Molecular Dynamics of Crystals

by

PAUL A. JANOWSKI

A dissertation submitted to the

Graduate School of New Brunswick

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

and

Graduate Program in Computational Biology and Molecular Biophysics

Written under the direction of

Prof. David A. Case

and

Prof. Darrin M. York

and approved by

_____

_____

_____

_____

New Brunswick, New Jersey

October, 2015

ABSTRACT OF THE DISSERTATION

MOLECULAR DYNAMICS OF CRYSTALS

By Paul A. Janowski

Dissertation Directors

David A. Case Ph.D and Darrin M. York Ph.D

We present a broad effort at the development of crystal simulation methodology and its application to benefit both macromolecular crystallography and molecular dynamics methods. Crystallography is the current method of choice for structural determination of biomolecules, but it is hampered by the inherently time and space-averaged nature of the experiment as well as methodological limitations that do not sufficiently account for the heterogeneous and dynamic nature of crystals. Molecular dynamics has proven itself as a method capable of probing the physics and chemistry of biomolecules on an atomic scale, but requires continued development of the underlying force field parameters to more accurately reproduce observables. Our effort has focused on developing the framework for molecular dynamics simulations of biomolecular crystals. We first present our methodology for performing crystal simulations and show how it is applied first to simple peptide crystals and then to increasingly complex biomolecular systems. We demonstrate the utility of crystal simulations for validation of molecular dynamics. Then we show the improvement to crystallographic methods that can be gained by incorporating molecular dynamics methods. Our work is of significant benefit to both the molecular dynamics and macromolecular crystallography communities and proposes specific approaches to integrate the two fields for the benefit of both.

Acknowledgements

I thank the Lord for the beautiful gift of His Creation, for the flowers, the mountains, the creatures and the fascinating molecular mechanisms of life that testify to His love. Thank you for the gift of reason and free will and for the desire of what is Good, True and Beautiful that drives me in all my pursuits. May we one day attain the fullness of Love.

I thank Maria who accompanied me throughout this path in person and in spirit. Thank you for your care, your patience, your inspiration. We have truly walked the past five years together in friendship. May we walk many more!

I thank my parents, Jolanta and Andrew, for their love in raising me to be the person I am. Thank you for teaching me what is important in life and what is not. I am forever indebted in the bond of filial love.

I thank my family and my friends for all their kindness and love, for all the good times spent together that will forever form part of the treasure of my memories. Without those wonderful times shared together, I would have been hard pressed to keep my sanity along the way.

I thank all my colleagues, my lab mates, my teachers and all the wonderful scientists I have had the pleasure to meet. Especially Prof. Karsten Krogh-Jespersen and Prof. Krzysztof Lewiński for transmitting the beauty of science through their lectures. Thank you for showing me how exciting science is!

I thank all of the good people I have met in life. Every encounter with each one of you has enriched me immeasurably. Only in relationship with others does man become truly man. I love you all.

I thank the excellent scientists I have had the pleasure and privilege of collaborating with over the last five years. I am fortunate to call many of you my friends. I have learned so much from all of you! Thank you for all of your support and encouragement. In no particular order: Bojan Zagrovic, Jurek Dobrucki, David Cerutti, Nigel Moriarty, James Holton, Chunmei Liu, Paul Adams,

Nathaniel Echols, Pavel Afonine, Thomas Terwilleger, Jane Richardson, Greg Warren, Brian Kelley, Timothy Giese, Jason Swails.

Finally, and in a special way, I thank my advisors, Prof. David Case and Prof. Darrin York.. You have challenged and inspired me constantly and at the same time have been caring guides along the path. I have learned from you not just about science but about what it means to be good persons. I remain forever grateful and in your debt!

To the loving memory of my Father who would have wanted to be here but is even closer than we can imagine.

Contents

Abstract of the dissertation ii

Acknowledgements iii

Contents v

List of Figures vi

List of Tables viii

Abbreviations used ix

Section I. Introduction 1

Chapter 1. Introduction and background 1

Section II. Developing molecular dynamics of crystals. 23

Chapter 2. Peptide Crystal Simulations Reveal Hidden Dynamics 23

Chapter 3. Improving Model Interpretation through Crystallographic Refinement and Molecular Dynamics Simulation 52

Section III. Applying Molecular Dynamics of Crystals to Proteins and Nucleic Acids 63

Chapter 4. Molecular Dynamics Simulation of Triclinic Lysozyme in a Crystal Lattice 63

Chapter 5. All-atom crystal simulations of DNA and RNA duplexes 91

Section IV. Improved crystallographic methods through crystal molecular dynamics 115

Chapter 6. Improved ligand geometries in crystallographic refinement using AFITT in Phenix 115

Chapter 7. Implementing molecular dynamics for improved crystallographic model refinement with Phenix and Amber 127

Bibliography 143

List of Figures

Figure 2-1: Three views of the simulated fav8 crystal lattice..... 27

Figure 2-2. Positional RMSDs of heavy atoms relative to the X-ray structure. 32

Figure 2-3: Superposition of the average simulated structure..... 34

Figure 2-4: Volume of the supercell..... 37

Figure 2-5. Left-hand plot: Comparison of computed atomic B-factors..... 39

Figure 2-6. Mean square displacements (MSD) of water molecules..... 41

Figure 2-7. Water densities in the channels..... 42

Figure 2-8. Water density observed in the 2.4 μs simulation 43

Figure 2-9. Mean residence times for each occurring water state..... 45

Figure 2-10. Correlation, as a function of measurement time..... 45

Figure 2-11. Experimental electron density of the Val B8 side chain..... 46

Figure 3-1. Atomic coordinate backbone RMSD 56

Figure 3-2. "Lattice" isotropic B-factors from each simulation 58

Figure 3-3. R-factor..... 59

Figure 4-1. Simulation setup of the HEWL supercell..... 65

Figure 4-2. RMSD for four different force field simulations..... 66

Figure 4-3. Comparison of secondary structure elements..... 70

Figure 4-4. Best-fit (top) and lattice (middle) and refined (bottom) Cα carbon RMSF..... 71

Figure 4-5. The averaged lattice fluctuations from each individual monomer..... 72

Figure 4-6. χ_1 angle side chain disorder..... 75

Figure 4-7. ASU center of mass movement..... 78

Figure 5-1. Two three dimension views..... 94

Figure 5-2. Positional RMSDs of all heavy atoms 96

Figure 5-3. Superpositions of the solution average structure..... 97

Figure 5-4. Plots of major-groove width for RNA 98

Figure 5-5. Conformational substates (BII) probability 100

Figure 5-6. RNA base pair step parameters 101

Figure 5-7. DNA base pair step parameters 102

Figure 5-8. Root-mean-square fluctuations 104

Figure 5-9. Lattice contacts 106

Figure 5-10. The average distances of O3'/A14–O2'/U4 and O2/U15–O2'/A25 107

Figure 5-11. Distances of center of mass between the interface residues 108

Figure 5-12. Same as Fig. 9 109

Figure 5-13. Cartoon view of the 1D23 crystal structure 111

Figure 5-14. Distances between the centers of mass 112

Figure 5-15. The center of mass position for each of the 32 duplexes 113

Figure 6-1. Ligand conformational energies from PDB-deposited models 121

Figure 6-2. Mogul validation 122

Figure 6-3. R-free distributions after refining the test set 123

Figure 6-4. Comparison of 8 randomly selected PDB structures 124

Figure 6-5. Comparison of 5 PDB structures containing ligand instances 124

Figure 6-6. Difference in run time 125

Figure 7-1. Phenix-Amber refinement improves model quality 131

Figure 7-2. Phenix-Amber refinement improves modelling of electrostatics 132

Figure 7-3. Phenix-Amber refinement improves model quality 134

Figure 7-4. Phenix-Amber reduces model overfitting 136

Figure 7-5: Left hand: violin plot distributions 139

List of Tables

Table 2-1. RMSD Values between Various Structures. 36

Table 3-1. Summary of performed simulations. 54

Table 3-2. Summary of structural and fluctuation characteristics. 56

Table 4-1. Molecular composition and basic statistics of the simulated systems. 65

Table 4-2. Average structure and average electron density statistics 68

Table 4-3. Interface behavior relative to deposited model 80

Table 5-1. Details of the simulations. 94

Table 5-2. Root mean square deviations (Å) from the deposited crystal structures 97

Table 5-3. Average α and γ angles (degrees) in crystal and solution simulations. 99

Table 5-4. The average twist (in degrees) for the RNA and DNA average structures 103

Table 5-5. Hydrogen bonds, van der Waals contacts and interactions 106

Table 5-6. van der Waals contacts and interactions between symmetry-related helices 110

Table 7-1: Summary of improvement obtained when refining 130

Abbreviations used

Å – Ångstrom;

AMBER – assisted model building with energy refinement

AMOEBA – atomic multipole optimizd energetics for biomolecular applications

BX – Biomolecular crystallography;

CIF – crystallographic information file;

fs – femtosecond;

FT – Fourier transform;

MD – Molecular dynamics;

μs – microsecond;

MMFF – Merck molecular mechanics force field;

ms – millisecond;

ns – nanosecond;

NVE – microcanonical ensemble;

NVT – canonical ensemble;

NPT – isobaric-isothermal ensemble;

PBC – periodic boundary conditions;

PDB – Protein Data Bank;

PME – particle mesh Ewald;

ps – picosecond;

Section I. Introduction

Chapter 1. *Introduction and background*

I distinctly remember my excitement when, during my initial visit to Rutgers University, Prof. David Case first mentioned the idea of applying molecular dynamics to simulate crystals and improve crystallographic methods. I had studied crystallography for two semesters in the course of my undergraduate degree in biophysics at Jagiellonian University in Krakow. The subject matter was eloquently taught by Prof. Krzysztof Lewiński, one of the best lecturers whose classes I've ever had the fortune of attending. I thank him for instilling in me a deep appreciation for x-ray crystallography. But he also succeeded in shrouding the topic in a sense of wonder and mystery: despite all my efforts at the time I was not able to grasp the essence of how a seemingly random arrangement of dots of varying intensity could be converted into the fascinating three dimensional mesh of electron density. I liked crystallography, but I also respected it and I feared it's mystery that had left me so stumped at university. Thus when Dr. Case presented the idea of molecular dynamics on crystals, I was excited: I could carry out my doctoral research on the development of molecular dynamics methods as I had wanted and I could at the same time get a second chance at figuring out this crystallography business.

I have thus very happily spent the last five years focused on our effort to simulate biomolecular crystals with molecular dynamics. The original question we asked ourselves was simple: what can be learned from molecular dynamics of crystals? This was quickly reformulated into the following four overarching questions that form the focus of this work:

1. What is the best way to carry out molecular dynamics of biomolecular crystals?
2. How can we use crystal simulations to improve molecular dynamics methods?
3. How can we use crystal simulations to improve crystallography methods?
4. What can we learn about real crystals from our simulations of crystals?

What follows is a brief introduction to the methods of crystallography and molecular dynamics, with special emphasis on aspects that relate directly to our work. We then discuss the goals and specific aims of this research and present the general organization of the dissertation before moving on to a presentation of the work in subsequent chapters.

1.1. Crystallography background[1]

Crystallography is a biophysical technique used to probe the three-dimensional atomic structure of molecules by analyzing the diffraction pattern of electromagnetic radiation on a crystal.[1]–[3] As the name implies, crystallography requires that billions of copies of the molecule of study arrange themselves in a regular repeating array which is, by definition, a crystal. When used to study the structure of biomolecules using x-rays, the method is referred to as macromolecular x-ray crystallography (MX). The fact that protein molecules can form crystals has been known for almost 150 years.[4] In general, crystal formation of biomolecules is promoted by slowly removing solvent from a solution of the protein of study. If the solvent is removed too quickly or if the solution is not of the required purity, the protein molecules will precipitate out of the solution and form an amorphous powder. However if the solution becomes supersaturated slowly the molecules may pack themselves in a regularly repeating array held together by non-covalent chemical interactions in a way that minimizes the overall energy of the solute. Finding the exact conditions under which a given biomolecule crystallizes can be very challenging and in many cases constitutes the crux of the crystallographic method.

Once crystallized, the regularly repeating array of the crystal acts as a diffraction grating when light is shined upon it. Diffraction refers in general to the physical behavior of waves as they impact

[1] There are many excellent books on the subject of crystallography. For the interested reader I particularly recommend:

- Rupp B. Biomolecular Crystallography. New York: Garland Science, Taylor & Francis Group; 2010.
- Blow D. Outline of Crystallography for Biologists. New York: Oxford University Press; 2002
- Hammond C. The Basics of Crystallography and Diffraction. New York: Oxford University Press; 1997.

objects or slits. Etimologically, the term was coined by Francesco Maria Grimialdi in 1665 in his *Physico mathesis de lumine, coloribus, et iride, aliisque annexis libri duo* and comes from the Latin diffringere meaning "to break up into pieces". In particular a regularly spaced array of slits or objects will cause the waves scattered off each object to interfere with each other. Wave crests lining up lead to constructive interference resulting in waves of higher amplitude, whereas when crests and troughs mix, destructive interference results in low amplitudes. Because of the dual nature of electromagnetic radiation, when light shines on diffraction grating it behaves like a wave and interference leads to the formation of bands (in the case of a one-dimensional diffraction grating) or spots (in the case of a two-dimensional diffraction grating). James Gregory's observation of the diffraction pattern of light shining through a bird feather in the late 17th century is regarded as the discovery of the first diffraction grating.

A crystal is a repeating array of objects and thus can naturally act as a diffraction grating. However, because the wavelength of visible light is much larger than the typical spacing between array planes in molecular crystals, the diffraction of light on molecular crystals is not observed. The breakthrough moment for x-ray crystallography came in 1912 during a conversation between Paul Peter Ewald and Max van Laue, when van Laue suggested that x-rays (discovered in 1895 by Wilhelm Roentgen) might have a shorter wavelength that would allow their diffraction on crystals to be observed.[5], [6] In 1912 van Laue recorded the first ever x-ray diffraction pattern on a copper sulfate crystal. Shortly thereafter the father-son pair of William Lawrence Bragg and William Henry Bragg formulated the law that describes the diffraction of x-rays on a crystal.[7], [8] The first diffraction pattern from a protein crystal was obtained by John Desmond Bernal and Dorothy Hodgkin using pepsin, and the first three-dimensional structure of a protein molecule solved using x-ray crystallography was myoglobin in 1958 by John Kendrew.[9] Van Laue, the Braggs and Kendrew all received Nobel Prizes for their work. In all thirteen Nobel Prizes have been award for work on or using crystallography.

The raw experimental data obtained in a crystallography experiment is a diffraction pattern. This pattern is obtained as a beam of x-rays is focused on a crystal and the x-ray photons scatter (diffract) off the electron clouds of the atoms that make up the crystal. For the work presented here it is crucial to understand that the diffraction pattern is not obtained in a single instant from single x-rays scattering off the crystal. Rather it is obtained over a significant period of time usually ranging from a few up to about several dozen minutes. The diffraction spots themselves require the constructive interference of an enormous number of x-rays to be observed. Furthermore the x-rays themselves diffract off the billion of molecules that make up the crystal. Thus crystallography is a time and space-averaged experiment.

The diffraction patterned obtained in the crystallography experiment contains two essential pieces of information. The first of these is the location and spacing of the diffraction spots. The spots appear on those vertices of an array called the reciprocal space lattice that intersect the Ewald sphere. The reciprocal space lattice and the Ewald sphere are mathematical constructs directly related to the parameters of the real space lattice of the crystal being studied. The appearance of diffraction spots can be described via the equation that is know as Bragg's Law (named for the younger of the two Braggs mentioned above). Essentially, diffraction spots can only form in locations where the x-rays arrive in phase (in more simple language, where the crests and troughs of the arriving x-ray waves are lined up with each other). The condition for this to happen is that the distances of the paths that all the arriving x-rays travel must all differ by an integral number of wavelengths of the x-rays. The resulting description of the formation of diffraction spots is Bragg's Law:

$$2d\ sin\theta = n\lambda \qquad \textbf{(Eq. 1)}$$

where d is the spacing between a given set of planes in the array, θ is the angle at which the x-rays impact the set of planes, *n* is a positive integer known as the order of reflection and λ is the wavelength of the x-rays. The lattice (spacing between planes) and wavelength are constant under normal experimental conditions. Thus they uniquely specify the angle at which the scattered x-rays

interact constructively and form a diffraction spot. Each spot thus corresponds uniquely to a specific set of planes in the array. Furthermore the angle is inversely proportional to the spacing. In other words smaller diffraction angles correspond to larger plane spacing in the lattice. Diffraction spots closer to the center of the diffraction pattern carry information about larger-scale features of the crystal. This is the basis for the concept of resolution: usually the diffraction pattern can only measured up to a certain radius away from the center: beyond that the angle of diffraction is too large and the spots too weak to be reliably recorded. Most importantly, by accurately measuring the location and spacing of the diffraction spots, one can deduce the spacing of the crystal's array and thus obtain the parameters of the crystal unit cell (the three box dimensions *a*, *b*, *c* and three box angles α,β,γ).

The other essential information in the diffraction pattern are the intensities of the diffraction spots. Whereas the location of the spots reveals the unit cell parameters of the crystal array, the intensities of the spots tell us about the actual distribution of scattering objects, i.e. atoms, within the crystal unit cell. The intensity of a wave (and thereby of the diffraction spot) is equal to the square of the wave's amplitude:

$$I = A^2 \qquad \textbf{(Eq. 2)}$$

We know from the previous discussion and Bragg's Law that an identical scattering object located at each lattice plane of a certain spacing *d* would produce an ideal constructive interference between x-rays and consequently a diffraction spot at the position corresponding to angle θ. But what happens if there is a second scattering objects located between the planes? The x-rays scattering off the object between the planes will arrive at the diffraction spot with a phase different from that of the rays scattering from the object on the plan. The resulting amplitude of the total x-ray wave arriving at the diffraction spot location will result from the sum of two waves: one resulting from the constructive interference caused by all the copies of the first object lying on the plane and the second resulting from the constructive interference caused by all the copies of the second object lying between planes.

Because the objects do not all lie integral distances of the scattering plane away from each other, the resulting waves that are summed are not all perfectly in phase. This results in an attenuation of the amplitude of the resulting wave or even in a complete disappearance of the diffraction spot.

If we treat the scattering electron density in the unit cell as continuous and divide it into infinitesimal sections *dx* along the scattering vector, the amplitude of the resulting diffraction spot can be obtained by integrating the partial x-ray wave scattered by each section *dx* of electron density:

$$\boldsymbol{F_h} = |F_h| \exp(i\varphi_h) = \int_0^1 \rho(x) \exp(2\pi i h x)\, dx \qquad \textbf{(Eq. 3)}$$

(Eq. 3) is presented for the one dimensional case but the generalization to three dimensions is straightforward. Each partial wave has an amplitude proportional to the electron density at *x* but with a phase relative to *x=0* of *2πhx*. The integration is performed over the unit cell vector *h* and the position *x* are described in fractional coordinates. ϱ(x) is the electron density at position *x*. $\boldsymbol{F_h}$ is called the structure factor and is a wave described by an amplitude $|F_h|$ and a phase φ_h. The intensity of the diffraction spot is related to the structure factor amplitude via (Eq. 2). Mathematically this equation is equal to the operation known as the Fourier Transform (FT). If the electron density is presented discontinuously as a set of scattering points (atoms) we obtain the discreet form of the structure factor equation:

$$\boldsymbol{F_h} = |F_h| \exp(i\varphi_h) = \sum_i^{atoms} f(i,h) \exp(2\pi i \boldsymbol{h} \cdot \boldsymbol{x_i})\, \boldsymbol{dx} \qquad \textbf{(Eq. 4)}$$

which we have now presented in three dimensions. *f* is the scattering contribution of atom *i* in the scattering direction corresponding to reflection ***h***. Conversely if we sum over each one of the diffracted waves (at each diffraction spot), we obtained the scattering electron density:

$$\rho(x) = \sum_h F_h \exp(-2\pi i h x) = \sum_h |F_h| \exp(-2\pi i h x + i\varphi_h) \qquad \textbf{(Eq. 5)}$$

Here again we present the one-dimensional form for pedagogical purposes. The summation is over all the diffraction spots of order *h*. This equation corresponds mathematically to the inverse Fourier Transform and is the inverse of the (Eq. 3). Thus we arrive at one of the fundamental concepts of x-ray crystallography: the electron density of the crystal unit cell is the inverse FT of the diffraction pattern.

Now let us examine what is needed to calculate the electron density. (Eq. 5) states that we need to perform a summation over each diffraction spot. For each spot we need the amplitude and phase of its corresponding structure factor. The amplitude is readily obtained as the square root of the intensity measured in the experiment, but unfortunately there is no information about the phase. This is known as the phase problem. Many ingenious methods exist to tackle the phase problem. However, assuming that a sufficiently good estimate of the phases is obtained from which a sufficiently good estimate of the electron density can be calculated, one can move on to the next part of the process, called refinement, that is of greater relevance in the context of the present work. In practice, the great majority of biomolecular structures are solved today by a technique called molecular replacement where a sufficiently good initial estimate of the electron density and phases is obtained by comparison to another similar molecule whose structure is already known.

Supposing that a fairly good estimate of the structure of the molecule has been obtained, one can move on to the next stage in the crystallography process which is called refinement (structural refinement, crystallographic refinement). Let us summarize what information we have at this stage. From the experiment we have the amplitudes of all the structure factors. If we also had the phases we would be able to calculate the electron density directly by (Eq. 5), but we usually don't have the phases. On the other hand we have an estimated structure of the molecule. This is referred to as a model and usually consists of a position relative to the crystal unit cell for each atom that we know makes up the molecule we are studying. From these atomic positions we can calculate the overall electron density of the model (the electron density is calculated by a mathematical function, usually a

sum of Gaussians, related to the number of electrons in the type of atom. The functions most commonly used today are the Cromer-Mann Gaussian functions.[10]) Then from the electron density of the model we can calculate amplitudes and phases via (Eq. 3). We now have a set of experimentally measured structure factor amplitudes which are commonly referred to as F_{obs}, and we have a set of structure factor amplitudes calculated from the current best estimate model of the molecule, which are commonly referred to as F_{calc}. We can now quantify how well the proposed model accounts for the experimental data (or alternatively, how well the experimental data describes the proposed model) by comparing F_{calc} to F_{obs}. This is usually done by a statistic known as the R-factor:

$$\boldsymbol{R} = \frac{\sum_{hkl}\left||F_{obs}| - |F_{calc}|\right|}{\sum_{hkl}|F_{obs}|} \qquad \textbf{(Eq. 6)}$$

where $|F_{obs}|$ and $|F_{calc}|$ are the amplitudes of the F_{obs} and F_{calc} set of structure factors respectively. We can envision the following process: given a starting model we calculate the R-factor. We can use the phases obtained from the model via (Eq. 3) together with the experimental amplitudes F_{obs} to calculate an electron density via (Eq. 5). Next we adjust the atomic positions of our model to better fit the electron density calculated from $F_{obs.}$ From there we calculate a new set of F_{calc} and a new R-factor. If the R-factor is better (lower) than the previous one, than the new model is better than the previous model. This iterative process of calculating the electron density using phases from the model and adjusting atomic positions of the model to fit the resulting density is called refinement.

