USE OF MOLECULAR MECHANICS FORCE FIELDS AND RISM DENSITIES TO IMPROVE MACROMOLECULAR MODELS

BY JONATHON GEORGE GRAY

A dissertation submitted to the

School of Graduate Studies

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Chemistry and Chemical Biology, Quantitative Biomedicine

Written under the direction of

David A. Case

and approved by

_____

_____

_____

_____

New Brunswick, New Jersey

October, 2020

ABSTRACT OF THE DISSERTATION

Use of molecular mechanics force fields and RISM densities to improve macromolecular models

by Jonathon George Gray

Dissertation Director: David A. Case

As RNA structures continue to be solved at a rapid pace, and as RNA has become a target for therapeutics and has been found to have many different functions other than basic nucleic acid functions, the importance of properly modelled structures continues to grow. With the conventional restraints used in crystallographic refinement, persistent outliers and errors crop up in the structures published. The use of AMBER-derived restraints in the PHENIX crystallographic refinement process has been proven to improve the structures of proteins modelled based on experimental data, and is implemented in RNA structures in this thesis. Further improvement of structural modelling can be made in solvent description. While the most accurate way to model solvent is through explicit solvent molecules in crystal MD simulations, it is also the most computationally expensive. Meanwhile, the faster implicit models, such as the Generalized-Born model, are approximate and can sometimes lead to improper secondary structure in macromolecules. The periodic 3D-RISM method, presented in Chapter 3 and an upcoming paper, calculates densities for each solvent entity. It is thought to be more accurate than general implicit methods, but is faster than explicit methods. In this thesis, crystal MD simulations and periodic 3D-RISM calculations are employed to study the solvation of RNA structures.

In Chapter 2, parallel refinements with conventional and AMBER-derived restraints in PHENIX on RNA molecules are presented. The resultant structures are analyzed via energy calculations and MolProbity analysis. The results show that in a data set of 21 structures, the AMBER restraints lead to improved electrostatic and non-bonded interactions over conventional restraints, which was expected, as this is the main improvement of AMBER restraints over conventional restraints. This leads to overall energetic improvement over the course of the data set, except for the structure at highest resolution. Also, this occurs with little concession to structure factors, as the r-free factors are very similar. There are increases in the r-work, but the r-gap, or the gap between the r-work and r-free factors (an indication of over-fitting when high), is generally the same if not decreased as compared to the conventionally restrained refinements. The geometric outliers are more numerous for AMBER-restrained structures, but analysis and testing of repetitive bond and angle outliers finds that this appears to be due to both a larger distribution of angles and bond lengths due to the interconnectedness of all the energy terms in AMBER, as well as a difference in the ideal values for these terms between AMBER and MolProbity. At low resolution, where the experimental data is poor and the need for external restraints is greatest, there is even greater improvement. This implies greater physical accuracy of the structures produced, and could lead to improved structural understanding.

In Chapter 3, the periodic 3D-RISM theory is presented. The existing 3D-RISM code for non-periodic systems has been expanded to periodic systems, which allows for the possibility of use in refinement description of solvent. Results are presented for experiments in proteins and RNA comparing refinement with the standard flat density, 3D-RISM results, and explicit MD solvent densities, which show that 3D-RISM improves r-factors over the standard density, while being improved upon by MD, which is more time-consuming. Results for different proteins are also presented showing that the number of water molecules produced via 3D-RISM calculations are all very similar to the numbers derived from crystal MD. Further work including calculation of 3D-RISM solvent throughout refinements may be the next step.

In Chapter 4, crystal MD simulations of three of the structures from the PHENIX

data set are presented as an opportunity to look at the dynamics of the structures as well as a baseline for testing the accuracy of 3D-RISM code presented in Chapter 3. The 3D-RISM calculations match fairly closely the number of water molecules found by MD. Through comparative 3D-RISM calculations, it is found that solvent composition has an effect on the number of ions produced to neutralize the solute. More sodium ions are used than potassium ions when used at the same concentration in conjunction with magnesium. As sodium's ionic radius is smaller than potassium's, it appears that this is due to size differences. The 3D-RISM code provides an appropriate, and less time-expensive, approximation of solvent description and interactions than the standard crystal MD simulations.

In Chapter 5, the sarcin/ricin domain of the ribosomal RNA of *E. coli*, a well-conserved domain across species with many structures in the PDB, is used as a test molecule for PHENIX refinements with AMBER restraints, periodic 3D-RISM single-point calculations, and minimizations with periodic 3D-RISM. These small structures all contain the same solute, so the differences in results should be resolution- or solvent-dependent. When analyzing the PHENIX results, there is a trend in energetics, specifically non-bonded interactions, toward greater improvement over conventional restraints with AMBER restraints as the resolution worsens. Different parameters were tested to determine what sets resulted in the fastest runs.

Acknowledgements

Chapter 3 is a collaborative work with George Giambaşu, Tyler Luchko, Darrin M. York, and David A. Case and is in preparation for submission for publication. Chapter 2 is the basis for an upcoming first-author paper. All original molecular images were created in UCSF Chimera, and all trajectory visualization was likewise performed in UCSF Chimera, which was developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco, and supported by NIH P41-GM103311[92].

I would have been completely unable to complete this work or even get to this point without the help of so many people. I'd like to thank them here, but there is not enough room to cover everyone. To all have helped me in my educational and personal growth, thank you.

Regarding specific people:

First, I'd like to thank Professor David A. Case for his support, his close mentorship, and his patience and understanding as I worked through learning a completely new sort of chemistry to perform all this work. I appreciate all the help and guidance. Thank you so much every piece of advice, answered email, and nudge. It has all helped me reach this point, and I'm forever grateful.

I'd also like to thank Shashidhar Rao and George Giambaşu for their close and intentional help as I entered the lab from an experimental background. Your one-on-one advice and encouragement helped me feel welcome, and helped me focus on completing my courses while also getting initiated to the lab and the work. You were so helpful, and I appreciate you both very much.

I want to thank my committee for being so supportive and focused on my growth as a scientist, my other fellow labmates for their emotional support, and my college

professors and high school chemistry teachers for motivating me to study chemistry all the way to this point. I especially want to thank Professor Daniel King, my analytical chemistry professor at Taylor University, who helped me throughout all of college: as a friend, a great teacher, and an amazing mentor. My high school AP chemistry teacher, Mr. John Thompson, was the stimulus toward me studying chemistry. He is a hilarious man and an enthusiastic, passionate teacher of chemistry. Thank you all.

Finally, I'd like to thank my family. My parents, Carlos and Teresa, and my brother, George, have been great support, pushing me to do my best when under their roof, and encouraging me from afar during my undergraduate and graduate studies. Thank you for the rides to practices, all of the backing, and the opportunity for a great future you provided me. George, who is also studying chemistry for his PhD, has been the best brother a man can ask for, and has shared his experiences with me and been a sounding board. Thank you so much George!

And of course, I could not mention my family without mentioning my wife, Grace. Grace, you have been supportive and loving for so long, even when I've flagged as a productive worker, at which points you've pushed me to be better than who I was being at the time. You're my best friend, my partner in goofiness, and the best part of my life. I'm so grateful to be yours, and I love you so much.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

As structure influences and informs function, it is important to have proper models of biological molecules. Generally speaking, there are two components to macromolecular structures: the macromolecule and the solvent environment. This dissertation will investigate methods implemented to improve both parts of macromolecular structures, with RNA as the test biological structure set.

1.1 RNA

mRNA has long been known as the intermediate molecule between DNA and protein[52]. In tRNA, we have a molecule that aids in that translation to protein from mRNA[99]. The other major type of RNA, rRNA (ribosomal) also plays a major part in translation[83], and has been found to be a member of a group of molecules known as ribozymes. These catalytic RNA molecules perform one of the more recently discovered uses of RNA molecules[62, 19]. Further uses of RNA include the escorting of Cas9 in CRISPR-Cas9 gene editing[54] and post-transcriptional modifications[40] and RNA silencing[32] in the forms of miRNA and siRNA, respectively.

With such a vast array of important functions and uses for RNA molecules, it is important to have proper structural models of these molecules. This dissertation will test out different computational methods for improving RNA structural models. A lot of the work performed with these methods has been performed on proteins, which are more diverse in terms of different types of building blocks available to make up the macromolecule, but have fewer backbone parameters than RNA. The flexibility and heterogeneity of the RNA backbone in terms of number of bonds, angles, and torsions, and even sugar puckers, could lead to issues unseen in work on protein structures. Also,

due to the sugar-phosphate backbone, RNA has a much more concentrated and higher overall charge than proteins do.

1.2 Crystallographic refinement

In the process of structural determination and model development from X-ray crystallography, refinement is the last part of the process. Before refinement can occur, however, molecules have to be isolated, crystals have to be grown, data has to be collected, densities have to be determined, and preliminary models must be developed.

In order to collect data for the development of electron density maps for structural models to be refined into, crystals of the structure must first be grown. This is a generally difficult process that is somewhat of a rate-limiting step for structure determination, as it is still a guess-and-check process to find the best conditions to grow crystals[37]. That is the case for proteins, but it is even more of an issue in nucleic acid structures, of which all of our structures in this thesis are made, where the strong backbone negative charge makes it difficult to form the crystal contacts[57] between neighboring molecules necessary for crystals to develop[66]. Once these crystals, which contain repeating unit cells of molecules related by symmetry operations, are grown, they are cryoprotected (usually) and eventually mounted for shooting with X-rays, which are diffracted by the electrons in the structure. This diffraction pattern is deconvoluted into electron density, into which a structural model based on known sequence of the structure is fit. From here, the structure is refined to improve fit to the electron density[37].

The general process of refinement involves minimizing a target function that compares the experimental structure factors and those calculated from the model. The amount of information able to be gleaned from the experimental data is insufficient, however. Also, using only the experimental data to fit the model can lead to overfitting. These are reasons why geometric restraints are used to supplement the experimental data in model refinement.

While there are many different refinement programs, PHENIX is the program that will be used in this dissertation. In *phenix.refine*'s coordinate refinement, the target

function to be minimized is generalized as:

$$T_{XYZ} = w * T_{exp} + T_{xyz\_restraints} \tag{1.1}$$

Each of these terms is a function of the atomic coordinates: T_{XYZ} is the target residual that *phenix.refine* aims to minimize, T_{exp} is a measurement of how the model structure factors match the experimental structure factors, $T_{xyz\_restraints}$ measures the model's fit to geometric restraints, and w is the generic weight factor that determines the relative weight placed on the experimental data and the geometric restraints. In conventional refinement, $T_{xyz\_restraints}$ involves the Engh and Huber and Conformation Dependent Library restraints[78].

In a recently published paper[78], AMBER force field restraints were introduced in place of the Engh and Huber restraints and used in refinements of protein structures in parallel with conventional refinements. The AMBER-restrained refinements improved structural models in the areas of clashscores, protein backbone torsion angles, electrostatic, hydrogen bond, and van der Waals interactions, and side chain rotamers. Particular improvement was seen in low resolution structures. While these sorts of improvements are expected in the RNA structures to be refined in this thesis, interactions with solvent and RNA backbone may affect the results due to the high negative charge of RNA, and the high number of cations needed to neutralize that charge. The high charge could possibly cause strong energetic interactions to overcome the anchoring of the structure in the experimental data.

1.3 Solvent models

As solvent makes up about 30-70% of volume in biomolecular crystals[73] and greatly facilitates interactions in active sites and with ligands[47, 93], it is important to properly model solvent. However, due to the vast amount of solvent compared to the number of macromolecules in modeled systems, and the fact that macromolecules already contain very large numbers of atoms, it can be difficult to model all the atoms in a unit cell. Thus, there are trade-offs to consider when deciding how to model solvent.

The highest-accuracy and most computationally intensive way is to model explicit solvent molecules in an atomistic way. This allows for direct calculation of forces on and due to each molecule and is the most comprehensive description of what is going on in the system, as solvent exists not as just some vast bath but as individual molecules that make up the bulk of the volume of the system. However, this way of modelling solvent requires calculations for all solvent molecules, resulting in much higher computational times than implicit models. Also, in the case of crystal MD simulations, which will be discussed later, the number of water molecules that properly simulate crystalline conditions must be reached through a guess-and-check series of test simulations, which also increase the amount of time spent reaching the desired end goal[21].

On the other end of the spectrum are implicit solvent models. These include various versions of the Generalized-Born and Poisson-Boltzmann models, and others. Implicit models have the benefit of being less computationally demanding, and thus faster, while having been developed to be good approximations of the explicit models. They also allow for better sampling of conformational space[34, 39, 2, 112, 3]. However, they are known to be a step down from explicit solvent models regarding accuracy, even resulting in inaccurate secondary structures within proteins[84, 103].

A middle-ground approach that will be tested in RNA structures later in this thesis is periodic 3D-RISM. The theory related to this method will be presented in Chapter 3, while calculations testing the accuracy of this approach are presented in both Chapters 3 and 4. 3D-RISM calculates densities of solvent entities (water and ions) chosen to solvate the macromolecule. An example of these densities as wire mesh, as compared to the placement of deposited solvent particles, can be seen in Figure 1.1. These densities can be used to try to place molecules, and numbers of particles in the system are able to be determined from these calculations. These calculations are able to reach an understanding of the solvent distribution in a system far faster than crystal MD simulations using explicit water molecules, and hopefully will lead to results that are fairly similar in terms of accuracy.

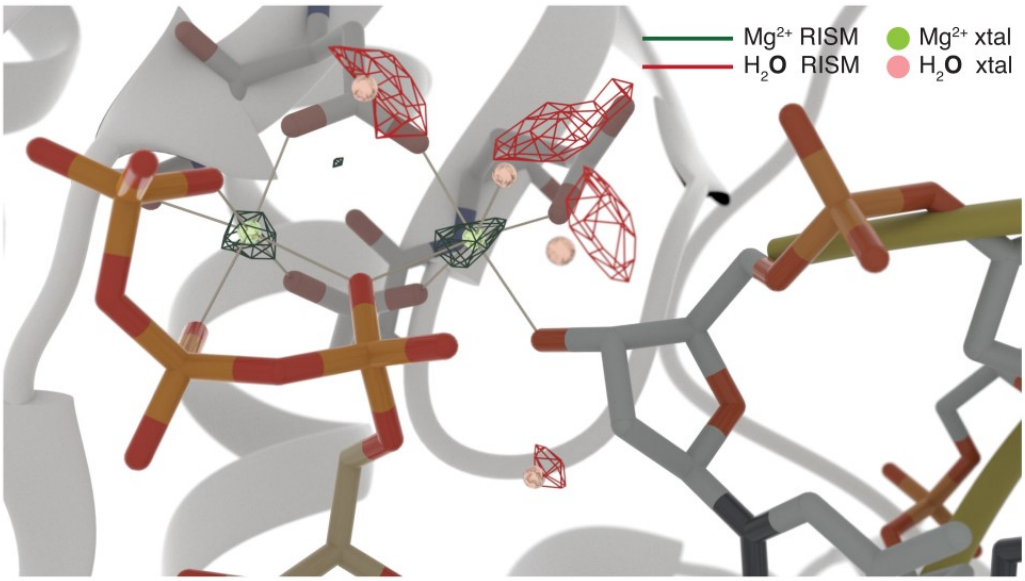

Figure 1.1: Computed Mg^{2+} and water densities (shown as dark green and red isodensity meshes, respectively) versus crystallographically resolved positions (shown as light green and pink spheres) near the protein−DNA interface of the catalytic site of polymerase h (PDB ID: 3mr2[15]). (Image and caption text from [42].)

1.4 Crystal MD simulations

MD simulations can be performed in a vast sea of solution, but they have been found to better match experimental data when performed with conditions used to mimic crystalline conditions[21]. It was found in simulations of both DNA and RNA duplexes that the crystal simulations result in structures with higher fidelity to the experimentally derived structures than those in solution simulations[68](further sources within article). The simulations performed in such conditions can provide complementary data of the dynamics of structures to the structural data gleaned from crystallographic experiments[68]. This is especially true in situations where molecules have to be in their inactive forms to be crystallized[89](for more sources and further information, check out this source). Further information about the setup of crystal MD simulations can be found in Chapter 4.

In the following work, RNA will be used as a test molecule type for different methods used in model description. Chapter 2 will investigate the effects of AMBER force fields being used as restraints in crystallographic refinement in PHENIX of RNA molecules. Chapter 3 will present theory of and data from tests of the new periodic 3D-RISM code, while Chapter 4 will look at solvation of three RNA structures through the established method of crystal MD simulations and the newly presented periodic 3D-RISM code. Finally, in Chapter 5, the sarcin/ricin domain is used as a target system for further analysis of all methods covered in this thesis.

Chapter 2

PHENIX refinement of RNA structures with AMBER force fields

2.1 Introduction

Since the turn of the millennium, the number of RNA structures solved has vastly increased. As these structures have been solved, the many functions of the different types of RNA molecules have begun to be discovered[27, 8, 97]. Self-catalytic RNA molecules, ribozymes, have been found to be possible evidence supporting the "RNA world" hypothesis[20, 18]. MicroRNAs have been studied as antitumor agents[16, 25, 69], and CRISPR, which is an RNA-guided technique, has expanded in use[94, 29, 9]. In order to make the best conclusions about function and use of these RNA structures, the models must be as physically accurate as possible. As the number of structures deposited in the PDB has increased, it has been determined that there is a persistence of geometric outliers in RNA models developed from crystallographic data[27, 98, 35]. These outliers not only arise due to low resolution in many RNA structures[27, 58], but also due to the use of geometric restraints that do not take attractive electrostatic interactions into account[79]. Different methods have been developed to try to improve upon the outliers found in RNA structures, including ERRASER.

ERRASER uses the Rosetta energy score function to guide real space refinement of the structure, individually modifying outlier bonds, angles, etc., to improve the overall structure. However, it is a time-consuming process, and requires another step of reciprocal space refinement within PHENIX at the end[27]. PHENIX has also introduced nucleic acid secondary structure restraints to maintain proper base pair hydrogen bonding, planarity, and stacking[53, 78]. This, however, does not include an energetic term, and forces the hydrogen bonding through restraint parameters.

Our interface of PHENIX and the AMBER force fields integrates energetics, especially the electrostatic term (an important factor in highly charged molecules, like RNA), with the electron density data to guide reciprocal space refinement. So far, this interface has been shown to guide reciprocal space refinement of protein structures with marked improvement over conventional refinement[78]. In this chapter, the effects of use of this interface on RNA structure refinement will be compared to conventional refinement, as well as other techniques designed to improve the results gained from refinement.

2.2 Theory and implementation

The "*import sander*" command in python allows for PHENIX to reach to AMBER and access the energies and forces through the *sander* python API. In each coordinate refinement step, the asymmetric unit is expanded to a unit cell, which is used by *sander* to calculate the energy gradients, and then those gradients are combined, in place of those developed by use of the conventional geometric restraints within PHENIX, with the gradients from the target function from the X-ray data. This combination of gradients is then used to change the coordinates through minimization here[78].

In *phenix.refine*'s coordinate refinement, the target function to be minimized is generalized as:

$$T_{XYZ} = w * T_{exp} + T_{xyz\_restraints} \tag{2.1}$$

Each of these terms is a function of the atomic coordinates: T_{XYZ} is the target residual that *phenix.refine* aims to minimize, T_{exp} is a measurement of how the model structure factors match the experimental structure factors, $T_{xyz\_restraints}$ measures the model's fit to geometric restraints, and w is the generic weight factor that determines the relative weight placed on the experimental data and the geometric restraints. In conventional refinement, $T_{xyz\_restraints}$ involves the Engh and Huber and Conformation Dependent Library restraints, while in AMBER refinement, this term actually ends up being the potential energy function in AMBER using the ff14SB protein force field[72], as well as the OL3 RNA force field[113].

The target function for AMBER-restrained refinement can be more specifically laid out as follows (with some tweaking of nomenclature):

$$T_{XYZ} = wxc\_scale * wxc * T_{exp} + wc * E_{AmberFF} \tag{2.2}$$

Here, wc is usually 1.0 in cases where restraints are being used and is set to 0 turn off restraints in high resolution cases, wxc is a "ratio of gradient norms" between the geometric and experimental target functions (normally fluctuates), and wxc_scale is a fixed value that consistently weighs the experimental target function against the restraint target function. Also, here, the restraint term has been labeled as $E_{AmberFF}$, as in the protein paper, to more explicitly show that the potential energy function has been implemented[78]. These weighting terms are important in this work, as to allow for Boltzmann weighting of AMBER restraints, we set up *phenix.refine* to set both wxc_scale and wxc to 1, and set wc=1.667. The experimental target function estimates $-LL$, the negative log of the likelihood of finding the data, given the current model; hence the likelihood is maximized when T_{exp} is minimized. For a Boltzmann distribution, the corresponding $-LL$ is $E_{AmberFF}/k_BT$, which is $1.667E_{AmberFF}$, for AMBER energies expressed in kcal/mol. (This "theoretical" value does not necessarily provide the best weighting of restraint and force field energies, but has been found to work fairly well for proteins, where the exact value of wc is not critical to the results. The *phenix.refine* code can also search for optimized weights, based on some empirical criteria, but this option was not explored here.)

2.3 Methods

2.3.1 Structure selection

In order to have a robust set of RNA structures to investigate and on which to test the PHENIX/AMBER interface, we started with the set of structures used to test ERRASER. After removing structures that were difficult to build due to size or difficult to parameterize units (ribosomal subunits, osmium-ion-containing structure, other building issues) and adding in 3 structures of an RNA mimic of the sarcin/ricin domain of the

23S - 28S ribosomal RNA, we arrived at a set of 21 structures ranging in resolution from 1.04 to 3.6 Å. The set includes the domain previously mentioned, as well as riboswitches, pseudoknots, ribozymes, and viral RNA domains, and some of the structures include the U1 small nuclear ribonucleoprotein A. For a description of each structure, see Table 2.1.

2.3.2 Structure preparation

Once the structures were chosen, they had to be prepared for refinement and use with AMBER's *sander* utility for minimization. This involves creating parameter and coordinate files for use in *sander*, a molecular dynamics engine within AMBER. This is done through the *AmberPrep* utility of PHENIX (assuming PHENIX has already been installed), preparing parameter and coordinate files for the whole unit cell, while also minimizing all atoms of the unit cell for 50 steps to straighten out any horrible geometric issues, as well as hopefully moving the structure away from any local minima and to a point where refinement could take the structure to the global minimum.

Before *AmberPrep* could be run, some of the structures had to be edited in order to avoid failures of the *AmberPrep* process. Any structures with 5'-terminal guanosine-monophosphates, labeled as a regular G, had to be labeled as GMP in order to be built with the phosphate group, as opposed to a typical 5'-guanosine without the phosphate. All 5'-terminal monophosphates were built into a new library as a part of this work and this library is now distributed as a part of the AmberTools20 package. Also, *mol2* and *frcmod* files for 5'-terminal GTP and GDP were built using parameters from Heather Carlson[74]. Other modified residues (5BU, CCC, etc.) were built in library or *mol2* files to allow for the proper parameters and proper connections (took CCC, built as a noncovalent ligand, and added proper connect atoms to connect to previous residue, etc.). Charges were also checked and modified, especially in modified terminal residues, to allow for integral charges on residues or, in the case of terminal residues, pairs of residues. Structures with modified residues and ligands are identified in Table 2.1.

PDB ID	Description	HETATMs
1q9a	sarcin/ricin domain from *E.coli* 23S rRNA	
1y0q	active group I ribozyme-product complex	SO4, SPK
2a43	luteoviral pseudoknot	
2gdi	thiamine pyrophosphate-specific riboswitch	CCC, GTP, TPP
2gis	SAM riboswitch mRNA regulatory element	GMP, IRI, SAM
2oiu	L1 ribozyme ligase circular adduct	
2pn3	Hep C IRES subdomain 2a	5BU
2pn4	Hep C IRES subdomain 2a	5BU
2qus	hammerhead G12A mutant pre-cleavage	GTP
2ygh	SAM-I riboswitch, with S-adenosylmethionine	GMP, SAM
3bo3	active site following exon ligation by group I intron	A23, GTP, U1
3e5e	SMK box (SAM-III) riboswitch with SAH	GTP, SAH
3f2q	FMN riboswitch with FMN	CCC, FMN, GTP
3gx5	*T. tencongensis* SAM-I riboswitch	GMP, SAM
3iwn	bacterial c-di-GMP riboswitch	C2E, U1
3mxh	c-di-GMP riboswitch from *V. cholerae*	C2E, GTP, U1
3r4f	prohead RNA	
3tzr	riboswitch complex from Hep C IRES	SO4, SS0
480d	sarcin/ricin domain from *E. coli* 23 S rRNA	
483d	sarcin/ricin domain from *E. coli* 23 S rRNA	
4fe5	xpt-pbuX guanine riboswitch aptamer domain	ACT, HPA, NCO

Table 2.1: This table shows the PDB ID, name, and chemical chemical component IDs of the noncovalent ligands, modified residues, or solvent molecules; the 3bo3, 3iwn, and 3mxh structures also include U1 proteins. Only 6 structures in the data set do not contain any of these molecules that need to be built with parameters not found explicitly within the AMBER force fields (1q9a, 2a43, 2oiu, 3r4f, 480d, and 483d).

Two structures had specific problems to address:

- 2oiu, the L1 ribozyme ligase circular adduct, required the creation of a *myuclinks* file used in the *tleap* utility of AMBER during the *AmberPrep* process to build the parameters of the structure and properly connect atoms to each other. In this *myuclinks* file, explicit bonds between the first and last residue in each chain had to be made in order to create the circular adduct, and the terminal residues had to be removed from the RNA force field being used to prevent them being built without these connections.

- With 3f2q, the flavin mononucleotide riboswitch bound to its ligand, a TER card had to be placed in the sequence where there is a gap in the density, and the *use_ reduce=False* parameter had to be used in the *AmberPrep* utility to prevent

a hydrogen from being added to O3' of the residue on the 5'-end of the gap and treating the residue as a 3'-terminal residue. This was also true for 1y0q.

- CIF files were also developed for any residues not included in PHENIX's data to provide energy terms within PHENIX. One specific residue where this is the case is GMP, which needed a specific CIF file developed for it as a terminal residue because it contains a proton on the phosphate group not found in the noncovalent ligand form of the molecule.

Once everything was prepared properly regarding any modified residues that needed to be bound in the RNA chain, '*phenix.AmberPrep xxxx.pdb minimise=amber_all*' was run (with the added *use_reduce=False* parameter in the cases mentioned) to build the parameter, coordinate, and order files needed to provide the required files for AMBER within the PHENIX refinement run, as well as providing a connection between the way AMBER and PHENIX each like to order atoms within residues. The *AmberPrep* utility runs the structure through *pdb4amber* to remove any multiple conformers and check for gaps in the structure and to determine residues that are not a part of protein or nucleic acid so that those can then be built (if normal) by *antechamber/eLBOW*. The *AmberPrep* script then creates an input script for *tleap* and runs it to fill out all the hydrogens and missing heavy atoms for residues and to build the parameter and coordinate files for the asymmetric unit. The coordinate file is then passed through *ChBox* to make sure the unit cell parameters at the end of the coordinate file are accurate. Then, the asymmetric unit is passed through *UnitCell* to propagate the asymmetric unit into a full unit cell using the symmetry operations provided in the PDB header, and the process is repeated (without the need to build the ligands again) for the unit cell, after which the 50-step minimization is performed, and the output coordinate file is combined with the parameter file using *ambpdb* to create a PDB file, which is pared down to the asymmetric unit, providing the *4phenix_xxxx.pdb* file used as the input in refinement, along with the unit cell parameter (*4amber_xxxx.prmtop*) and minimized coordinate (*4amber_xxxx.rst7*) files.

2.3.3 Refinement

As the focus of this chapter is to test the effectiveness of AMBER restraints within PHENIX refinement compared to conventional Engh and Huber restraints, all of the prepared structures were refined in PHENIX with parallel refinements with and without AMBER restraints. First, a conventional refinement was performed with a flag to generate r-free flags for each structure in order to provide a reflections file with all of the needed r-free flags built in. This output mtz file was then used as the input reflection data for the parallel refinements for that particular structure. Each refinement script contained the same parameters other than whether or not AMBER restraints were called for. Each refinement was 10 macrocycles long, and called for refinement of individual coordinates in reciprocal space, individual B-factors, and occupancies. In the AMBER refinements, the script contained the command line arguments for the input parameter, coordinate, and order files, as well the argument to print AMBER energies in the output log file, as well as the set up of the weighting of the AMBER restraints compared to the reflection data (*wxc* fixed to 1.0, *wc* set to 1.667; indicated in Equation 2). For the nonstandard residues and ligands, the CIF files had to be included as input files on the first line of the command in the script. In conventional refinements for 2oiu and 3f2q, custom restraint files had to be made to induce bonding of the terminal residues to form the circular ligase and to bond the terminal GTP to the rest of the structure, respectively.

2.3.4 Analysis

In order to compare the results of refinement with and without AMBER restraints, a set of statistics were chosen as comparison points: structure factors, clashscore, geometric outliers, suiteness score, and AMBER energy. Clashscores are per-1000-atom numbers of strong clashes between atoms in the structures[26]. The geometric outliers include bond length, bond angle, sugar pucker, and torsion suite outliers. While bond length and bond angle outliers require no definition here, sugar puckers and torsion suites bear explanation. The ribose sugars in RNA nucleotides have two main types of pucker, as

they are not planar molecules. These are C3'-endo and C2'-endo[26, 81], and outliers are defined in MolProbity when there are outliers in e or d torsion angles, or if "the perpendicular distance between the C1'-N1/9 glycosidic bond vector and the following (3') phosphate" is less than 2.9 Å for C3'-endo puckers and is greater than 2.9 Å for C2'-endo puckers (as the opposite is correlated to the proper puckers: >2.9 Å for C3'-endo and <2.9 Å for C2'-endo)[26]. Regarding torsion suites, it has been found that the RNA backbone torsions, when analyzed from sugar to sugar (7 torsions, d-e-z-a-b-g-d), result in a set of "distinct 'rotameric' backbone conformers"[26], or torsion suites[81, 101, 26]. The outliers and suiteness score are determined by the *Suitename* utility. Outliers are suites that exist outside the boundary parameters of the 7 torsions within the any of the defined suite conformers[101, 26, 111]. There are differing numbers of consistent suite conformers determined through analysis of deposited structural data, from 42 to 52 [81, 101, 26, 111]. Finally, the suiteness score is a composite of the suiteness scores for each torsion suite, where 1.0 is a perfect fit to the mean torsion angles for the labeled conformer, and scores run all the way to 0.01 at the extremes of the range for each torsion within a suite. A suiteness score of 0 results from torsions outside of any suite conformer's ranges, and, thus, indicates an outlier, or a suite that has yet to be discovered as a consistent suite within the data [101].

In order to get to these statistics, only a few things needed to be done. To get all of these data points other than the AMBER energy, MolProbity[35] was called from within the PHENIX command line, with '*phenix.molprobity output.pdb *.cif*' (if there were any CIF files for ligands needed to perform the refinement; make sure there are no files with the *.cif* file name ending in the directory other than those for ligands, etc.; alternatively, each *.cif* file can be named individually in the command). The output contains a python file to point out clashes, etc., in *coot*, and a text file with information about the geometry outliers and other statistics. All of the statistics we were interested in were taken from this output file. The structure factors and clashscore were found in the summary at the end of the file, while the geometry outliers and suiteness score were all found in the RNA validation portion of the output file. In the case of a structure where there were no torsion suite outliers, *phenix.rna_validate* was run in order to get

the suiteness score because it did not show up in the MolProbity output file unless there were suite outliers.

To get the AMBER energy, one could find energy values throughout the output of the AMBER refinement. However, in order to get an energy value for the conventional refinement output structure, as well as providing a consistent way to get energy values for all structures, the output structures were stripped of their water molecules, mono- and divalent cations, and fairly simple anions (sulfate, acetate, etc.). The structures were also stripped of their hydrogen atoms using *pdb4amber*, as the riding hydrogens consistently provided large energy penalties. Then, these stripped structures were run through *AmberPrep* with 100 steps of minimization with the *minimise=amber_h* flag, with adjustment of the parameters in the AmberPrep.py file to restrain the heavy atoms using *ibelly=1* as opposed to *ntr* restraints. (The standard build of the AmberPrep.py code uses standard harmonic potential restraints to hold the chosen atoms in very similar locations, but as we did not want a change of heavy atom location at all, the code was changed to use belly-type dynamics to hold the heavy atoms in place without any movement at all. The *minimise=amber_h* flag calls for only hydrogen atoms to be minimized.) The unit cell parameter and coordinate files made from this *AmberPrep* were used to run a 0-step MD run in *sander* with periodic 3D-RISM. Running *sander* with a step number of 0 allows for a calculation of the energy at the beginning of the "run," without any MD being performed. The 3D-RISM calculation allows for the screening out of some of the electrostatic energy by the calculated solvent density. This helps to focus the results on the macromolecular contribution to the energy and provide a true representation of the solvent electrostatics, as the electrostatics resulting directly from the refinement involve only a few solvent molecules deposited in the structures, and thus do not provide an accurate description of the electrostatics in the system and the interaction between the macromolecules and the solvent in the unit cell. Each component of the energy as printed out in the *sander* output of each method had the value for the conventional refinement subtracted from the AMBER-restrained refinement value and was then divided by the number of macromolecular residues in the structure to provide a comparative, normalized per-residue value. The more negative the value,

the more energetically favorable the AMBER-refined structure was. Further analysis was performed by visual analysis of the structures.

2.4 Data and discussion

2.4.1 Example with 2oiu

For first-time users of the command-line interface for PHENIX, the process of building and refining a structure could easily be confusing. Also, some modifications need to be made with certain structures before they go through this process. In order to aid in reproducibility of results and provide instruction for further use of this interface, a structure had to be chosen as an example. PDB 2oiu[102] is the L1 Ribozyme Ligase circular adduct, which required extra work to be performed to make sure the circular nature of the RNA molecules was retained. Thus, it made a perfect choice for a demonstration. The following is a walkthrough of the entire process used in this chapter for building and refining this structure.

2.4.1.1 Accessing, modifying, and creating needed files

The PDB file and *mtz* files were fetched from the PDB using the *phenix.fetch_pdb* command with the *--mtz* flag in order to download the reflections file. This could have also been done by accessing *www.rcsb.org* on the Internet, searching for the structure in question via its PDB ID, clicking on the "Download Files" section, and finding the desired files, but the command was used both to keep in line with the command-line nature of the work, and because of the simplicity of using a single command. The point-and-click method of downloading is useful if an error occurs with your terminal's connection to the PDB and your work is time-sensitive. After running the *fetch_pdb* command, the working directory contained the deposited PDB file for 2oiu as well as the experimental reflections file.

```
jgg75@casegroup5:~/2oiu/phenix_refinement/example$ phenix.fetch_pdb 2oiu --mtz
Model saved to /home/jgg75/2oiu/phenix_refinement/example/2oiu.pdb
```

```
...
Converted structure factors saved to 2oiu.mtz
```

Then, because the ligase structure is a circular adduct, the leaprc.RNA.OL3 file, the file that sets up the RNA force field used by default in the PHENIX/AMBER interface and found in $PHENIX/conda_base/dat/leap/cmd/, was modified to remove the residue name mapping for 5'- and 3'-terminal residues. This way, the "terminal" residues, called such to indicate they are the residues at the beginning and end of the chain designation in the PDB file, would not be built as actual terminal residues, since there is no beginning or end of a circle. The following is the section of the standard leaprc.RNA.OL3 file that enumerates the names for terminal RNA nucleotides.

```
#
#       Define the PDB name map for the nucleic acids
#
addPdbResMap {
  { 0 "G" "G5"  } { 1 "G" "G3"  }
  { 0 "A" "A5"  } { 1 "A" "A3"  }
  { 0 "C" "C5"  } { 1 "C" "C3"  }
  { 0 "U" "U5"  } { 1 "U" "U3"  }
```

In the modified leaprc.RNA.OL3, which was copied into the working directory to supersede the standard file when loaded into *LEaP* (via *tleap*), the same section was edited to appear as follows to avoid the building of terminal nucleotides.

```
#
#       Define the PDB name map for the nucleic acids
#
addPdbResMap {
  {"G" "G"  }
  {"A" "A"  }
  {"C" "C"  }
  {"U" "U"  }
```

Without the residue mapping for the terminal residues, those "terminal" residues are read as mid-chain residues with connections at both the phosphorus and O3' atoms.

