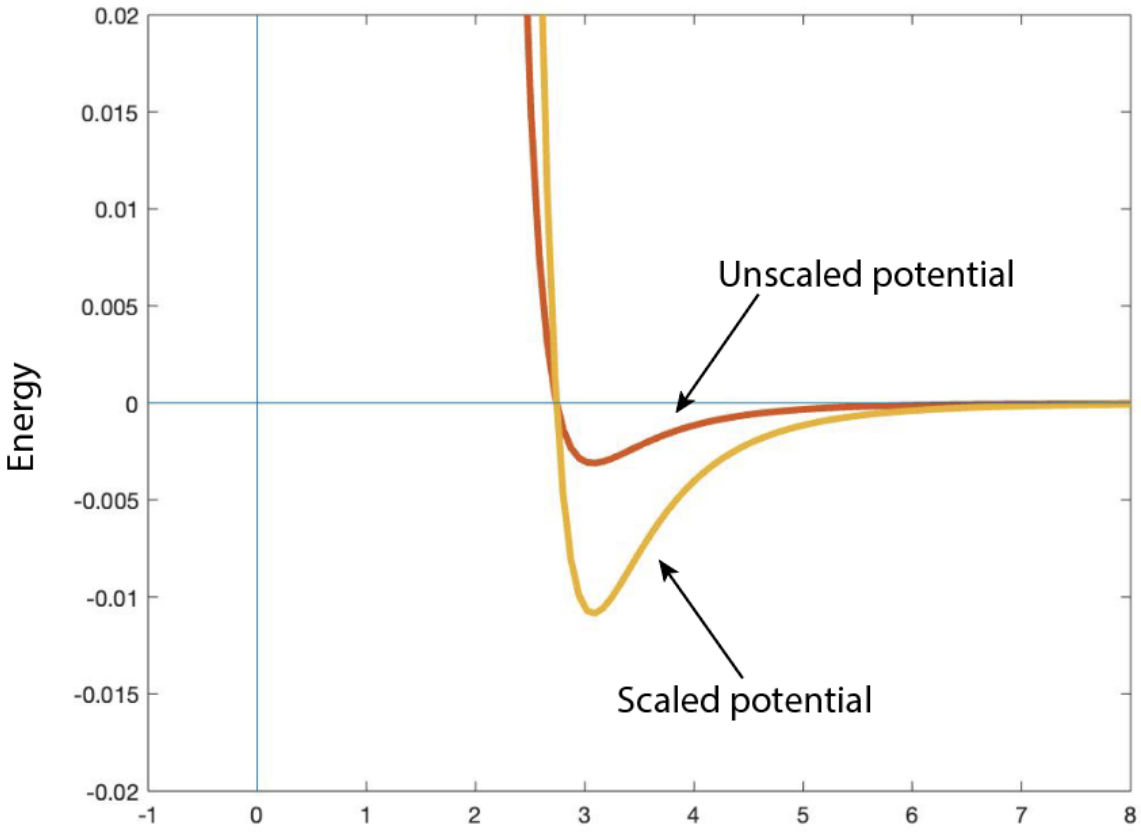


Figure 2.2: To observe the unbinding event in relatively short simulations, the interactions between the ligand atoms and water atoms are scaled up by the scaling factor $\alpha'_{ij} = 3.5$. Shown are the scaled potential (yellow) and the unscaled potential (orange).

2.2.4 Analysis of molecular dynamics simulations

The primary focus of the analysis is on the contacts between ligand and the RNA. For each pose, a reference file containing the native contacts was generated using the equilibrated structure of the pose and a cutoff value of 3.5 Å. If the distance between a pair of atoms in the ligand and RNA is within the cutoff value, 3.5 Å, the atoms are considered in contact with each other. To calculate the contacts from the simulation trajectories, the contact tool from MoleTools https://github.com/atfrank/MoleTools was used. To study the unbinding of the ligand, the mean number of reference contacts that remains at each frame were calculated over the 20 independent trajectories, and this is defined as the unbinding profile of the pose. The unbinding profile was then fitted to a single exponential decay function, $\langle M(t) \rangle = Ae^{(-t/\tau)} + B$. Where $\langle M(t) \rangle$ is the average number of the contacts that remains at time t. And the t is the timeframe of the trajectory. Three parameters were extracted from the exponential fitting and were used to evaluate the stability of the poses (Figure 2.3).

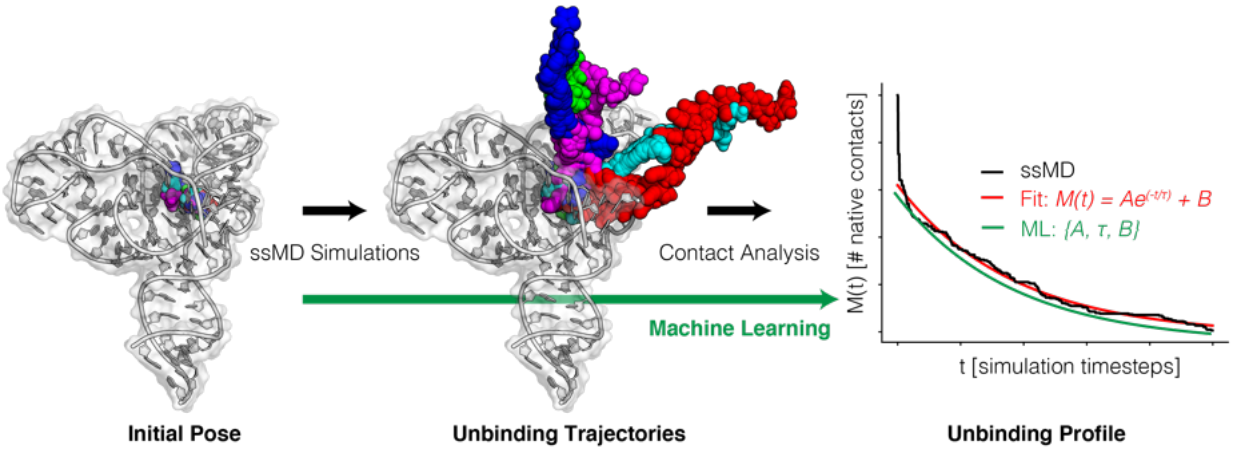


Figure 2.3: Illustration of the simulation-based framework employed in this study. For individual poses, we generate a series of unbinding trajectories using the selectively scaled MD (ssMD) technique. We then apply contact analysis to the resulting trajectories and, from them, compute an unbinding profile, defined as the number of native contacts that persist as a function of time ($\langle M(t) \rangle$). We then fit the unbinding profile to a single-exponential function to extract a set of parameters (A, τ and B) that compactly describe the unbinding profile. Here, we assess whether we could use any of these parameters to correctly identify near-native poses. In an attempt to bypass costly validation simulations, we also explore whether we could learn from simulations and predict, directly from the structure of the initial pose, the parameters of the unbinding profile (green arrow).

2.3 Results and Discussion

2.3.1 Simulation-Derived Unbinding Profiles Resolve Native-Like Poses from Decoys

This work was motivated by our interest in developing methods that can enhance RNA-ligand pose prediction, that is, methods capable of accurately discriminating native poses from non-native (decoy) poses. If one assumes that the native poses of RNA-ligand complexes correspond to the most stable RNA-ligand poses, it seems likely that near-native poses will exhibit slower dissociation rate compared to non-native poses. Based on this assumption, we sought to assess whether native-like poses show distinct unbinding profiles from non-native poses. To address this, enhanced molecular dynamics (MD) simulations were used to generate all-atom unbinding trajectories of diverse poses for a set of five benchmark RNA-ligand complexes (Figure 2.4a). The enhanced sampling method used was selectively scaled MD (ssMD) simulations61. In ssMD simulations, water-ligand interactions are strengthened such that dissociation trajectories can be observed over the course of relatively short simulations, and without any a priori assumption about the details of the escape paths of the ligand.

For each pose in each of the five RNA-ligand system, a set of 20 independent unbinding trajectories were generated using the ssMD method. From the trajectories, we then extracted the time series of all the reference contacts present in the initial equilibrated pose. Across the 20 replicas, we then computed the average number of reference contacts that present at each time step in the saved trajectories. The resulting unbinding profile was then fitted to a single exponential decay function $\langle M(t) \rangle = Ae^{(-t/\tau)} + B$.

For the five systems, we carried out receiver-operating-characteristic (ROC) analysis to generate ROC plots when poses are ranked using either A, B, or τ. The parameters A, B, and τ generated for each pose of the five systems are shown in Table 2.2, 2.3, 2.4, 2.5 and 2.6, respectively. And the plots of the exponential decay for each pose are shown in Figure A.1, A.2, A.3, A.4 and A.5, respectively. The ROC curve is the plot of true positive rate

versus false positive rate at different threshold values. In the ROC plots, one should focus on the area under the curve (AUC), an AUC approaching 1 corresponds to instances where a given metric (in this case A, B, or τ, respectively) can perfectly separate a set of objects with binary labels, in this case the "native-like" and "non-native/decoy". By contrast, an AUC value near 0.5 correspond to what would be expected if labels were randomly assigned to the poses.

In Figure 2.4b, we show ROC plots for the five test cases when using τ to rank poses. Among the three parameters generated from the exponential fit, τ exhibited the highest resolving power, as demonstrated by the AUC values (Figure 2.5). Overall the AUC_τ values ranged between 0.56 and 0.92, with a mean of 0.80 (Table 2.7). In contrast, the mean AUC_A and AUC_B were 0.41 (Figure 2.5a) and 0.72 (Figure 2.5c), respectively. Shown in Figure 2.4c is the comparison between the native poses and the best poses among the top 1, 3 and 5 when they are ranked by τ. The RMSD between the native pose and the pose with the highest τ (top 1 prediction) ranged between 0.44 and 8.36 Å, with a mean of 3.21 Å. By comparison, the average best RMSD across the five systems was 1.03 Å and 0.00 Å when considering the top 3 and 5 ranked poses (Figure 2.4c and Table 2.7). As such, across the five test cases, we could accurately identify near-native RNA-ligand poses by ranking them based on their τ derived from the ssMD unbinding simulations and identifying at most, the top five poses. It is worth noting that the pose-prediction AUC values obtained using the rDock interaction score term were lower than those obtained when ranking poses using ssMD-derived τ. The ROC plots for the five test cases when using the rDock interaction score were shown in Figure 2.6. The average AUC from the rDock interaction score term was 0.76 (Figure 2.6), compared with the 0.80 value obtained when using τ (Figure 2.4). As such, ssMD-derived τ lead to enhanced pose prediction.

Pose Number	**A**	τ	**B**
0	0.635	366.820	0.015
1	0.757	21.055	0.003
2	0.722	320.598	0.078
3	0.559	258.628	0.045
4	0.711	379.913	0.024
5	0.958	2.630	0.001
6	0.524	52.526	0.015
7	0.954	7.619	0.000
8	0.635	110.635	0.008
9	0.333	223.299	0.230
10	0.648	71.896	0.011
11	0.669	53.897	0.007
12	0.830	26.362	0.013
13	0.530	124.186	0.024
14	0.599	54.019	0.010
15	0.485	74.052	0.017
16	0.618	149.651	0.010
17	0.791	34.520	0.011
18	0.589	61.622	0.009

Table 2.2: The simulation-derived parameters A, τ and B for system 2B57. Pose number 0 corresponds to the native pose, and the following poses correspond to the non-native poses generated by rDock.

Pose Number	**A**	τ	**B**
0	0.643	90.839	0.024
1	0.572	134.261	0.021
2	0.899	19.599	0.027
3	0.634	79.296	0.005
4	0.502	59.322	0.040
5	0.961	1.486	0.001
6	0.644	49.176	0.022
7	0.926	35.619	0.009
8	0.778	52.192	0.001
9	0.610	18.345	0.047
10	0.575	88.167	0.026
11	0.692	56.491	0.019
12	0.676	145.519	0.033
13	0.625	33.617	0.004
14	0.605	142.536	0.016
15	0.557	17.686	0.009
16	0.605	51.490	0.039
17	0.638	61.791	0.023
18	0.475	37.420	0.016
19	0.755	80.241	0.014

Table 2.3: The simulation-derived parameters A, τ and B for system 2XNW. Pose number 0 corresponds to the native pose, and the following poses correspond to the non-native poses generated by rDock.

Pose Number	**A**	τ	**B**
0	0.708	25.560	0.002
1	0.776	248.143	0.012
2	0.896	4.145	0.000
3	0.936	4.814	0.001
4	0.906	27.755	0.001
5	0.681	51.668	0.013
6	0.935	3.897	0.001
7	1.059	3.657	0.000
8	0.865	9.547	0.002
9	1.039	2.500	-0.000
10	1.010	1.554	-0.000
11	1.081	2.947	-0.000
12	1.039	2.325	0.000
13	0.657	24.452	0.008

Table 2.4: The simulation-derived parameters A, τ and B for system 3NPN. Pose number 0 corresponds to the native pose, and the following poses correspond to the non-native poses generated by rDock.

Pose Number	**A**	τ	**B**
0	0.864	10.834	0.002
1	0.964	3.253	0.000
2	0.908	6.473	0.000
3	0.914	6.880	0.000
4	1.045	8.291	0.001
5	0.858	10.261	0.001
6	0.862	7.140	0.001
7	0.995	0.601	0.000
8	0.924	23.719	-0.000
9	0.931	3.893	0.000
10	0.874	15.436	0.002
11	1.025	3.551	0.001
12	0.921	4.629	0.001
13	0.974	1.656	0.001
14	1.021	5.803	-0.000
15	1.036	4.133	0.000
16	0.978	4.216	0.001
17	0.856	9.640	-0.001
18	1.157	4.416	-0.001

Table 2.5: The simulation-derived parameters A, τ and B for system 4XWF. Pose number 0 corresponds to the native pose, and the following poses correspond to the non-native poses generated by rDock.

Pose Number	**A**	τ	**B**
0	0.528	113.574	0.054
1	0.631	71.400	0.058
2	0.618	117.064	0.004
3	0.568	51.331	0.002
4	0.442	315.734	0.088
5	0.702	17.051	0.007
6	0.398	211.814	0.015
7	0.506	43.541	0.003
8	0.546	16.083	0.001
9	0.536	53.373	0.027
10	0.606	11.598	0.010
11	0.802	11.545	0.003
12	0.862	9.671	0.000
13	0.576	54.592	0.010

Table 2.6: The simulation-derived parameters A, τ and B for system 5KPY. Pose number 0 corresponds to the native pose, and the following poses correspond to the non-native poses generated by rDock.

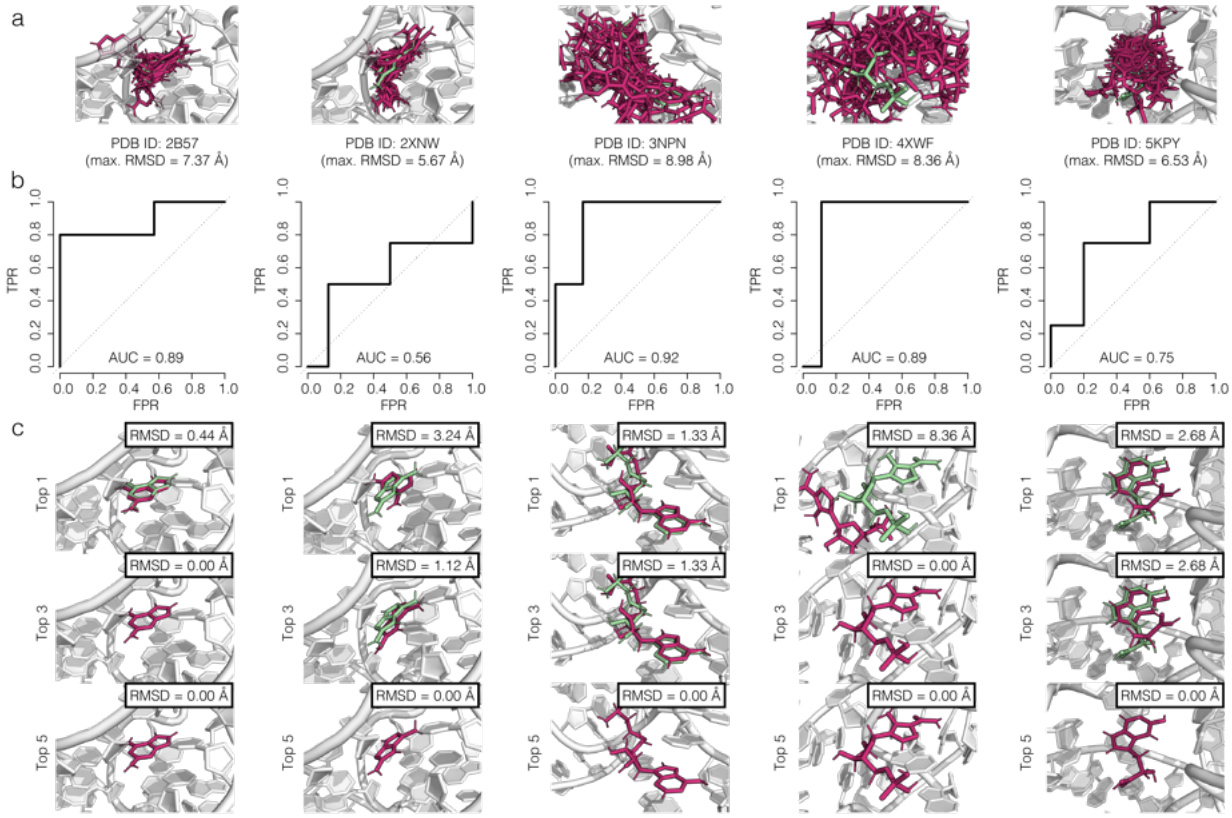


Figure 2.4: (a) Overlay of poses for each of the five RNA-ligand complexes we examined in this study. (b) Corresponding ROC plots for discriminating native-like poses from non-native (decoy) poses based on the characteristic decay times (τ_{ssMD}) estimated from ssMD simulations of individual poses. (c) Comparison between the native (green) and best poses (pink) among the top 1, 3, and 5 poses, when we sort poses by τ_{ssMD}.

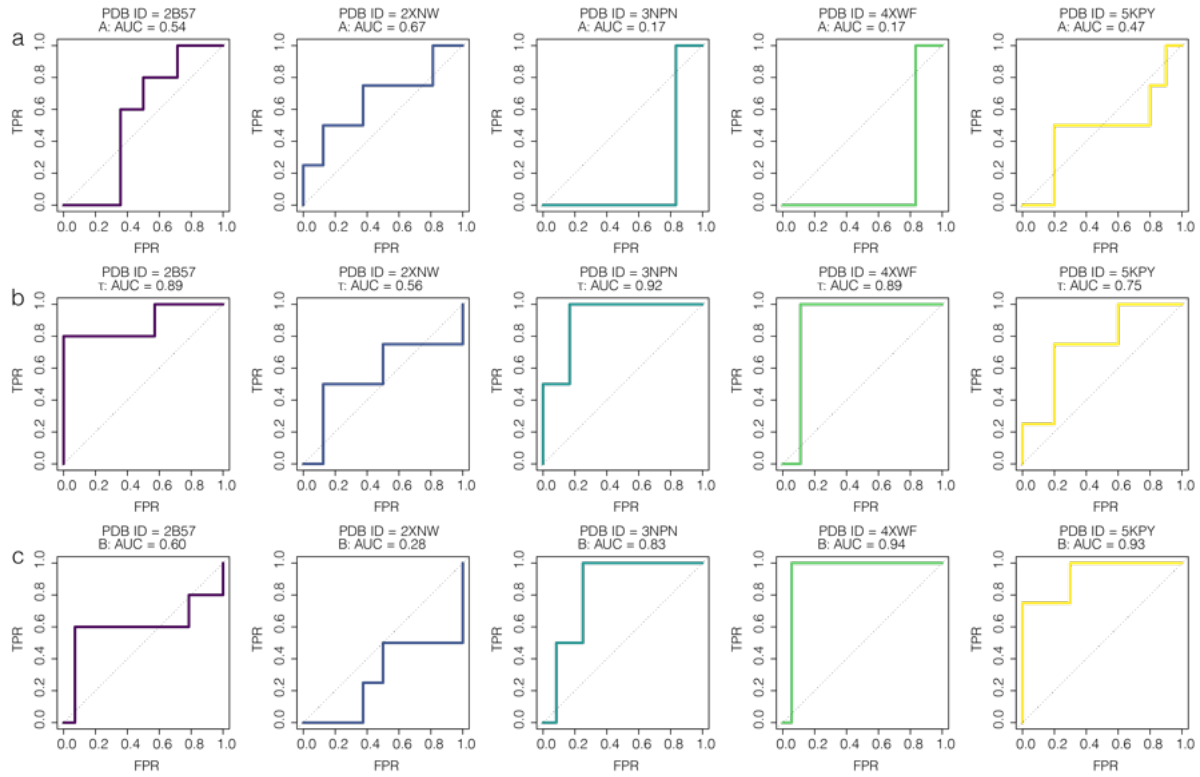


Figure 2.5: ROC plots when using the ssMD-derived exponential fit parameters (a) A, (b) τ, and (c) B of five RNA-ligand complexes, here identified using their PDB IDs.

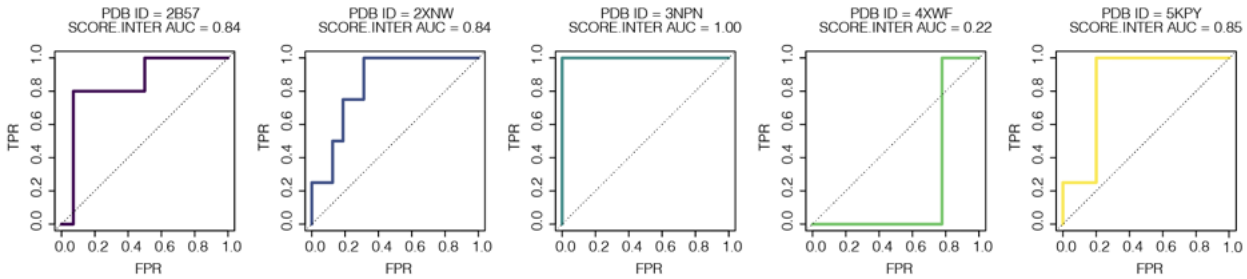


Figure 2.6: ROC plots when ranking poses using the rDock interaction scores of five RNA-ligand complexes, here identified using their PBD IDs.

System	**Description**	N	**AUC**	**min RMSD** (Å)
2B57 (2.15 Å)	Purine riboswitch bound to 2,6-Diaminopurine	19	0.89	0.44/0.00/0.00
2XNW (1.50 Å)	Purine riboswitch bound to triazolo-triazole-diamine	20	0.56	3.24/1.12/0.00
3NPN (2.79 Å)	SAH riboswitch bound to S-adenosylhomocysteine	14	0.92	1.33/1.33/0.00
4XWF (1.80 Å)	ZTP riboswitch bound to ZMP	19	0.89	8.36/0.00/0.00
5KPY (2.00 Å)	5-hydroxytryptophan aptamer bound to 5-hydroxy-L-trypotophan	14	0.75	2.68/2.68/0.00
–	–	17.2	0.80	3.21/1.03/0.00

Table 2.7: Pose-prediction accuracy metrics when using τ estimated from ssMD-derived unbinding profiles to rank RNA-ligand poses. The resolution of each structure is included in the parenthesis next to the PDB ID. N is the number of poses, and the min RMSD listed is the lowest RMSD when considering the top 1/3/5 poses. The average value for N, AUC, and min RMSD are listed at the bottom row.

2.3.2 Structure-based models can robustly predict simulation-derived unbinding profiles

Though based on a small collection of RNA-ligand complexes, the results reported above indicated that estimating the characteristic decay scale of unbinding during enhanced MD simulations like ssMD might be a viable strategy for identifying near-native poses. To validate these initial observations, we could carry out ssMD simulations on additional RNA-ligand systems. Unfortunately, ssMD simulations are computationally demanding. For example, generating 20 ssMD unbinding trajectories used to construct the unbinding profiles for a single pose required 20×9 hours = 180 hours of com using one GeForce RTX 2080 TI GPU and eight cores with the CHARMM/OpenMM interface. Therefore, in lieu of direct validation via additional ssMD simulations, we next attempt to parameterize a structure-based model that we could use to estimate the unbinding parameters directly from RNA-ligand poses.

To explore whether we could predict ssMD-derived unbinding profiles directly from the structure, we trained a structure-based regression model using the partial least squares (PLS) approach [88]. Briefly, for each pose of the five RNA-ligand complexes in our data set, we generated molecular (pose) fingerprints [89] and then paired these fingerprints with the exponential fitting parameters from the unbinding profiles computed from the ssMD simulations initialized from that pose. Individual PLS models were then used to predict each parameter (A, τ and B) from the pose fingerprints.

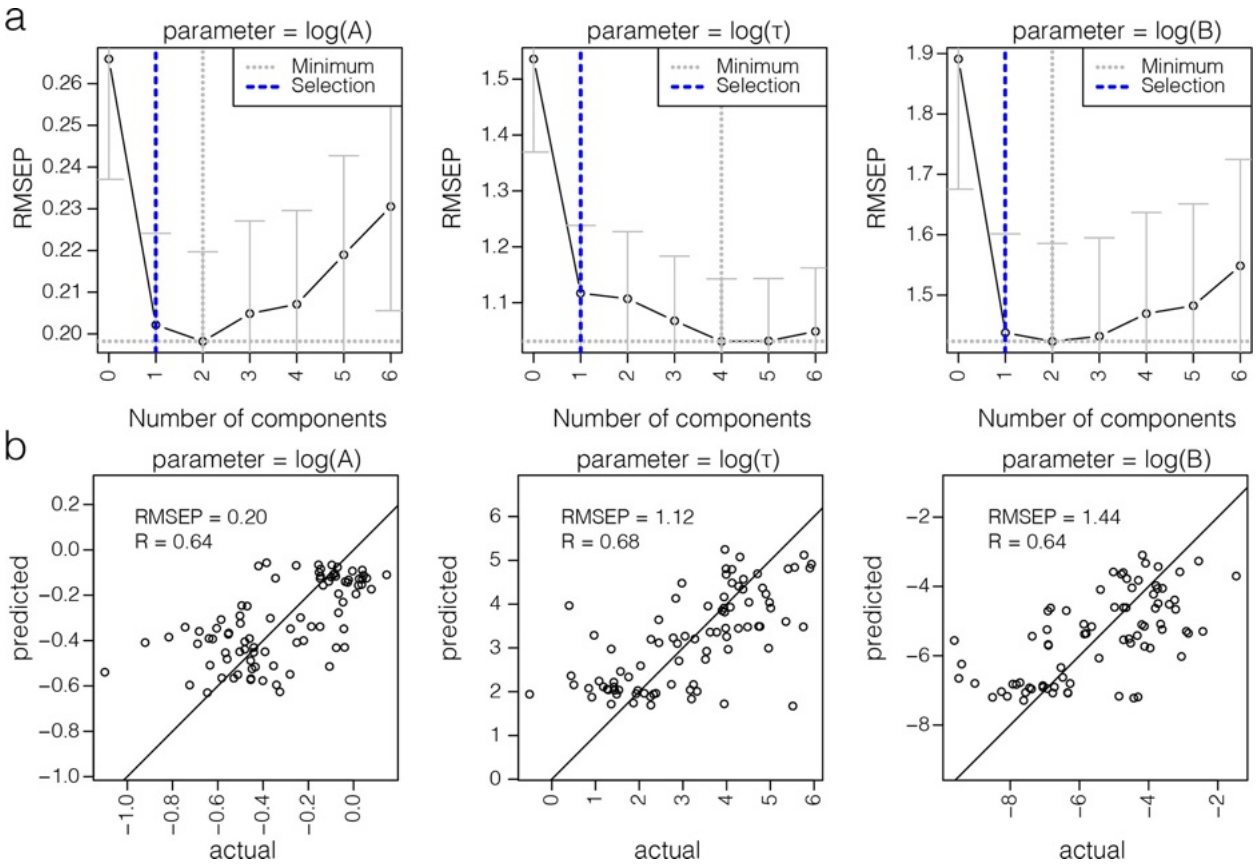


Figure 2.7: Results we obtained when using partial least squared (PLS) to predict parameters extracted from ssMD unbinding profiles (A, τ, and B) directly from the structure of individual poses. (a) Cross-validation RMSEP as function of the number of components in the PLS model used to predict $\log(A)$ (left), $\log(\tau)$ (middle), and $\log(C)$ (right). Gray dotted lines correspond to the absolute minimum RMSEP observed, and blue dotted lines indicate the optimal selected number of components. We chose the latter using the one-sigma rule as the smallest number of components that exhibited an RMSEP within 1σ of the minimum RMSEP. The error bars capture the variance in the estimated RMSEP over the leave-one-out cross-validation experiments. (b) Scatter plots comparing actual (ssMD-derived) $\log(A)$ (left), $\log(\tau)$ (middle), and $\log(C)$ (right) with those predicted using optimal number of components. Results correspond to those from leave-one-out cross-validation experiments. And we have annotated the scatter plots with the RMSEP and correlation coefficient (R).