Refinement is a complex and one could carry out the process just described by hand for a very long time and not obtain any significant improvement. Fortunately refinement can be formulated mathematically as a non-linear optimization problem and solved via one of many known mathematical algorithms. In the most basic formulation a least squares residual between the observed and calculated structure factor amplitudes is minimized:

$$\boldsymbol{E} = \sum_{hkl} (|F_{obs}| - |F_{calc}|)^2 \qquad \textbf{(Eq. 7)}$$

However, in practice this problem is often not well-defined because of the low ratio between the observed data (the set of structure factor amplitudes) and the parameters to be estimated via the optimization (the set of x,y,z coordinates of all the atoms in the asymmetric unit of the crystal) combined with the various sources of noise and error inherent in the x-ray diffraction experiment. Therefore several approaches exist to increase the data to parameter ratio. For example one can decrease the number of parameters to be refined by ignoring some set of atomic coordinates such as the hydrogens. Alternatively, one can increase the set of "observed" data by incorporating previous knowledge about the structure of molecules into the equation. For example, we know that an sp3 carbon-carbon bond should have a length of 1.54Å. This knowledge imposes a set of restraints on the final solution set of atomic positions in the molecule. Thus the residual to be minimized becomes:

$$\boldsymbol{E} = wE_{x-ray} + E_{chem} = w\sum_{hkl} (|F_{obs}| - |F_{calc}|)^2 + \sum_{restraints} (|r_0 - r_{calc}|)^2 \qquad \textbf{(Eq. 8)}$$

Here the x-ray term E_{x-ray} corresponds to the same residual as in (Eq. 7). The chemistry term E_{chem} (also sometimes called stereochemistry or geometry term) corresponds to the summed residual over all the restraints where r_0 is the target value of the restraint and r_{calc} is the value of the restraint in the proposed model. w is a relative scaling weight that is adjusted in the refinement procedure to adjust the relative weight between the x-ray and the restraint term. The restraints used can be obtained from a variety of previously known information about the chemical structure of molecules but most commonly include knowledge of bond lengths, angles and torsions. The most popular crystal refinement programs in use today apply a set known as the Engh & Huber restraints

which were derive from survey of accurate small molecule crystal structures from the Cambridge Crystallographic Database.[11], [12]

In practice several additional levels of complexity are present in modern refinement programs. First, the least squares formulation of the residual to be minimized is most often replaced with a maximum likelihood formulation. This allows for a statistical treatment of observation and restraint probabilities. The chemistry restraints can be incorporated as *a priori* knowledge in a Bayesian formulation. Statistical probability estimates can then be obtained on the resulting parameters. Furthermore, by incorporating this statistical knowledge a large degree of the model bias present in the calculated electron density maps due to the use of phases obtained from the model can be removed.[13] Second, sophisticated mathematical algorithms such as the Limited memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS) algorithm are implemented to optimize the residual based on gradients of its component terms. Third the equation for calculating the structure factors from the model is often more complex than the integral shown in (Eq. 3) as it includes contributions from overall anisotropy and fluctuations and from the contribution of the unmodeled bulk solvent atoms that don't show up distinctly in the experimental electron density. Lastly, crystallographic refinement usually proceeds in stages where the refinement of the x,y,z positions of the atoms in the asymmetric unit is just one stage. Other parameters that affect the calculated structure factors are refined in the other stages. Arguably the most important of these are the B-factors.[14], [15] Where the x,y,z coordinates describe the mean positions of the atom in the structure, B-factors describe how that atoms' instantaneous position fluctuates around that mean. A significant portion of that oscillation can be ascribed to thermal fluctuations. Thus B-factors are often also referred to as temperature factors. B-factors can be isotropic (describing a spherical isotropic fluctuation around the mean position and leading to a single additional parameter to be refined per atom) or anisotropic (describing a three dimensional elliptical oscillation, requiring a symmetric 3x3 tensor and thus 6 additional parameters to be refined per atom). In real space B-factors act like a convolution of a

Gaussian function with the electron cloud around the mean position of an atom, effectively smearing out that atom's electron density. The equation for the FT of the electron density thus becomes:

$$\boldsymbol{F_h} = |F_h| \exp(i\varphi_h) = \sum_{i}^{atoms} f(i,h) \exp(2\pi i \boldsymbol{h} \cdot \boldsymbol{x_i}) \exp(-B_i(\sin\theta/\lambda)^2) \quad \textbf{(Eq. 9)}$$

Importantly the isotropic B-factor is related to the physical mean displacement $\langle\mu\rangle$ of the atom around its mean position by the following equation:

$$\mathrm{B} = \frac{8}{3}\pi\langle\mu\rangle^2 \quad \textbf{(Eq. 10)}$$

Other stages within the refinement process include refinement of bulk solvent contribution, overall anisotropic scaling parameters, atomic occupancies and alternate conformations, rigid body motion and translation-libration-screw (TLS) parameters. A full completion of each of the stages of refinement is usually referred to as a macrocycle. A complete solution of a crystallographic structure usually requires many macro-cycles of refinement interspersed with stages of manual adjustment of the structure to better fit the electron density.

The end result of refinement and of the crystallography experiment in general is a complete three dimensional structure of the atoms in the molecule as it is found in the crystal of study. As of July 4th, 2015, there were 110071 biomolecular structures in the Protein Data Bank (PDB)[16] of which 98000 had been solved by x-ray crystallography. This represents 89% of all solved biomolecular atomic structures making x-ray crystallography by far the most important contributor of data to structural biology.

1.2. Molecular Dynamics Background[2]

[2] There are many excellent books on molecular dynamics. For the interested reader I particularly recommend:

- Allen M, Tildesley D. Computer Simulations of Liquids. Oxford University Press; 1989.
- Cramer C. Essentials of Computational Chemistry. John Wiley & Sons, Inc.; 2004.

Molecular dynamics is a computational technique that aims at analyzing the internal dynamics of a physical multi-body system such as a liquid, a gas or a molecule.[17]–[22] The was first developed by B.J. Alder and T.E. Wainright[23] and independently by A. Rahman[24] in the late 1950's and early 1960's. It was originally invented as a method to study hard sphere collisions in statistical physics, but quickly grew in its application to other fields. The first simulation of a protein was a study of bovine pancreatic tripsin inhibitor by McCammon et al. in 1977.[25] Today molecular dynamics simulations are routinely performed for a wide variety of applications ranging from biophycis and chemistry to atmospheric sciences and astrophysics. Molecular dynamics can be used to obtain both a time resolved detailed view of the dynamics of the system as well as to calculate thermodynamic statistical averages over the system of study.

Molecular dynamics models the system of study as a set of balls connected by springs. Dynamics of the system is obtained by applying Newtonian physics. By Newton's second law of motion we have

$$\boldsymbol{F} = m\ddot{\boldsymbol{x}} \qquad \textbf{(Eq. 11)}$$

where $\boldsymbol{F}$ is the force on a body and $\ddot{\boldsymbol{x}}$ is the second derivative of the position which is the acceleration on that body induced by the force. Because acceleration is the first derivative of velocity and the second derivative of position and can be related to the former two by:

$$\dot{\boldsymbol{x}} = \ddot{\boldsymbol{x}}t + \dot{\boldsymbol{x}}_{\boldsymbol{0}} \qquad \textbf{(Eq. 12)}$$

- Frenkel D, Smit B. Understanding Molecular Simulation: from algorithms to applications. San Diego: Academic Press; 2001.

$$\boldsymbol{x} = \frac{\ddot{\boldsymbol{x}}t^2}{2} + \dot{\boldsymbol{x}}_{\boldsymbol{0}}t + \boldsymbol{x}_{\boldsymbol{0}} \qquad \textbf{(Eq. 13)}$$

given initial positions and velocities, one can integrate the acceleration at a given time to obtain new velocities and positions. By (Eq. 11) to obtain the acceleration, one needs the force, but the force is known to be minus the gradient of the potential energy of the system:

$$\boldsymbol{F}(\boldsymbol{x}) = -\nabla U(\boldsymbol{x}) \qquad \textbf{(Eq. 14)}$$

Thus, by calculating the potential energy of a system with respect to the coordinates of the bodies that make up the system, one can take the gradient of the potential energy with respect to a specific body's position to obtain the force acting on that particle. From there an updated set of velocities and coordinates of the body can be obtained by integrating the laws of motion (Eq. 12) and (Eq. 13). By applying this to all bodies in the system at a given time and by iterating the process over subsequent moments in time a "movie-like" trajectory of the dynamics of the system can be obtained.

We now discuss how to calculate the potential energy of the system. The potential energy equation can take many forms depending on the system being studied. In the case of biomolecular systems, the most common molecular dynamics software packages in use today (Amber[26], CHARMM[27], NAMD[28], Gromacs[29]) use a similar potential function. In the case of Amber, which is the program used in this work, the potential energy function has the form

$$U(\boldsymbol{x}) = \sum_{bonds} k(r - r_0)^2 + \sum_{angles} k(\theta - \theta_0)^2 + \sum_{torsions} k(1 - \cos(n\varphi)) + \sum_{\substack{non-bonded \\ pairs\ i,j}} \left(\frac{A_{ij}}{r_{ij}^{12}} + \frac{B_{ij}}{r_{ij}^{6}}\right) + \sum_{\substack{non-bonded \\ pairs\ i,j}} \frac{q_i q_j}{r_{ij}} \qquad \textbf{(Eq. 15)}$$

The terms in the potential energy equation correspond to bond, angle, torsion or dihedral angle, Lennard-Jones or van der Waals interaction and electrostatic interaction energies respectively. The bond term makes it immediately clear why in molecular dynamics the many-body system is treated via a "beads on springs" model: the bond energy is calculated as the square of the deviation of the current bond length r from the ideal or target bond length r_0 times a constant k which is equivalent to Hook's law for the potential energy of a spring displaced from equilibrium. Angles and torsion angles are treated similarly with the torsion term incorporating the trigonometric function to account for a periodicity of at most 2π. The fourth term of the equation accounts for quantum repulsive and dispersive forces, sometimes known as van der Waals forces. These interactions result from the repulsion of electrons from each other as two atoms draw near to each other (why two atoms cannot overlap) and from the relatively weak attraction between atoms due to instantaneous anisotropy in the electrostatically charged electron clouds as two atoms are separated from each other. The mathematical form of this term is known as the Lennard-Jones potential and has been found to model the repulsive/dispersive interactions sufficiently well. The final term accounts for the Coulomb force electrostatic interactions between the charges of individual atoms in the system.

Examination of the potential function reveals what is needed to run a molecular dynamics simulation. First, one requires starting coordinates of the atoms in the system. These are necessary to calculate the distances between atoms pairs as well as bonds, angles and torsions. Sometimes the velocities are also provided but if not they can be assigned from a Boltzmann distribution at a given temperature. Second, one needs to know which atoms are connected by bonds. This allows for the summations over all atoms connected by bonds, angles or torsions as well as all remaining pairs of non-bonded atoms. This information is referred to as the topology of the system. Finally, one requires the parameters that go into the potential energy function. These include the ideal bond lengths, angles and torsion measures, the Lennard-Jones parameters for different types of atoms as well as the electrostatic charges of atoms required to calculate the Coulomb interaction. This

collective set of parameters used to calculate the potential energy function given a set of atomic positions and topology is known as a force field.

The greater bulk of effort at developing and improving the accuracy of molecular dynamics simulations goes into deriving better sets of force field parameters. Force field parameters are derived by fitting simulated properties to calculations obtained through *ab initio* quantum methods or to experimental measurements of thermodynamic or spectroscopic properties. Several force fields exist in the Amber program. The most recent force field as of this writing is the ff14SB Amber force field which is an elaboration of earlier Amber force fields.[30], [31] Other available force fields include Amber ff14ipq[32], CHARMM36[33], OPLS as well as the AMOEBA[34], [35] polarizable force field that allows for changes to atomic partial charges as the simulation proceeds.

This generally simple outline of molecular dynamics is made slightly more complex by a multitude of enhancements mostly aimed at improving the accuracy and/or the computational efficiency of the simulation. First, a straightforward run of molecular dynamics replicates the thermodynamic microcanonical ensemble where the number of particles, volume of the system and total energy of the system are constant (NVE). However by adding computational algorithms to maintain a specified temperature or a specified pressure in the system, the canonical constant particles, volume and temperature (NVT) and the isobaric-isothermal constant particles, pressure and temperature (NPT) ensembles can be enforced. These temperature and pressure monitoring algorithms are called thermostats and barostats and the most common algorithms in use today include the Berendsen[36], Langevin[37] and Monte Carlo algorithms. The NPT ensemble is of particular importance in our work as it allows the system volume to fluctuate under constant temperature and pressure as is the case with a physical crystal in experimental conditions.

Second, there are enhancements aimed at increasing the efficiency of the molecular dynamics calculations. Each cycle of calculating the potential energy and its gradients with respect to atomic positions, integrating the equations of motion and updating the atomic velocities and positions is

called a step. The length of the step can be regulated by specifying the value of t in (Eq. 12) and (Eq. 13). The rules of statistical thermodynamics govern the rate of occurrence of events of interest at the atomic/molecular scale. Some events, such as the rotation of an amino acid side chain around a torsion angle, may occur often, on a nanosecond time scale. Other events, such as the complete folding of a protein require orders of magnitude more time, usually on the microsecond timescale. Therefore, it is desirable in molecular dynamics to simulate a length of time sufficient to be make it probable that the event of interest will occur within the simulated time window. However, one cannot simply make the time step larger: if the time step is larger than the time scale of the fastest events simulated by the force field, the integration of positions and velocities will proceed in leaps without responding in time to the effects of these events and resulting in severe instabilities in the system. Normally the fastest events observable in the simulations are bond length vibrations that occur on a femtosecond scale. Thus the time step employed in a typical simulation is usually 1 or 2 femtoseconds. By constraining the fastest bond vibrations which are those involving hydrogen atom bonds, to constant values using specifically designed Lagrange multiplier based algorithms (for example SETTLE[38], SHAKE[39] and RATTLE[40]), time steps can sometimes be increased up to about 5 femtoseconds.

Molecular dynamics thus proceeds in steps, in which the greatest amount of computational time is spent on calculating the potential energy function. As mentioned, there is a need to simulate time lengths long enough to observe events of interest. This is called the sampling problem in molecular dynamics: when an event of interest is not observed in a simulation one can only speculate whether the event does not happen due to the actual physics and chemistry of the system or if it does happen but we have simply not simulated for a long enough time. The problem is further complicated by inaccuracies in the force fields as well as physical limitations in the accuracy of floating point operations on modern day computers: these small inaccuracies tend to add up as the simulation progresses often leading to instability before the target time scale is reached. The first protein simulation in 1977[25] was 8.8 picoseconds (ps) long. The longest simulations to date have attained

the millisecond time scale[41], but the great majority of simulations performed today range from 10s to 100s of nanoseconds (ns).

Since the great bulk of calculation time is spent on the potential energy function, a number of approaches aim increase the efficiency of this part of the method. For example, the three bonded term calculations can easily be split up between several computer processers and thus calculated in parallel. Because the Lennard-Jones potential decreases very quickly with distance ($1/r^6$), a cut-off can be introduced to only calculate the energy over pairs of atoms that are sufficiently close together. This is a significant savings as the number of pairs of atoms increases as N^2. A number of sophisticated algorithms, especially parallel computing ones, exist for efficiently maintaining and updating the list of particle pairs within the cut-off distance.[27], [42], [43] The same cannot be done for electrostatics, which decreases much more slowly with distance ($1/r$).[44] Fortunately however, when dealing with a periodic system, an algorithm called Ewald summation is able to accurately calculate the electrostatic energy by decomposing the interactions into short-range and long-range terms and calculating the Fourier transformed long-range terms in reciprocal space. In 1993 Thomas Darden and Darrin York devised a method called Particle Mesh Ewald (PME) that is able to calculate the Ewald sum in significantly faster time by spreading the charge density on a three dimensional grid.[45]

Finally, we mention that molecules are very rarely simulated in molecular dynamics *in vacuo*. Not only is this unrealistic as molecules usually are not encountered in nature in solitary confinement, but it also would lead to sever artefacts on account of the high energy of electrostatic interaction between charged moieties within proteins and nucleic acids. Thus the standard approach in molecular dynamics is to immerse the molecule of interest in a water box composed usually of several thousand water molecules surrounding the protein. There are many sets of force field parameters for water of which the most popular ones are spc-e[46], tip3p, tip4p[47] and tip4p-ew[48]. However, even such a system would not be successful as the box of water with protein would itself be located in a vacuum leading all of the waters to fly away from each other into space. Thus an algorithm called periodic

boundary conditions (PBC) is used. This essentially consists of replicating the simulated box out infinitely in space in all directions. In other words copies of the box itself are placed all around it so that when a particle flies out of the box on one side, an identical particle flies in on the other side. This not only surrounds the simulated box with virtual matter preventing an exploding artefact but also essentially creates a periodic system, thus enabling us to use the PME algorithm to calculate electrostatics.

1.3. Goals and outline of reasearch

Both macromolecular x-ray crystallography and molecular dynamics have proven to be extremely valuable methods in the biophysical arsenal. However, both also suffer from several limitations. It is the overarching idea in the present work that molecular dynamics simulations of crystals can contribute to resolving some of these limitations.

Crystallography suffers from several sources of both systematic and random experimental error.[49]–[51] In many cases this error and the innate qualities of the system being studied result in very low resolution structures that preclude the accurate determination of structural details. For example, often crystallographers will be forced to eliminate atoms, side chains or even several-residue long fragments from the final structure. Simulations of crystals, if found to be reliable, would present the possibility of modelling these sections. Information from simulation could be combined with the indeterminate low-resolution information from experiment to accurately resolve these features. Second, by it's very nature, the diffraction experiment is time and space averaged as the diffraction spots of the x-rays result from the summation of x-rays diffracting from the billions of copies of the asymmetric unit in the crystal during exposures of up to several dozen minutes at a time. The general approach to analyzing the data has been traditionally been to find a single static structure that best interprets the experimental data. However, it has been recognized that this approach is limited at best.[52] Biomolecular crystals are not mathematical entities in which the crystal lattice is perfectly maintained and each unit cell is an identical copy of all the other ones. In fact, there is a degree of

heterogeneity and variation within the structures of the molecules from one unit cell to the other and furthermore crystals are dynamic which a rich variety of motion still able to occur within the lattice. Thus a fuller interpretation of the experimental data and a complete structural understanding of a crystal would require a representation of the conformational ensemble represented by the various copies of the molecule in the crystal and of the dynamics being sampled by them. Recent efforts by several groups have aimed at moving the state of the art towards such a more integral understanding of crystals. For example, ensemble refinement has been proposed where the calculated structure factors used in refinement are taken over a best-fit ensemble of structures rather than a single model.[53], [54] Room temperature and cryogenic crystallography of the same crystal have been compared against each other to reveal differences that point at structural heterogeneity and dynamics.[55] Computational algorithms have been developed to find alternate conformations in an automated manner and to discover networks of alternate conformations that point toward large scale conformational changes in the molecules within the crystal.[56]–[58] Also, recent attempts have been made to develop methods aimed at interpreting the diffuse scattering (x-ray scattering outside of the Bragg peaks) than contains information about disorder within the crystal.[59]–[61] Molecular dynamics is well poised to contribute to this field. By simulating multiple independent copies of the crystal unit cell over extended periods of time the dynamics and disorder inherent in the crystal can be studied. Information gained, if judged sufficiently accurate, can then be applied to the experimental data resulting in an ensemble-based interpretation of the experiment that agrees better with experimental results then a single static structure model. Furthermore, crystal simulations can be used to investigate the fundamental physics of crystals such as the non-covalent interactions that hold crystals together or the free-energy barriers of conformational transitions within crystals. Also, in itself the molecular dynamics force fields that have been steadily improving in accuracy and reliability over the years, can be incorporated in crystallography refinement schemes to provide an improved set of priors (chemistry restraints). Lastly, molecular dynamics simulations can be used to create high quality synthetic crystallography data sets. These can be used for testing, validation and

improvement of crystallography methods as they provide simulated experimental data for which the exact solution is known ahead of time, since we have direct access to all measurable quantities from the simulation.

Molecular dynamics can equally well be served by research on simulations of crystals. The standard practice in molecular dynamics is to simulate the molecule of interest in a solvated environment, i.e. surrounded by several thousand water molecules in the simulated box. This makes good sense first because the biomolecule should not be simulated in vacuum for reasons mentioned above and second because the solvated state best replicates the native state of the molecules. Nevertheless, the solvated approach to simulations also presents several drawbacks. First, because the solvated environment is different from the crystalline environment of the experiment, it is not possible to directly compare simulation results to experimental results. As mentioned, one of the greatest obstacles for molecular dynamics simulations is the development of accurate force field parameters. This process is usually accomplished by assessing the quality of simulation results with a set of new force field parameters. Crystal simulations can prove to be very useful in this regard because they allow for direct validation of the molecular dynamics simulations against experimental data. Structural averages and fluctuations can be directly compared against the structures and B-factors of experiment. Moreover, the average electron density and consequently the structure factor amplitudes and intensities can be calculated from a crystal simulation and compared directly against the raw data of the experiment. A second benefit of crystal simulations is related to the sampling problem. Molecular dynamics suffers from the requirement of sufficient computational resources and time to accurately sample the simulated system for the events of interest. Molecular dynamics simulations can help alleviate the problem. In a typical crystal set-up the ratio of protein or nucleic acid atoms to solvent atoms is much greater than in a solvated simulation. The crystal does not require a buffer of solvent to surround the system but is rather made up mostly of the independent copies of the molecule of interest and solvent is only used to simulate the mother liquor found in the interstices of the crystal. Thus in a crystal simulation a relatively greater portion of time is spent on

calculations pertaining to the biomolecule. Plus there are many independent copies of the molecule being simulated at the same time thus further greatly increasing the degree of sampling. In this way crystal simulations can serve to help overcome both the sampling and validation problems in molecular dynamics.

We have thus introduced briefly the two methods of biomolecular crystallography and molecular dynamics in a way that should permit the general understanding of the work that follows. We continue with a specific presentation of the research carried out on the various aspects of molecular dynamics of crystals. Section II is methodological and deals with the development of methods for all-atom crystal molecular dynamics simulations. It investigates how to set up and carry out crystal simulations as well as how to analyze them given the unique qualities of the produced data. Chapters 2 and 3 also demonstrate a proof of concept that information from molecular dynamics crystal simulations can be directly used to enhance our knowledge about crystals and to aid in the interpretation of experimental data. Chapter 2 first appeared in the *Journal of the American Chemical Society* as "Peptide Crystal Simulations Reveal Hidden Dynamics Pawel A. Janowski, David S. Cerutti, James Holton, and David A. Case. Journal of the American Chemical Society 2013 135 (21), 7938-7948."[62] Chapter 3 is being prepared for submission.

In Section III, we apply crystal simulations to larger and more relevant systems. Chapters 4 and 5 examine various aspects of molecular dynamics force field validation and indicate possible further paths for improving force fields based on data obtained from crystal simulations. Chapter 4 first appeared in *Protein Science* as "Molecular Dynamics Simulation of Triclinic Lysozyme in a Crystal Lattice. Pawel A. Janowski, Chumei Liu, Jason Deckman, David A. Case. Protein Science 2015."[63] Chapter 5 first appeared in *Biochimica et Biophysica Acta* as "All-atom crystal simulations of DNA and RNA duplexes. Chunmei Liu, Pawel A. Janowski, David A. Case. Biochimica et Biophysica Acta 2015 1850(5), 1059-1071."[64]

Section IV presents the application of crystal molecular dynamics to improve crystallography methods. In Chapter 6 a molecular dynamics force field is implemented to accurately model protein ligands and small molecules in macromolecular crystals thus leading to chemically more accurate ligand geometries. Chapter 7 presents an integration of the Amber molecular dynamics software package with Phenix software for crystallographic refinement. Incorporation of molecular dynamics of crystals directly in biomolecular crystal refinement leads to improved structural models and better agreement with experimental data. Chapter 6 has been submitted for publication and is currently under review at *Acta Crystallographica D*. Chapter 7 is being prepared for submission.

Section II. Developing molecular dynamics of crystals.