The other key to building the circular RNA molecule is explicitly connecting the "end" residue's O3' atom to the phosphorus of the "beginning" residue. The *myuclinks* file was developed to force the bonding of the 5'- and 3'-"ends" of the RNA molecules to make them circular within the *AmberPrep* process. If a file exists within the working directory with the filename *myuclinks*, it will be identified by the *AmberPrep* run and included in the *tleap* input file for building of the unit cell parameter and coordinate files. This includes any special commands that one would like to include, such as the 4 *bond* commands below, used to connect the "head" and "tail" of the 4 RNA molecules in the unit cell.

```
logFile leap.log
source leaprc.protein.ff14SB
source leaprc.DNA.OL15
source leaprc.RNA.OL3
source leaprc.water.tip3p
source leaprc.gaff2
set default nocenter on
set default reorder_residues off
x = loadpdb 2oiu_4tleap_uc.pdb
bond x.71.O3' x.1.P
bond x.142.O3' x.72.P
bond x.246.O3' x.176.P
bond x.317.O3' x.247.P
set x box { 45.29  100.018  71.93 }
saveAmberParm x 2oiu_uc.prmtop 2oiu_uc.rst7
quit
```

2.4.1.2 Building of the structure

Once any pre-building preparation was finished, the actual building of the structure was performed by running the *phenix.AmberPrep* command with the flag calling for AMBER minimization of all atoms. The following is the output, which shows the whole process of the production of the 4phenix_2oiu.pdb file, as well as the 4amber_2oiu.prmtop, 4amber_2oiu.rst7, and 4amber_2oiu.order files as described in subsection 2.3.2:

```
jgg75@casegroup5:~/2oiu/phenix_refinement/example$
phenix.AmberPrep 2oiu.pdb minimise=amber_all
================================================
Running pdb4amber on 2oiu.pdb
================================================
================================================
Summary of pdb4amber for: 2oiu.pdb
================================================
----------Chains
The following (original) chains have been found:
P
Q
---------- Alternate Locations (Original Residues!))
The following residues had alternate locations:
None
---------- Missing heavy atom(s)
None
================================================
Setting up library files for non-standard residues
================================================
================================================
Preparing asu files and 4phenix_2oiu.pdb
================================================
| ~> /home/jgg75/phenix-1.16-3546/build/../conda_base/bin/tleap
-f 2oiu_asu_tleap_input_run
Checking output filenames
  file : 2oiu_asu.prmtop
  file : 2oiu_asu.rst7
| ~> /home/jgg75/phenix-1.16-3546/build/../conda_base/bin/ChBox
-c 2oiu_asu.rst7 -o 2oiu_asu.rst7 -X 45.29 -Y 100.018 -Z 71.93 -al 90.0 -bt 104.42 -gm 90.0
============================================================
Preparing unit cell files: 4amber_2oiu.prmtop and 4amber_2oiu.rst7
============================================================
================================================
Running pdb4amber on 2oiu_4tleap_uc1.pdb
================================================
================================================
Summary of pdb4amber for: 2oiu_4tleap_uc1.pdb
================================================
----------Chains
The following (original) chains have been found:
P
Q
```

```
a
b
c
d
e
---------- Alternate Locations (Original Residues!))
The following residues had alternate locations:
None
---------- Missing heavy atom(s)
None
| ~> /home/jgg75/phenix-1.16-3546/build/../conda_base/bin/tleap
-f 2oiu_uc_tleap_input_run
Checking output filenames
  file : 2oiu_uc.prmtop
  file : 2oiu_uc.rst7
| ~> /home/jgg75/phenix-1.16-3546/build/../conda_base/bin/ChBox
-c 2oiu_uc.rst7 -o 2oiu_uc.rst7 -X 45.29 -Y 100.018 -Z 71.93 -al 90.0 -bt 104.42 -gm 90.0
4amber_2oiu.prmtop
================================================
Minimizing input coordinates.
================================================
| ~> /home/jgg75/phenix-1.16-3546/build/../conda_base/bin/sander
-O -i 2oiu_amber_all.in -p 4amber_2oiu.prmtop -c 4amber_2oiu.rst7
-o 2oiu.min.out -ref 4amber_2oiu.rst7 -r 4amber_2oiu.min.rst7
checking special positions in 4phenix_2oiu.pdb
================================================
Done. Four new files have been made:
     4phenix_2oiu.pdb
     4amber_2oiu.rst7
     4amber_2oiu.prmtop
     4amber_2oiu.order
================================================
Example
  phenix.refine 4phenix_2oiu.pdb use_amber=True \
    amber.topology_file_name=4amber_2oiu.prmtop \
    amber.coordinate_file_name=4amber_2oiu.rst7 \
    amber.order_file_name=4amber_2oiu.order \
    ....(other refinement keywords here).....
```

While pre-building preparation was made to properly build the circular RNA molecules, it was still important to check that the bonds were actually made in the output structure files. If not, the O3' and P atoms in the "tail" and "head" residues

would be free to move away from each other, and the structure would not be properly refined. To check for the existence of these bonds, *parmed* was run on the unit cell *prmtop* and *rst7* files. The *printBonds* command was performed on all O3' and P bonds, and the bonds in question appeared in the output (truncated below; the bonds in question have P as atom 1), so the workflow was continued.

```
jgg75@casegroup5:~/2oiu/phenix_refinement/example$ parmed 4amber_2oiu.prmtop
ParmEd: a Parameter file Editor
Loaded Amber topology file 4amber_2oiu.prmtop
Reading input from STDIN...
> loadCoordinates 4amber_2oiu.rst7
Adding coordinates to 4amber_2oiu.prmtop from 4amber_2oiu.rst7
> printBonds @O3' @P
             Atom 1              Atom 2       R eq   Frc Cnst   Distance     Energy
     34  O3' (  OS)      35    P (   P)     1.6100   230.0000     1.6051     0.0055
      1    P (   P)    2297  O3' (  OS)     1.6100   230.0000     1.5985     0.0305
     68  O3' (  OS)      69    P (   P)     1.6100   230.0000     1.6049     0.0061
    101  O3' (  OS)     102    P (   P)     1.6100   230.0000     1.6007     0.0200
    132  O3' (  OS)     133    P (   P)     1.6100   230.0000     1.5927     0.0687
...
   2298    P (   P)    4594  O3' (  OS)     1.6100   230.0000     1.5592     0.5945
...
   4680    P (   P)    6976  O3' (  OS)     1.6100   230.0000     1.5985     0.0305
...
   6977    P (   P)    9273  O3' (  OS)     1.6100   230.0000     1.5592     0.5945
...
> quit
Done!
```

A conventional refinement was performed to develop an *mtz* file with r-free flags. As the downloaded reflections files for most of the chosen structures did not include r-free flags, these were generated for each structure using the below input file. The "*refinement.input.xray_ data.r_free_flags.generate=True*" option generated those r-free flags and put them into the output *mtz* file with the prefix "cdl_start". This file was then moved to 2oiu_data.mtz right after the generation of the flags at the beginning of the refinement using the *mv* command, as shown in the input file, subsequent running of the input file, and the selected portions of the output:

```
run_cnew.sh
```

```
#!/bin/sh
#   "standard"  CDL refinement
if [ "$#" -ne 1 ]; then
   echo "Usage:  run_cdl.sh <pdb-id>"
   exit 1
fi
phenix.refine  \
   4phenix_$1.pdb $1.mtz \
   c_beta_restraints=False discard_psi_phi=False \
   strategy=individual_sites+individual_adp+occupancies \
   flip_symmetric_amino_acids=True \
   refinement.target_weights.optimize_xyz_weight=True \
   refinement.input.xray_data.r_free_flags.generate=True \
   refinement.main.number_of_macro_cycles=10 \
   prefix=cdl_start serial=1 \
   write_geo=False cdl=True
/bin/mv cdl_start_data.mtz $1_data.mtz
jgg75@casegroup5:~/2oiu/phenix_refinement/example$ ./run_cnew.sh 2oiu
================================ X-ray data ================================
...
Generating a new array of R-free flags.
Miller array info: R-free-flags
Observation type: None
Type of data: bool, size=19049
Type of sigmas: None
Number of Miller indices: 19049
Anomalous flag: False
Unit cell: (45.29, 100.018, 71.93, 90, 104.42, 90)
Space group: P 1 21 1 (No. 4)
Systematic absences: 0
Centric reflections: 691
Resolution range: 31.3421 2.60002
Completeness in resolution range: 0.994155
Completeness with d_max=infinity: 0.993429
Test (R-free flags) flag value: 1
Number of work/free reflections by resolution:
                                        work  free  %free
  bin  1: 31.3444 -  5.5920 [1886/1946]  1707   179   9.5%
  bin  2:  5.5920 -  4.4426 [1918/1936]  1719   199  10.4%
  bin  3:  4.4426 -  3.8822 [1913/1917]  1719   194  10.1%
  bin  4:  3.8822 -  3.5278 [1922/1924]  1737   185   9.6%
  bin  5:  3.5278 -  3.2752 [1890/1896]  1695   195  10.3%
  bin  6:  3.2752 -  3.0823 [1913/1926]  1733   180   9.4%
```

```
  bin  7:  3.0823 -  2.9280 [1914/1916]  1711   203  10.6%
  bin  8:  2.9280 -  2.8007 [1878/1880]  1694   184   9.8%
  bin  9:  2.8007 -  2.6929 [1907/1909]  1718   189   9.9%
  bin 10:  2.6929 -  2.6000 [1908/1911]  1704   204  10.7%
                                overall 17137  1912  10.0%
...
No array of experimental phases found.
Writing MTZ file:   /home/jgg75/2oiu/phenix_refinement/example/cdl_start_data.mtz
...
```

A look at the files in the working directory shows a cdl_start_data.mtz that is an output file, but 2oiu_data.mtz appears at the beginning of the refinement run, showing that it is the input reflections file with the r-free flags generated:

```
-rw-r--r-- 1 jgg75 case   369885 Oct 22 14:10 4phenix_2oiu.pdb
-rw-r--r-- 1 jgg75 case      585 Oct 22 14:27 parmed.log
-rw-r--r-- 1 jgg75 case   459656 Oct 22 14:52 2oiu_data.mtz
-rw-r--r-- 1 jgg75 case    33040 Oct 22 14:52 cdl_start_001.eff
-rw-r--r-- 1 jgg75 case   383872 Oct 22 15:33 cdl_start_001.pdb
-rw-r--r-- 1 jgg75 case   423733 Oct 22 15:33 cdl_start_001.cif
-rw-r--r-- 1 jgg75 case  1232480 Oct 22 15:34 cdl_start_001.mtz
-rw-r--r-- 1 jgg75 case    33218 Oct 22 15:34 cdl_start_002.def
-rw-r--r-- 1 jgg75 case   108744 Oct 22 15:34 cdl_start_001.log
```

This *mtz* file was then used as the input *mtz* file for the parallel AMBER and conventional refinements to provide consistent reflection data and proper r-free flags.

2.4.1.3 Parallel refinements with AMBER and conventional geometric restraints

As the focus of this chapter was to compare the resultant structures from refinement with AMBER force field restraints and conventional restraints, the next step was to perform those refinements with the output files from *AmberPrep*. The AMBER refinement was run using the following input file. This script is set up to allow running of subsequent runs using the previous AMBER refinement's output PDB file as the input for the next run by changing the serial number of the run. It uses the *mtz* file with the generated r-free flags as the common reflections file with the conventional refinement, and calls for use of AMBER restraints, and thus *sander* minimization,

using the 4amber_2oiu.prmtop and 4amber_2oiu.rst7 files to provide AMBER atom types and coordinates for use in *sander*. It also calls for the printing of AMBER energies from within each cycle of *sander* minimization. To run the refinement, the command at the bottom of this section was executed, where 2oiu is the PDB ID used to call for use of the proper input files in this generalized script, and 1 is the serial number, making sure that the script calls for a first refinement of 2oiu using the 4phenix_2oiu.pdb file as the input file and results in amber_001.pdb as its output structure file.

```
#!/bin/bash
#  run a "standard Amber refinement
if [ "$#" -ne 2 ]; then
   echo "Usage:  run_amber.sh <pdbid> <serial no.>"
   exit 1
fi
new=$(printf '%03d' "$2")
old=$(printf '%03d' $(($new - 1)) )
if [ "$new" -eq 1 ]; then
  inpdb="4phenix_$1.pdb"
else
  # inpdb="amber_$old.pdb"
  inpdb="4phenix_$1.pdb"
fi
phenix.refine  \
   $inpdb $1_data.mtz \
   c_beta_restraints=False discard_psi_phi=False \
   strategy=individual_sites+individual_adp+occupancies \
   refinement.main.number_of_macro_cycles=10 \
   flip_symmetric_amino_acids=True nqh_flips=True \
   refinement.target_weights.optimize_xyz_weight=False \
   fix_wxc=1.0 wc=1.6667 \
   use_amber=True \
   amber.topology_file_name=4amber_$1.prmtop \
   amber.coordinate_file_name=4amber_$1.rst7 \
   amber.order_file_name=4amber_$1.order \
   print_amber_energies=True \
   prefix=amber serial=$new \
   write_geo=False --overwrite cdl=True
grep 'Amber total' amber_$new.alog | tail -1 > lastenergy
/bin/mv amber_$new.alog amber_$new.log
```

```
cat lastenergy >> amber_$new.log
/bin/rm -f lastenergy
jgg75@casegroup5:~/2oiu/phenix_refinement/example$ ./run_amber.sh 2oiu 1
```

The following is an excerpt from the AMBER refinement output log. It shows the main difference in AMBER and conventional refinements: the use of *sander* in coordinate refinement. The AMBER energies for each minimization step in the first macrocycle are shown, as well as the starting r-factors and clashscore, etc.

```
============================== Initializing AMBER ==============================
  topology    : 4amber_2oiu.prmtop
  atom order  : 4amber_2oiu.order
  coordinates : 4amber_2oiu.rst7
  Amber total: -52743.46 bonds (n=10080):   662.20 angles (n=17736):  2057.28
  diheds (n=36328):  6480.96 elec.: -58554.83 vdW: -3389.07
...
====================== XYZ individual (reciprocal space) ======================
    R-FACTORS           RMSD     CLASH  RAMA  ROTA CBET WEIGHT    TARGETS
 work  free  delta  bonds  angl                                  data   restr
  Amber total: -52743.46 bonds (n=10080):   662.20 angles (n=17736):  2057.28
  diheds (n=36328):  6480.96 elec.: -58554.83 vdW: -3389.07
27.10 27.20   0.10  0.015   2.2    1.7   0.0   0.0    0   none  0.077 -0.6738
  ...
  Amber total: -54417.19 bonds (n=10080):   624.96 angles (n=17736):  2192.23
  diheds (n=36328):  6603.02 elec.: -60527.26 vdW: -3310.15
26.38 28.20   1.81  0.015   2.5    1.1   0.0   0.0    0  1.000  0.071 -0.7123
Legend:
  - first line corresponds to starting state (before refinement)
  - R-factors reported in percent
  - delta is Rfree-Rwork in percent
  - CLASH is all-atom MolProbity clashscore
  - ROTA is percent of side-chain rotamer outliers
  - RAMA is percent of Ramachandran plot outliers
  - CBET is number of Cbeta deviations
  - WEIGHT is relative weight between X-ray (or neutron) target and restraints
  - TARGETS: the values of X-ray (or neutron) and restraints target functions
```

In Figure 2.1, the graph shows the AMBER total energy value found in the AMBER refinement output log over each step of minimization throughout 10 macrocycles of differing refinements (a couple of the "steps" are just the statement of the starting energy

and the ending energies at the beginning and end of the refinement). While this particular refinement used a combination of least-squares and maximum-likelihood targets, the graphs in the figure used all least-squares, all maximum-likelihood, and least-squares with optimization of the target weights, respectively. (Least-squares target functions are used in refinements that seek to minimize the squared weighted difference between experimental and model structure factors[17, 88, 1], while maximum-likelihood targets are used to refine structures to a point that maximizes the likelihood of having observed the experimental data given the structural model[88, 1]). In each case, the spikes at the beginning of each macrocycle are due to PHENIX's adjustment of hydrogens, which usually does not handle hydrogens in an energetically favorable manner. Also, with the weight optimization graph, the numerous spikes are due to each macrocycle involving running minimization with differing settings for the weighting options to find the best structure.

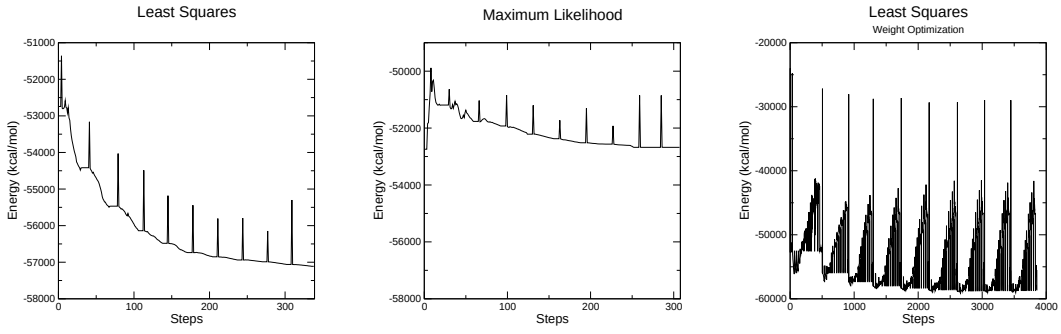

Figure 2.1: Graphs of the AMBER total energy throughout refinements of 2oiu using least-squares, maximum-likelihood, and least-squares with weight optimization minimizers. The energy appears to reach the lowest value with least-squares with weight optimization, but takes a lot longer to complete refinement.

Among these targets, both of the least-squares refinements result in much lower energies than the maximum-likelihood refinement does. This occurs with a large trade-off in r-factors, as the r-work/r-free for the maximum-likelihood refinement is 0.1913/0.2409, while the lowest for either least-squares run is 0.2204/0.2589 for the non-weight-optimization refinement. This is likely due to the setup for the *wc* variable: this appears to work well for maximum-likelihood while not doing as well for least-squares refinements. In fact, in the standard AMBER refinements used in this chapter, the majority of macrocycles

use maximum-likelihood and wc=1.667.

The ending energies for all three refinements in the figure are only reached after 8-10 macrocycles, indicating that 5 macrocycles would not be a good standard number of macrocycles to use with AMBER force field restraints in order to reach convergence. The 10-macrocycle standard used in this chapter appears to be a good place to start. The energy profile for the weight optimization run results in a similar endpoint as least-squares with a fixed wc value, but with an even lower ending energy, but higher r-factors. This weight optimization works to find sets of weight values that optimize multiple factors, including r-factors and geometric outliers. This appears to also optimize the ability of the least-squares target to lead to a lower minimum.

The next step, although it was usually run simultaneously, was to perform the conventional refinement of the same structure. Because the bonds completing the circle of the RNA molecules were not stored in the PDB file but in the *prmtop* file, which is not used as input for the conventional refinement, a custom restraints file had to be created and included in the input to force those bonds to exist and keep the atoms within the proper distance from each other. The restraints file (2oiu.eff) is seen below. Unlike the *myuclinks* file, this file only includes 2 bond additions, because the input PDB file only contains the asymmetric unit.

```
refinement.geometry_restraints.edits {
   bond {
        action = *add
        atom_selection_1 = chain P and resid 1 and name P
        atom_selection_2 = chain P and resid 71 and name O3'
        distance_ideal = 1.61
        sigma = 0.015
        slack = None
   }
}
refinement.geometry_restraints.edits {
   bond {
        action = *add
        atom_selection_1 = chain Q and resid 1 and name P
        atom_selection_2 = chain Q and resid 71 and name O3'
        distance_ideal = 1.61
        sigma = 0.015
```

```
            slack = None
        }
    }
```

The conventional refinement script below calls for the same parameters of refinement as the AMBER refinement, other than the use of AMBER restraints and the fixed weighting of the geometric restraints versus the x-ray data. As mentioned above, the other main difference here is that the *prmtop* and *rst7* files were not used as input for conventional refinement.

```
#!/bin/sh
#   "standard"  CDL refinement
if [ "$#" -ne 1 ]; then
   echo "Usage:  run_cdl.sh <pdb-id>"
   exit 1
fi
phenix.refine  \
   4phenix_$1.pdb $1_data.mtz 2oiu.eff\
   c_beta_restraints=False discard_psi_phi=False \
   strategy=individual_sites+individual_adp+occupancies \
   flip_symmetric_amino_acids=True nqh_flips=True \
   refinement.main.number_of_macro_cycles=10 \
   prefix=cdl_restrained serial=1 \
   write_geo=False cdl=True
```

The command below runs the conventional refinement script. As with the AMBER refinement, 2oiu is used as an argument along with the command running the script. Below that is the coordinate refinement portion of the first macrocycle from the conventional refinement output log file. No AMBER energies are printed, as there is no call to *sander* for minimization. While the starting structures are the same for AMBER and conventional refinements (the r-factors, clashscore, etc., are all the same), the targets themselves are already different at the beginning, likely due to the difference in geometric restraints and the different weighting scheme used in the AMBER refinement. Also, the ending structures in the macrocycle were already noticeably different, as observed through the r-factors.

```
jgg75@casegroup5:~/2oiu/phenix_refinement/example$ ./run_cdl.sh 2oiu
...
```

```
====================== XYZ individual (reciprocal space) ======================
    R-FACTORS           RMSD     CLASH  RAMA  ROTA CBET WEIGHT    TARGETS
 work  free  delta  bonds  angl                                    data   restr
27.10 27.20   0.10  0.015   2.2    1.7   0.0   0.0    0   none  0.077  1.2702
25.21 27.28   2.07  0.009   1.4   11.1   0.0   0.0    0  9.170  0.065  0.0943
Legend:
  - first line corresponds to starting state (before refinement)
  - R-factors reported in percent
  - delta is Rfree-Rwork in percent
  - CLASH is all-atom MolProbity clashscore
  - ROTA is percent of side-chain rotamer outliers
  - RAMA is percent of Ramachandran plot outliers
  - CBET is number of Cbeta deviations
  - WEIGHT is relative weight between X-ray (or neutron) target and restraints
  - TARGETS: the values of X-ray (or neutron) and restraints target functions
```

2.4.1.4 Geometric analysis of the parallel refinements

The resulting output structures from the parallel AMBER and conventional refinements were then analyzed using *phenix.molprobity* to look at the RNA geometric outliers and more general PDB statistics. If the structure contained modified residues or ligands, the CIF files would need to be included in the command line before the "$>$". As there were no ligands or modified residues in this structure, this option was not used. The selected sections of the output shown below are the sections that were focused on for comparison of the refinement methods across the data set. The geometric outliers examined were only those found in the RNA via the RNA validation portion. These could also be found by using *phenix.rna_validate*, but that command's output does not also include the overall PDB statistics. As seen below, in 2oiu, the bond length, bond angle, and pucker outliers were more numerous in AMBER refinement, but the torsion suites were actually improved in the AMBER-refined structure as compared to conventional refinement. Within the PDB statistics, the r-factors were fairly similar, but the AMBER clashscore was much improved over that of the conventionally refined structure.

```
jgg75@casegroup5:~/2oiu/phenix_refinement/example$
phenix.molprobity amber_001.pdb > amber.out
...
```

```
================================ RNA validation ================================
                  ----------Backbone bond lengths----------
  ...
  10/1986 bond outliers present
                   ----------Backbone bond angles----------
  ...
  225/3116 angle outliers present
                       ----------Sugar pucker----------
  ...
  3/142 pucker outliers present
                 ----------Backbone torsion suites----------
  ...
  10 suites triaged and 0 incomplete leaving 132 suites
  19/142 suite outliers present
  Average suiteness: 0.528
================================== Summary ==================================
  Ramachandran outliers =   0.00 %
                favored =   0.00 %
  Rotamer outliers      =   0.00 %
  C-beta deviations     =     0
  Clashscore            =   1.97
  RMS(bonds)            =   0.0188
  RMS(angles)           =   2.84
  MolProbity score      =   2.11
  Resolution            =   2.60
  R-work                =   0.1943
  R-free                =   0.2388
  Refinement program    = PHENIX
jgg75@casegroup5:~/2oiu/phenix_refinement/example$
phenix.molprobity cdl_restrained_001.pdb > cdlrest.out
================================ RNA validation ================================
                   ----------Backbone bond angles----------
  ...
  1/3116 angle outliers present
                 ----------Backbone torsion suites----------
  ...
  13 suites triaged and 0 incomplete leaving 129 suites
  22/142 suite outliers present
  Average suiteness: 0.522
================================== Summary ==================================
  Ramachandran outliers =   0.00 %
                favored =   0.00 %
  Rotamer outliers      =   0.00 %
```

```
C-beta deviations     =     0
Clashscore            =   9.13
RMS(bonds)            =   0.0083
RMS(angles)           =   1.57
MolProbity score      =   2.64
Resolution            =   2.60
R-work                =   0.1945
R-free                =   0.2320
Refinement program    = PHENIX
```

2.4.1.5 Energy Analysis

As the addition of energetic terms to the restraints is the most important improvement provided by interfacing PHENIX and AMBER, it was important to compare the energies of the output structures. In order to do so, the output structures had the water molecules and magnesium ions removed (and all CONECT cards, etc., that referenced them), and then also the hydrogen atoms. The riding hydrogens used in conventional refinement were found to be very energetically unfavorable, resulting in very large differences between refinement types. As hydrogens are not placeable via electron density in all but the very highest resolution structures, it is fair to say that riding hydrogens are mainly used to minimize r-factors, and thus are not a useful part of analyzing improvements upon conventional refinement via AMBER-restrained refinement. It is assumed that, because energetic terms are used in AMBER-restrained refinement, the hydrogen atoms in the AMBER output structure are more energetically favorable than the conventionally refined ones, and thus can be dismissed to get a better view of the effects of AMBER restraints on refinement of the heavy atoms of the macromolecules. The hydrogen atoms were removed from each output structure using the *pdb4amber* utility with the *-y* flag, which removes all hydrogen atoms.

```
jgg75@casegroup5:~/2oiu/phenix_refinement/example/energy$
pdb4amber -i amber_001.pdb -o amber_001_noH.pdb -y
```

The output PDB from this command, as expected, resulted in a structure without hydrogen atoms, as seen below.

```
ATOM      1  P   G   P   1      -2.494 -23.551 -16.078  1.00 81.50           P
ATOM      2  OP1 G   P   1      -1.020 -23.651 -16.003  1.00 94.82           O
ATOM      3  OP2 G   P   1      -3.365 -23.953 -14.968  1.00 60.76           O
ATOM      4  O5' G   P   1      -2.824 -22.049 -16.621  1.00 73.28           O
ATOM      5  C5' G   P   1      -1.897 -21.405 -17.486  1.00 75.67           C
ATOM      6  C4' G   P   1      -2.380 -20.115 -18.173  1.00 72.60           C
ATOM      7  O4' G   P   1      -3.507 -20.296 -19.028  1.00 75.69           O
ATOM      8  C3' G   P   1      -2.733 -19.009 -17.194  1.00 79.95           C
ATOM      9  O3' G   P   1      -1.564 -18.263 -16.846  1.00 87.51           O
ATOM     10  C2' G   P   1      -3.736 -18.218 -18.057  1.00 76.18           C
ATOM     11  O2' G   P   1      -3.112 -17.364 -19.007  1.00 72.83           O
ATOM     12  C1' G   P   1      -4.484 -19.304 -18.788  1.00 67.97           C
ATOM     13  N9  G   P   1      -5.639 -19.773 -17.976  1.00 63.69           N
ATOM     14  C8  G   P   1      -5.775 -20.802 -17.053  1.00 72.58           C
ATOM     15  N7  G   P   1      -6.979 -20.951 -16.559  1.00 61.00           N
ATOM     16  C5  G   P   1      -7.710 -19.938 -17.187  1.00 61.01           C
ATOM     17  C6  G   P   1      -9.110 -19.572 -17.102  1.00 62.03           C
ATOM     18  O6  G   P   1     -10.022 -20.054 -16.422  1.00 60.69           O
ATOM     19  N1  G   P   1      -9.446 -18.492 -17.911  1.00 62.24           N
ATOM     20  C2  G   P   1      -8.558 -17.828 -18.714  1.00 71.60           C
ATOM     21  N2  G   P   1      -9.038 -16.754 -19.313  1.00 69.42           N
ATOM     22  N3  G   P   1      -7.246 -18.143 -18.847  1.00 69.67           N
ATOM     23  C4  G   P   1      -6.890 -19.212 -18.050  1.00 65.46           C
ATOM     24  P   G   P   2      -1.258 -17.683 -15.343  1.00 88.23           P
```

The next step is to again modify the *myuclinks* file with the proper input file names and the proper atoms for the *bond* command, as there are no water molecules or magnesium ions in this structure, resulting in differing residue numbers. The modified *myuclinks* file is below.

```
...
x = loadpdb amber_001_noH_4tleap_uc.pdb
bond x.71.O3' x.1.P
bond x.142.O3' x.72.P
bond x.213.O3' x.143.P
bond x.284.O3' x.214.P
set x box { 45.29  100.018  71.93 }
saveAmberParm x amber_001_noH_uc.prmtop amber_001_noH_uc.rst7
quit
```

To perform the 0-step MD run in *sander*, the unit cell *prmtop* and *rst7* files were required to be built from the output structure PDB files. The *phenix.AmberPrep* run

here for each of the output structures requires the added flag of *use_reduce=False* to prevent *reduce* from adding hydrogens within *pdb4amber* (the default). Without this flag, the rebuilding of the conventional refinement output structure usually resulted in some error where *reduce* tried to add a hydrogen atom to a residue that did not require that particular hydrogen atom, causing *tleap* to fail. Including this flag resulted in all hydrogens being added by *tleap* as required by each residue type. Also, as these hydrogen atoms were added purely based on parameters specifically made for atom types, not including any real interactions with the rest of the structure, the *minimise=amber_h* flag was used to result in minimization of just the new hydrogen atoms. This was only done after modifying the AmberPrep.py code to change the restraint method for restraining the heavy atoms during this minimization. The use of *ntr* restraints was not strong enough to keep the heavy atoms in place without fluctuation, and thus *ibelly* was used. The *bellymask*–unlike the *restraintmask*, which describes the set of atoms to be restrained–describes the atoms to be allowed to move. Below are the *sander* minimization inputs for the original and modified AmberPrep.py codes, respectively.

```
inputs = {"amber_h" : """Initial minimization
  &cntrl
   ntwx   = 0, ntb    = 1, cut    = 9.0,     nsnb   = 10,
   ntr    = 1, restraint_wt = 50.0, restraintmask ='!@H=',
   imin   = 1, maxcyc = 1000, ncyc   = 200, ntmin  = 1, ntxo = 1,
  /
  """,
inputs = {"amber_h" : """Initial minimization
  &cntrl
   ntwx   = 0, ntb    = 1, cut    = 9.0,  nsnb   = 10,
   ibelly = 1, restraint_wt = 50.0, bellymask ='@H=',
   imin   = 1, maxcyc = 100, ncyc   = 200, ntmin  = 2, ntxo = 1,
  /
  """,
```

The command below was used to build the AMBER-refined output structure. The output to the terminal was similar to that of the original *AmberPrep* run for 2oiu, so that is omitted here.

```
jgg75@casegroup5:~/2oiu/phenix_refinement/example/energy$
```

```
phenix.AmberPrep amber_001_noH.pdb minimise=amber_h use_reduce=False
```

Once the *prmtop* and *rst7* files were created for both refinement output structures, they were run through *sander* for a 0-step MD calculation with the following command and input file.

```
sander -O -i energy.in -p 4amber_amber_001_noH.prmtop
-ref 4amber_amber_001_noH.rst7 -c 4amber_amber_001_noH.rst7
-o amber_001_noH_en.out  -xvv ~/rismKCL.xvv
unit cell
 &cntrl
    ntx=1, ntpr=1, ntwx=0, ntwr=0
    ioutfm=1
    imin=0, drms=1E-4, nstlim=0, maxcyc=0,
    ig=314159
    ntb=1
    irism=1
    cut=9.0
 /
 &rism
    periodic='pme'
    closure='kh'
    buffer=1, grdspc=0.5,0.5,0.5
    solvcut=9.0
    npropagate=0
    mdiis_del=0.5, mdiis_nvec=10, tolerance=1e-6
    apply_rism_force=0  /
```

The command called the MD engine *sander*, with a call to overwrite existing output files with the same names (*-O*), use the *prmtop* and *rst7* files for the parameters and coordinates, output the MD output into a file called amber_001_noH_en.out, and use the file rismKCL.xvv as the bulk solvent description file for the RISM calculation. This file uses 100 mM KCl in water as the neutralizing salt to balance the negative charge of the RNA, and screens out some of the electrostatic energy improvements found in the AMBER structure due to its electrostatic interactions with the deposited solvent molecules. The input file was set up to use constant volume periodic boundary conditions, the particle mesh Ewald method of calculating the electrostatics for the full unit cell with periodic boundary conditions, and the Kovalenko-Hirata closure for the RISM calculation.

To get to the energy data used in this chapter, the output files (as seen in part below) were used to find the energy values for each energy component. For each of these components, including the total energy (EPtot), the conventional output was subtracted from the AMBER output and then divided by the number of macromolecular residues in the unit cell. The manipulation of the following data (AMBER output first, then conventional) as mentioned above results in the numbers found in the 2oiu line in Table 2.2. As can be seen both by looking at the raw data, and the normalized difference values in the table, the AMBER-refined structure was favorable across all of the energy components except for the bond angles and the RISM energy. Each of the labels below corresponds to a different component of the total energy. Etot is really the total energy, but in this case, since there is no kinetic energy (EKtot), the total potential energy (EPtot) is the same value as the total energy. BOND, ANGLE, and DIHED are the energy terms associated with energy penalties related to deviations from ideal bond lengths and angles and dihedral angles, respectively. 1-4 NB is the non-bonded term for interactions between the first and last atoms in a dihedral angle, 1-4 EEL is the electrostatic interaction term for those atoms, VDWAALS is the van der Waals term, EELEC is the electrostatics term, and ERISM is the energy associated with the RISM solvent distribution.

```
amber_001_noH.out
...
NSTEP =        0   TIME(PS) =       0.000  TEMP(K) =     0.00  PRESS =     0.0
Etot   =    -54692.4826  EKtot   =         0.0000  EPtot      =    -54692.4826
BOND   =       824.6225  ANGLE   =      2681.7445  DIHED      =      6604.8747
1-4 NB =      2846.3056  1-4 EEL =    -31537.9503  VDWAALS    =     -6098.3011
EELEC  =    -21496.7750  ERISM   =     -8517.0035  RESTRAINT  =         0.0000
cdl_001_noH.out
...
NSTEP =        0   TIME(PS) =       0.000  TEMP(K) =     0.00  PRESS =     0.0
Etot   =    -52971.2120  EKtot   =         0.0000  EPtot      =    -52971.2120
BOND   =      1095.5831  ANGLE   =      2139.3299  DIHED      =      6657.0875
1-4 NB =      2981.0591  1-4 EEL =    -30461.0156  VDWAALS    =     -5038.2771
EELEC  =    -20451.6281  ERISM   =     -9893.3507  RESTRAINT  =         0.0000
```

2.4.2 Structure factors, energy, and clashscores

As the goal of this research was to determine the effects of AMBER force field restraints within refinement as compared to conventional refinement restraints, it was important to define statistics that would be good indicators of improvement and would be considered relevant. Three of the main statistics chosen were the structure factors (and gap between them, the r-gap; this is an indicator of overfitting if large), the clashscores, and the energy. Structure factors and clashscores are standard statistics used in analysis of macromolecular models developed from crystallographic data and refined with refinement software. The energy was chosen because energetic terms are included in the restraints by using AMBER force fields. Therefore, the energy is an important statistic, as it would be a check on whether AMBER-restrained refinement is doing what is expected. (Brief aside: all trendlines in these graphs are used to guide the eyes, not as accurate regression lines.)

When looking at those three sets of statistics data, overall improvement using AMBER restraints over conventional restraints can be surmised, but is not quite obvious. When looking at r-work values (Figure 2.2) for refinements with AMBER versus conventional restraints, the conventional refinements generally result in values that are better than those for AMBER refinements by 1 or 2 percentage points. However, this may be due to overfitting, as r-free values are fairly similar except at very low resolution, where the electron density being modeled provides more room for physically incorrect structures to fit the data. This overfitting possibility is also backed up by the decreased r-gap by AMBER refinement as compared to conventional refinement. It was not expected that AMBER-restrained refinement would result in improvements of structure factors as compared to conventional refinement, as conventional refinement generally had a higher weight on the crystallographic data in the target function. However, the fact that the AMBER r-free values are as close as they are to those for the conventional refinement implies that the difference in how the two restraint sets fit the models into the data is small enough that the possible trade-off for improvement of the other statistics would be worthwhile.