Shown in Figure 2.7 are cross-validation root mean square error of prediction (RMSEP) for each exponential fit parameter as a function of the number of components included in the predictor. Using the optimal number of PLS components, the cross-validation RMSEP for $\log(A)$, $\log(\tau)$ and $\log(B)$ was 0.20, 1.12, and 1.44, respectively. Also, the correlation coefficient R was 0.64, 0.68, and 0.64, respectively. Thus, we could train reasonably robust models to predict the exponential fitting parameters obtained from the ssMD data directly from the structure of individual RNA-ligand poses.

2.3.3 Structure-derived unbinding profiles help resolve native-like poses from non-native decoys

Since we have shown that τ estimated from the ssMD simulations (τ_{ssMD}) could be used to resolve poses (Figure 2.4), we asked whether we could also identify native or native-like poses from decoys using the τ values estimated by our PLS model (τ_{PLS}).

In Figure 2.8, we show the ROC plots obtained when using τ_{PLS} to rank poses across ten additional RNA-ligand test cases. We note that these complexes and poses were distinct from those used to train the PLS model. With the exception of PDB ID: 4NYA (AUC = 0.36), the AUC values were all greater than 0.75, and the mean value was 0.80 (Table 2.8). An explanation about the test case 4NYA is presented in Figure 2.9. The RMSD of the best poses among the top 1, 3 and 5 poses were typically less than 3.00 Å. Across the ten RNA-ligand complexes, the mean RMSD among the top 1, 3 and 5 was 2.86, 2.66 and 2.63 Å, respectively. Interestingly, when ranking these poses using the rDock RNA-ligand interaction score, we observed lower AUCs and higher minimum RNSDs (Table 2.9). These results demonstrated that the τ_{PLS} computed directly from structure could be used to resolve poses. Moreover, they imply that the structure-kinetic trends observed during our ssMD simulations could be captured by the structure-based model and transferred to similar yet unrelated RNA aptamers to guide pose prediction.

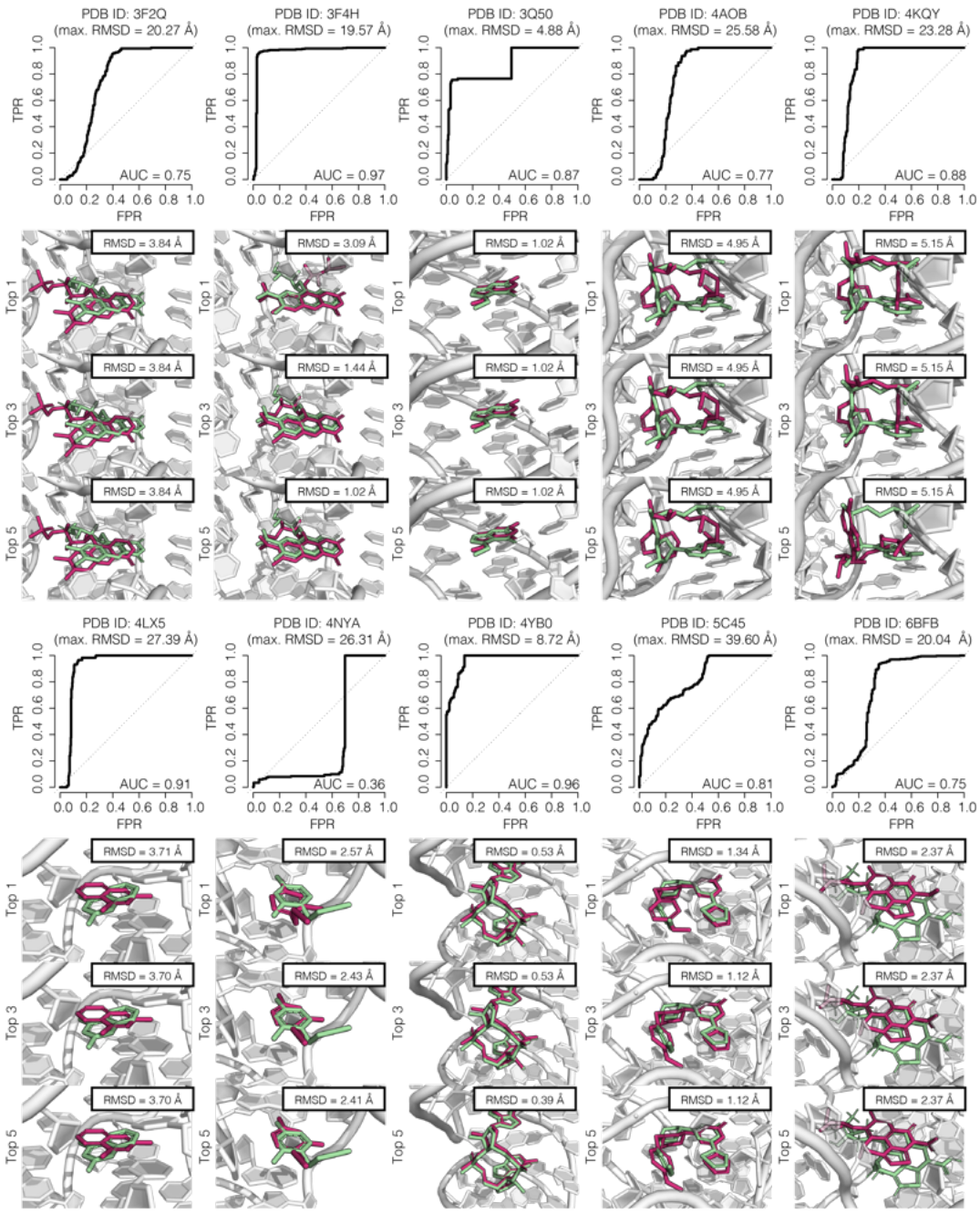


Figure 2.8: ROC plots for discriminating native-like poses from non-native (decoy) poses based on the characteristic decay times predicted directly from the structure of individual poses using our structure-based PLS model (τ_{PLS}). Below each ROC plot, we show comparisons between the native (green) and best poses (pink) among the top 1, 3, and 5 poses, when poses are sorted by τ_{PLS}.

System	Description	N	AUC	min RMSD (Å)
3F2Q (2.95 Å)	FMN riboswitch bound to FMN	550	0.75	3.84/5.50/5.05
3F4H (3.00 Å)	FMN riboswitch bound to Roseoflavin	550	0.97	3.09/2.00/1.77
3Q50 (2.75 Å)	Class I PreQ1 riboswitch to PreQ1	500	0.87	1.02/1.02/1.03
4AOB (2.95 Å)	SAM-I riboswitch bound to SAM	600	0.77	4.95/8.91/7.49
4KQY (3.02 Å)	yitJ S box/SAM-I riboswitch bound to SAM	600	0.88	5.15/5.31/5.38
4LX5 (2.13 Å)	M6" riboswitch bound to PPDA	550	0.91	3.71/3.71/3.71
4NYA (2.65 Å)	thiM riboswitch bound 5-(azidomethyl)-2-methylpyrimidin-4-amine	450	0.36	2.57/2.58/2.56
4YB0 (2.12 Å)	3',3'-cGAMP riboswitch bound with c-di-GMP	350	0.96	0.53/0.54/0.49
5C45 (2.93 Å)	FMN riboswitch bound to Ribocil-B	550	0.81	1.34/1.22/1.31
6BFB (2.82 Å)	FMN riboswitch bound to WG-3	550	0.75	2.37/2.59/2.68
–	–	525	0.80	2.86/2.66/2.64

Table 2.8: Pose-prediction accuracy metrics when using structure-based τ_{PLS} to rank RNA-ligand poses. The resolution of each structure is included in the parenthesis next to the PDB ID. N is the number of poses, and the min RMSD listed is the lowest RMSD when considering the top 1/3/5 poses.

System	Description	N	AUC	min RMSD (Å)
3F2Q (2.95 Å)	FMN riboswitch bound to FMN	550	0.95	2.44/2.44/1.27
3F4H (3.00 Å)	FMN riboswitch bound to Roseoflavin	550	0.94	1.61/1.61/1.58
3Q50 (2.75 Å)	Class I PreQ1 riboswitch to PreQ1	500	0.36	4.58/4.58/4.58
4AOB (2.95 Å)	SAM-I riboswitch bound to SAM	600	0.63	5.61/5.60/5.60
4KQY (3.02 Å)	yitJ S box/SAM-I riboswitch bound to SAM	600	0.54	17.54/6.16/6.16
4LX5 (2.13 Å)	M6" riboswitch bound to PPDA	550	0.41	3.07/3.03/3.02
4NYA (2.65 Å)	thiM riboswitch bound 5-(azidomethyl)-2-methylpyrimidin-4-amine	450	0.32	2.81/2.81/2.81
4YB0 (2.12 Å)	3',3'-cGAMP riboswitch bound with c-di-GMP	350	0.98	0.45/0.43/0.43
5C45 (2.93 Å)	FMN riboswitch bound to Ribocil-B	550	0.90	5.72/5.67/5.57
6BFB (2.82 Å)	FMN riboswitch bound to WG-3	550	0.81	8.38/1.90/1.90
–	–	525	0.68	5.22/3.42/3.29

Table 2.9: Pose-prediction accuracy metrics when using the rDock interaction scores to rank RNA-ligand poses. The resolution of each structure is included in the parenthesis next to the PDB ID. N is the number of poses, and the min RMSD listed is the lowest RMSD when considering the top 1/3/5 poses.

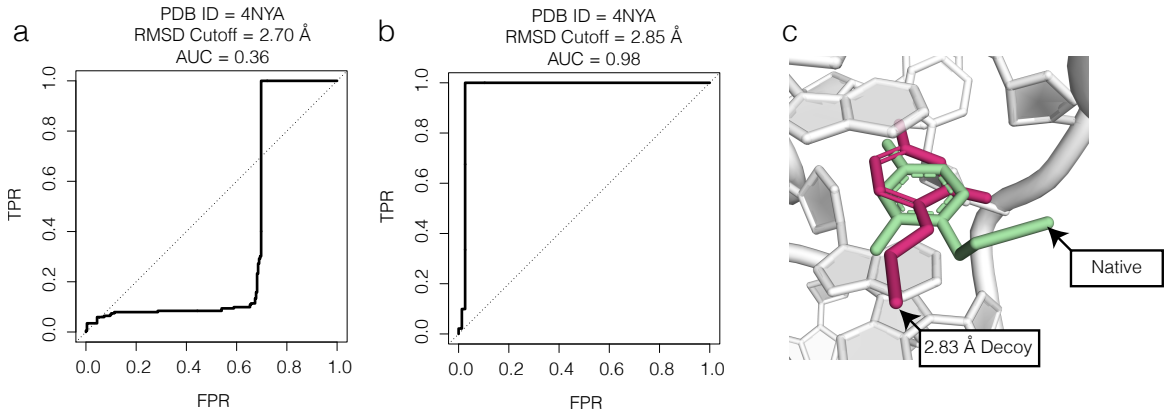


Figure 2.9: ROC plots for the 4NYA test case when using a 2.70 Å (a) and 2.85 Å (b) RMSD cutoff to distinguish native from non-native (decoy) poses. Shown in (c) is a closeup view of the native pose and a pose with 2.83 Å RMSD. With 2.70 Å RMSD cutoff, such poses would be considered non-native though their location overlaps with the native pose, albeit in a flipped conformation. The featurization approach we used is likely insensitive to such subtle differences thus yielding similar estimates for τ as poses with RMSD less than $\sim$2.80 Å. This, in part, explains the low AUC observed when using a 2.70 Å RMSD cutoff as such poses would be considered non-native.

2.4 Conclusion

We have showed that contact analysis of RNA-ligand unbinding trajectories from ssMD simulations could help identify native-like poses from non-native poses. This method is based on the comparison of the unbinding profiles of each individual poses. When the unbinding profiles were fitted into single-exponential decaying functions, the unbinding profiles for near-native poses tend to have longer characteristic decay times (τ) compared to non-native poses. Therefore, τ could be used to identify native-like poses from decoys. Based on these observations, we then showed that a simple PLS regression model trained with the ssMD-derived parameters and the molecular fingerprints of the RNA-ligand complexes could rapidly reproduce the properties of the unbinding profiles directly from a distance-based fingerprint of a given pose. Though we found that the unbinding parameters estimated from either ssMD simulations or a PLS regression model are useful in identifying native-like RNA-ligand poses, there are a several caveats that must be considered.

First, for the initial ssMD analysis, 20 independent ssMD simulations for each RNA-ligand complexes were carried out. Even with a small set of five RNA-ligand complexes with no more than 20 poses for each systems, the total number of simulations amounted to 1720, each 10 ns and requiring around 15480 hours of computing. Since our initial ssMD analysis was limited to five RNA-ligand complexes, we cannot rule out that analysis on a larger data set may yield different results. However, the fact that we could build a PLS regression model that also showed similar resolving power on a distinct collection of poses suggests that the trend we observed in the ssMD analysis is robust.

Second, though PLS models can accurately extracting trends with fewer samples and many features, it is likely that increasing the size of the training set will give a better performing and more generalized model. In this regards, we are currently expanding out database with ssMD trajectories of diverse RNA-ligand poses.

Third, our analysis was restricted to ligand sensing riboswitches because they represent a subset of RNAs that are highly structured and for which X-ray crystal structures are

available. Though riboswitches can be important antibacterial drug target, there is an emerging interest in other RNAs such as human mRNAs and miRNAs as therapeutic targets. However, little to no experimental data are available for such RNAs interacting with drug-like molecules. Thus, it remains an open question whether the results generalized to other RNA-ligand systems.

Fourth, although the τs predicted from the PLS model showed strong resolving power when ranking poses, the absolute τ_{PLS} values had a much narrower range compared to the τs estimated from the ssMD trajectories that used to train the model. For example, the mean τ_{ssMD} for the native-like poses for the five training samples was 153.8 $\times$ 101 ps, whereas the mean τ_{PLS} was only 70.5 $\times$ 101 ps. And across the 10 complex systems used to test the PLS model, the τ_{PLS} only ranged between 9.0 $\times$101 ps and 18.0 $\times$101 ps, with a mean of 11.1 $\times$ 101 ps. Therefore, though the relative values τ_{PLS} retained the resolving power of τ_{ssMD}, their absolute values may be less meaningful.

Despite these limitations, the results indicate that characterizing ligand unbinding profiles from simulations is a viable way for identifying native-like poses. Moreover, by training regression models that learn from these simulations, one might be able to identify the native pose while bypassing the computationally expensive simulations. We envision that such regression models will be useful in docking studies of RNA-ligand complexes. Within the context of such studies, consensus scores can be generated by combining the τs predicted using structure-based regression models with scores from other pose prediction strategies[89, 90, 91]. One could then use these consensus scores to identify high-confidence poses that can be used as input for binding energies estimation and other structure-based virtual screening tasks.

CHAPTER 3

Pre-Equilibrium Kinetics May Underlie the Differential Activity of Selective Binders of the FMN-dependent Riboswitch

Elizabeth Tidwell, Markos Koutmos, and Aaron T. Frank conceived project, Indrajit Deb and Yichen Liu carried out simulations, Indrajit Deb, Yichen Liu, and Aaron T. Frank analyzed the results.

Abstract

In the previous chapter, selectively scaled molecular dynamics (MD) simulations were used to assist the prediction of native poses for RNA aptamer-ligand complexes. In this chapter, the application of selectively scaled MD simulations to the study of unbinding kinetics will be discussed. In particular, we focused on the flavin mononucleotide (FMN)-dependent riboswitch, which selectively binds FMN and regulates the biosynthesis of FMN. Two synthetic small molecules, namely, ribocil-B and WG-3, have been discovered to bind to the aptamer domain of the FMN-dependent riboswitch, but show different ability to inhibit bacterial growth through transcription termination. Ribocil-B can bind to the FMN riboswitch

and terminates transcription, while WG-3 does not terminate transcription. The underlying mechanism of the different activity is not well understood. Here we apply selectively scaled MD (ssMD) simulations and kinetic modeling to elucidate the differential activity of the ligands. Our simulations and kinetic modeling predict that, compared to ribocil-B, WG-3 quickly unbinds from the aptamer domain. The unbinding kinetics indicates that on the transcriptional timescale, WG-3 is unable to stabilize the FMN riboswitch into a conformational state required for transcription termination, which may explain why WG-3 is unable to inhibit bacterial growth. The dominant unbinding pathways of WG-3 and ribocil-B are distinct, which is likely due to the solvent accessibility of the ligands. This work suggests that the characterization of ligand unbinding kinetics is a reliable way for selecting active binders of kinetically controlled riboswitches.

3.1 Introduction

Riboswitches are structured elements located in the 5'-untranslated region (UTR) of bacterial mRNAs, regulating the expression of downstream gene in response to the binding of ligand [92]. Riboswitches typically regulate genes involved in the biosynthesis or export of their cognate ligand, which is essential for maintaining cellular homeostasis. Riboswitches contain an aptamer domain that selectively binds to its cognate ligand and an expression platform that undergoes conformational changes upon ligand binding. Since riboswitches are found predominantly in bacteria and control the expression of genes that are essential for proper cellular function, they are considered promising antibacterial drug targets [34, 37].

Here we focused on the flavin mononucleotide (FMN)-dependent riboswitch, which controls the FMN biosynthesis pathway [93]. This riboswitch is an OFF-switch, and the binding of FMN induces the termination of the downstream gene, resulting in a reduced level of FMN biosynthesis. FMN is an essential metabolite and disruption of the synthesis of FMN can

inhibit bacterial growth. Therefore, the FMN riboswitch can be used as a model system to study the druggability of structured RNA. Ribocil-B and WG-3 were identified as small molecules that binds to the FMN riboswitch, and they occupy the same binding pocket as the cognate ligand FMN [68]. WG-3 stabilizes a conformation that is globally similar to the ribocil-B riboswitch conformation, with RMSD = 0.63 Å. Even though the two ligands stabilize similar conformations, the binding of ribocil-B can terminate transcription and inhibit bacterial growth, while WG-3 is not able to inhibit bacterial growth [94].

In this work, we attempted to rationalize the difference in activity between ribocil-B and WG-3 by examining whether the repositioning of A48 affects the unbinding of ribocil-B and WG-3 from the aptamer domain of the FMN riboswitch. We first applied selectively scaled MD simulations (ssMD) [59] to accelerate the unbinding of the ligands (ribocil-B, WG-3, FMN) from the FMN riboswitch. Then we applied kinetic modeling to estimate the off-rates (k_{off}) at the unscaled potential. We focused on the off-rates because the FMN riboswitch is kinetically controlled [35], which indicates that it operates under non-equilibrium conditions. Although the equilibrium dissociation constant K_{D} is directly related to the binding affinity and serve as an important parameter to study the ligand-receptor interaction, equilibrium is rarely reached in vivo. Therefore, the equilibrium dissociation constants of the ligands could not explain their difference in activity. The estimated off-rates suggested that the WG-3 unbinds more rapidly than ribocil-B, which may explain why WG-3 is unable to inhibit bacterial growth. We also explored the dominant release pathways of ribocil-B and WG-3 and the solvent accessibilities of the ligands in the vicinity of the exit channel.

3.2 Methods

3.2.1 Selectively scaled MD simulations

The unbinding trajectories were generated by Indrajit Deb.

Selectively scaled molecular dynamics (ssMD) simulations were conducted to study the unbinding of the ligands ribocil-B, WG-3 and FMN from the aptamer domain of the FMN riboswitch [59] The starting structures of the FMN riboswitch in complex with ribocil-B, WG-3 and FMN were obtained from PDB using accession code 5C45 [68], 6BFB [94], and 2YIE [95], respectively. The simulations were carried out using the CHARMM simulation package [83], and the input files were generated using the CHARMM-GUI webserver [96] The CHARMM36 nucleic acid force field [97] was used to simulate RNA and the CHARMM general force field (CGenFF) was used for ligand [84] The RNA-ligand complexes were solvated in a cubic TIP3 water box [98] The minimum distance between the solute and the edge of the box was set to 10 Å. Potassium ions were added to neutralize the system. Each system was energy minimized by 200 steps of steepest descent (SD) followed by 500 steps of adopted basis Newton-Raphson (ABNR) algorithms. Then the systems were equilibrated in NVT conditions for 2 ns with a timestep of 1 fs. The nucleobase heavy atoms were harmonically restrained with a force constant of 0.1 kcal/mol/Å^2, and the heavy atoms of the ligand and the RNA backbones were harmonically restrained with a force constant of 1.0 kcal/mol/Å^2. The equilibrated structures were used as the starting structures for production simulations. To accelerate the unbinding, the Lennard-Jones potentials associated with all ligand-water pairs were selectively scaled with the scaling factor, α (where $\alpha = 1/\alpha'_{ij}$, and α'_{ij} were set to 3.25, 3.50 and 3.75. respectively) [59]. The well-depths (ϵ_{ij}) of the Lennard-Jones potential for all water-ligand atom type pairs were scaled by α'_{ij} using the NBFIX facility in CHARMM.

$$V_{ij} = \epsilon_{ij} \times \alpha'_{ij} \left[\left(\frac{r_{ij,\min}}{r_{ij}} \right)^{12} - \left(\frac{r_{ij,\min}}{r_{ij}} \right)^{6} \right] \tag{3.1}$$

All energy minimizations, equilibrations and production run simulations were carried out using graphical processing unit (GPU) through CHARMM/OpenMM interface [86]. Periodic boundary conditions (PBC) were applied during all the simulations. Non-bounded interactions were calculated using 16.0 Å cutoff and long range electrostatics were treated using the particle mesh Ewald (PME) method [99]. The SHAKE algorithm was used to constrain

hydrogen atoms for both equilibration and production simulations [87]. For production simulations were carried out in NPT ensembles, Langevin dynamics were propagated at 303.15 K with a timestep of 2 fs when Monte Carlo (MD) barostate was turned on at a reference pressure of 1 atm and sampled at a frequency of 15. The Langevin friction coefficient was set to 10 ps^{-1} for both equilibration and production simulations. For each of the system, 50 independent 50 ns long trajectories were generated.

3.2.2 Contact Analysis

The trajectories were processed by contact analysis. The contacts between the heavy atoms of ligand and RNA were calculated using the contact tool from MoleTools toolset https://github.com/atfrank/MoleTools. For the three ligands, the time-dependent unbinding profiles were determined from contact data using in-R script. The equilibrated structures were used to generate reference contacts. These contacts were defined as the contacts that present in the equilibrated structure between the ligand and the RNA. If a pair of atoms in RNA and ligand were within 5 Å, they were considered as in contact. If the minimum distance among the native contacts was greater than 10 Å, the ligand was considered as unbound. For each system, 50 time-dependent unbinding profiles were generated for each of the trajectories, representing the bound or unbound state of the ligands as a function of simulation time.

3.2.3 Kinetic Models

The maximum unbinding rate were computed to estimate the off-rates ($k_{\text{off}}(\alpha)$) of WG-3, ribocil-B, and FMN. The maximum unbinding rate is defined as the maximum average number of unbinding trajectories per unit time [100]. The averaging was carried out over 50 ssMD trajectories and using the time-dependent unbinding profiles derived from the contact

analysis.

For WG-3 and ribocil-B, the $k_{\text{off,rel}}(\alpha)$ relative to that of FMN were computed (where $\alpha = 1/\alpha'_{ij}$, and α'_{ij} were set to 3.25, 3.50 and 3.75, respectively). We then used the Kramers'-based rate extrapolation (KRE) [59] to estimate the $\log(k_{\text{off,rel}})(\alpha = 1)$ by performing linear extrapolations on the logarithm(10) scale to estimate the $\log(k_{\text{off,rel}})(\alpha = 1)$, which gives the rate on the unscaled potential. For FMN and ribocil-B, the experimentally determined koff is known ($k_{\text{off,FMN}} = 1.5 \times 10^{-4}\text{s}^{-1}$ and $k_{\text{off,ribocil-B}} = 4.4 \times 10^{-3}\text{s}^{-1}$, respectively)[68], we determined the correlation factor of ∼0.54 from $\log(k_{\text{off,ribocil-B}}/k_{\text{off,FMN}})$ divided by the estimated $\log(k_{\text{off,rel,ribocil-B}})(\alpha = 1))$. This correlation factor was further used to correct the estimated $\log(k_{\text{off,rel,WG-3}})(\alpha = 1)$ and bring it to the experimental scale, which is our estimated $k_{\text{off,WG-3}}$. Then we constructed a simplified kinetic model for WG-3 using the experimentally measured dissociation constant ($K_{\text{D,WG-3}} = 130$ nM) [94] and the estimated $k_{\text{off,WG-3}}$. Using the experimental dissociation constant and the estimated off-rate, the $k_{\text{on,WG-3}}$ was calculated as $k_{\text{off,WG-3}}/K_{\text{D,WG-3}}$. For ribocil-B, the kinetic model was built based on the experimentally measured $k_{\text{on,ribocil-B}}$ (7.5×10^5 $\text{M}^{-1}\text{s}^{-1}$) and $k_{\text{off,ribocil-B}}$ ($4.4 \times 10^{-3}\text{s}^{-1}$), respectively [68]. The kinetic models were used to simulate the pre-equilibrium unbinding of WG-3 and ribocil-B. The numeric solutions were carried out in R using the differential equation solver, deSolve [101].

3.2.4 Selection of representative trajectories

In addition to recovering the kinetics of ligand unbinding, it is also crucial to compare the unbinding pathways of the ligands to better understand their difference in activity. The first step in the investigation of the unbinding pathway is to select the trajectories that represent the overall trend of unbinding. The contacts data were used to select the representative trajectory. For each system, the average contacts between the ligand atom and RNA were calculated across the 50 replicas. For each individual trajectory, the difference between the

distance for each atom pair and the average is calculated and summed. For each system, the trajectory with the smallest sum of differences compared with the average is selected as the representative trajectory. The calculations involved in the selection of representative trajectories were conducted in R.

3.3 Results and Discussion

3.3.1 WG-3 rapidly unbinds from the FNM aptamer compared to ribocil-B

To understand the different ability of WG-3 and ribocil-B to inhibit bacterial growth, we sought to build kinetic models describing the binding of both WG-3 and ribocil-B to the aptamer of FMN riboswitch. The experimentally measured K_{D} for WG-3 and ribocil-B are 130 nM and 6.6 nM [68, 94]. However, since the FMN riboswitch is kinetically controlled, the K_{D} values do not completely explain the differential abilities of WG-3 and ribocil-B to inhibit bacterial growth. The higher K_{D} of WG-3 relative to that of ribocil-B could be due to either a lower on-rate ($k_{\mathrm{on,WG-3}}$) or a higher off-rate ($k_{\mathrm{off,WG-3}}$), and both could impact the ability of WG-3 to trigger the ligand-dependent switching of the FMN riboswitch. Although $k_{\mathrm{on,ribocil-B}}$ and $k_{\mathrm{off,ribocil-B}}$ measurements have been reported for the ribocil-B binding to the aptamer domain, corresponding values are not available for WG-3.

First, we built a kinetic model for the binding/unbinding of WG-3 to/from the aptamer of FMN riboswitch, and compared it to that for ribocil-B. For WG-3 the kinetic model was constructed using the experimentally measured K_{D} and the estimated $k_{\mathrm{off,WG-3}}$ from ssMD simulations using the Kramer's-based rate extrapolation (KRE) approach (Figure 3.1A). Then we estimated $k_{\mathrm{on,WG-3}}$ ($k_{\mathrm{on,WG-3}} = k_{\mathrm{off,WG-3}}/K_{\mathrm{D,WG-3}}$) (Figure 3.1B). Using selectively scaled MD simulations to accelerate rare unbinding events, we were able to observe unbinding in as little as 10 ns of simulated time.