Chapter 2. *Peptide Crystal Simulations Reveal Hidden Dynamics*[3]

2.1. Abstract

Molecular dynamics simulations of biomolecular crystals at atomic resolution have the potential to recover information on dynamics and heterogeneity hidden in X-ray diffraction data. We present here 9.6 μs of dynamics in a small helical peptide crystal with 36 independent copies of the unit cell. The average simulation structure agrees with experiment to within 0.28 Å backbone and 0.42 Å all-atom RMSD; a model refined against the average simulation density agrees with the experimental structure to within 0.20 Å backbone and 0.33 Å all-atom RMSD. The R-factor between the experimental structure factors and those derived from this unrestrained simulation is 23% to 1.0 Å resolution. The B-factors for most heavy atoms agree well with experiment (Pearson correlation of 0.90), but B-factors obtained by refinement against the average simulation density underestimate the coordinate fluctuations in the underlying simulation where the simulation samples alternate conformations. A dynamic flow of water molecules through channels within the crystal lattice is observed, yet the average water density is in remarkable agreement with experiment. A minor population of unit cells is characterized by reduced water content, 3_{10} helical propensity and a gauche(-) side-chain rotamer for one of the valine residues. Careful examination of the experimental data suggests that transitions of the helices are a simulation artifact, although there is indeed evidence for alternate valine conformers and variable water content. This study highlights the potential for crystal simulations to detect dynamics and heterogeneity in experimental diffraction data as well as to validate computational chemistry methods.

2.2. Introduction

[3] Reproduced with permission from P. A. Janowski, D. S. Cerutti, J. M. Holton, and D. A. Case, "Peptide crystal simulations reveal hidden dynamics," *J. Am. Chem. Soc.*, vol. 135, no. 21, pp. 7938–7948, 2013. Copyright 2013 American Chemical Society.

X-ray crystallography has played the essential role in the development of the field of structural biology. In doing so, the conventional focus of biomolecular X-ray crystallography has been on identifying a single structure to represent the molecule that best explains the collected diffraction data. Yet, it is well established that biomolecules, both in solution and in crystal, are highly dynamic objects which populate an ensemble of structurally heterogeneous states.[55] Information on this dynamicity and heterogeneity is "hidden" in the experimental data set which, by its nature, is essentially time and space averaged.[65] In recent years, several attempts have been made to develop methods to mine the experimental data for information on dynamics and structural heterogeneity in the protein.[53], [58] Here we present a further advance in this direction by employing the power of all atom, explicit solvent, molecular dynamics (MD) simulations of crystals to gain a more exact and time-resolved picture of the inner dynamics of a peptide crystal. Crystallographic refinements against the computed average electron density are critically compared against refinements against the experimental density.

The potential of computer simulations to extend our understanding of the motions of biomolecules beyond the experimental images offered by X-ray crystallography or NMR experiments has driven the application of computational techniques to problems in structural biology. It is now feasible to simulate protein systems containing hundreds of residues for microseconds of real time. Commensurate with improved simulation algorithms and computer hardware, the molecular models have been scrutinized for their dynamic, equilibrium thermodynamic, and structural characteristics. In many respects, the models perform realistically,[66]–[69] but by pressing the models to jump from reproducing known results to correctly predicting new data,[70], [71] the models also show signs of overfitting and reduced transferability. Peptide and protein crystals offer a rich set of experimental data and the opportunity to subject molecular models to tests in which the time-averaged positions and fluctuations of atoms are known.

The applicability of simulations to the interpretation and even improvement of X-ray data sets is a goal on the horizon. More immediately, efforts have been focused on tailoring simulations to match

crystallization conditions and devising appropriate analyses to directly compare molecular models with crystallized biomolecules.[72]–[74] Crystallographic data have also been used to validate computational results in many forms.[75] One of the challenges of simulating crystals lies in the necessity to extrapolate the unknown crystal solvent content. It remains a high priority to select systems with as little uncertainty in the crystallization solution as possible. Our previous simulations[76]–[78] were among the longest crystal simulations performed at the time, but even with 8–12 independent copies of the unit cell and hundreds of nanoseconds of simulation, some of the most interesting parameters, such as the persistence of hydrogen bonds and density of material near crystallographic water sites, were not sufficiently converged to determine whether the simulation matched the X-ray data.

In this study we present simulations of the crystallized decapeptide hereafter referred to as "fav8".[79] The sequence of this synthetic peptide favors helix formation and aromatic intermolecular interactions. Furthermore, the crystal is exceptionally dry, with only four waters placed in the experimental electron density, and no room for disordered "bulk" solvent. As we show in the results, the unit cell volume is correctly maintained by including only the four crystal water molecules. The ability to simulate the entire fav8 decapeptide crystal lattice with certainty about its material composition for microseconds enables us to compare the simulation and the X-ray diffraction data in unprecedented detail. We perform several simulations, the longest of which reached 2.4 μs, of an extended fav8 lattice comprising 36 independent unit cells—in all, roughly 10 times the simulation length of our previous simulations with 10 times the number of independent unit cells. The results give a much clearer picture of the time-averaged solvent density, solvent diffusion within the peptide lattice, and hydrogen bonding for maintaining peptide structure.

2.3. Methods

2.3.1. Preparation of the Simulation Supercell

Atomic coordinates were taken from the cif format file in the Supporting Information of the publication that reported the molecule's structure.[79] This is a synthetic decapeptide (sequence Boc-Aib-Ala-Phe-Aib-Phe-Ala-Val-Aib-Ome) designed to fold in a helical conformation with aromatic t-stacking interactions between phenylalanine rings of separate monomers in its crystallized form. In the decapeptide, Aib (α-aminoisobutyryl) is a nonstandard amino acid (alanine modified by methylation of the Cα hydrogen) and Boc (N-tert-butoxycarbonyl) and Ome (O-methyl ester) are terminal blocking groups. The peptide formed crystals in the P1 space group, with one asymmetric unit (ASU) per triclinic unit cell of dimensions a = 10.802, b = 16.361, c = 17.853 Å, α =116.405°, β = 95.535°, and γ = 93.164°. The ASU consists of two nonequivalent decapeptides, referred to as monomer A (residues A1–A10) and monomer B (residues B1–B10) as well as four crystallographic water oxygen positions. The diffraction experiment was carried out at a temperature of 294 K. The major structural features of the unit cell include phenylalanine side chain π- and t-stacking interactions, as discussed in the original publication; four crystallographic water molecules lie within hydrogen-bonding distance of each other and of the N- and C-termini of adjacent decapeptides.

A "supercell" of 4 × 3 × 3 unit cells was created by using the PropPDB module of the Amber11 package,[80] measuring 43.208 × 49.083 × 53.559 Å and comprising 72 copies of the fav8 decapeptide. Views of the supercell along the three crystal vectors are shown in Figure 2-1. Inspection of the supercell shows that crystal packing places the crystallographic waters clusters in interstices, connected to one another with little steric hindrance between adjacent unit cells forming channels along the a vector of the crystal lattice.

In previous all-atom crystal simulations,[76]–[78] solvent that was unaccounted for in the X-ray data was added to the simulation supercell until the experimental volume of the crystal was accurately reproduced by MD at the temperature and pressure of the crystal growth conditions. Furthermore, different species of solvent were added in proportions to mimic the composition of the crystal mother liquor. In the case of fav8, initial equilibration and trial MD production runs reproduced

crystal lattice parameters accurately without any additional solvent. Therefore, we performed all production runs with only the molecules found in the original cif file.

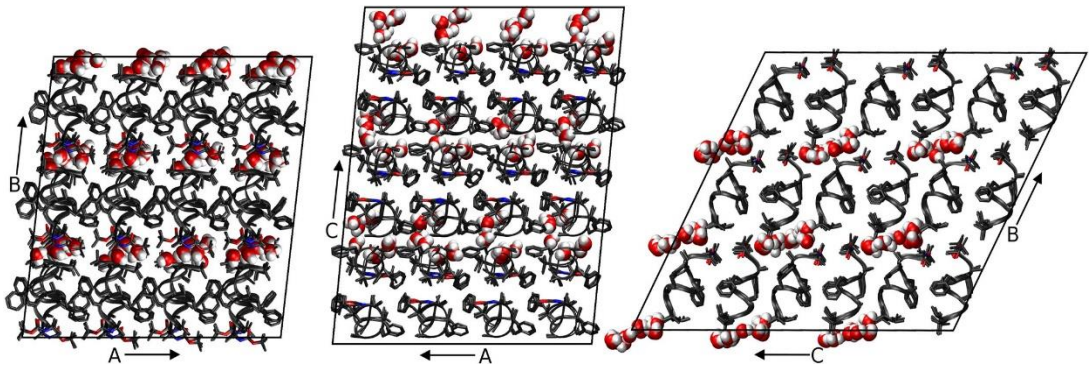

Figure 2-1: Three views of the simulated fav8 crystal lattice. 36 unit cells are stacked in a 4 × 3 × 3 arrangement in the triclinic supersystem; each unit cell comprises two fav8 decapeptide helices arranged roughly parallel to one another. Each view looks down one axis of the lattice, and borders of the simulated system are marked in black lines. The peptide backbone is shown in ribbons or in licorice form in the case of Aib and terminal blocking residues. Water molecules are illustrated in space-filling form; we find that the water forms continuous channels running through the lattice along the a axis.

2.3.2. MD Simulations

Protonation of the peptide structure and construction of molecular topology and coordinate files for the crystal supercell was done using the tleap module of Amber11 and Reduce.[81] The peptide in the simulation supercell was modeled using parameters of the Amber ff99SB force field[30] and the TIP3P water model.[82] The Boc, Aib, and Ome residues are not found in the standard Amber force field, but we obtained charges for these residues using RESP fitting[83] and took other parameters from similar compounds described by ff99SB; details are in the Supporting Information.

System optimization, equilibration, and production dynamics were performed using the PMEMD module of AMBER11. When the system volume was allowed to vary, constant pressure was maintained by a Berendsen barostat[36] with isotropic pressure scaling (at the time this study was conducted, anisotropic scaling was not available in Amber for a triclinic box. This feature has since

been added). Constant temperature was maintained during all dynamics with a Langevin thermostat[37] (collision frequency of 1/ps) at the experimental crystal diffraction temperature of 294 K. To avoid artifacts arising from the reuse of the same random number sequences,[84] a different random number generator seed was used each time a simulation was restarted. Force calculations were performed with a 9.0 Å real space cutoff in the context of periodic boundary conditions, smooth particle-mesh Ewald electrostatics,[45], [85] and a homogeneity assumption for long-range van der Waals contributions. The SHAKE[39] and SETTLE[38] algorithms were used to constrain the lengths of bonds to hydrogen and the internal geometry of rigid water molecules, respectively.

System equilibration was carried out using the following scheme: First, the conformations of peptide residues, including added hydrogens, were relaxed via 100 steps of steepest-descent optimization followed by 900 steps of conjugate gradient optimization with 256 kcal/(mol-Å2) position restraints applied to solvent molecules. Next, the entire system was optimized in the same manner but with no restraints. Initial restrained dynamics were performed at constant volume for 50 ps with a 1.0 fs time step and 256 kcal/(mol-Å2) restraints on all peptide heavy atoms, followed by another 225 ps of restrained dynamics at a 1.5 fs time step during which restraints were gradually reduced to 4.0 kcal/(mol-Å2). Next, restrained dynamics were performed at a pressure of 1 bar for 400 ps using a 2 fs time step as restraints on peptide heavy atoms were gradually relaxed from 4.0 to 0.0625 kcal/(mol-Å2). Unrestrained production dynamics were propagated at a 2 fs time step, matching the final phase of equilibration in which all restraints had been reduced to zero.

Production simulations were carried out on clusters of 48 core 2.2 GHz Opteron CPUs provided by the Rutgers BioMaPS High-Performance Computing facility and also on a private cluster of serial GPUs. A total of 4 simulations were propagated for 1.6–2.4 μs each.

2.3.3. Analysis of Data

Data analysis was carried out using in-house scripts and the Amber11 ptraj module for MD trajectory analysis. Two root-mean-square deviation (RMSD) metrics which we refer to as "ASU RMSD" and "lattice RMSD" were calculated using the Kabsch algorithm.[86], [87] They are described briefly in section 2.4.1, and more details can be found in ref. [78]. Secondary structure was determined using the DSSP[88] algorithm. Experimental electron density maps were calculated from experimental intensities kindly provided by S. Aravinda and P. Balaram, coordinates and anisotropic displacement parameters found in the Supporting Information of Aravinda et al., 2003 by zero-cycle unrestrained maximum likelihood refinement using Refmac.[89] Molecular refinement was performed with Phenix.[90], [91] The Visual Molecular Dynamics (VMD) program[92] and ccp4mg[93] were used for visualization and image generation. Approaches to calculating B-factors are described in section 2.4.1.

To calculate average simulation electron density and structure factors, an evenly spaced selection of 4000 snapshots was taken from the final 2 μs of the longest of our simulation trajectories, amounting to 144 000 conformations of the ASU. Electron density maps were generated directly from each of these conformations using the CCP4 program SFALL.[94], [95] For each map-generation run, all 36 unit cells for the given time point were included in the calculation using a unit cell repeat that was an integral reduction of the simulation cell. For any given time point in the simulation the B-factors of all the atoms are formally zero, but this presents certain problems in calculating electron density because the constant "c" term in the conventional Cromer–Mann reciprocal-space atomic form factor tables[10] becomes a Dirac δ-function in real space. This results in a singularity when plotting the electron density onto a grid for the fast Fourier transform calculation of the structure factors.[96] To avoid this singularity, a B-factor of 15 was assigned to all atoms (large enough to avoid aliasing errors) before calculating the electron density maps. Despite the slightly different cells (due to simulation in the NPT ensemble, see section 2.4.1), all of these maps were calculated to have the same number of grid points: 96 × 108 × 120.

Structure factors were calculated from each of these maps, and the translation needed to optimally superimpose each time point in the simulation onto the published structure was determined by deconvolution in reciprocal space. This was necessary because the "origin" is not restrained and drifts slowly throughout the simulation, so that averaging electron density in real space (or structure factors with phases in reciprocal space) would eventually "blur" itself down to a constant (the average electron density of the crystal), driving all structure factors to zero. Specifically, the complex structure factors calculated from the published atomic coordinates were divided by the complex structure factors obtained from the electron density of the simulation time point. The map calculated from these "quotient" structure factors is the correlation function of the two parent maps, and the tallest peak in this map is located at the optimal translation to "align" them.

After determining these optimal shifts, the atoms from each simulation time point were translated appropriately, and the electron density maps recalculated. The average of all these electron-density maps was then taken, and a final Fourier transform was computed to obtain the expected structure factors of a single crystal mosaic domain comprised of all 144 000 ASUs represented in the trajectory. The CCP4 program CAD was used to remove the contribution of the B-factor = 15 from the structure factors. The R-factor of these simulation structure factors with the observed structure factors was calculated after applying an optimal scale and B-factor with the CCP4 program SCALEIT.[94], [96]

2.4. Results and Discussion

Dynamics of the fav8 peptide crystal lattice were analyzed on the microsecond time scale in a system comprising 36 unit cells stacked 4 × 3 × 3. Simulations were run in quadruplicate (one 2.4 μs trajectory, and three additional 1.6 μs trajectories). The simulated system retained the unit cell angles and aspect ratios of the crystal due to the isotropic pressure rescaling of cell dimensions, but the corresponding atoms in each of the 36 unit cells were otherwise allowed to move independently. In addition to structural comparisons, we computed isotropic B-factors for all peptide heavy atoms and

again found close agreement with the experiment. Finally, we turned our attention to dynamics of water molecules and found them to migrate between different unit cells, indicating that the electron density of water molecules in the fav8 crystal arises from many distinct molecules interchanging positions during the experiment.

2.4.1. Comparison to Experimental Structure

It is less straightforward than one might think to quantify the agreement between a crystal lattice simulation and the refined structure inferred from X-ray diffraction data. Unit cell volume, positional RMSD, average unit cell structure, and thermal vibrations provide a strong set of indicators as to the simulation's accuracy. Positional RMSD was measured in two distinct ways. First, we define "ASU RMSD" as

$$ASU\ RMSD = \sqrt{\frac{1}{M}\sum_{i=1,M}\left[\frac{1}{N}\sum_{j=1,N}\left|r_{i,j} - r^{*}_{i,j}\right|^{2}\right]} \qquad \textbf{(Eq. 16)}$$

where the inner summation runs over N atoms, the outer summation runs over M ASUs, $r_{i,j}$ is the position vector of an atom in the simulation snapshot, $r^{*}_{i,j}$ is the experimentally determined position vector of that atom, and the statistic is calculated after rotational and translational alignment of the backbone heavy atom coordinates in each ASU against the crystal fav8 structure using the Kabsch algorithm.[86] This RMSD, which was computed for backbone and side-chain atoms (with provisions for the symmetry of atoms in Phe rings and the Boc terminus), accounts for all disorder arising from bending and distortion of individual fav8 monomers and disorder arising from changes in the contacts between the pair of monomers that composes each ASU. Second, we compute a "lattice RMSD" which follows the same formula as the ASU RMSD; however in this case ASUs are not aligned in the traditional manner. Instead, ASU's are superimposed by first center of mass aligning each supercell and then reversing the translational space group operations by which the simulation supercell was constructed. The center of mass alignment is necessary due to translational

drift of the origin of the supercell, since its potential energy is translationally invariant. This metric captures rigid-body librations of the peptides in the unit cell and lattice distortion between fav8 monomers in different unit cells, since atoms in different unit cells are not constrained to move in any symmetric fashion. Figure 2-2 plots these RMSD measurements over the course of the 2.4 μs trajectory. If one focuses on a much shorter time scale, the RMSD of both the backbone and of the side-chain atoms appears to converge to 0.5/0.7 Å after as little as 20 ns of dynamics, but Figure 2-2 shows that these metrics rise suddenly at 400 ns to 0.6/0.75 Å, levels which are maintained for the remainder of the simulation. (Convergence of the other three trajectories is illustrated in Figure S1; backbone and side-chain RMSDs in these simulations are comparable to that of the 2.4 μs trajectory.) Also after roughly 400 ns, backbone lattice RMSD converges to about 0.75 Å. RMSD adds in quadrature, and therefore these results indicate that there is an approximately equal contribution to overall RMSD from intra- and intermolecular distortions. All further analyses were performed after discarding the first 400 ns of simulation.

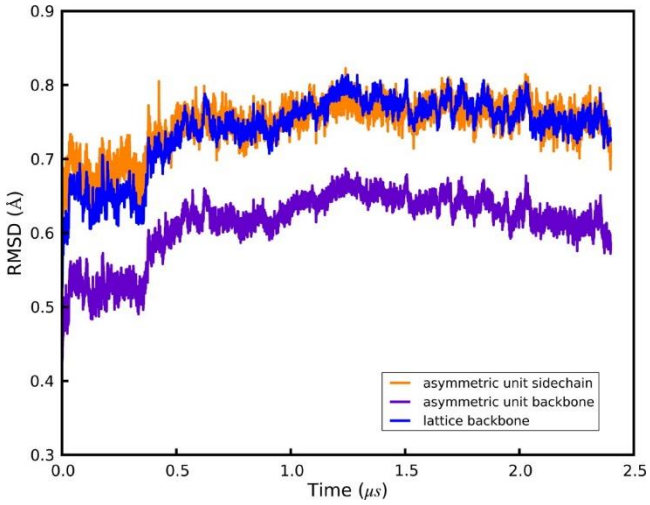

Figure 2-2. Positional RMSDs of heavy atoms relative to the X-ray structure. Details of each metric are given in the main text. All quantities are plotted over the course of a 2.4 μs simulation, and plots for three additional 1.6 μs simulations are given in the Supporting Information. Purple: ASU RMSD for backbone (N,CA,C) atoms. Orange: ASU RMSD for side-chain heavy atoms. Blue: lattice RMSD for backbone atoms.

The crystallographic raw data are a diffraction pattern that is the averaged result over time and over three-dimensional space of the repeating unit cell. To set our analysis in line with the experimental results, we calculated an average structure of the simulated unit cells using the same reverse symmetry operations and Phe/Boc atom equivalencies that had been used to compute lattice and ASU RMSDs. A superposition of the resulting average structure with the X-ray result is shown in Figure 2-3. The RMSD of backbone and side chain heavy atoms for this average structure is 0.32/0.45 Å, which is much lower than the RMSD of the individual snapshots cited above. Thus, structural deviations can occur at instantaneous snapshots of the simulation, while the time-averaged structure maintains close similarity to the X-ray model, as is consistent with a dynamic interpretation of the crystal. In the average structure, monomer A agrees nearly perfectly (0.15/0.17 Å backbone/side chain RMSD) with the refined X-ray structure, and only the C-terminus of monomer B (residues B8–B10) is seen to deviate significantly (residues A1-B7 0.20/0.21 Å, residues A1-B8 0.21/0.38 Å, residues A1-B9 0.29/0.44 Å; indicating disorder in only the side chain of residue B8 and in both backbone and side chain of residues B9/B10). As shown in Figure S3, the deviations in monomer B are in fact confined to a subset of 9 of the 36 unit cells. The average heavy atom RMSD of monomer B in this subset is 0.84 Å, while in the remaining cells it is 0.51 Å (for comparison, the average RMSD of monomer A in all cells is 0.23 Å). Furthermore, if the C-terminus (residues B8–B10) is removed from the calculation, the RMSD of the subset of 9 unit cells drops from 0.84 to 0.63 Å and for the remaining cells from 0.51 to 0.23 Å, identical to the average RMSD of monomer A (0.23 Å). As is evident in Figure 2-3 and is discussed more fully below, the simulation reflects an ensemble of two structural populations characterized by differences at the C-terminus of monomer B.

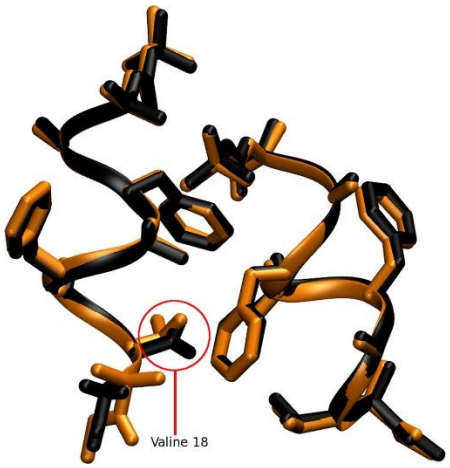

Figure 2-3: Superposition of the average simulated structure (black) against the structure refined from diffraction data (orange). The first decapeptide (monomer A) matches the X-ray data closely; and monomer B deviates in the side-chain conformation of its Val residue and in the helicity of its C-terminal backbone residues B6–B10.

Direct comparison of electron densities provides a more useful criterion for a structural comparison of the simulation against experiment, since it is X-ray scattering from an average density that determines the intensities of the observed diffraction peaks. For this, we calculated the electron density of 4000 evenly spaced snapshots taken from the simulation trajectory, amounting to 144 000 conformations of the ASU. The electron densities were optimally aligned to the crystallographic origin, as described in section 2.3, to account for the slow drift of the origin during the simulation. The average of all these electron-density maps was then taken, and a final Fourier transform computed to obtain the expected structure factors of a single crystal mosaic domain comprised of 144 000 ASUs. Comparison to the observed structure factors using the CCP4 program SCALEIT[94], [96] resulted in best-fit scale = 1.09 and B = −0.7948, indicating that the overall Wilson B-factor of the real crystal was remarkably similar to that predicted by the simulation. The R-factor of these calculated structure factors with the observed structure factors was 28% to 1.0 Å resolution and 21% to 2.0 Å resolution. After applying the 4-σ intensity cutoff traditionally employed when computing R-factors for small molecules, the agreement of our simulation-averaged structure factors with observed structure factors was 23% to 1.0 Å and 20% to 2.0 Å. This is remarkably good

agreement considering that the observed structure factors were not used to bias the simulation run, qualifying this R-factor as not just an R-free[97] but as the R-vault statistic proposed by Kleywegt.[98] Given the clearly anomalous behavior of the C terminus of the B chain in the simulation, some disagreement with the observed structure factors is expected, so the close agreement of the observed structure factors with those predicted by averaging over this unbiased MD simulation is remarkable.