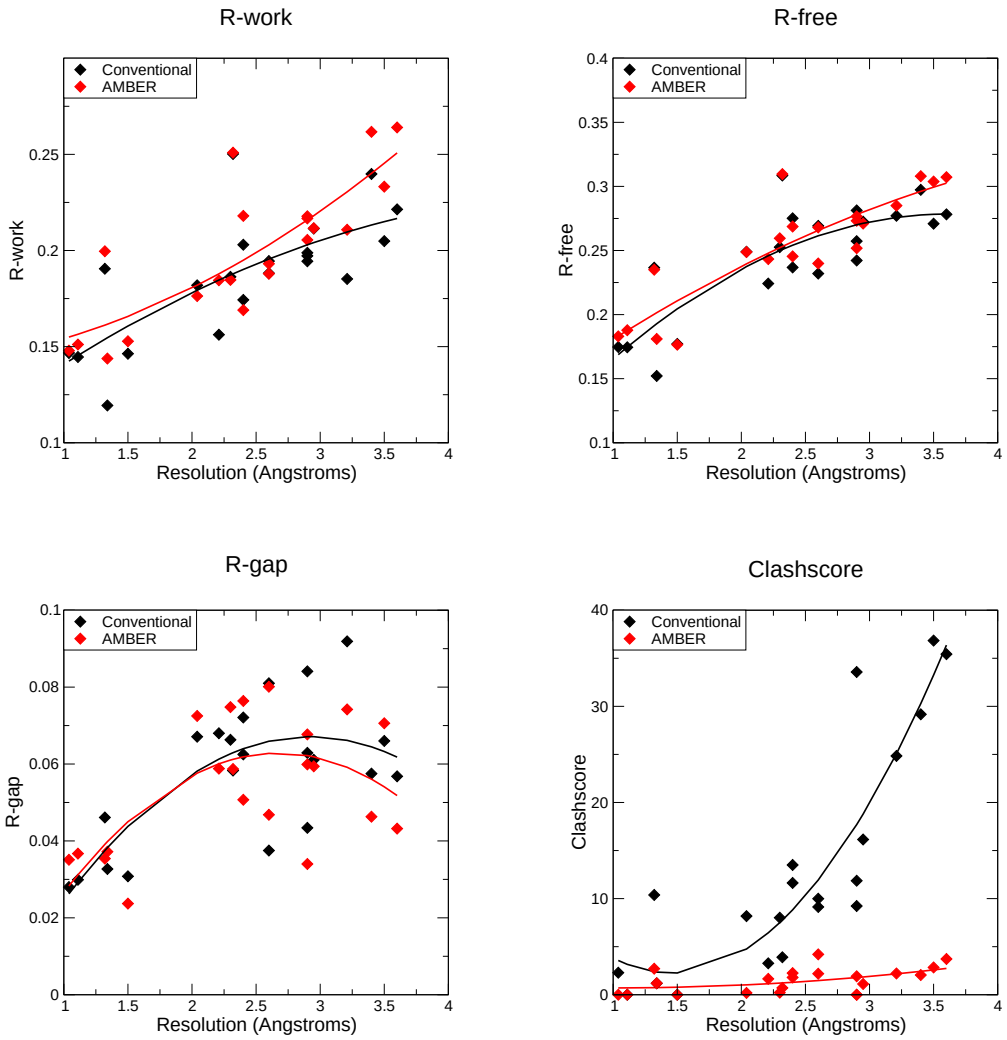

Figure 2.2: Graphs comparing the structure factors, r-gap, and clashscore for the AMBER-restrained and conventionally restrained refinements. Structure factors are similar for both sets of restraints, and the r-gap is somewhat improved for low-resolution AMBER structures. The largest difference occurs in clashscores, where AMBER-restrained refinements result in nearly similar clashscore values across the data set, while conventional refinement results in very high clashscores at low resolution.

With clashscores and energies, this is an obvious trade-off (Figure 2.2 and Tables 2.2 and 2.3). The clashscores for very high resolution structures are very low regardless of what restraints are used, due to the fact that the resolution is high enough to properly distinguish between atoms and place them in proper locations for bonds that do not result in clashes. The rest of the structures show far lower clashscores for the structures refined with AMBER restraints as compared to their conventional refinement counterparts. This was to be expected, as the use of AMBER restraints included the use of a van der Waals term in the energy calculation, putting a very large energy penalty on clashes. Also, the gap at the lowest resolutions got larger between conventional refinement clashscores and AMBER refinement clashscores. This matched the expectation that improvement would be greater at low resolution, where there is greater room for error in placing the model in the density and a greater chance of physical inaccuracy of the structure.

Regarding the energy, as one might expect when calculating AMBER energy with one set of restraints developed to minimize AMBER energies and another set of restraints not scored against the AMBER force field, the AMBER-restrained refinements resulted in far better total AMBER energy values than those with conventional restraints, as seen in Tables 2.2 and 2.3.

PDB	Res(Å)	B	A	D	E	F	W	V	R	EFR	TOT
1q9a	1.04	1.17	1.07	0.13	0.10	-2.03	-0.24	-0.54	0.36	-1.56	0.02
483d	1.11	0.48	0.24	0.13	-0.55	-1.96	-0.27	-0.55	0.83	-1.68	-1.65
4fe5	1.32	0.76	-0.34	-0.17	-1.89	-3.06	-0.72	-1.11	2.60	-2.35	-3.92
2a43	1.34	0.69	0.22	-0.14	-0.54	-3.21	-0.48	-0.81	1.37	-2.39	-2.91
480d	1.50	1.16	0.16	-0.01	-0.45	-3.14	-0.52	-0.63	1.40	-2.19	-2.03
2gdi	2.04	-1.95	0.39	0.14	-2.44	-3.94	-0.36	-2.38	3.63	-2.75	-6.91
3tzr	2.21	0.75	-2.10	0.24	0.22	-3.51	-0.70	-2.46	1.59	-1.70	-5.98
2pn4	2.32	0.17	0.85	-0.01	-3.92	-3.45	-0.19	-2.89	4.54	-2.83	-4.90
2qus	2.40	-0.73	1.33	0.22	-3.19	-3.34	-0.06	-4.18	3.53	-3.01	-6.43
3gx5	2.40	-2.17	-1.05	-1.16	-1.65	-3.95	-1.17	-4.78	3.70	-1.91	-12.24
2oiu	2.60	0.80	2.15	-0.12	-3.73	-3.73	-0.44	-3.54	4.92	-2.54	-3.70
2ygh	2.60	-2.04	-0.27	-0.42	-3.95	-3.62	-0.57	-3.86	4.73	-2.84	-10.00
2gis	2.90	-2.81	-2.97	-1.37	-2.92	-3.42	-2.07	-12.05	5.06	-1.28	-22.54
2pn3	2.90	-1.34	-0.44	-0.64	-6.71	-3.35	-0.60	-5.20	6.86	-3.21	-11.44
3e5e	2.90	-5.81	-2.18	-0.62	-4.42	-5.13	-1.37	-6.54	7.63	-1.91	-18.44
3f2q	2.95	-0.40	0.61	-1.04	-0.46	-5.85	-1.16	-6.21	5.09	-1.21	-9.42
3r4f	3.50	0.10	1.01	-0.08	-9.53	-2.84	-1.02	-7.16	8.90	-3.47	-10.61
1y0q	3.60	-0.29	0.38	0.13	-35.00	-4.29	-2.23	-17.98	36.33	-2.96	-22.95

Table 2.2: Per-nucleotide energy difference values (kcal/mol) between AMBER- and conventionally restrained PHENIX refinements for all components of the structures that contain RNA without the U1 protein. The conventional value is subtracted from the AMBER value, and, therefore, the more negative the value, the more favorable the AMBER-restrained structure is. As the resolution worsens, the improvement upon conventional refinement by use of AMBER-derived restraints becomes greater. (B=bond, A=angle, D=dihedral, E=electrostatics, F=1-4 electrostatics, W=1-4 non-bonded, V=van der Waals, R=RISM, EFR=E+F+R, TOT=EP Tot)

PDB	Res(Å)	B	A	D	E	F	W	V	R	EFR	TOT
3mxh	2.30	-1.80	0.35	0.19	-6.58	-2.72	-0.66	-3.00	7.36	-1.93	-6.85
3iwn	3.20	-0.95	-0.04	-0.54	-18.47	-3.26	-1.47	-11.99	18.87	-2.86	-17.85
3bo3	3.40	-2.95	-1.50	-0.74	-15.10	-4.87	-2.67	-9.33	18.35	-1.62	-18.81

Table 2.3: Per-residue energy difference values (kcal/mol) between AMBER- and conventionally restrained PHENIX refinements for all components of the structures that contain RNA with the U1 protein. The conventional value is subtracted from the AMBER value, and, therefore, the more negative the value, the more favorable the AMBER-restrained structure is. As the resolution worsens, the improvement upon conventional refinement by use of AMBER-derived restraints becomes greater. (B=bond, A=angle, D=dihedral, E=electrostatics, F=1-4 electrostatics, W=1-4 non-bonded, V=van der Waals, R=RISM, EFR=E+F+R, TOT=EP Tot)

The majority of the improvement, and the trend in greater improvement at lower resolution, appears to be due to the van der Waals contributions, which also accounts

for the greater improvement in clashscores at low resolution. These tables were developed by subtracting the energy values from the conventional structure from those of the AMBER structure and dividing by the number of macromolecular residues in the unit cell to provide normalization. With the inclusion of electrostatic interactions and the aforementioned van der Waals term, the emphasis on the AMBER restraints in the target function for refinements with AMBER increased the likelihood that the structure would be more energetically favorable than one using a stronger emphasis on electron density data and a restraint set without energy terms. The RISM energy was surprisingly more favorable for the conventionally refined structure. It also screened some of the electrostatics, as can be seen in the column for the combined electrostatics, 1-4 electrostatics, and RISM term. While not all structures are more favorable with AMBER restraints (1q9a is slightly more favorable with conventional refinement, likely due to the resolution being so high that most atoms are visible in the electron density, and, thus, the density provides very energetically favorable bond lengths, angles, etc.), there is a general trend of the total potential energy becoming more favorable as resolution worsens (Figure 2.3), which was both expected and hoped for as a major improvement due to implementation of the AMBER force fields. As mentioned above regarding clashscore improvement at low resolution, as the accuracy of the experimental data decreases, the refinements that are more strongly tethered to general physical restraints and the poor experimental data are less able to provide energetically favorable structures than the AMBER-restrained refinements where energetic terms are able to steer the structures away from bad clashes and physically inaccurate geometries.

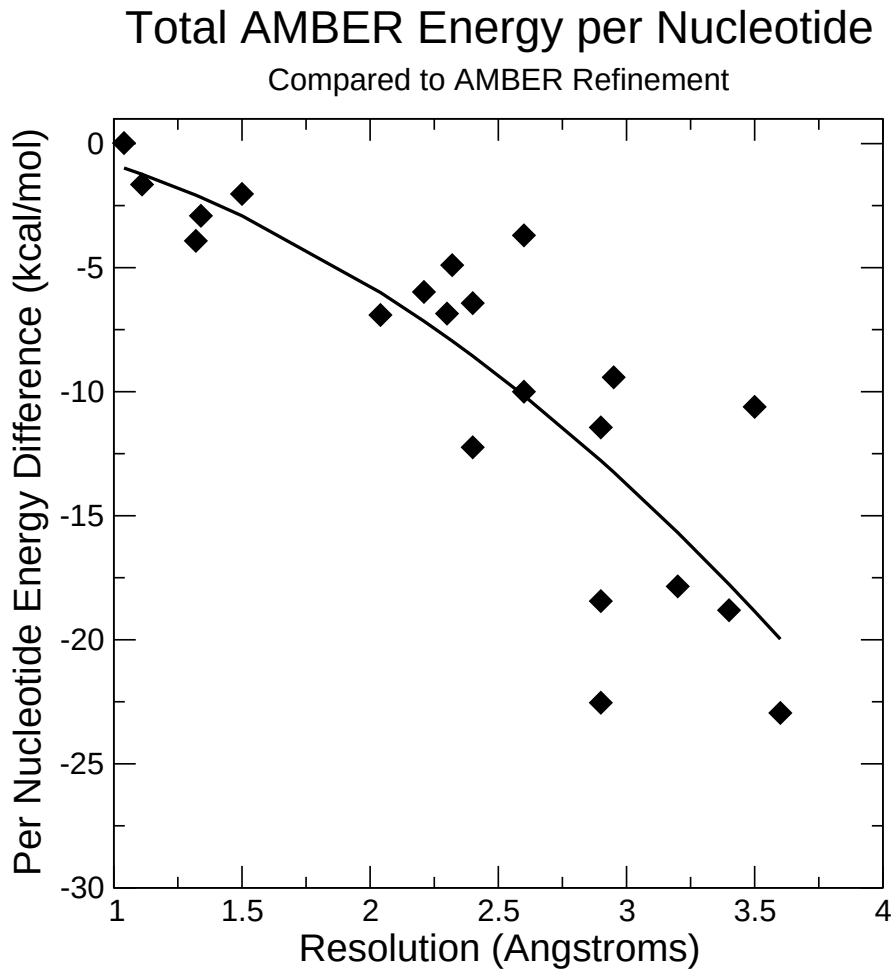

Figure 2.3: Graph of per-residue energy difference values (kcal/mol) for the total potential energy (EPtot) between AMBER- and conventionally restrained PHENIX refinements . The conventional value is subtracted from the AMBER value, and, therefore, the more negative the value, the more favorable the AMBER-restrained structure is. As the resolution worsens, the improvement found by using AMBER-derived restraints becomes more pronounced.

In order to compare the refinement restraint sets, structure factors, clashscores, and energy were chosen as benchmark statistics. Conventional restraints led to output structures with better structure factors than those for the AMBER-restrained refinement outputs. However, the r-free values for AMBER-restrained refinements are very close to those for the conventionally refined structures, and the slight differences are offset by the

vast improvements on energy and clashscores. Also, the gap between the r-work and r-free values, the r-gap, is lesser for AMBER-restrained structures than for conventionally restrained structures for the most part, indicating that there may be some overfitting of the conventionally restrained structures. Both clashscores and energy values exhibit a trend of increased difference between the conventionally refined and AMBER-restrained structures as the resolution worsens, which shows that the interface of PHENIX and AMBER provided the improvement that was expected in the target resolution area.

2.4.3 Differences in structures

Because AMBER force fields were implemented to improve the structural models developed in refinement, it was important to not only look at the energetics, but the structures themselves. This was done both visually and through RMSD calculations using *cpptraj*. With the large differences in energy, especially at low resolution, between conventional and AMBER refinements, one might expect the structures to be vastly different. On the other hand, due to the fact that the structural models were being fit to the same experimental data, one might also assume there would not be very large global structural differences. Therefore, it was a worthwhile analysis to conduct.

When heavy atom RMSD calculations comparing the solvent-stripped output PDB files from conventional and AMBER refinements (which contain one asymmetric unit each) were performed, the largest RMSD was around 0.6 Å. As seen in Figure 2.4, the majority of the structures actually had an RMSD of half that or less. There is a general increase in RMSD as the resolution worsens, as would be expected as the energy difference increased with resolution worsening.

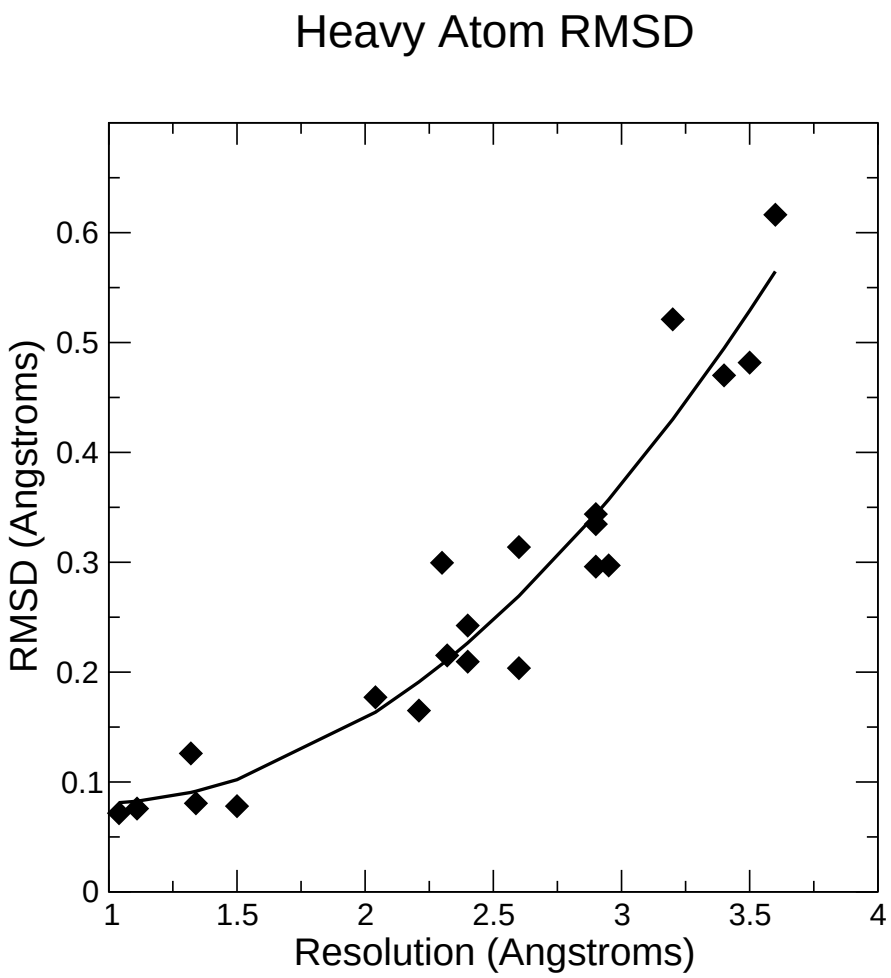

Figure 2.4: Graph of RMSD between conventionally and AMBER-refined structures. The RMSD increases as the resolution worsens, indicating that there are greater differences between the output structures as the quality of the data worsens.

When the structures that were above 0.4 Å regarding RMSD were examined (3iwn, 3bo3, 3r4f, and 1y0q; Figures 2.5, 2.6, 2.7, and 2.8, respectively), a general trend of slight global differences in the backbone of the RNA was found. There was also a slight difference in the ligands in structures that contained ligands, as seen in Figures 2.5 and 2.8. These images make it clear that while these differences were slight, they could very well be large enough to make a difference in choice of restraints worthwhile.

Here, it was found through both RMSD analysis and visual analysis that there were

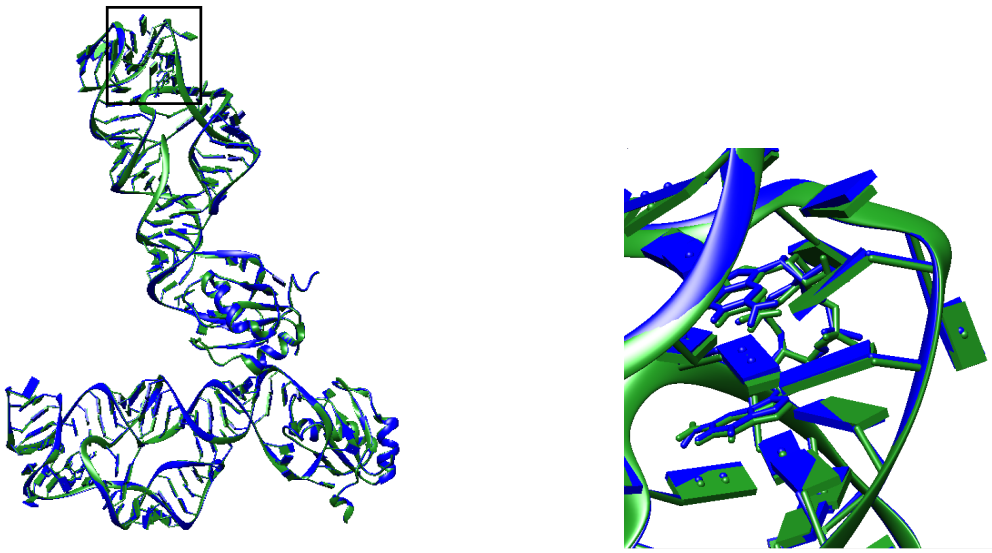

Figure 2.5: Images of global (left) and local (right) differences in the AMBER- (blue) and conventionally refined (green) structures of 3iwn[63]. The local image is a rotated look at the boxed area in the global image. Very little changes on the global scale, but there are some slight, but significant, changes at the local level, specifically in the ligand and bases around it.

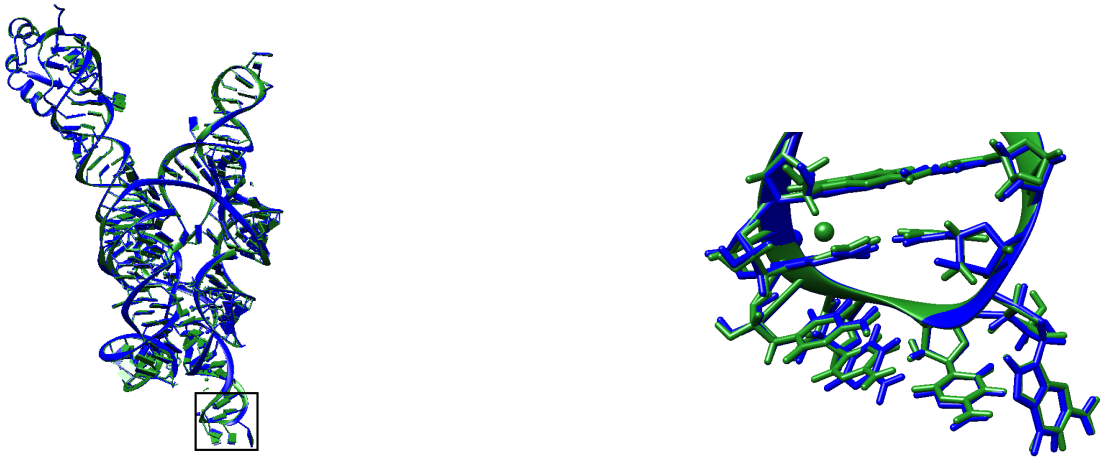

Figure 2.6: Images of global (left) and local (right) differences in the AMBER- (blue) and conventionally refined (green) structures of 3bo3[67]. The local image is a look at the boxed area in the global image. Very little changes on the global scale, but there are some slight, but significant, changes at the local level, specifically in the bases. There is also deviation in the magnesium ion locations, but these are not included in RMSD calculations.

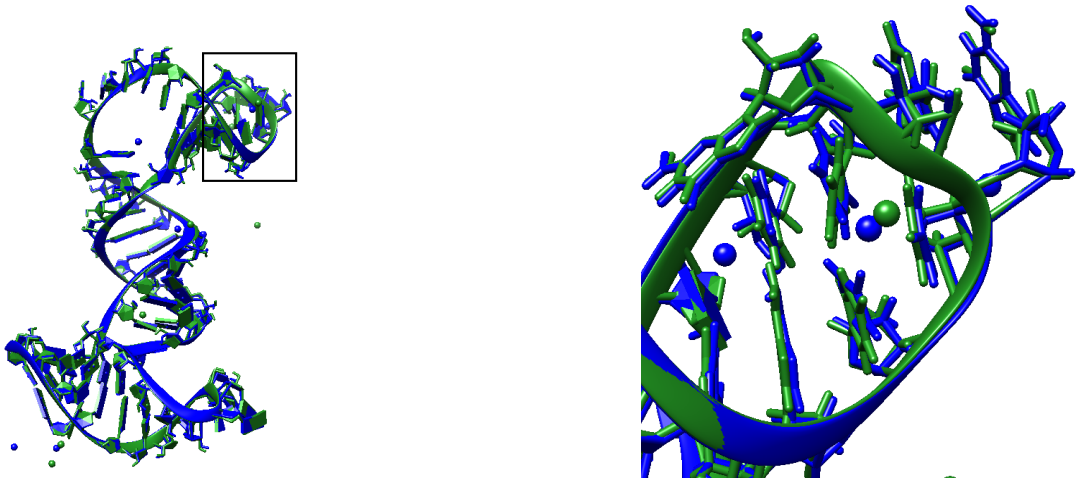

Figure 2.7: Images of global (left) and local (right) differences in the AMBER- (blue) and conventionally refined (green) structures of 3r4f[36]. The local image is a look at the boxed area in the global image. Very little changes on the global scale, but there are some slight, but significant, changes at the local level, specifically in the bases. Also, while not included in RMSD calculations, the magnesium ions are very differed in location between the structures.

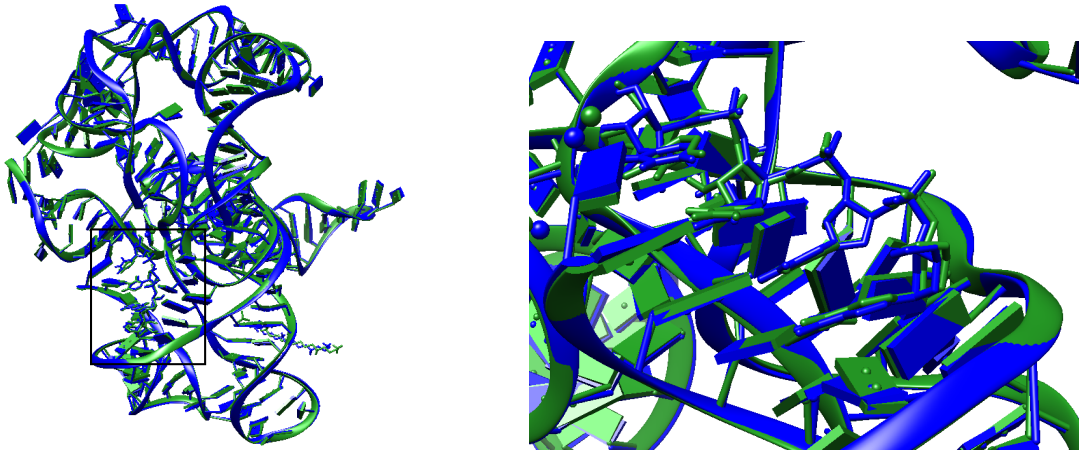

Figure 2.8: Images of the global (left) and local (right) differences in 1y0q[44] between the AMBER-(blue) and conventionally refined (green) structures. This is the structure with the largest RMSD between AMBER- and conventionally refined structures. A lot of the deviation appears to show up in the ligand and bases, as seen in the local image.

small global differences in the structures made by the differing restraint sets in refinement. However, especially at low resolution, these small differences were significant, especially when looked at regarding their effects at the local level.

2.4.4 Geometric outliers

As mentioned before, the statistics chosen to compare the output structures from the 2 different restraint sets included the energy, the clashscores, the structure factors, and the geometric outliers. While the energy differences and clashscores were in AMBER's favor and the structure factors are similar enough to accept the slight worsening of the agreement with experimental data for the improvements in energy and clashscore, the geometric outliers and suiteness scores were less complimentary to refinement using AMBER restraints (Table 2.4). This was surprising, as it seemed to follow reason that using energy terms to guide in the refinement would lead to bond lengths, angles, and torsion angles that would be most energetically favorable and, thus, ideal.

PDB ID	Resolution	Bond	Angle	Pucker	Suite	Suiteness
1q9a	1.04	0.028	0.10	0	0	-0.009
483d	1.11	0	0.046	0	0.0093	0.034
4fe5	1.32	0.056	0.082	0.0037	0	0.056
2a43	1.34	0	0.026	0	0	0.072
480d	1.50	0.028	0.037	0.0093	0	-0.015
2gdi	2.04	0.20	0.36	0.0047	0.0047	0.035
3tzr	2.21	0	0.13	0.0069	0	0.055
3mxh	2.30	0.11	0.79	0.022	0.011	0.061
2pn4	2.32	0.017	0.24	0.0028	-0.0028	0.034
2qus	2.40	0.16	1.93	0.015	-0.015	0.044
3gx5	2.40	0.0066	0.052	0.0053	-0.0027	-0.077
2oiu	2.60	0.035	0.87	0.011	-0.011	0.026
2ygh	2.60	0.012	0.11	0.0026	-0.0013	0.03
2gis	2.90	0.0040	0.076	-0.0013	-0.0040	-0.045
2pn3	2.90	0	0.060	0	0	0.001
3e5e	2.90	0.011	0.054	0.0012	-0.0012	0.06
3f2q	2.95	0.034	0.25	0.0030	-0.012	-0.046
3iwn	3.20	0.015	0.39	0.017	-0.0081	0.02
3bo3	3.40	0.0084	0.26	0.0023	-0.0034	0.026
3r4f	3.50	0.00095	0.12	0.0057	-0.0057	-0.019
1y0q	3.60	0.0054	0.24	0.0027	0.00054	0.107

Table 2.4: Per-nucleotide difference values for geometric statistics, and difference in suiteness score. The more negative the value, the more favorable AMBER restraints are for that particular geometric statistic. Generally speaking, the conventional refinements result in better geometric outlier numbers. At lower resolution, the AMBER-restrained refinements produce better suite outlier numbers.

The table shows a comparative look at the RNA geometric outlier numbers across the whole data set. The outlier difference values are created by subtracting the number of outliers of that type for the conventional refinement from the number for AMBER refinement, and then dividing that number by the number of nucleotides in that structure's unit cell. The suiteness difference value is the normalized difference found when subtracting the AMBER-restrained structure's value from the conventional refinement output value. (This opposite order for subtraction was set up to keep in order with the setup for energy differences: a negative difference means that AMBER is more favorable.)

A few trends appear in the geometric data as follows. First, all of the conventional structures compare favorably to the AMBER structures regarding bond outliers and angle outliers, as indicated by the positive normalized differences, even to the point in 2qus where there are 2 angle outliers per nucleotide more in the AMBER-restrained refinement output structure than in the conventional structure. This consistent difference was a little concerning, so it was dug into, with results in the next section. While there does not appear to be a resolution-dependent trend in the angle outlier data, it does seem the case that the normalized bond outlier difference decreases at lower resolution. When individual structure outlier numbers are studied, the AMBER-restrained refinement output structures have a lower proportion of bond length outliers at lower resolution, indicating that at low data quality, the AMBER restraints have less trouble providing bond lengths that match MolProbity's ideal bond lengths.

Second, sugar pucker outlier numbers are also generally better for conventional refinement, while the suite outliers are generally the same at high resolution, and the AMBER-refined structures were better for most of the rest of the structures, albeit only by one or two total outliers. Due to the fact that each suite spans from one d to the next, there are far fewer suites than bonds or angles, and thus a difference of 1 or 2 outliers results in a much larger proportional difference than it would in the bond or angle outlier differences.

Finally, the suiteness scores for the parallel refinements do not really appear to have a trend as to which restraint set results in a better score. For the most part,

the conventional refinement structures have better suiteness scores, but there are 6 structures where the AMBER refinement resulted in better suiteness scores, with no real resolution dependence for these "trends". The only real trend is that 4 of the 6 structures with better suiteness scores for the AMBER refinement have better suite outlier numbers for the AMBER refinement as well. It is surprising that far more of the structures with better suite outlier numbers for AMBER refinements have worse suiteness scores than their conventional counterparts when looked at at face value connecting the two statistics. However, suiteness is not just a measure of outliers, but is a measure of all the suites and how they fit the ideal torsions for their defined suites. Thus, while the conventional refinement structures showed more outliers at low resolution, the whole of the structures had better overall suites than those in the AMBER output.

As the geometric statistics chosen to analyze the differences between the output structures for each restraint type were observed, three things were found. First, the bond and angle outliers were consistently fewer for conventional restraints. Second, the torsion suite outliers were actually better for AMBER-restrained refinements at low resolution, but only by 1 or 2 outliers. Finally, while the suite outliers were worse for conventional refinement, the sugar puckers and the overall suite quality (suiteness score) for the conventional refinements were generally better than those for the AMBER-restrained refinements. These statistics indicated that the conventional restraints resulted in structures that better fit MolProbity's ideal values for the statistics studied. However, it seemed unlikely that AMBER restraints would result in such vastly poor structures regarding bonds and angles. Thus, examples were studied in the next section.

2.4.5 Investigation of outlier difference

When looking at output from MolProbity, it was determined that AMBER-restrained refinements resulted in consistently worse bond and angle outliers than their conventionally restrained counterparts. Within these outliers, there were some fairly consistent bonds and angles that appeared from structure to structure within AMBER-restrained refined structures. Some of these consistent outliers were examined more closely.

When analyzing these consistent outliers, it was found that the ideal values for some

of these bonds and angles were different from MolProbity to AMBER. To see whether this was the reason for the number of outliers, one particular bond and one particular angle were chosen to be studied. First, the AMBER ideal measurements were modified to see if that affected the overall number of outliers for the chosen bond and angle. In this case, the P-O5' bond was chosen, and the O3'-P-O5' bond angle was chosen. The ideal value for the bond length was set to 1.66 Å (while the standard AMBER ideal is 1.61 Å and the MolProbity ideal is 1.59 Å) within the RNA.OL3 force field in AMBER, and 2ygh[104] was rebuilt for refinement using *AmberPrep*. For another test on the effects of the parameters, the force constant was doubled from 230 to 460 and the structure was re*AmberPrep*ped using this modification. Then, AMBER and conventional refinements were performed with the rebuilt structures. Further *AmberPrep*ping of the output structures without minimization provided *prmtop* and *rst7* files for *parmed* analysis (using the *printBonds* command) of all the P-O5' bond lengths in the starting and output structures from each refinement. The values from *parmed* output were used to make histograms for the original refinement, the ideal length modification, and the force constant doubling. Further modification of the parameters to make the AMBER ideal equal to the ideal in MolProbity was also performed, and the resulting *AmberPrep*ped structure was also refined and analyzed. MolProbity analysis showed a change in the number of outliers, and a change in which bonds/angles were outliers and whether the outliers were higher or lower than the ideal (Table 2.5, Figure 2.9).

RT	Ideal	FC	Out	A Out	A<MP	A>MP	B Out	B<MP	B>MP
PR	1.61	230	11	0	0	0	0	0	0
CV	1.61	230	1	0	0	0	0	0	0
AM	1.61	230	10	1	1	0	1	1	0
PR	1.66	230	14	2	0	2	1	0	1
CV	1.66	230	1	0	0	0	0	0	0
AM	1.66	230	14	4	0	4	2	0	2
PR	1.61	460	11	0	0	0	0	0	0
CV	1.61	460	1	0	0	0	0	0	0
AM	1.61	460	7	0	0	0	0	0	0
PR	1.59	230	11	0	0	0	0	0	0
CV	1.59	230	1	0	0	0	0	0	0
AM	1.59	230	14	3	3	0	3	3	0

Table 2.5: Table looking at P-O5' (bond A) bond outliers as a result of different modifications to the AMBER ideal bond length (Ideal) and bond force constant (FC) for bonds of the P-OS bond types. The other bond of that type, O3'-P (bond B), was also examined (MP=MolProbity ideal bond length of 1.59 Å for P-O5', 1.61 Å for O3'-P). As the AMBER ideal value is changed, the location of outliers relative to the MolProbity ideal changes, indicating that the differences in ideal values affects the number of outliers. Increasing the force constant removes all outliers of A and B, indicating that the distribution of bonds might be narrowing. (RT=refinement type, PR=pre-refinement, CV=conventional, AM=AMBER, Out=bond outliers)

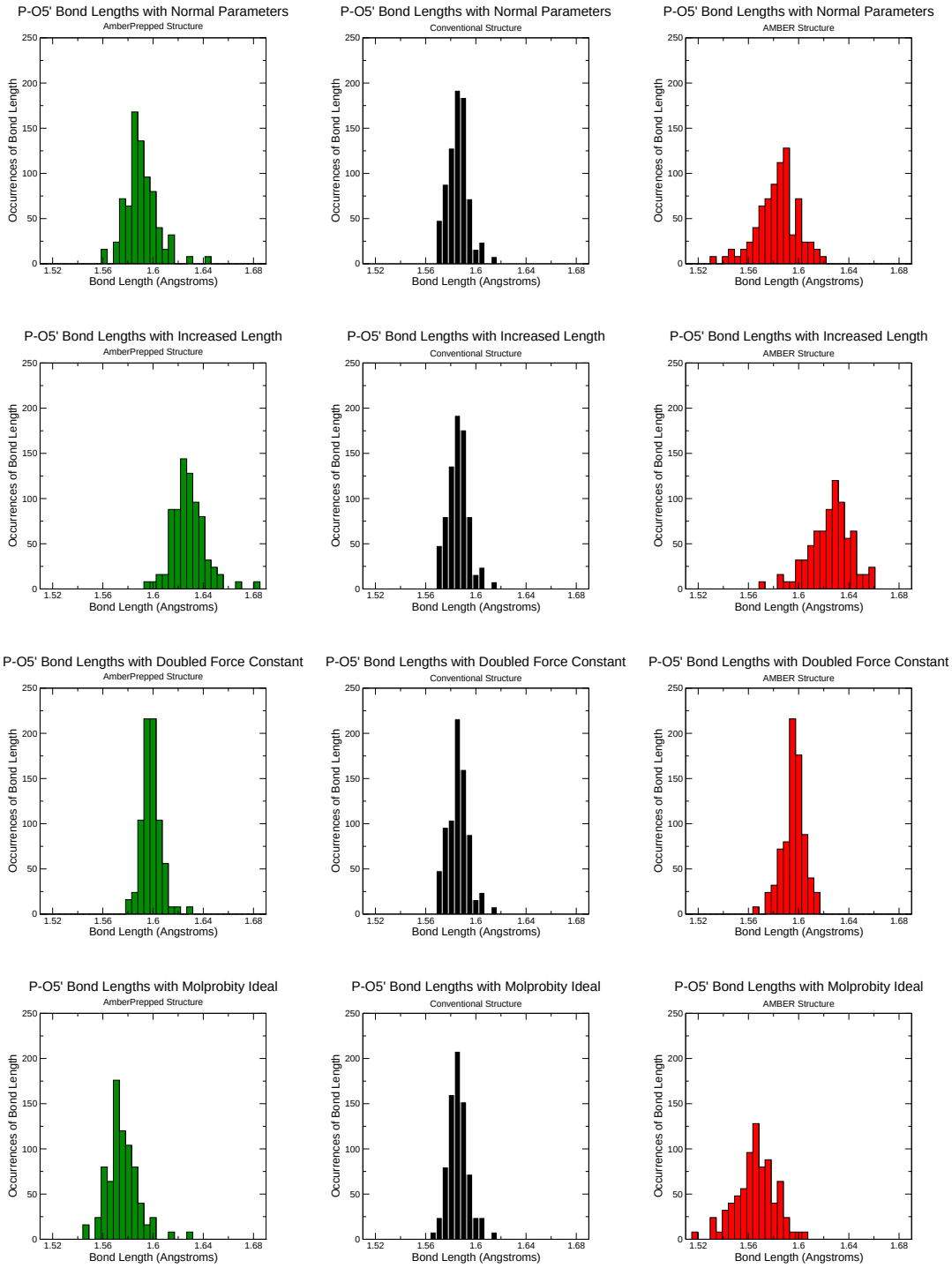

Figure 2.9: Histograms showing the distribution of all P-O5' bond lengths in the *AmberPrep*ped structure and the refinement output structures when prepared and refined with the standard AMBER parameters, longer AMBER ideal bond length, doubled bond force constant, and AMBER ideal bond length equal to MolProbity's (from top to bottom). AMBER-restrained refinements resulted in wider distributions with centers that shifted as the AMBER ideal was modified. The distribution narrowed with an increase in force constant. Regardless of the AMBER ideal, the conventional refinement distribution was nearly the same.