For ribocil-B, the kinetic model was constructed using experimentally measured $k_{\text{on,ribocil-B}}$ and $k_{\text{off,ribocil-B}}$ [94]. The kinetic models were then used to explore the pre-equilibrium kinetics of WG-3 and ribocil-B on the time scale of riboswitch transcription (Figure 3.1C). For the FMN riboswitch from the ribDEATH operon of Bacillus subtilis, the elapsed time between transcription of the aptamer domain and the formation of the terminator helix is around 70 s [35], we simulate ligand unbinding from the kinetic models over an elapsed time of 100 s. The kinetic simulations reveal that during Δt = 70 s, the fraction of the aptamer that bound to WG-3 rapidly decay to < 0.10. In contrast, the fraction is $>$ 0.60 for ribocil-B during the same period. WG-3 has a high off rate ($k_{\text{off,WG-3}} = 4.7 \times 10^{-2}$ s^{-1}) compared to ribocil-B, $k_{\text{off,ribocil-B}} = 4.4 \times 10^{-3}$ s^{-1} the maximum estimated off-rate constants for the ligands are shown in Figure 3.2, where WG-3 rapidly unbinds compared to ribocil-B and the cognate ligand FMN. This off-rate relative to the corresponding on-rate provides a possible explanation of why ribocil-B is active in inhibiting bacterial growth while WG-3 is inactive. WG-3 is a less effective "molecular fuse" compared to ribocil-B on the time scale of RNA folding,, as it is unable to lock the FMN aptamer into its *holo* conformation for transcription termination.

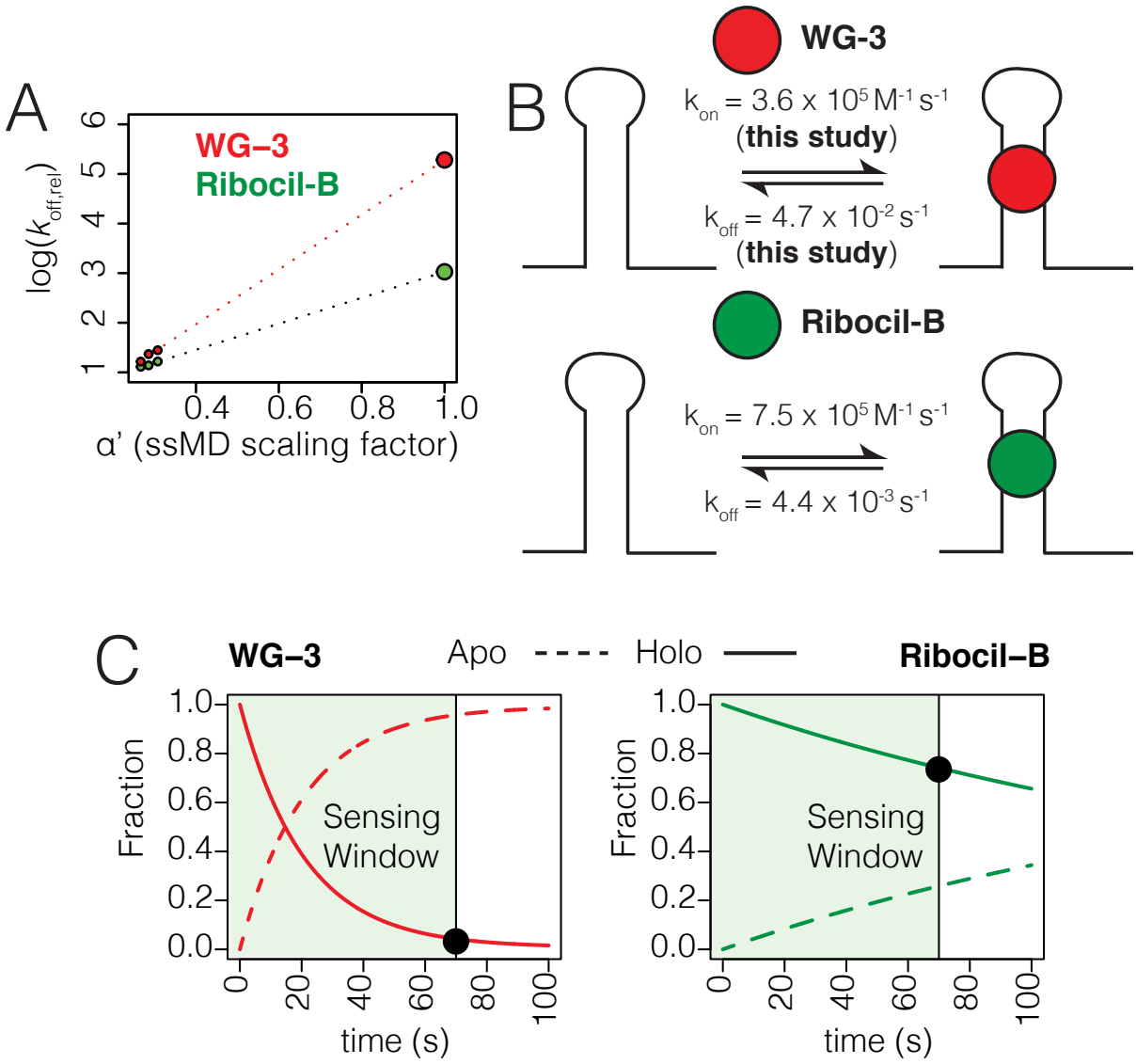


Figure 3.1: (A) ssMD-derived relative estimates of the off-rate ($k_{\text{off}}(\alpha)$) for WG-3 and ribocil-B as a function of scaling factor, α (where $\alpha = 1/\alpha'_{ij}$ and α'_{ij} set to 3.25, 3.50, and 3.75), used during the simulations. Results at unscaled potential (i.e., α=1 or equivalently $\alpha' = 1$) were obtained by extrapolating the $k_{\text{off}}(\alpha)$ derived from short ssMD simulations. (B) Kinetic models for binding/unbinding of WG-3 (top panel) and ribocil-B (bottom panel) to and from the FMN-aptamer domain. For WG-3, k_{on} and k_{off} were derived from simulations; k_{on} was computed as $k_{\text{on}} = k_{\text{off}}/K_{\text{D}}$ using experimental K_{D} and k_{off} estimated from our simulations via extrapolation shown in (A). For ribocil-B, k_{on} and k_{off} correspond to previously reported experimental values. (C) Pre-equilibrium dynamics of WG-3 (left) and ribocil-B (right), simulated using the kinetics models that are shown in (B). The Sensing Window is the time elapsed between the aptamer domain's transcription and the terminator helix's formation.

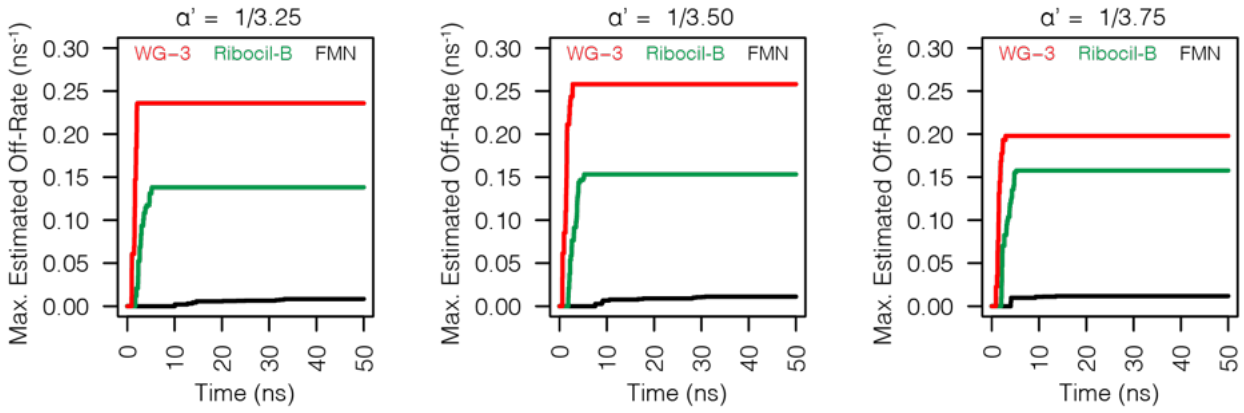


Figure 3.2: The maximum off-rate constants lead to specific time courses for WG-3, ribocil-B and FMN at the three scaling values.

3.3.2 WG-3 and ribocil-B exhibit distinct ssMD escape paths

To examine the possible reasons for the differences in off-rate constants (k_{off}) and on-rate(k_{on}) of WG-3 and ribocil-B, we analyzed the release pathways sampled during the ssMD simulations. Overall, we observed two dominant release pathways which we refer to as pathway I (Figure 3.3A) and pathway II (Figure 3.3B). Pathway I was found to be the dominant release pathway for WG-3 and pathway II was the dominant release pathway for ribocil-B. We inspected the initial structures of the ssMD simulations (equilibrated structures) to rationalize the different release pathways. For the WG-3 system (Figure 3.3A), the solvent density is higher in exit channel I relative to exit channel II, with exit channel II obstructed by residue A48. In contrast, for ribocil-B (Figure 3.3B), the solvent density in the exit channel I was lower than that near exit channel II, and exit channel I was instead obstructed by A48. The observations suggest that the residue A48 serves as a 'gate' that simultaneously controls solvent access to the binding pocket via exit channels I and II, which

determines the release pathways for WG-3 and ribocil. The WG-3 and ribocil-B are blocked at different sides of the residue A48, which results in different solvent accessibility at the two exit channels. Since the dissociation of ligands is closely related to the interaction with the solvent, their different release pathways could be explained by their positions relative to residue A48.

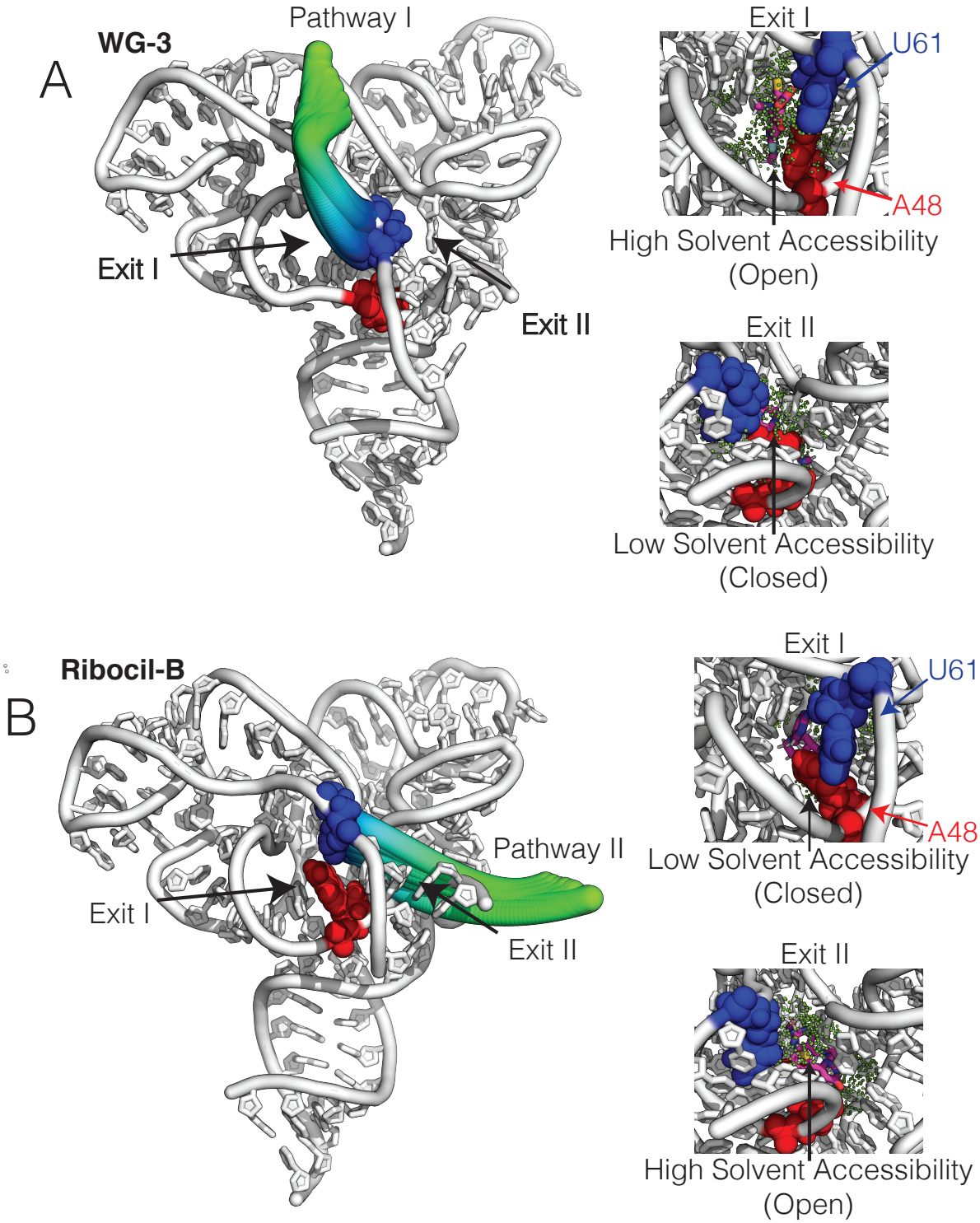


Figure 3.3: Representative release pathways for (A) ribocil-B and (B) WG-3 from the aptamer of the FMN-dependent riboswitch. A48 and U61 are shown as red and blue space filling models, respectively.

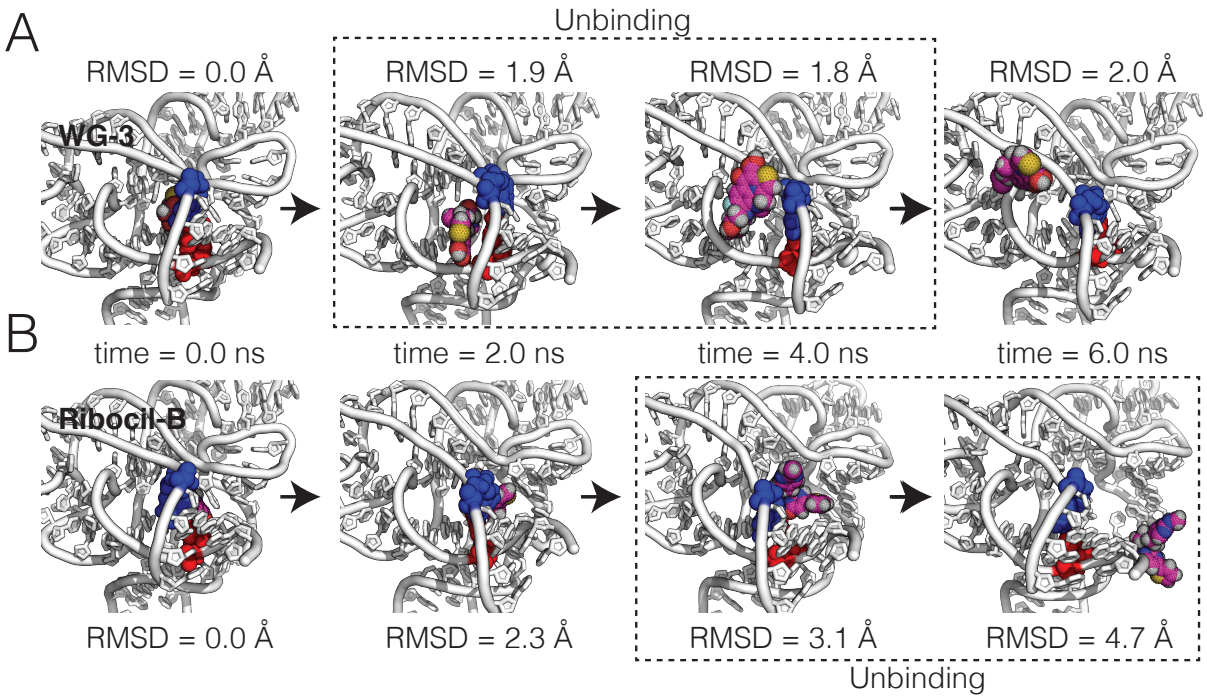


Figure 3.4: Representative unbinding trajectories. Snapshots along the representative unbinding trajectories of (A) WG-3 and (B) ribocil-B. Highlighted in the dashed box is the interval during which unbinding occurs. Indicated below each snapshot is the RMSD relative to the initial structure along the unbinding trajectories.

3.3.3 Unbinding of ribocil-B induces larger scale conformational changes compared to that of WG-3

To gain deeper insights into the molecular events that are likely to accompany the release of WG-3 and ribocil-B, the root-mean-squared deviations (RMSDs) relative to the starting structure was computed using the representative unbinding trajectories for both ribocil-B and WG-3 (Figure 3.4). For both systems, we observed an increase in the RMSD of the aptamer along the representative trajectories. Interestingly, the increase in RMSD for ribocil-B is larger than that for WG-3, indicating that the conformational change accompanying the unbinding of ribocil-B was larger than that associated with the unbinding of WG-3.

3.4 Conclusions

Several bacterial genes are regulated by riboswitches through the binding of metabolites and their analogs. Since riboswitches regulate the genes related to essential metabolites and are not found in humans, they are promising targets for developing antibacterials [36, 37]. Designing molecules that target such riboswitches relies on the understanding of the molecular recognition characteristics, and an understanding of the interplay between the activity of riboswitches and their binding affinity and binding kinetics.

In this study, all-atom explicit-solvent molecular dynamics simulations were employed to study the unbinding of two selective binders of the FMN riboswitch, ribocil-B and WG-3. Two dominant release pathways were identified, characterized by escape through two distinct exit channels in the aptamer domain of the FMN riboswitch. The two exit channels are gated by a single residue, A48, whose position dictates which channel is open and which escape path is more likely. The inspection of the 50 independent trajectories shows that the majority of the trajectories for each system escape through the exit found for the representative trajectory. For ribocil-B, the ligand exits through Exit I in only two trajectories. For WG-3, only 5 trajectories exit through exit II. Analysis of the representative unbinding trajectories suggests that the aptamer domain undergoes a larger conformational change during the release of ribocil-B, explaining the faster off-rate for WG-3 ($k_{\mathrm{off,WG-3}}$) compared to that of ribocil-B($k_{\mathrm{off,ribocil-B}}$). The rapid unbinding of WG-3 relative to ribocil-B may explain why WG-3 is inactive as an antibacterial. That is, Even though WG-3 can bind to the aptamer domain of FMN, by the time the expression platform is transcribed the ligand has already dissociated from the aptamer domain so the partitioning between anti-terminator/terminator proceeds as if no ligand was present.

The results presented here suggest that the binding kinetics, in particular the dissociation rate or residence time relative to the duration of the sensing window, which is defined as the time elapse between transcription of the aptamer domain and the formation of the terminator helix, may determine whether a given ligand can stabilize the aptamer in its holo

conformation long enough to terminate transcription or translation. Therefore, for kinetically controlled riboswitches that operate under non-equilibrium conditions, the relative unbinding kinetics can be used to as an important parameter for the design of small molecule probes that target those riboswitches. Enhanced simulations such as the ssMD method used in this study can be applied to rapidly explore the relative unbinding kinetics of a series of compounds, and the contact data derived from simulations can be used to guide the selection of hits with lower off rates on their target riboswitches. Although this work provides a possible explanation for the differential activitiy of the binders of FMN riboswitch, there are several limitations that should be considered. First, in our kinetic model, the off-rates at unscaled potential were obtained through extrapolation using the simulation-derived off-rates at only three scaling factors. Since there are only three scaling factors and those factors span a relatively small range, the errors associated with this estimation should be considered. To obtain better estimation of the unscaled off-rates, ssMD simulations should be conducted using more scaling factors, ideally span a wider range. Second, the FMN riboswitch is the only system studied in this work, to test whether the framework could be applied to other systems, the workflow should be tested on more kinetically controlled riboswitch systems and their corresponding binders. Third, to better test our methods in simulating kinetics, binding simulations of longer duration could be tested to see if the simulation derived on-rate match the experimental values.

CHAPTER 4

Structural Basis of Partitioning Between Termination and Anti-termination of the Fluoride Riboswitch

The content presented in this chapter was acquired in close collaboration with Emily Ellinger and Adrien Chauvier from the Walter Lab and Jason Porta from the Ohi Lab. The cryo-EM density maps and initial structures were provided by Jason Porta, biochemical assays were carried out by Emily Ellinger and Adrien Chauvier. Yichen Liu carried out molecular dynamics flexible fitting simulations. Yichen Liu and Emily Ellinger carried out analysis.

Abstract

In bacteria, riboswitches function to regulate the expression of a significant number of genes by controlling early events in transcription. Investigating the communication between a nascent riboswitch, embedded at the 5'end of a messenger RNA (mRNA), and the transcribing RNA polymerase (RNAP) is therefore crucial for understanding the mechanism of regulation. An important process in transcription is transcriptional pausing, which is known

to typically be initiated and stabilized by a consensus DNA pause sequence. Unlike other riboswitches, the fluoride-sensing riboswitch is not followed by a consensus pause sequence, yet stabilizes pausing and favors termination in the absence of its ligand, the fluoride ion, while mediating pause release and anti-termination in the presence of ligand. In this study, we combined biochemical assays, cryo-electron microscopy, and molecular dynamics flexible fitting to study the coupling between the *Bacillus cereus* fluoride-sensing *crcB* riboswitch and transcription termination of bacterial RNAP. In our structures, we observed that ligand binding initiates a cascade of events leading to global conformational changes of RNAP, which promotes the continuation of transcription. Our work reveals a mechanism for RNA dependent transcription pausing and release.

4.1 Introduction

Riboswitches are structured regulatory elements located at the 5'-untranslated region (5'-UTR) of mRNA, regulating gene expression in bacteria to maintain homeostasis, which is essential for the proper function of cells. Riboswitches selectively bind a ligand that then triggers a conformational change of the expression platform, either promoting or terminating the expression of downstream genes [28, 30, 102, 103, 104]. Co-transcriptional folding of riboswitches is crucial for their function, as they have to bind their ligand and operate during synthesis by RNA polymerase (RNAP)[105].

In *E. coli* and *B. subtilis*, a common 16-nucleotide sequence was discovered to establish stabilizing interactions with the RNAP active site, which prevents further nucleotide addition [106]. Transcriptional pauses play essential roles such as aiding in transcription or translation factor recruitment, ensuring the proper folding of RNA, and coordinating transcription-translation coupling [107, 108, 109, 110]. The pausing behavior can be further modulated by RNA structures and regulatory factors. For example, the *his* pause involves the formation of a hairpin structure in the exit channel that prolongs pausing, which is

further stabilized upon the binding of NusA [108, 111]. Previous studies have revealed a coupling mechanism between the pre-queuosine (preQ1) riboswitch and bacterial RNAP [112, 113] where the pseudoknot structured riboswitch enhances pausing at a transcriptional consensus sequence in the absence of its ligand, preQ1, through interactions with the RNAP within the exit channel, and stabilizes a global inactive state of the RNAP [113]. However, this consensus sequence does not exist for other riboswitches such as the fluoride riboswitch, implying a solely RNA-driven mechanism for gene regulation. In the fluoride-free state, the formation of a terminator hairpin in the expression platform is favored, which triggers transcription termination. Upon binding to the fluoride ion, the terminator hairpin is disrupted and transcriptional read-through promotes the downstream expression of proteins that export the toxic fluoride ion. To understand this mechanism of transcriptional regulation, we investigated the fluoride-sensing riboswitch from *Bacillus cereus*.

Structural information is often central to an understanding of the mechanism of biomolecular machines. Although structures of the isolated fluoride riboswitch in both its ligand bound and unbound states have been resolved, there are no structures of the riboswitch in complex with the RNAP and DNA template where regulation occurs. To elucidate the mechanism by which riboswitch folding promotes the switch from transcription termination to read-through, we used cryo-EM to solve the structures of Pause Elongation Complexes (PEC) stalled prior to the termination point in the fluoride free and bound states. We further applied molecular dynamics flexible fitting to build structurally informed models of the incompletely resolved RNA in both states. Our reconstructions reveal that a ligand-induced shift in riboswitch pose outside the RNAP exit channel triggers global conformational changes of the enzyme, promoting pause release. The global rearrangements of the protein upon ligand binding result in increased interactions between the RNAP and RNA within the RNA-DNA hybrid region, favoring a transcription-competent enzyme. Combined with biochemical assays, we propose an RNA-based model in which co-transcriptional folding and structural rearrangement modulate RNAP processivity.Our results illuminate the mech-

anism of communication between the nascent transcript and the macromolecular machinery of the cell that could be leveraged to modulate gene expression.

4.2 Methods

4.2.1 RNAP expression and purification

The wild-type RNAP was expressed and purified by lab members from the Walter Lab.

The expression and purification of *E. coli* RNAP were described in a previous study114. *E. coli* RNAP was obtained from the *E. coli* strain BLR λDE3. Cells were grown in Luria medium with kanamycin (25 μg ml^{-1}) and incubated at 37 °C on a platform shaker (250 rpm) until an apparent optical density (OD600) of 0.5 to 0.6. Protein expression was induced through the addition of IPTG (1 mM). The cells were incubated for 3-4 hours at 37 °C and 200 rpm and centrifuged (3,400 g, 15 minutes, 4 °C). Lysis buffer (50 mM Tris-Cl, pH 8.0, 2 mM EDTA, 1 mM β-mercaptoethanol, 5% v/v glycerol, 300 mM NaCl, 1 mM DTT, 0.25 mL of 10 mg PMSF per mL in ethanol, and 0.5 mL of a protease inhibitor cocktail (1 mg antipain, 1 mg aprotonin, 31.2 mg benzamide, 0.5 mg chymostatin, 0.5 of leupeptin, and 0.1 mg of pepstatin per mL in ethanol). The resuspended cells were sonicated. All purification steps were performed at 4 °C. Polyethylenimine (average molecule weight 60 K; Acros Organics, catalog no. 17857) was added to the sonicated cells to a final concentration of 0.6 % with gentle stirring and centrifuged (20,000 g, 15 min) to recover the precipitate. The pellet was gently resuspended in 25 mL of TGEDZ buffer (10 mM Tris-Cl pH 8.0, 1 mM DTT, 0.1 mM EDTA, 5 % glycerol, 5 µM $ZnCl_2$) at 500 mM NaCl and then centrifuged (20,000 g, 15 min). The pellet was resuspended in 25 mL of TGEDZ buffer at 1 M NaCl and centrifuged (20,000 g, 15 min). The RNAPs was precipitated by the slow addition of solid ammonium sulfate to 37% w/v final while stirring gently and left overnight. The RNAPs

were centrifuged (20,000 g, 15 min) and then redissolved in 20 mL of buffer A (50 mM NaH_2PO_4, 300 mM NaCl, 1 mM DTT, 5% glycerol, pH 8.0) containing 10 mM imidazole. The sample was loaded onto a 10 mL Ni^{2+}-NTA agarose resin (Qiagen) column, then washed with 50 mL of buffer A containing 20 mM imidazole, and eluted with buffer A containing 250 mM imidazole. Collected fractions containing RNAP were identified through a Bradford assay and subsequently combined and dialyzed with 2 L TGEDZ buffer containing 200 mM NaCl for 3-4 hours. The RNAP was loaded onto a heparin-sepharose column (5 mL HiTrap) at 3 mL min^{-1} by an AKTA purifier (GE Healthcare), then washed with 25 mL of TGEDZ buffer containing 200 mM NaCl, and eluted with TGEDZ buffer containing 500 mM NaCl. The purified RNAPs were dialyzed into storage buffer (20 mM Tris-Cl pH 8.0, 250 mM NaCl, 20 μM $ZnCl_2$, 1 mM $MgCl_2$, 0.1 mM EDTA, 1 mM DTT, 25 % glycerol) and stored at -80 °C.

4.2.2 Preparation of elongation complex for cryo-EM

Adrien Chauvier performed the complex preparation.