We next refined the fav8 coordinates against the structure factors from the simulation density, which yielded an R-work/R-free of 9.6%/12.1%. This is higher than the reported experimental R-factor of 8%[79] primarily because the simulated crystal has more disorder than the experimental one, as discussed below. This refinement represents an "expected refined structure given the simulation density" and is arguably the best vehicle for making structural comparisons between theory and experiment, since X-ray scattering is determined by the average electron density and not by any average of the coordinates themselves. Table 2-1 presents RMSD statistics between this model and coordinates obtained by refinement against experimental density and by the more common procedure of simply averaging the coordinates over the simulation snapshots. (For consistency we use results from our rerefinement against experimental data; the RMSD of our rerefined structure vs the one originally deposited is 0.04/0.05 Å backbone/side chain.) The RMSD of the simulation-refined model to the experiment-refined model was 0.21/0.30 Å backbone/side chain, which is lower than the values (0.28/0.44 Å) obtained by coordinate averaging. Furthermore, calculation of the mean obtained by comparing simulation snapshots against each of these three structures yield higher RMSDs showing that while instantaneous simulation coordinates can differ to a greater degree from the refined model, the overall simulation average remains close. Therefore, a simulation-refined model provides a good representation of the average simulation structure while avoiding the geometric irregularities incurred with the more commonly employed coordinate averaging.

	Exp. refined	Sim. refined	Sim. average	Average snapshot
Experiment refined	0.0/0.0	0.205/0.301	0.283/0.423	0.462/1.180
Simulation refined		0.0/0.0	0.129/0.282	0.387/1.121
Simulation average			0.0/0.0	0.372/0.822

Table 2-1. RMSD Values between Various Structures. The statistics in each box are the backbone (first) and the all heavy atom RMSDs. Terminal capping residues were excluded from the calculation. "Experiment-refined" is the model obtained from refinement of fav8 against the experimental density in Phenix. "Simulation refined" is the structure obtained by refinement against the simulation average density. "Simulation average" is the structure composed of the mean coordinates of each atom over the entire length of the 2.4 μs simulation. The last column presents the average backbone/side chain RMSD of all simulation snapshots against the single structure for that row.

One global parameter which indicates how well a crystal lattice simulation is reproducing the crystal is its volume. In previous work, we have sought to reproduce this parameter arbitrarily to within 0.3% of the experimental result[78] and found that the choice of simulation models has a significant impact on the outcome.[76] As before, our simulations were performed in an NPT ensemble using a Berendsen barostat and Langevin thermostat. The experimental volume of 2795.8 $Å^3$ was maintained at a mean of 99.89 ± 0.003% of experiment (Figure 2-4 and S2). It is noteworthy that this was achieved without the addition of extra water molecules or other solvent. The fav8 X-ray structure is of high resolution, and the unit cell itself is very compact, but perhaps most importantly the unit cell is very dry for a proteinaceous crystal.

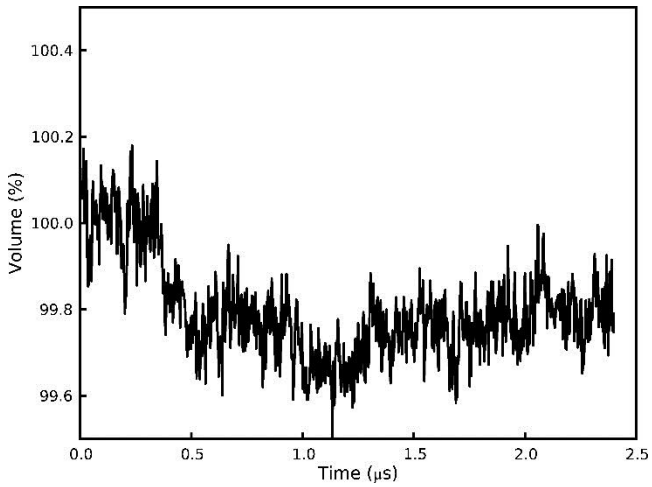

Figure 2-4: Volume of the supercell over the course of a 2.4 μs simulation. Following an initial settling, the system volume reaches an equilibrium value roughly 0.2% below the volume of the unit cell observed by X-ray diffraction. Instantaneous fluctuations of the volume have amplitudes of an additional 0.2%.

Crystallographic B-factors may be loosely interpreted as indicators of the thermal motion occurring in a crystal structure, but it is more accurate to say that B-factors can arise both from movements of the individual atoms within an ASU (intra-ASU or "local" disorder) as well as from rigid-body librations and lattice distortion (inter-ASU or "global" disorder). Isotropic B-factors are related to the mean-squared fluctuations of atoms around their average position by the formula:[14]

$$B = \frac{\langle u^2 \rangle \times 8\pi^2}{3} \qquad \textbf{(Eq. 17)}$$

where $\langle u^2 \rangle$ is the three-dimensional mean square deviation and B is the thermal isotropic B-factor. In crystallographic refinement models, an atom that is posited to be responsible for the surrounding electron density must exhibit a distribution of positions; this distribution is estimated from the available electron density, and the mean squared fluctuations of the distribution then imply a B-factor. The difference between contributions to the B-factors arising from "local" and "global" disorder, which can be discriminated by MD, is related to the difference between calculations of ASU and lattice RMSD.[77] We computed B-factors for the 2.4 μs simulation using both methods as

described in ref. [78]. Briefly, "RMSD B-factors" are calculated by first translationally and rotationally fitting each snapshot of each ASU during the trajectory to the crystal ASU and then calculating mean positions and positional variance. "Reverse symmetry" B-factors are calculated by reversing the translational space group operations by which the simulation supercell was constructed to align each snapshot of each ASU but without any translational/rotational fitting to minimize structural RMSD. The former method thus calculates positional variance stemming from intra-ASU fluctuations, while the latter also takes account of contributions from rigid-body librations and lattice distortion (i.e departure from crystal symmetry in the relative positions of the ASUs to each other). The computed B-factors are compared to the X-ray model in the left-hand side of Figure 2-5. If global disorder is removed from the calculation ("RMSD B-factors"), the simulation would underestimate the B-factors of most atoms. However, when disorder from rigid body libration is included in the B-factor estimates ("reverse symmetry B-factors"), the results for monomer A are in much better agreement with experiment (backbone B-factor RMSD 0.66 vs 1.73 for reverse symmetry and RMSD B-factors, respectively). Similar results are observed for monomer B except for C-terminal residues B6–B10. These residues undergo changes in their helical state that are coupled to water motion in the crystal lattice (discussed in detail in the following section). The right-hand side of Figure 2-5 presents the B-factors obtained from refinement against the average simulation density. These are generally in close agreement with the "reverse symmetry" B-factors that directly reflect the mean square fluctuations of the coordinates among the simulation snapshots. The refinement-derived and coordinate fluctuation-derived B-factors agree less well in the C-terminus of monomer B, where the simulation samples two different structural conformations. Whereas the coordinate-based B-factor statistic includes the large fluctuations between the two conformations, the refinement algorithm only models one conformation, but its B-factors underestimate the actual magnitude of fluctuations in the underlying simulation. The underlying disorder that is then not reflected in the B-factors gives rise to a higher R-work/R-free statistic. Five cycles of occupancy refinement with an alternate conformation for residues 15–20, reflecting the minor population found in the simulation, reduced R-work/R-free to

7.7%/9.2% (9.6%/12.1% without the alternate conformation) and converged to a relative occupancy of 71%/29% for the major and minor population of the ensemble, in close agreement with the relative ensemble populations of 72%/28% derived directly from the simulation.

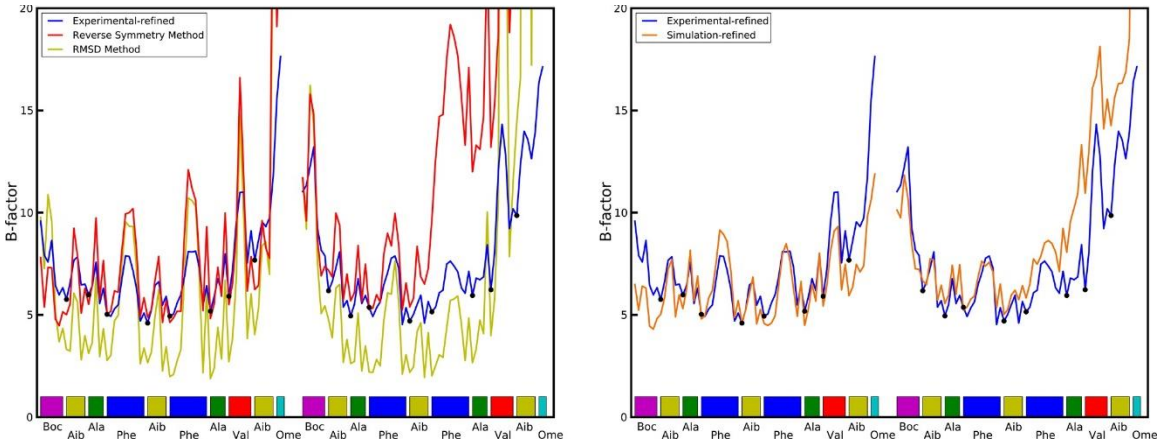

Figure 2-5. Left-hand plot: Comparison of computed atomic B-factors obtained over the course of the 2.4 μs trajectory to experimental data. "RMSD" B-factors only account for intra-ASU fluctuations and consistently underestimate experimental values. "Reverse symmetry" B-factors account for both local and global (inter-ASU) fluctuations and more closely match experiment. See text for further explanation of the two methods. Right-hand plot: Comparison of B-factors obtained from refinement against the experimental density and against the simulation average density. Cα atoms are indicated with dots.

2.4.2. Crystal Solvent Dynamics

While the RMSD of the peptide converges very quickly in the simulation, the RMSD of the solvent does not converge even after >2 μs of simulation. A visualization of the crystal reveals that the packing of the crystal is such that "channels" for water molecules are formed within the crystal. These channels are co-linear with lattice vector a and provide little steric hindrance for waters to move between adjacent unit cells. The waters are seen to rapidly diffuse between unit cells through the channels. A careful inspection of the trajectory reveals that the water molecules do not flow smoothly through the channels but rather make sudden hops between positions in adjacent unit cells.

Trajectory frames were recorded every 10 ps, and in this time water molecules are sometimes seen to move by several angstrom.

A diffusion constant was calculated for the water from a linear fit of the cumulative mean square displacement of the waters from their initial position using the Einstein diffusion equation for one dimension:

$$D = \frac{\sum_i [r_i(t + \Delta t) - r_i(t)]^2}{2\Delta t} \quad \textbf{(Eq. 18)}$$

Plots of the mean square displacement in each direction of space, shown in Figure 2-6 and S8, do indeed demonstrate that the water is dynamic along the channels, while it is restrained from moving in other directions by the channel walls. The channels can be estimated to be 3–4 Å wide based on a converged mean square displacement of about 12 Å^2 in the directions perpendicular to the channel axis. Diffusion along the channel in the four simulations ranged from 1 × 10–8 to 3.4 × 10–8 cm^2/s, with a mean diffusion rate of 2.5 × 10–8 cm^2/s calculated after discarding the first 400 ns of each trajectory for equilibration. This is roughly 2000 times slower than the reported 5.2 × 10–5 cm^2/s diffusion constant of TIP3P water[99] and 1000 times slower than the experimental diffusion constant of liquid water at the same temperature; the waters are dynamic in the simulation, but movement through the channels is constricted. Some variability in water diffusion is evident, as a function of time, in each of the four simulations and particularly in the 2.4 μs trajectory; over the first 400–500 ns, a diffusion constant of 3.6 × 10–8 cm^2/s could be calculated, but the rate abruptly changed to 1.0 × 10–8 cm^2/s thereafter. A possible connection between these abrupt changes and the disorder in the C-terminus of monomer B is explored later in this section.

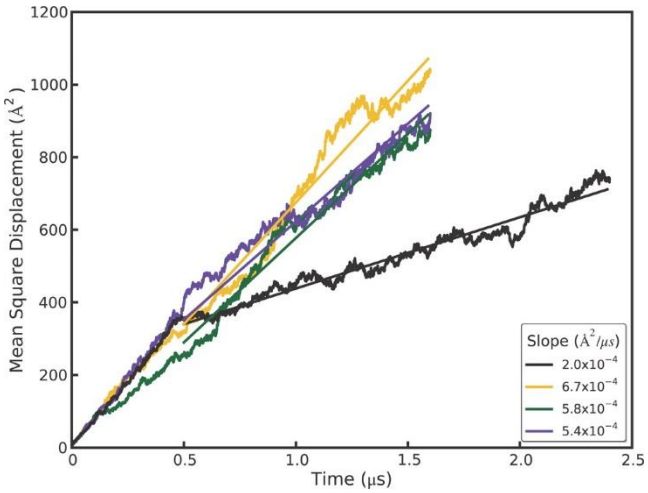

Figure 2-6. Mean square displacements (MSD) of water molecules over the course of three 1.6 μs and one 2.4 μs simulation trajectories. The slope of the linear fit used to compute the diffusion coefficient is shown in the box.

Further analysis in Figure 2-7 shows that the water molecules occupy several distinct sites within each unit cell along a channel. Hydrogen bonding between water molecules or to peptide backbone atoms is expected to be the primary determinant of these energy minima. Although the average number of waters per unit cell is set to be four in our simulation, the water dynamics produce a heterogeneous population of individual unit cell states; at any given time unit cells may contain as few as zero and as many as eight water molecules. A histogram of the water states occurring throughout the simulation (Figure 2-7) shows that although 4 is the average state, 5 is in fact the most populous water state. A direct comparison of the cumulative water density from simulation (Figure 2-8, left panel) to the experimental electron density (Figure 2-8, right panel) reveals close correspondence between the simulation and X-ray data. In both the simulated and experimental structures, two crystallographic waters are located centrally within a compact and spherical lobe of the simulated density, while the other two crystallographic waters are located on smeared, dumbbell-shaped regions of density. Correspondingly, these waters also have 3 times higher experimental B-factors. Both images also reveal a fifth area of water density. No specific water was attributed to this density in the X-ray structure, but a partial water occupancy at this position is indicated by the frequently occurring

5-water state (Figure 2-7) and is consistent with the experimental electron density. Furthermore, a meticulous strategy of free refinement of water occupancy identified 17 putative water peaks and converged to a total of 61 electrons or 6 water molecules altogether. The final R-work/R-free statistics for this model were 4.1%/5.8% compared to 6.5%/9.2% for refinement of the 4-water model. Therefore, although exchange of the water molecules between unit cells is not directly reflected in the refined fav8 structure, a model in which exchanges and migration occur continuously is fully consistent with the X-ray diffraction data and leads to improved agreement with the observed structure factors.

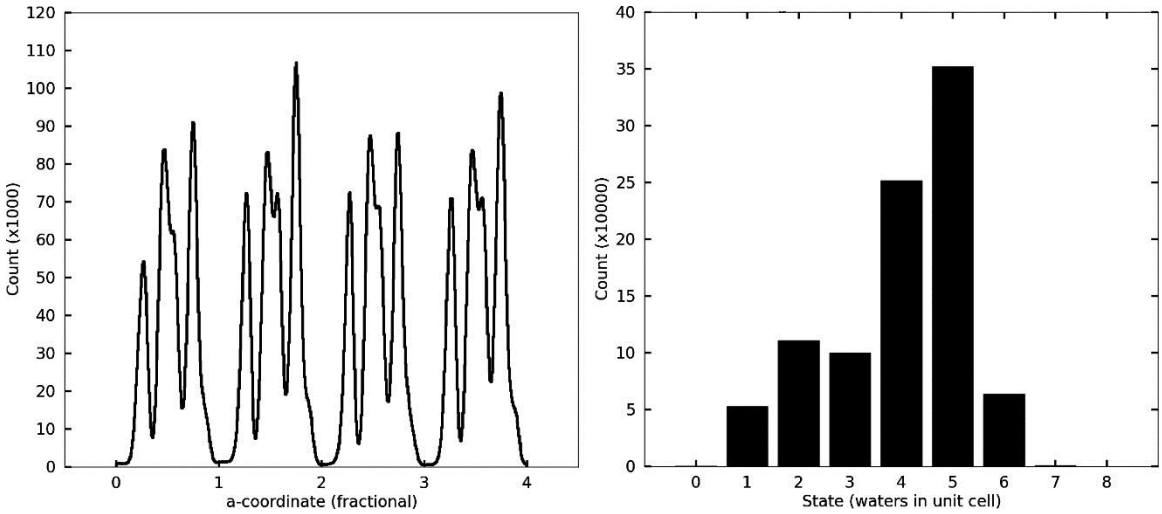

Figure 2-7. Water densities in the channels observed in simulations. The left-hand panel depicts the density of waters as a function of the a crystal vector coordinate, summed over all nine channels running across the simulation box. The abscissa is numbered according to unit cell fractional coordinates. The right-hand panel plots a histogram of times which each unit cell in the simulation was observed to be associated with a particular number of waters during the 2.4 μs trajectory.

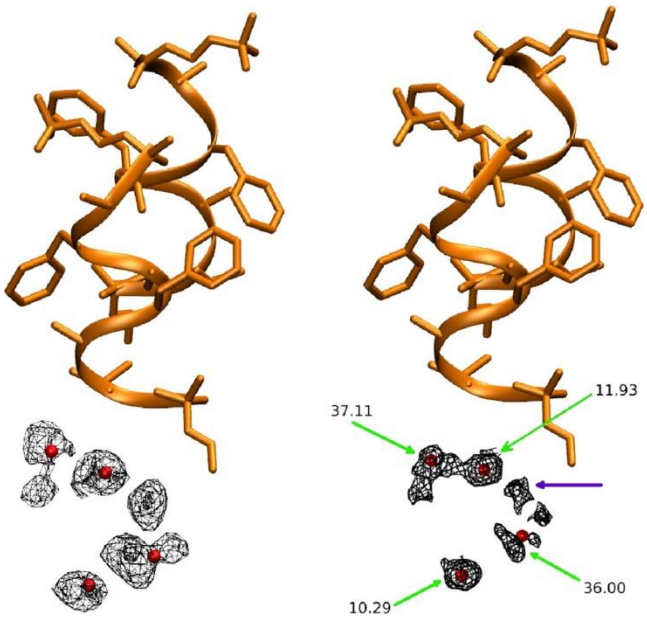

Figure 2-8. Water density observed in the 2.4 μs simulation, obtained by using crystal symmetry operations to superimpose all simulated waters onto a single unit cell. Crystallographic peptide is shown in orange and crystallographic water oxygens as red spheres. Left-hand panel shows the simulated water density (mesh encloses 90% of water density), right-hand panel shows the electron density obtained by X-ray diffraction (2mFo–DFcalc map at 0.8σ). Green arrows point to crystallographic waters and indicate their experimental B-factors, and purple arrow shows a fifth lobe of water density (see text). Produced with VMD and ccp4 mg.

To investigate the tendency of unit cells to take on varying amounts of water, residence times were calculated for each of the water states. We used different smoothing windows to eliminate noise, but regardless of the smoothing window, the one water state exhibits by far the longest residence time (Figure 2-9). Closer examination of individual water cells revealed that unit cells were rarely occupied by only a single water, but when such dry states did occur, they tended to persist for hundreds of nanoseconds or even indefinitely. A visual inspection of the trajectory revealed that these dry unit cells undergo a conformational change upon acquiring the defect, strongly associated with two other characteristics: elevated propensity for a 3_{10} helical conformation in monomer B and the χ1 dihedral of Val B8 flipping to gauche(-). By creating a vector of zeros (state absent) and ones (state present) for all unit cells and all frames of a trajectory, the Pearson correlation coefficients

between various states can be computed. Over the course of the 2.4 μs trajectory, the dry state correlates with monomer B 3_{10} helicity by a coefficient of 0.986 and with the Val B8 gauche(-) rotamer state by a coefficient of 0.965, and the correlation between the Val B8 gauche(-) rotamer state and monomer B helicity is 0.967 (see Figure S10). It is difficult to determine whether one of these characteristics leads to another, but we can quantify the time by which the correlations develop. If the correlation between states A and B is 0.95 over a period of 2 μs but only 0.3 when averaged over many short intervals of 10 ns, it can be said that state A or B does not lead to the other within 10 ns, although the two are associated in the long term. Formally, we computed the Pearson correlations between the three states over windows of up to 100 ns from all trajectories using the formula

$$\frac{1}{w}\sum_{k=1,w}\frac{cov(x,y)}{\sqrt{cov(x)cov(y)}} \quad \textbf{(Eq. 19)}$$

For a given window size, the summation runs over all the nonoverlapping windows in the trajectory, and $cov(x,y)$ denotes the covariance of the vectors x and y. The elements of x and y are the average values of the given characteristic in the window for each of the unit cells. As shown in Figure 2-10 the correlation between monomer B 3_{10} helicity and the dry state rapidly approaches its long-term asymptotic correlation, whereas the other two correlations take much longer to develop, implying that C-terminal helicity and wet or dry unit cell states are tightly coupled, whereas the gauche(-) Val B8 rotamer conformation may be favored by monomer B 3_{10} helicity or the dry state but is not a gating motion leading to either.

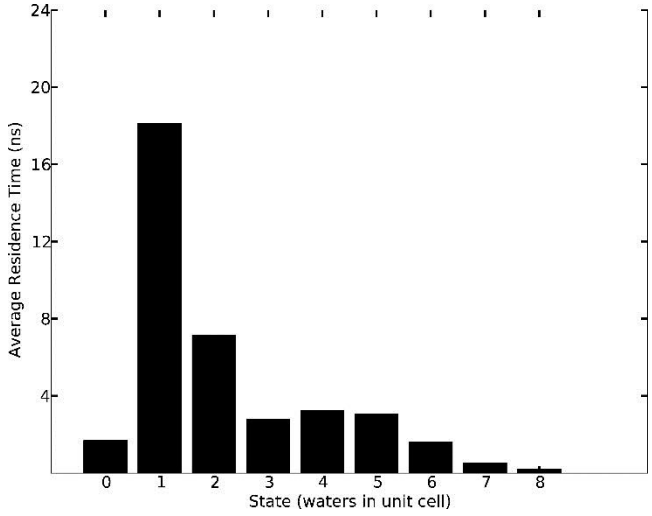

Figure 2-9. Mean residence times for each occurring water state over the course of the 2.4 μs trajectory. The one and two water states, though much less frequent than other states (cf. Figure 2-7), exhibit very long residence times, in some cases extending into hundreds of nanoseconds.

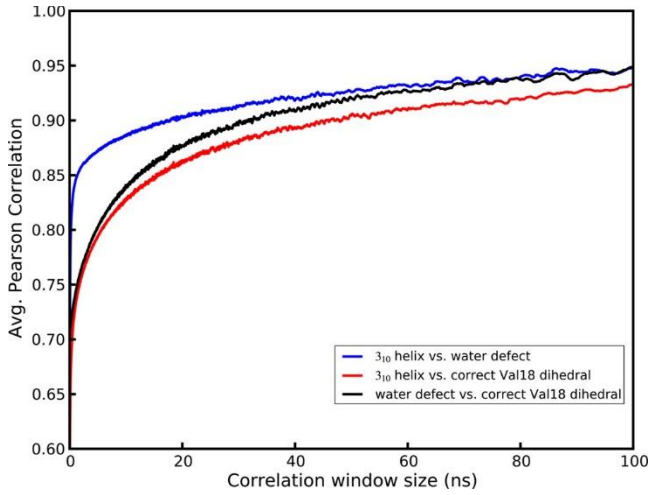

Figure 2-10. Correlation, as a function of measurement time, between the presence of a Val B8 gauche(-) rotamer, 1- or 2-water defects, and 3_{10} helical conformation. A conformational change of monomer B helicity is found to be more strongly connected to water defects than either condition is to the Val B8 rotamer state.