(It should be noted that "outliers" in this context are bonds with lengths that are 4 sigma or more away from the MolProbity value.) With the original refinement, the only outlier for P-O5' bonds was actually found to be below the MolProbity ideal, even though the AMBER ideal is higher than the MolProbity ideal. When the ideal length was increased, the number of outliers in AMBER refinement and the *AmberPrep*ped structure increased, and all of the P-O5' and O3'-P bond outliers were longer bonds than the MolProbity ideal value. When the ideal length was decreased to be equal with the MolProbity ideal, there were also more outliers for AMBER refinement than there were with the standard AMBER ideal bond length value, and the number of O3'-P bond length outlier also increased, and all of these outliers were also below the MolProbity ideal. This was a surprise, but maybe it should not have been, as the histograms (Figure 2.9) showed that in general, the range of the distribution of bond lengths was wider for AMBER refinement. This appears to have been due to an inherent lower level of weight placed on the restraint for this bond length in AMBER as compared to conventional refinement. With doubling of the AMBER force constant for this bond, that range decreased, and got rid of all P-O5' and O3'-P bond length outliers. Another key point is that the conventional refinement's distribution was roughly the same every time, even from different starting points. This makes sense, as the restraints being used for these refinements were the same every time.

As the histograms show, what appeared to happen was that the distribution of the values was wider with AMBER refinement due to the complexity of the energy equation and the number of different factors involved, and, while the distribution was not necessarily centered on the AMBER ideal value, the center of the distribution increased in bond length in the refinement with the increased AMBER ideal value. This led to the higher-end bonds getting moved further away from the MolProbity ideal, and thus made them the new outliers. When the MolProbity ideal was used as the AMBER ideal bond length, the center of the distribution shifted lower than the normal AMBER ideal (because the MolProbity ideal is lower than the AMBER ideal), but because the distribution was wider than that for the conventional refinement, there were still outliers. It appears that the overall reason for the larger number of outliers is due to the weaker

restraint weight (force constant) on the bond length in AMBER as compared to the conventional restraints, as evidenced by the disappearance of all P-O5' bond length outliers and the narrowing of the distribution when the force constant was doubled. The interconnectedness of all of the energy terms in AMBER allows for play in the bond lengths to provide for an energetically favorable and physically accurate structure at the expense of perfectly ideal bond lengths.

In the instance of looking at bond angle outliers, the O3'-P-O5' angle was modified in three separate ways: increasing the ideal angle from 102.6 degrees to 112.6 degrees, doubling the force constant from 45 to 90, and increasing the ideal angle to the MolProbity ideal of 104 degrees. After structure preparation and refinement as described above, the data was analyzed by *AmberPrep*ping the output structures and running *parmed* (*printAngles*) on the resultant *prmtop* and *rst7* files (Table 2.6).

RT	Ideal	FC	Out	A Out	A<MP	A>MP
PR	102.6	45	71	7	6	1
CV	102.6	45	3	0	0	0
AM	102.6	45	85	26	26	0
PR	112.6	45	65	1	0	1
CV	112.6	45	3	0	0	0
AM	112.6	45	79	8	0	8
PR	102.6	90	61	1	1	0
CV	102.6	90	3	0	0	0
AM	102.6	90	66	9	9	0
PR	104.0	45	67	5	4	1
CV	104.0	45	3	0	0	0
AM	104.0	45	68	16	16	0

Table 2.6: Table looking at O3'-P-O5' (angle A) bond angle outliers as a result of different modifications to the AMBER ideal bond angle (Ideal) and angle force constant (FC) for bond angles of the OS-P-OS bond types (MP=the MolProbity ideal value of 104 degrees for this angle). As with bond lengths, adjusting the force constant appeared to decrease the number of A outliers, likely due to a tighter adherence to the AMBER ideal value. Also, as the ideal value was increased, the outliers were all greater than the MolProbity ideal, as opposed to less than the MolProbity ideal when refined with the normal AMBER ideal value. (RT=refinement type, PR=pre-refinement, CV=conventional, AM=AMBER, Out=angle outliers)

As seen in the histograms in Figure 2.10, the distribution of the angle measurements for AMBER was much wider than that for the conventional refinement, as well as having

a different median of the distribution. The median of the AMBER distributions is always below the ideal value. (As with the bond lengths, I'm a little unsure why that is.) As with the bond length distributions, when the force constant was doubled, the range of the distribution was decreased. In this case, the distribution was still wider than the conventional refinement's, but that was likely due to the force constant for this angle, even when doubled, being much lower than that for the bond length.

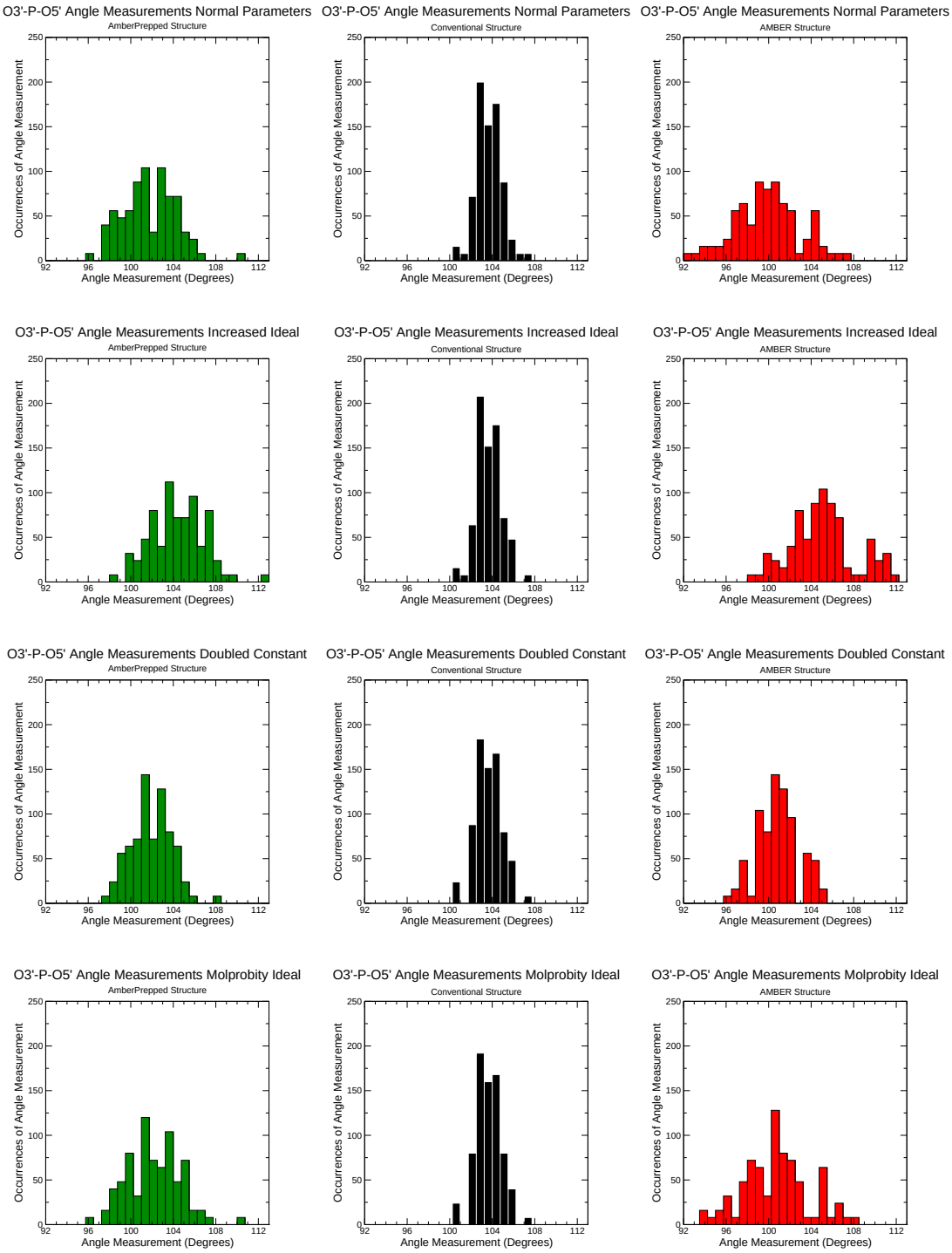

Figure 2.10: Histograms looking at O3'-P-O5' bond angles with differing ideal values and force constant values for the *AmberPrep*ped, conventional refinement, and AMBER refinement structures. As with Figure 2.9, the AMBER distributions were wider than the conventional ones, with shifting centers based on AMBER ideal, and narrowing of the distribution with an increase in the force constant.

As seen in Table 2.6, the O3'-P-O5' outliers for the AMBER output with the original parameters were all at measurements less than the MolProbity ideal. When the AMBER ideal angle measurement was increased, the whole distribution shifted higher in terms of angle measures while staying roughly the same in terms of distribution of angles. This led to a decrease of O3'-P-O5' outliers, and all of those outliers being at measures greater than the MolProbity ideal. This is due to the upper end of the distribution being shifted out of the 4 sigma range from the MolProbity ideal, and the outliers at the lower end of the original refinement being shifted into that 4 sigma range on the lower side of the MolProbity ideal value. The doubled force constant also resulted in a similar decrease of the outliers in question due to the shrinking of the distribution width, and all of those outliers were still on the lower side of the MolProbity ideal, likely due to the lower AMBER ideal angle measure for this angle type. Finally, with the AMBER ideal set to be equal to the MolProbity ideal of 104 degrees, the number of outliers of the selected angle decreased, and all of the outliers were still on the lower side of the MolProbity ideal. Even with the same ideal value as MolProbity, the AMBER-restrained refinement resulted in outliers of the angle in question, indicating that, as in the case of the bond length outliers, the outliers come from a combination of a much lower restraint weight on the angle measurement than conventional refinement, as well as a difference in the ideal value used by AMBER as opposed to the Engh and Huber restraints.

As the geometric outliers in Subsection 2.4.4 showed, the bonds and angles in AMBER-restrained refinement output structures provided far more outliers than their conventionally restrained counterparts. It was important to understand this in order to be able to determine if the differences were due to systematic errors or just freedom of movement inherent to the complex nature of AMBER force fields. To examine this, one consistent bond outlier (P-O5') and one consistent angle outlier (O3'-P-O5') were chosen for testing. For each measurement, the AMBER ideal value and the force constant were modified to test their effects on the number of outliers of that particular measurement, which side of the MolProbity ideal those outliers fell on, and the width of the distribution of that measurement. Due to the fact that in both cases, the number of outliers decreased as the force constant was doubled and when the AMBER ideal

value was changed to that of MolProbity, it was determined that the outliers in these cases were due to a combination of the greater flexibility allowed in these measurements when using AMBER force fields as compared to conventional restraints and (to a lesser extent) the difference in the ideal values between AMBER and MolProbity. Thus, at least for these 2 measurements, it is fair to say that the outliers are less an indication of poor structure quality within AMBER and more of a result of the interconnected nature of all parts of a structure that is less of a consideration via conventional restraints.

2.4.6 Hydrogen bonds

One final area of analysis that could point to the usefulness or lack thereof of AMBER-restrained refinement is hydrogen bonding. Presumably, AMBER energy terms would lead to increased hydrogen bonding due to the energetic favorability of such interactions. To look into this, specifically within the confines of base pairing within the RNA molecules, the *nastruct* command in *cpptraj* was performed at differing hydrogen bond cutoff distances to make sure that differences in hydrogen bond or base pair numbers were due to an actual lack of interaction, as opposed to just a very sharp distance cutoff. Table 2.7 shows the difference in base pairs and hydrogen bonds found in AMBER and conventional refinements for the first asymmetric unit of all the structures in the data set. Situations where a base in one asymmetric unit "base paired" with a base in another asymmetric unit were not counted as base pairs. Certain bases had base pairs with multiple bases within the first asymmetric unit. These were counted. Negative differences are indicative of AMBER having a larger number of base pairs or hydrogen bonds, and positive differences occur when the conventional refinement resulted in more base pairs or hydrogen bonds.

It was found to be the case that from high to mid-range resolution, there is no real difference in hydrogen bonding for base pairs. At low resolution, the hydrogen bonding and base pairing are generally better for AMBER-restrained refinements, but there are still not consistently large differences. This is likely due to the lack of large structural changes, as found in the RMSD calculations. In most cases where there was a difference in base pairs, but not hydrogen bonds, it was due to a base pair existing

PDB	Res(Å)	Cut=3.5 Å		Cut=3.6 Å		Cut=3.7 Å	
		BP	HB	BP	HB	BP	HB
1q9a	1.04	0	0	0	0	0	0
483d	1.11	0	0	0	0	0	0
4fe5	1.32	0	0	0	0	0	0
2a43	1.34	0	0	0	0	0	0
480d	1.50	0	0	0	0	0	0
2gdi	2.04	0	0	0	0	0	0
3tzr	2.21	-1	0	-1	0	-1	0
3mxh	2.30	0	0	0	0	0	0
2pn4	2.32	0	0	0	0	0	0
2qus	2.40	0	0	0	0	0	0
3gx5	2.40	1	0	1	0	1	0
2oiu	2.60	0	-2	0	-1	0	-1
2ygh	2.60	-1	-4	-1	-4	-1	-4
2gis	2.90	0	0	0	0	0	0
2pn3	2.90	0	0	0	0	0	0
3e5e	2.90	-1	0	-1	0	-1	-2
3f2q	2.95	-1	0	-1	0	-1	0
3iwn	3.20	2	-1	2	-3	2	-2
3bo3	3.40	1	-7	1	-4	1	-4
3r4f	3.50	0	1	0	1	0	-1
1y0q	3.60	-3	-6	-2	-6	0	-3

Table 2.7: The differences in the numbers of base pairs (BP) and hydrogen bonds (HB) in conventional and AMBER refinements at different hydrogen bond cutoff distances. Little is found in the way of differences at high- and mid-resolution, but in low-resolution structures improvements are found in hydrogen bonding with AMBER-restrained refinement.

without hydrogen bonds within the existing hydrogen bond parameters. In particular, 3e5e is a structure where the AMBER-restrained refinement resulted in a structure with one more base pair than that for the conventional refinement, but that base pair had no hydrogen bonds associated with it until the hydrogen bond cutoff was at 3.7 Å, and then it was deemed to have 1 hydrogen bond, and another base pair that both structures had in common picked up a hydrogen bond in the AMBER-restrained structure as well. Through further analysis (not shown in this table), that same base pair appeared in the conventional refinement output when the cutoff was set to 3.8 Å, along with 1 more hydrogen bond, and the structures became even in hydrogen bonds and base pairs at the 4.0 Å cutoff for the conventionally refined structure (did not go to that cutoff with AMBER structure). In 2oiu, the differing hydrogen bond numbers evened out by setting the cutoff for the conventionally refined structure at 4.0 Å. Only one structure (3r4f) resulted in fewer hydrogen bonds in the conventionally refined structure than in the AMBER-refined structure with cutoffs raised up to 4.0 Å for the conventional structure, and in that case, it was only 1 hydrogen bond different. So, while the general numbers show improvement with AMBER with the default cutoff and values near it, the interactions are not necessarily absent from the conventionally refined structures, but may just be weaker.

Finally, as the images for the low resolution structures with ligands showed changes in ligand orientation, it seemed important to see how the hydrogen bonding of those ligands was different between conventional and AMBER refinements. The results of the *hbond* command in *cpptraj* for the structures with ligands are found in Table 2.8. In almost all of the structures, the AMBER-refined structure has more hydrogen bonds involving the ligand, indicating more favorable energetic interactions with the ligands, and thus better binding. These numbers were found by counting the hydrogen bonds to the ligands in the first asymmetric unit for these structures, as found in the solute_avg_*.dat, solvent_avg_*.dat, and bridge_*.dat output by the command (specific parameters set up to get it to work this way), where * indicates the restraint type (AMBER or conventional). The cutoff from acceptor to donor heavy atom was the default 3.0 Å, and in cases where there were vast differences, the cutoff was extended

incrementally up to 4.0 Å to see whether the interaction was there, just at a longer distance. This was only done for 4fe5 and 2gdi, as it stood to reason at high resolution that if the hydrogen bonds existed in the AMBER-refined structure, they should probably do so in the conventional structure, even if at a slightly higher distance. In both cases, none of the missing hydrogen bonds appeared in the data files after extending the cutoff. This could either be due to slight differences in angles outside of the cutoff range for the hydrogen bond angle or due to a slight enough difference in these local areas, both in the ligand itself and the surrounding binding pocket atoms and solvent molecules, that the additive difference as compared to the AMBER structure resulted in the interactions not existing. This was examined for 4fe5, and in Figure 2.11, it can be seen that the missing hydrogen bonds found in the AMBER-restrained refinement output resulted from the change in orientation of a water molecule due to the influence of the electrostatics term in AMBER.

PDB ID	Resolution (Å)	AMBER	Conventional
4fe5	1.32	6	4
2gdi	2.04	11	8
3tzr	2.21	3	3
3mxh	2.30	17	11
3gx5	2.40	6	5
2ygh	2.60	8	6
2gis	2.90	8	4
3e5e	2.90	7	7
3f2q	2.95	7	6
3iwn	3.20	7	8

Table 2.8: Comparison of hydrogen bonds involving non-covalent ligands in AMBER- and conventionally refined structures. Generally, there are more hydrogen bonds in the AMBER-refined structures, but surprisingly the lowest-resolution structure in this group has 1 more hydrogen bond in the conventionally refined structure.

In the one structure where the conventionally refined structure has more hydrogen bonds with the ligand(s) than the AMBER-restrained refinement output structure, 3iwn, there was only a difference of one hydrogen bond. When the distance cutoff from hydrogen bond acceptor and hydrogen bond donor heavy atom was adjusted to 3.2 Å for both structures, they each had 13 hydrogen bonds, although they were different hydrogen bonds in some cases. For example, a hydrogen bond from chain B, residue

A 101's H62 atom (donated by N6) to O2' from ligand C2E in chain B exists within the default cutoff in the conventionally refined structure, but does not show up in the AMBER output structure at the default cutoff, when the number of hydrogen bonds is equal at a cutoff of 3.2 Å for both, or even at a cutoff of 4.0 Å for the AMBER output. Actually, a different hydrogen bond between the ligand and A 101 exists in the AMBER output at 3.2 Å cutoff between the N1 atom of A 101 and HN21 (donated by N2) from the C2E in chain B.

The hydrogen bonding of these structures seemed a reasonable measure of the effects of AMBER energy favorability being a driving force in the refinement of the AMBER-restrained refinements. The base-pair hydrogen bonding and ligand hydrogen bonding were studied in *cpptraj*, and the results were somewhat underwhelming. While there appeared to be reasonably significant differences in the ligand hydrogen bonding, (which, if substantiated, could result in better ligand binding and binding pocket understanding, and thus lead to the use of AMBER in refinement for drug discovery purposes, where improved description of binding site interactions is of premium importance,) the overall difference in base-pair hydrogen bonding was underwhelming. The higher-resolution half of the data set showed little to no differences at all, and the rest of the structures showed that, while the AMBER-refined structures resulted in more hydrogen bonds near the default cutoff distance, when the cutoff distance was expanded, conventional refinement was able to retain almost all of those hydrogen bonds. It seems at least encouraging, however, that there are decreased hydrogen bonding distances in those cases for the AMBER-restrained outputs, indicating stronger hydrogen bonds. It bears further examination as to the possible benefits of AMBER restraints in this particular area.

2.5 Conclusions

A study of the use of the PHENIX/AMBER interface on RNA structures has been presented here in comparison to conventional PHENIX refinement. The introduction of the AMBER force fields into PHENIX refinement provided marked improvement across the set of 21 RNA-containing structures in this study, especially in terms of energy and

clashscores, while giving away little in terms of r-free factors. The emphasis on van der Waals' and electrostatic energies from AMBER, which are not considered in conventional refinement restraints, in refining the structures led to structures that had far fewer clashes and had more energetically favorable, and physically accurate, interactions, including hydrogen bonds. While there were obvious improvements, the changes in the structures were generally not very large, as evidenced by a maximum RMSD of around 0.6 Å. The geometric statistics for the structures showed that conventional refinement resulted in structures that had fewer bond, angle, and sugar pucker outliers. However, it was found in a test case that P-O5' bonds and O3'-P-O5' bond angles, which were persistent outliers in AMBER-refined structures, were found to be outliers because of the difference in ideal values for these measures between MolProbity and AMBER, as well as an overall greater flexibility allowed in these two measures for AMBER-refined structures based on the force constant being used and the interplay of so many different energetic terms allowing for the finding of the overall best structure. Therefore, it was determined that these outliers were again a modest trade-off for the improvement in interactions being modeled with the AMBER-restrained refinement outputs.

Maybe more important than the general trend of improvement for AMBER-refined structures is the improvement in low resolution structures. As the resolution decreased, a marked trend of increase in energetic difference toward AMBER favorability, improvement in clashscores, improvement in the r-gap, and RMSD between differently refined structures (showing that AMBER found a considerably different, and likely better, structure) appeared, showing greater improvement at low resolution, where the lack of experimental data necessitates greater reliance on restraints. Also, hydrogen bonding between bases appeared to be slightly improved as compared to conventional refinement, with a few hydrogen bonds that were at least 0.3 Å closer in the AMBER-restrained output. As many RNA structures are solved at low resolution[27, 58], these improvements at low resolution should incentivize the use of this novel interface.

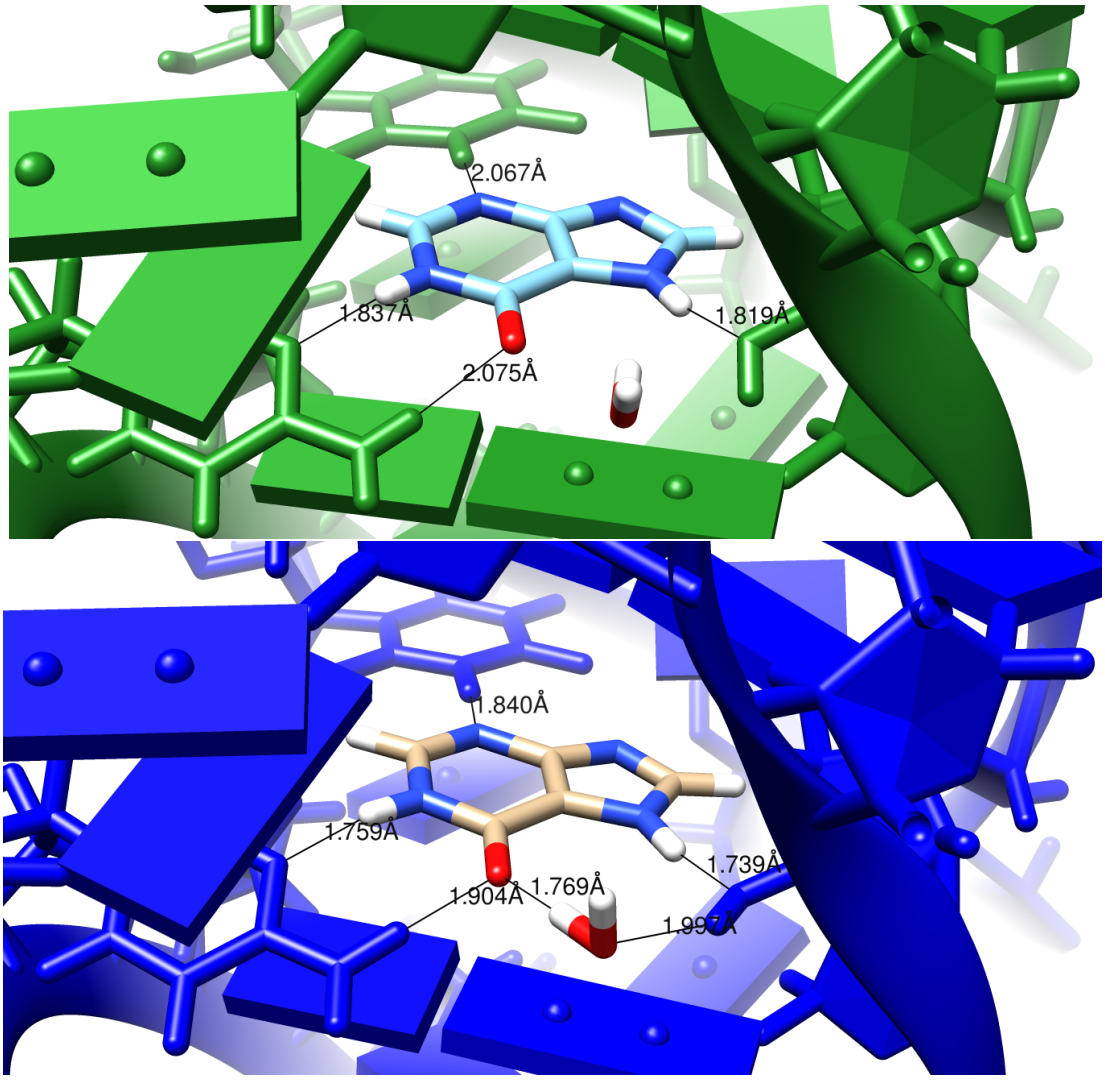

Figure 2.11: Images comparing conventionally and AMBER-restrained (green and blue, respectively) refinement outputs for 4fe5[10]. The measurements are from donor to acceptor, not donor heavy atom to acceptor. The difference here is in distances and also the additional hydrogen bonds involving a nearby solvent molecule in the AMBER image. The ligand being examined is hypoxanthine (HPA).

Chapter 3

Integral equation models for disordered solvent in macromolecular crystals

3.1 Summary

[This is a paper co-authored by George M. Giambaşu, Darrin M. York and David A. Case. My contribution resides in Table 3.2, where the three structures tested are from the data set in Chapter 2, as well as in the work that led up to the production of Figure 3.2.]

X-ray scattering measurements from macromolecular crystals are influenced by the solvent environment, but conventional refinement techniques use only very simplified models for water molecules (and other solvent species) that form the bulk of the solvent environment. Here we examine solvent distributions for a variety of crystals, computed using molecular dynamics or with a newly-developed periodic version of the integral equation (3D-RISM) codes in AMBER. Bragg intensities for both MD and RISM solvent models are in better agreement with experiment at all resolution ranges than are intensities computed using the default "flat" solvent model in the *refmac5* refinement programs, with the greatest improvement in the 1.5 to 2.5 Å range. Localized ions or water molecules match known behavior in many cases, and the total number of ions and water molecules can (in principle) be compared to crystal density measurements and atomic emission spectroscopy. The 3D-RISM solvent distributions can be derived in seconds (for unit cells that are roughly 50 Å on a side), and could be updated regularly during the course of crystallographic refinement. The new 3D-RISM codes provide the numerically accurate gradients required to use 3D-RISM as an implicit solvent model, and yield charge neutrality for charged solutes; this is an important consideration for

nucleic acids, where most or all of the counterions are part of the "disordered" solvent. Prospects for improving accuracy and incorporating integral equation models into crystallographic refinement are discussed.

3.2 Introduction

Ions and water molecules have been long known to play crucial roles in governing biomolecular stability and function. Elucidating how ions and water molecules distribute themselves around the solutes should provide valuable insights in the mechanism of how those molecules function, and also provide experimental tests for theoretical predictions. However, there are few methods that directly probe the positions of ions and water molecules around macromolecules. In solution, counts of the total number of excess waters and ions around a macromolecule can be obtained from atomic emission spectroscopy,[7, 41] small-angle X-ray scattering,[85, 86, 75, 82] and measurements of partial molar volumes.[22, 24, 23, 106] These techniques, however, give relatively little information about the distribution of water and ions in the vicinity of a biomolecule.

In principle, much more detailed information is available from X-ray diffraction studies on biomolecular crystals, and it is common to include some number of "bound" (or localized) waters and ions in a refined atomic model that has been optimized to fit observed scattering intensities. These locations are typically identified as features in a difference density map that satisfy criteria for both intensity (percent occupation) and geometry. Since it is common for biomolecular crystals to consist of 30-70% solvent,[73] locations that are favorable in a crystal lattice are also likely to be favorable in solution, and taking account of such positions can be an important component in computational schemes that analyze ligand-binding geometries and affinities.[47, 93] However, the "bound" solvent molecules generally make up only a small fraction of the total solvent; the remainder is typically modeled as a flat distribution, usually with density and B-factor components that are adjusted to optimize the fit of the total model to observed intensities. The limitations of such a flat-density model are thought to contribute to the "R-factor gap", which reflects the nearly universal observation that differences between computed and observed intensities in macromolecular crystallography are much greater

than the experimental uncertainties, prompting searches for better models.[50]

In this paper, we explore integral equation (3D-RISM) models for the solvent distribution in molecular crystals of proteins and nucleic acids. We present results from a newly-developed periodic version of the existing non-periodic 3D-RISM models in AMBER,[70, 71] as well as from molecular dynamics simulations of crystals with an explicit solvent. Particular attention is paid to the way in which charged solutes are handled, since there is more than one way to ensure electroneutrality of the entire unit cell, that is, to ensure that the distribution of ions in the solvent cancels the net charge of the solute.

3.3 Reference Interaction Site Model for periodic systems

The core principle of RISM is to find the single particle density distributions that minimize the excess chemical potential in response to an external potential arising from a molecular "solute". The basic idea, and the approximations involved, have been discussed many times,[70, 96] and we only give a brief summary here. In principle, the distribution of solvent molecules around a (fixed) solute is a six-dimensional quantity, describing the translation and orientations of the solvent molecules. The 3D-RISM formalism reduces these to three dimensions by decomposing polyatomic solvents (water molecules here) into atomic contributions, such that the resulting solvent density distributions contain only a spatial dependence, $\rho_\gamma(\mathbf{r})$, and can be represented by scalar densities on 3D grids. Here, the solvent index γ would range over H and O sites in water, and over mobile atomic ions such as Na^+ and Cl^-.

An Ornstein-Zernike-like equation relates the total correlation function, $h_\gamma(\mathbf{r}) = g_\gamma(\mathbf{r}) - 1$, and direct correlation function, $c_\gamma(\mathbf{r})$, through a convolution (denoted by $*$):

$$h_\gamma^{OZ}(\mathbf{r}) = \sum_\alpha c_\alpha(\mathbf{r}) * \chi_{\alpha\gamma}(r) \tag{3.1}$$

Here, $\chi_{\alpha\gamma}(r)$ is the site-site solvent-susceptibility of solvent sites α and γ and describes the orientationally averaged bulk properties of the solvent. These values are pre-computed (generally by a "1D-RISM" approach) for the reference solvent using the

dielectrically consistent RISM (DRISM) integral equation [91, 90]. Eq. 3.1 is augmented by a 3D closure relation:

$$h_\gamma^{closure}(\mathbf{r}) = \exp\left\{-\beta u_\gamma(\mathbf{r}) + h_\gamma^{OZ}(\mathbf{r}) - c_\gamma(\mathbf{r}) + b_\gamma(\mathbf{r})\right\} - 1 \tag{3.2}$$

where $b_\gamma(\mathbf{r})$ is the bridge function, which is only known as an infinite series of functionals and is always subject to some approximation[46]. Among the many closure relations that have been developed here we use family of closures related to the hypernetted chain (HNC) closure [80] where the bridge function is simply set to zero. HNC produces good results for ionic [51, 45, 95] and polar systems [49, 48] and has an exact, closed form expression for the excess chemical potential [105]. Since HNC solutions are often difficult to converge, one can use intermediaries such as the so-called partial series expansion of order-n (PSE-n) [56] of HNC as a Taylor series expansion when the exponent in 3.2 is positive:

$$h_\gamma^{\mathrm{PSE}-n}(\mathbf{r}) = \begin{cases} \exp\{t_\gamma(\mathbf{r})\} - 1 & t_\gamma(\mathbf{r}) < 0 \\ \displaystyle\sum_{i=1}^{n} \frac{t_\gamma(\mathbf{r})^i}{i!} & t_\gamma(\mathbf{r}) \geq 0 \end{cases} \tag{3.3}$$

$$t_\gamma(\mathbf{r}) = -\beta u_\gamma(\mathbf{r}) + h_\gamma^{OZ}(\mathbf{r}) - c_\gamma(\mathbf{r}).$$

where HNC is the limiting case as $n \to \infty$. As for HNC, the PSE-n family of closures have an exact, closed form expression for the chemical potential. The form of this approximation has a major impact on the convergence of calculations as well as on resulting thermodynamic quantities and correlation functions.

The goal of the self-consistent 3D-RISM procedure can be viewed as finding a direct correlation function $c_\gamma(\mathbf{r})$ such that h_γ^{OZ} and $h_\gamma^{\mathrm{closure}}$ become identical at all grid points to within some (fairly tight) tolerance. In existing, non-periodic, implementations, the convolution required in Eq. 3.1 is carried out via fast Fourier transforms in a rectangular box surrounding the solute, and additional terms that account for solvent outside of the artificial box are added to this. The periodic codes described here are simpler because there is no "external" region to account for. Key differences are that the electrostatic

and Lennard-Jones potentials that appear in Eq. 3.3 need to take periodic boundary conditions into account, and that some special considerations are needed, when the solute has a net charge, to ensure charge neutrality for each unit cell. These are taken up in the next two sections.

3.3.1 Computing the periodic solute potential.

The closure functional equation requires the mapping of the solute potential onto regular grids that cover the entire unit cell; there is one grid for each type of solvent site. The workflow closely follows what is done in molecular dynamics simulations that use the Particle Mesh Ewald (PME) procedure.[33, 38] Lennard-Jones interactions between solute atoms and all solvent types are calculated at each grid point using a distance cutoff (default is 9 Å) and the minimum-image convention. The same procedure is used for the short-range part of the electrostatic energy, where the bare Coulomb interaction is replaced by $\mathrm{erfc}\left(\beta\left|\mathbf{r}-\mathbf{r}_i\right|\right)/\left|\mathbf{r}-\mathbf{r}_i\right|$. Here $\mathbf{r}$ is the position of a solute atom, and $\mathbf{r}_i$ a point on the grid. The remaining, "long-range" part of the Coulomb interaction is handled via a fast Fourier transform procedure, in direct analogy to what is done for molecular dynamics calculations:[33, 38]

1. Interpolate the solute atomic charges to the Cartesian grid. We choose the smooth PME (SPME) approach, which uses a cardinal b-spline to interpolate the source charge to the grid. The b-spline interpolation has a roughly Gaussian character at high polynomial orders, and has the desirable trait that integration of its weights over the region of interpolation equals unity.

2. Convert the source charge grid from real space to frequency space using a FFT.

3. Convolute the source charge grid with the electrostatic interaction Green function. In frequency space the convolution is a simple multiplication, and the electrostatic interaction potential is k^{-2} .

4. Compute the electrostatic potential on the grid by converting the convoluted kernel from frequency space to real space using an inverse FFT.

At this point, the solute potential is ready, and the next step is to handle the solvent.

3.3.2 Solving the 3D-RISM equations.

As noted above, "solving" the 3D-RISM equations amounts to finding a c_γ function (for each solute site γ) that minimizes $\triangle c_\gamma \equiv h^{closure} - h^{OZ}$ at all points on the grid. Calculations are initialized with a guess for each c_γ, which is typically chosen to be uniformly zero, although a user-provided starting point can accelerate convergence. Each self-consistent procedure cycle begins with computing h_γ^{OZ} in the reciprocal space, followed by a switch to the real space where $h_\gamma^{closure}$ is computed, and ending by modifying the current guess for c_γ using an MDIIS procedure based upon $\triangle c_\gamma$. This cycle is repeated until $\triangle c_\gamma$ reaches a pre-determined threshold, which is typically 10^{-10} if gradients are to be used (for minimization or dynamics), and $10^{-?}$ if one just needs thermodynamic parameters or solvent distribution functions.