The RNA for the fluoride-PEC was transcribed in vitro with the T7 RNAP and gel purified. DNA oligonucleotides were ordered from Integrated DNA technologies. The nucleic scaffold containing tDNA, ntDNA and RNA was assembled at 1:1:1 ratio in buffer B (50 mM Tris-HCl pH 7.5, 100 mM KCl) at 90 °C for 2 minutes followed by 10 minutes at 37 °C, and an additional 10 minutes at 23 °C. The *E. coli* RNAP was added to the nucleic scaffold at a molar ratio of 1:3 in assembly buffer (50 mM Tris-HCl pH 7.5, 100 mM KCl, 1 mM $MgCl_2$) and incubated at 37 °C for 15 minutes. CHAPSO (Sigma Aldrich, catalog no. 82473-24-3) was added to a final concentration of 8 mM and the entire complex was then concentrated with a centrifugal filtration column (Amicon, 100 kDa molecular weight cutoff). Grid preparation of the complex was at 5-6 mg mL^{-1} and the fluoride ion was added

to the complex at a final concentration of 2.5 mM for 5 minutes at room temperature when present.

4.2.3 Time pausing assays

The time pausing assays of ligand bound and unbound fluoride with NusG were performed by Emily Ellinger. The transcriptional pause assays of RNAP triple mutant with fluoride were performed by Adrien Chauvier.

A DNA template containing the T7A1 promoter and the fluoride sensing riboswitch from *B. cereus* was generated using an overlapping PCR strategy. Directly following the promoter region was a stretch of 25 nucleotides that did not contain any thymine nucleotides apart from the +2 position that allowed for the initiation of transcription with the ApU dinucleotide. The template was generated by PCR using the 'T7A1-fluoride-RNA0p' forward nucleotide and the complementing reverse oligonucleotides. The formation of the halted complexes was performed using a 25 μM ATP/CTP mix, 50 nM α-[32P]GTP (3,000 Ci mmol^{-1}), 10 μM ApU dinucleotide primer (Trilink, catalog no. O-31004) and 50 nM DNA template in transcription buffer (20 mM Tris-HCl, pH 8.0, 20 mM NaCl, 20 mM $MgCl_2$, 14 mM 2-mercaptoethanol, 0.1 mM EDTA). 100 nM *E. coli* RNAP holoenzyme (expressed and purified in our lab) was added to bind to the promotor region during the incubation period of 10 minutes at 37 ℃. The complex was purified with a G50 gel filtration column (GE Healthcare, catalog no. 27533001) to remove free nucleotides. A mixture of rNTPs (10 μM each), heparin (450 μg ml^{-1}), and sodium fluoride when present (5 μM) were added to the complex to complete single round transcription. The reaction mixture was incubated at 37 ℃ and aliquots were quenched at various timepoints in 2x volume of loading buffer (95% formamide, 1 mM EDTA, 0.1% SDS, 0.2% bromophenol blue, 0.2% xylene cyanol). Aliquots were denatured at 95°C for 2 minutes before loading 6 μl each onto a denaturing 8 M urea,

6% polyacrylamide sequencing gel. The gel was dried and exposed to a phosphor screen (typically overnight), that was then scanned on an Amersham Typhoon PhosphorImager (GE Lifesciences). Gel images were analyzed with ImageLab (Bio-Rad) software.

4.2.4 Time-pausing analysis

Time-pausing analysis for the wild-type RNAP was conducted by Emily Ellinger. Time-pausing analysis for the triple mutant RNAP was conducted by Adrien Chauvier.

The fraction of the RNA pause species compared with the total amount of RNA for each time point was calculated to determine the half-life of transcriptional pausing. The results were analyzed with pseudo-first order kinetics to extract the half-life. For each determination the global background signal was subtracted. Error bars in transcription quantification represent the standard deviation of the mean from independent replicates.

4.2.5 Molecular dynamics flexible fitting simulations

Molecular dynamics flexible fitting (MDFF) simulation is a simulation method used to fit atomic coordinates into cryo-EM density maps [114, 115]. In MDFF, external forces proportional to the gradient of the density map are added during MDFF simulations.

The density map is converted into the potential U_{EM}, so that the high-density regions in the density map correspond to the minima on the potential energy surface. Forces proportional to the gradient of the potential are applied to fit the atoms into the energy minima of the potential energy surface. The potential U_{EM} is defined as:

$$U_{EM}(R) = \Sigma_j \omega_j V_{EM}(r_j) \tag{4.1}$$

where j denotes the atoms, ω_j are the atomwise weights, and the MDFF potential map V_{EM} is given by:

$$V_{EM}(r) = \begin{cases} \xi[1 - \frac{\Phi(r)-\Phi_{thr}}{\Phi_{max}-\Phi_{thr}}] \text{ if } \Phi(r) \geq \Phi_{thr}, \\ \xi \qquad\qquad\qquad\qquad \text{if } \Phi(r) < \Phi_{thr}. \end{cases} \tag{4.2}$$

Where ξ is the force scaling factor, $\Phi(r)$ is the EM density at position r. Φ_{max} is the maximum of the density map and Φ_{thr} is a density threshold. The density threshold is used to eliminate the EM data corresponding to the solvent contribution. The guiding force of MDFF that fit the structure to the density map is defined as:

$$f_i^{EM} = -\frac{\partial}{\partial r_i} U_{EM}(R) = -w_i \frac{\partial}{\partial r_i} V_{EM}(r_i) \tag{4.3}$$

The initial structure of the system consisted of RNA polymerase, template and non-template DNA, and a 9-residue long downstream RNA-DNA hybrid formed by riboswitch RNA and template DNA. To obtain the complete structure for the simulations, a linker was generated in PyMOL [116] and attached the linker to the 5' end of the RNA present in the initial structure for both ligand bound and ligand unbound systems. Then a 51-residue aptamer domain of fluoride riboswitch (PDB 4ENC) [33] was attached to the 5' end of the linker. Since the cryo-EM density map did not have enough density to fit the entire riboswitch, the first 44 residues of the aptamer domain were removed from the structure. For the ligand bound system, the fluoride ion was not included because the binding pocket was removed with the first 44 residues. The hydrogens in the linker were removed using PyMOL [116]. CHARMM-GUI web server was used to add hydrogens and terminal patches [83]. To remove the clashes in the structures obtained from CHARMM-GUI, energy minimization was applied with 100 steps of steepest descent and 500 steps of adopted basis Newton-Raphson method with a gradient tolerance of 0.01. The energy minimization was run in vacuum using CHARMM. The nonbonded list was generated using a cutoff of 16 Å. The Lennard-Jones and electrostatic interactions were treated using the switching function. The

backbone heavy atoms of protein and nucleic acid were harmonically restrained with a force constant of 1.0 kcal/mol/Å^2. The heavy atoms of the protein side chain and the base of the nucleic acid were harmonically restrained with a force constant 0.1 kcal/mol/Å^2. For the energy minimization, the CHARMM36 force field was used for protein and nucleic acid [84].

The models obtained after energy minimization and the cryo-EM density maps were used to set up the MDFF simulations. For cryo-EM density maps, we picked the cluster with the most density in the aptamer domain to conduct the simulations. For the ligand unbound state, cluster 2 was used and for ligand bound state, cluster 0 was used. We picked these two clusters because they have the most density available for the riboswitch, which allow us to fit more residues. The densities associated with the exit channel and the aptamer domain for each cluster are shown in Figure 4.1A and the zoomed view of the density is shown in Figure 4.1B. The input files for the simulations were generated using the VMD software [117]. Among the input files, restrain files were generated to preserve the secondary structure of the protein, DNA and RNA to avoid overfitting during the simulations. For MDFF simulations, 1000 steps of energy minimization were performed followed by the production run. The CHARMM36 force field was used for protein and nucleic acid. For production runs, the simulations were run in vacuum for 2 ns in NAMD [118]. The time step for the simulation was set to 1 fs. The temperature was maintained at 300 K using the Langevin thermostat. During the simulations, the coordinates of RNA polymerase, template and non-template DNA and the 9-nucleotide RNA base paired with DNA were held fixed by setting the B factor to 1 in the input file. The upstream 12 residues of the riboswitch were allowed to move to fit it into the density map. For MDFF simulations, the atomwise weights ω_j were set to the atomic masses and the scaling factor ξ was set to 0.3. For both ligand bound and ligand unbound systems, 100 MDFF trajectories were generated independently, and the global cross-correlation coefficients were calculated using the MDFF plugin in VMD to select the best model. For both of the systems, the structure with the highest global cross-correlation coefficient was selected for further analysis.

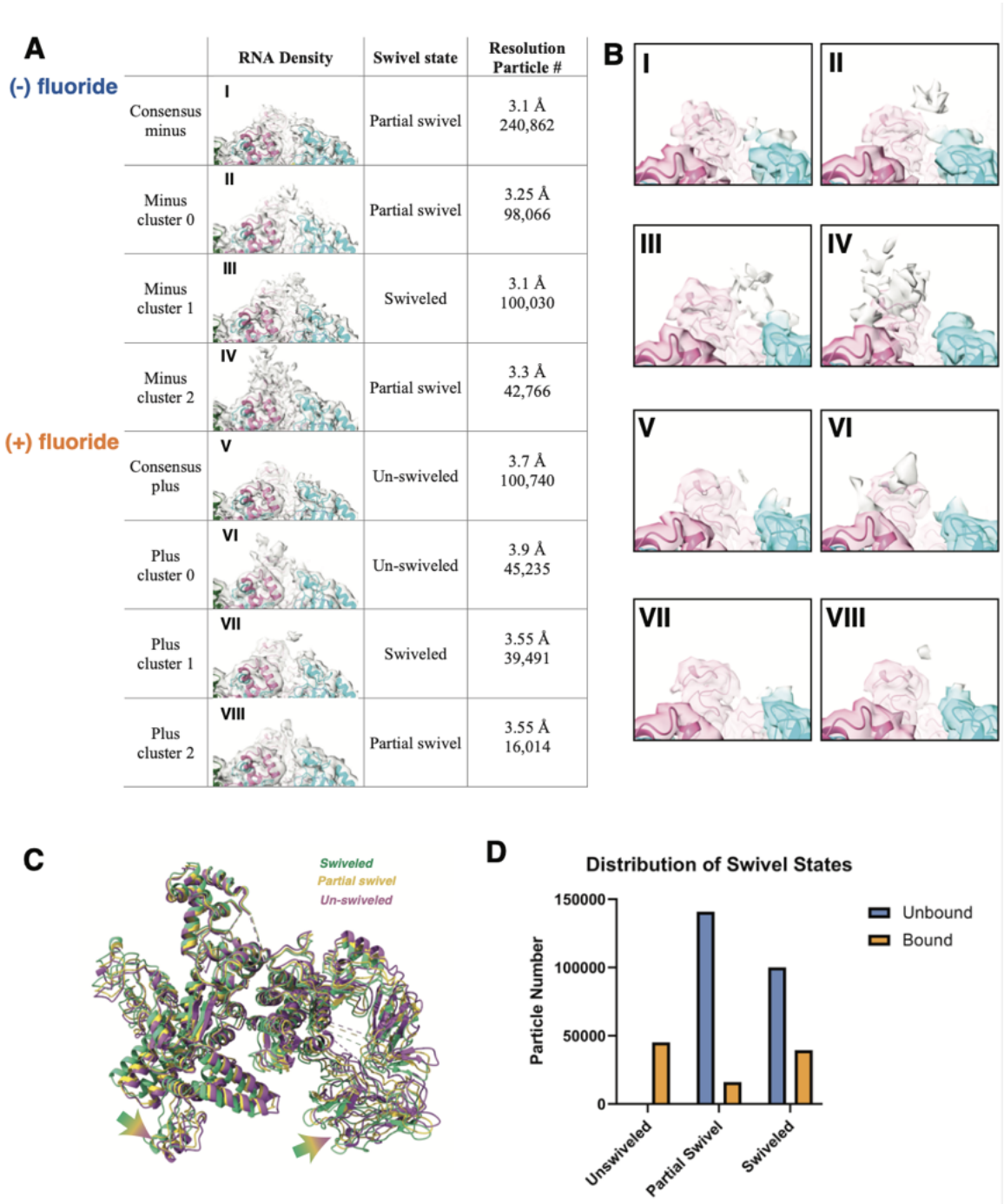


Figure 4.1: Comparison of the densities available for each cluster and the distribution of swivel state. (A) The densities of the RNA in each cluster and the description of the corresponding swiveled state and resolutions. The β subunit is colored in cyan and the β' subunit is colored in pink. (B) Zoomed view of densities available for riboswitch. (C) Illustration of the swiveled, partial swivel and unswiveled states of the RNAP. (D) Distribution of swivel states. The distribution shows that the ligand bound state favors the unswiveled state and the ligand-unbound state favors the swiveled state.

4.3 Results

4.3.1 The in vitro assembled fluoride-PEC is functional

The fluoride-sensing riboswitch contains multiple pause sites with the U72 pause located directly upstream of the terminator, potentially acting as the final checkpoint before partitioning between termination and anti-termination (Figure 4.2A,C)[119, 120]. In the context of PEC-72, the Paused Elongation Complex arrested at U72, stabilization of the pseudoknot structure within the RNAP exit channel is favored compared to the complete terminator hairpin [119]. The *E. coli* RNAP was chosen to investigate the mechanism because it can recognize these pauses. The fluoride-PEC was assembled through the reconstitution of a transcription bubble comprised of DNA and RNA oligonucleotides (Figure 4.2B). The fluoride-PEC was then purified through size-exclusion chromatography. Then the particles were frozen in the cryo-EM buffer containing 8 mM CHAPSO to allow random orientations of the particles. The buffer components are not likely to interfere with the termination behavior of the riboswitch [113, 121].

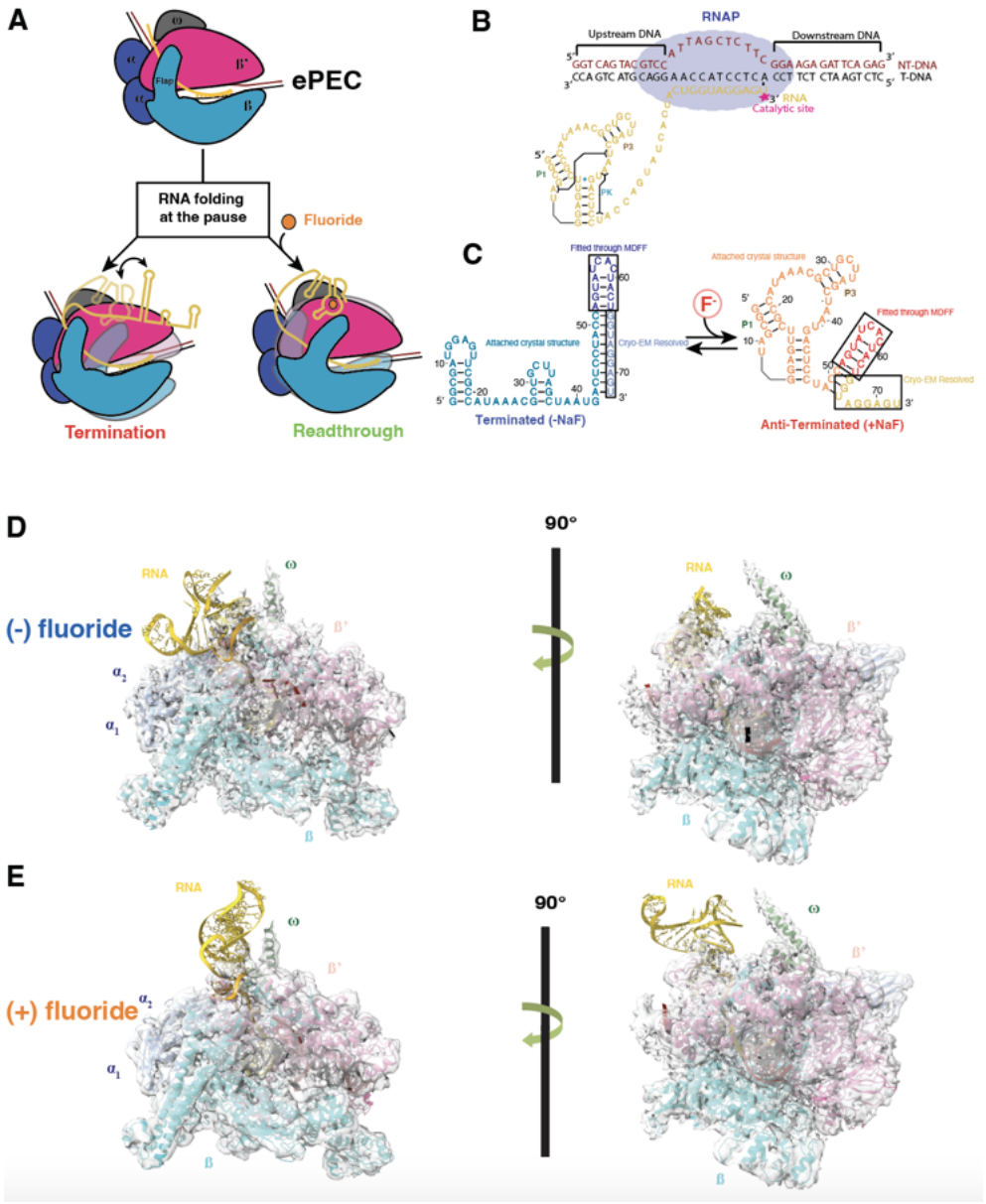


Figure 4.2: (A) Illustration of the structural alteration of the fluoride riboswitch in response to the binding of fluoride ligand. (B) Nucleic acid scaffold used for cryo-EM data acquisition. The secondary structure of the riboswitch and the sequence of the template and non-template DNA are shown. (C) Secondary structure of the ligand-bound and ligand-unbound structure. The regions fitted by MDFF and resolved by cryo-EM are boxed. (D) The model after attaching the aptamer domain of the fluoride riboswitch in the ligand unbound state. The region fitted through MDFF is rendered in gold, the attached aptamer is rendered in yellow. The cryo-EM density map rendered as transparent surface. (E) The model after modeling the aptamer domain of the fluoride riboswitch in the ligand bound state. The region fitted through MDFF is colored in gold, the attached aptamer is colored in yellow. The cryo-EM density map rendered as transparent surface.

4.3.2 Data collection

A single consensus three-dimensional reconstruction was determined for each the ligand-unbound and -bound PEC-72u and PEC-72b, respectively each reconstruction shows details of the active site and the 9-nucleotide RNA-DNA hybrid. The initial model was built by placing the *E. coli* RNAP(PDB ID: 6ALF) [122] in its elongation complex into the EM density map. The resolution of the RNAP ranged from 3.3-11 Å for PEC-72u and 3.8-11 Å for PEC-72b.

To study the conformational heterogeneity, the particles were submitted to 3D variability analysis (3DVA) and found to generate 3 clusters for each PEC. To be able to fit more riboswitch nucleotides into the cryo-EM density map, we picked the clusters with the most densities available for the riboswitch to conduct MDFF simulations. For PEC-72u, the structure from cluster 2 was used for MDFF simulations with resolution 3.3 Å. For PEC-72b, the structure from cluster 0 was used for MDFF simulations with resolution 3.9 Å.

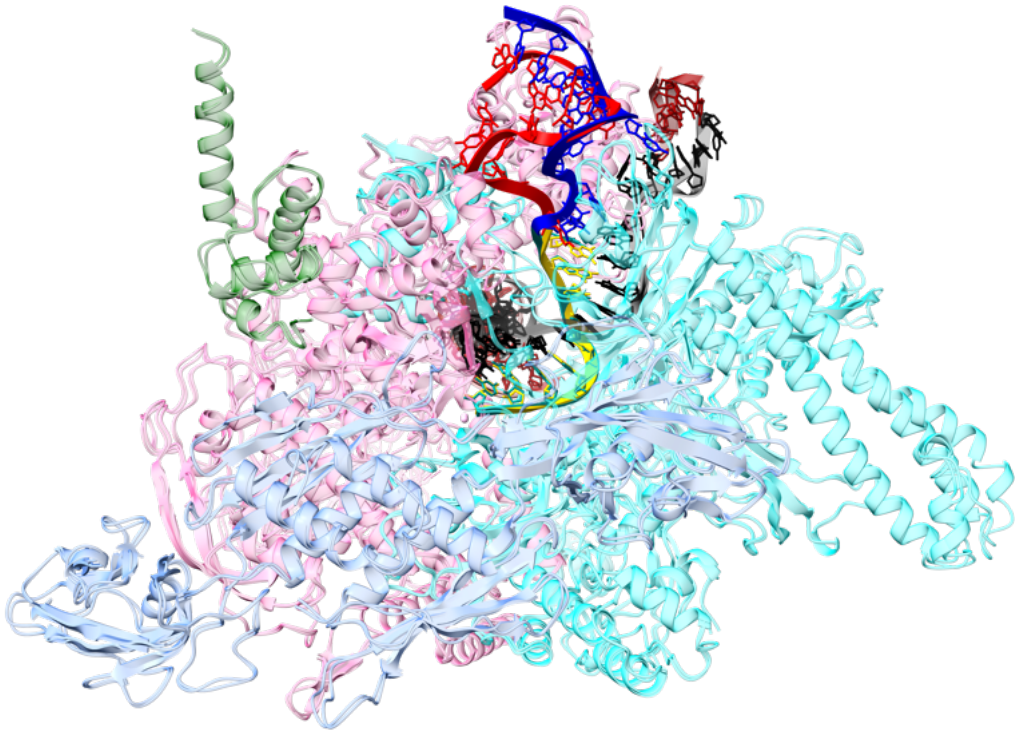
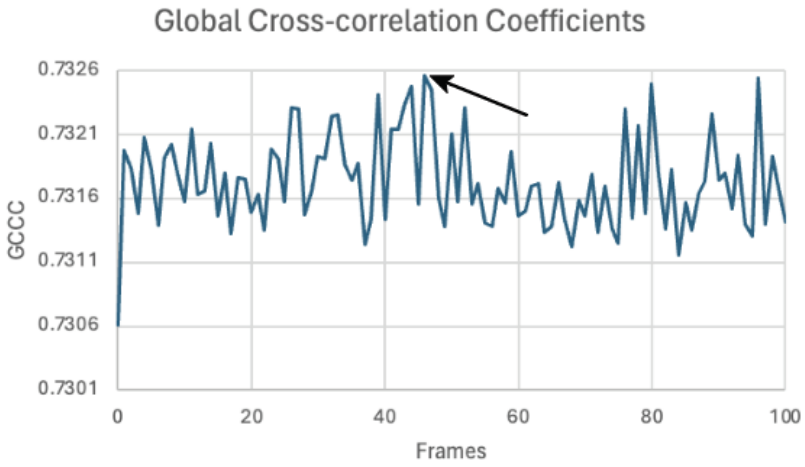


Figure 4.3: RNAP in complex with the MDFF fitted region of the riboswitch. The fitted residues for the ligand bound state are colored in red. The fitted residues for the ligand unbound state are colored as blue. The RNAP and template/non-template DNA are rendered as transparent cartoons.The non-template DNA is colored in dark red; the template DNA is colored in black. The α subunit of the RNAP is colored in corn flower blue, the β subunit is colored in cyan, the β' subunit is colored in pink, and the ω subunit is colored in dark green.

4.3.3 Ligand binding induces riboswitch rotation and opens the RNAP exit channel

The fluoride riboswitch is located at the solvent accessible region and is conformationally flexible, so its structure could not be confidently resolved using cryo-EM. To generate a model of the entire complex, we fitted 12 residues upstream of the RNA-DNA hybrid of the riboswitch using molecular dynamics flexible fitting (MDFF) simulations [114, 115] using the cryo-EM density map. The 12 residues include 7 residues located in the exit channel and 5 residues of the aptamer domain (Figure 4.3, Figure 4.2C). For 100 independent trajectories generated for each PEC-72u and PEC-72b, the global cross-correlation coefficients (GCCC) were calculated using the MDFF plugin in VMD[117]; the best-fit models were then selected based on the GCCC. The structure with the highest GCCC from each system was selected for subsequent structural analysis. For the ligand free structure, the GCCC ranges from 0.7306 to 0.7326 (Figure 4.4A), which means the accuracy of the fit is reasonably good. For the ligand bound structure, the GCCC is even better and ranges from 0.8080 to 0.8118 (Figure 4.4B).

A



B

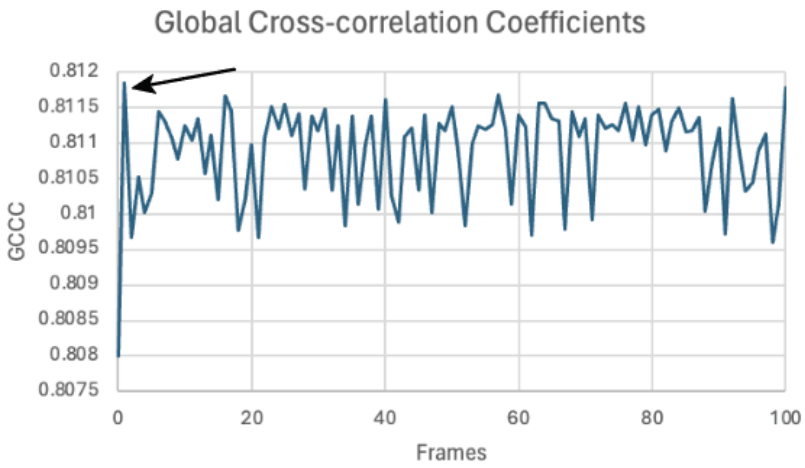


Figure 4.4: (A). Plot of the global cross-correlation coefficients for the 100 structures generated for the ligand-unbound state. For the ligand-unbound state, trajectory 46 was used as the model as it has the highest GCCC among all of the fitted structures. (B) Plot of the global cross-correlation coefficients for the 100 structures generated for the ligand-bound state. For the ligand-bound state, trajectory 1 was used as the model as it has the highest GCCC among all of the fitted structures.

Since the GCCC values of the 100 structural models span a very small range, it is crucial to validate our model by analyzing more structural models generated for each system, as the GCCC values indicate that there is no significant differences in terms of the quality of the fitting. Therefore, we analyzed the top 20 models for each system. For the top 20 models, the MDFF fitted region of the riboswitch formed a cluster for each system, and the two clusters centered around the model with the highest GCCC for each system. This result indicates that our selection of model is reasonable since there is no significant difference in the top 20 models generated for each system.

The structures fitted through MDFF showed distinct conformations between the ligand bound and unbound states of the fluoride riboswitch (Figure 4.2). To quantify the differences between the two states, we generated a plane for both ligand unbound and ligand bound state aptamer and calculated the angle between the two planes using Chimera (Figure 4.5). The ligand bound riboswitch rotates 72.31° compared to the ligand unbound state, and the nucleotides A52 – G53 (the first two residues of the MDFF fitted region) of the two states are oriented in different directions (Figure 4.6 A). These experimentally informed structural models were then used as a foundation to graft on the remainder of a previously determined crystal structure (PDB: 4ENC) [33]. Attaching the aptamer domain provides complete models that provide evidence that binding of the fluoride ion to the riboswitch induces a rotation of the RNA and expands the exit channel (Figure 4.6).