In our simulations, there appear to be two structural subpopulations of unit cells. The major population, about 75% of the cells, maintains the crystallographic C-terminal α-helical conformation, a wet unit cell with 3–5 water molecules, but puts the side-chain of Val B8 in a noncrystallographic

trans conformation. The minor population of unit cells displays increased propensity for a 3_{10} helical conformation in monomer B, leading to high B-factors and higher positional RMSD in these residues, and retains only one or two waters per unit cell; the minor population also places the Val B8 side-chain in its crystallographic gauche(-) rotamer. The disagreement in average structure and B-factors leads us to conclude that the minor population is an artifact of the calculation. For the Val B8 rotamer, however, both the Fo–Fc map and a Ringer[57] plot shown in Figure 2-11 provide evidence of a minor trans conformation for Val B8 in the original fav8 data. Furthermore, the trans conformation is the favored conformation of valine generally,[100] so the preponderance of this state in our simulations is unsurprising. Evidence for the occurrence of the alternate valine rotamer in the crystal is provided by occupancy refinement of the model with two alternate conformers. Standard anisotropic refinement of the model with and without the alternate valine conformer produced an R-work/R-free of 4.11%/5.84% (without the alternate trans rotamer) and 3.89%/5.53% (with the alternate rotamer). The occupancy of the trans/gauche(-) rotamer refined to 74%/26% ± 2%, which is the reverse of that seen in the 2.4 μs simulation (32%/68%), suggesting that the relative energy of the gauche(-) conformation is about 1 kcal/mol too negative in the simulation, but that finding both conformers present is to be expected.

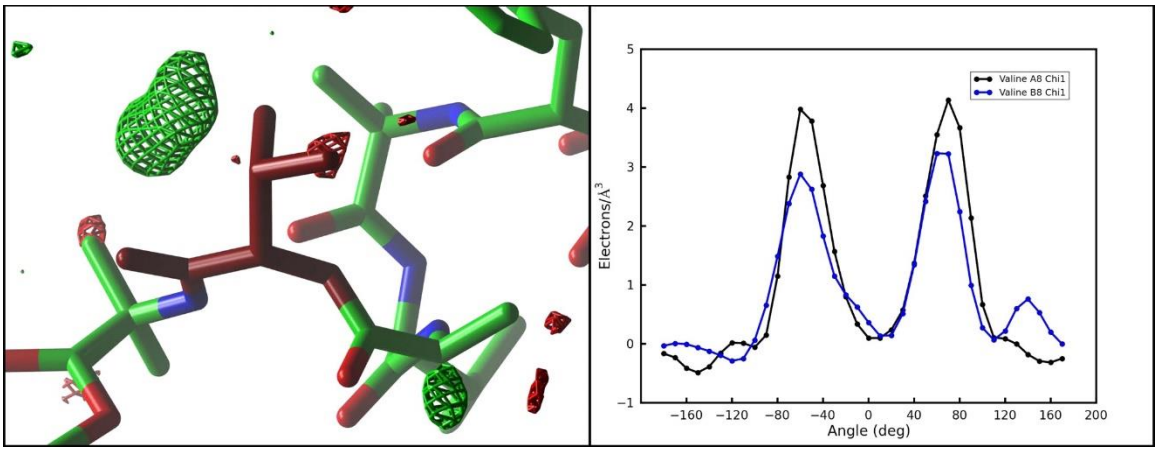

Figure 2-11. Experimental electron density of the Val B8 side chain reveals evidence for partial occupancy of the trans rotamer that is preferentially sampled in our simulations. The left-hand panel shows the Fo–Fc map sampled on a 0.50 Å3 grid and contoured at 4.0 (green) and −4.0 (red) in the

vicinity of Val B8 (burgundy). The valine side chain is seen in the experimentally determined gauche(-) rotamer. A region of positive density indicates the missing alternate trans rotamer sampled in our simulation. Image generated with ccp4 mg. The right-hand panel shows the output of Ringer(47) for the χ1 angle of Val A8 (black) and B8 (blue). An additional peak in the latter case points to the presence of a partially occupied trans rotamer in the electron density.

The coupling between C-terminal helicity and the dry states offers a possible explanation for the sudden shifts in the water diffusivity seen in Figure 2-6. During the 2.4 μs simulation, after about 400 ns of dynamics, 9 of the unit cells in the crystal enter a prolonged 1 water defect state. The supercell has nine water channels, and dry cell defects are distributed one per channel. Near the end of the trajectory, from 2 to 2.4 μs, some cells are seen to escape the water defect: with only 6 dry unit cells remaining, water diffusivity increases by almost 2-fold. These observations indicate that sampling of the one water defect corresponds to slowing of the water flow in a given channel. Moreover, two concurrent water defects are very rarely observed in one water channel. We hypothesized that because the water defect corresponds strongly to 3_{10} helical sampling and because the 3_{10} helix is a more tightly wound but longer helix, it could be jutting into the channel to sterically impede water movement at that point. Effectively it would serve as a block in the channel which would reduce overall water diffusion. However, expelling waters at the defect site would force them into adjacent cells and inhibit other cells from drying along that particular channel.

2.5. Conclusions

We present here results of 6 simulations of a peptide crystal composed of 36 unit cells in a P1 crystal system. Our results offer some of the most detailed agreement to date between a simulation and the diffraction data taken from a biomolecular crystal. In all, the peptide crystal supercell was simulated for 9.6 μs. Our results show that the Amber ff99SB force field coupled with a TIP3P water model maintains the integrity of the crystal structure very well. Volume, RMSD, and average structure all agree well with experiment. Remarkable B-factor agreement is obtained, except for the final residues of the second peptide. Both the aromatic t-stacking and hydrogen-bonding interactions

that stabilize crystal packing are maintained. Methodologically, refinement against the average simulation density yields an optimal representation of the average simulation structure and avoids the pitfalls of the more commonly employed coordinate averaging over the simulation trajectory. Calculation of B-factors from coordinate fluctuations yields close agreement with B-factors from crystallographic refinement only when global disorder and lattice distortion effects are accounted for. On the other hand, B-factors from refinement are found to underestimate coordinate fluctuations where the simulation samples an alternate conformation.

The simulation also provided a glimpse into the hidden dynamics of the crystal. The atomic motions seen in the simulation can be placed into three broad categories. Most of the peptide atoms vibrate around a single average structure (with amplitudes well-described by the experimental atomic displacement parameters.) Atoms at the end of the second peptide visit alternate conformations, and the B-factors obtained by refinement against a single structural model underestimate the extent of this motion. (Some evidence for the alternate conformations is present in the observed electron density, but the simulation appears to exaggerate their importance.)

Water molecules observed in the X-ray structure are not bound to any particular unit cell but rather exchange positions frequently within unit cells and between neighboring cells along solvent channels. The time scale of the simulations permits measurements of this diffusion as well as correlation of protein motion and structural heterogeneity resulting from the migratory crystal defects in unit cells. The dynamic nature of the solvent produces a heterogeneous population of water states with individual unit cells at any given time containing anywhere from zero to eight water molecules. A five water state is seen to occur most frequently, and a fifth lobe of water density is observed corresponding to electron density found in the experimental diffraction data. Somewhat larger defects are also observed in which unit cells dry to only a single water molecule, and these defects appear to slow the diffusion of water throughout entire channels. This transient variability in solvent content offers a reasonable model of the true crystal lattice—the average density of simulated water recovers the crystallographic density with remarkable precision. While traditional crystal

refinement to a single ASU gives no indication of water hopping or variation in water content between cells, it is known that mean residence times of single water molecules are short (microseconds even for waters buried deep within a protein cavity).[101]–[103] This behavior is explicitly revealed here by the MD simulations. Moreover the simulations lead to the identification of additional water positions and improved refinement statistics (R-work/R-free), thus demonstrating the potential utility of all-atom crystal simulations in the interpretation of experimental electron density. We thus provide evidence for the potential of MD to contribute additional structural information to the interpretation of crystallographic data that would otherwise remain lost.

An ensemble of two structurally different populations of unit cells is observed. About 25% of the unit cells are characterized by increased 3_{10} helical propensity, decreased water content (containing only 1 or 2 waters) and occupancy of the gauche(-) $\chi 1$ rotmer of Val B8. These three characteristics are highly correlated over the course of the microsecond long simulations, but it is unclear which of them might be the driving factor. Because 3_{10} propensity is not seen in the sequentially identical monomer A, we believe that this behavior is not driven by the valine dihedral but rather must be caused by factors external to the monomer itself. The water channel at the C-terminus provides a spatial opening for the tighter but longer 3_{10} helix to form, and variations in water content or close contacts with an Aib 5 side chain in monomer A can affect hydrogen bond-stabilizing interactions in the helix. Nevertheless, the presence of this conformational ensemble is only partly consistent with the experimental data, which leads us to believe that part of this observation is a simulation artifact. Careful examination of the experimentally derived electron density and refinement of a model with an alternate conformation does indeed support the presence of a minor population of the alternate valine rotamer. This is consistent with recent results from the Ringer program,[57] showing that 18% of a test set of PDB structures contained unidentified alternate conformations. As discussed above, the simulated water density also closely tracks the diffraction data. However, the disagreement in B-factors observed in the C-terminus of monomer B indicates that a simulation artifact is present. There is also no substantial evidence in the experimental

electron density for the presence of both 3_{10} and α-helical varieties of the second monomer. Thus we conclude that the observed correlation between the unit cell water content, the Val B8 rotamer, and the helical conformation of the molecule is an artifact of the simulation. This is valuable information for further work on improved force field models for MD. A fine equilibrium of protein–protein and protein–solvent interactions drives the formation of the various types of helices,[104] and we suspect that further fine-tuning of hydrogen-bond treatment and solvent parameters in current force field models is necessary.[105], [106]

Thus the development of all-atom crystal simulations requires continued work. More simulations on both small and large structures are needed. We are also continuing investigation of the fav8 peptide with simulations of varying water content as well as simulations using the all-atom AMOEBA[34], [35], [107] force field to elucidate the interactions leading to the alternate unit cell population. Taken together, our results demonstrate that MD simulations of crystals possess strong potential as both a tool for validating next generation force fields against experimental data and as a powerful tool for extricating additional information about biomolecular structure and dynamics from diffraction data.

2.6. Supporting Information

Parameters for nonstandard amino acids; simulation input files for Amber programs; additional figures analyzing RMSD, volume, and water displacement over time. This material is available free of charge via the Internet at http://pubs.acs.org.

2.7. Acknowledgement

Experimental diffraction data of the fav8 decapeptide was kindly provided by S. Aravinda and P. Balaram. We thank Darrin York, Huanwang Yang, and Joe Marcotrigiano for helpful discussions. This work was supported in part by NIH grant GM 45811 and by a Rutgers Presidential Fellowship to P.A.J. J.M.H. is also supported by the National Institutes of Health GM073210, GM082250, and

GM094625 and the Integrated Diffraction Analysis Technologies (IDAT) program under contract no. DE-AC02-05CH11231 with the U.S. Department of Energy.

Chapter 3. *Improving Model Interpretation through Crystallographic Refinement and Molecular Dynamics Simulation*

3.1. Abstract

Molecular dynamics of crystals is useful both as a diagnostic tool of force field accuracy and to provide additional information for crystallographic data interpretation. In a previous study[62] of a small peptide crystal we demonstrated excellent agreement with experimental data but also discrepancies which led us to predict a different representation of solvent in the model. We now present the results of seven 500ns simulations of the crystal supercell with varying solvent content as well as 75 ns of simulation using the AMOEBA polarizable force field. We see that both structural (RMSD, R-factor) and dynamic (B-factors) agreement with experimental data improves as water is added to the system and then deteriorates again past a certain point. Important structural insights from the previous study such as the dynamic flow of solvent through crystal interstices and side chain conformational heterogeneity are maintained while elements postulated to be artifacts, such as persistent dry states and an elevated B-factor region around the C-terminus of monomer B disappear. All results are wholly consistent with our predictions from the previous study. Furthermore, we make methodological contributions by demonstrating reproducibility of crystal simulation data and validating the NVT ensemble approach to crystal simulations. Our results lead to a more accurate understanding of the physical crystal and confirm the potential of crystal molecular dynamics methods for validating experimental refinement results.

3.2. Introduction

Crystallography is fundamental to the study of biomolecular structure with over 90% of the models in the Protein Data Bank[16] having been solved by crystallographic methods. However, the typical approach in crystallography has been to present a single static model of atomic coordinates that best agrees with experimental diffraction data. This is expected because the diffraction experiment is a time and space-averaged technique.[65] Nevertheless, because physical crystals are

both spatially heterogeneous and temporally dynamic[55], a more complete understanding of biomolecules entails finding a way to move past the single static representation. A number of research efforts in recent years have moved in this direction.[53], [58] One such approach is through molecular dynamics (MD) simulations of crystals.[64], [76]–[78], [108]

In recent work[62], an MD simulation of the crystalline form of a small synthetic decapeptide molecule, referred to as fav8[79], provided insight into the heterogeneous and dynamic landscape sampled by individual molecules in the crystal. Statistically averaged quantities such as atomic coordinates, atomic displacement parameters (ADPs) and average electron density were shown to be in good agreement with experimental data, but time-resolved snapshots of the simulation offered previously undiscovered details about the crystal. Water molecules were shown to exchange dynamically between adjacent unit cells in a hop-scotch fashion. A valine side chain was discovered to sample additional rotameric conformations that were not easily discernable in the experimental data. The MD results informed a re-refinement of the model resulting in an R-free[97], [109] drop from 9.2% to 5.8%. It was furthermore postulated that i) the crystal unit cell contains 6 water molecules (2 more than the original deposited model); ii) large atomic positional fluctuations around the C-terminus of the second monomer in the unit cell were artefacts of the simulation; iii) likewise that a high correlation (Pearson correlation coefficient %) between unit cell solvent content, a valine rotamer and the secondary structure helix type was an artefact of the simulation.

To test these predictions we performed a series of additional MD simulations of the fav8 crystal with varying amounts of solvent. The results are presented below. Six separate simulations of the fav8 crystal consisting of a 4x3x3 arrangement of unit cells with 4, 4.5, 5, 5.5, 6 and 6.5 water molecules per crystallographic unit cell. Additional waters were added manually to the asymmetric unit in non-sterically-clashing locations followed by energy minimization and MD equilibration to allow waters to settle into energetically favorable positions. Production of each simulation was carried out for 500ns. The first of these, 4 waters per unit cell, reproduces the same conditions as the simulation reported in the previous study and was carried out to test reproducibility of the published

results. All simulation protocol parameters were as those reported in Ref. 1. Briefly, all simulations consisted of 4x3x3 unit cells with explicit solvent (TIP3P[82]), periodic boundary conditions and parameters provided by the Amber ff99SB[30] force field. The only difference was that we ran all simulation in a canonical ensemble (NVT) and monitored pressure fluctuations instead of volume to assess conformance with experimental crystal data. The canonical ensemble was used because other work (to be published in a separate paper) indicates that this is preferred for crystal MD simulations, leading to more straightforward analysis of simulation results and is computationally more efficient. To confirm that the switch to a canonical ensemble does not impact our results, we also performed an additional isobaric/isothermal (NPT) ensemble simulation of the 5 water per unit cell model and compared results. We also carried out a simulation of the original model (4 waters per unit cell) using the polarizable AMOEBA[34], [35], [107] force field. MD was carried out using Amber12[80] and all analysis was done using a combination of AmberTools[26] (CPPTRAJ[110] and the XtalAnalysis package) and in-house scripts. A summary of simulations is shown in Table 3-1.

Name	Force field	No. H2O	Length	Ensemble
4water_JACS2013	ff99SB	4	1000ns	NPT
4water	ff99SB	4	500ns	NVT
4.5water	ff99SB	4.5	500ns	NVT
5water	ff99SB	5	500ns	NVT
5.5water	ff99SB	5.5	500ns	NVT
6water	ff99SB	6	500ns	NVT
6.5water	ff99SB	6.5	500ns	NVT
4water_AMOEBA	AMOEBA	4	75ns	NVT
5water_NPT	ff99SB	5	500ns	NPT

Table 3-1. Summary of performed simulations. The first column provides the name used to refer to the simulations in the text. 4water_JACS2013 refers to the simulation described in Ref. 1 and is added here for comparison. The third column provides the number of water molecules per unit cell in the simulated system. NPT is constant pressure/constant temperature ensemble (isothermal/isobaric). NVT is constant volume/constant temperature (canonical).

3.3. Results

3.3.1. Impact of additional solvent on molecular structure.

Addition of solvent immediately results in significantly improved agreement of atomic positions with the experimental model (Figure 3-1). The 4water simulation has a mean instantaneous backbone root mean square deviation (RMSD) of 0.60 Å and an all atom RMSD of the average simulation structure of 0.39 Å, while in the 4.5water simulation the RMSD statistics drop to 0.49 Å and 0.39 Å respectively and in the 5water simulation they further decreases to 0.45 Å and 0.38 Å respectively (Table 3-2). Further addition of solvent leaves the RMSD statistics essentially unchanged. RMSD of simulations with additional water also converges quickly, whereas RMSD of the 4water simulation appears unconverged even after 500ns. Comparison of best-fit (only accounts for internal atomic positional variation) vs. lattice (includes variation due to both intra-molecular and inter-molecular or lattice motion) indicates that the improvement in RMSD can be ascribed to structural changes within monomers and not to the relative position of monomers to each other within the crystal lattice. As a whole, the data indicates that significantly better agreement with experiment is obtained with 5 or more waters per unit cell.

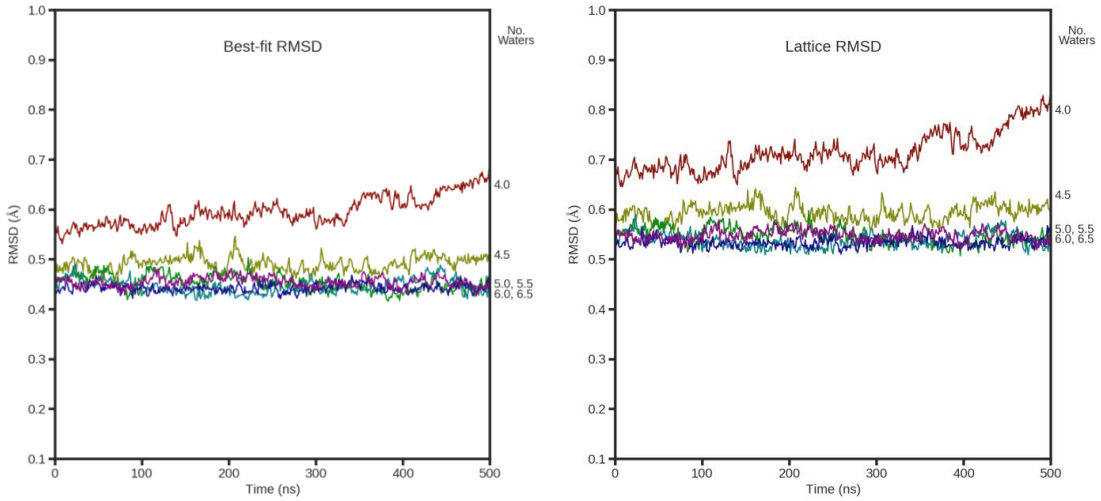

Figure 3-1. Atomic coordinate backbone RMSD for the simulations with varying solvent content. Lines are labeled along the right vertical axis. Left panel shows "best-fit" RMSD. Right panel shows "lattice" RMSD. RMSD statistics computed for peptide only.

	Avg. bbone rmsd	Avg. sdch rmsd	Resid 1-16 rmsd	Resid 1-18 rmsd	Resid 1-8, 10-18 rmsd	Mean instantaneous best-fit bbone rmsd	B-factor RMSD
4water_JACS2013	0.32	0.45	0.19	0.21	0.21	0.55 ±0.04	4.48
4water	0.31	0.39	0.18	0.21	0.21	0.60 ±0.03	4.91
4.5water	0.26	0.37	0.15	0.17	0.17	0.49 ±0.02	1.54
5water	0.25	0.38	0.14	0.16	0.16	0.45 ±0.02	1.36
5.5water	0.26	0.38	0.14	0.16	0.16	0.45 ±0.02	1.45
6water	0.25	0.37	0.15	0.16	0.16	0.44 ±0.01	1.51
6.5water	0.26	0.39	0.16	0.18	0.18	0.46 ±0.01	1.47
4water_AMOEBA	0.35	0.36	0.20	0.22	0.22	0.61 ±0.07	2.75
5water_NPT	0.25	0.38	0.14	0.16	0.17	0.46 ±0.03	1.33

Table 3-2. Summary of structural and fluctuation characteristics. Simulations are labeled according to the names assigned in Table 3-1. All RMSD values are compared to the experimental model. 1st column – the backbone atom RMSD of the average structure from the entire simulation. 2nd column – the side chain atom RMSD of the average structure. 3rd-5th columns – backbone RMSD of the average structure including only the specified residues. 6th column – mean instantaneous best-fit backbone RMSD over the entire simulation. 7th column – RMSD of the B-factor values from

simulation compared to the refined model. Tail residues were excluded from this statistic as well as the side chain residues of valine B8 (multiple rotamers) in order to show trends more clearly.

3.3.2. Impact of additional solvent on atomic fluctuations.

The previously published simulation showed very good agreement of atomic B-factors with experiment for the first monomer in the unit cell and for the N-terminus half of the second monomer, but there was significant divergence of B-factor values with a "hump" of excessively high fluctuations around the C-terminus of the second monomer. Addition of solvent leads to an radical improvement in B-factor agreement where the "hump" is virtually eradicated with 4.5 waters per unit cell (Figure 3-2). With 5 waters per unit cell the agreement of B-factor values with experiment is close to perfect. However, as even more water is added, simulated B-factor values fall and underestimate the experimental values. This is likely due to a "freezing" effect as atoms are locked into place and movement is prevented by excessive packing of solvent matter into the crystal interstices. Furthermore, we observe that addition of solvent disrupts the highly correlated behavior between a valine rotamer, secondary structure propensity of the second monomer and unit cell solvent content that was reported previously for the 4water_JACS simulation. Importantly, however, the valine side chain continues to sample two rotamers in accordance with experimental data as was previously shown. Thus, fluctuation analysis indicates that optimal agreement with experiment is obtained with 5 waters per unit cell and that the B-factor hump and correlated behavior observed previously were artefacts, as was postulated.

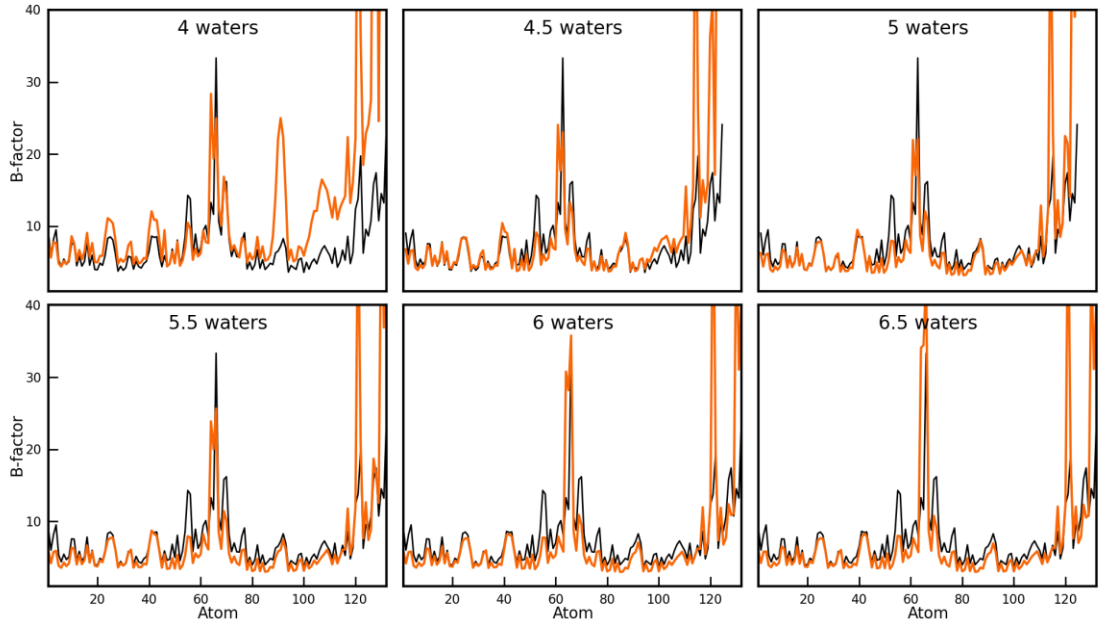

Figure 3-2. "Lattice" isotropic B-factors from each simulation (orange) compared to experimental model (black). B-factors are shown for all heavy atoms.