This procedure is complicated when charged solutes are used: here one wants the solute net charge to be neutralized by the converged ion distribution of the solvent. However, as long as the reference solvent is neutral, the h^{OZ} distribution arising from Eq. 3.1 will also be neutral. This is, of course, a problem when the solute charge is non-zero. We have considered two ways to address this problem: the first, adopted by Kovalenko and Hirata for non-periodic 3D-RISM[61], modifies the OZ total correlation function to ensure that the solvent charge exactly balances the solute charge. The second model removes the restraint that the bulk solvent be neutral.

3.3.2.1 (a) The Ornstein-Zernike equation generates an electro-neutral solvent

The total solvent charge can be obtained from the integral of the charge distribution of all solution species ($Q_\lambda \rho_\lambda g_\lambda(\mathbf{r})$) over the volume of the system, in this case the volume of the unit cell. The solvent species distributions are generated using the OZ equation:

$$Q_{solvent}^{OZ} = \int_{V_{cell}} \sum_\gamma Q_\gamma \rho_\gamma \left[h_\gamma^{OZ} + 1\right] d\mathbf{r}$$

If we restrict to the situations where the solution composition is chosen to satisfy the neutrality condition, ie $\sum_\gamma Q_\gamma \rho_\gamma = 0$ and replace h_γ^{OZ} using Eq 3.1, we get:

$$Q_{solvent}^{OZ} = \int_{V_{cell}} \sum_\gamma Q_\gamma \rho_\gamma \sum_\alpha c_\alpha * \chi_{\alpha\gamma} d\mathbf{r}$$

which can be rearranged using linearity and distributivity properties of convolution to read:

$$Q_{solvent}^{OZ} = \int_{V_{cell}} \sum_\alpha c_\alpha * \sum_\gamma Q_\gamma \rho_\gamma \chi_{\alpha\gamma} d\mathbf{r}$$

and can be further re-written using the integration property of convolutions as:

$$Q_{solvent}^{OZ} = \sum_\alpha \left(\int_{V_{cell}} c_\alpha d\mathbf{r} \right) \left(\int_{V_{cell}} \sum_\gamma Q_\gamma \rho_\gamma \chi_{\alpha\gamma} d\mathbf{r} \right)$$

Remembering the definition of solvent-solvent susceptibilities:

$$\begin{aligned} \chi_{\alpha\gamma} &= \omega_{\alpha\gamma} + \rho_\alpha h_{\alpha\gamma} \\ \omega_{\alpha\gamma} &= \delta\left(r - l_{\alpha\gamma}\right) \end{aligned} \tag{3.4}$$

and consequently that $\int \sum_\gamma Q_\gamma \rho_\gamma \chi_{\alpha\gamma} d\mathbf{r} = 0$, ie the sum over the excess number of particles (γ) with respect to a specific species (α) equals to the total charge of the bulk solvent model (which is chosen here to be zero, electro-neutrality) which nullifies each term of the sum over α irrespective of the values of the direct correlation function, c_α. So, the OZ equation when used in conjunction with susceptibilities derived for a electro-neutral solvent, will always lead to an electro-neutral three dimensional solvent distribution.

3.3.2.2 (b) Extending the RISM equations to achieve charge neutral periodic systems

RISM and, in general, molecular solvation theories, due to their formulation in the grand canonical ensemble, should be able to build in the necessary excess of ionic charges to a charged solute such that the final system is electro-neutral. As shown above, irrespective of the values of c_γ the OZ equation insures that the solvent and not the system will be neutralized. To address this problem, we follow the basic idea used by Kovalenko and Hirata for the non-periodic problem,[59, 60] and modify the total correlation functions to impose system neutrality, ie:

$$\sum_\gamma Q_\gamma \rho_\gamma \int h_\gamma^{OZ,corr} d\mathbf{r} + Q_{solute} = 0 \tag{3.5}$$

where Q_γ and ρ_γ are the charge and reference concentrations of the solvent sites, $h^{OZ,corr}$ is a corrected form of the total correlation functional, which we propose to take the form:

$$h_\gamma^{OZ,corr}(\mathbf{r}) = h_\gamma^{OZ}(\mathbf{r}) - Q_\gamma \phi \tag{3.6}$$

that adds an additional term dependent on the charge of the solution site. Hence

$$\sum_\gamma Q_\gamma \rho_\gamma \int \left[h_\gamma^{OZ}(\mathbf{r}) - Q_\gamma \phi\right] d\mathbf{r} + Q_{solute} = 0$$

Remembering that $\sum_\gamma Q_\gamma \rho_\gamma \int h_\gamma^{OZ}(\mathbf{r})\, dr = 0$, we can solve for ϕ:

$$\phi = \frac{Q_{solute}}{V_{cell} \sum_\gamma Q_\gamma^2 \rho_\gamma} \tag{3.7}$$

which, as such, becomes constant and does not need to be updated at each iteration. The correction to h_γ^{OZ} applies only to charged species of the solvent (ions) and not on neutral components (such as water).

Algorithm 1 shows how the RISM equations are solved to reach self-consistency, as described at the beginning of Subsection 3.3.2.

Algorithm 1 Periodic algorithm with h_{OZ} shifted; assumes HNC closure for simplicity.

$c_0 = 0$; $\delta c_0 = 999$.
while $\delta c_i > thresh$:
  $\hat{c}_{i-1} = FT\left[c_{i-1}\right]$
  $\hat{h}_{OZ,i} = \hat{c}_{i-1} * \hat{\chi}_{vv}$
  $h_{OZ,i} = FT^{-1}\left[\hat{h}_{OZ,i}\right]$
  $h_{OZ,i} = h_{OZ,i} - Q_\gamma \phi$
  $h_{HNC,i} = Exp\left[-U_{PME} + h_{OZ,i} - c_{i-1}\right] - 1$
  $\delta c_i = h_{HNC,i} - h_{OZ,i}$
  $c_i = MDIIS[c_i, \delta c_i]$
  $i++$

3.3.2.3 (b) Employing non-neutral bulk solvent models.

A second approach to system neutralization abandons the constraint that the solvent be net neutral. This approach won't work for non-periodic systems, since the (physical) bulk solvent conditions are satisfied (by construction) at a large distance from the solute. But for periodic systems, there is no point in space that is not close to a solute molecule, and hence no clear reason to require that the reference solvent be neutral. We encounter the same sort of decision here with modifying h^{OZ} in the integral equation approach discussed above: for (say) a negative solute, should one increase the concentration of cations, or decrease the concentration of anions, or take some combination of these two? To begin exploration, we looked at the sarcin-ricin RNA, which has a net solute charge of -104. Calculations using the modified OZ correlation (described above) suggested that very few anions would be expected in the water channels here, so we decided to explore reference solvents that just had water and varying concentrations of Na^+ ions. An ion concentration of 2.5M (using the pse2 closure) resulted in having 104.02 Na+ ions per unit cell, which compares well with the ion counts of 107.41 Na+ and 3.41 Cl- when using a (neutral) 1.0 M NaCl reference solvent and the OZ modifications.

3.3.3 Computing forces on the solute atoms

Under certain circumstances (apparently, that there is a closed form for the chemical potential, independent of the path of thermodynamic integration), KH argue that the

gradient of the excess chemical potential can be written in the following form:

$$\mathbf{f}\left(\mathbf{R}_i\right) \equiv -\nabla_{R_i} \triangle\mu = \sum_{\gamma} \rho_{\gamma} \int d\mathbf{r} g_{\gamma}\left(\mathbf{r}\right) \nabla_i U_{\gamma}\left(\mathbf{r}-\mathbf{R}_i\right) \tag{3.8}$$

where the electrostatic part of U_γ is $U_\gamma^{el} = Q_\gamma \phi_{solute}^{el}$, which is the electrostatic potential of the periodic solute that can be computed using lattice sums, such as Ewald sum or PME. ϕ_{solute}^{el} can thus be separated in a real and reciprocal space terms term.

$$\begin{aligned} \mathbf{f}^{el}\left(\mathbf{R}_i\right) &= \int \left(\sum_{\gamma} Q_\gamma \rho_\gamma g_\gamma\left(\mathbf{r}\right)\right) \left[\nabla_i \phi_{solute}^{el,short} + \nabla_i \phi_{solute}^{el,long}\right] d\mathbf{r} \\ &= \int \rho\left(\mathbf{r}\right) \left[\nabla_i \phi_{solute}^{el,short} + \nabla_j \phi_{solute}^{el,long}\right] d\mathbf{r} \end{aligned} \tag{3.9}$$

The real (short range) term is in part a convolution ($\rho * \phi_{solute}^{el,short}$) with a short-range kernel

$$\phi_{solute}^{el,short}\left(\mathbf{r}\right) = \sum_j \frac{Q_i \mathrm{erfc}\left(\beta\left|\mathbf{r}-\mathbf{R}_i\right|\right)}{\left|\mathbf{r}-\mathbf{R}_i\right|} + const.$$

that can be carried out in the real space: where the constant terms have been ignored, and with which the short range electrostatic field component can be computed by taking the derivative of the above:

$$\nabla_i \phi_{solute}^{el,short} = Q_i \left(-\frac{2\beta}{\sqrt{\pi}} \exp\left(-\beta^2\left|\mathbf{r}-\mathbf{R}_i\right|^2\right) + \frac{\mathrm{erfc}\left(\beta\left|\mathbf{r}-\mathbf{R}_i\right|\right)}{\left|\mathbf{r}-\mathbf{R}_i\right|}\right) \frac{\mathbf{r}-\mathbf{R}_i}{\left|\mathbf{r}-\mathbf{R}_i\right|^2}$$

The contributions of this term are evaluated using a minimum image convention and using a cutoff and can be computed simultaneously with the LJ part. The reciprocal (long range) term can be obtained considering first a simple case where the charge density interacts with a single Gaussian positioned at $\mathbf{R}_i$. The interaction energy is: $E_i^{el} = Q_i \int \rho\left(\mathbf{r}\right) \phi_{Gaussian}\left(\mathbf{r}-\mathbf{R}_i\right) d\mathbf{r}$, which can be recast as a convolution:

$$E_i^{el} = Q_i \rho * \phi_{Gaussian}\left(\mathbf{R}_i\right) \tag{3.10}$$

where the analytical form of $\phi_{Gaussian}(\mathbf{r}) = \mathrm{erf}(\beta|\mathbf{r}|)/|\mathbf{r}|$; in practice it is represented on the reciprocal grid and recycled from the initiation step where the solute potential is mapped on the regular grid. Therefore, the electrostatic force acting on the atom i could be written as:

$$\mathbf{f}_i^{el} = Q_i \nabla_i \rho * \phi_{Gaussian} \tag{3.11}$$

showing that one has to carry out a single FFT based convolution between the solvent charge density and the electrostatic potential of a Gaussian distribution centered at the origin, followed by an estimation of the force at the atomic position using a spline derivative, numerical differentiation in the real space.

We are examining whether this equation still holds if option (a) above is chosen, where h^{OZ} is modified to force neutrality.

3.4 Results

3.4.1 Solvent distribution in molecular crystals

3.4.1.1 Comparison with X-ray scattering factors

How we can compute scattering from the solvent distributions coming from 3D-RISM, and compare to experiment, and cross-compare to simple, flat bulk solvent models is discussed in Section 3.6. Results below show that refinement with RISM density and the solvent distributions from MD show improvement over the standard flat model for solvent. The MD distributions result in the best r-factors, but r-factors for refinements using RISM densities become closer to the MD results as resolution worsens. See Fig. 3.1 and Tables 3.1 to 3.2.

3.4.1.2 Total number of waters

It would be of considerable utility to know the total number of waters per unit cell.[50, 64]. This parameter is also useful in testing how well 3D-RISM is working. Table 3.3 gives the number of solvent water molecules per protein chain for several examples. In

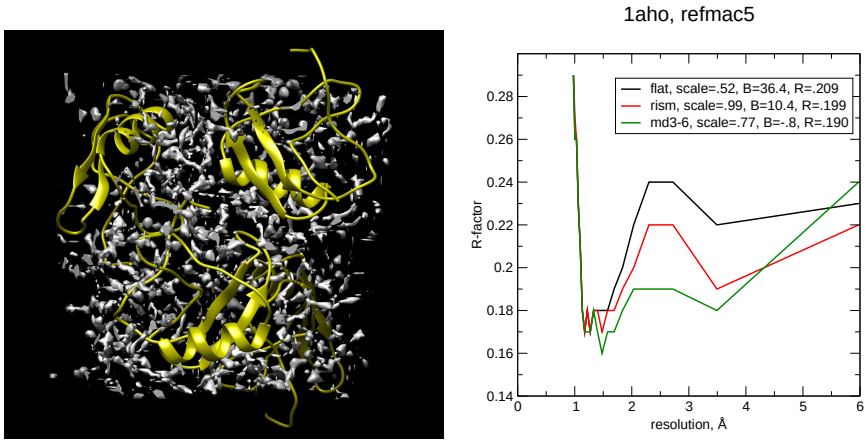

Figure 3.1: Water density in 1aho.

Protein	scorpion toxin	scorpion toxin[a]	GB3	myoglobin	lysozyme	lysozyme	cyclophilin
PDB ID/resol.	1aho/0.96	1aho/0.96	2igd/1.10	1bzr/1.15	4lzt/0.95	2lzt/1.97	4yul/1.42
flat (Refmac)	.209/.214	.178/.190	.220/.233	.200/.208	.196/.205	.167/.216	.201/.224
3D-RISM	.197/.211	.158/.174	.213/.224	.194/.206	.190/.197	.154/.201	.185/.202
explicit MD	.189/.198	.144/.167	.191/.209	.186/.192	.191/.202	.153/.214	.172/.185

Table 3.1: Bulk solvent models with a single protein configuration; each block shows R/Rfree after 40 cycles of *refmac5* refinement. There is an average drop in R of 0.019 between flat and explicit MD, and an average drop of 0.011 between flat and 3D-RISM. (a) results for 1aho using alternate conformers.

RNA	2a43	480d	2qus
flat (Refmac)	.223/.261	.192/.216	.206/.255
3D-RISM	.208/.229	.175/.208	.186/.234

Table 3.2: Bulk solvent models with a single RNA configuration; each block shows R/Rfree after 40 cycles of *refmac5* refinement. There is an average drop in Rfree of 0.017 between flat and 3D-RISM, slightly larger than the value of 0.011 found for proteins in Table 3.1.

protein	PDB ID	solvent %[a]	MD[b]	cSPCE_kh[c]	cSPCE_pse3[d]	NaCl_pse3[e]	f000[f]	find_F000
lysozyme[g]	4lzt	34	284	271	288	284	312	312±8
lysozyme[h]	5l9j	44	419	408	422	412	453	558±19 (!)
cyclophilin-A	4yul	56	971		963	960	995	
thaumatin	4el7	59	1274			1256	1334	
scorpion toxin	1aho	41	195	180		189	202	

Table 3.3: Predicted number of solvent molecules per protein chain. (a) solvent fraction reported by *phenix.f000*; (b) from MD simulation with SPCE water, see text; (c) 3D-RISM result using the KH closure; (d) 3D-RISM result using the PSE3 closure; (e) as in (d), but using 0.1 M NaCl solvent, rather than pure water; (f) prediction from *phenix.f000* assuming the default solvent electron density of 0.35 $e/\text{Å}^3$; (g) triclinic; (h) tetragonal

all protein structures, the RISM water counts are all very similar to the crystal MD counts, indicating the high accuracy of the periodic code. In fact, at least one RISM calculation for each structure is the closest prediction to the MD results, performing better than the built-in method from PHENIX.

3.4.2 Using 3D-RISM as an implicit solvent model for biomolecular crystals

The next figure compares the average structures from (Fig. 3.2) 100 ps of simulation of the sarcin-ricin RNA unit cell, using either the 3D-RISM result (left) or no solvent term (right), to the experimentally-determined and deposited structure. As can be seen, there is less deviation from the deposited structure in the simulation with RISM forces than in the simulation without solvent. This indicates that the macromolecular interactions with solvent stabilize the structure, and that simulations with RISM forces do a better job of replicating experimental conditions than simulations with no solvent description. More work should be done, both with different structures, as well as to compare these results with those found from crystal MD simulations using explicit solvent molecules. Within AMBER, this RISM implementation is the only implicit model of solvent with the ability to replicate periodic systems like macromolecular crystals.

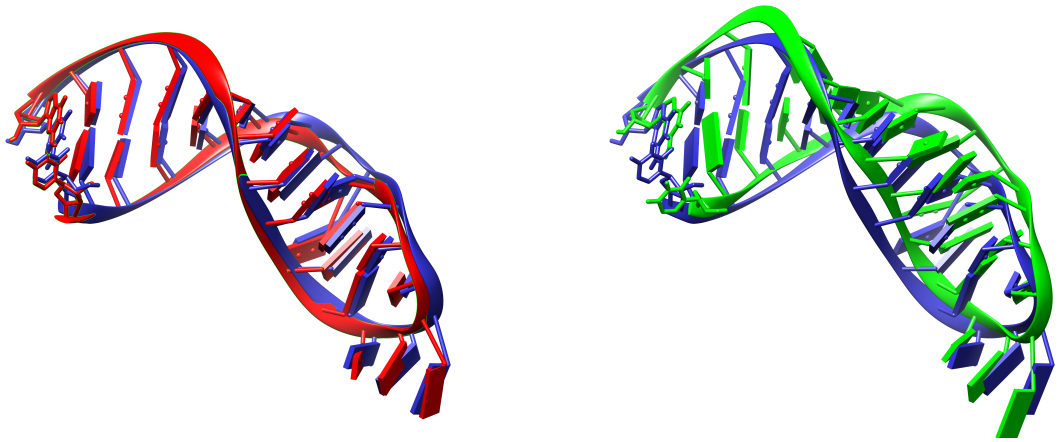

Figure 3.2: Blue: experimental average structure from X-ray crystallography (PDB ID 480d); red: average structure from a 3D-RISM crystal simulation; green: average structure from a crystal simulation with no solvent correction.

3.4.3 Thermodynamics in the infinite dilution regime

The examples discussed above dealt with molecular crystals, where solute molecules are in contact with their images in neighboring unit cells, and the solvent volume is fairly small. Another application might be to a single (dilute) solute surrounded by a buffer of solvent. As the size of the unit cell increases, such a calculation should approach the infinite dilution, non-periodic limit that has traditionally been assumed in 3D-RISM applications. As noted above, these traditional calculations actually employ a regular periodic grid in the vicinity of the solute (to enable convolutions to be carried out via fast Fourier Transforms), and add in estimates of the "asymptotic" contributions from solvent outside the grid. Here we study the box-size dependence of periodic 3D-RISM calculations that have a single solute molecule at the origin. We show that extrapolation of thermodynamic quantities to an infinite box size can be readily carried out, yielding results that are in close agreement with those from existing non-periodic codes. See Fig. 3.3

The slope of this line at large box sizes is just $q^2\zeta/2L$. This implies that a correction to infinite box size can be made from a single periodic calculation, provided that it is large enough to be in the linear regime. Hence there is a possibility of using periodic calculations to replace non-periodic ones, with particular advantages for gradients, but

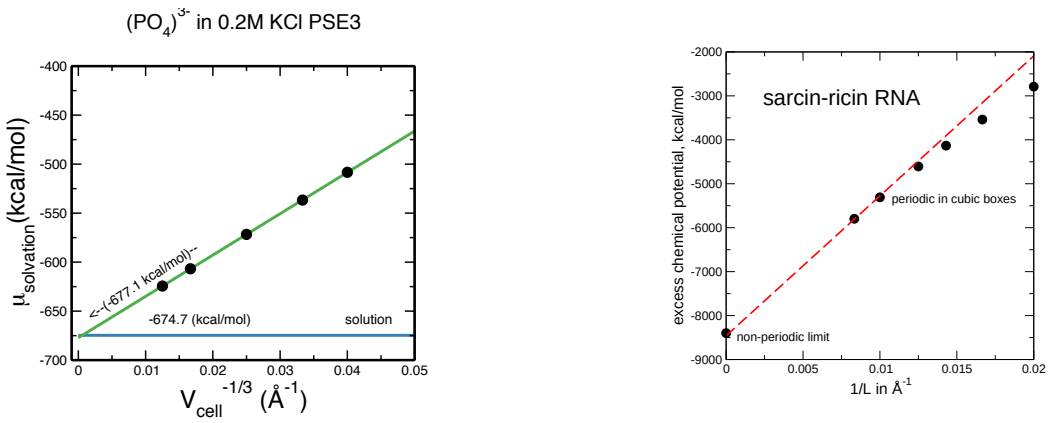

Figure 3.3: Variation of solute chemical potential with respect to periodic cell size (black dots and green linear fit) and comparison with solution case (blue line).

also potential timing advantages for single-point calculations. Clearly, the size of the unit cell in the periodic calculations needs to be large enough to capture the energetic consequences of perturbations in the solvent, and this minimal "buffer size" likely depends on the nature of both the solute and solvent. It seems likely that the required buffer size for non-periodic calculations also has to be large enough to describe these solvent perturbations, although it is possible that the (approximate) asymptotic contributions are accurate enough to allow somewhat smaller explicit grids. This will be a subject of future study.

3.5 Conclusions

Water molecules and ions around biomolecules often play a crucial role in function. Analysis of the solvent distributions in biomolecular crystals can provide an important check on the accuracy of computational models. Here we have presented sample results for a variety of small proteins and RNAs, derived from explicit solvent molecular dynamics calculations, and from a newly-developed periodic integral equation (3D-RISM) code. The predicted solvent distributions can be compared to experiment in a variety of ways: by looking at the locations of ordered waters and ions that can be identified in density maps derived from X-ray crystallography; by comparing computed and observed Bragg intensities; and (potentially) by comparing predicted and measured crystal

densities (which reflect the total number of water and ions per unit cell).

3.6 Computational details

The six macromolecule crystal structures being refined have Protein Data Bank (PDB) IDs 1aho (scorpion toxin protein), 1bzr (whale myoglobin), 2igd (protein G IGG-binding domain II), 2lzt (lysozyme), 4lzt (hen egg white lysozyme), and 4yul / 3k0n (Cyclophilin A - CypA).

Refinement calculations were performed using Refmac5. The refinement procedure requires two input structures: the solvent density distribution (as calculated by the solvent model) and the energy minimized solute structure. During refinement, the solvent density is held constant (except for overall scaling and overall B-factors, which are refined), and the atomic positions and B-factors of the solute are modified to achieve best agreement with the observed diffraction intensities. The final R-factor is obtained after 40 refinement cycles.

All 3D-RISM solvent model calculations were performed with the Kovalenko-Hirata (KH) closure on a uniform 0.35 Å spaced grid. A 10^{-6} correlation function convergence error tolerance was enforced at each MD time step. The periodic 3D-RISM implementation directly produced solvent distributions which were used in Refmac refinement calculations.

The periodic 3D-RISM implementation used here is part of AmberTools (versions 19 and above), an open source collection of molecular simulation software. The implementation was based upon an existing non-periodic RISM code that was primarily developed by Tyler Luchko, David Case, and Andriy Kovalenko [70]. Extensions to periodic systems were spearheaded by Jesse Johnson and George Giambaşu, and a more complete description of the codes is given elsewhere.[55]

Chapter 4

Studies of RNA structures using 3D-RISM and explicit solvent crystal MD

4.1 Introduction

As mentioned in Chapter 2, more and more RNA molecules have been solved in recent years, and as such, our understanding of the roles RNA has played and does play in the world has expanded[27, 8, 97, 20, 18], and ideas for leveraging RNA therapeutically have gained traction[16, 25, 69, 94, 29, 9]. As such, structural accuracy is of the utmost importance. Along with structural accuracy, however, proper understanding of solvent interactions with RNA and distribution of water molecules and ions is also paramount.

Many interactions at active sites and in ligand binding are mediated by water molecules and ions[47, 93], and tertiary structure of RNA molecules can be supported by electrostatic interactions with nearby water molecules and ions[110, 76]. The water molecules and ions involved in these kind of interactions are generally well placed in models from X-ray crystallography[6, 109, 12, 13, 5]. However, as solvent can make up from 30-70% of the volume in biomolecular crystals[73, 4, 28], there is a great deal of solvent that is not represented on a molecular basis in crystallographic models. In fact, crystallographic refinement software generally uses a flat solvent distribution to model this disordered solvent. This flat model is modified via its parameters (B-factors and density) to improve the model's fit to experimental data on a global scale, but does not properly describe the vast majority of the solvent in the crystal[78]. Methods other than X-ray crystallography, such as atomic emission spectroscopy[7, 41], small-angle X-ray scattering[85, 86, 75, 82], and partial molar volume measurement[22, 24, 23, 106], can provide ion and water molecule numbers for the bulk solvent in solution, but provide

little to no information regarding the solvent molecules near the biomolecules. Computational methods may be the best way to determine solvent information both at the local interaction level and the bulk solvent level. Both 3D-RISM calculations and crystal MD simulations are methods suitable to providing bulk solvent characterization, one implicitly and the other explicitly, and to placement and description of nearby solvent molecules and their energetic interactions with the biomolecules. Crystal MD simulations, however, have been an established method for longer than 3D-RISM.

Crystal MD simulations, which have been around since at least the 1980s [108], mimic crystalline environments with explicit solvent molecules describing the entire unit cell's solvent, and could ideally provide a fuller description of solvent left out of crystallographic models. The insistence on keeping unit cell dimensions for the size of the box held under periodic boundary conditions in such a simulation leads to a more accurate description of crystal solvent and macromolecular dynamics than can be provided by typical solution simulations. Not only can these simulations provide a more diverse description of bulk crystal solvent than standard flat densities, but they can also help provide structural dynamical information to describe diffuse scattering observed in crystallographic experiments. This method is still developing and improving as technology improves[21] (for further information, see this source).

3D-RISM, on the other hand, is a much more recent development as a method to describe and predict how ions and water molecules solvate macromolecules. The distributions of the different solvent species are found as atomic spatial distributions in three dimensions from an approximate solution to an equation similar to the Ornstein-Zernike equation. The equation used relates the total correlation and direct correlation functions to each other, but in order to get the solution, a closure equation that relates the two must be used. This method has the benefit over other non-explicit models of being developed with explicit water as its starting point and having statistical mechanics as its basis (see Chapter 3).

3D-RISM, as with all non-explicit models, has the advantage of being far more time efficient than explicit water simulations to arrive at a description of the solvent. However, as the solutions sought are approximations, it might be questionable whether,

especially with a new implementation, the description arrived at is accurate. In this paper, both 3D-RISM and crystal MD simulations will be used to probe the solvent in 3 different RNA structures.

4.2 Methods

4.2.1 Structure selection and preparation

Three structures with a good range of resolution (1.34-2.6 Å), while still relatively high resolution for RNA structures, and fairly simple structures (no complex modifications) were chosen (PDB IDs 2a43[87], 2oiu, 3tzr). To prepare the structures for crystal MD simulations, the crystallographic water molecules were removed from the structures, and *prmtop* and *rst7* files were prepared for the unit cell of each structure. This was done using *phenix.AmberPrep*, and then the output *prmtop* and *rst7* files were combined using *ambpdb* in AMBER to build unit cell PDB files. These were then run through a script called addtobox.sh, which used the *AddToBox* utility to add enough sodium ions to neutralize the structures and a test number of water molecules to provide a stable pressure in the crystal during simulation (the test starting numbers were chosen after 3D-RISM calculations of the structures). The structure preparation was more of a cycle, as this addition of water molecules had to be done multiple times to get the proper pressures throughout the simulation process.

In order to be prepared for RISM calculations, the structures had all deposited waters and monatomic ions removed. These structures, with their existing ligands and crystallization solvent molecules (as listed in Table 4.1), were then run through *phenix.AmberPrep* without minimization in order to prepare topology and coordinate files for the unit cell to be handled as input for the calculations.

PDB ID	Resolution (Å)	Non-RNA Entities
2a43	1.34	None
3tzr	2.21	SO4, SS0
2oiu	2.60	None

Table 4.1: The three structures used as input for RISM calculations, with their respective small molecules and ions. Only 3tzr has any non-RNA entities.

4.2.2 Crystal simulation parameters and process

For crystal MD simulations of these structures, at each number of water molecules tested out, a 500-step minimization using the steepest descent method was performed with constant volume periodic boundary conditions, followed by 20 ns of restrained equilibration (*restraint_ wt*=10.0) up to the experimental temperature at constant pressure periodic boundary conditions with an artificially high constant pressure, and then 20 ns each of restrained simulation at restraint weights of 10.0, 1.0, and 0.1, followed by at least 60 ns of unrestrained simulation in order for the system to converge. The restraints were set on all heavy atoms in the systems, other than those in the water molecules and the sodium ions. The unrestrained simulations were only performed once the proper number of water molecules was settled upon. This was done by checking the average pressures at the end of each of the restrained runs: water numbers were changed if the absolute value of the average pressure was higher than 500, 200, and 100 for the 10.0, 1.0, and 0.1 restraint weight runs, respectively. All of these calculations, performed in *pmemd* using the *cuda_ SPFP* implementation in AMBER18 on a GPU (except for the minimization, which used *cuda_ DPFP*), used SHAKE on the hydrogen atoms in the systems.

4.2.3 rism3d.snglpnt parameters

For *rism3d.snglpnt* calculations of these structures, differing sets of parameters were used, both in trying to converge the calculations and in order to test differing sets of ions and concentrations. All calculations were performed with grid spacing of 0.35 Å in each dimension, with a cutoff of 20 Å and solvent cutoff of 9 Å, with up to 10 previous iterations used for predicting the next one (*mdiis_ nvec*=10) and a maximum of 10,000 steps for convergence at each closure. The calculations were all performed with 2 different closures, starting with Kovalenko-Hirata and ending with PSE2 with tolerances of 0.1 and 0.000001, respectively. Concentrations of salts were chosen arbitrarily in some cases, and at experimental levels in others.

4.2.4 Analysis

Multiple routes of analysis were charted, both numerical and visual in nature. For the crystal MD simulations, the first line of analysis was already performed, checking on average pressures to make sure the boxes were stable in the simulations. Once the proper numbers of waters were settled upon for the simulations (and marked down), to monitor the convergence of the unrestrained simulations, the RMSD for the entire crystal lattice of the RNA base heavy atoms, the asymmetric unit RNA base heavy atoms, and all heavy atoms of the RNA in an asymmetric unit were analyzed using the XtalAnalyze, GetBfactors, and XtalPlot scripts in *AmberTools/src/xtalutil/Analysis/*, with modification to allow for the differences in atom names from protein to RNA (base heavy atoms were used as the backbone, as the bases should stay pretty stable due to the base pairing interactions) and C1' atoms used in place of the CA atoms found in proteins. When the RMSD leveled out, the simulations were considered to be converged, and further analysis was performed.

At this point, the scripts used above were performed on the finished simulations. In all uses of these scripts, the trajectories, starting *rst7* files, and *prmtop* files were stripped of all atoms but the RNA and MG ions in *cpptraj*, and only the unrestrained trajectories were used for analysis. (Currently, what use the output plots for B-factors and RMSD will be put to is unseen.) Further, easier, analysis of the RMSD was performed using *cpptraj*, loading in the stripped *prmtop*, starting *rst7* (as the reference), and unrestrained trajectories and running the *rms* command, with fitting and mass-weighting, for all RNA heavy atoms and for just the base heavy atoms, in the entire unit cell and each individual asymmetric unit. Further RMSD analysis involved using the average structure asymmetric unit backbone (base) heavy atom and all RNA heavy atom RMSD from the XtalAnalyze script's use of reverse symmetry placement of all of the asymmetric units on top of each other and averaging of the structures and comparing to the starting asymmetric unit and placing these numbers in a table. Further analysis concerning water molecule counts was performed to provide predictions using crystal parameters in three different ways. First, the *phenix.f000* command was performed on

the 4phenix PDB files made for 3D-RISM use using a *mean_ solvent_ density* of 0 and again with it set to 0.35 (the default). The resultant F(0,0,0) values indicate an estimate of the number of electrons without and with solvent molecules, respectively. Finding the difference between these values, and dividing by 10 electrons per water molecule, provides the F(000) WP values in Table 4.2. Second, the SC WP values were found by using the solvent content percentage on the "Experimental" tab for each structure on the PDB[14, 11], multiplying that by the unit cell volume, and then dividing by 30 $Å^3$per water molecule. Finally, the PSC WP predicted value comes from a similar calculation to the SC WP, just using the fraction value given by *phenix.f000* as the solvent content percentage. Visual analysis of the structures from these simulations was performed by creating images in Chimera overlaying the average asymmetric unit structure on top of the starting asymmetric unit structure to visually inspect and depict the difference in the average output structure and the starting structure.

Analysis of the 3D-RISM calculations was mostly numerical, looking at the output for each calculation, and comparing numbers of waters and ions found with differing solvent setups. One other measure recorded was the excess chemical potential for RISM, which is equivalent to the ERISM value in the energy calculations for the PHENIX structures. Finally, the data for both crystal MD simulations and 3D-RISM calculations were compared to demonstrate the effectiveness of 3D-RISM in solvating the structures when compared to the more established explicit solvent method (crystal MD). The ion counts and water counts were compared.

4.3 Data and discussion

4.3.1 Crystal simulation results

As the standard to which the 3D-RISM results were to be compared, it was important to successfully perform and analyze the crystal MD simulations. This included looking at the ion and water molecule counts, as well as the RMSD of the RNA molecules in the simulations to the starting structures from before minimization. The following data will show that the simulations converged properly, as well as showing the standards for

numbers of water molecules for 3D-RISM results to be compared to.

PDB	Res(Å)	DMG	WAT	F(000)	SC	PSC	NA	P(bar)	UCV(Å^3)
2a43	1.34	12	3125	3125	2476	2976	126	-81.2	135062.4
3tzr	2.21	24	1549	1506	1028	1434	112	103.9	87240.9
2oiu	2.60	14	7968	8086	6745	7670	256	-34.0	315564.8

Table 4.2: Water and ion counts needed to neutralize and stabilize the unit cell for each structure, as well as the average pressure in the longest unrestrained run for each and predictions of water molecule numbers via different methods (Res=resolution, DMG=deposited magnesium ions in unit cell, WAT=water molecules in simulation, F(000)=predicted number of waters using *phenix.f000*, SC=predicted number of waters using the PDB solvent content percentage, PSC=predicted number of waters using PHENIX solvent content percentage from *phenix.f000*, NA=sodium ions in simulation, P=average pressure in longest unrestrained run, UCV=deposited unit cell volume). All numbers found in the simulations are generally higher than the predictions, and the F(000) predictions are closest to the simulation results.

As can be seen in Table 4.2, the number of water molecules in these unit cells were relatively proportional to the unit cell volume, and the structures were all neutralized with only positive ions added. The MD numbers were generally higher than the predictive values, which was somewhat surprising, although the F(000) predicted values are all rather close to the number of waters found in the MD simulations, with the largest difference being 118 waters in 2oiu, which is only 1.48% of the total number found in the MD simulation for this structure, while the difference for 3tzr is 2.78% of that structure's MD waters. The prediction for 2a43, at very high resolution, was exactly the same as the number found. All three structures' longest unrestrained runs had average pressures within about 100 bar of 0, meaning there was not much effort of the unit cell to try to expand or contract from its deposited cell dimensions.

As mentioned above, the RMSD of the structures were analyzed in multiple ways. First, the output from XtalAnalyze of the unrestrained simulations included base heavy atom and all RNA heavy atom RMSD for the average structure of all the asymmetric units as compared to the starting asymmetric unit. These values appear in Table 4.3, and show that there are similar changes from the starting structure to the average structure for 2a43 and 3tzr, while the average structure is vastly different from the starting structure for 2oiu. This result, while initially surprising, became less so when seeing

the RMSD found in previous simulations was around 4.2 Å for the docked conformation of the structure[43]. Another interesting point to consider with 2oiu is that the two monomers in the starting asymmetric unit are not identical: one is in a mimic of the active state (docked confirmation) of the ligase ribozyme, while the other is in an inactive state[102]. Further analysis of the monomers of this structure was performed due to this possible reason for the large RMSD.

PDB	Res(Å)	Sim Time(ns)	BHA(Å)	HA(Å)
2a43	1.34	160	1.0498	1.0333
3tzr	2.21	340	1.5022	1.4596
2oiu	2.60	60	4.1526	4.4531

Table 4.3: Unrestrained simulation times (Sim Time), average structure asymmetric unit base heavy atom RMSD (BHA), and average structure asymmetric unit heavy atom RMSD (HA). Typical RMSD values are found for 2a43 and 3tzr, while 2oiu has relatively high values that are in line with other published results.