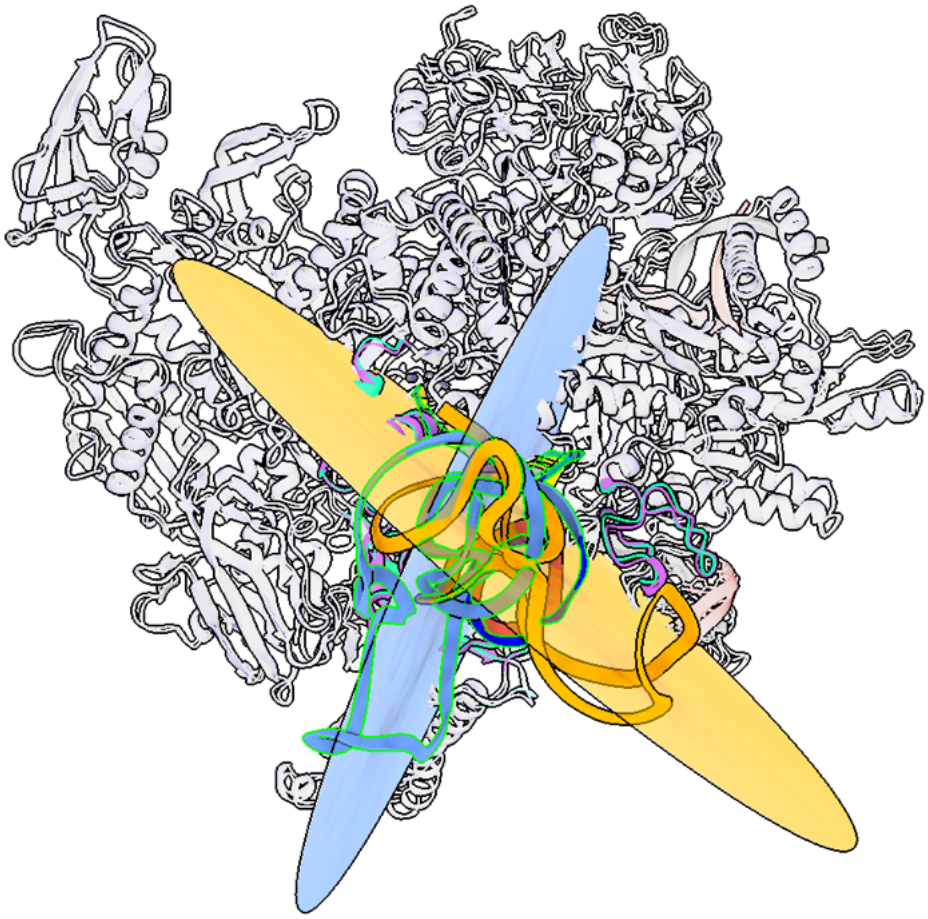


Figure 4.5: The plane generated in Chimera to quantify the twist upon ligand binding. The angle between the ligand-unbound aptamer (blue) and the ligand-bound (gold) aptamer was calculated.

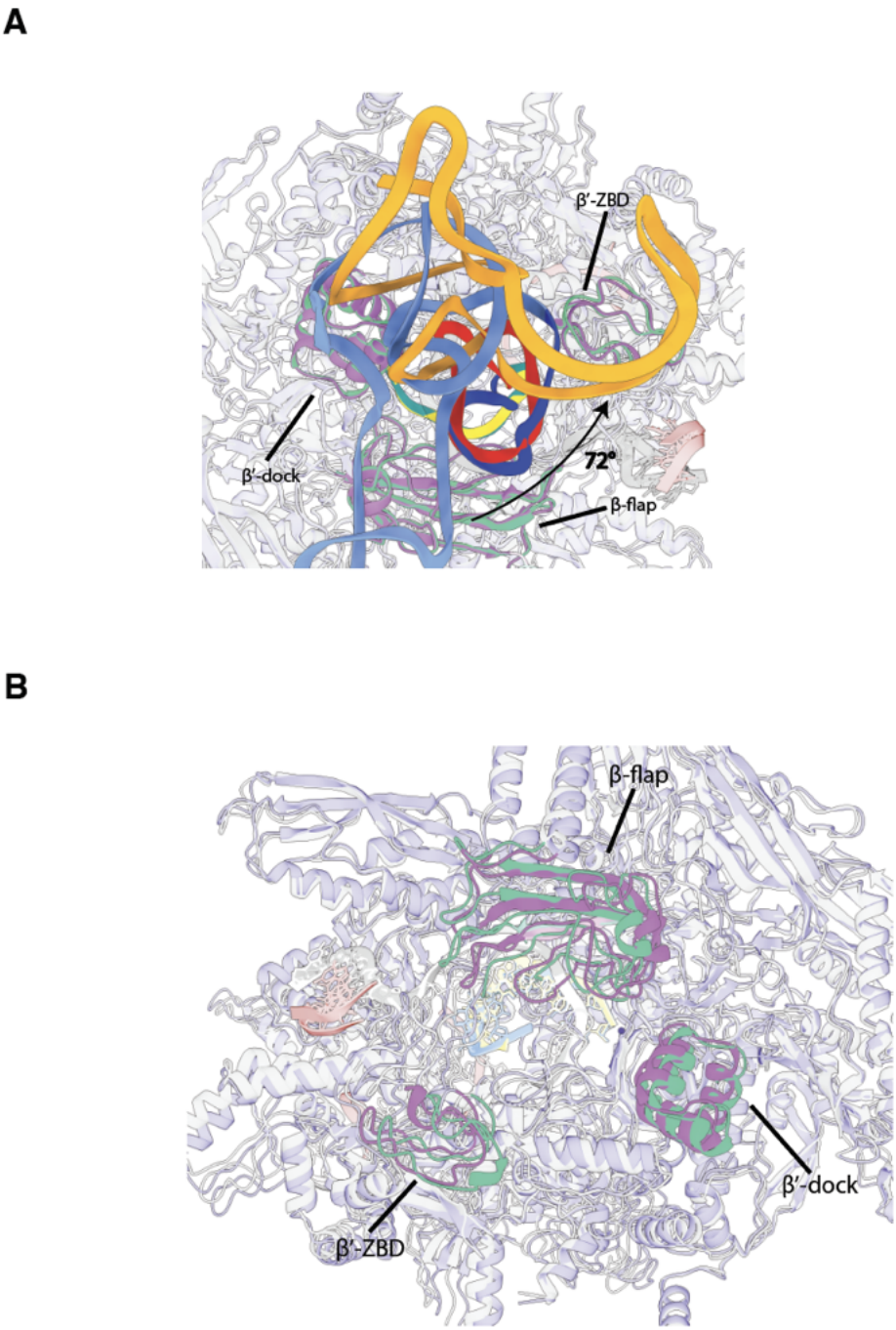


Figure 4.6: (A) View from the outside of the RNAP exit channel with the attached aptamer domain of the ligand bound (gold) and unbound (blue) riboswitch. A twist of the aptamer domain is observed. (B) The exit channel of the ligand-unbound complex compared to the ligand-bound complex. The β-flap, β'-dock and β'-ZBD domains are highlighted for the ligand-bound (purple) and -unbound (green) complexes. The β'-ZBD domain of the ligand-bound complex shifts and opens the exit channel compared to the ligand-unbound complex.

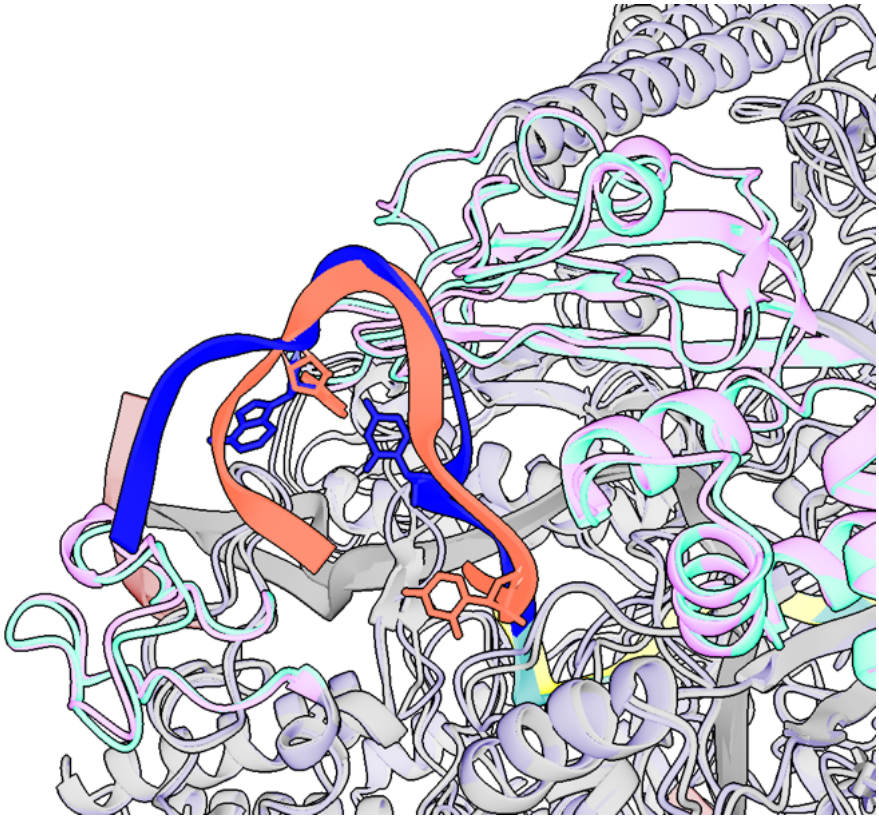


Figure 4.7: Distance between residue A55 and U60 for the ligand bound (red) and ligand unbound (blue) system. Nucleotides A55 and U60 are shown. The residues A55 and U60 are closer to each other in the ligand free system.

In addition to quantifying the rotation of the aptamer domain, we also calculate the distance between the residue A55 and U60. The distance was calculated from the distance of the phosphorus atom of each residue. The two residues are expected to form a base pair in the ligand bound and ligand free system. For the top 20 models, the average distance between A55 and U60 for the ligand bound system is 18.42 Å. And for the ligand free system, the average distance is 15.92 Å(Figure 4.7). The distances show that for both systems, there is a tendency to form the stem-loop structure and the shorter average distance between A55 and U60 in the ligand free state suggests that there is a stronger tendency to form base pair between A55 and U60 without the ligand. Although we fitted in a single strand of RNA, the MDFF fitted structures reveal the tendency to form stem-loops for both systems.

4.3.4 Global conformational changes within the PEC

The global motions of the RNAP are difficult to pinpoint in the bound (PEC-72b) and unbound (PEC-72u) and consensus structures because the movement of RNAP is heterogeneous. To determine distinct conformations of the RNAP, the particles were submitted to three-dimensional variability analysis (3DVA) in cryoSPARC. 3DVA is a powerful tool for analyzing the discrete and continuous heterogeneity in single particle cryo-EM datasets[113]. From this analysis, 20 structural intermediates were calculated. They show that the RNAP predominantly resides in a swiveled state in the ligand-unbound state, which is known to be associated with a paused state of RNAP [121, 123, 124] (Figure 4.8). In the swiveled state, the RNAP structural modules (clamp, dock, shelf, SI3, and C-terminal segment of the β' subunit) are rotated about an axis that is perpendicular to the plane defined by the helical axis of the DNA-RNA hybrid and the downstream DNA[113]. The range of the swiveling is up to 5°. In contrast, in the fluoride bound state the swiveling motion is less pronounced, up to 2°. The bound state favors the unswiveled state similar to the active elongation complexes. The swiveling of the unbound state has a maximum of 5°, comparing to ∼1.5° for the *his*-PEC [125]. The swiveling motion of the unbound state fluoride-PEC in the absence of transcription factor is comparable to the textithis-PEC in the presence of NusA (4.5°-6°)[125]. NusA is a transcriptional pausing factor that binds to RNA hairpins and stabilizes pausing behavior during transcription [126, 127].

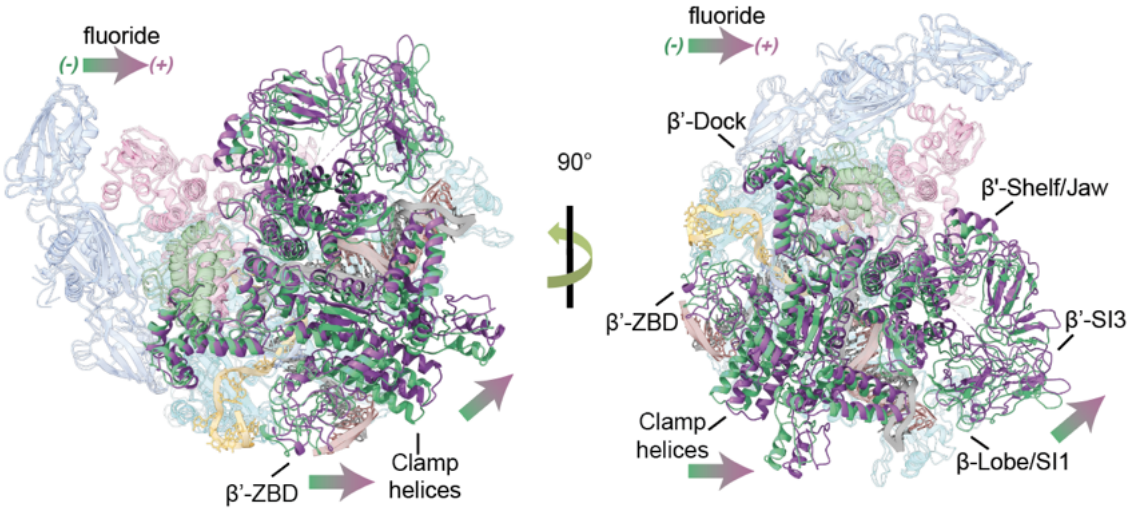


Figure 4.8: Comparison of the ligand-unbound (green) and ligand-bound (purple) states of RNAP, which are predominantly swiveled and unswiveled, respectively.

To validate the swiveled state of the unbound complex, in vitro time-pausing assays of the *crcB*-PEC72 in the presence of the transcriptional factor N-utilizing substance G protein (NusG) were performed. NusG functions as elongation factor in some Gram-positive bacteria such as *E. coli* and functions as a termination factor in some Gram-negative bacteria such as Bacillus subtilis [128, 129, 130, 131]. NusG binds to the tip of RNAP clamp and keep the clamp closed during transcription, which stabilizes the unswiveled state and prevents termination, leading to an increased transcription rate[121]. NusG is known to preferentially bind to the unswiveled state [121]. In the absence of ligand, it was observed that the pause half-life of the U72 pause is unaffected in the presence of NusG, suggesting that the RNAP is in the swiveled state and the NusG is unable to bind to the RNAP. The results for the time pausing assays are shown in Figure 4.9.

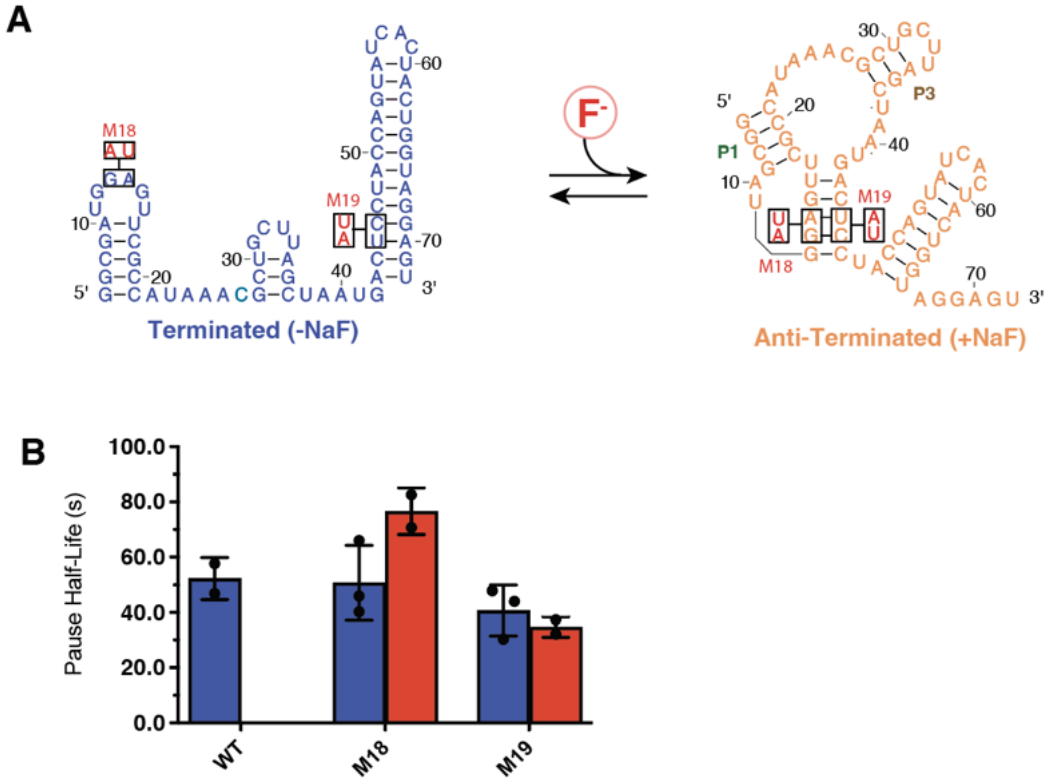


Figure 4.9: Time pausing assays for the wild-type and mutants RNAP in the presence of NusG. The ligand-unbound state is colored in blue and the ligand-bound state is colored in red.

4.3.5 Ligand binding alters RNA-protein interactions

Upon ligand binding, we observed an increase in interactions between the RNA within the RNA-DNA hybrid and the RNAP, particularly in the β' domain (Figure 4.10 A,B). Similar trend was observed for the RNA near the exit channel. The ligand-bound state forms more interactions that the ligand unbound state. The ligand-bound state forms 15 interactions in total while the ligand-unbound structure only forms 10 interactions. For example, an interaction equivalent to that between K1306 and U60 in the ligand bound state was previously observed in both the ligand unbound and bound states of the preQ1 riboswitch [113] and is absent from the unbound state of PEC-72u. The increase in interactions upon ligand binding and the differences in conformations of the exit channel show that ligand binding induces a twist in RNA that forms more interactions in both the RNA-DNA hybrid and exit channel with the RNAP, the increase in interaction upon ligand binding is consistent with the observation that the ligand-bound state favors the less swiveled motion, which stabilizes a structure that allows transcription to continue. The results are also consistent with the shift in the β'-ZBD domain as shown in Figure 4.6, where in the fluoride bound state, the β'-ZBD domain moves away from the RNA, providing more space for the RNA to exit (Figure 4.6B). Intriguingly, a shift within the active site was not observed in contrast to the que-PEC complex [113].

To better visualize the interactions between riboswitch and the exit channel, a space filling model with the MDFF-fitted residues was generated for the ligand-bound structure and the ligand-unbound structure (Figure 4.11). For the ligand-bound state, the RNA is closer to the β'-dock domain while for the ligand-unbound state, there is no interactions between the β'-dock domain and the RNA.

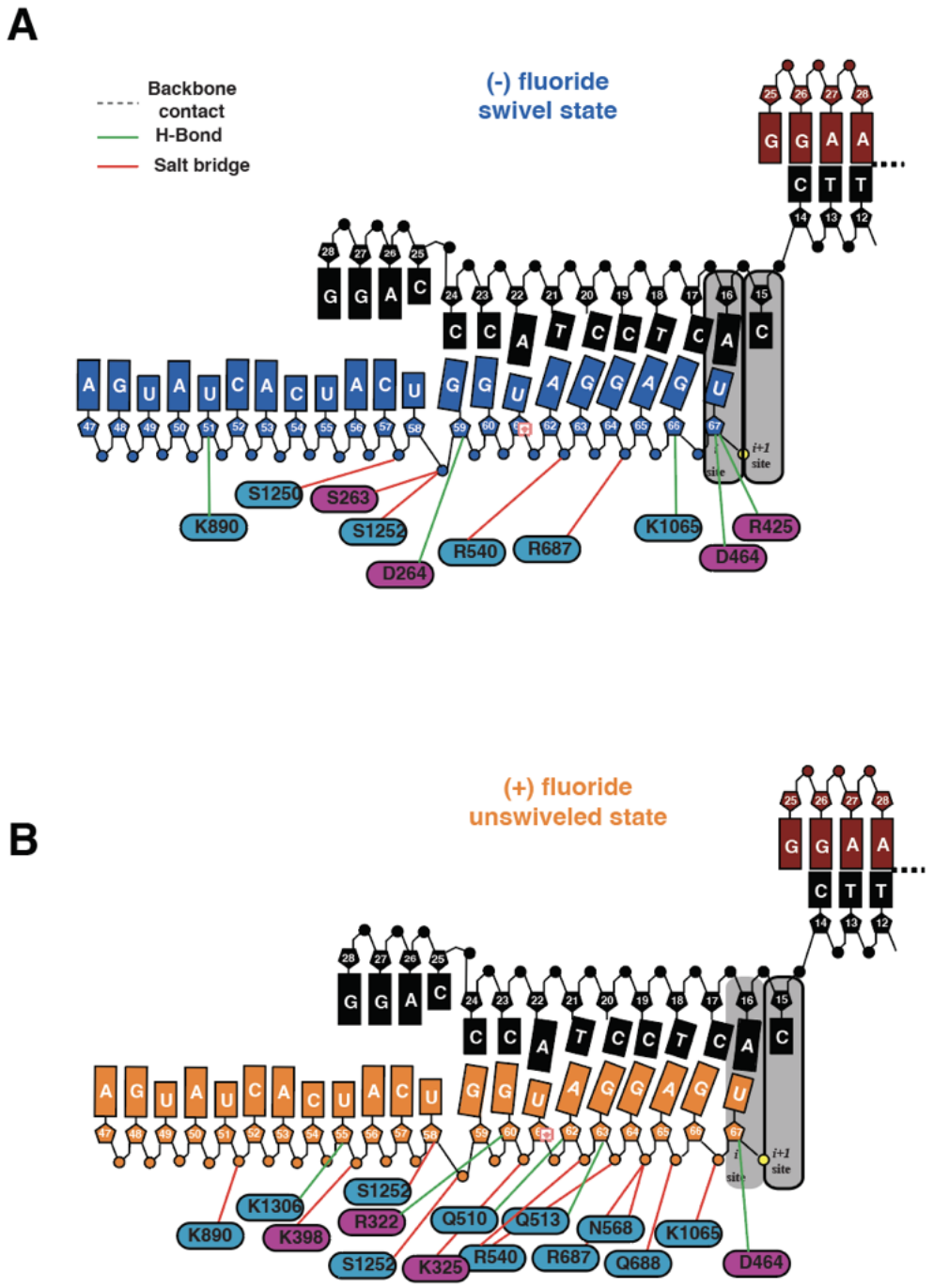


Figure 4.10: Interactions between the riboswitch and the RNAP. (A). Interactions between the RNA residues in the MDFF-fitted region and the RNA-DNA hybrid with the RNAP for the ligand-unbound state. (B). Interactions between the RNA residues in the MDFF-fitted region and the RNA-DNA hybrid with the RNAP for the ligand-bound state.

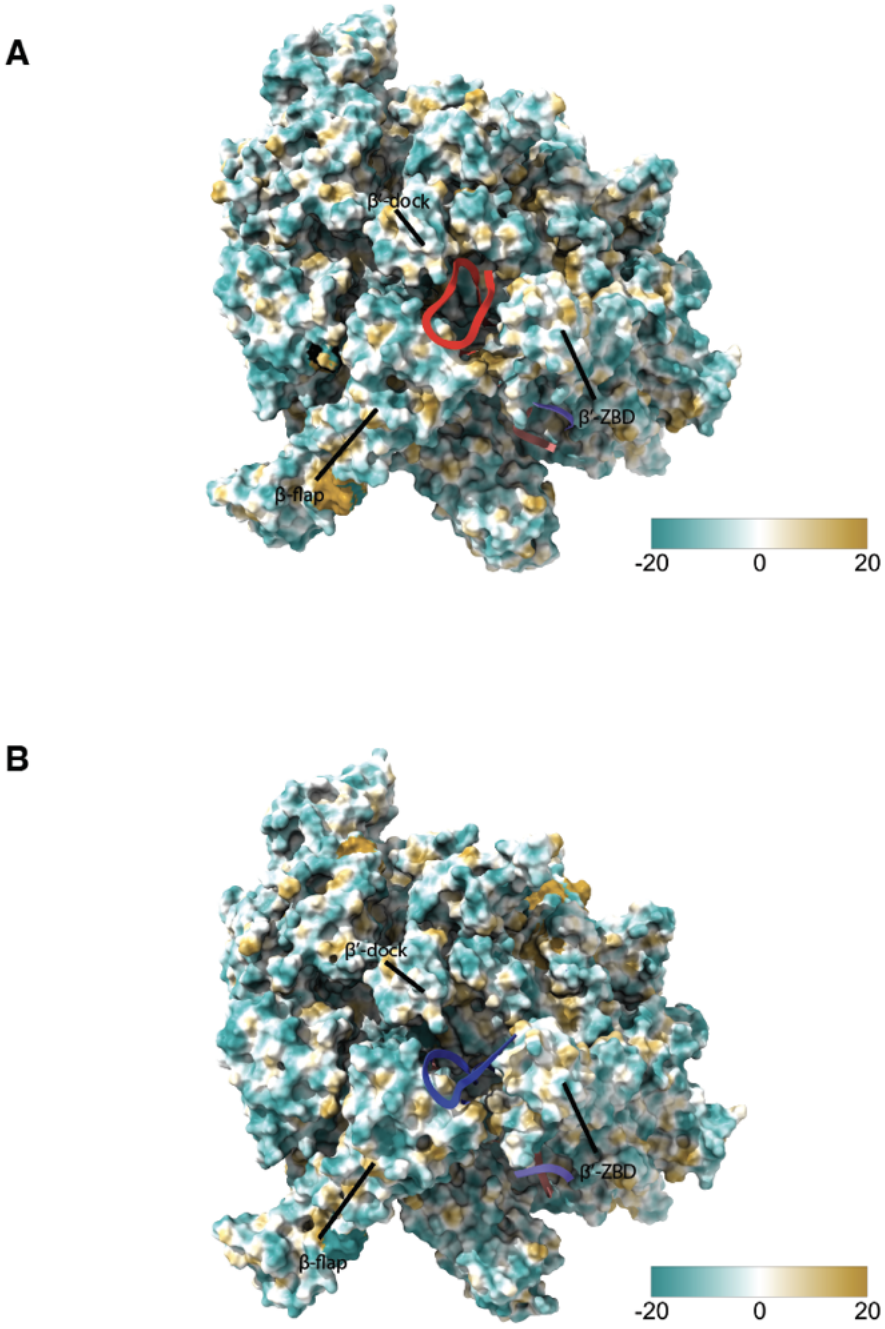


Figure 4.11: (A) Space filling model for the fluoride bound complex. (B) Space filling model for the ligand unbound complex. The RNA moves away from the β'-dock domain compared to the ligand-bound state.

4.3.6 Time pausing assays validate switching behavior of the fluoride riboswitch

The SiM-KARTS experiments for M18 was performed by Emily Ellinger. The SiM-KARTS experiments for M19 was performed by Adrien Chauvier.

To better understand the switching behavior of the fluoride riboswitch, time pausing experiments were conducted to explore the relationship between the formation of the pseudoknot and termination destabilization. We utilized 2 previously studied mutant fluoride riboswitches, M18, which promotes the termination structure, and M19 which destabilizes both the terminator and pseudoknot structure (Figure 4.9A)[28, 132]. In addition, Single Molecule Kinetic Analysis of RNA transient Structure (SiM-KARTS) experiment were performed to monitor the binding of a probe to the terminator structure. The U72 pause remains robust in the presence of fluoride for both of the mutants, indicating the stabilization of the pseudoknot structure is necessary precursor that leads to pause escape (Figure 4.9A, B).

4.4 Discussion

Riboswitches play essential roles in the regulation of bacterial gene expression [133, 134, 135, 136], rendering them significant model systems for studying the mechanism of transcription elongation and termination. For example, understanding the structural basis of transcription termination can provide insights for developing novel antibacterial drugs that counter antibiotic resistance[30, 110, 111].

Our work leveraged cryo-EM and single molecule methods to study the RNAP pausing induced by RNA structure alone of the *crcB* fluoride-sensing riboswitch PEC. We report structures of the fluoride-PEC in the ligand unbound and ligand bound states at resolutions

of 3.3 Å and 3.9 Å, respectively. MDFF simulations were utilized to resolve the flexible region of the RNA around the RNAP exit channel. We compared the global conformational rearrangements of the RNAP in response to ligand binding and found that the swiveled state in favored in the absence of ligand and the unswiveled state is favored in the presence of ligand. Based on the model fitted by MDFF simulations, we observed a 72° twisting of the RNA away from the β'-ZBD domain in the exit channel upon ligand binding. The twist of the RNA widened the exit channel and provides more space for RNA to escape. The shift of the ZBD away from the exit channel has been observed in both the que-PEC and his-PEC as associated with an unswiveled movement [113, 121]. The communication between the ZBD domain and the fluoride riboswitch was verified through in-vitro transcription reactions with a previously reported ZBD triple mutant [112]. The decreased pause half-life of the ligand bound state of the RNAP mutant shows that the riboswitch structure in the exit channel is no longer stabilized in its terminator form, as the decrease in half-life is related to pause release and read-through.