3.3.3. Impact of additional solvent on average electron density.

In the previously published simulation the average simulation electron density was calculated and agreed with the experimental density with an R-factor of 23% to 1.0 Å resolution. Considering the unbiased nature of the simulation and it's length this was an impressive result and akin to the R-vault statistic proposed by Kleywegt et al.[98] More recently we have obtained similar statistics from other crystal simulations. For example, a triclinic lysozyme simulation(*publication in preparation*) yields an R-factor of 24.9 to 4.0 Å but only 41.2 to 1.0 Å, and a DNA decamer crystal simulation[64] yields 55.2 to 4.0 Å and 64.7 to 1.0 Å. This shows that the simulations really are unbiased and that the result obtained with fav8 is not expected but rather indicative of the high degree of agreement between the computational and experimental result. Presently, for the fav8 simulations with added solvent, the R-factor is seen to first decrease as waters are added to the model and then begins to increase again with a clear minimum reached with 5 waters per unit cell (R-factor 21.1% to 1.0 Å, Figure 3-3). This further confirms the previous conclusion that 5 waters per unit cell is the correct quantity.

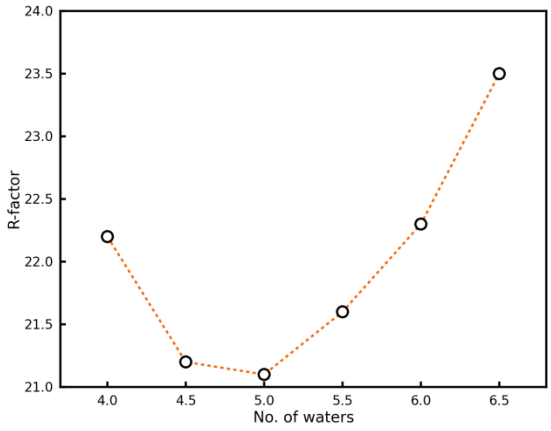

Figure 3-3. R-factor between experimental structure amplitudes and amplitudes calculated from the average simulation electron density. Structure factors up to 1.0 Å resolution were included in the calculation.

3.3.4. Methodology for crystal MD simulations.

The work presented here has provided some important methodological insights for future MD simulations of crystals. First, we ran the 5 water simulation using both a canonical (NVT) and an isothermal/isobaric (NPT) ensemble. Recently, we have seen that NVT simulations are preferable: consistency with experiment can be verified by monitoring system pressure while post-simulation analysis is simplified and artefacts are avoided by the fact that unit cell parameters are constrained to experimental values. Comparison of the 5water NPT and NVT simulations reveals almost identical results on all counts including RMSD, B-factors, electron density and R-factor, water diffusion rates, structural heterogeneity, etc. Second, the work presented here included a new simulation of the 4 water model. Comparison to the previously published simulation also reveals virtually identical results in all statistical quantities. Thus our findings indicate that crystal simulation results are reproducible and that the NVT approach is valid.

3.3.5. AMOEBA polarizable force field simulations.

To confirm the previously published findings and help discern between physical water behavior from simulation artefact, we carried out a 75ns simulation of the 4 water model using the

AMOEBA[34], [35], [107] force field. AMOEBA is a polarizable force field that allows for real-time adjustment of partial charges during the course of the simulation and thus leads to more accurate treatment of molecular polarizability and non-bonded interactions. RMSD of the average simulated structure reached similar accuracy, compared to experiment, as the ff99SB simulation: 0.36 Å for backbone atoms and 0.52 Å for all heavy atoms. Water continued to translocate dynamically across unit cells as in the ff99SB simulation, but slower, at about half the diffusion rate. The valine B8 rotamer continued to sample the alternate rotameric conformation at 30% frequency, in agreement with both ff99SB and experimental results. On the other hand, the large B-factor "hump" at the C-terminus of the second monomer was not reproduced (although B-factors remained high in places, above those for the simulations with additional waters), and we found no correlation between the persistent dry states of unit cells, the valine rotamer and the secondary structure propensity of the 2nd monomer. These results are all in agreement with original postulates from the previous work and confirm the results contained therein.

3.4. Discussion

Our results lend strong credibility to the findings from the initial fav8 crystal simulation. After adjusting solvent content, dynamic movement of water molecules through channels formed by crystal packing is still observed while time and space-averaged observables are in agreement with experimental values. Rotameric heterogeneity of a valine side chain is also observed, in accordance with the previous simulation and validated by experimental data. On the other hand behavior that was postulated to be an artefact of the simulation because it did not agree with experimental observation (elevated B-factors around C-terminus of second monomer and high correlation between low solvent content in particular unit cells, 3_{10} helical propensity in the second monomer and propensity towards the gauche(-) rotamer of one of the valines) completely disappears. The results allow us to conclude that the physical fav8 crystal contains 5 waters per unit cell, one more than was originally modelled.

These results are significant for a number of reasons. First, they highlight the potential utility of an interplay between direct analysis of experimental diffraction data and computational simulation of the crystal. The two approaches can be complementary, each one in turn informing the other and both leading to a more accurate understanding of the crystal. Second, they underscore the care that must be taken in both crystallographic model building and refinement and MD simulations. Both methods are prone to over or under-interpretation of data features and validation using an orthogonal method, such as the approach taken here, can be useful. Third, they show that artefacts in MD simulations are not just due to force field error. Even an accurate force field can lead to artefacts when the initial conditions of the system are modeled incorrectly. Fourth, they show that a highly dynamic and heterogeneous structural landscape of the crystal can nevertheless wholly agree with average statistics and high resolution diffraction data. Fifth, it underscores the importance of solvent for crystal packing and stability.

This last point merits additional consideration. Considerable research effort is currently being spent on crystal packing prediction and *de novo* crystal construction. Our work shows that reducing the amount of solvent in a crystal interface by one molecule can lead to significant instability including changes in side chain rotamers, secondary structure propensity and dynamics within the crystal. It is known that hydrogen bonding networks mediated by water can lend considerable energetic stability.[cite] Therefore, we suggest that future efforts at both the *a posteriori* analysis and *a priori* prediction of crystal structure should pay particular attention to the role of solvent molecules at crystal interfaces.

Lastly, we show that on the level of direct comparison to experimental data, the R-factor of the simulation-derived structure factor amplitudes with the experimental amplitudes also improves as the simulation improves. However, it is intriguing that even with the eradication of artefacts generated by the inaccurate starting conditions and consequent excellent agreement of RMSD and B-factors to the experimental observables, the R-factor, though smaller, still hovers around 21%. On the one hand, investigation of the difference electron density map between the simulation and experiment can

suggest further avenues of correction for the simulation. On the other hand, the fact that average coordinates and coordinate fluctuations agree so well with experiment indicates the relative sensitivity of the R-factor to even slight discrepancies between model and experiment. Recently Holton et al.[111] investigated the possible underlying sources of the R-factor gap. Further investigation is necessary in this direction.

Section III. Applying Molecular Dynamics of Crystals to Proteins and Nucleic Acids

Chapter 4. *Molecular Dynamics Simulation of Triclinic Lysozyme in a Crystal Lattice*[4]

4.1. Introduction

Molecular dynamics (MD) simulations of protein and nucleic acid crystals are poised to offer significant contributions to two fields: experimental crystallography and computational chemistry. Crystallographic methods have played an immense role in providing detailed biomolecular structural information and have been fundamental in the development of our understanding of the structure-function relationship. At the same time, crystallographic models can display an overreliance on static representations of biomolecular structure, despite the fact that biomolecules are both dynamic and heterogeneous.[112]–[118] Current models for protein function increasingly rely on an ensemble-based view where a statistical distribution of conformations exhibiting fluctuations around energy minima is modified upon binding events. This ensemble-based heterogeneity and dynamic behavior is also present in biomolecular crystals[52], [119]. Efforts in recent years have sought to elucidate and account for these aspects of crystals that are often hidden in the time and space averaged diffraction data.[53], [54], [56], [58], [120]–[122] Molecular dynamics simulations of crystals can contribute to this effort. Past work has shown that MD is in principle capable of accurately reproducing experimental diffraction data while offering a time resolved glimpse of the hidden inner life of crystals.[62]

Simulations of biomolecular crystals also provide an excellent arena for validation of the procedures and force fields used in such simulations.[76], [123] Crystal simulations have a long history[124]–[130], but convergence is slow (as we illustrate below), it can be difficult to model

[4] Reproduced with permission from P. A. Janowski, C. Liu, J. Deckman, and D. A. Case, "Molecular Dynamics Simulation of Triclinic Lysozyme in a Crystal Lattice " *Protein Science,* 2015. DOI: 10.1002/pro.2713. Copyright 2015 John Wiley & Sons.

disordered solvent, and modeling lattice disorder requires simulations that encompass many unit cells. We have developed a methodology for all-atom molecular dynamics of biomolecular crystals employing modern force fields (with explicit solvent and ions) to represent the interactions within crystals.[62], [64], [76]–[78] To help guide future work we have undertaken an evaluation of four modern force fields on the molecular dynamics of a protein crystal. HEWL, an enzyme of 129 amino acids, was chosen as the host crystal, since it is one of the most commonly studied proteins. A number of experimental studies have been carried out to investigate HEWL crystal packing and flexibility via structural and dynamic properties.[131]–[135] Several earlier computational studies focused on conformational differences in solution and in the crystalline environment[60], [136], [137] as well as solvent and ion mobility in crystals[138], [139]. We constructed a supercell composed of 12 unit cells of triclinic hen egg-white lysozyme (HEWL, PDB:4LZT)[140] with explicit solvent. In total we performed more than 9 μs of molecular dynamics sampling of the crystal lattice equivalent to more than 100 μs sampling of the lysozyme monomer. The results offer insight into both the strong and weak points of the current force fields and more generally into the accuracy of results that can be expected from crystal simulations using popular current force fields.

4.2. Results

The crystal supercell was set up as shown in Figure 4-1 and described in detail in Methods. In all we performed the set of simulations shown in Table 4-1. Whenever not specifically identified below, ff99SB and ff14SB refer to the first μs of each simulation (for consistency with the other 1 μs simulations).

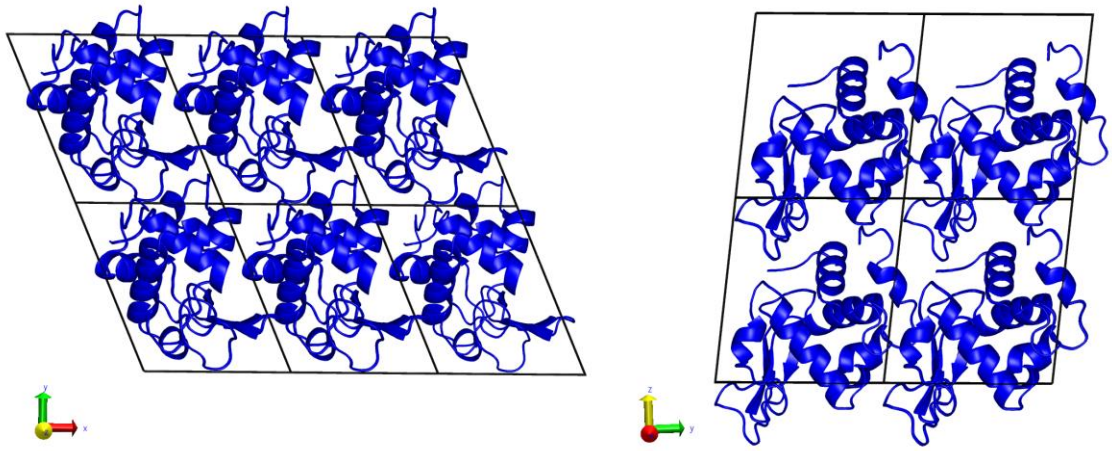

Figure 4-1. Simulation setup of the HEWL supercell. The P1 space group unit cell was extended three times along the crystallographic *a* axis and two times each along the *b* and *c* axes. Addition of solvent is described in Table 4-1.

	Exp.	ff99SB	ff14SB	ff14ipq	C36	ff14SB_solv
Protein molecule	--	12	12	12	12	1
Force field		Amber ff99SB	Amber ff14SB	Amber ff14ipq	Charmm 36	Amber ff14SB
Solvent model	--	tip3p	tip4p-ew	tip4p-ew	tip3p	tip4p-ew
Water (H_2O)	55.5 M	3358	3358	3278	3268	9375
Acetate (CH_3COO^-)	100 mM	39	39	39	39	-
Nitrate (NO_3^-)	250 mM	91	91	91	91	-
Sodium (Na^+)	250 mM	22	22	22	22	8(Cl^-)
Total atoms	--	34265	34265	34025	33995	39468
Density (g/cm^3)		1.281	1.281	1.274	1.273	1.019
Equilib. length (ns)	--	160	160	180	180	28
Production length (ns)	--	3000	3000	1000	1000	1000
Mean press. (bar)	1	129±323	-25±322	-22±304	82±310	0.8±141
Mean trajectory RMSD (Å)	--	0.75/ 1.27	0.65/ 1.15	0.76/ 1.25	0.78/ 1.37	0.95/ 1.47

Table 4-1. Molecular composition and basic statistics of the simulated systems. Last line provides the mean instantaneous backbone atom/all heavy atom RMSD after optimal alignment to the deposited model as a reference over the course of each trajectory.

4.2.1. Structure

The root-mean-square deviation (RMSD) is a measure of structural similarity between two sets of atomic coordinates (Figure 4-2). Following previous work [62], [76], two types of RMSD metric were calculated. "Best-fit RMSD" is calculated by rotating and translating each monomer snapshot to

optimize agreement with the experimental structure. "Lattice" RMSD is calculated by using the crystal symmetry and translation operations to move all snapshots to the same unit cell, but without rotational-translational optimization for a best fit. Thus, best-fit RMSD only includes contributions from intra-molecular fluctuations, whereas lattice RMSD includes both intra-molecular fluctuations and also contributions from the inter-molecular motion of protein monomers relative to each other within the crystal lattice. As expected, crystal simulations give lower RMSD values vs. experiment compared to the solvated molecular dynamics simulation (Table 4-1, Figure 4-2). The closest agreement to the experimental model is obtained with ff14SB.

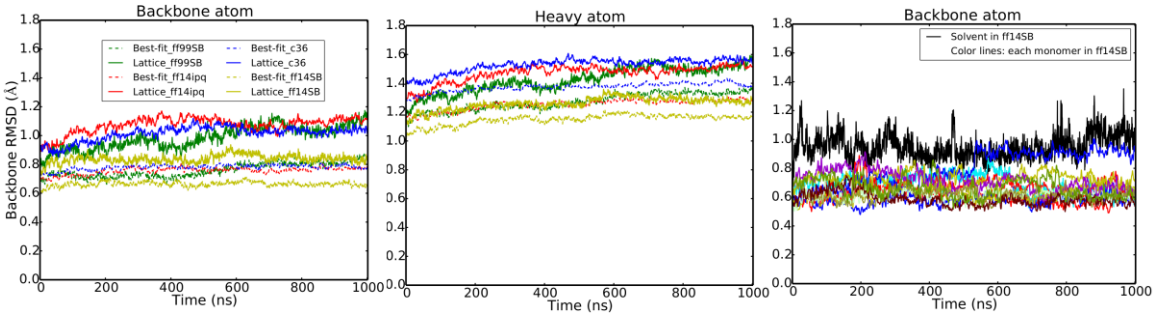

Figure 4-2. RMSD for four different force field simulations and comparison with solution simulation. Left hand panel shows backbone atom RMSD; middle panel shows all heavy atom RMSD. Dotted lines show best-fit and solid lines show lattice RMSD (see text and Ref. [62] **for more details). Black line in right panel represents the best-fit backbone RMSD of the liquid state simulation (ff14SB_solv). Colored lines show the best-fit backbone RMSD of each monomer in the ff14SB crystal simulation. The final 1000 ns of each simulation are shown (first 160-180ns of each simulation were discarded to allow the systems to equilibrate).**

Furthermore, the RMSD results indicate that convergence towards equilibrium in crystal simulations is slower than in typical solvated simulations. Results of several additional microseconds of simulation to test equilibration times and reproducibility of results are presented in the Supplementary Material. In particular, we find that the ff99SB crystal simulation does not reach convergence until more than 1 μs of simulation (Suppl. Material Figure 2). On the other hand ff14SB RMSD converges after about 250ns.

For each simulation we also calculated the average protein structure and its RMSD to the experimental model. Results are shown in Table 4-2. Consistent with previous crystal simulations, RMSDs of the average structures are significantly lower than those from a similar solution simulation and are also lower than the average instantaneous RMSD discussed above. This is not surprising, as the RMSD of an ensemble average must be smaller than the average of the corresponding instantaneous RMSDs;[141] both sorts of statistic are commonly used (here and elsewhere) in characterizing ensembles of structures. The instantaneous RMSD in the solution simulation also deviates more, sometimes making short-lived excursions of about 0.3 Å above the mean value. By comparison (Figure 4-2, third panel) crystal simulations have lower and more stable RMSD without the excursions seen in solution. Among the force fields, ff14SB and ff14ipq simulations most closely reproduce experimental data (0.37/0.79 backbone/heavy atom best-fit RMSD for ff14SB and 0.40/0.77 respectively for ff14ipq). Interestingly, ff14ipq average structure heavy atom RMSD is the lowest of the four simulations even though the instantaneous heavy atom RMSD for ff14ipq (Figure 4-2) was consistently higher. As in previous studies[62], [64] a larger degree of conformational variation is sampled by the molecules at any given moment in the simulation (instantaneous backbone RMSD 0.65-0.78, Table 4-1) even though the average coordinates are much closer to the experimental values (average backbone RMSD 0.37-0.47, Table 4-2).

	Simulation				Refinement			
	Backbone RMSD[1]	**Heavy atom RMSD[2]**	**Elec. density R-factor[3]**	**Map CC[4]**	**R-work[5]**	**R-free[5]**	**Backbone RMSD[6]**	**Heavy atom RMSD[7]**
ff99SB	0.41	0.84	46.4	0.547	18.5	19.9	0.39	0.56
ff14ipq	0.40	0.77	86.3	0.511	22.5	26.0	0.37	0.67
C36	0.47	1.00	83.7	0.515	18.3	19.7	0.46	0.79
ff14SB	0.37	0.79	46.6	0.588	15.1	16.0	0.38	0.79

1 Backbone atom RMSD of the average simulated structure compared to the experimental model.
2 Heavy atom RMSD of the average simulated structure compared to the experimental model.
3 R-factor of the amplitudes from the simulation average electron density and the experimental model.
4 Map correlation coefficient of the simulation average electron density map and the experimental model map after optimal translation using phenix.get_cc_mtz_pdb and phenix.get_cc_mtz_mtz.
5 R-work and R-free statistics after fifteen macrocycles of standard refinement and one round of manual rebuilding of the experimental model into the simulation average electron density.
6 Backbone atom RMSD of the model refined into the simulation average electron density and the experimental model.
7 Heavy atom RMSD of the model refined into the simulation average electron density and the experimental model.

Table 4-2. Average structure and average electron density statistics calculated directly from simulation and after refinement of the experimental model into the simulation average electron density. All RMSD statistics are in units of angstrom (Å).

Simulations of crystals permit comparison directly against observed experimental data. We calculated the average electron density and corresponding structure factors from each simulation (using the asymmetric unit alignment and electron density averaging methods outlined in Ref. [62]). Comparison against the experimental model (Table 4-2) is consistent with RMSD conclusions: ff14SB agrees more closely with experiment. We furthermore refined the experimental model against the average electron density from each simulation by a limited procedure of 10 automated macrocyles of reciprocal space coordinate and isotropic B-factor refinement in *phenix.refine*[91], followed by a limited manual rebuilding in COOT[142] aimed at removing the most flagrant disagreements with electron density, and followed by a further 5 macrocyles of standard automated refinement. The resulting models are arguably the most representative structures of each simulation[62] and avoid the structural artifacts of direct coordinate averaging. RMSDs of the resulting structures (Table 4-2) are consistent with those of the average simulation structures, with backbone atom RMSD varying between 0.37Å and 0.46Å. R_{free} values for the refinements against simulation density vary from 16.0% and 19.9% for ff14SB and ff99SB to 19.7% for C36 and 26.0% for ff14ipq. Interestingly, the

R_{free} statistic obtained for the ff14SB simulation is close to the experimental R_{free} for PDB:4LZT which was 14.7%. It should be noted that the refinements performed here were limited and without refinement of anisotropic atomic displacement parameters. This was done for consistency in order to place the refinements against the density obtained with each of the four force fields on comparable footing. A more exact refinement approach with the use of anisotropic atomic displacement parameters and ordered solvent would likely bring the ff14SB R_{free} very close to the experimental value.

The secondary structure of lysozyme (shown in Figure 4-3) has three β-strands (β1, β2, β3), four α-helices (α1, α2, α3, α4), and four 3_{10} helices (G1, G2, G3, G4). Figure 4-3 shows mean stability of these secondary structure elements during the simulations. All of the force fields consistently maintain the α-helical structures, although there is a uniform tendency to unravel the helix termini, particularly the C-terminus, in favor of turn or 3_{10} helical conformations. The situation is markedly different for the 3_{10} helices. Helix G1 is poorly maintained by all of the force fields, from about 25% of the time with ff14SB to about 10% of the time with C36. Helix G3 is well maintained by all the Amber force fields but unravels about 50% of the time with C36 in favor of a turn conformation. A similar situation occurs with Helix G4, although here there is also a tendency of all force fields to lose the N-terminus (maintained 80% of the time with ff14SB but only 10% of the time with C36). β-sheet structures are well maintained by all of the force fields but most strongly by C36. In summary, α-helices tend to be understabilized at termini (ff99SB and ff14ipq understabilize the most; ff14SB and C36 the least) and all force fields tend to understabilize 3_{10} helices (C36 and ff99SB the most, ff14ipq and ff14SB the least). These results may provide helpful insights into future development of the respective force fields.

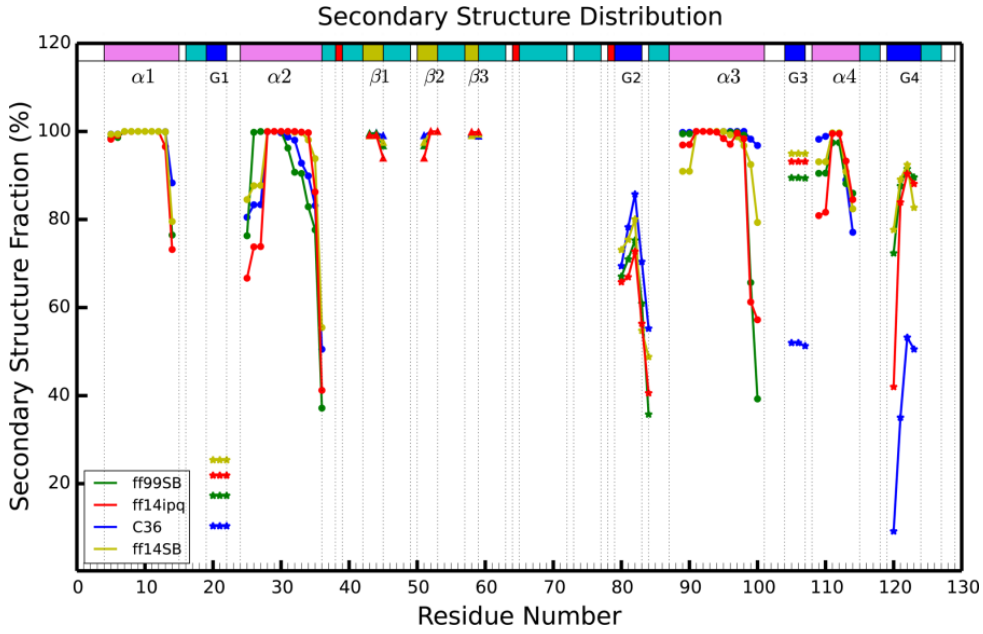

Figure 4-3. Comparison of secondary structure elements during each of four different force field simulations. Residue numbers are on the x-axis and percentage of simulation time spent in a particular type of secondary structure is on the y-axis. Each simulation is represented by a different color and each type of secondary structure is shown by a different geometric figure (α, β, and G are alpha helix, beta sheet, 3_{10} helix, respectively).