When looking at the entire unrestrained simulation heavy atom RMSD for 2a43, the entire unit cell RMSD jumped from 2.0 to 2.5 Å in the first 60 ns and stayed relatively steady for the rest of the run. Generally speaking, the asymmetric units followed the same trend, but at lower levels (started between 1.0 and 1.7 Å, ended between 1.3 and 2.0 Å). The only real exception to that was the 6th asymmetric unit, which grew up to about 2.75 Å by the 60-ns mark. After this, it leveled off back around the rest of the asymmetric units. When looking at the simulation around this time, the 6th asymmetric unit started to become very unstable regarding the sugar-phosphate backbone, and the 3'-end of the structure spread apart from the mid-chain loop that base paired to it. There's also a large fluctuation in residue 12 (labeled this way in the deposited structure, in analysis it becomes residue 10 as the deposited structure starts with 3), the base of which sticks out of the base pairing groove and rotated freely during this portion of the simulation. This asymmetric unit became more stable as the simulation continued, and, thus, the RMSD stabilized. When looking at this structure visually in Figure 4.1, the differences were across the board, especially with all of the non-base-paired bases, and even one of the two magnesium ions changed locations relative to the RNA molecules in a rather large way (even though the magnesium ions were not included in the atom

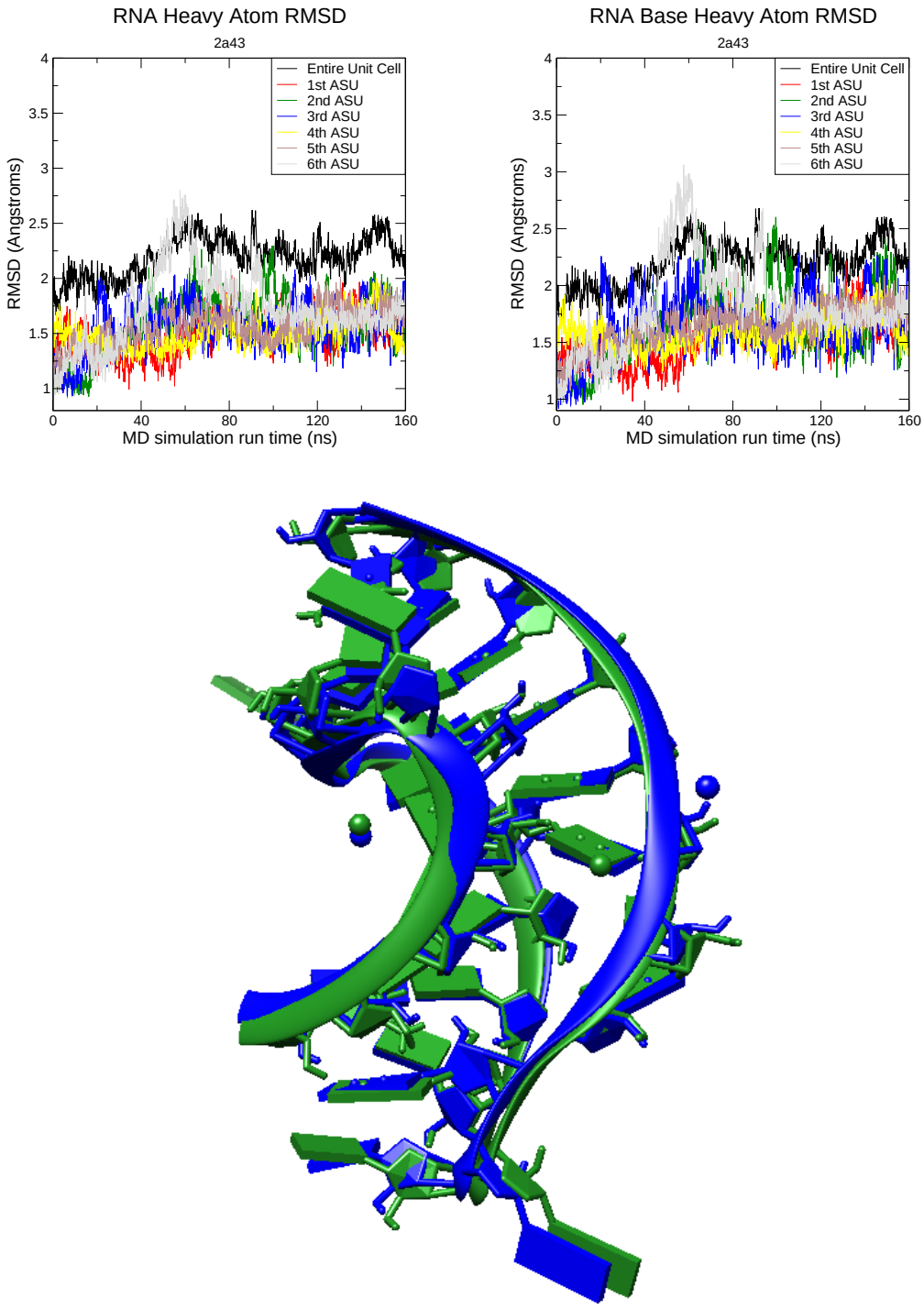

Figure 4.1: Heavy-atom and base heavy-atom RMSD analysis, as well as visual analysis of 2a43 crystal simulation. The structural image is an overlay of the starting asymmetric unit PDB file (green) and the average coordinate PDB file from the simulation (blue). The RMSD between these structures is in Table 4.3. The RMSD throughout the simulation stays relatively low, other than the sixth asymmetric unit, which has a large spike in the middle of the simulation and restabilizes. There is some backbone deviation in the average asymmetric unit structure, but a lot of the differences appear to take place in the bases.

mask for RMSD).

When looking at the data from the 2oiu simulation in Figure 4.2, one might expect somewhat higher RMSD values due to this structure being at the lowest resolution of the 3 structures studied in this chapter. However, the RMSD was far and away much worse than expected purely when compared to the other structures. The unrestrained simulation heavy atom and base heavy atom RMSD were relatively the same, ranging from 3.25 to 4.75 Å for the unit cell, with both of the asymmetric units fluctuating between 2.5 and 4.5 Å. It was somewhat surprising that with such a high RMSD, the simulation was stable enough to be considered converged even in just the first 60 ns of prolonged unrestrained simulation.

As mentioned in analysis of Table 4.3, the average structure RMSD was not far from the docked state RMSD found in simulations, and there are two different states in the two monomers in each asymmetric unit. To better look into how these two different states affect the overall RMSD, each monomer's RMSD was analyzed in a similar way to the asymmetric unit RMSD. In Figure 4.3, the unit cell RMSD graphs are the same as in Figure 4.2, but the other curves are the RMSD for each monomer throughout the simulation. Generally speaking, the 1st monomer in each asymmetric unit is lower in RMSD than the 2nd monomer, around 2 to 3 Å, while the 2nd monomers were much closer to the overall unit cell RMSD. While it was not surprising that their were differences in the RMSD from monomer to monomer when the monomers were in different conformations, it was surprising to compare which monomers were higher in RMSD. When looking at the literature [102, 43], one finds that the 1st monomer of each asymmetric unit is the undocked conformation, and the 2nd asymmetric unit is the docked conformation, and the RMSD in simulations in [43] is higher for the undocked conformation, which is not the case here. The structural image in Figure 4.2, an overlay of the average asymmetric unit structure on the starting asymmetric unit structure, further shows the difference in the monomers over the course of the simulation. The monomer in the bottom portion of the image is the 1st monomer in the asymmetric unit, and this has a much better fit, minus one or two free bases floating at different angles outside of the groove, while the top monomer is the 2nd, showing

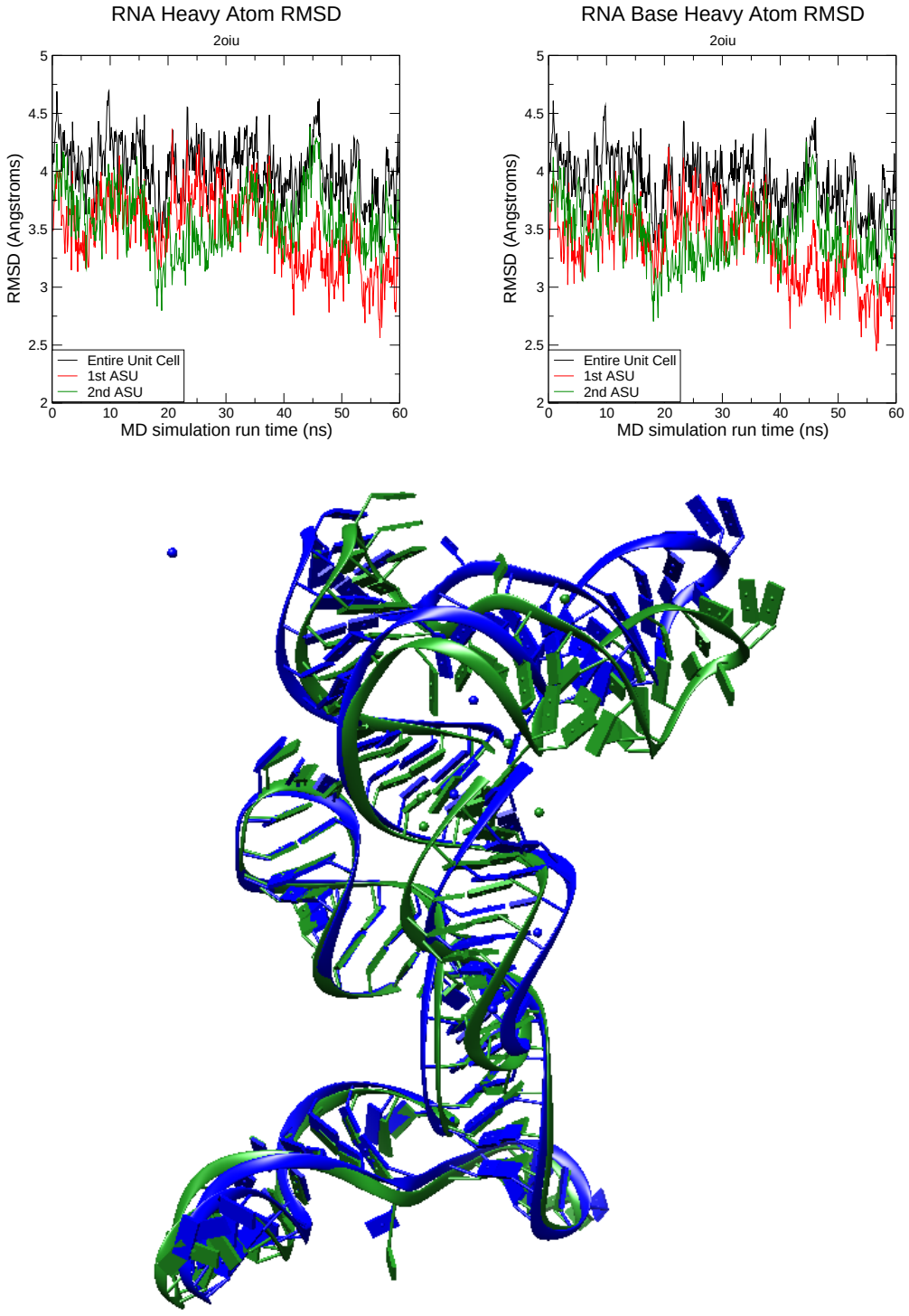

Figure 4.2: RMSD and visual analysis of 2oiu crystal simulation. The RMSD for this structure is rather high, but matches what is found in the literature for this structure. Visual analysis of the starting structure (green) and the average asymmetric unit structure (blue) from the simulation shows both massive deviation in the backbone and bases, especially in the upper monomer as depicted.

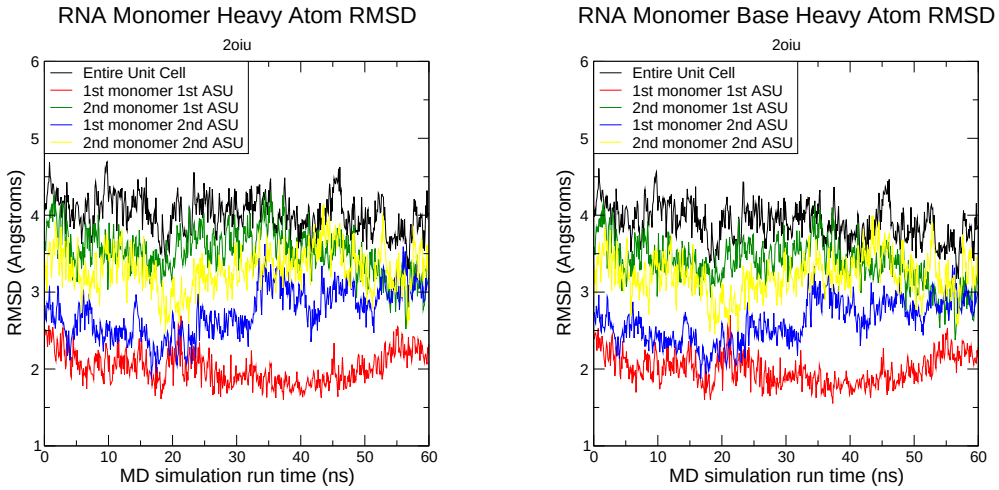

Figure 4.3: RMSD analysis of 2oiu on a per-monomer basis. The first monomer in each asymmetric unit shows much less deviation than the second monomer for most of the simulation, which is a surprise as compared to the literature on this structure.

the large RMSD difference from the starting structure seen in the monomer RMSD graphs. When analyzing the unrestrained simulation of 3tzr, the original 60 ns of simulation showed a non-converged simulation, with the whole unit cell's RMSD (and especially the 4th asymmetric unit) on the rise. The RMSD continued to rise slightly, until around the 240-ns mark, at which point it leveled out for the last 100 ns. The asymmetric units were all between 1.5 and 2.25 Å in the converged section in terms of all RNA heavy atom RMSD, while the unit cell RMSD stayed between 2.0 and 2.4 Å. This shows a dynamism to the structure similar to that of 2a43, despite the fact that the average asymmetric unit structure RMSD was nearly 0.5 Å higher for 3tzr. The base heavy atom RMSD was even a bit higher than the all heavy atom RMSD, with the asymmetric units more spread out, between 1.5 and around 2.5 Å, and the overall unit cell RMSD closer to 2.5 Å. The RMSD using both sets of atoms is fairly similar to that of 2a43, a higher resolution structure with fewer residues per asymmetric unit, which was surprising, even with the difference in number of asymmetric units . One possible reason for this similarity in stability could be due to a larger number of base pairs within the structure of the asymmetric unit of 3tzr as compared to that of 2a43. This seems to be borne out by the average structure-original asymmetric unit overlay

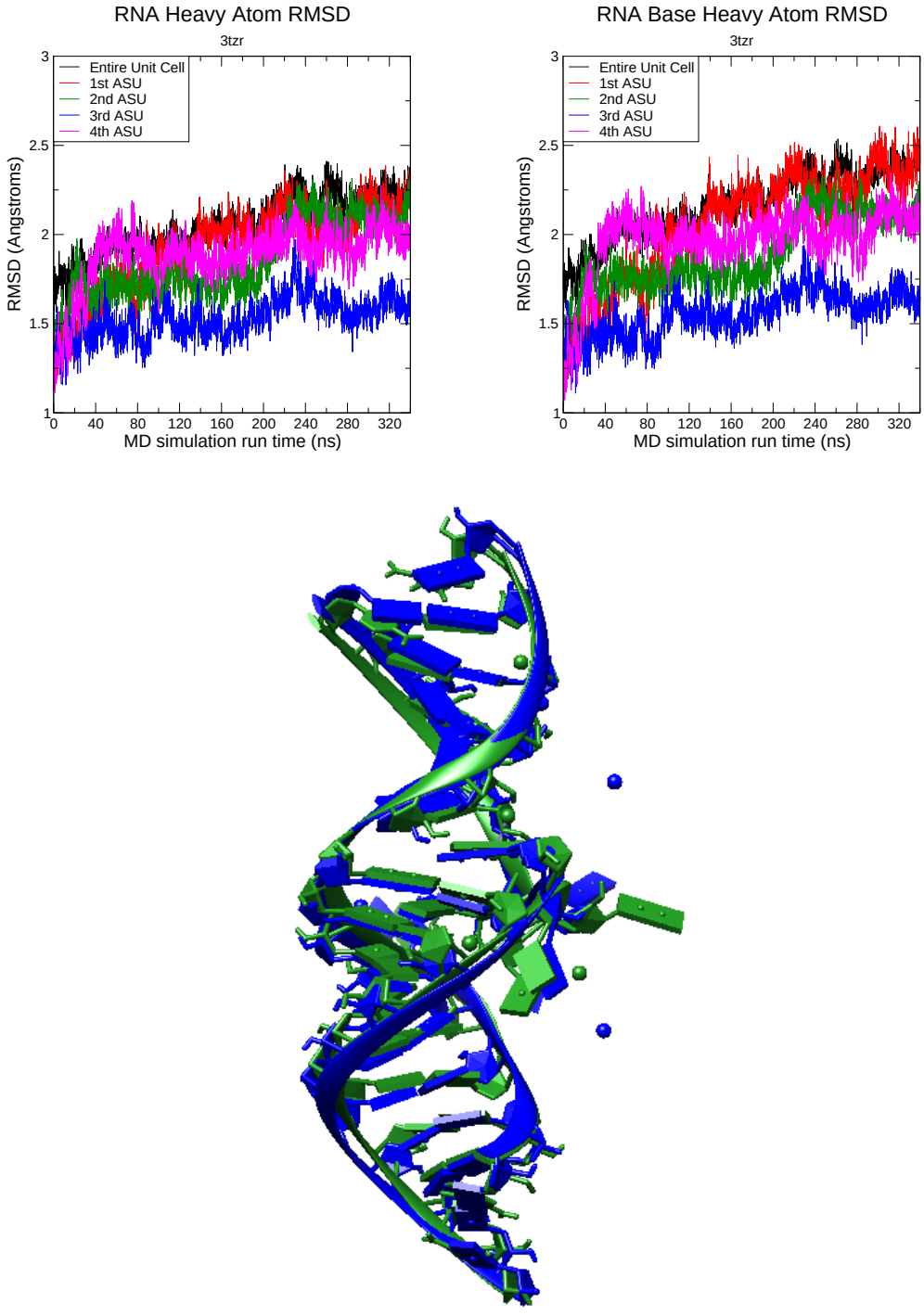

Figure 4.4: RMSD analysis of 3tzr crystal simulation. This simulation took the longest to converge, but the RMSD is relatively the same as that found for 2a43. As with 2a43, the comparison of the average structure to the starting structure shows some differences in the backbone, but most of the deviation occurs in the bases.

image in Figure 4.4, as most of the bases in base pairs seem to be fairly well-overlapped, with the main differences appearing in the sugar-phosphate backbone at the termini of the RNA, as well as the few bases that are not base paired and lie outside the groove.

As was hoped, the simulations for all three structures resulted in convergence, with average pressures that indicated the box was set up properly with the right number of water molecules. While the simulations all seemed stable, the asymmetric units in each structure were not uniform in their dynamics throughout the simulation, with each structure having a spread of RMSD across the asymmetric units (or monomers for 2oiu). The RMSD for 2oiu was much higher than for either of the other 2 structures, but that was expected based on other simulations of the same structure in the literature. Based on the images comparing the average structure to the starting structure, it appeared that most of the structural difference resided in bases that floated outside of the base-pairing groove and in residues near the termini of the chains. There were persistent differences in the bases, but there still appeared to be base pairs. One difference between these simulations and crystalline conditions was the choice to neutralize the RNA with only sodium cations. This will surely be found to be different in 3D-RISM output.

4.3.2 3D-RISM solvation results

As the focus of this chapter is to examine the ability of the 3D-RISM code to replicate and approximate crystallographic solvent conditions as found in crystal MD simulations, it was important to perform calculations with differing solvent concentrations, including conditions approximating crystallization conditions. It also was decided to be relevant to determine what effects differing concentrations of monovalent and divalent ions had on the solvent predicted. In order to analyze the effects of these differences, numbers of ions and water molecules were recorded and compared from the output of the different calculations.

The selections of solvent species and concentrations in Table 4.4, besides the last choice for each structure, were mostly arbitrary in their choice. The concentrations were generally chosen as ones that would likely be in relevant ranges for crystals and ones that already had solvent files created for them, and the ions were chosen to see the

PDB	Solv	mdel	MG	Wat	NA	K	CL	MG%	ERISM
2a43	A	0.3	60	3166	0	31	1	79.5	-4527.65
	B	0.3	47	3129	57	0	1	62.3	-4537.14
	C	0.3	40	3111	71	0	0	53.0	-4595.51
	D	0.5	74	3162	6	0	5	96.1	-4361.46
3tzr	A	0.5	68	1608	0	24	0	85.0	-2515.11
	B	0.5	53	1568	54	0	0	66.3	-2562.08
	C	0.4	45	1550	70	0	0	56.3	-2592.64
	E	0.2	62	1580	37	0	0	77.0	-2637.33
2oiu	A	0.5	110	8066	0	66	2	76.9	-13274.28
	B	0.5	86	8001	114	0	3	60.1	-13275.14
	C	0.5	73	7965	139	0	1	51.2	-13390.55
	F	0.5	121	8048	47	0	4	83.7	-13333.59

Table 4.4: Calculation results for differing concentration setups for the three structures. The last setup for each is a modified version of the concentration of ions in the crystallization solvent, using Na for all monovalent cations (except for in the case of high concentration, such as the high molar concentration of Li in 2oiu), Mg for the divalents, and Cl to neutralize the cations. The water counts for each structure are rather in agreement with each other, while the ion counts seem to have a dependence both on relative concentrations of the two salts involved and on ionic radius. (Solvent A=20 mM $MgCl_2$, 140 mM KCl; Solvent B=20 mM $MgCl_2$, 140 mM NaCl; Solvent C=10 mM $MgCl_2$, 100 mM NaCl; Solvent D=100 mM $MgCl_2$, 50 mM NaCl, simulation of experimental conditions for 2a43; Solvent E=7.5 mM $MgCl_2$, 30 mM NaCl, simulation of experimental conditions for 3tzr; solvent F=35 mM $MgCl_2$, 75 mM NaCl, simulation of experimental conditions for 2oiu; mdel=*mdiis_del*, Solv=solvent)

differences in how ion distribution was affected by differing monovalent cations and by changes in relative monovalent cation and divalent magnesium ion concentrations.

The column for *mdiis_del* shows the value set to handle the step size of MDIIS, which helps accelerate the iterative solving for the direct correlation function. Generally, 0.5 was the starting point, but decreasing the step size was necessary in some cases to aid convergence. The next columns bear the weight of the data, showing to-the-nearest-whole-particle numbers for ions and water molecules determined by the density calculated. While some of these numbers of ions may not add up to a neutralizing charge when compared to the charge of the rest of the system, this is due to the rounding, and the overall charge of the densities calculated always neutralized the system. Further analysis included the MG%, which is the percentage of positive charge supplied by magnesium ions as rounded to the nearest whole number of ions.

A few trends appeared in this data. As one would expect, as concentrations of

particular ions were increased, the ion count for those ions increased, and the opposite trend occurred when those concentrations were decreased. Generally speaking, it looks as if the driving force, when the monovalent salt concentration and magnesium concentration both changed, was the ionic radius of the monovalent salt, and more the change in relative concentration between the two ions than really the absolute concentration of either salt. As there was a switch (within structures, comparing solvent setups) in which monovalent salt was used, from potassium to sodium, there was a large increase in the number of monovalent cations (and a corresponding decrease in magnesium ions). This was somewhat surprising, as there was a lower absolute concentration of sodium (100mM) than that of potassium (140mM); however, the relative concentration of sodium is higher with the lower magnesium concentration (10mM) than the potassium with a higher magnesium concentration (20mM). Also, the ionic radius of sodium is much smaller than that of potassium, allowing for more of it to fit into channels that were too small for Mg ions. To check to see what the exact cause for this was, a calculation was performed for each structure with equal concentrations of magnesium chloride (20mM) and sodium chloride (140mM) to those in the magnesium chloride and potassium chloride calculations. In each case, there are fewer sodium ions than in the corresponding calculation with 10 mM magnesium and 100 mM sodium, but there were still more sodium ions than potassium ions in corresponding calculations. This indicates that the larger number of magnesium ions in the potassium calculation than the 10 mM magnesium, 100 mM sodium calculation is both due to an increase in magnesium concentration, and the ionic radius difference between potassium and sodium.

Regarding the number of water molecules, for each structure, the difference between the highest number found and the lowest number found is at most 101 water molecules, with the largest percentage difference (range divided by lowest number of waters) being 3.74%. The percentage difference for the largest absolute difference is 1.27%. This lack of large differences should not be surprising, as the water "concentration" in each solvent file was basically the same, with small differences by decimal points, all around 55 M. It does appear as though the water molecule numbers were somewhat dependent on ion counts. For all three structures, the RISM calculation with the lowest water molecule

count had the fewest magnesium ions, and the calculations with the second fewest number of waters were the calculations with the second fewest number of magnesium ions. Actually in 3tzr, the relationship between magnesium ion count (and inversely, monovalent cation count) and water molecules has no exceptions: the greater number of magnesium ions, the greater the number of water molecules. In the other two structures, what breaks this trend is chloride ions: the calculation that has the second-most water molecules has the most magnesium ions, but also has the most chloride ions (5 in 2a43, 4 in 2oiu). This appears to indicate that in general, as more of the positive charge was supplied by monovalent cations with larger ionic radii than the magnesium ions and taking up more space, it prevented more water molecules from inhabiting that space. Also, the relatively large ionic radius of chloride ions, when they were included in the calculation, and the added number of cations needed to neutralize their charge, led to fewer water molecules being able to inhabit the volume they took up. This trend is more easily seen in the MG% column, where in each structure, as the percentage of positive charge provided by magnesium ions increases, the number of water molecules also increases, except in the cases where the numbers of chloride ions are higher, where the water counts are a bit lower, even though the MG% is higher. This all indicates that water molecule counts are heavily influenced by volume occupation by ions.

One last area of focus here is the number of chloride ions in each calculation. Out of twelve total calculations in the table, only seven had any chloride ions resulting from the calculation. None of these calculations resulted in more than five chloride ions. This was surprising on the surface, considering the concentrations of chloride needed to neutralize the cations in solution. However, considering the high negative charge on the RNA molecules in a relatively small space, as well as the large ionic radius of chloride ions, it was not surprising to see many of the calculations having no chloride ions, especially in the smallest unit cell (and a structure already containing large sulfate ions), 3tzr, where none of the calculations resulted in any chloride ions in the unit cell. The numbers of chloride ions are lower for the calculations that had more monovalent cations, as well, hence the lack of chloride ions in the 2a43 calculation with the largest number of monovalent cations. This further indicates the importance of the ionic radius

of the cations used to neutralize the RNA in what other particles can be in the unit cell, as the monovalent cations have larger ionic radii and decrease the already limited volume available for chloride ions to inhabit.

One last piece of 3D-RISM analysis was performed by looking at the ion densities provided by the calculations for 2a43. In this case, the solvent setup was solvent D, or the experimental conditions (100mM Mg, 50mM Na). In Figure 4.5, the particle density was compared to the Laplacian for the same particle density as provided by *metatwist*. The main difference between the density and Laplacian is that the Laplacian shows a more condensed prediction of ion placement as compared to the density itself, for all ions and oxygens from the water molecules. For the magnesium portion of the calculation, there are far more locations found than for the sodium portion. This was expected by looking at the ion counts in Table 4.4. Also, the magnesium locations are almost all along the sugar-phosphate backbone of the RNA, as expected for such a negatively-charged portion of a molecule. The sodium locations are far more spread out. Water oxygen locations are far more numerous and do cover some of the deposited locations. No areas of high density of any ion or atom are found in the deposited magnesium ion locations. This was surprising, as such high-resolution data would likely result in properly placed ions, or at least proper electron density locations that may be misassigned. Further analysis of this result should be performed.

In order to assess the new 3D-RISM code, 4 different calculations were performed on each of 3 structures, 3 of which were the same calculation for each structure, with the 4th trying to mimic the experimental conditions in which the crystals were grown. The numbers of water molecules, which for select calculations in Table 4.4 will be used for comparison of 3D-RISM to crystal MD simulations, were fairly similar, with no more than a 3.74% spread from the lowest to highest number of waters within a structure. A trend in numbers of ions (and water molecules) found in the unit cell was found to coincide with ionic radius of the ions used in the calculation. Most notably, in calculations with equal magnesium concentrations and equal, but higher, concentrations of monovalent cations, with sodium in one calculation and potassium in the other, the number of magnesiums was always higher in the potassium calculation, seemingly in

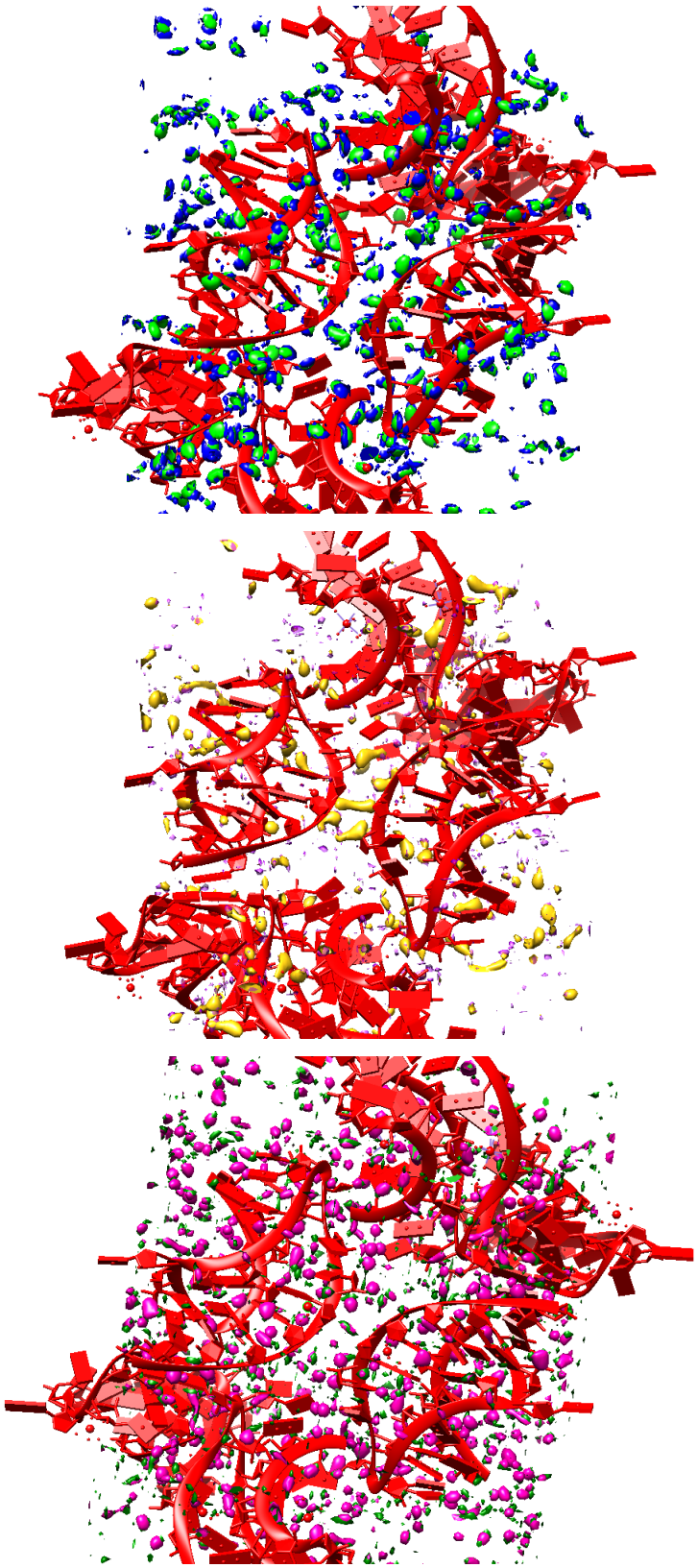

Figure 4.5: Solvation results for 2a43 with magnesium, sodium, and water respectively. The densities for the three are shown in blue, purple, and dark green, respectively, while the Laplacians are seen in light green, orange, and pink, respectively. In all three cases, the Laplacian locations are more condensed, and as expected, there are far more water and magnesium locations than those for sodium.

order to provide neutralization of the solute without creating clashes due to the larger ionic radius of potassium as compared to both sodium and magnesium. This ionic radius point was somewhat of a surprise, and the magnesium counts in all calculations were somewhat surprising due to numbers of magnesiums found in the deposited structures, but this will be addressed in the next subsection. Visual analysis of the results for 2a43 are found in Figure 4.5, where it was seen that particle densities follow the numerical results, and ionic locations make sense electrostatically. The main issue to be analyzed in the future is that no density appeared to exist where the deposited magnesium ions were placed.

4.3.3 Comparison between methods

As the focus of this chapter was to provide a comparison of the more established crystal MD simulations and the 3D-RISM code developed by current and former members of the Case and York labs and presented in Chapter 3 of this thesis, this subsection is the crux of the chapter. That comparison was best made by looking at the numbers of water molecules found for each method in each structure. As the ions for the MD simulations were arbitrarily chosen by using just the deposited magnesium ions and only using sodium ions to neutralize the unit cell, ion counts became a possible source of explanation for differences in water molecule numbers. Final comparison and conclusions about the efficacy of the new 3D-RISM code were made.

Table 4.5 shows the comparative data for the crystal simulations, which used deposited magnesium ions and added sodium ions to neutralize the solute, along with explicit water molecules; and two different 3D-RISM calculations for each structure: one that used 10 mM magnesium chloride and 100 mM sodium chloride, and one for each structure that tried to do a good job of mimicking the experimental settings of the crystal growing for these structures. The focus was to see whether the 3D-RISM calculations did a good job of approximating the explicit MD solvent description, especially the number of water molecules.

As can be seen in the table, in each case, the experimental conditions RISM calculations resulted in higher water molecule counts than in the crystal MD simulations,

PDB	Res(Å)	Solv	RISM Solvent	Wat	MG	NA	CL
2a43	1.34	MD	N/A	3125	12	126	0
		RISM	C	3111	40	71	0
		RISM	D	3162	74	6	5
3tzr	2.21	MD	N/A	1549	24	112	0
		RISM	C	1550	45	70	0
		RISM	E	1580	62	37	0
2oiu	2.60	MD	N/A	7968	14	256	0
		RISM	C	7965	73	139	1
		RISM	F	8048	121	47	4

Table 4.5: Comparison of differing RISM methods and explicit water crystal MD simulations regarding solvent description. There is reasonable agreement regarding water numbers, but differences in ion counts. (Solv=solvent method)

while the 10mM $MgCl_2$, 100mM NaCl calculations had numbers of water molecules much closer to, if not lower than, the MD number. MD simulations always had more sodium ions than the RISM calculations. The first point about the water molecule counts appears to be intertwined with the second point about the ion counts. As mentioned in subsection 4.3.2, it appears that as the number of magnesium ions increased, the number of water molecules that were able to occupy volume in the unit cell increased. As the crystal MD simulations used only the deposited magnesium ions, which are usually ones close to the RNA that are coordinating some residues or a conserved water molecule and have strong enough scattering to have density in the structure, it meant that the sodium ions bore the weight of neutralizing the RNA, resulting in much higher sodium ion counts for the MD simulations. As these simulations had much higher sodium ion counts and far lower magnesium ion counts than the RISM calculations, that meant that they had less volume for water molecules to fill. The calculations with the largest difference in water molecules from the MD have very high magnesium ion counts, possibly leading to their high volume available for water molecules. This is possibly the cause in the differences in water molecule counts between the methods.

That being said, the key conclusion to draw from this table is that the RISM calculations are approximating the number of water molecules in the unit cell fairly well. The largest difference in water molecule count from MD to RISM was 80 waters, in a system where that was 1.00% of the number of waters in the MD simulation. The

largest percentage difference was in 3tzr, and was 2.00%. This structure also has the fewest waters, no matter the method, out of all of them, meaning that a difference of 31 waters, as seen here, results in a 2.00% difference, but in 2oiu would be less than 1%.

Despite these minute water molecule count differences, it appears that 3D-RISM has approximated the number of water molecules in the unit cell fairly well in comparison with crystal MD simulations, while also taking much less time to reach this goal. These 3D-RISM calculations took on the order of 5 to 50 minutes when run in parallel (they took a bit longer in serial calculations, which are required to properly output the density maps), while the MD simulations took on the order of 80 to 110 minutes per 20-nanosecond run, with an undetermined number of MD runs of that length needed to determine the exact proper number of water molecules, and then more runs needed to fully converge the simulation.

One more way to compare the crystal MD/deposited structures and RISM is the placement of ions. It does no good to predict numbers of ions used to neutralize a system if the ions that are predicted do a horrible job replicating the ion locations determined by experiment in the localized areas near the RNA. While analyzing the results in 2a43 (found in Figure 4.5), the ion locations did not match, but further work looking into this structure and the others needs to be performed to really provide any definitive understanding of the accuracy of the ion placement.

Before concluding discussion of these results, it is important to consider the ion count differences between the crystal MD simulations and 3D-RISM calculations. Due to the fact that only well-diffracting ions are placed in crystal structures, and the fact that these ions do not provide enough positive charge to neutralize the highly negatively-charged RNA molecules, it can be assumed there are more cations in the bulk solvent within the crystal unit cell, and the distribution of these cations between monovalent and divalent cations is unknown. These periodic 3D-RISM calculation results show that there is a possibility that the divalent to monovalent ratio is far higher than assumed for the crystal MD simulations in this chapter. If correct, these predictive results could provide a better understanding of the bulk solvent and better fit to scattering data when used in refinements. As alluded to in the following paragraph, crystal MD simulations

with divalent and monovalent cation counts as found in the experimental 3D-RISM calculations are being performed to study the effects on RNA stability as well as on water molecule counts and bulk solvent description.