The MDFF fitted models also suggest that for both ligand bound and unbound systems, there is a tendency to form stem-loop starting at the A55-U60 base pair. However, the tendency is stronger in the absence of ligand, which indicates a more stable stem-loop structure in the absence of ligand and this may result in the stabilization of the terminator hairpin. The binding of ligand then disrupt the terminator hairpin by weakening the base pair A55-U60, as illustrated by the longer distance between the two residue in the ligand bound state, favoring the formation of anti-terminator.

The ligand unbound state favors the swiveled state, and this is validated by in vitro assays in the presence of NusG. Since an increase in the transcription leads to more read-through, preventing NusG binding in the structure promotes termination. The hypothesis is that NusG is unable to bind to the ligand unbound fluoride-PEC and the half-life of U72 is unaffected. Additionally, the observation of a steric clash between the upstream DNA in the ligand unbound fluoride-PEC and the NusG binding site (Figure 4.12) suggests that

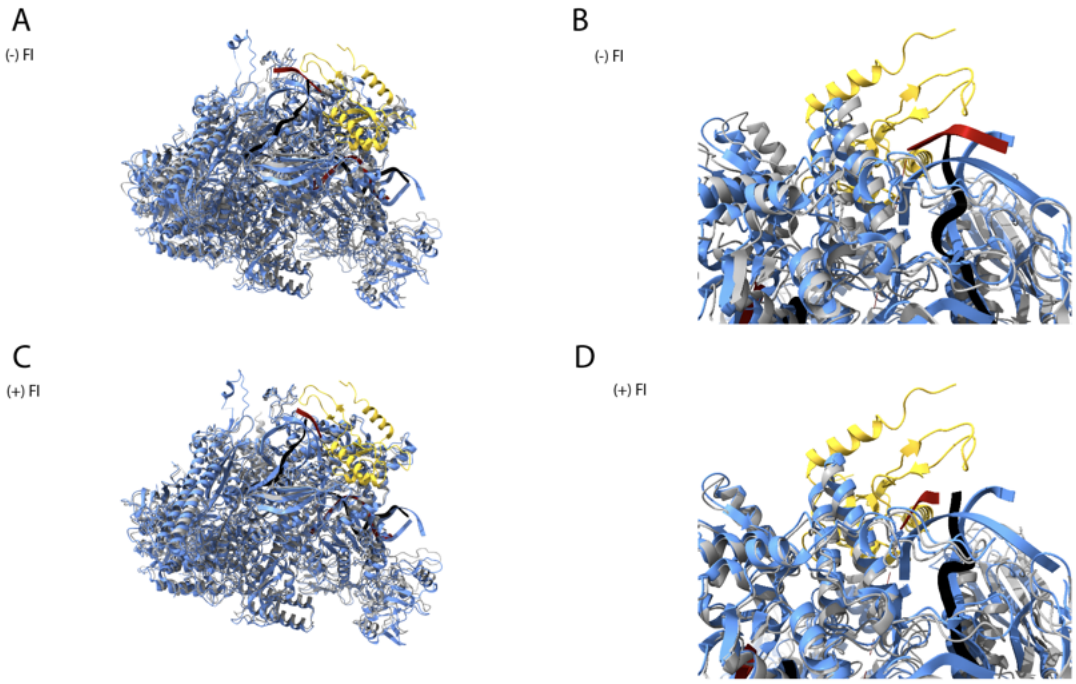


Figure 4.12: Overlaid structure of E. coli RNAP (blue) in complex with NusG (yellow)123 and the fluoride-PEC structure resolved by cryo-EM (gray). (A) NusG binds to the ligand-bound state fluoride-PEC. (B) Zoomed view of NusG near the template/non-template DNA for the ligand-bound state fluoride-PEC. (C) NusG binds to the ligand unbound state fluoride-PEC. (D) Zoomed view of NusG near the template/non-template DNA for the ligand-unbound state fluoride-PEC which shows a steric clash.

the swiveled state of the RNAP stabilized by the ligand-free riboswitch supersedes NusG in determining termination. The conformation of the upstream DNA also shifts upon ligand binding in response to the unswiveling as previously observed [121].This shift of the DNA removes the steric clashes and allows NusG to bind to the RNAP and promote readthrough.

The ligand unbound fluoride-PEC shows pronounced swiveling motion compared to other riboswitches.. NusA has been shown to nearly double the pause half-life of U72 in the fluoride-PEC in the absence of ligand, indicating that the ligand unbound state of the fluoride-PEC favors the swiveled state that NusA preferentially binds [137, 138]. We hypothesize that in the presence of the ligand, the recruitment of transcriptional factors does not need to be as tightly regulated because we do not expect there to be a strongly stabilized formation of terminator structure in the RNA [137]. The pausing behavior observed when transcriptional factors are present show that the fluoride riboswitch shares similarities with a Class I pause. The swiveling motion induced by the transcript without transcription factor is comparable to the Class I pause with NusA.

In summary, the binding of ligand triggers a cascade of events from the movement of the riboswitch and the expansion of exit channel, which then alters the global conformational ensemble of the RNAP towards reversal of the swivel stabilized by the ligand-free riboswitch to regulate gene expression.. Until this alteration of the global state, transcriptional factors like NusG are unable to bind and promote read-through. It was observed that there are less interactions between the riboswitch and the RNAP in the ligand unbound state, suggesting a less stable interaction and disengagement of the nascent RNA and the RNAP. To continue the transcription, the binding of fluoride increases the number of interactions between riboswitch and RNAP, which stabilizes the conformation of RNAP toward the conformation required to pull the RNA out of the exit channel. The interface between the riboswitch and the RNAP may be a promising target developing antibacterial drugs [30].

CHAPTER 5

Summary and Future Outlook

5.1 Summary

RNAs are important mediators and regulators in genetic information transfer and are thus crucial for the function of all biological systems. Therefore, accurately predicting the structures of RNA and understanding its structure-function relationships of RNAs are important for the study of biological processes.

In chapter 2, a method that leverages enhanced molecular dynamics (MD) simulations and machine learning was developed to predict the native-like poses of RNA aptamer-ligand systems. Unbinding profiles were generated from simulation trajectories of five RNA-ligand systems and used to predict the native-like poses. Partial least square regression models were trained using the structures and the simulation-derived parameters to predict the native-like poses for new RNA-ligand systems. The model outperformed rDock in resolving the native-like poses from non-native decoys, providing a fast and robust strategy for pose prediction.

In chapter 3, we investigated the unbinding kinetics of three FMN riboswitch-ligand systems using selectively scaled MD simulations. Contact analysis was conducted and the unbinding rate of the three ligands were calculated. Two distinct unbinding pathways were also identified for different ligands, which enhanced our understanding of the differential activities between the ligands. The methods used in this chapter can be applied to a larger set of RNA-ligand systems to screen the candidate ligand for antibacterial drugs.

In chapter 4, biochemical assays, single particle cryo-electron microscopy, and molecular dynamics flexible fitting were used to investigate the coupling between the fluoride-sensing *crcB* riboswitch and the termination of *E. coli* RNAP transcription. We generated models for the ligand bound and unbound state fluoride riboswitch using the cryo-EM densities and MDFF simulations. We then analyzed the interactions between the riboswitch and the RNAP and the global conformational dynamics of the complex. We then proposed a solely RNA-dependent mechanism for transcription regulation for the fluoride riboswitch. This mechanism is potentially applicable to other riboswitches that do not rely on a consensus pause sequence for the determination of termination or readthrough. Our findings also provide insights into the design of antibiotics targeting the interface of RNAP and the riboswitch, which is considered a promising antibacterial drug target [30, 34]. With deeper understanding of the interactions of RNAP and the riboswitch, is possible to design small molecules that alters the interactions between RNAP and riboswitches in a therapeutically desirable manner.

Overall, the work presented in this dissertation enhanced the understanding of the structure prediction, kinetics, and structure-functional relationships of riboswitch-containing complexes. The methods presented can also be applied to a broader range of systems such as other RNA containing complexes and protein systems.

5.2 Future Directions

This dissertation presented the application of molecular dynamics simulations to the study of structure, kinetics and the mechanisms of riboswitch-containing complexes.

In chapter 2, a robust model was developed for predicting the native-like poses of riboswitch-ligand systems. However, this study only focused on the aptamer of riboswitches and their ligands. Whether similar workflows can be applied to other systems such as microRNA-ligand and protein-ligand systems needs further investigation. Simulations can be carried out on more diverse systems to test if this workflow works as well for different systems. To further improve the model, enhanced MD simulations, such as selectively scaled MD (ssMD) simulations should be conducted on more RNA aptamer-ligand systems when the structures are available, and test whether the characteristic decay time derived from the simulation can still be used to predict the native poses. When the simulation trajectories are available, it is also important to generate unbinding profiles from more RNA-ligand systems to increase the training dataset, in order to generate models with better resolving power. In addition to enlarging the training set, it is worth testing different machine learning frameworks to identify the best-performing model. Methods such as deep neural network and random forest have been applied to study biological system [91, 139], those models can also be trained and benchmarked using the available structures for RNA-ligand or protein-ligand structures. Other than pose prediction, properties such as the binding affinities can be calculated from MD simulations and used to train machine learning models to predict those properties from structure, which can then be applied to virtual screening of candidate drugs.

In chapter 3, selectively scaled MD simulations were applied to accelerate the unbinding of ligand and the contact data generated from the unbinding trajectories was used to estimate the kinetics of unbinding. The binding kinetics of the three systems explains their differential ability in inhibiting bacterial growth. Only three riboswitch-ligand systems were studied in this work. It would be helpful to conduct similar analyses on a broader range of ligands, such as including more ligands that bind to the FMN riboswitch or other riboswitch systems, to

see if the relationship between unbinding kinetics and activity retains for a larger dataset. Besides generating unbinding trajectories for other identified binders of FMN riboswitch, the methods developed in Chapter 2 can also be used to assess the unbinding kinetics of RNA-ligand systems. The simulation-derived parameter τ predicted directly from structure was found to be useful in finding the near-native poses (poses that are more stably bind to the ligand and less likely to dissociate). For the FMN riboswitch, it is helpful to predict the simulation-derived parameters directly from the structures, and then rank the unbinding rate based on the predicted parameter.

In chapter 4, our proposed mechanism indicates an increase in interactions between the riboswitch and the RNAP upon ligand binding, explaining the promotion of readthrough upon ligand binding. However, the interactions were calculated using static models and do not contain much information about the dynamics of the complex. One thing that could be done is to conduct molecular dynamics simulations on the RNA-DNA-RNAP complexes with the attached RNA aptamer, for both ligand-bound and ligand unbound states. Using the trajectories generated from the MD simulations, it would be beneficial to perform contact analysis of the residues of interest to see how the contacts change over time. This analysis can be used to complement the contacts identified through Chimera, providing more insights into the coupling between the ligand binding and the conformational dynamics. For the interactions that are likely to be important for the partition of pausing and read-through, we could mutate the corresponding protein residues and conduct molecular dynamics simulations on the mutated RNAP in complex with the riboswitch. From such simulations, we can then compare the contact change between the wild-type and the mutants in order to test our hypothesis about the importance of those residues.

APPENDIX A

Supporting Information for Chapter 2

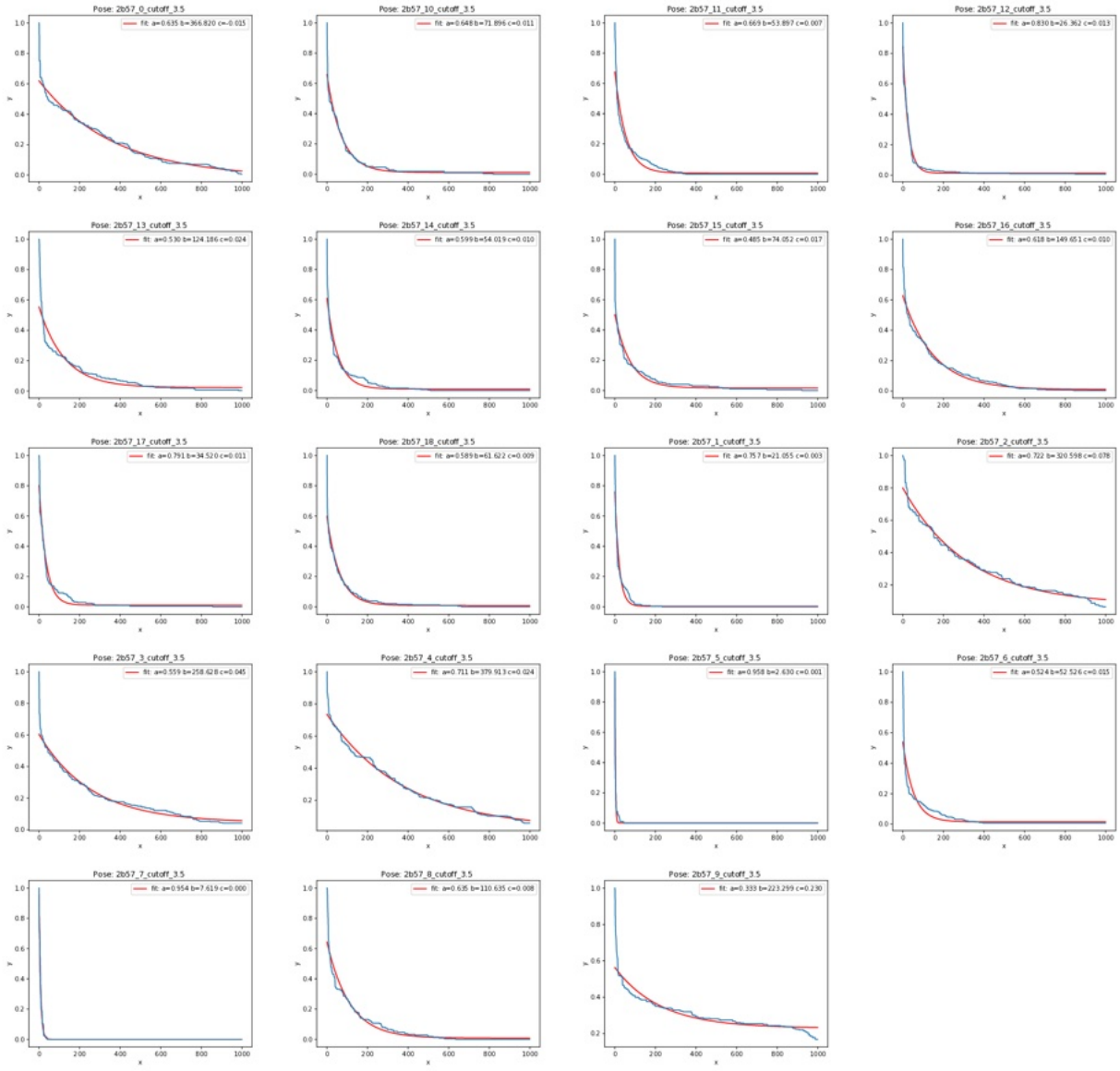


Figure A.1: Plot of the unbinding profile and the exponential curve for each pose of system 2B57. The x-axis corresponds to the timeframe of the trajectory, the y-axis corresponds to the proportion of native contacts that remains. The three parameters A, τ and B correspond to a, b, and c in each plot.

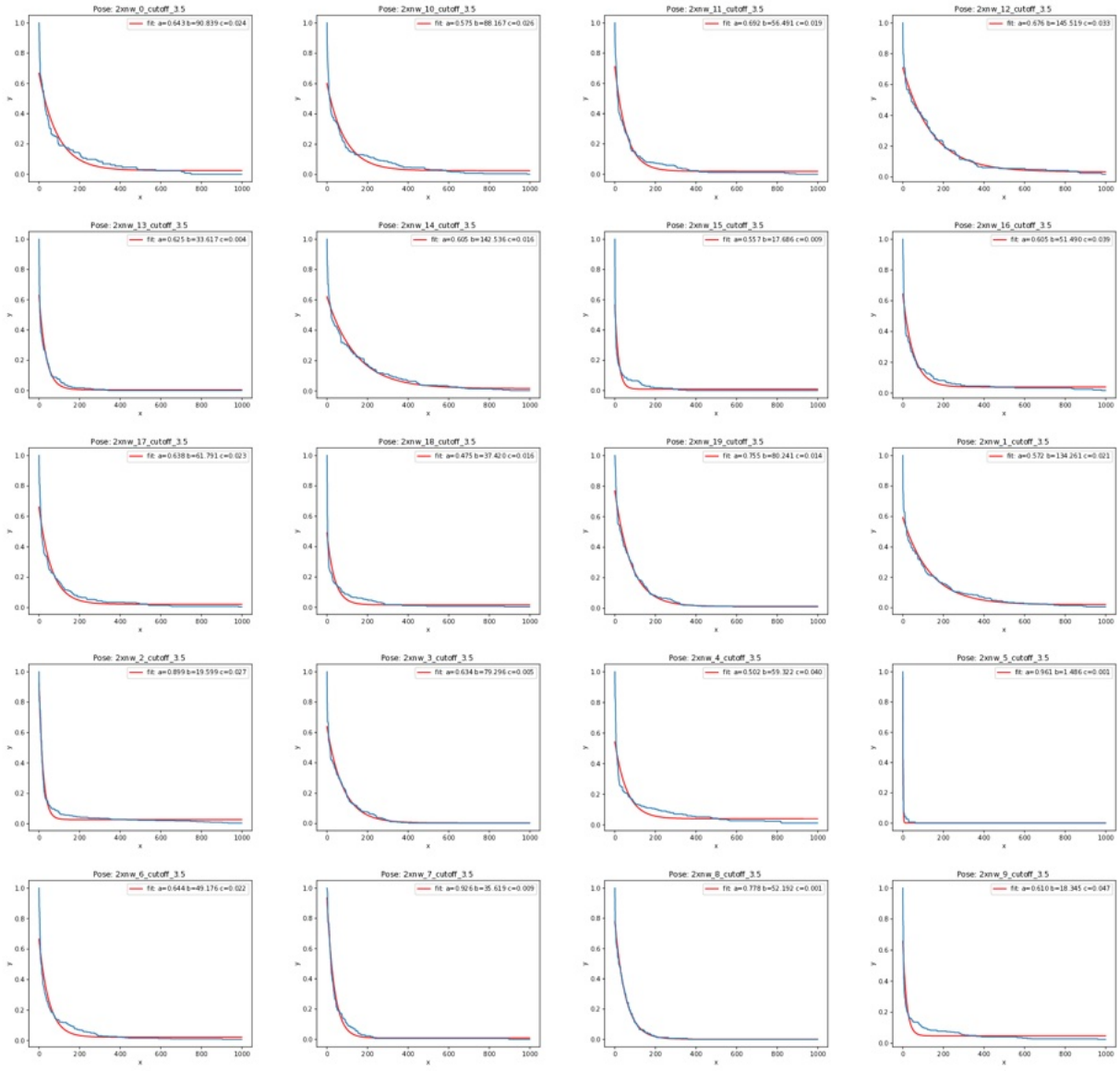


Figure A.2: Plot of the unbinding profile and the exponential curve for each pose of system 2XNW. The x-axis corresponds to the timeframe of the trajectory, the y-axis corresponds to the proportion of native contacts that remains. The three parameters A, τ and B correspond to a, b, and c in each plot.

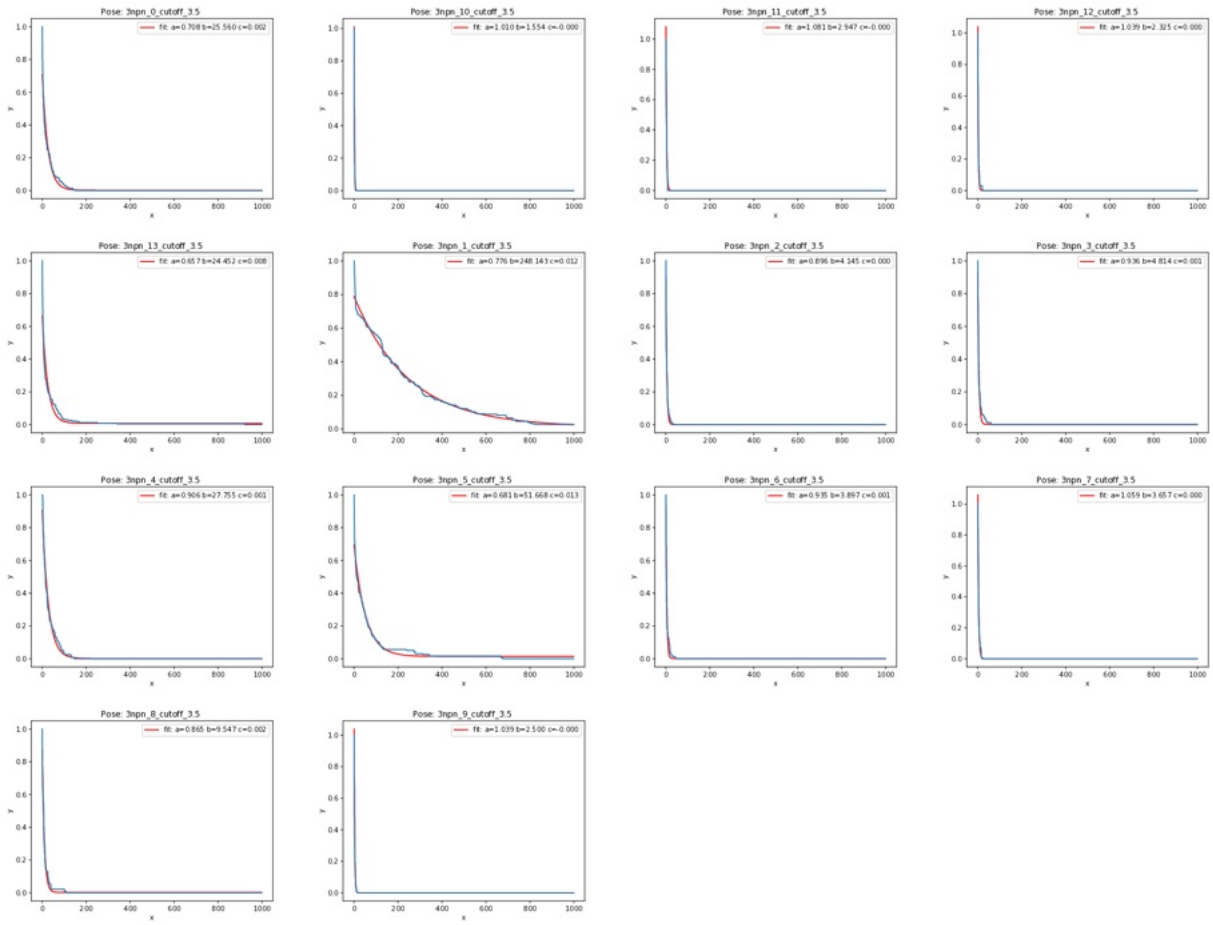


Figure A.3: Plot of the unbinding profile and the exponential curve for each pose of system 3NPN. The x-axis corresponds to the timeframe of the trajectory, the y-axis corresponds to the proportion of native contacts that remains. The three parameters A, τ and B correspond to a, b, and c in each plot.

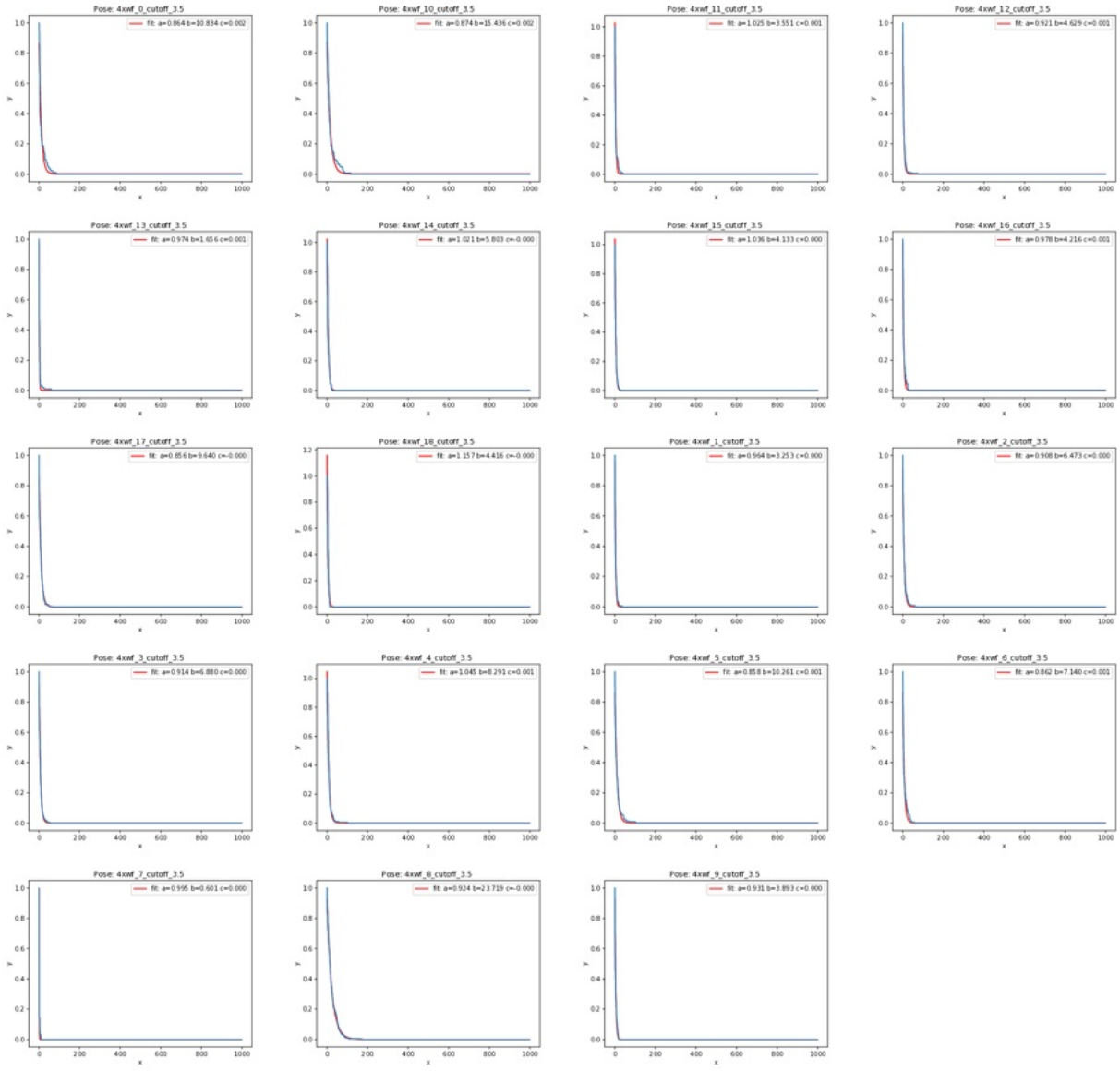


Figure A.4: Plot of the unbinding profile and the exponential curve for each pose of system 4XNW. The x-axis corresponds to the timeframe of the trajectory, the y-axis corresponds to the proportion of native contacts that remains. The three parameters A, τ and B correspond to a, b, and c in each plot.