4.2.2. Fluctuations

In addition to structural accuracy of the average atomic coordinates in the simulation, it is important to consider the fluctuations around those mean positions. Furthermore, atomic root mean square fluctuations (RMSF) can be directly compared to the atomic B-factors determined during the crystallographic structural refinement. As in previous work[62], [76] we calculated two sets of fluctuations, "best-fit" which account for intra-molecular fluctuations in the atomic positions and "lattice" which also include contributions from inter-molecular fluctuations in the crystal lattice. "Best-fit" fluctuations are calculated by first rotationally and translationally fitting all monomer snapshots to a reference structure to minimize RMSD, finding the average coordinates of that set of fitted snapshots and then calculating fluctuations around that average; this monitors intramolecular atomic movement around a mean position. "Lattice" fluctuations are calculated by first aligning each

supercell snapshot by center of mass and then applying the crystal symmetry and lattice translation operations to bring all monomers into the space of a common asymmetric unit. No RMSD-minimizing rotational translational fitting is applied, thus preserving the contribution of lattice distortion during the simulation.

These two sets of fluctuations are presented in the two top panels of Figure 4-4, and fluctuations derived from refinement of the model against the average electron density derived from the simulation is presented in the bottom panel. Experimental root mean squared fluctuations have been calculated from the deposited B-factors using the relation: $B = \frac{8}{3}\pi^2 * RMSF^2$. Backbone and per-residue RMSF values from all of the simulations correlate modestly (Pearson correlation 0.76-0.85) with the experimental set (Suppl. Material Table 2), above the typical range of 0.5-0.7 previously reported in MD simulations [143]–[150]. Correlations with ff14SB and ff14ipq are slightly higher than with C36 and ff99SB. For example ff14SB and ff14ipq exhibit all heavy atom correlations of 0.79 and 0.78 while C36 and ff99SB have correlations of 0.71 and 0.70 respectively.

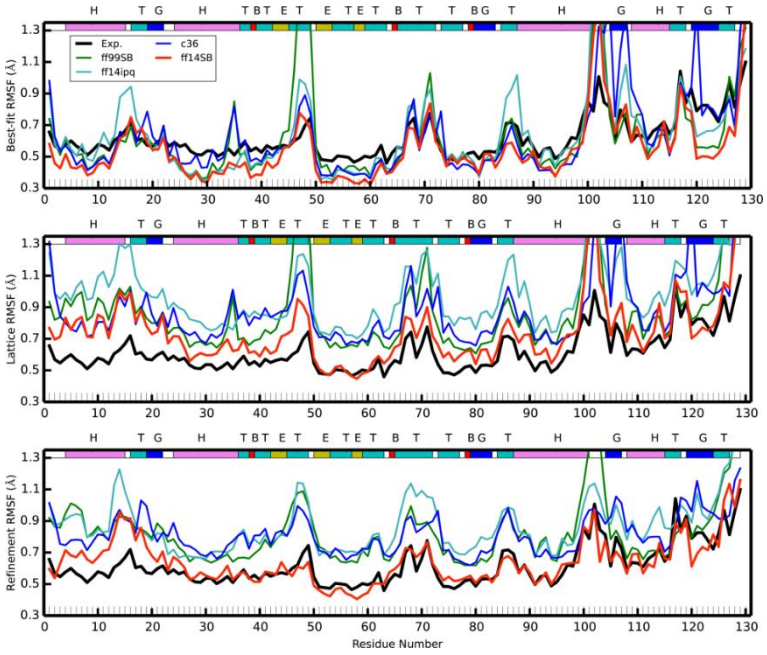

Figure 4-4. Best-fit (top) and lattice (middle) and refined (bottom) Cα carbon RMSF for the four crystal simulations and compared to experiment. Colored lines correspond to each of the four

simulations (red: ff14SB, blue: C36, cyan: ff14ipq, green: ff99SB); black shows the experimental results. The colored band across the top describes the secondary structure (T: turn, E: β-sheet, H: α-helix, G:310 helix, B:isolated bridge).

The best-fit RMSF underestimates the baseline of the experimental fluctuations. This is to be expected as the experimental fluctuations contain contributions from various sources, both static and dynamic disorder, which are eliminated when naively performing the best-fit RMSF calculations. On the other hand, lattice fluctuations for all simulations overestimate the experimental fluctuations. We suggest that this is due to the lattice distortion effect described in the next section of this paper. The ff14SB RMSF are closest to experiment, followed by ff99SB, C36 and ff14ipq. Interestingly, when we calculate lattice fluctuations for each monomer individually (Suppl. Material Fig 10) and average the resulting fluctuations, we obtain results that match experiment very closely (Figure 4-5). It remains to be seen whether this is a coincidental result for this system only or if this will be a consistent result across other crystal simulations as well.

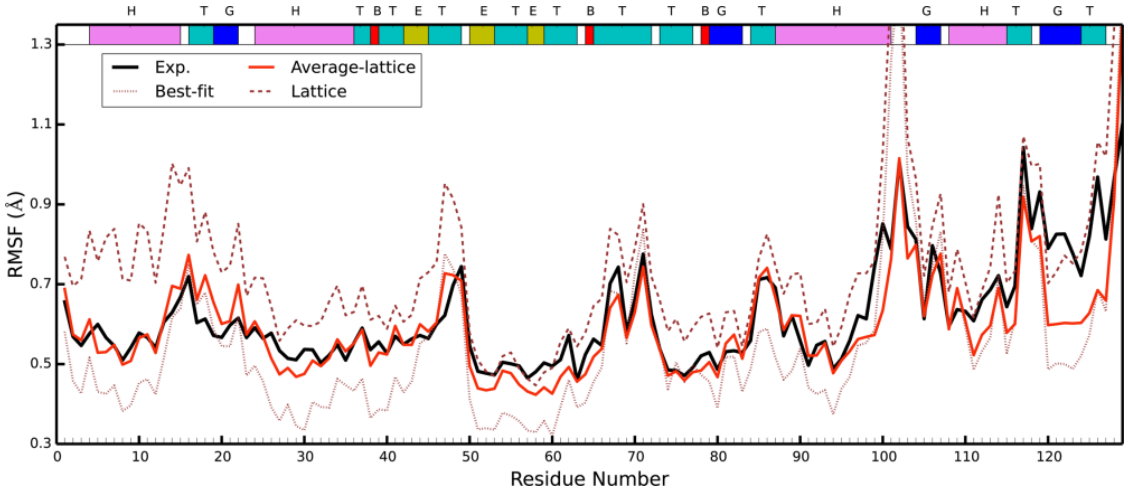

Figure 4-5. The averaged lattice fluctuations from each individual monomer in the ff14SB simulation (shown in red). Lattice RMSF were calculated for each of the 12 monomers and then averaged. Experimental results are shown in black. The best-fit (brown dots) and lattice(brown dashes) fluctuations are those of ff14SB found in Figure 4-4 and are shown here for reference.

Experimental RMSF peaks (regions of high fluctuation) are recapitulated in the simulations, but the RMSF peaks derived from simulation are significantly higher. In part this may be because refined B-factors are known to underestimate atomic fluctuations[108], [151] while the simulations may be revealing the true extent of the fluctuations in the physical crystal. On the other hand force field inaccuracies can lead to structural molecular and lattice instability producing higher than experimental fluctuations. Thus, it may be posited that the true fluctuations in these regions are to be found somewhere between the refined and the simulation-derived values. Fluctuations obtained from refinement against the simulation electron density (Figure 4-4, bottom panel) appear to confirm this conjecture: the ff14SB refined fluctuations are generally lower than the ff14SB lattice fluctuations and in excellent agreement with experimental results, whereas refined fluctuations from the other three force fields, while lower than the corresponding lattice fluctuations, are still higher than the experimental result. Because we find ff14SB to preserve the crystal lattice and structure of the protein with higher integrity than the other force fields (see next section), this does indicate that the higher than refined "real" fluctuations are due to both a limitation of the refinement algorithm[108], [151] and excessive fluctuations resulting from inaccurate force fields. This insight will be treated in more detail in an upcoming publication. Furthermore, we see that all of the fluctuation peaks occur at helix termini (regions around residues 88, 100, 105) or at extended turn loops (around residues 16, 49, 70, 115). As discussed previously, our simulations tend to under-stabilize helix termini; this could lead to the higher fluctuations we observe.

4.2.3. Side-chain disorder

An analysis of side chain disorder (*cf.* Suppl. Material Table 3) reveals that χ_1 angle distributions behave similarly across the four crystal simulations as well as the solution simulation. For each residue we computed the percentage of trans, gauche minus and gauche plus (t, g-, g+ respectively) conformers. About half of the residues display at least some disorder (major χ_1 rotamer population < 99%), but we focused on residues where the major χ_1 conformation was sampled less than 80% of the time (Figure 4-6). The number of these "multimeric" residues was ff14SB: 24, ff99SB: 31,

ff14ipq: 28, C36: 26 and ff14SB_solv: 28. While there is a common set of 56 residues that are not multimeric in any simulation, there are only 9 residues that are consistently multimeric across all four simulations. This could be indicative of insufficient sampling. In all there are 50 unique residues that are multimeric in at least one simulation. Out of these 35 have polar or charged side chains. The share of polar/charged multimeric side chains varies in the simulations and is 66% in ff14SB, 71% in ff99SB, 79% in ff14ipq, 88% in C36 and 75% in ff14SB_solv. Thus, while the number of total multimeric residues is approximately the same, C36 (and to a lesser extent ff14ipq,) tend toward more frequent heterogeneity in charged and polar side chains than in hydrophobic side chains. Only two of the 50 distinct multimeric residues are buried residues (solvent accessible surface area is 0 $Å^2$) and only 11 have an accessible surface area of less than 50 $Å^2$. Thus the great majority of the multimeric residues are surface residues and all but the buried two are involved in contacts at crystal interfaces.

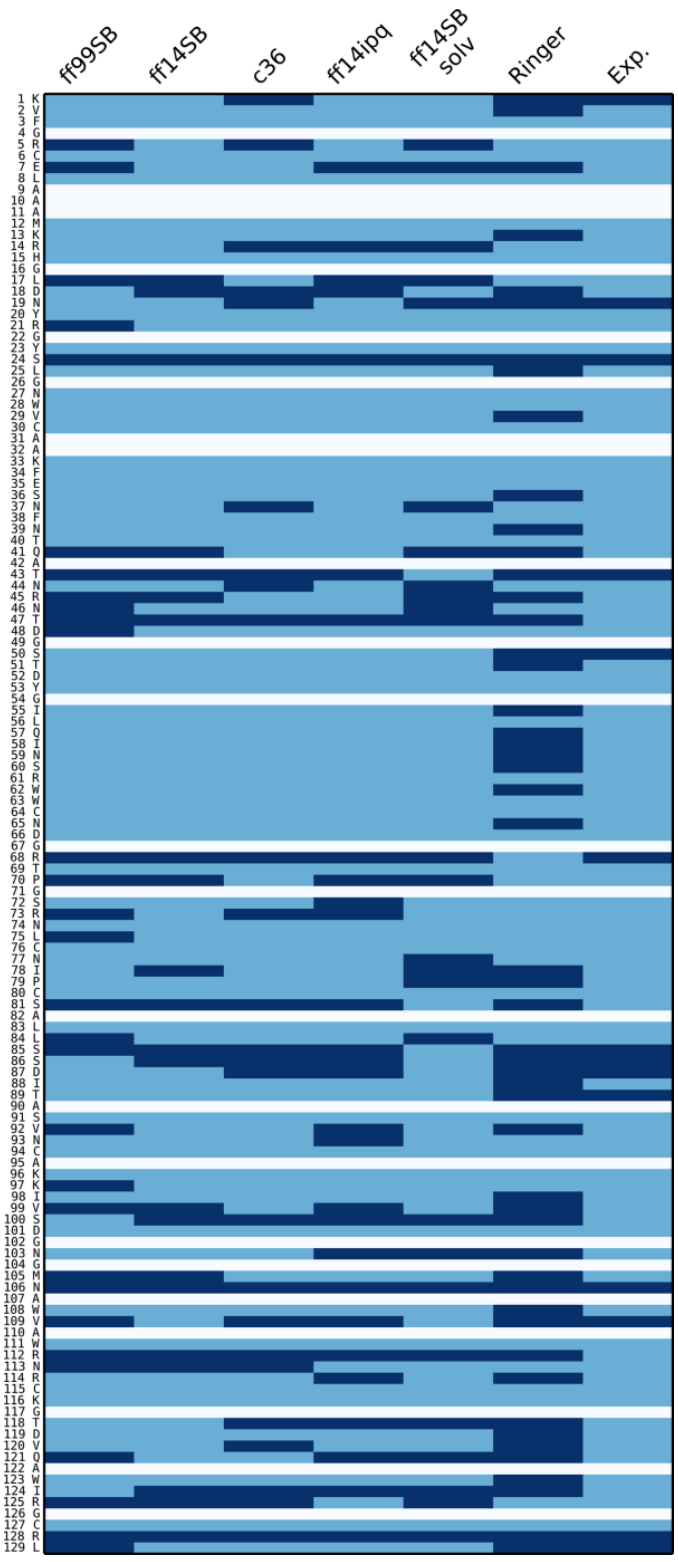

Figure 4-6. χ_1 angle side chain disorder in each of the simulations. In each simulation (first five columns), each residue (rows) is classified as either multimeric (dark blue) or non-multimeric (light

blue). White rows indicate alanine or glycine residues. See text for explanation on the classification method. For the Ringer column, dark blue means Ringer predicted more than one χ_1 rotamer, light blue if Ringer predicted only one. Experimental column is dark blue if the side chain was modeled with an alternate conformer in the 4LZT deposition, light blue otherwise.

Among the crystal simulations, the force fields disagree on the most populated rotamer of a given residue in 22 instances out of a possible 106. However, between ff14SB and ff14SB_solv, we find only 7 residues where major rotamer preferences differ, suggesting that force field differences play a stronger role in rotamer disorder than solvent/crystal environment. We also performed a Ringer analysis[57] of the experimental electron density and model and found weak correlation between the simulation multimeric residues and those identified as containing alternate conformations by Ringer. Out of the 50 multimeric residues, Ringer identifies 26 as having secondary χ_1 peaks. However, of the 56 non-multimeric simulation residues, Ringer finds 19 residues with secondary χ_1 peaks. In other words, in some cases our simulations find side chain disorder that is not supported by experimental evidence and in other cases we fail to find disorder that can be predicted from the experimental data. Again, this could be due to insufficient sampling or force field deficiency.

In summary, a significant amount of side chain rotamer disorder is sampled by the simulations. The χ_1 rotamer disorder is consistent among the crystal simulations, although C36 and ff14ipq tend to sample rotamer disorder of polar/charged residues more frequently. The amount of disorder in the solution simulation does not appear to be higher than in the crystal simulations. Most disordered side chains are charged or polar and almost all lie on the surface of the protein and are involved in crystal contacts within the lattice.

4.2.4. Crystal lattice disorder

We next attempted to characterize the disorder in the crystal lattice. An analysis of monomer movement (Figure 4-7) indicates that all of our simulations exhibit a small progressive deterioration of the ideal crystal lattice. The centers of mass (COM) of the independent unit cells (each containing

a single lysozyme molecule) explore regions close to but slightly away from the location of the COM in an ideal crystal lattice (Figure 4-7, top-left). The mean *instantaneous* distance (in Å) from the ideal position in the crystal lattice, averaged over all the monomers and all snapshots, is 0.31 for ff14SB, 0.42 for ff99SB, 0.53 for ff14ipq and 0.43 for C36. The mean distance from the ideal crystal position of each monomer's *average* center of mass position is 0.20 for ff14SB, 0.31 for ff99SB, 0.47 for ff14ipq and 0.35 for C36. The degree of lattice deterioration appears to be force field dependent (Figure 4-7, top-right), with ff14SB showing the least deterioration and ff14ipq the greatest. Deterioration appears to increase with simulation time and then level off: a comparison of the three microseconds of the ff14SB simulation shows that the mean ASU distance from ideal is 0.20 during the first microsecond, 0.24 during the second microsecond, and 0.23 during the third microsecond (Figure 4-7, bottom left and bottom right). In ff99SB, the mean ASU distance from ideal is 0.31 during the first microsecond, 0.38 during the second microsecond and 0.36 during the third microsecond. The movement of the ASU centers of mass within the lattice appears to be stochastic and does not appear to follow any specific pattern, such as a monotonic movement away from the ideal crystal lattice position. Some ASU's (e.g. #5 and #6 in Figure 4-7 bottom panels) do progressively move away from the crystal ideal, whereas others (e.g. #4) move away and then return during the first and second microsecond of the simulation respectively; still others (e.g. #2) move away first along in one direction and then "swing around" to move away in a different direction.

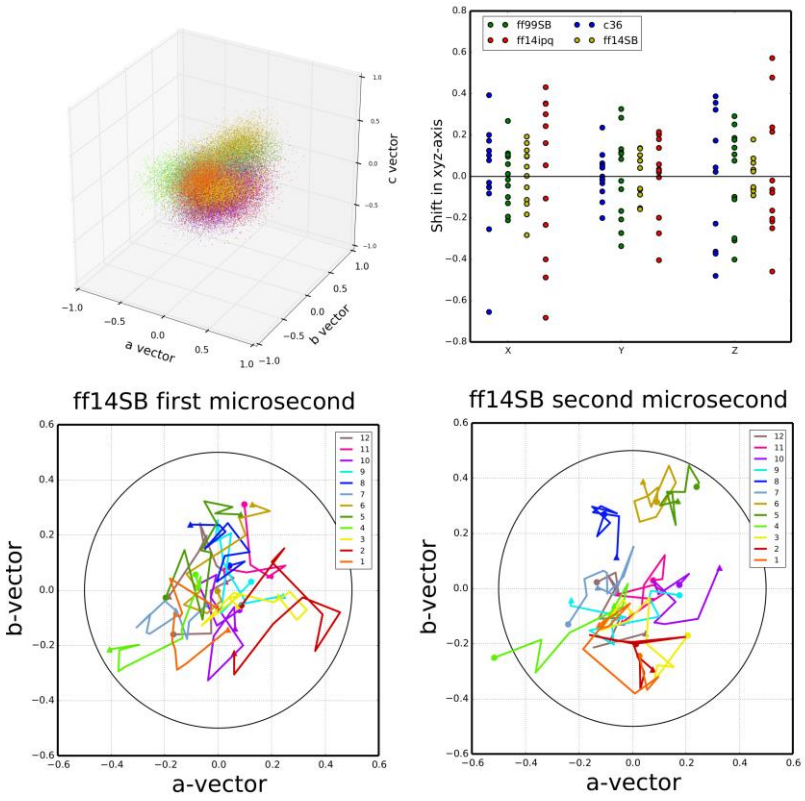

Figure 4-7. ASU center of mass movement relative to ideal crystal lattice positions. Upper-left: cumulative plot of the center of mass of each ASU relative to the ideal lattice position at each time point in the simulation. Points are colored by each independent copy of the ASU in the system (12 independent ASU's). ff14SB is shown, data for the other simulations can be found in the Supplementary Material. Upper-right: Mean distance of each independent ASU relative to the ideal crystal position along each of the crystal system axes (a, b, c). Lower-left: mean position of each ASU's center of mass plotted over intervals of 100ns over the course of the first microsecond of simulation. Starting position (<t=0-100ns>) indicated by a circle and ending position (<t=900-1000ns>) indicated by a triangle. Data shown for the ff14SB simulation; similar plots for the other simulations available in the Supplementary Material. Lower-right: similar plot for the second microsecond of the ff14SB simulation. A circle is drawn at 0.5 Å from the ideal center of mass in both plots.

To further characterize the changes in the crystal lattice during the simulation, we investigated the behavior of the crystal interfaces. Triclinic lysozyme contains 6 unique crystal interfaces

{x+1,y,z}, {x,y+1,z}, {x,y,z+1}, {x+1,y+1,z}, {x+1,y,z+1} and {x,y+1,z-1} which we will refer to here as X, Y, Z, XY, XZ and YZ respectively. There are 12 independent copies of each interface in the simulated supercell. We calculated the relative distance between the centers of mass of lysozyme monomers across crystal interfaces. (see Table 4-3 below and Figure 7 in Supplementary Material). Behavior across the interfaces is variable. In particular, the distance between interfaces X, XY and XZ is on average very close to experiment, whereas interfaces Y, Z and YZ tend to come slightly closer together (by about 0.25Å). These results are consistent for all simulations. However, maintaining the interface distance close to the experimental value may be artificially imposed by the periodic boundary conditions: when one set of monomers move apart, another set must necessarily move closer together. Therefore, it may be more informative to look at the deviations from experiment of individual ASU's (Suppl. Material Figure 7). Here we observe a variety of behaviors between simulations. For example, across the XZ interface, the greatest interface distance change for ff14SB is 0.5Å while for ff14ipq most of the monomers spend most of the time at interface distance changes greater than 0.5Å and the largest deviations are of more than 1.5Å. In general, the greatest fluctuations are observed for interfaces XY and XZ. ff14SB exhibits the smallest fluctuations of all the force fields.

Interface		X	Y	Z	XY	XZ	YZ	Total
Avg. Δdist of monomer COM across interface	ff14SB	0.01	-0.1	-0.08	0.01	0	-0.13	0.33
	ff14ipq	0.01	-0.14	-0.16	0.01	0	-0.33	0.65
	C36	0.01	-0.03	-0.06	0.01	0	-0.16	0.27
	ff99SB	0.01	-0.09	-0.10	0.01	0.01	-0.19	0.41
Avg. no. of contacts	Exp	9	15	14	11	5	4	58
	ff14SB	10.69	14.45	15.69	11.14	7.54	5.10	64.61
	ff14ipq	11.87	18.39	13.76	12.33	6.01	4.61	66.97
	C36	11.02	16.13	13.43	12.31	6.54	4.57	64.00
	ff99SB	11.69	14.74	13.90	11.51	7.69	4.79	64.32
Crystal contacts maintained (strong/weak)	ff14SB	7 (6/1)	9(4/5)	11(5/6)	10(7/3)	5(3/2)	4(4/0)	46
	ff14ipq	8(7/1)	14(7/7)	8(3/5)	9(8/1)	2(1/1)	3(2/1)	44
	C36	9(4/5)	9(6/3)	7(3/4)	8(6/2)	3(2/1)	3(2/1)	39
	ff99SB	9(4/5)	10(5/5)	11(4/7)	9(8/1)	5(3/2)	4(2/2)	48
Crystal contacts lost	ff14SB	2	6	3	1	0	0	12
	ff14ipq	1	1	6	2	3	1	14
	C36	0	6	7	3	2	1	19
	ff99SB	0	5	3	2	0	0	10
New contacts	ff14SB	1	4	3	0	3	1	12
	ff14ipq	4	7	4	0	3	1	19
	C36	2	7	4	3	2	1	19
	ff99SB	2	2	1	0	1	1	7
Interface area	Exp	268	394	379	239	168	113	
	ff14SB	319	347	348	252	248	112	
	ff14ipq	326	396	298	277	196	109	
	C36	314	399	337	266	227	112	
	ff99SB	303	374	323	264	243	109	
Avg. no of H-bonds (crystal/newly formed)	Exp	5	2	3	2	1	0	13
	ff14SB	2.14/0.81	0.14/0.91	0.74/1.65	1.03/0	0.61/0	0	4.66/3.37
	ff14ipq	2.34/1.71	0.29/1.85	0.34/2.68	1.02/0	0.14/0.25	0	4.13/6.49
	C36	1.77/0	0.63/1.35	0.70/1.99	0.84/0	0.36/0.16	0	4.30/3.50
	ff99SB	2.40/0.88	0.78/0.69	0.73/0.93	1.08/0	0.47/0.15	0	5.46/2.65

Table 4-3. Interface behavior relative to deposited model for each of four different simulations.Six columns correspond to each of the six unique crystal interfaces in the model. Each simulation contained 12 independent copies of each interface. The top block shows the relative change in distance between the centers of mass of the interfacing residues. The second block shows the average number of contacts made between unique residues across interface. The third block shows the number of residue contacts from the experimental model that were maintained in each simulation.