As the comparison was made between the solvent descriptions achieved by crystal MD simulations and differing solvent setups for 3D-RISM, the water molecule counts and ion counts were analyzed. It was found that the water molecule numbers found across methods, while not identical, were similar enough to be considered properly approximated, with the largest percentage difference being 2.00% from the explicit molecule count in the MD simulation for 3tzr to the highest number found by 3D-RISM for that structure. For all three structures, it was found that the MD water molecule count was never the highest of the three methods. This appears to have been due to the large number of sodium ions required to neutralize the RNA when only using the deposited magnesium ions, which are likely not the only ones in the crystal, just the ones that resulted in enough scattering to provide density for the ions to be placed into. As the sodium ion counts decreased into the RISM output, the numbers of magnesium ions were higher due to their smaller ionic radii, and it appears this was the reason for the larger water numbers in the higher RISM outputs. For further confirmation of the accuracy of 3D-RISM, as well as study of the effects on RNA stability and bulk solvent description, similar numbers of magnesium ions to those found in the 3D-RISM calculations are being introduced into crystal MD simulations, and the resultant number of water molecules will be determined from those simulations as they were with the ion distributions in Table 4.5. If the water molecule numbers rise, it will be further confirmation of the relation of water counts with ionic radii of the ions used to neutralize the solute. As it is, the 3D-RISM results appear to match the crystal MD simulation results sufficiently and provide possibly interesting insight into the distribution of neutralizing positive charge among divalent and monovalent cations throughout the crystal unit cell.

4.4 Conclusions

The focus of this chapter was to investigate ways in which the solvent in a crystal can be described. The two specific methods studied here were crystal molecular dynamics

simulations and 3D-RISM using the code developed by current and former members of the Case and York labs. The more particular focus of this chapter was to compare said 3D-RISM code to the standard of crystal simulations. This was done through comparison of water molecule counts and ion counts from the MD simulations and 3D-RISM calculations performed on three RNA structures–2a43, 3tzr, and 2oiu–at 1.34, 2.21, and 2.6 Å resolution, respectively, in relatively simple systems. The MD simulations and 3D-RISM calculations were analyzed separately, and then compared to check for accuracy of the new code.

For all three structures, crystal simulations reached convergence, albeit at different lengths of unrestrained simulation. For 2a43 and 3tzr, the overall unit cell RMSD values for all RNA heavy atoms and the base heavy atoms were rather similar from structure to structure, while there were some differences in these values for the asymmetric units from structure to structure and within each structure. While these RMSD values were very similar, the average asymmetric unit structure's RMSD from the starting asymmetric unit for 3tzr was nearly 0.5 Å higher than that for 2a43. The similarity of the full unit cell RMSD values in light of this average structure RMSD difference appears to rise from the larger number of asymmetric units in 2a43, allowing for propagation of differences in each asymmetric unit to add up to a greater difference as opposed to that found from fewer asymmetric units. The third structure, 2oiu, had much larger RMSD values than those found for either of these structures. The values were actually similar to, if not lower, than those found in simulations in the literature, however. The one oddity was that the undocked conformation monomers were lower in RMSD than the docked conformers, which was opposite the results found in Giambaşu, et al.[43] The overall takeaways from these simulations were that stable water molecule numbers had been found to stabilize the crystal dimensions for the boxes in the simulations, and that the structures in general showed some fluctuation in the base pairs, but the largest differences occurred at the termini of the chains and in the bases that floated outside of the base-pairing grooves.

For the 3D-RISM results from all three structures, a few conclusions were reached. First, it was found that the numbers of water molecules across different solvent setups

for each structure stayed within 3.74% of each other, indicating that the calculations are able to be consistent regarding the number of water molecules, but that there is some fluctuation as the ionic concentrations and species are changed. Second, the ionic radii of the ions used affected both the number of water molecules predicted and the balance of divalent and monovalent ions. When potassium, the largest cation used in terms of ionic radius, was used, far more magnesium ions were predicted than when sodium was the monovalent cation, and more waters were found in the unit cell. Third, along with ionic radii, the relative concentrations of ions played more of a role in determining the distribution of positive charge among the magnesium ions and monovalent cations. Even in cases where the magnesium concentration was very reduced relative to other calculations for the same structure, the number of magnesium ions could be higher than for other calculations if the sodium concentration decreased by a larger ratio than the magnesium concentration did.

After these points were discovered, the crux of this chapter was reached: how did the 3D-RISM calculations compare to the more-established crystal MD simulations? For each structure, the data from the simulations were compared to the data from the experimental conditions 3D-RISM calculations and a simple two-salt magnesium and sodium calculation with the same concentrations from structure to structure. In all three cases, the experimental 3D-RISM calculations resulted in higher numbers of water molecules found than the crystal simulations, although the numbers were all within 2.00% of the MD water count. None of the differences were larger than 80, indicating good agreement with the MD solvent description. When including ideas about ionic radii and which ions appear in the system, it might be possible to account for some of these differences in water count, as the only magnesium ions in the simulations were the deposited ones, whereas each of the 3D-RISM calculations resulted in far more magnesium ions than what was deposited in the PDB. However, this was not a trend that held up across all RISM/MD comparisons. 3D-RISM calculations here had the added bonus of being much faster than crystal MD simulations at arriving at a stable number of water molecules to solvate the system.

While it was found that the 3D-RISM calculations approximated the water counts

well, further work could be done to substantiate this claim. First, performing simulations with added magnesium ions to meet the magnesium values found in the 3D-RISM calculations could further confirm the idea that the number and size of different ions affected the water counts. If correct, my hypothesis would indicate that the number of waters found in the simulation would be higher as the number of smaller cations used increased. This work could also provide more insight on the ionic makeup of the bulk solvent in the crystal unit cell. Second, comparing the placement of ions in the RISM density outputs to the deposited ions would indicate the ability of 3D-RISM to properly place ions based on the electrostatics and steric factors of the structures. Work to look at this was performed on 2a43, and there is not enough data from this one structure to properly get a sense of how the code compared to deposited ion placements. Finally, the expansion of this study to more structures would be helpful to further support the accuracy of the code.

Chapter 5

Sarcin/ricin domain as a case study for RNA simulations

5.1 The sarcin/ricin domain

The sarcin/ricin domain of the ribosomal RNA, or the SRL RNA, contains the largest conserved region of the ribosomal RNA across all species[30, 107, 31, 65]. It is actually a well-known, well-studied example of RNA containing an internal loop, the type of which contained in this domain is known as a bulged-G motif[30, 107, 31, 65]. This domain was actually one of the very first structures discovered containing the bulged-G internal loop motif[30, 107]. This motif contains a "bulged G" that forms a base triple[30, 31]. On each side of this triple, non-canonical base pairs exist, followed by canonical ones[30, 107, 31, 65]. The domain also contains a GAGA tetraloop[30, 107, 31, 65].

The sarcin/ricin domain of the ribosomal RNA is an important binding site for elongation factors involved in translation[30, 107, 31, 65]. Actually, binding of the elongation factors prevents chemical changes to this segment of the RNA[77, 107]. The translational process is disrupted by the toxins for which this domain is named when either one (or a mimic) breaks a covalent bond, preventing binding of the elongation factors. This shuts down the formation of proteins[30, 107, 31, 65], leading eventually to cell death[107, 30]. What is remarkable is that synthetic mimics of this domain have been found to simulate both form and function of the native structure, including acting as binding partner for the elongation factor EF-G and as target for sarcin and ricin action[30, 31].

As this domain contains a well-conserved sequence across species, plays such an important role in translation and the life cycle, and is relatively small, it has been an oft-studied structure. As such, there are 537 structures in the PDB that result from a query for "sarcin/ricin", the majority of which have been solved at or below 2.5 Å[11].

Of the 537, 32 are from some form of *E. coli*, and 20 of those are actually EM structures that are of whole ribosomal subunits, while the other 12 are crystallographic structures. The three structures that are chosen for this chapter, however, do not show up in this set of *E. coli* structures due to classification issues, so there are likely more. The size, general high resolution but range of resolutions, number, and understanding of these structures lend to them the quality of being good test structures for comparisons of methods across resolutions.

In this chapter, the three different sarcin/ricin domain structures from *E. coli* will be used to test PHENIX refinement with and without AMBER restraints at different resolutions of the same structure. They will also be used to test out the effects on ion and water distributions developed from 3D-RISM of different resolutions for structures of the same sequence, as well as what effects there are from differing concentrations of the same ions. Finally, one of the structures, due to its small number of atoms, was used as a test structure for minimizations with 3D-RISM.

5.2 Methods

5.2.1 Structure selection and description

The three structures chosen, PDB ID 1q9a[30], 480d[31], and 483d[31], all come from *E. coli* 23 S rRNA, and are all the wild type sequence of the sarcin/ricin domain of the rRNA. They are all the same 27-mer RNA, with data collection for 1q9a and 483d (at 1.04 and 1.11 Å resolution, respectively) performed at 100 K, while the data was collected for 480d (1.50 Å) at 295 K. The three structures are all free of ions and crystallization solvent, having only water molecules (differing numbers) in the structure besides the single RNA molecule in the asymmetric unit. The unit cells are all slightly different in size, as can be seen in the Volume column of Table 5.7. As these are all high-resolution RNA structures, with the same sequence and no complicating factors in the structures (difficult to build ligands or modified residues, etc.), they make a great set of structures to compare the effects of data resolution, water molecule count, and unit cell volume on structure refinement with and without AMBER restraints and on

solvent description in 3D-RISM. The small size of the structure also provides a great test case for testing out different parameters with minimizations with 3D-RISM solvent description and their effects on runtime.

5.2.2 Refinement methods and parameters

The refinement data from Chapter 2 for these three structures were collected as described in Chapter 2. In this chapter, the data are presented again, but are analyzed by focusing only on these three structures and searching for trends in energy values or crystallographic or geometric statistics as a result of resolution or number of water molecules in the structure. Further analysis of the RMSD between the structures was also performed.

5.2.3 Solvation studies of sarcin/ricin

The structures were prepared for 3D-RISM calculations by first removing all of the water molecules, and then using *phenix.AmberPrep* to build the necessary unit cell parameter and coordinate files. 3D-RISM calculations using *rism3d.snglpnt* were then performed for all three structures using grid spacings of 0.35 in all three dimensions, to the kh and pse2 closures to tolerances of 0.1 and 0.000001, respectively. The calculations were performed at two different solvent conditions, which appear in Table 5.7.

Analysis of the results was performed by comparing the RISM charges and counts of magnesium ions, potassium ions, chloride ions, and water molecules. As the structures were the exact same solutes, it was expected that there would be very little, if any, differences from structure to structure at a certain solvent concentration setup. Having no ions in their structures, though, it was to be interesting to see what ions might exist in the crystals due to their differing crystallization conditions.

5.2.4 Time-step tracking of minimization with periodic 3D-RISM

As periodic 3D-RISM calculations are being implemented in minimizations and MD runs and the periodic RISM forces are being used to guide said simulations, it is important to

have a good sense of what combinations of parameters and numbers of processors provide the most time-efficient calculations. In order to gain this knowledge, the PDB-REDO structure (chosen as PDB-REDO's versions of these structures were being studied) of 1q9a was prepared in the same way as all three structures were prepared in subsection 5.2.3–by removing all the water molecules and then running the structure through *AmberPrep*. Many different short (10-20 steps) minimizations were performed on the structure using differing numbers of cores and processors (with *mdiis_del=0.40*), as well as differing minimization engines and RISM parameters such as *npropagate* (number of old solutions used to predict the next solution; can speed up calculations), *mdiis_del* (MDIIS step size), *grdspc* (grid spacing), and *mdiis_nvec* (vectors MDIIS uses). Differing closure options were also tested for time steps, but all other parameters were tested in minimizations where the RISM setup called for kh and pse2 closures to be converged to 0.1 and 0.000000001 tolerances, respectively. These RISM calculations were performed using 0.100M KCl as the solvent, with the xvv file made in *rism1d* to the pse2 closure. Average time per step and RISM energy values were recorded to determine which combinations of parameters were fastest and to check whether the resultant RISM calculations were in line with the other ones.

5.3 Data and discussion

5.3.1 Refinement data and trends

As these three structures are only different, regarding what residues they contain, in terms of number of water molecules, but were all solved at different resolutions, it is an interesting case study to analyze these structures and the resultant AMBER and conventional output structures from parallel refinements in PHENIX. Presumably, the main differences leading to how AMBER restraints versus conventional restraints affect energy differences would be the differences in solvent and resolution-dependent differences in the structures. Presumably, also, there would be little to no difference in the geometric statistics at the small difference in resolution.

In Table 5.1, the energy data from Table 2.2 for these three structures are presented

again for close focus on the three related structures. In all three, the bonds and angles in the conventionally refined structures provided more favorable bond and angle energies, but the angles do trend more toward them being almost energetically interchangeable relative to those in the AMBER-restrained structures. This is likely due to the less defined electron density that results from the data collected in the room temperature structure. The dihedral angles are all relatively similar energetically between refinement methods.

PDB	Res(Å)	B	A	D	E	F	W	V	R	EFR	TOT
1q9a	1.04	1.17	1.07	0.13	0.10	-2.03	-0.24	-0.54	0.36	-1.56	0.02
483d	1.11	0.48	0.24	0.13	-0.55	-1.96	-0.27	-0.55	0.83	-1.68	-1.65
480d	1.50	1.16	0.16	-0.01	-0.45	-3.14	-0.52	-0.63	1.40	-2.19	-2.03

Table 5.1: Per-nucleotide energy values (kcal/mol) for each of three sarcin/ricin structures used in Chapter 2. The electrostatic and non-bonded interactions, as well as the total energy, appear to be more favorable for AMBER restraints as the resolution worsens. (B=bond, A=angle, D=dihedral, E=electrostatics, F=1-4 electrostatics, W=1-4 non-bonded, V=van der Waals, R=RISM, EFR=E+F+R, TOT=EP Tot)

Regarding non-bonded interactions, the AMBER-restrained structures for all three structures generally were more favorable, with a slight trend toward a greater improvement in these areas with decreased resolution, especially in 1-4 electrostatics and 1-4 non-bonded interactions. The only of these non-bonded areas where there was a lack of improvement due to AMBER restraints was in electrostatics for 1q9a. This was likely due to the high resolution of the data leading to the conventionally restrained structure sticking very closely to the data, and the resultant structure being represented well energetically.

The RISM energy is more favorable for conventionally refined structures, and becomes moreso as resolution worsens. This is due to the screening of electrostatics and 1-4 electrostatics that the RISM potential does. That combined set of interactions trends toward increased AMBER favorability, thus explaining the reverse trend for RISM energy. As the water molecules are all removed before these calculations, the effects of the numbers of water molecules are not very easily established here.

The overall energy difference between refinement methods does appear to have a trend here based on resolution. The overall energy difference between AMBER- and

conventionally restrained refinement output structures increases toward higher AMBER favorability as the resolution worsens. This was also observed over the course of the whole structure set in Table 2.2.

When looking at the geometric statistics in Table 5.2 and how the differences between the AMBER- and conventionally restrained refinement output structures relate to resolution or number of water molecules, there did not appear to be a real trend in any of the statistics. Part of that was due to the fact that it is hard to discern trends between values for three structures, but part of that is truly that no real trends showed themselves. The r-work improvement by conventional refinement as compared to AMBER-restrained refinement increased as resolution worsened, likely due to the greater likelihood of AMBER restraints to try to correct energetic issues and take the structure away from the experimental data a slight bit more. However, it might be due to a lesser chance of AMBER over-fitting the data than the conventional refinement, as evidenced by the trend toward a smaller r-gap for AMBER-restrained refinements at lower resolution.

PDB	Res(Å)	R-work	R-free	R-gap	Clash	Bond	Angle	Pucker	Suite	Suiteness
1q9a	1.04	0.0012	0.0083	0.0071	-2.29	0.028	0.102	0	0	-0.009
483d	1.11	0.0065	0.0133	0.0068	0	0	0.046	0	0.009	0.034
480d	1.50	0.0065	-0.0006	-0.0071	0	0.028	0.037	0.009	0	-0.015

Table 5.2: Differences in crystallographic and geometric statistics between conventionally and AMBER-restrained refinement output structures. Bond, angle, pucker, and suite values are differences in numbers of outliers of that type per nucleotide. All other values are absolute differences. Negative values indicate more favorable values for AMBER-restrained refinement, positive values indicate favorable conventional refinements. There are no cohesive trends found in this data relative to resolution or numbers of waters.

The only other real trend that emerges is in the angle outliers per nucleotide difference, where, as the resolution worsens, the difference in number of outliers decreases. This is due to, surprisingly enough, fewer angle outliers in the AMBER structures as the resolution decreases. This is likely due to, again, the greater freedom allowed by the less defined electron density at lower resolution, and thus more room for both energetically favorable and data-fitting structural details such as angles.

One more level of structural analysis performed for these structures not performed

for the full PHENIX data set was cross-structure and cross-method RMSD analysis. In Table 5.3, the heavy atom RMSD of the RNA molecules in the first asymmetric unit was calculated using different PDB IDs and different structures for those PDB IDs as the reference structures (D stands for deposited, C is the conventional refinement output, and A is the AMBER refinement output). For the first column of RMSD in the table, the output structures for each PDB ID are the reference and input structures. The differences between the output structures at each PDB ID are roughly the same for each structure. However, the conventional output structures move slightly less far away from the deposited structures than the AMBER structures do.

PDB	Res(Å)	A to C	C to D	A to D
1q9a	1.04	0.072	0.117	0.139
483d	1.11	0.076	0.084	0.115
480d	1.50	0.078	0.102	0.110

Table 5.3: RMSD (Å) measurements for the three structures, looking at differences between the AMBER (A) and conventional (C) refinement output and deposited (D) structures for a particular PDB ID. For example, the 0.072 value in A to C for 1q9a indicates that the heavy atom RMSD for the AMBER output structure for 1q9a from the conventional output structure is 0.072 Å. All three structures show little deviation amongst their refinement output structures, while there is greater deviation from refined structures to the deposited. There is also greater deviation from the AMBER-refined structures to the deposited ones as compared to conventional to deposited.

PDB	Res(Å)	D1q9a	D483d	D480d
1q9a	1.04	0	0.231	0.380
483d	1.11	0.231	0	0.324
480d	1.50	0.380	0.324	0

Table 5.4: RMSD (Å) measurements for the three structures, looking at differences between the deposited (D) structures from the other PDB IDs to a particular PDB ID. For example, the 0.231 value in D1q9a for 483d indicates that the heavy atom RMSD for the deposited structure for 483d from the deposited structure for 1q9a is 0.231 Å. As expected, the greatest deviation is from the lowest-resolution structure to the highest-resolution structure.

When doing cross-structure RMSD (seen in Tables 5.4, 5.5, and 5.6), a very slight trend that might be inferred appeared, showing that the refinements bring the structures slightly closer together than the deposited structures are, likely due to the fact that the structures are all the same regarding the physical source material. The values found

PDB	Res(Å)	C1q9a	C483d	C480d
1q9a	1.04	0	0.198	0.334
483d	1.11	0.198	0	0.299
480d	1.50	0.334	0.299	0

Table 5.5: RMSD (Å) measurements for the three structures, looking at differences between the conventional (C) refinement output structures from the other PDB IDs to a particular PDB ID. For example, the 0.198 value in C1q9a for 483d indicates that the heavy atom RMSD for the conventional output structure for 483d from the conventional output structure for 1q9a is 0.198 Å. The values here are less than in Table 5.4, indicating that the refinement brings these structures closer together than their deposited predecessors. Also, again the lowest- and highest-resolution structures have the greatest deviation.

PDB	Res(Å)	A1q9a	A483d	A480d
1q9a	1.04	0	0.176	0.322
483d	1.11	0.176	0	0.307
480d	1.50	0.322	0.307	0

Table 5.6: RMSD (Å) measurements for the three structures, looking at differences between the AMBER (A) output structures from the other PDB IDs to a particular PDB ID. For example, the 0.176 value in A1q9a for 483d indicates that the heavy atom RMSD for the AMBER output structure for 483d from the AMBER output structure for 1q9a is 0.176 Å. There is generally less deviation here than in either Table 5.5 or Table 5.6, indicating that the use of AMBER restraints draws these structures, all the same solute, closer together, as they should be fairly similar. Again, the greatest deviation is from lowest- to highest-resolution structures.

for the deposited structures compared to each other are nearly 0.05 Å larger than for the AMBER or conventional refinement when comparing 1q9a to 480d, and there is a similar, but smaller difference for 1q9a to 483d and 483d to 480d. However, it is such a small difference that it is more than likely that it is not a real trend, but more of an interesting coincidence.

While it was expected there might be effects due to the larger number of water molecules in the higher resolution structure than the lower resolution structures, and that there would likely be trends in many of the statistics analyzed related to resolution, there only appeared to be trends related to resolution, and not a lot of them. This is likely due to the small number of structures making it hard to observe trends in data, and also due to the fact that little of the analysis performed indicates anything related to the number of water molecules. The few trends found in data based on the resolution

were mostly energy related, showing that conventionally restrained bonds and angles were more energetically favorable at high resolution, while the electrostatics and non-bonded interactions were far more energetically favorable at lower resolution than their conventionally refined counterparts. No real trends were found that were any different than those found in Chapter 2, and the statistics did not show any particular similarity either that one might expect from structures that are as similar as these are.

5.3.2 Solvation results: ion and water counts

As there were no ions in any of the deposited structures for these threes structures, it seemed likely to be a good idea to perform 3D-RISM calculations using different solvent setups, using some of the ions found in the experimental crystallization setup. Also, since there were more water molecules in the higher resolution structures, it seemed like a good idea to analyze how that occurred, considering that the structures are all roughly the same. These structures were also all high resolution and small, and thus seemed like opportunities to get very well resolved solvent densities.

There are a few conclusions to draw from the calculations that were performed. First, the ion counts are the same within rounding error for each PDB ID with the same solvent conditions. This was not surprising due to the solute charge being the same and the structure being roughly the same. Second, as in Chapter 4, as ion concentrations change, the ion distribution changes, albeit not linearly. As the potassium concentration increased (and the magnesium concentration decreased), the charge stayed the same, but the distribution of positive charge between the magnesium and potassium ions changed. However, even though the magnesium concentration was cut in half and the potassium concentration was doubled, the resultant changes in ion counts are not a strict doubling or halving. The changes in ion counts are relative, not absolute in terms of quantity.

Finally, the water molecule counts, among results for the same calculation type, did not follow the trend of the deposited structures, where the number of water molecules decreased from high resolution to low resolution. Here we see the highest resolution having the second most water molecules for each calculation, 483d (in the middle in terms of resolution) with the fewest, and the lowest resolution structure having the

most water molecules. It makes sense that the trend for deposited structures would be the way it is, due to the fact that the highest resolution data would be most likely to have the largest number of high density areas in the solvent region corresponding to defined water molecules. This trend for the RISM results, though, was somewhat surprising due to the nature of the solutes being the same, as well as the solvent setups being the same. While a difference of 50 or 60 waters from crystal MD to 3D-RISM is not a problem, having that large of a difference for the same calculations with very similar structures is not likely, unless there's another reason. The only possible reason seems to be a difference in the structures that was not noticed. At this point, the crystal dimensions were checked, and the crystal sizes for these three structures actually were different regarding edge lengths of the unit cell. While the space group and angles for the unit cell were the same, these edge differences resulted in differences in volume as seen in Table 5.7. When taking into account that a water molecule takes up roughly 30Å^{3}of volume [100], a theoretical calculation of water molecule difference based on the difference in volume (TWD) was performed using the equation $(V_{480d} - V_{smaller})/30\mathring{A}^3$, where $V_{smaller}$ is the volume of the smaller unit cell of the two structures being compared (480d has the largest unit cell of the three structures). The numbers in the TWD column were compared to the AWD, or actual water difference, from the calculation with the largest number of waters (480d for both solvent types), and were very similar. This indicates that the difference in numbers of water molecules in these calculations was due to the difference in volume between unit cells for these structures. It is very likely that the true numbers of water molecules in the unit cells in these crystals follow the trend seen in the RISM data (increasing with volume) as opposed to that seen in the deposited structures (increasing with better resolution).

These structures, which had no ions in their deposited files in the PDB and contain the same solute, were great choices to be analyzed via 3D-RISM calculations. The results showed that the distribution of ions using a specific solvent setup ended up being the same for each structure, which was not very surprising. Also, as seen in Chapter 4, changes in concentrations in the solvent resulted in expected relative changes in ion distribution without matching the exact change in concentrations. When the potassium

PDB	Res(Å)	Solv	Charge	Wat	AWD	MG	K	CL	Vol	TWD
1q9a	1.04	G	104.0	1286	56	49	7	0	66770	59
		H	104.0	1260	57	38	29	0		
483d	1.11	G	104.0	1277	65	49	6	0	66361	73
		H	104.0	1251	66	38	29	0		
480d	1.50	G	104.0	1342	0	49	7	0	68549	0
		H	104.0	1317	0	37	29	0		

Table 5.7: Comparison of 3D-RISM results using different solvent concentrations. The water number differences seem to match the differences expected based on volume differences, while the ions using the same solvent are the same within rounding error. (Solv=solvent; TWD=theoretical water difference from largest unit cell water molecule count, using 30 $Å^3$as the volume per water molecule, and comparing only those calculations with the same concentrations; AWD=actual water difference from number found for 480d, the structure here with the largest unit cell; Vol=volume; Solvent G=20mM $MgCl_2$, 50mM KCl; Solvent H=10mM $MgCl_2$, 100mM KCl)

concentration doubled and the magnesium concentration was halved, the number of potassiums increased and the number of magnesiums decreased. However, the changes were not strictly the same as the changes in concentration. This shows that there's not much of quantitative relationship to be found in particle counts versus concentrations other than the relative change in particle counts, which follows the relative change in ion concentrations. The more surprising result was that the difference in number of water molecules from PDB ID to PDB ID was rather large for structures that were so similar. However, once the size of the unit cells was analyzed, it became clear that the difference in water molecules was likely due to the difference in volume for the unit cells. The next step to really tease out how well these calculations worked, as well as seeing similarities and differences between these structures, would be to check the ion densities and water densities and see where the ions and waters line up from structure to structure, and whether any of the water density blobs line up with the deposited water molecules in these structures.

5.3.3 Time step results

As previously mentioned, these small structures are of a good size to provide a test case for minimization with 3D-RISM solvent description and forces. The smaller number of atoms results in fewer degrees of freedom in each calculation, allowing for quicker

calculations, while still showing differences based on parameters. As it so happens, the structure chosen for these minimizations is not technically one that was used in other experiments in this chapter: it is the PDB-REDO structure for 1q9a as opposed to the deposited structure. The waters were stripped and the structure was run through *AmberPrep*, thus providing the same necessary starting structure files for RISM as in the RISM calculations above. This was purely an arbitrary choice of continuing with calculations of a structure that had been experimented with and seemed to converge fairly quickly for RISM steps. No atom or residue changes were made in the solute of this structure, just adjustments of residues to improve density fit by re-refinement and rebuilding of the structure. While this likely results in differences in the ERISM and time steps as compared to the deposited structure, the relative differences for differing parameters within use of this structure should not be different.

In Table 5.8, the comparisons performed are of differing numbers of steps and differing minimization methods, even without RISM for a benchmark (3 NR). For minimizations where the minimization method was *ntmin=2*, this equates to the use of the steepest descent minimization method. For *ntmin=3*, this results in minimizations using XMIN, which requires a submethod choice, of which TNCG, or the "optionally LBFGS-preconditioned Truncated Newton Conjugate Gradient" method, was chosen. This particular method actually results in multiple substeps being performed per cycle of minimization, meaning that far more steps are performed than just 100 for the minimizations in this table that used this method. The time/step was calculated using the number of these substeps used, as opposed to 100 cycles, as the denominator for *ntmin=3* rows in the tables.

As can be seen for this table, the number of seconds per step is different for each method, with it being the most for the 10-step *ntmin=2* minimization, while the minimization without RISM is obviously the fastest. Of those with RISM, the *ntmin=3* minimization is faster than those for *ntmin=2*. However, when looking into the amount of time spent just setting up the minimization and RISM calculations in the output, it seems likely that most of this difference is due to the larger number of steps. For instance, it does not seem likely that just by changing the number of minimization steps,

ntmin	steps	time/step (s)	ERISM
2	10	122.4	-559.2
2	100	114.8	-600.0
3	100	109.0	-1110.8
3 NR	100	2.3	N/A

Table 5.8: Timing and RISM excess chemical potential for single-core, 16-processor minimizations in *sander.MPI*, with differing minimization methods and step numbers to look at their effects. Minimizer choice for *ntmin=3* is TNCG, and each step requires multiple neutralizing steps (this is the number, not 100, used for the time per step calculation). Here, for minimizations with RISM, the fastest was with *ntmin=3*; however, this is likely due to the larger number of "real" steps averaging out the startup time. (NR=no RISM)

the step speed would change within the use of the same minimizer. The difference in the first two rows is mainly due to the startup time of the minimization being averaged out over more steps. The same likely holds true for the jump from *ntmin=2* to *ntmin=3*, as there are far more substeps in the *ntmin=3* than just 100 cycles, and thus, that startup time gets further averaged out with the larger number of steps. Nothing conclusive can really be drawn from this table, other than that the RISM energy decreases further with more steps of minimization in this case, and that minimization without calculating and applying RISM forces is far faster than with the calculation and application of said forces.

In Table 5.9, comparisons are made between 10-step minimizations using the steepest descent minimization method. On the *perceval* cluster, since shuttered, there were plenty of resources for running minimizations in parallel. The *sander.MPI* utility was used with differing numbers of nodes and processors to see what combination was the fastest for RISM minimizations of this structure. While it was expected that the larger number of processors used, the faster the minimization would progress, this was not the case. The sweet spot appeared to be 16 processors on 1 node, resulting in a speed of about 2 minutes per step. It was somewhat odd to see slight differences in the RISM energy, although the minimization would not necessarily be the same every time it is run. One interesting point is that when these minimizations are run, the output prints a warning message if the number of processors is not a power of 2, indicating that it is likely that this might be an issue that causes the 20-processor run to take longer

than the 16-processor run. Similar runs were performed for the TNCG minimization method, with the results appearing in Table 5.10. Again, the fastest combination was 1 node, 16 processors, and here it was about 13 seconds faster per minimization step, but that is over the course of 146 substeps as compared to 10 steps for the steepest descent method. It therefore seems likely that the difference is due to the greater number of steps over which the setup time for the RISM grid is averaged. Also, the greater number of substeps results in the overall time of minimization being much longer for TNCG, and it is less controllable than the steepest descent method. Further calculations were performed with the steepest descent method.

nodes	processors	time/step (s)	ERISM
1	1	1166.5	-558.9
1	2	614.2	-558.9
1	4	365.5	-559.2
1	8	215.4	-559.2
1	10	181.7	-559.4
1	16	122.4	-559.2
1	20	130.4	-559.6
2	16	165.7	-559.2
2	20	172.9	-559.6
2	32	199.8	-559.2
2	40	288.6	-559.6
3	30	264.2	-559.4
3	60	315.0	-559.6
4	32	268.2	-559.2
4	40	277.5	-559.6
4	64	415.4	-559.1
4	80	383.3	-559.5
8	64	357.4	-559.1

Table 5.9: Timing and RISM excess chemical potential for 10-step minimizations in *sander.MPI* with differing sets of cores and processors, using *ntmin=2* as the minimization method. Surprisingly, the fastest combination is just 1 node with 16 processors.

All of the above minimizations were performed with *mdiis_del* values of 0.40. In order to see what effect this step size has on time, these minimizations were performed with the steepest descent method and varying *mdiis_del* values. This is one of the parameters adjusted when convergence of the 3D-RISM calculations fails, so it was not surprising to see some of these calculations fail. The RISM energy appears in the 0.10

nodes	processors	time/step (s)	ERISM
1	10	165.1	-683.3
1	16	109.5	-717.0
1	20	125.9	-662.3
2	20	164.4	-732.4
2	40	264.8	-698.7
3	30	165.0	-683.9
3	60	284.7	-713.1
4	40	262.9	-730.6
4	80	302.4	-625.0

Table 5.10: Timing and RISM excess chemical potential for 10-step minimizations in *sander.MPI* with differing sets of cores and processors, using *ntmin=3*, TNCG as the minimization method. As with *ntmin*=2, the fastest combination is 1 node with 16 processors.

row because the minimization failed on the 8th step of the minimization. The other failures occurred before even the first step of minimization could be finished. The use of *mdiis_del=0.50* resulted in the fastest minimization time, as seen in Table 5.11.

mdiis_del	time/step (s)	ERISM
0.10	failed	-557.9
0.20	180.3	-559.2
0.30	135.5	-559.2
0.40	122.4	-559.2
0.50	115.5	-559.2
0.60	failed	N/A
0.70	failed	N/A
0.80	failed	N/A
0.90	failed	N/A

Table 5.11: Timing and RISM excess chemical potential for 10-step minimizations in *sander.MPI* with 1 core, 16 processors, using *ntmin=2*, and differing values for *mdiis_del*. Most options here resulted in failure to converge, while the fastest runs were with an *mdiis_del* value of 0.50.

Other parameters tested can be seen in Tables 5.12 and 5.13. In the first of these two, grid spacing was adjusted, looking at the effects of widening or narrowing the grid from the 0.35 setting used in all of the previous minimizations. Also, in this case, 20 processors and an *mdiis_del* value of 0.40 were used for arbitrary reasons. As expected, the smaller the grid spacing, the longer it took to perform each RISM calculation. However, it was surprising that the largest grid spacing actually ended up being much

slower than slightly smaller grids. This appears to be due to a much longer time in the FFT portion of the calculations. In the second table, the parameters being tested are the number of iterations within the RISM calculation (*mdiis_nvec*) used to help find the next solution and the number of previous RISM solutions used to make the next RISM guess (*npropagate*). The value that was used in all previous minimizations for *mdiis_nvec* was 10, and *npropagate* was set to 0. In the table, we find that using any number of iterations less than 5 results in failure of convergence, likely due to the freedom provided by fewer iterations being saved to stray away from the tolerance being sought. There was very little difference in time per step for 5 iterations used versus 10 iterations. When regarding the number of previous solutions used to guide the guess for the next RISM calculation, it appears that saving at least one solution to guide the next guess had a marked impact on speed, but using a second solution in conjunction with the first did not, likely due to the fact that these were 10-step minimizations, and waiting until two solutions were saved meant that the second solution did not have the benefit of an educated guess, slowing the minimization down a little bit. Over the course of a longer minimization, that brief slowdown on the second step could very likely be made up for by the increased speed of a greater number of subsequently faster steps.

grid spacing (Å)	time/step(sec)	ERISM
0.25	254	-558.2
0.35	129	-559.6
0.40	75	-560.5
0.50	215	-557.1

Table 5.12: Timing and RISM excess chemical potential for 10-step minimizations in *sander.MPI* with 1 core, 20 processors, using *ntmin=2*, *mdiis_del=0.40*, and differing values for grid spacing. The grid spacing value is in each direction, creating a grid that is cubic with the given side length. As expected, as the grid spacing was smaller, the minimizations took longer, as there were more grid points to evaluate. Surprisingly, the largest grid spacing also resulted in a long time step.

Due to the relatively small size of 1q9a, any version of this structure, such as the PDB-REDO version here, would make a great test case for longer calculations made shorter by the smaller size. As such, the PDB-REDO version of the 1q9a structure was used in many different *sander.MPI* minimizations with periodic RISM forces. Differing numbers of nodes and processors were tested, finding that regardless of minimization

mdiis_nvec	*npropagate*	time/step (s)	ERISM
1	0	failed	N/A
2	0	failed	N/A
3	0	failed	N/A
5	0	129	-559.6
10	0	130	-559.6
10	1	101	-559.6
10	2	106	-559.6

Table 5.13: Timing and RISM excess chemical potential for 10-step minimizations in *sander.MPI* with 1 core, 20 processors, using *ntmin=2*, and differing values for *mdiis_nvec* and *npropagate*. Parameters for *mdiis_del* and grid spacing were set to 0.40 and 0.35 (in each direction), respectively. Using too few vectors resulted in failure to converge, while including a previous solution for guessing the next (*npropagate>0*) resulted in a shortening of the time step.

method chosen, 1 node and 16 processors was the fastest combination, despite the fact that it was expected that the speed would scale with number of processors used. "Sweet spots" were found regarding *mdiis_del* (0.50) and grid spacing (0.40) values, although these values were not generally tested at the same time. Finally, it was found that using *npropagate* (ie, setting it to something other than 0) sped up the minimization. While having more solutions used in making an initial guess did not make the minimization any faster in this test, it seems likely that over the course of a longer minimization, the added saved solutions could speed up the minimization even more.