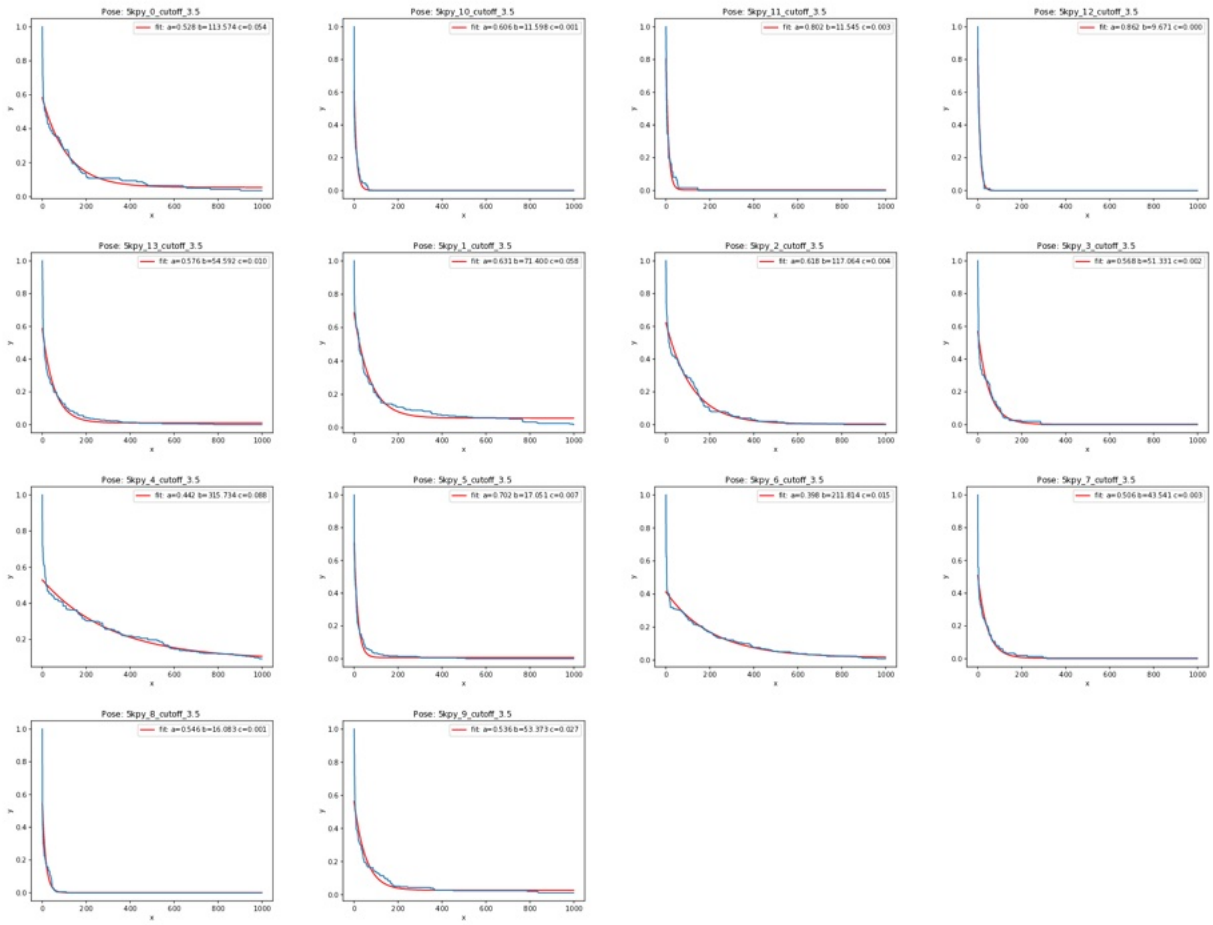


Figure A.5: Plot of the unbinding profile and the exponential curve for each pose of system 4XNW. The x-axis corresponds to the timeframe of the trajectory, the y-axis corresponds to the proportion of native contacts that remains. The three parameters A, τ and B correspond to a, b, and c in each plot.

APPENDIX B

Study of the Undocking Pathway of the Manganese Riboswitch

The content present in Appendix B is done in collaboration with Rajeev Yadav, a former Walter Lab member.

B.1 Introduction

Maintaining the homeostasis of manganese ions is crucial for the survival and virulence of many bacteria [140]. The *yybP-ykoY* manganese (Mn)-sensing riboswitch controls the expression of genes related to the homeostasis of manganese ions [140, 141]. The structure of the Mn riboswitch has been previously characterized [142]. The Mn-sensing riboswitch contains four helices, P1, P2, P3 and P4, respectively. The pairs P1-P2 and P3-P4 form two coaxially stacked superhelices tethered by a four-way junction, docked together through the internal L1 and L3 loops. The L1 and L3 loops are highly conserved, indicating their significance in the proper function of the riboswitch. The Mn riboswitch contains two metal ion binding sites; one specifically binds to Mn^{2+} whereas the other is nonspecific and binds to any divalent ion, including both Mn^{2+} and Mg^{2+}[141]. The secondary structure of the Mn-sensing riboswitch is shown in Figure B.1.

It has been shown that the P1.1 helix of the aptamer domain of the Mn riboswitch is mutually exclusive with the transcription terminator [143], indicating the Mn-sensing riboswitch functions as an ON-switch. The binding of the ligand disrupts the terminator structure and promotes the expression of downstream genes that encode a predicted P-type II ATPase, which exports excess Mn^{2+} ions from *L. lactis* and *S. pneumoniae* [140, 142].

Although the folding of the Mn-sensing riboswitch was identified as following an induced-fit mechanism, where the riboswitch is in its unfolded state, and ligand binding induces folding [144], the undocking pathway of the Mn riboswitch has not yet been characterized. Since the Mn-sensing riboswitch appears in many human and plant pathogens, understanding the mechanism of its undocking is crucial for the design of therapeutics that target bacteria with a Mn-sensing riboswitch. Here, we explore the global conformational dynamics of the Mn-sensing riboswitch from *L. lactis*. Optical tweezers and force spectroscopy experiments were conducted to study the conformational changes of the Mn riboswitch upon undocking. Experimental results showed that for the undocked state of the Mn riboswitch, the distance between the P3.2 and P1.1 increased by $\sim$6.8 nm, and at the transition state, the distance between the two ends has increased by $\sim$3.4 nm. Based on the preliminary results from these experiments, we applied steered molecular dynamics simulations and conventional molecular dynamics simulations to gain a better understanding of the undocking pathway of Mn-sensing riboswitch.

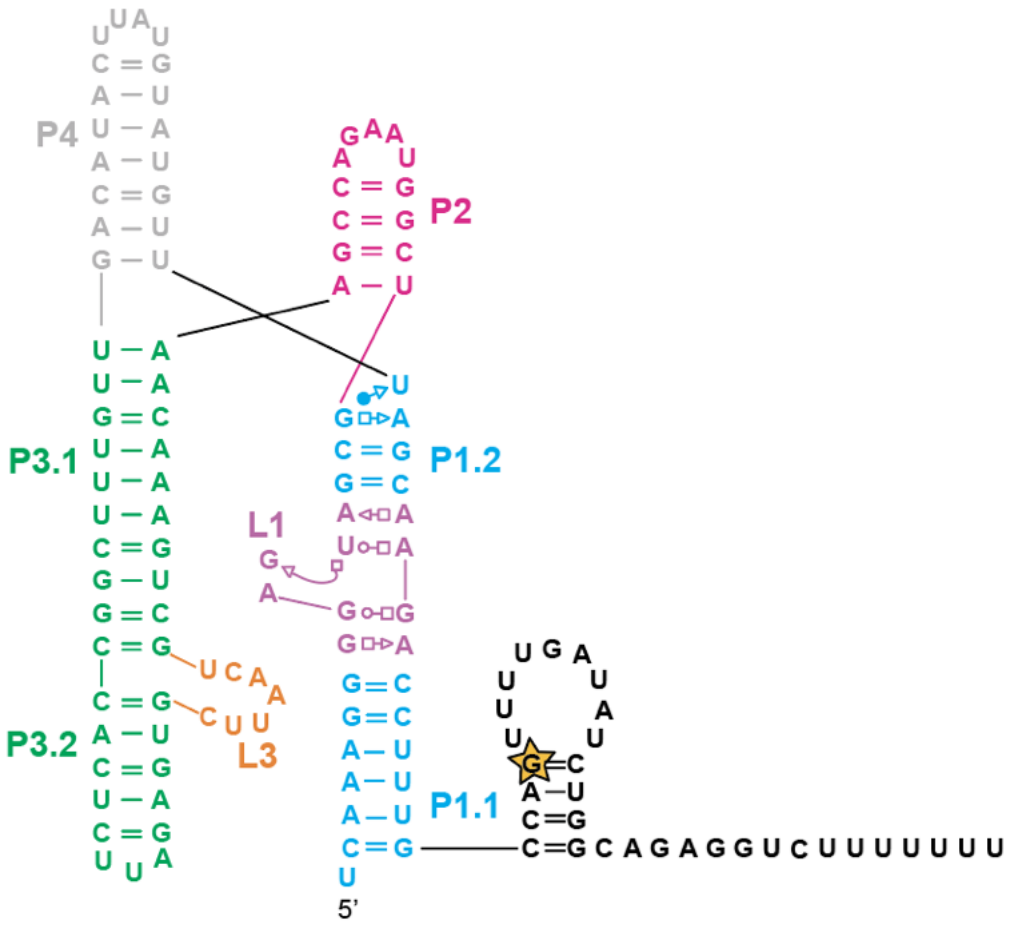


Figure B.1: The secondary structure of the Mn-sensing riboswitch. This figure is generated by Adrien Chauvier. The P1 (blue)-P2 (pink) and P3 (green)-P4 (gray) form two superhelices connected by the L1 (purple) and L3 (orange) loops. The L1 and L3 loops form the binding pocket for metal ions.

B.2 Methods

B.2.1 System preparation

The crystal structure of the Mn-sensing riboswitch was obtained from the Protein Data Bank with the PDB ID: 4Y1I, with a resolution of 2.85 Å [142]. This structure corresponds to the ligand bound state of the Mn riboswitch. The solvent and excess ions not located at the binding pocket were removed from the structure. Only the Mn^{2+} and Mg^{2+} located at the binding pocket were kept for the ligand bound state of the riboswitch. Since the crystal structure corresponds to two identical Mn-riboswitch structures in the crystal lattice, only one was kept for structure preparation. The sequence of this structure was edited in PyMOL[116] to be consistent with the construct used for experiments, which matches the secondary structure and sequence shown in Figure B.1. For the ligand unbound structure, the Mn^{2+} and Mg^{2+} at the binding pocket were removed. The sequence editing involved the addition, deletion, and the mutation of several residues. To connect the residues and remove steric clashes after sequence editing, the coordinate files obtained were submitted to the CHARMM-GUI webserver [83] to generate the input files for energy minimization. The energy minimization was applied with 100 steps of steepest descent and 500 steps of adopted basis Newton-Raphson method with a gradient tolerance of 0.01. The energy minimization was run in vacuum using CHARMM. The structures obtained after minimization were inspected to make sure there are no breaks in the RNA.

The minimized coordinate files for the ligand bound and unbound Mn-sensing riboswitch were then submitted to the CHARMM-GUI webserver [83] to generate the input files for subsequent MD simulations. For the ligand bound system, the topology and parameter files were required for the manganese ion for the successful generation of input files. The topology and parameter files were generated using the parameters obtained from a previous study [145]. The CHARMM36 nucleic acid force field was used to simulate the dynamics of the RNA. The ligand bound and unbound systems were solvated in a cubic TIP3P water

box with a minimum distance of 10 Å between the solute and the boundary of the water box. Sodium and chloride ions were added to neutralize the solvated system.

B.2.2 Molecular dynamics simulations

Each simulation of the Mn riboswitch consists of an equilibration run and a production run. The equilibration run and production run were carried out using the CHARMM/OpenMM interface [86]. The equilibration simulations were carried out under NVT conditions, and the hydrogen atoms were constrained using the SHAKE algorithm [87]. Harmonic restraints were applied to the backbone heavy atoms and the nucleobase heavy atoms with force constant 1.0 kcal/mol/Å^2 and 0.1 kcal/mol/Å^2, respectively. The systems were equilibrated for 2 ns with time step set to 1 fs, so the equilibration simulation consisted of 2,000,000 steps.

Then the equilibrated structures were then used as the starting structures for the production runs. For such a production run, the SHAKE algorithm was used to constrain hydrogen atoms. No harmonic restraints were applied to the heavy atoms during production simulations. NPT conditions were used to carry out the production simulations. For each system (ligand bound and ligand unbound), 20 independent 20-ns simulations were carried out. The time step of the production simulations was set to 2 fs, so a 20 ns simulation consists of 10,000,000 steps. Coordinates of the system were written out to a trajectory file every 5000 steps, so each trajectory contains 2000 frames. The trajectories were used for subsequent analysis.

B.2.3 Steered MD simulations

To mimic the optical tweezers setup, constant velocity steered MD simulations were employed to pull the P1.1 helix of the Mn riboswitch to move the riboswitch to its undocked state. The phosphate atom of the G51 was kept fixed and the phosphate atom of C96 was set

as the SMD atom to be pulled. In constant velocity steered MD simulations, the SMD atom is pulled at a constant speed through a virtual spring. The steered MD simulations were carried out in NAMD [118]. The CHARMM36 nucleic acid force field was used to simulate the dynamics of RNA and the force field parameters obtained from a previous study [145] were used to simulate the manganese ions. The temperature of the simulations was set to 303.15 K. The simulation length was set to 10 ns with a time step of 2 fs, so the total number of steps was 5,000,000. Coordinates of the system were saved to a trajectory file every 10000 steps, so the trajectory contains 500 frames.

Multiple force constants and pulling velocities were tested. To make sure the hypothesized 6.8 nm increase in distance between P3.2 and P1.1 helices was reached in the distance change during the relatively short simulations, we picked the force constant 15 kcal/mol/Å^2 and the velocity 1 nm/ns. The direction of the force is the vector that connects the fixed atom and the SMD atom.

B.2.4 Analysis of the trajectories

To compare the dynamics of the ligand bound and unbound state of the riboswitch, the root mean squared deviations (RMSDs) of each nucleotide were calculated for each frame of the trajectory. The RMSDs were calculated with PyMOL[116]. For both the ligand bound and unbound systems, the average RMSD for each residue as a function of simulation time steps was calculated across the 20 replicas and plotted. We compared the RMSDs for each residue for the ligand bound and unbound riboswitch states and the residues with significant difference in RMSD between the ligand bound and unbound state were identified. For the trajectories obtained from steered MD simulations, the distance between the P3 and P1 helices were measured.

B.3 Results

B.3.1 The P2 helix and the L3 loop are more flexible in the ligand unbound state

The RMSDs of each nucleotide of the Mn-sensing riboswitch were calculated. For each nucleotide, we compared the RMSD in the ligand bound and unbound state. For most of the residues, the average RMSD in the ligand bound and unbound state are comparable. However, several residues were identified to have larger RMSDs in the ligand unbound state of the Mn-sensing riboswitch, suggesting that those residues are more flexible in the ligand unbound state. The residues identified are A22, A23, G24, A25, C26, A43, U44, and U45. Among the eight nucleotides, the A22, A23, G24, A25, and C26 are within the P2 helix, and the residues A43, U44, and U45 are within the L3 loop. Plots of the RMSDs in both states are shown in Figure B.2. The RMSDs for the ligand bound state are colored in red and the RMSDs for the ligand unbound state are colored in blue.

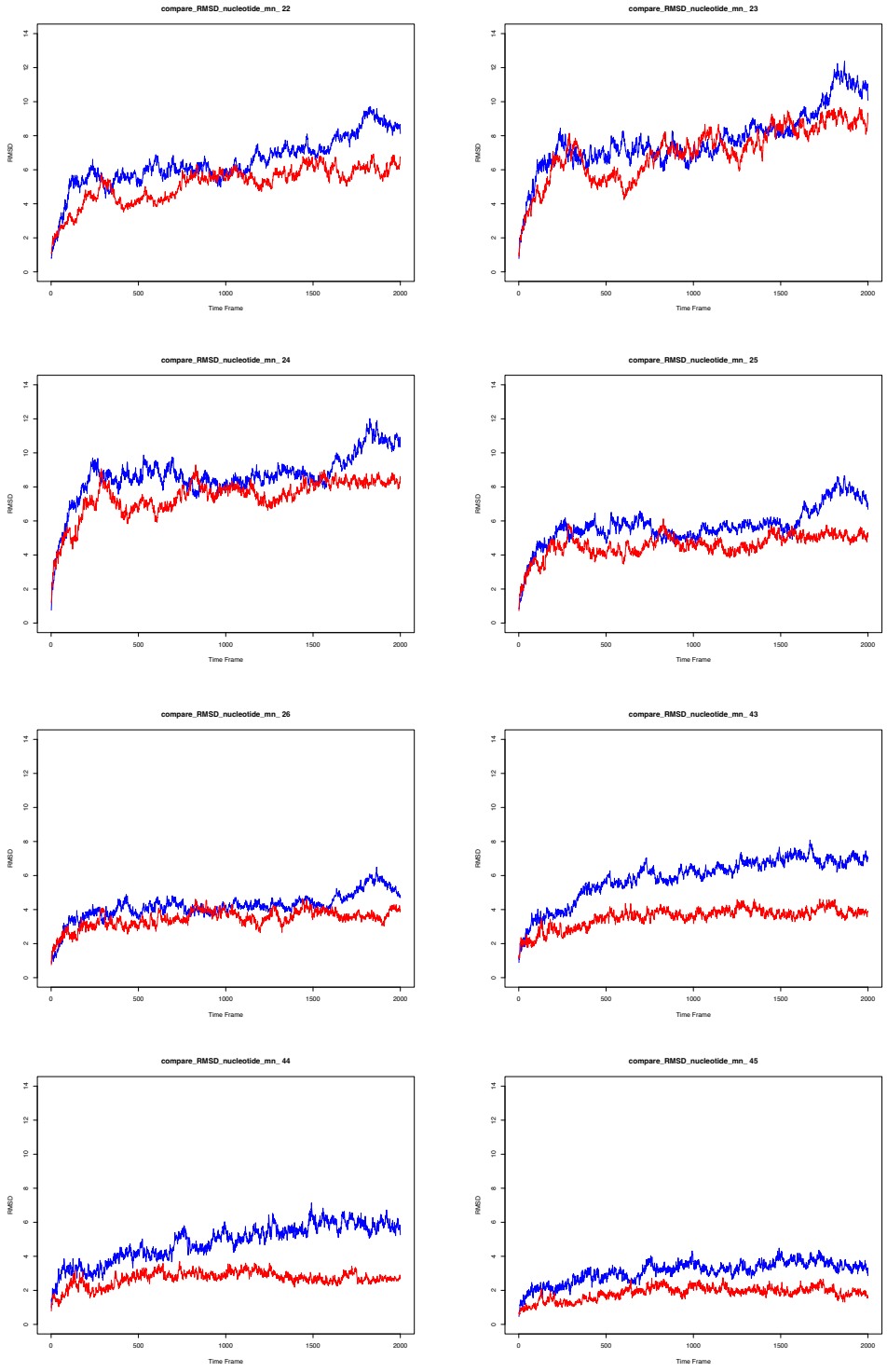


Figure B.2: Plot of the RMSDs of the nucleotides A22, A23, G24, A25, C26, A43, U44, and U45. The RMSDs of the ligand bound riboswitch are colored in red, and the RMSDs for the ligand unbound riboswitch are colored in blue.

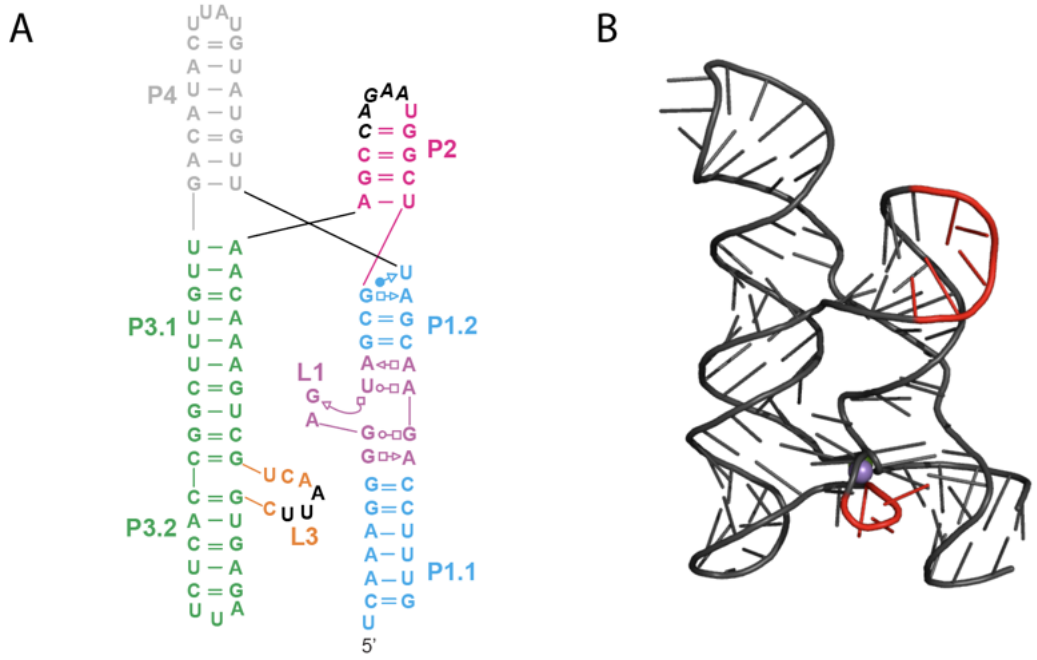


Figure B.3: (A). Secondary structure of the Mn-sensing riboswitch. Residues with larger RMSDs in the ligand unbound state are colored in black. (B). 3D structure of the Mn-sensing riboswitch. The residues with larger RMSDs in the ligand unbound state are colored in red.

B.3.2 Melting of the riboswitch was observed during steered MD simulations

For the trajectory obtained from steered MD simulations, the base pairs in the P1.1 helix were broken, and the helix structure was disrupted. The breaking of the base pairs occurs when the distance between the fixed atom and the SDM atom were increase by around 1.3 nm. At this point, the two metal ions, magnesium and manganese ions were still docked in the binding pocket. When the end-to-end extension between the P3 and P1 increased to around 3.4 nm and 6.8 nm, which corresponds to the predicted extension for the transition state and the undocked state based on force spectroscopy experiments, the Mn^{2+} and Mg^{2+} did not show a tendency to leave the binding pocket either. This suggests that the two metal ions have a strong stabilizing effect for the loop-loop interaction at the binding sites.

B.4 Discussion

The RMSDs calculated from the MD simulations indicate that the nucleotides in the L3 loop show the most significant difference between the ligand bound and ligand unbound states of the Mn riboswitch, which indicates the significance of the highly conserved loop-loop interactions[142] in the selective binding of ligands and maintaining docked state structure. The increase in flexibility of those residues in the unbound state can be explained by the removal of the positively charged metal ligands. In the presence of ligand, the positively charged ligands form stabilizing interactions with the negatively charged RNA backbone. After the removal of the metal ions, the negatively charged RNA backbones of the L1 and L3 loop repel each other to give increased flexibility compared to the ligand bound state. Five nucleotides in the P2 helix also showed increased flexibility in the ligand unbound state (Figure B.3), indicated by the larger RMSDs (Figure B.2). Although the five nucleotides are not in the vicinity of the binding site, the increased flexibility may suggest potential allosteric effects during the ligand unbinding.

The steered MD simulations showed an increase in the distance between helices P3 and P1.1. However, the base pairs in the P1 helix were broken during steered MD simulations. Even when the P1 helix was extended beyond the distance suggested by experiments, the L1 and L3 loops were still intact, preventing undocking. To make it possible to observe the undocking, the parameters for the metal ions could be scaled to reduce the stabilizing effect.

To better understand the undocking pathways, especially the structure of the transition state, starting structures that resemble the proposed transition state could be used to start the MD simulations. Long MD simulations to the microsecond time scale could also be carried out for the ligand bound and unbound systems, which may allow the observation of a conformational change due to the unbinding of the metal ions. Since the steered MD trajectories show that the interaction between the metal ligand and the loop structures of the riboswitch is too strong to allow unbinding, it would be helpful to conduct steered MD simulations for structures without ligand, as it can reduced the interactions at the L1 and

L3 loop and provide structural information about the dynamics during unbinding.

BIBLIOGRAPHY

[1] Francis Crick. Central dogma of molecular biology. *Nature*, 227(5258):561–563, 1970.

[2] John S Mattick and Igor V Makunin. Non-coding rna. *Human molecular genetics*, 15(suppl_1):R17–R29, 2006.

[3] Fabrício F Costa. Non-coding rnas: meet thy masters. *Bioessays*, 32(7):599–608, 2010.

[4] IHGSC. International human genome sequencing consortium. *Nature*, 431:931, 2004.

[5] Baby Santosh, Akhil Varshney, and Pramod Kumar Yadava. Non-coding rnas: biological functions and applications. *Cell biochemistry and function*, 33(1):14–22, 2015.

[6] David P Bartel. Micrornas: genomics, biogenesis, mechanism, and function. *cell*, 116(2):281–297, 2004.

[7] Scott M Hammond. An overview of micrornas. *Advanced drug delivery reviews*, 87:3–14, 2015.

[8] Richard Giegé. Toward a more complete view of trna biology. *Nature structural & molecular biology*, 15(10):1007–1014, 2008.

[9] Armen R Karapetyan, Coen Buiting, Renske A Kuiper, and Marcel W Coolen. Regulatory roles for long ncrna and mrna. *Cancers*, 5(2):462–490, 2013.

[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] Julia D Ransohoff, Yuning Wei, and Paul A Khavari. The functions and unique features of long intergenic non-coding rna. *Nature reviews Molecular cell biology*, 19(3):143–157, 2018.

[12] Wayne A Decatur and Maurille J Fournier. rrna modifications and ribosome function. *Trends in biochemical sciences*, 27(7):344–351, 2002.

[13] Heath D Murray, David A Schneider, and Richard L Gourse. Control of rrna expression by small molecules is dynamic and nonredundant. *Molecular cell*, 12(1):125–134, 2003.

[14] Brian J Paul, Wilma Ross, Tamas Gaal, and Richard L Gourse. rrna transcription in escherichia coli. *Annu. Rev. Genet.*, 38:749–770, 2004.

[15] Sean R Eddy. Non–coding rna genes and the modern rna world. *Nature Reviews Genetics*, 2(12):919–929, 2001.

[16] Thomas X Lu and Marc E Rothenberg. Microrna. *Journal of allergy and clinical immunology*, 141(4):1202–1207, 2018.

[17] N Bushati and SM Cohen. microrna functions. annu rev cell dev biol. 2007.

[18] Stefanie Jonas and Elisa Izaurralde. Towards a molecular understanding of microrna-mediated gene silencing. *Nature reviews genetics*, 16(7):421–433, 2015.

[19] Minju Ha and V Narry Kim. Regulation of microrna biogenesis. *Nature reviews Molecular cell biology*, 15(8):509–524, 2014.

[20] Sayda M Elbashir, Javier Martinez, Agnieszka Patkaniowska, Winfried Lendeckel, and Thomas Tuschl. Functional anatomy of sirnas for mediating efficient rnai in drosophila melanogaster embryo lysate. *The EMBO journal*, 20(23):6877–6888, 2001.

[21] Derek M Dykxhoorn, Carl D Novina, and Phillip A Sharp. Killing the messenger: short rnas that silence gene expression. *Nature reviews Molecular cell biology*, 4(6):457–467, 2003.

[22] Martin D Jansson and Anders H Lund. Microrna and cancer. *Molecular oncology*, 6(6):590–610, 2012.

[23] Kelly Kruger, Paula J Grabowski, Arthur J Zaug, Julie Sands, Daniel E Gottschling, and Thomas R Cech. Self-splicing rna: autoexcision and autocyclization of the ribosomal rna intervening sequence of tetrahymena. *cell*, 31(1):147–157, 1982.

[24] Martha J Fedor and James R Williamson. The catalytic diversity of rnas. *Nature reviews Molecular cell biology*, 6(5):399–412, 2005.

[25] Philip C Bevilacqua and Rieko Yajima. Nucleobase catalysis in ribozyme mechanism. *Current opinion in chemical biology*, 10(5):455–464, 2006.

[26] Randi M Jimenez, Julio A Polanco, and Andrej Lupták. Chemistry and biology of self-cleaving ribozymes. *Trends in biochemical sciences*, 40(11):648–661, 2015.

[27] Maumita Mandal and Ronald R Breaker. Gene regulation by riboswitches. *Nature reviews Molecular cell biology*, 5(6):451–463, 2004.

[28] Ronald R Breaker. Riboswitches and the rna world. *Cold Spring Harbor perspectives in biology*, 4(2):a003566, 2012.

[29] Andrew D Garst, Andrea L Edwards, and Robert T Batey. Riboswitches: structures and mechanisms. *Cold Spring Harbor perspectives in biology*, 3(6):a003533, 2011.

[30] Emily Ellinger, Adrien Chauvier, Rosa A Romero, Yichen Liu, Sujay Ray, and Nils G Walter. Riboswitches as therapeutic targets: promise of a new era of antibiotics. *Expert Opinion on Therapeutic Targets*, 27(6):433–445, 2023.

[31] Alexander Serganov and Evgeny Nudler. A decade of riboswitches. *Cell*, 152(1):17–24, 2013.

[32] Laurène Bastet, Adrien Chauvier, Navjot Singh, Antony Lussier, Anne-Marie Lamontagne, Karine Prévost, Eric Massé, Joseph T Wade, and Daniel A Lafontaine. Translational control and rho-dependent transcription termination are intimately linked in riboswitch regulation. *Nucleic acids research*, 45(12):7474–7486, 2017.

[33] Aiming Ren, Kanagalaghatta R Rajashankar, and Dinshaw J Patel. Fluoride ion encapsulation by mg2+ ions and phosphates in a fluoride riboswitch. *Nature*, 486(7401):85–89, 2012.

[34] Vipul Panchal and Ruth Brenk. Riboswitches as drug targets for antibiotics. *Antibiotics*, 10(1):45, 2021.

[35] J Kenneth Wickiser, Wade C Winkler, Ronald R Breaker, and Donald M Crothers. The speed of rna transcription and metabolite binding kinetics operate an fmn riboswitch. *Molecular cell*, 18(1):49–60, 2005.

[36] Elaine R Lee, Kenneth F Blount, and Ronald R Breaker. Roseoflavin is a natural antibacterial compound that binds to fmn riboswitches and regulates gene expression. *RNA biology*, 6(2):187–194, 2009.

[37] Kenneth F Blount and Ronald R Breaker. Riboswitches as antibacterial drug targets. *Nature biotechnology*, 24(12):1558–1564, 2006.

[38] David W Staple and Samuel E Butcher. Pseudoknots: Rna structures with diverse functions. *PLoS biology*, 3(6):e213, 2005.

[39] MS Smyth and JHJ Martin. x ray crystallography. *Molecular Pathology*, 53(1):8, 2000.

[40] SH Kim, FL Suddath, GJ Quigley, A McPherson, JL Sussman, AHJ Wang, NC Seeman, and Alexander Rich. Three-dimensional tertiary structure of yeast phenylalanine transfer rna. *Science*, 185(4149):435–440, 1974.

[41] Ryland W Jackson, Claire M Smathers, and Aaron R Robart. General strategies for rna x-ray crystallography. *Molecules*, 28(5):2111, 2023.

[42] Boris Fürtig, Christian Richter, Jens Wöhnert, and Harald Schwalbe. Nmr spectroscopy of rna. *ChemBioChem*, 4(10):936–962, 2003.

[43] Almudena Ponce-Salvatierra, Astha, Katarzyna Merdas, Chandran Nithin, Pritha Ghosh, Sunandan Mukherjee, and Janusz M Bujnicki. Computational modeling of rna 3d structure based on experimental data. *Bioscience reports*, 39(2):BSR20180430, 2019.

[44] Hong-Wei Wang and Jia-Wei Wang. How cryo-electron microscopy and x-ray crystallography complement each other. *Protein Science*, 26(1):32–39, 2017.

[45] Haiyun Ma, Xinyu Jia, Kaiming Zhang, and Zhaoming Su. Cryo-em advances in rna structure determination. *Signal Transduction and Targeted Therapy*, 7(1):58, 2022.

[46] Kalli Kappel, Kaiming Zhang, Zhaoming Su, Andrew M Watkins, Wipapat Kladwang, Shanshan Li, Grigore Pintilie, Ved V Topkar, Ramya Rangan, Ivan N Zheludev, et al. Accelerated cryo-em-guided determination of three-dimensional rna-only structures. *Nature methods*, 17(7):699–707, 2020.

[47] Martin Karplus and Gregory A Petsko. Molecular dynamics simulations in biology. *Nature*, 347(6294):631–639, 1990.

[48] Kurt Binder, Jürgen Horbach, Walter Kob, Wolfgang Paul, and Fathollah Varnik. Molecular dynamics simulations. *Journal of Physics: Condensed Matter*, 16(5):S429, 2004.

[49] Scott A Hollingsworth and Ron O Dror. Molecular dynamics simulation for all. *Neuron*, 99(6):1129–1143, 2018.

[50] Sereina Riniker, Jane R Allison, and Wilfred F van Gunsteren. On developing coarse-grained models for biomolecular simulation: a review. *Physical Chemistry Chemical Physics*, 14(36):12423–12430, 2012.

[51] Soumil Y Joshi and Sanket A Deshmukh. A review of advancements in coarse-grained molecular dynamics simulations. *Molecular Simulation*, 47(10-11):786–803, 2021.

[52] Anuj Chaudhri, Isidro E Zarraga, Tim J Kamerzell, J Paul Brandt, Thomas W Patapoff, Steven J Shire, and Gregory A Voth. Coarse-grained modeling of the self-association of therapeutic monoclonal antibodies. *The Journal of Physical Chemistry B*, 116(28):8045–8057, 2012.

[53] John MA Grime, James F Dama, Barbie K Ganser-Pornillos, Cora L Woodward, Grant J Jensen, Mark Yeager, and Gregory A Voth. Coarse-grained simulation reveals key features of hiv-1 capsid self-assembly. *Nature communications*, 7(1):11568, 2016.

[54] Olgun Guvench and Alexander D MacKerell. Comparison of protein force fields for molecular dynamics simulations. *Molecular modeling of proteins*, pages 63–88, 2008.

[55] Rafael C Bernardi, Marcelo CR Melo, and Klaus Schulten. Enhanced sampling techniques in molecular dynamics simulations of biological systems. *Biochimica et Biophysica Acta (BBA)-General Subjects*, 1850(5):872–877, 2015.

[56] Aaron T Frank and Ioan Andricioaei. Reaction coordinate-free approach to recovering kinetics from potential-scaled simulations: application of kramers' rate theory. *The Journal of Physical Chemistry B*, 120(33):8600–8605, 2016.

[57] Hideki Tsujishita, Ikuo Moriguchi, and Shuichi Hirono. Potential-scaled molecular dynamics and potential annealing: effective conformational search techniques for biomolecules. *The Journal of Physical Chemistry*, 97(17):4416–4420, 1993.

[58] Sergei Izrailev, Sergey Stepaniants, Barry Isralewitz, Dorina Kosztin, Hui Lu, Ferenc Molnar, Willy Wriggers, and Klaus Schulten. Steered molecular dynamics. In *Computational Molecular Dynamics: Challenges, Methods, Ideas: Proceedings of the 2nd International Symposium on Algorithms for Macromolecular Modelling, Berlin, May 21–24, 1997*, pages 39–65. Springer, 1999.

[59] Indrajit Deb and Aaron T Frank. Accelerating rare dissociative processes in biomolecules using selectively scaled md simulations. *Journal of chemical theory and computation*, 15(11):5817–5828, 2019.

[60] Doris A Schuetz, Mattia Bernetti, Martina Bertazzo, Djordje Musil, Hans-Michael Eggenweiler, Maurizio Recanatini, Matteo Masetti, Gerhard F Ecker, and Andrea Cavalli. Predicting residence time and drug unbinding pathway through scaled molecular dynamics. *Journal of Chemical Information and Modeling*, 59(1):535–549, 2018.

[61] Phuc-Chau Do, Eric H Lee, and Ly Le. Steered molecular dynamics simulation in rational drug design. *Journal of Chemical Information and Modeling*, 58(8):1473–1482, 2018.

[62] Barry Isralewitz, Mu Gao, and Klaus Schulten. Steered molecular dynamics and mechanical functions of proteins. *Current opinion in structural biology*, 11(2):224–230, 2001.

[63] Aaron T Frank, Sean M Law, and Charles L Brooks III. A simple and fast approach for predicting 1h and 13c chemical shifts: toward chemical shift-guided simulations of rna. *The Journal of Physical Chemistry B*, 118(42):12168–12175, 2014.

[64] Aaron T Frank, Sean M Law, Logan S Ahlstrom, and Charles L Brooks III. Predicting protein backbone chemical shifts from cα coordinates: extracting high resolution experimental observables from low resolution models. *Journal of Chemical Theory and Computation*, 11(1):325–331, 2015.

[65] Yichen Liu and Aaron T Frank. Using selectively scaled molecular dynamics simulations to assess ligand poses in rna aptamers. *Journal of Chemical Theory and Computation*, 18(9):5703–5709, 2022.

[66] Shuai Wei, Charles L Brooks III, and Aaron T Frank. A rapid solvent accessible surface area estimator for coarse grained molecular simulations. *Journal of Computational Chemistry*, 38(15):1270–1274, 2017.

[67] Hasane Ratni, Martin Ebeling, John Baird, Stefanie Bendels, Johan Bylund, Karen S Chen, Nora Denk, Zhihua Feng, Luke Green, Melanie Guerard, et al. Discovery of risdiplam, a selective survival of motor neuron-2 (smn2) gene splicing modifier for the treatment of spinal muscular atrophy (sma), 2018.

[68] John A Howe, Hao Wang, Thierry O Fischmann, Carl J Balibar, Li Xiao, Andrew M Galgoci, Juliana C Malinverni, Todd Mayhood, Artjohn Villafania, Ali Nahvi, et al. Selective small-molecule inhibition of an rna structural element. *Nature*, 526(7575):672–677, 2015.

[69] Olga Fedorova, G Erik Jagdmann Jr, Rebecca L Adams, Lin Yuan, Michael C Van Zandt, and Anna Marie Pyle. Small molecules that target group ii introns are potent antifungal agents. *Nature chemical biology*, 14(12):1073–1078, 2018.

[70] Colleen M Connelly, Tomoyuki Numata, Robert E Boer, Michelle H Moon, Ranu S Sinniah, Joseph J Barchi, Adrian R Ferré-D'Amaré, and John S Schneekloth Jr. Synthetic ligands for preq1 riboswitches provide structural and mechanistic insights into targeting rna tertiary structure. *Nature communications*, 10(1):1501, 2019.

[71] Laura R Ganser, Megan L Kelly, Daniel Herschlag, and Hashim M Al-Hashimi. The roles of structural dynamics in the cellular functions of rnas. *Nature reviews Molecular cell biology*, 20(8):474–489, 2019.

[72] Peter Daldrop, Francis E Reyes, David A Robinson, Colin M Hammond, David M Lilley, Robert T Batey, and Ruth Brenk. Novel ligands for a purine riboswitch discovered by rna-ligand docking. *Chemistry & biology*, 18(3):324–335, 2011.

[73] Andrew C Stelzer, Aaron T Frank, Jeremy D Kratz, Michael D Swanson, Marta J Gonzalez-Hernandez, Janghyun Lee, Ioan Andricioaei, David M Markovitz, and Hashim M Al-Hashimi. Discovery of selective bioactive small molecules by targeting an rna dynamic ensemble. *Nature chemical biology*, 7(8):553–559, 2011.

[74] Andrea Cavalli, Giovanni Bottegoni, Caterina Raco, Marco De Vivo, and Maurizio Recanatini. A computational study of the binding of propidium to the peripheral anionic site of human acetylcholinesterase. *Journal of medicinal chemistry*, 47(16):3991–3999, 2004.

[75] Hernan Alonso, Andrey A Bliznyuk, and Jill E Gready. Combining docking and molecular dynamic simulations in drug design. *Medicinal research reviews*, 26(5):531–568, 2006.

[76] Andrea Spitaleri, Sergio Decherchi, Andrea Cavalli, and Walter Rocchia. Fast dynamic docking guided by adaptive electrostatic bias: The md-binding approach. *Journal of chemical theory and computation*, 14(3):1727–1736, 2018.

[77] Takako Sakano, Md Iqbal Mahamood, Takefumi Yamashita, and Hideaki Fujitani. Molecular dynamics analysis to evaluate docking pose prediction. *Biophysics and physicobiology*, 13:181–194, 2016.

[78] Laura J Kingsley, Juan Esquivel-Rodríguez, Ying Yang, Daisuke Kihara, and Markus A Lill. Ranking protein–protein docking results using steered molecular dynamics and potential of mean force calculations. *Journal of computational chemistry*, 37(20):1861–1865, 2016.

[79] Nicola Scafuri, Miguel A Soler, Andrea Spitaleri, and Walter Rocchia. Enhanced molecular dynamics method to efficiently increase the discrimination capability of computational protein–protein docking. *Journal of Chemical Theory and Computation*, 17(11):7271–7280, 2021.

[80] Daria B Kokh, Tom Kaufmann, Bastian Kister, and Rebecca C Wade. Machine learning analysis of τramd trajectories to decipher molecular determinants of drug-target residence times. *Frontiers in molecular biosciences*, 6:36, 2019.

[81] 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.

[82] Michael Feig, John Karanicolas, and Charles L Brooks III. Mmtsb tool set: enhanced sampling and multiscale modeling methods for applications in structural biology. *Journal of Molecular Graphics and Modelling*, 22(5):377–395, 2004.

[83] Sunhwan Jo, Taehoon Kim, Vidyashankara G Iyer, and Wonpil Im. Charmm-gui: a web-based graphical user interface for charmm. *Journal of computational chemistry*, 29(11):1859–1865, 2008.

[84] Kenno Vanommeslaeghe, E Prabhu Raman, and Alexander D MacKerell Jr. Automation of the charmm general force field (cgenff) ii: assignment of bonded parameters and partial atomic charges. *Journal of chemical information and modeling*, 52(12):3155–3168, 2012.

[85] Bernard R Brooks, Charles L Brooks III, Alexander D Mackerell Jr, Lennart Nilsson, Robert J Petrella, Benoît Roux, Youngdo Won, Georgios Archontis, Christian Bartels, Stefan Boresch, et al. Charmm: the biomolecular simulation program. *Journal of computational chemistry*, 30(10):1545–1614, 2009.

[86] Peter Eastman, Jason Swails, John D Chodera, Robert T McGibbon, Yutong Zhao, Kyle A Beauchamp, Lee-Ping Wang, Andrew C Simmonett, Matthew P Harrigan, Chaya D Stern, et al. Openmm 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS computational biology*, 13(7):e1005659, 2017.

[87] Jean-Paul Ryckaert, Giovanni Ciccotti, and Herman JC Berendsen. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of computational physics*, 23(3):327–341, 1977.

[88] Ron Wehrens and B-H Mevik. The pls package: principal component and partial least squares regression in r. 2007.

[89] Sahil Chhabra, Jingru Xie, and Aaron T Frank. Rnaposers: machine learning classifiers for ribonucleic acid–ligand poses. *The Journal of Physical Chemistry B*, 124(22):4436–4445, 2020.

[90] Yuyu Feng and Sheng-You Huang. Itscore-nl: an iterative knowledge-based scoring function for nucleic acid–ligand interactions. *Journal of Chemical Information and Modeling*, 60(12):6698–6708, 2020.

[91] Filip Stefaniak and Janusz M Bujnicki. Annapurna: A scoring function for predicting rna-small molecule binding poses. *PLoS computational biology*, 17(2):e1008309, 2021.

[92] Ali Nahvi, Narasimhan Sudarsan, Margaret S Ebert, Xiang Zou, Kenneth L Brown, and Ronald R Breaker. Genetic control by a metabolite binding mrna. *Chemistry & biology*, 9(9):1043–1049, 2002.

[93] Wade C Winkler, Smadar Cohen-Chalamish, and Ronald R Breaker. An mrna structure that controls gene expression by binding fmn. *Proceedings of the National Academy of Sciences*, 99(25):15908–15913, 2002.

[94] Noreen F Rizvi, John A Howe, Ali Nahvi, Daniel J Klein, Thierry O Fischmann, Hai-Young Kim, Mark A McCoy, Scott S Walker, Alan Hruza, Matthew P Richards, et al. Discovery of selective rna-binding small molecules by affinity-selection mass spectrometry. *ACS chemical biology*, 13(3):820–831, 2018.

[95] Quentin Vicens, Estefanía Mondragón, and Robert T Batey. Molecular sensing by the aptamer domain of the fmn riboswitch: a general model for ligand binding by conformational selection. *Nucleic acids research*, 39(19):8586–8598, 2011.

[96] Jumin Lee, Xi Cheng, Sunhwan Jo, Alexander D MacKerell, Jeffery B Klauda, and Wonpil Im. Charmm-gui input generator for namd, gromacs, amber, openmm, and charmm/openmm simulations using the charmm36 additive force field. *Biophysical journal*, 110(3):641a, 2016.

[97] Elizabeth J Denning, U Deva Priyakumar, Lennart Nilsson, and Alexander D Mackerell Jr. Impact of 2-hydroxyl sampling on the conformational properties of rna: update of the charmm all-atom additive force field for rna. *Journal of computational chemistry*, 32(9):1929–1943, 2011.

[98] William L Jorgensen, Jayaraman Chandrasekhar, Jeffry D Madura, Roger W Impey, and Michael L Klein. Comparison of simple potential functions for simulating liquid water. *The Journal of chemical physics*, 79(2):926–935, 1983.

[99] Tom Darden, Darrin York, and Lee Pedersen. Particle mesh ewald: An n log (n) method for ewald sums in large systems. *The Journal of chemical physics*, 98(12):10089–10092, 1993.

[100] Rosalind J Allen, Patrick B Warren, and Pieter Rein Ten Wolde. Sampling rare switching events in biochemical networks. *Physical review letters*, 94(1):018104, 2005.

[101] Karline Soetaert, Thomas Petzoldt, and R Woodrow Setzer. Solving differential equations in r: package desolve. *Journal of statistical software*, 33:1–25, 2010.

[102] Kumari Kavita and Ronald R Breaker. Discovering riboswitches: the past and the future. *Trends in biochemical sciences*, 48(2):119–141, 2023.

[103] Sha Gong, Yanli Wang, Zhen Wang, and Wenbing Zhang. Co-transcriptional folding and regulation mechanisms of riboswitches. *Molecules*, 22(7):1169, 2017.

[104] Catherine E Scull, Shiba S Dandpat, Rosa A Romero, and Nils G Walter. Transcriptional riboswitches integrate timescales for bacterial gene expression control. *Frontiers in molecular biosciences*, 7:607158, 2021.

[105] Adrien Chauvier and Nils G Walter. Regulation of bacterial gene expression by non-coding rna: It is all about time! *Cell Chemical Biology*, 2024.

[106] Matthew H Larson, Rachel A Mooney, Jason M Peters, Tricia Windgassen, Dhananjaya Nayak, Carol A Gross, Steven M Block, William J Greenleaf, Robert Landick, and Jonathan S Weissman. A pause sequence enriched at translation start sites drives transcription dynamics in vivo. *Science*, 344(6187):1042–1047, 2014.

[107] Tao Pan and Tobin Sosnick. Rna folding during transcription. *Annu. Rev. Biophys. Biomol. Struct.*, 35:161–175, 2006.

[108] Irina Artsimovitch and Robert Landick. The transcriptional regulator rfah stimulates rna chain synthesis after recruitment to elongation complexes by the exposed nontemplate dna strand. *Cell*, 109(2):193–203, 2002.

[109] Ivan Gusarov and Evgeny Nudler. The mechanism of intrinsic transcription termination. *Molecular cell*, 3(4):495–504, 1999.

[110] Surajit Chatterjee, Adrien Chauvier, Shiba S Dandpat, Irina Artsimovitch, and Nils G Walter. A translational riboswitch coordinates nascent transcription–translation coupling. *Proceedings of the National Academy of Sciences*, 118(16):e2023426118, 2021.

[111] Cathleen L Chan and Robert Landick. Dissection of the his leader pause site by base substitution reveals a multipartite signal that includes a pause rna hairpin. *Journal of molecular biology*, 233(1):25–42, 1993.

[112] Julia R Widom, Yuri A Nedialkov, Victoria Rai, Ryan L Hayes, Charles L Brooks, Irina Artsimovitch, and Nils G Walter. Ligand modulates cross-coupling between riboswitch folding and transcriptional pausing. *Molecular cell*, 72(3):541–552, 2018.

[113] Adrien Chauvier, Jason C Porta, Indrajit Deb, Emily Ellinger, Katarina Meze, Aaron T Frank, Melanie D Ohi, and Nils G Walter. Structural basis for control of bacterial rna polymerase pausing by a riboswitch and its ligand. *Nature Structural & Molecular Biology*, 30(7):902–913, 2023.

[114] Ryan McGreevy, Ivan Teo, Abhishek Singharoy, and Klaus Schulten. Advances in the molecular dynamics flexible fitting method for cryo-em modeling. *Methods*, 100:50–60, 2016.

[115] Leonardo G Trabuco, Elizabeth Villa, Eduard Schreiner, Christopher B Harrison, and Klaus Schulten. Molecular dynamics flexible fitting: a practical guide to combine cryo-electron microscopy and x-ray crystallography. *Methods*, 49(2):174–180, 2009.

[116] Warren L DeLano et al. Pymol: An open-source molecular graphics tool. *CCP4 Newsl. Protein Crystallogr*, 40(1):82–92, 2002.

[117] William Humphrey, Andrew Dalke, and Klaus Schulten. Vmd: visual molecular dynamics. *Journal of molecular graphics*, 14(1):33–38, 1996.

[118] James C Phillips, Rosemary Braun, Wei Wang, James Gumbart, Emad Tajkhorshid, Elizabeth Villa, Christophe Chipot, Robert D Skeel, Laxmikant Kale, and Klaus Schulten. Scalable molecular dynamics with namd. *Journal of computational chemistry*, 26(16):1781–1802, 2005.

[119] Irina O Vvedenskaya, Hanif Vahedian-Movahed, Jeremy G Bird, Jared G Knoblauch, Seth R Goldman, Yu Zhang, Richard H Ebright, and Bryce E Nickels. Interactions between rna polymerase and the "core recognition element" counteract pausing. *Science*, 344(6189):1285–1289, 2014.

[120] Adrien Chauvier, Frédéric Picard-Jean, Jean-Christophe Berger-Dancause, Laurène Bastet, Mohammad Reza Naghdi, Audrey Dubé, Pierre Turcotte, Jonathan Perreault, and Daniel A Lafontaine. Transcriptional pausing at the translation start site operates as a critical checkpoint for riboswitch regulation. *Nature communications*, 8(1):13892, 2017.

[121] Jin Young Kang, Rachel Anne Mooney, Yuri Nedialkov, Jason Saba, Tatiana V Mishanina, Irina Artsimovitch, Robert Landick, and Seth A Darst. Structural basis for transcript elongation control by nusg family universal regulators. *Cell*, 173(7):1650–1662, 2018.

[122] Jin Young Kang, Paul Dominic B Olinares, James Chen, Elizabeth A Campbell, Arkady Mustaev, Brian T Chait, Max E Gottesman, and Seth A Darst. Structural basis of transcription arrest by coliphage hk022 nun in an escherichia coli rna polymerase elongation complex. *Elife*, 6:e25478, 2017.

[123] Chengjin Zhu, Xieyang Guo, Philippe Dumas, Maria Takacs, Mo'men Abdelkareem, Arnaud Vanden Broeck, Charlotte Saint-André, Gabor Papai, Corinne Crucifix, Julio Ortiz, et al. Transcription factors modulate rna polymerase conformational equilibrium. *Nature communications*, 13(1):1546, 2022.

[124] Seungha Hwang, Paul Dominic B Olinares, Jimin Lee, Jinwoo Kim, Brian T Chait, Rodney A King, and Jin Young Kang. Structural basis of transcriptional regulation by a nascent rna element, hk022 put rna. *Nature Communications*, 13(1):4668, 2022.

[125] Jin Young Kang, Tatiana V Mishanina, Yu Bao, James Chen, Eliza Llewellyn, James Liu, Seth A Darst, and Robert Landick. An ensemble of interconverting conformations of the elemental paused transcription complex creates regulatory options. *Proceedings of the National Academy of Sciences*, 120(8):e2215945120, 2023.

[126] Ivan Gusarov and Evgeny Nudler. Control of intrinsic transcription termination by n and nusa: the basic mechanisms. *Cell*, 107(4):437–449, 2001.

[127] Smarajit Mondal, Alexander V Yakhnin, Aswathy Sebastian, Istvan Albert, and Paul Babitzke. Nusa-dependent transcription termination prevents misregulation of global gene expression. *Nature microbiology*, 1(1):1–8, 2016.

[128] Bing Wang and Irina Artsimovitch. Nusg, an ancient yet rapidly evolving transcription factor. *Frontiers in Microbiology*, 11:619618, 2021.

[129] Jin Young Kang, Tatiana V Mishanina, Robert Landick, and Seth A Darst. Mechanisms of transcriptional pausing in bacteria. *Journal of molecular biology*, 431(20):4007–4029, 2019.

[130] Bin Liu and Thomas A Steitz. Structural insights into nusg regulating transcription elongation. *Nucleic acids research*, 45(2):968–974, 2017.

[131] Zachary F Mandell, Reid T Oshiro, Alexander V Yakhnin, Rishi Vishwakarma, Mikhail Kashlev, Daniel B Kearns, and Paul Babitzke. Nusg is an intrinsic transcription termination factor that stimulates motility and coordinates gene expression with nusa. *Elife*, 10:e61880, 2021.

[132] Kyle E Watters, Eric J Strobel, Angela M Yu, John T Lis, and Julius B Lucks. Cotranscriptional folding of a riboswitch at nucleotide resolution. *Nature structural & molecular biology*, 23(12):1124–1131, 2016.

[133] Simon Blouin, Jérôme Mulhbacher, J Carlos Penedo, and Daniel A Lafontaine. Riboswitches: ancient and promising genetic regulators. *ChemBioChem*, 10(3):400–416, 2009.

[134] Alexander Serganov and Dinshaw J Patel. Ribozymes, riboswitches and beyond: regulation of gene expression without proteins. *Nature Reviews Genetics*, 8(10):776–790, 2007.

[135] Brian J Tucker and Ronald R Breaker. Riboswitches as versatile gene control elements. *Current opinion in structural biology*, 15(3):342–348, 2005.

[136] Alexey G Vitreschak, Dimitry A Rodionov, Andrey A Mironov, and Mikhail S Gelfand. Riboswitches: the oldest mechanism for the regulation of gene expression? *TRENDS in Genetics*, 20(1):44–50, 2004.

[137] Adrien Chauvier, Pujan Ajmera, Rajeev Yadav, and Nils G Walter. Dynamic competition between a ligand and transcription factor nusa governs riboswitch-mediated transcription regulation. *Proceedings of the National Academy of Sciences*, 118(47):e2109026118, 2021.

[138] Xieyang Guo, Alexander G Myasnikov, James Chen, Corinne Crucifix, Gabor Papai, Maria Takacs, Patrick Schultz, and Albert Weixlbaumer. Structural basis for nusa stabilized transcriptional pausing. *Molecular cell*, 69(5):816–827, 2018.

[139] Kaili Wang, Renyi Zhou, Yifan Wu, and Min Li. Rlbind: a deep learning method to predict rna–ligand binding sites. *Briefings in Bioinformatics*, 24(1):bbac486, 2023.

[140] Julia E Martin, My T Le, Nabin Bhattarai, Daiana A Capdevila, Jiangchuan Shen, Malcolm E Winkler, and David P Giedroc. A mn-sensing riboswitch activates expression of a mn2+/ca2+ atpase transporter in streptococcus. *Nucleic acids research*, 47(13):6885–6899, 2019.

[141] Krishna C Suddala, Ian R Price, Shiba S Dandpat, Michal Janeček, Petra Kührová, Jiří Šponer, Pavel Banáš, Ailong Ke, and Nils G Walter. Local-to-global signal transduction at the core of a mn2+ sensing riboswitch. *Nature Communications*, 10(1):4304, 2019.

[142] Ian R Price, Ahmed Gaballa, Fang Ding, John D Helmann, and Ailong Ke. Mn2+-sensing mechanisms of yybp-ykoy orphan riboswitches. *Molecular cell*, 57(6):1110–1123, 2015.

[143] Jeffrey E Barrick, Keith A Corbino, Wade C Winkler, Ali Nahvi, Maumita Mandal, Jennifer Collins, Mark Lee, Adam Roth, Narasimhan Sudarsan, Inbal Jona, et al. New rna motifs suggest an expanded scope for riboswitches in bacterial genetic control. *Proceedings of the National Academy of Sciences*, 101(17):6421–6426, 2004.

[144] Hsuan-Lei Sung and David J Nesbitt. Single-molecule fret kinetics of the mn2+ riboswitch: evidence for allosteric mg2+ control of "induced-fit" vs "conformational selection" folding pathways. *The Journal of Physical Chemistry B*, 123(9):2005–2015, 2019.

[145] Youngdo Won. Force field for monovalent, divalent, and trivalent cations developed under the solvent boundary potential. *The Journal of Physical Chemistry A*, 116(47):11763–11767, 2012.