5.4 Conclusions

Due to its high conservation across species and small size, the sarcin/ricin domain of the ribosomal RNA of *E. coli* provided a great test structure in order to have varying resolution and solvent content. This variety among structures with the same solute provided for a nice, small set of structures on which to test the PHENIX/AMBER interface, periodic 3D-RISM singlepoint calculations, and periodic 3D-RISM minimizations.

During analysis of the previous PHENIX refinements of 1q9a, 483d, and 480d, it was found that while it was expected there might be differences in the statistics of the structures based on the number of waters, there were no observed effects when comparing AMBER- and conventionally restrained refinements. Resolution-dependent

effects seen in Chapter 2 across the broader data set appeared here as well, however, where the overall and interaction (electrostatics, non-bonded) energy values became more favorable for AMBER-restrained refinements as the resolution worsened. There were few trends, if any, found in the geometric data. This could very easily be due to the small number of structures, so further work including more sarcin/ricin domain structures (ideally wild-type like these) could be useful in searching for trends. Also, looking at absolute values, instead of method-dependent comparison values, for statistics could be useful, especially in determining the effects of the number of water molecules in these structures. Further analysis of electron-density fit would likely be the best way to judge the effects of the number of placed water molecules in each structure.

When these structures were solvated using periodic 3D-RISM singlepoint calculations with differing concentrations of potassium and magnesium ions, it was expected that, due to the similarities in the solute of each structure, and looking at no other information, it was expected that the water molecule and ion counts would be the same with a particular solvent setup. However, the number of water molecules was different for each structure within a particular ion concentration setup. When the unit cell dimensions were examined, it was found that 480d was a larger unit cell than either of the other two structures. The number of water molecules was found to increase as the unit cell volume increased. This difference in volume appears to be the cause of the trend, as the differences in numbers of water molecules were found to be very similar to the numbers found by dividing the differences in unit cell volume by 30 $Å^3$per water molecule. Regarding ion counts, it was found that the number of magnesium ions and potassium ions found for each structure were the same (within rounding error) in calculations with the same concentrations for each ion. This was to be expected, as there were no ions found in the deposited structures, and the solute charge and structure were the fairly much the same for each structure. However, when the ion concentrations were changed, as found in Chapter 4, the resultant changes in ion counts were not linear with the change in concentrations. The ion counts did change in accordance with the relative concentration changes, however. Further work on these structures could be performed

to see what differences would be found in ion placement and water counts at experimental crystallization conditions (which are different for the three structures), which have thus far led to non-converged calculations, as well as examining the similarities and differences in ion placement and water placement across the three structures with the already converged calculations. Also, it would be interesting to see what effects placement of ions and/or waters from RISM density might have on refinement of these three structures in both conventionally and AMBER-restrained refinements.

The use of the PDB-REDO structure of 1q9a as a test structure for minimizations with periodic RISM forces allowed for many tests to be run within relatively short periods of time. These tests tried out different combinations of nodes, processors, minimization methods, *mdiis_del*, *mdiis_nvec*, *npropagate*, and grid spacing. It was found that running these minimizations with 1 node and 16 processors was the fastest combination of hardware used, despite the expectation that speed would scale with processor usage. Further work to analyze this could be performed with longer minimizations on each of the combinations of processors and nodes. Regarding the minimization methods, it was hard to determine whether the steepest descent method or the TNCG method was faster at performing these minimizations with periodic RISM forces, as the number of steps was not really equal. The best way to truly determine which is faster would be to use the steepest descent method for a number of steps equal to the number of neutralization steps used by TNCG in a "10-step" minimization in order to directly compare these methods and give the same amount of steps for the initial setup time to be averaged over. When using the steepest descent method, different levels of *mdiis_del*, *mdiis_nvec*, *npropagate*, and grid spacing were tested to find those that worked fastest. However, some of these tests were performed with the non-optimal values for the other parameters and processors, so it would be ideal to truly determine the ideal options by testing them all at their optimal levels and really test across ranges of all the values. It was interesting to find that using at least one previous solution (*npropagate>0*) sped up finding the solution in the next RISM step, but that using more than one solution did not appear to speed it up any faster than using just one. Ideally, running a longer minimization could help us truly see if that is the case, or if that was just an artifact

caused by only running 10 steps of minimization and taking more than one of those to build up a memory of previous solutions to use slowed enough of the minimization down to outweigh the benefit that could be seen over more steps.

The small set of structures used here, along with the shorter minimizations performed, provided a brief overview of how these techniques are working on structures with the same structure. However, any real trends that might appear will likely need more structures to provide space for those trends to show. Also, longer minimizations, as most minimizations are, and different solvent conditions, such as the experimental conditions, for 3D-RISM calculations, would better allow us to fully grasp what is going on in these structures as we study them.

Chapter 6

Conclusions and future directions

This accumulation of five years of work covers many different topics, and also leaves large amounts of work performed in that time out. As work on larger structures revealed issues in how nucleic acids were being handled, the focus of this work shifted to RNA and how it was handled by the different methods involved in this work, as a lot of the focus for previous work with these techniques had been with proteins. Other work was performed in collaboration with other labs, but did not amount to anything publishable. There was, of course, still plenty to cover.

In Chapter 1, background information about RNA and methods of interest was provided as a primer for understanding the work performed. This was a short introduction, and discussion of the PHENIX/AMBER interface and its use in RNA structures was entered into in Chapter 2. Here, it was found that the use of AMBER force fields to restrain the geometry of the structures provided the added benefit of electrostatic and non-bonded energies to the restraints guiding refinement, providing more energetically favorable structures, as expected when scoring structures refined with and without AMBER restraints against the AMBER force fields. Also, the inclusion of non-bonded terms improved clashscores for these structures. However, AMBER-restrained refinement led to more geometric outliers as determined by MolProbity. One test bond and one test angle were investigated to determine why there were such persistent outliers for AMBER-restrained refinement output structures. As AMBER ideal bond lengths and angles for those chosen were modified both in terms of the ideal value and the force constant, it was determined that the large number of outliers, at least for the chosen bonds and angles, were due to both differing ideal values for AMBER and MolProbity and AMBER's force constants allowing for more freedom in bonds and angles to allow for

the best structure energetically to be reached by including electrostatic and non-bonded terms. Essentially, the AMBER refinements led to larger distributions of bond angles and lengths, allowing for more measurements to be outside the ±4-sigma range demarcating "real" outliers. There were also some slight improvements in hydrogen bonding found with AMBER-restrained refinements due to the energetic favorability of hydrogen bonds. The best improvement upon conventional refinement found in this chapter was that of the low-resolution structures. The structures reached by AMBER-restrained refinements of the low-resolution structures were highly energetically favorable to their conventionally refined counterparts, and far superior regarding clashscores. Further work could be done to cement the superiority of AMBER restraints in PHENIX refinement of RNA structures by increase the size of the data set. Also, test sugar pucker outliers and suite outliers could be chosen to analyze the reason for increased numbers of such outliers in AMBER-restrained refinements and whether or not it is for the same reasons found for bond lengths and angles.

In Chapter 4, new periodic 3D-RISM code, as presented in Chapter 3, was tested on three of the RNA structures from the data set in Chapter 2 and compared to the standard for solvent description, crystal MD simulations. The MD simulations were stable and settled on reasonable numbers of water molecules. Periodic 3D-RISM single-point calculations were performed for each structure with four different solvent setups, testing the effects of different ion types and concentrations on water molecule and ion counts produced by the code. It was found that all of the water counts produced by the differing solvent setups were within 3.74% of the lowest count for within a PDB ID, and the largest gap was 101 waters, in a structure where that translated to 1.27% of the smallest number of waters. It was found that ionic radii seemed to have an effect on both ion counts and water counts, as a general trend of higher water counts with higher numbers of magnesium ions was found. As more of the positive charge was accounted for by smaller, divalent magnesium ions, there appeared to be more volume available to be occupied by water molecules. Also, when comparing solvents with sodium as opposed to potassium, there were more sodium ions than potassium ions at the same concentrations, likely due to the smaller ionic radius of sodium. Relative concentrations were also

found to have more of an effect on ion distribution than the absolute concentrations. As the 3D-RISM calculations containing magnesium and sodium were compared to the crystal MD results, the main conclusion was that the water counts found through 3D-RISM were very close to the MD counts. The largest percentage difference was 2.00%, and the largest gap was only 80 waters. This is very promising as to the ability of the periodic 3D-RISM calculations to approximate closely the results of the far more time-expensive MD simulations. These calculations also resulted in interesting divalent to monovalent cation ratios, indicating that our understanding of the bulk solvent and the way we model it in crystal MD simulations may need tweaking. Visual analysis of the placement of ion and water density was performed on 2a43 with crystallization solvent conditions, showing that the majority of the magnesium density surrounded the negatively-charged portions of the sugar-phosphate backbone. However, the deposited magnesium sites were not replicated. Further analysis of the effectiveness of these calculations would involve more visual checking like this for all the structure and solvent combinations to see how well the output ion and water densities matched placement of the deposited ions and water molecules, especially near the RNA molecules. Also, closer comparison of the crystal MD simulations and 3D-RISM calculations could be performed by placing the number of magnesium ions found in 3D-RISM calculations into simulations to see what effect the change in the divalent ion:monovalent ion ratio has on the number of water molecules in the unit cell, and on the overall solvent description and RNA stability within the unit cell..

In Chapter 5, the highly conserved sarcin/ricin domain of the ribosomal RNA was chosen as a test structure for different methods, due to its small size and the high resolution of some of the structures. The three structures analyzed here were already included in the PHENIX/AMBER interface data set in Chapter 2. The data from their refinements was analyzed to see what effects the differing numbers of deposited water molecules and the resolutions of the structures had on refinement results. Only resolution appeared to have an effect on improvements due to the use of AMBER restraints in terms of energy, and no real trends appeared to exist in geometric outlier data. Regarding how periodic 3D-RISM singlepoint calculations turned out for these

three structures, it was found that different numbers of water molecules filled the unit cells of the structures, but not in the same order as they did in the deposited structures. The higher the resolution of the deposited structure, the more water molecules there were in the deposited structure. However, when it came to the 3D-RISM results, the highest number of water molecules was found in the unit cell with the highest volume, which actually happened to be the structure with the lowest resolution. The differences in number of water molecules appeared to be due to the difference in volume, as they were roughly the same as the theoretical differences found when assuming that a water molecule fills 30 $Å^3$of volume. The ion counts, on the other hand, were the same within rounding error for each set of concentrations, and did not change linearly with the change in concentrations. This was not surprising, as the solutes were the same, and thus had the same charge that needed to be neutralized. Finally, the PDB-REDO version of 1q9a was used as a test case for short minimizations in *sander.MPI* with periodic 3D-RISM forces. Testing a bunch of different sets of parameters, certain "sweet spots" for parameters were found, but further work needs to be performed to really decide exactly what the best options are for the best and fastest results, likely in longer minimizations. Further work to get a better grasp of what is going on in these structures with PHENIX would likely include comparing absolute energies and geometric statistics across resolution and water molecule counts as opposed to looking at the differences between methods for each structure. Also, trends are hard to dissect from only three structures, so other sarcin/ricin structures could be chosen to be refined to give more data through which to comb for trends. In the case of the 3D-RISM results for these structures, the next steps should be to test solvent setups that give a semblance of the experimental crystallization conditions for the three structures (as this was tried, but proper convergence failed; also, they have different experimental conditions) and to compare the placement of the ion and water densities from structure to structure.

Overall, the work presented in this dissertation covers a wide range of ways to look at RNA structures and hopefully provide insights in the future. The energetic and clash-score improvements found by implementing AMBER force field restraints in PHENIX refinement should lead to more physically accurate RNA structures, including better

ligand binding interactions, hopefully leading to better understanding of interactions for RNA-interacting therapeutics and of ribozymes and the like. Periodic 3D-RISM calculations can hopefully provide a better understanding of water- and ion-coordinated interactions within structures and with ligands, again leading to better understanding of therapeutic interactions. Ideally, further use of periodic 3D-RISM, both in singlepoint calculations and MD could even lead to better overall descriptions of bulk solvent scattering and thus improved agreement of models with experimental data. While I hope that these visions of what this work can lead to become realities, there is still plenty of work needed to get there.

References

[1] Pavel V Afonine, Ralf W Grosse-Kunstleve, and Paul D Adams. A robust bulk-solvent correction and anisotropic scaling procedure. *Acta Crystallographica Section D: Biological Crystallography*, 61(7):850–855, 2005.

[2] Rommie E Amaro, Xiaolin Cheng, Ivaylo Ivanov, Dong Xu, and J Andrew McCammon. Characterizing loop dynamics and ligand recognition in human-and avian-type influenza neuraminidases via generalized born molecular dynamics and end-point free energy calculations. *Journal of the American Chemical Society*, 131(13):4702–4709, 2009.

[3] Ramu Anandakrishnan, Aleksander Drozdetski, Ross C Walker, and Alexey V Onufriev. Speed of conformational change: comparing explicit and implicit solvent molecular dynamics simulations. *Biophysical journal*, 108(5):1153–1164, 2015.

[4] Klas M Andersson and Sven Hovmöller. The protein content in crystals and packing coefficients in different space groups. *Acta Crystallographica Section D: Biological Crystallography*, 56(7):789–790, 2000.

[5] P Auffinger and E Westhof. Roles of hydration on the structure and dynamics of nucleic acids. *Water Management in the Design and Distribution of Quality Food*, pages 165–198, 1999.

[6] Pascal Auffinger and Eric Westhof. Water and ion binding around rna and dna (c, g) oligomers. *Journal of molecular biology*, 300(5):1113–1131, 2000.

[7] Y. Bai, M. Greenfeld, K.J. Travers, V.B. Chu, J. Lipfert, S. Doniach, and D. Herschlag. Quantitative and Comprehensive Decomposition of the Ion Atmosphere around Nucleic Acids. *J. Am. Chem. Soc.*, 129:14981–14988, 2007.

[8] N. Ban, P. Nissen, J. Hansen, P.B. Moore, and T.A. Steitz. The complete atomic structure of the large ribosomal subunit at 2.4 ï¿œ resolution. *Science*, 289:905–920, 2000.

[9] Rodolphe Barrangou, Christophe Fremaux, Hélène Deveau, Melissa Richards, Patrick Boyaval, Sylvain Moineau, Dennis A Romero, and Philippe Horvath. Crispr provides acquired resistance against viruses in prokaryotes. *Science*, 315(5819):1709–1712, 2007.

[10] Robert T Batey, Sunny D Gilbert, and Rebecca K Montange. Structure of a natural guanine-responsive riboswitch complexed with the metabolite hypoxanthine. *Nature*, 432(7015):411, 2004.

[11] H. Berman, K. Henrick, H. Nakamura, and J.L. Markley. The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucl. Acids Res.*, 35:D301–D303, 2007.

[12] Helen M Berman. Hydration of dna. *Current Opinion in Structural Biology*, 1(3):423–427, 1991.

[13] Helen M Berman. Hydration of dna: take 2. *Current Opinion in Structural Biology*, 4(3):345–350, 1994.

[14] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne. The Protein Data Bank. *Nucl. Acids Res.*, 28:235–242, 2000.

[15] Christian Biertümpfel, Ye Zhao, Yuji Kondo, Santiago Ramón-Maiques, Mark Gregory, Jae Young Lee, Chikahide Masutani, Alan R Lehmann, Fumio Hanaoka, and Wei Yang. Structure and mechanism of human dna polymerase η. *Nature*, 465(7301):1044–1048, 2010.

[16] Marina Bousquet, Guoqing Zhuang, Cong Meng, Wei Ying, Patali S Cheruku, Andrew T Shie, Stephanie Wang, Guangtao Ge, Piu Wong, Gang Wang, et al. mir-150 blocks mll-af9–associated leukemia through oncogene repression. *Molecular Cancer Research*, 11(8):912–922, 2013.

[17] AT Brünger, M Karplus, and GA Petsko. Crystallographic refinement by simulated annealing: application to crambin. *Acta Crystallographica Section A: Foundations of Crystallography*, 45(1):50–61, 1989.

[18] Thomas R Cech. The efficiency and versatility of catalytic rna: implications for an rna world. *Gene*, 135(1-2):33–36, 1993.

[19] Thomas R Cech. The ribosome is a ribozyme. *Science*, 289(5481):878–879, 2000.

[20] Thomas R Cech, Daniel Herschlag, Joseph A Piccirilli, and AM Pyle. Rna catalysis by a group i ribozyme. developing a model for transition state stabilization. *Journal of Biological Chemistry*, 267(25):17479–17482, 1992.

[21] David S Cerutti and David A Case. Molecular dynamics simulations of macromolecular crystals. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 9(4):e1402, 2019.

[22] T.V. Chalikian. Volumetric properties of proteins. *Annu. Rev. Biophys. Biomol. Struct.*, 32:207–235, 2003.

[23] T.V. Chalikian. On the molecular origins of volumetric data. *J. Phys. Chem. B*, 112:911–917, 2008.

[24] T.V. Chalikian and R.B. Macgregor, Jr. Nucleic acid hydration: a volumetric perspective. *Phys. Life Rev.*, 4:91–115, 2007.

[25] Jennifer A Chan, Anna M Krichevsky, and Kenneth S Kosik. Microrna-21 is an antiapoptotic factor in human glioblastoma cells. *Cancer research*, 65(14):6029–6033, 2005.

[26] Vincent B Chen, W Bryan Arendall, Jeffrey J Headd, Daniel A Keedy, Robert M Immormino, Gary J Kapral, Laura W Murray, Jane S Richardson, and David C

Richardson. Molprobity: all-atom structure validation for macromolecular crystallography. *Acta Crystallographica Section D: Biological Crystallography*, 66(1):12–21, 2010.

[27] F.-C. Chou, P. Sripakdeevong, S.M. Dibrov, T. Hermann, and R. Das. Correcting pervasive errors in RNA crystallography through enumerative structure prediciton. *Nature Methods*, 10:74–76, 2013.

[28] Maksymilian Chruszcz, Wojciech Potrzebowski, Matthew D Zimmerman, Marek Grabowski, Heping Zheng, Piotr Lasota, and Wladek Minor. Analysis of solvent content and oligomeric states in protein crystalsâdoes symmetry matter? *Protein Science*, 17(4):623–632, 2008.

[29] Le Cong, F Ann Ran, David Cox, Shuailiang Lin, Robert Barretto, Naomi Habib, Patrick D Hsu, Xuebing Wu, Wenyan Jiang, Luciano A Marraffini, et al. Multiplex genome engineering using crispr/cas systems. *Science*, 339(6121):819–823, 2013.

[30] C.C. Correll, J. Beneken, M.J. Plantinga, M. Lubbers, and Y.L. Chan. The common and the distinctive features of the bulged-G motif based on a 1.04 A resolution RNA structure. *Nucl. Acids Res.*, 31:6806–6818, 2003.

[31] C.C. Correll, I.G. Wooland, and A. Munishkin. The two faces of the Escherichia coli 23 S rRNA sarcin/ricin domain: The structure at 1.11 A resolution. *J. Mol. Biol.*, 292:275–287, 1999.

[32] Hassan Dana, Ghanbar Mahmoodi Chalbatani, Habibollah Mahmoodzadeh, Rezvan Karimloo, Omid Rezaiean, Amirreza Moradzadeh, Narges Mehmandoost, Fateme Moazzen, Ali Mazraeh, Vahid Marmari, et al. Molecular mechanisms and biological functions of sirna. *International journal of biomedical science: IJBS*, 13(2):48, 2017.

[33] T. Darden, D. York, and L. Pedersen. Particle mesh Ewald–an Nlog(N) method for Ewald sums in large systems. *J. Chem. Phys.*, 98:10089–10092, 1993.

[34] Laurent David, Ray Luo, and Michael K Gilson. Comparison of generalized born and poisson models: energetics and dynamics of hiv protease. *Journal of Computational Chemistry*, 21(4):295–309, 2000.

[35] Ian W Davis, Laura Weston Murray, Jane S Richardson, and David C Richardson. Molprobity: structure validation and all-atom contact analysis for nucleic acids and their complexes. *Nucleic Acids Research*, 32:W615–W619, 2004.

[36] Fang Ding, Changrui Lu, Wei Zhao, Kanagalaghatta R Rajashankar, Dwight L Anderson, Paul J Jardine, Shelley Grimes, and Ailong Ke. Structure and assembly of the essential rna ring component of a viral dna packaging motor. *Proceedings of the National Academy of Sciences*, 108(18):7357–7362, 2011.

[37] Jan Drenth. *Principles of protein X-ray crystallography*. Springer Science & Business Media, 2007.

[38] U. Essmann, L. Perera, M.L. Berkowitz, T. Darden, H. Lee, and L.G. Pedersen. A smooth particle mesh Ewald method. *J. Chem. Phys.*, 103:8577–8593, 1995.

[39] Michael Feig. Kinetics from implicit solvent simulations of biomolecules as a function of viscosity. *Journal of chemical theory and computation*, 3(5):1734–1748, 2007.

[40] Witold Filipowicz, Suvendra N Bhattacharyya, and Nahum Sonenberg. Mechanisms of post-transcriptional regulation by microrna: are the answers in sight? *Nature reviews genetics*, 9(2):102–114, 2008.

[41] M. Gebala, G.M. Giambasu, J. Lipfert, N. Bisaria, S. Bonilla, G. Li, D.M. York, and D. Herschlag. Cation-Anion Interactions within the Nucleic Acid Ion Atmosphere Revealed by Ion Counting. *J. Am. Chem. Soc.*, 137:14705–14715, 2015.

[42] G. Giambasu, D.A. Case, and D.M. York. Predicting Site-Binding Modes of Ions and Water to Nucleic Acids Using Molecular Solvation Theory. *J. Am. Chem. Soc.*, 141:2435–2445, 2019.

[43] George M Giambaşu, Tai-Sung Lee, Carlos P Sosa, Michael P Robertson, William G Scott, and Darrin M York. Identification of dynamical hinge points of the l1 ligase molecular switch. *RNA*, 16(4):769–780, 2010.

[44] Barbara L Golden, Hajeong Kim, and Elaine Chase. Crystal structure of a phage twort group i ribozyme–product complex. *Nature structural & molecular biology*, 12(1):82–89, 2005.

[45] Jean Pierre Hansen and Ian R McDonald. Statistical mechanics of dense ionized matter. iv. density and charge fluctuations in a simple molten salt. *Physical Review A*, 11(6):2111, 1975.

[46] Jean-Pierre Hansen and Ian Ranald McDonald. *Theory of simple liquids: with applications to soft matter*. Academic Press, 2013.

[47] Thomas Hermann. Rational ligand design for rna: the role of static structure and conformational flexibility in target recognition. *Biochimie*, 84(9):869–875, 2002.

[48] Fumio Hirata, B Montgomery Pettitt, and Peter J Rossky. Application of an extended rism equation to dipolar and quadrupolar fluids. *The Journal of Chemical Physics*, 77(1):509–520, 1982.

[49] Fumio Hirata and Peter J Rossky. An extended rism equation for molecular polar fluids. *Chemical Physics Letters*, 83(2):329–334, 1981.

[50] J.M. Holton, S. Classen, K.A. Frankel, and J.A. Tainer. The R-factor gap in macromolecular crystallography: an untapped potential for insights on accurate structures. *FEBS J.*, 281:4046–4060, 2014.

[51] J.J. Howard, G.C. Lynch, and B.M. Pettitt. Ion and Solvent Density Distributions around Canonical B-DNA from Integral Equations. *J. Phys. Chem. B*, 115:547–556, 2011.

[52] François Jacob and Jacques Monod. Genetic regulatory mechanisms in the synthesis of proteins. *Journal of molecular biology*, 3(3):318–356, 1961.

[53] S. Jain, D.C. Richardson, and J.S. Richardson. Computational methods for RNA structure validation and improvement. *Meth. Enzymol.*, 558:181–212, 2015.

[54] Wenyan Jiang, David Bikard, David Cox, Feng Zhang, and Luciano A Marraffini. Rna-guided editing of bacterial genomes using crispr-cas systems. *Nature biotechnology*, 31(3):233, 2013.

[55] B.J. Johnson, W.E. Antholine, S.V. Lindeman, M.J. Graham, and N.P. Mankad. A One-Hole Cu4S Cluster with N2O Reductase Activity: A Structural and Functional Model for CuZ*. *J. Am. Chem. Soc.*, 138:13107–13110, 2016.

[56] S.M. Kast and T. Kloss. Closed-form expressions of the chemical potential for integral equation closures with certain bridge functions. *J. Chem. Phys.*, 129:236101, 2008.

[57] Ailong Ke and Jennifer A Doudna. Crystallization of rna and rna–protein complexes. *Methods*, 34(3):408–414, 2004.

[58] Kevin S Keating and Anna Marie Pyle. Semiautomated model building for rna crystallography using a directed rotameric approach. *Proceedings of the National Academy of Sciences*, 107(18):8177–8182, 2010.

[59] A. Kovalenko and F. Hirata. Potentials of mean force of simple ions in ambient aqueous solution. I Three-dimensional reference interaction site model approach. *J. Chem. Phys.*, 112:10391–10402, 2000.

[60] A. Kovalenko and F. Hirata. Potentials of mean force of simple ions in ambient aqueous solution II. Solvation structure from the three-dimensional reference interaction site model approach, and comparison with simulations. *J. Chem. Phys.*, 112:10403–10417, 2000.

[61] Andriy Kovalenko and Fumio Hirata. Potentials of mean force of simple ions in ambient aqueous solution. i. three-dimensional reference interaction site model approach. *The Journal of Chemical Physics*, 112(23):10391–10402, 2000.

[62] 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.

[63] Nadia Kulshina, Nathan J Baird, and Adrian R Ferré-D'Amaré. Recognition of the bacterial second messenger cyclic diguanylate by its cognate riboswitch. *Nature structural & molecular biology*, 16(12):1212, 2009.

[64] P.T. Lang, J.M. Holton, J.S. Fraser, and T. Alber. Protein structural ensembles are revealed by redefining X-ray electron density noise. *Proc. Natl. Acad. Sci. USA*, 111:237–242, 2014.

[65] N. Leontis and E. Westhof. A Common Motif Organizes the Structure of Multhelix Loops in 16 S and 23 S Ribosomal RNAs. *J. Mol. Biol.*, 283:571–584, 1998.

[66] Lina Lin, Jia Sheng, and Zhen Huang. Nucleic acid x-ray crystallography via direct selenium derivatization. *Chemical Society Reviews*, 40(9):4591–4602, 2011.

[67] Sarah V Lipchock and Scott A Strobel. A relaxed active site after exon ligation by the group i intron. *Proceedings of the National Academy of Sciences*, 105(15):5699–5704, 2008.

[68] C. Liu, P.A. Janowski, and D.A. Case. All-atom crystal simulations of DNA and RNA duplexes. *Biochim. Biophys. Acta*, 1850:1059–1071, 2015.

[69] Jun Lu, Gad Getz, Eric A Miska, Ezequiel Alvarez-Saavedra, Justin Lamb, David Peck, Alejandro Sweet-Cordero, Benjamin L Ebert, Raymond H Mak, Adolfo A Ferrando, et al. Microrna expression profiles classify human cancers. *Nature*, 435(7043):834–838, 2005.

[70] T. Luchko, S. Gusarov, D.R. Roe, C. Simmerling, D.A. Case, J. Tuszynski, and A. Kovalenko. Three-Dimensional Molecular Theory of Solvation Coupled with Molecular Dynamics in Amber. *J. Chem. Theory Comput.*, 6:607–624, 2010.

[71] T. Luchko, I.S. Joung, and D.A. Case. Integral equation theory of biomolecules and electrolytes. In T. Schlick, editor, *Innovations in Biomolecular Modeling and Simulation, Volume 1*, pages 51–86. Royal Society of Chemistry, London, 2012.

[72] J.A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K.E. Hauser, and C. Simmerling. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theor. Comput.*, 11:3696–3713, 2015.

[73] Brian W Matthews. Solvent content of protein crystals. *Journal of molecular biology*, 33(2):491–497, 1968.

[74] K.L. Meagher, L.T. Redman, and H.A. Carlson. Development of Polyphosphate Parameters for Use with the AMBER Force Field. *J. Comput. Chem.*, 24:1016–1025, 2003.

[75] S. Meisburger, S. Pabit, and L. Pollack. Determining the Locations of Ions and Water around DNA from X-Ray Scattering Measurements. *Biophysical Journal*, 108:2886–2895, 2015.

[76] Vinod K Misra and David E Draper. On the role of magnesium ions in rna stability. *Biopolymers: Original Research on Biomolecules*, 48(2-3):113–135, 1998.

[77] Danesh Moazed, James M Robertson, and Harry F Noller. Interaction of elongation factors ef-g and ef-tu with a conserved loop in 23s rna. *Nature*, 334(6180):362–364, 1988.

[78] Nigel W Moriarty, Pawel A Janowski, Jason M Swails, Hai Nguyen, Jane S Richardson, David A Case, and Paul D Adams. Improved chemistry restraints for crystallographic refinement by integrating the amber force field into phenix. *Acta Crystallographica Section D: Structural Biology*, 76(1):51–62, 2020.

[79] N.W. Moriarty, D.E. Tronrud, P.D. Adams, and P.A. Karplus. A new default restraint library for the protein backbone in Phenix: a conformation-dependent geometry goes mainstream. *Acta Cryst.*, D72:176–179, 2016.

[80] Tohru Morita. Theory of classical fluids: Hyper-netted chain approximation, i: formulation for a one-component system. *Progress of Theoretical Physics*, 20(6):920–938, 1958.

[81] L.J.W. Murray, W.B. Arendall, D.C. Richardson, and J.S. Richardson. RNA backbone is rotameric. *Proc. Natl. Acad. Sci. USA*, 100:13904–13909, 2003.

[82] H.T. Nguyen, S.A. Pabit, L. Pollack, and D.A. Case. Extracting water and ion distributions from solution x-ray scattering experiments. *J. Chem. Phys.*, 144:214105, 2016.

[83] Harry F Noller. Structure of ribosomal rna. *Annual review of biochemistry*, 53(1):119–162, 1984.

[84] A. Okur, L. Wickstrom, M. Layten, R. Geney, K. Song, V. Hornak, and C. Simmerling. Improved Efficiency of Replica Exchange Simulations through Use of a Hybrid Explicit/Implicit Solvation Model. *J. Chem. Theory Computat.*, 2:420–433, 2006.

[85] S.A. Pabit, K.D. Finkelstein, L. Pollack, and D.E. Herschlag. Using anomalous small angle X-ray scattering to probe the ion atmosphere around nucleic acids. *Meth. Enzymol.*, 469:391–410, 2009.

[86] S.A. Pabit, S.P. Meisburger, L. Li, J.M. Blose, C.D. Jones, and L. Pollack. Counting Ions around DNA with Anomalous Small-Angle X-ray Scattering. *J. Am. Chem. Soc.*, 132:16334–16336, 2010.

[87] Pradeep S Pallan, William S Marshall, Joel Harp, Frederic C Jewett, Zdzislaw Wawrzak, Bernard A Brown, Alexander Rich, and Martin Egli. Crystal structure of a luteoviral rna pseudoknot and model for a minimal ribosomal frameshifting motif. *Biochemistry*, 44(34):11315–11322, 2005.

[88] N.S Pannu and R.J. Read. Improved structure refinement through maximum likelihood. *Acta Cryst.*, A52:659–668, 1996.

[89] Maria T Panteva, Thakshila Dissanayake, Haoyuan Chen, Brian K Radak, Erich R Kuechler, George M Giambaşu, Tai-Sung Lee, and Darrin M York. Multiscale methods for computational rna enzymology. In *Methods in enzymology*, volume 553, pages 335–374. Elsevier, 2015.

[90] John Perkyns and B Montgomery Pettitt. A site–site theory for finite concentration saline solutions. *The Journal of chemical physics*, 97(10):7656–7666, 1992.

[91] JS Perkyns and B Montgomery Pettitt. A dielectrically consistent interaction site theory for solventâelectrolyte mixtures. *Chemical Physics Letters*, 190(6):626–630, 1992.

[92] Huang CC Couch GS Greenblatt DM Meng EC Ferrin TE. Pettersen EF, Goddard TD. Ucsf chimera–a visualization system for exploratory research and analysis. *J Comput Chem*, 25(13):1605–12, Oct 2004.

[93] Anna Pyle. Metal ions in the structure and function of rna. *JBIC Journal of Biological Inorganic Chemistry*, 7(7-8):679–690, 2002.

[94] F Ann Ran, Patrick D Hsu, Jason Wright, Vineeta Agarwala, David A Scott, and Feng Zhang. Genome engineering using the crispr-cas9 system. *Nature protocols*, 8(11):2281, 2013.

[95] JC Rasaiah, DN Card, and JP Valleau. Calculations on the "restricted primitive model" for 1–1 electrolyte solutions. *The Journal of Chemical Physics*, 56(1):248–255, 1972.

[96] E.L. Ratkova, D.S. Palmer, and M.V. Fedorov. Solvation Thermodynamics of Organic Molecules by the Molecular Integral Equation Theory: Approaching Chemical Accuracy. *Chem. Rev.*, 115:6312–6356, 2015.

[97] John F. Atkins Raymond F. Gesteland, Thomas Cech, editor. *The RNA World: The Nature of Modern RNA Suggests a Prebiotic RNA World*. Cold Spring Harbor Laboratory Press, 3rd edition edition, 2006.

[98] Randy J. Read, Paul D. Adams, W. Bryan Arendall III, Axel T. Brunger, Paul Emsley, Robbie P. Joosten, Gerard J. Kleywegt, Eugene B. Krissinel, Thomas LÃŒtteke, Zbyszek Otwinowski, Anastassis Perrakis, Jane S. Richardson, William H. Sheffler, Janet L. Smith, Ian J. Tickle, Gert Vriend, and Peter H. Zwart. A new generation of crystallographic validation tools for the protein data bank. *Structure*, 19(10):1395–1412, October 2011.

[99] Alexander Rich and UL RajBhandary. Transfer rna: molecular structure, sequence, and properties. *Annual review of biochemistry*, 45(1):805–860, 1976.

[100] Frederic M Richards. The interpretation of protein structures: total volume, group volume distributions and packing density. *Journal of molecular biology*, 82(1):1–14, 1974.

[101] Jane S Richardson, Bohdan Schneider, Laura W Murray, Gary J Kapral, Robert M Immormino, Jeffrey J Headd, David C Richardson, Daniela Ham, Eli Hershkovits, Loren Dean Williams, et al. Rna backbone: consensus all-angle conformers and modular string nomenclature (an rna ontology consortium contribution). *RNA*, 14(3):465–481, 2008.

[102] Michael P Robertson and William G Scott. The structural basis of ribozyme-catalyzed rna assembly. *Science*, 315(5818):1549–1553, 2007.

[103] D.R. Roe, A. Okur, L. Wickstrom, V. Hornak, and C. Simmerling. Secondary Structure Bias in Generalized Born Solvent Models: Comparison of Conformational Ensembles and Free Energy of Solvent Polarization from Explicit and Implicit Solvation. *J. Phys. Chem. B*, 111:1846–1857, 2007.

[104] Kersten T Schroeder, Peter Daldrop, and David MJ Lilley. Rna tertiary interactions in a riboswitch stabilize the structure of a kink turn. *Structure*, 19(9):1233–1240, 2011.

[105] Sherwin J Singer and David Chandler. Free energy functions in the extended rism approximation. *Molecular physics*, 55(3):621–625, 1985.

[106] I. Son, Y. Lai Shek, D.N. Dubins, and T.V. Chalikian. Hydration Changes Accompanying Helix-to-Coil DNA Transitions. *J. Am. Chem. Soc.*, 136:4040–4047, 2014.

[107] AA Szewczak and PB Moore. The sarcin/ricin loop, a modular rna. *Journal of molecular biology*, 247(1):81–98, 1995.

[108] W.F. Van Gunsteren and M. Karplus. Protein dynamics in solution and in a crystalline environment. *Biochemistry*, 21:2259–2274, 1982.

[109] Eric Westhof. Water: an integral part of nucleic acid structure. *Annual review of biophysics and biophysical chemistry*, 17(1):125–144, 1988.

[110] Eric Westhof. Structural water bridges in nucleic acids. In *Water and biological macromolecules*, pages 226–243. Springer, 1993.

[111] Christopher J Williams, Jeffrey J Headd, Nigel W Moriarty, Michael G Prisant, Lizbeth L Videau, Lindsay N Deis, Vishal Verma, Daniel A Keedy, Bradley J Hintze, Vincent B Chen, et al. Molprobity: More and better reference data for improved all-atom structure validation. *Protein Science*, 27(1):293–315, 2018.

[112] Bojan Zagrovic and Vijay Pande. Solvent viscosity dependence of the folding rate of a small protein: distributed computing study. *Journal of computational chemistry*, 24(12):1432–1436, 2003.

[113] M. Zgarbova, M. Otyepka, J. Sponer, A. Mladek, P. Banas, T.E. Cheatham, and P. Jurecka. Refinement of the Cornell et al. nucleic acids force field based on reference quantum chemical calculations of glycosidic torsion profiles. *J. Chem. Theory Comput.*, 7:2886–2902, 2011.