This number is further subdivided (in parentheses) into strong contacts (on average found in 10 or more unit cells at a time) and weak contacts (on average found in more than 6 but less than 10 unit cells at a time). The fourth block shows the number of residue contacts from the experimental model that were lost in each simulation (found in less than 6 unit cells at a time). The fifth block shows the number of new interface residue contacts formed in each simulation that were not present in the experimental model ("formed" means that they were present on average in more than 7 unit cells at a time). The sixth block shows the average surface area in each simulation. The last block shows the average number of hydrogen bonds across each interface. The first number corresponds to hydrogen bonds that exist also in the experimental model and the second number to new hydrogens bond formed in the simulation.

Another factor that could account for lattice disintegration could be the inaccurate modeling of crystal contacts and hydrogen bonds across the crystal interfaces. To further characterize the relocation of monomers inside the crystal lattice, we performed a detailed analysis of crystal interface contact residues and hydrogen bonds (Table 4-3; see also Suppl. Material for detailed presentation of all bonds and contacts: Tables 4-7). Use of various cut-off values for assigning contacts yielded similar conclusions. For the results printed here a "contact" is defined as two residues belonging to different ASU's with at least two heavy atoms within 3.2 Å of each other. We note that the average number of contacts per interface, compared to the number of contacts in the experimental structure, is slightly higher in all four force fields, ranging from 64 to 67 compared to 58 in the experimental structure. This includes both contacts found in the experimental structure and new contacts formed during the simulations, indicating that a rearrangement of contacts takes place. We classified the contacts found in the experimental structure into *strong* (found on average in more than 10 of the 12 independent interface copies during the course of the simulation), *weak* (found on average in more than 6 of the 12 interfaces) or *broken* (found on average in less than 6 of the 12 interfaces). We also identified *new* contacts, not present in the experimental structure, if a contact occurred on average in 7 or more of the 12 interfaces during the simulation. We see that ff99SB maintains more crystal contacts than the other simulations (48 vs. 46, 44 and 39 for ff14SB, ff14ipq and C36 respectively).

ff99SB and ff14SB also create fewer new contacts (7 and 12, respectively) than ff14ipq and C36 (19 in both cases). This indicates that ff99SB and ff14SB result in less rearrangement of interface contacts.

On a per-interface basis, results vary and patterns are less clear. For example, ff14SB loses more crystal contacts than the other force fields do at the X and Y interfaces but fewer than the other force fields at the Z, XY, XZ and YZ interfaces. C36 does well at the X interface but loses many contacts at the Y and Z interfaces. In general, all interfaces average more contacts per interface in all simulations than the number of contacts found in the crystal structure. The exception to this is interface Z. In three of the simulations this interface has slightly fewer contacts on average (13.90 for ff99SB, 13.76 for ff14ipq, 13.43 for C36) than the crystal, which has 14 contacts. Only ff14SB has more, with 15.69 Z interface contacts. Of the contacts that are maintained by the simulations, the most stable contacts tend to be hydrophobic interactions, whereas polar and electrostatic interactions tend to be less stable. This could indicate that the force fields model hydrophobicity well, but that some electrostatic-based effects are too weak compared to alternative interactions with waters.

A "hydrogen bond" is defined here as a nitrogen or oxygen with a covalent hydrogen on one of the atoms and a distance between the two heavy atoms of less than 3.2 Å (no angle cut-off was used). For ff14SB, we first analyzed hydrogen bonds within the active site and compared them to those identified from an analysis of the experimental electron density by Held and van Smaalen[132]. We found that the simulation reproduces the same set of hydrogen bonds and with similar relative strengths as those reported in the cited work: reported strong bonds between Ala31 and Glu35, between Asn44 and Asp52 and between the side chain of Asp52 and one of the side chains of either Asn44, Asn46 or Asn59 are all consistently maintained (>75%). A reported weaker bond between Glu35 and Ala110 is also less common in the simulations (<30%). This shows that intra-molecular hydrogen bonds (at least in the active site) are maintained by the force field in a manner that is consistent with experimental density.

On the other hand, crystal interface hydrogen bonds are not stably maintained in the simulation: there are 13 interface H-bonds in the deposited model: 2 across the Y interface, 3 across Z, 5 across X, 2 across XY, and 1 across XZ. In general, H-bond occupancy statistics are remarkably similar across all of the force fields, with the same bonds being the most strongly maintained or most likely to be broken across all simulations; (See Table 3 in Supplementary Material for stabilities of specific hydrogen bonds in the simulations). Of the five hydrogen bond interactions across the X interface, only three in ff14SB and ff99SB (45@O–77@ND2; 114@NH2–18@O; 114@NH1–16@O), two in ff14ipq (45@O–77@ND2; 114@NH2–18@O), and one in C36 (45@O–77@ND2) are preserved more than 50% of the time. In all simulations, the crystal hydrogen bond 45@O–77@ND2 is maintained more than 50% of the time. The other crystal hydrogen bonds across interface Y and Z are not preserved well: all the crystallographic hydrogen bonds are preserved less than 50% of the time. For interface XZ and XY there are fewer crystal hydrogen bonds compared with that in interfaces X, Y and Z. However, the crystal hydrogen bond 116@NZ–77@OD1 across interface XY is generally maintained in all simulations. Furthermore, the same rearrangements of hydrogen bonding are seen to occur in all cases, such as the Y interface Arg21@NH2-Asp66@O breaking in preference of Arg21@O-Arg68@NH1/NH2 with Arg21 switching roles from H-bond donor to acceptor and Asn19@ND2--Ser81@O breaking in preference of Asn19@ND2--Leu84@O and Asn19@OD1--Gln41@NE2. However, this particular rearrangement occurs more frequently in ff14ipq, C36 and ff14SB than in ff99SB. Across the Z interface, all four force fields completely break the H-bond Ser100@OG-Leu128@NH1 found in the experimental model but involving the terminal Leu128. However, Phe3@O-Arg73@NH1, is almost completely broken in ff14ipq and C36, but is more strongly maintained in ff99SB and ff14SB (respectively 35% and 37% of the time). On average ff14SB and ff99SB tend to maintain more of the hydrogen bonds found in the experimental model (on average 4.66 and 5.46 experimental H-bonds vs. 4.13 and 4.30 for ff14ipq and C36 respectively), and to create fewer new H-bonds not found in the experimental model (on average 3.37 and 2.65 new bonds vs. 6.49 and 3.50 new bonds for ff14ipq and C36 respectively; Table 4-3).

Nevertheless, the results do not allow us to draw definite conclusions about H-bond behavior that could explain the varying degrees of crystal lattice degradation observed in the different force fields.

4.3. Discussion

Our simulations of a triclinic lysozyme crystal with explicit solvent reproduce experimental structural results well, both in regards to atomic mean positions and fluctuations. In terms of atomic RMSD, ff14SB performs particularly well (0.37/0.79 Å backbone/heavy atom RMSD), but none of the crystal simulations produce RMSD deviations of more than 0.50/1.00 Å backbone/heavy atom. These results are encouraging considering that this degree of structural divergence is on par within the deviations seen between independent crystal structures of triclinic lysozyme. For example the backbone RMSD between PDB:3LZT/4LZT is 0.28 Å[140] and between PDB:1V7T/4LZT it is 0.37 Å[152]. These results are maintained even for simulations up to 3 μs in length. Thus, even if one were to possess a "perfect" force field, it might not necessarily produce smaller structural deviations. Atomic fluctuations are also generally consistent with experiment, with Pearson correlations ranging from 0.77 to 0.85 for backbone atoms. Correlations are slightly better when including the effect of both dynamic (intramolecular) and static (lattice) disorder. Direct comparison of average electron density against experimental measurements and refinement reinforces these conclusions: ff14SB and ff99SB perform slightly better than the other two force fields and refinement results produce similar RMSD results (e.g. 0.37/0.67 Å backbone/heavy atom for ff14SB). Fluctuation obtained from refinement against the simulation electron density are lower than the "true" lattice fluctuations in all force fields, but in particular show excellent agreement in the case of ff14SB, possibly indicating a lower degree of artifactual disorder in this simulation. The improved performance of the newer ff14SB is encouraging in that is implies that force field development is progressing in the right direction. The ff14SB force field differs from earlier Amber force fields (such as ff99SB) in terms of torsion preferences of certain side chains. It is likely that these side chain torsional preferences are important in yielding structures that more consistent with the crystal density. Studies on a wider variety of proteins would be needed to establish this is a general trend. Lastly, a slight unraveling of

helix termini is common to all of the force fields. It is possible this is a physical phenomenon that is masked in the electron density by experimental error or averaging effects, but the systematic nature of the small differences that we see, which extend to almost all 3_{10} and α-helical segments, suggests limitations in the force fields used here as a more likely contributor.

We noted above that crystal simulations of proteins are not new. In 2000, Stocker *et al.* compared simulations of lysozyme in solution and in an orthorhombic crystal.[137] It is representative of improvements in computer speed and dynamics algorithms that the earlier results (which had four monomers in a single unit cell) were carried out for 2ns, compared to 3 μs in the present study. The differences between solution and crystal are remarkably similar to those seen here (compare Figure 1 of Ref. [137] to Figure 4-2 here), but improvements in force fields are also evident: the instantaneous Cα atom deviation of the earlier crystal simulation from experiment was about 1.3 Å, compared to 0.7 Å here for ff14SB. A 20 ns study of tetragonal lysozyme[138] showed Cα atom deviations ranging from 1.1 Å for Amber ff03, to 1.6 Å for OPLS-AA, to 4.0 Å for GROMOS96 and most likely would have been higher had those simulation been extended to sample time scales on par with the current results. It remains to be seen how much of the difference between those studies and this one stems from differences in the packing of the crystal space group (previous studies used tetragonal; we use triclinic). Nevertheless, in this sort of structural comparison, there is a clear trend in going from GROMOS96(43A1)[153] [developed in 1996], to ff99SB[30] [developed in 2006] to ff14SB [developed in 2014]. The ff14SB results for fluctuations are also in remarkably good agreement with B-factors refined from a room-temperature crystal study, as shown in Figure 4-4 and Figure 4-5.

Our results also provide information on limitations of current MD force fields. *First*, some atomic fluctuations are too high compared to experimental results. These fluctuations correspond to regions of the structure that are solvent exposed and involved in crystal contacts. It is known that refined B-factors tend to underestimate the true atomic fluctuations[108], [151], but large differences between individual asymmetric units, deterioration of secondary structure and changes in the crystal lattice indicate that structural instability during the simulation also contributes. *Second*, secondary

structure analysis also indicates that fluctuation and structural differences can be attributed to inaccurate modeling of hydrogen bonds. In particular it may be that 3_{10} helices are understabilized by C36 and ff14ipq. *Third*, we observe a slight but progressive distortion of the crystal lattice that grows as the simulations progress, but that is actually quite small, especially for the best performing force field (ff14SB average ASU center of mass 0.20 Å from ideal lattice position). This deterioration is not affected by system pressure or small variations in the amount of solvent. A rearrangement of contacts and bonding networks across the crystal interfaces occurs during the simulations, but no clear correlation between that and the degree of lattice deterioration was discovered. Further analysis of the factors contributing to this lattice distortion, such as the implications of the size of explicit supercell, as well as a potential implementation of crystal molecular dynamics that restrains monomer center of mass to idealized crystal positions are two possible future areas of investigation. Such an approach would complement recent efforts at obtaining accurate MD trajectories by means of electron density based restraints[154].

As we noted above, times required to reach equilibrium are longer than those typically required for solution simulations. This is not surprising considering the somewhat constrained nature of the crystal lattice that hinders solvent rearrangement. On the other hand, it does not appear that more conservative equilibration schemes (Suppl. Material Figure 1) using longer (up to 500 ns) heating and restraint protocols lead to different results. However, crystal simulations allow for independent sampling of multiple unit cells, which enhances the sampling of protein configurations. In the past computational resources often restricted crystal simulation studies to single unit cells[60], [136]. The current approach can help re-examine those findings and identify possible artifacts resulting from periodic boundary conditions imposed on a single unit cell. Further methodological investigation is needed to find the best way to harness and consolidate this information.

One of the most exciting potential contributions of crystalline MD is that it provides a detailed synthetic data set for probing crystal refinement applications. Refinement of the lysozyme structure against the observed simulation electron density yields an R-free factor that is on par with

experimental results (16.7% ff14SB refinement without alternate conformations vs. 14.7% experimental result), but it remains to be seen whether the same factors are responsible for that similar level of disagreement. A more detailed analysis of this "R-factor gap"[111] will be presented in a separate publication.

4.4. Methods

4.4.1. Preparation of the simulation supercell

Atomic coordinates were taken from Protein Data Bank[16] entry 4LZT[140]. This structure of hen egg-white lysozyme was solved in a triclinic P1 space group at 295 K. Alternate conformations were removed, in each case keeping only the major conformer. His15 was set to the protonated state consistent with its experimental pK_a of 4.5-4.6[140]. A "supercell" of 3 x 2 x 2 unit cells measuring 81.72 x 63.74 x 68.46 Å and containing 12 copies of the lysozyme molecule was created by using the *PropPDB* module of the AmberTools[80] package (Figure 4-1). Solvent conditions followed the strategy described earlier[77], [78]: we retained all of the experimentally determined solvent positions (except for minor alternate conformers) which included 134 water, 7 nitrate and 3 acetate molecules. We used the AmberTools *AddToBox* program to add additional acetate, nitrate and sodium ions to both neutralize the system and replicate crystal liquor concentrations. Three of the additional acetates and one nitrate were placed in the positions identified in the cryogenic structure (PDB:3LZT[140]). Test simulations of about 10ns in the NPT ensemble after equilibration were performed in order to find the amount of solvent that best matched the experimental volume of the crystal. Details of the simulations are given in Table 4-1.

4.4.2. Molecular dynamics simulations

Protonation of the protein and construction of molecular topology and coordinate files for the crystal supercell were done using the *tleap* module of AmberTools and *Reduce*.[81] Acetate ions were modeled with parameters derived using the IPOLQ method.[155] Nitrate ion parameters were taken from Ref. [156]. All other parameters were taken from the corresponding force field's standard

parameters. The force fields used were ff99SB[30], ff14SB, ff14ipq[32] and CHARMM 36[157] (C36). The TIP4P-Ew parameters[48], [158] were used for the water model with the corresponding Joung/Cheatham parameters set for Na^+ ions.[159]

System optimization, equilibration, and production dynamics were performed using the PMEMD module of Amber14[26]. When the system volume was allowed to vary (during equilibration only), constant pressure was maintained by a Berendsen barostat[36] with isotropic pressure scaling and a time constant of 1.0 ps. Constant temperature was maintained with a Langevin thermostat[37] (collision frequency of 1/ps) at the experimental crystal diffraction temperature of 295 K. Force calculations were performed with a 9.0 Å real space cutoff in the context of periodic boundary conditions, smooth particle-mesh Ewald electrostatics[45], [85] and a homogeneity assumption for long-range van der Waals contributions. The SHAKE[39] and SETTLE[38] algorithms were used to constrain the lengths of bonds to hydrogen and the internal geometry of rigid water molecules, respectively. A 2 fs timestep was used.

To test the amount of solvent necessary to replicate experimental volume, the equilibration scheme of Ref. [62] was used, followed by approximately 10 ns of unrestrained dynamics propagated in the isothermal/isobaric ensemble. Volumes over the trajectory were compared to experimental volume and the systems shown in Table 4-1 were chosen for production.

For equilibration, non-crystallographic solvent positions were relaxed via 100 steps of steepest descent optimization followed by 900 steps of conjugate gradient optimization with 256 kcal/(mol-Å^2) position restraints applied to protein and crystallographic solvent molecules. Next the conformations of protein residues, including added hydrogens, were relaxed using the same minimization algorithm and with restraints applied to all solvent molecules. A third round of coordinate optimization followed in the same manner but with no restraints. Next, initial restrained dynamics were performed at constant volume for 1 ns with 10 kcal/(mol-Å^2) restraints on all protein, acetate and nitrate atoms, as the system was heated to the experimental temperature of 295K.

Restraints were then relaxed with 4 ns of 10 kcal/(mol-Å^2) restraints on the same atoms, 6 ns of 1 kcal/(mol-Å^2) and 12 ns of 0.1 kcal/(mol-Å^2) restraints. Unrestrained dynamics were then propagated in the NVT ensemble with a two femtosecond timestep for 1140-1160ns. Only the final 1000ns of each simulation were used for the analysis presented here.

A parallel solution simulation placed a single monomer in a box of 9375 TIP4P-EW water and 8 Cl^- ions in a rectangular box of dimension 60 x 67 x 74 Å. We performed 28 ns of equilibration, with gradually decreasing constraints on the protein atoms, followed by 1 μs of production simulation, using a 1 fs time step, a Langevin thermostat with a collision time of 1 ps^{-1}, and a weak-coupling barostat with a time constant of 5 ps.

4.4.3. Analysis of data

Data analysis was carried out using a combination of in-house scripts and the AmberTools cpptraj[110] module. Two root mean square deviation (RMSD) metrics referred to here as "best-fit superposition RMSD" and "lattice-fit RMSD" were calculated using the Kabsch algorithm[86], and two B-factor metrics were calculated as described below and in greater detail in Ref. [62]. Secondary structure was determined using the DSSP[160] algorithm. Simulation average electron density was calculated as described in detail in Ref. [62] using md2map, part of the crystal simulation analysis toolkit in AmberTools and making use of CCP4 programs.[94] The maps were truncated at a resolution of 0.95 Å, corresponding to the experimental result. The Visual Molecular Dynamics (VMD) program[92], PyMOL[161] and matplotlib[162] were used for visualization and image generation.

Refinements against the simulation average electron density were carried out via 10 automated macrocyles of reciprocal space coordinate and isotropic B-factor refinement in phenix.refine[91], followed by a limited manual rebuilding in COOT[142], followed by a further 5 macrocyles of standard automated refinement. To ensure consistency of results and eliminate possible contributions stemming from differences in refinement protocols and software, we repeated a refinement of the

lysozyme model against the original experimental structure factor amplitudes (deposited in the PDB) using a similar protocol. Results were in very close agreement to the deposited model (backbone RMSD 0.04, B-factor Pearson correlation 0.97, Rfree difference .003, see also Suppl. Material Figure 15). Thus details of the refinement protocol play a minor role, and we have chosen to make all comparisons below against the deposited model.

4.5. Supplementary material

Results of extended equilibration protocols, convergence of longer simulations, crystal simulation reproducibility and ff12SB simulations. Analysis of individual monomers. B-factor correlations. Rotamer and hydrogen bond populations. Raw simulation trajectories are available directly from the authors upon request.

4.6. Acknowledgements

We acknowledge NIH grant GM103297 for financial support. PJ acknowledges Rutgers University Presidential Fellowship for support. We thank James Holton and Thomas Terwilliger for many helpful comments. The authors declare no conflict of interest.

Chapter 5. *All-atom crystal simulations of DNA and RNA duplexes*[5]

5.1. Abstract

Background: molecular dynamics simulations can complement experimental measures of structure and dynamics of biomolecules. The quality of such simulations can be tested by comparisons to models refined against experimental crystallographic data. *Methods:* we report simulations of DNA and RNA duplexes in their crystalline environment. The calculations mimic the conditions for PDB entries 1D23 [d(CGATCGATCG)2] and 1RNA [(UUAUAUAUAUAUAA)2], and contain 8 unit cells, each with 4 copies of the Watson–Crick duplex; this yields in aggregate 64 μs of duplex sampling for DNA and 16 μs for RNA. *Results:* the duplex structures conform much more closely to the average structure seen in the crystal than do structures extracted from a solution simulation with the same force field. Sequence-dependent variations in helical parameters, and in groove widths, are largely maintained in the crystal structure, but are smoothed out in solution. However, the integrity of the crystal lattice is slowly degraded in both simulations, with the result that the interfaces between chains become heterogeneous. This problem is more severe for the DNA crystal, which has fewer inter-chain hydrogen bond contacts than does the RNA crystal. *Conclusions:* crystal simulations using current force fields reproduce many features of observed crystal structures, but suffer from a gradual degradation of the integrity of the crystal lattice. *General significance:* the results offer insights into force-field simulations that test their ability to preserve weak interactions between chains, which will be of importance also in non-crystalline applications that involve binding and recognition.

5.2. Introduction

RNA and DNA molecules play an important role in many biological processes, and an understanding of their structure and dynamics is indispensable for a complete understanding of their function. Molecular dynamics simulations can offer a detailed complement to experiment, and

[5] Reproduced with permission from C. Liu, P. A. Janowski, and D. A. Case, "All-atom crystal simulations of DNA and RNA duplexes." *Biochimica et Biophysica Acat*, vol. 1850, no. 5, pp. 1059-1071, 2014. Copyright 2014 Elsevier.

nucleic acid simulations in a crystal environment have long been used to test simulation methods[163], [164]. A "modern" era began in the mid 1990s with the introduction of Ewald-based methods to simulate long-range electrostatic interactions.[165] At about that time, simulations of Watson–Crick paired duplexes in both solution and in the crystal lattice provided evidence that simulations remained stable for a few nanoseconds without the need for artificial restraints[166]–[168], and that, not surprisingly, the duplex structure in the crystal lattice environment more closely resembles the experimental crystal structures than do simulations in a solution environment. Subsequent studies in a crystalline lattice employed longer time scales and different force fields, reaching broadly similar conclusions.[169]–[174] Advances in computing speed have spurred a new round of biomolecular crystal simulations (mostly for proteins)[62], [76], [77], [108], [138], [139], [154], [175], that use larger simulation cells and pay increased attention to the properties of the crystal lattice, as well as to the structural characteristics of individual chains.

Here we report results of molecular dynamics simulations of duplexes of DNA (PDB ID: 1D23[176]) and RNA (PDB ID: 1RNA[177], [178]) for 2.0 and 0.5 μs; both crystals are in the P212121 space group, with four duplexes per unit cell. The periodic unit in the simulations is a "supercell" containing 8 unit cells, so the simulations contain 32 copies of each duplex. Parallel simulations of a single duplex in water (neutralized by Na+ counterions) are reported for comparison to the crystal simulations. Since a large number of solution simulations of DNA and RNA helices have previously been reported[165], [179]–[186], our primary emphasis here is on the properties of the crystal lattice and the ability of current simulation methods to reproduce such behavior. The simulation results explore the details of stability of the duplexes and crystal packing effects in μs dynamics; this time scale has been explored in earlier solution simulations of DNA duplexes.[184], [185]

The main results of this paper are as follows: first, the average duplex structure from the crystal simulation more closely resembles the experimental crystal structure than the average structure in a solution simulation. Some details of the crystal structure, such as groove widths, are averaged out in

solution simulations but not in the crystal, whereas other sequence-dependent variations are maintained. These latter include the characteristic alternating pattern of base pair roll/twist in alternating A–U sequences in RNA and the BI/BII pattern in the backbone torsion angles ($\varepsilon - \zeta$) in DNA[187]–[189]. As in previous simulations, the backbone conformation presents more of a challenge to simulations than does base positioning. However, for both DNA and RNA, the integrity of the crystal lattice is slowly degraded, and a full equilibration is not achieved. The interchain contacts that hold these lattices together are less extensive than is typical for protein crystals, and are likely to be influenced by ions in ways that are poorly represented here. A partial "melting" of the crystal lattice takes place at seemingly random points in the supercell, such that the 32 equivalent interfaces become rather heterogeneous, with some quite close to their original (experimental) configurations and others are displaced in ways that are irreversible on the time scales sampled here.

5.3. Models and computational methods

Coordinates for the initial state of the RNA and DNA came from PDB entries: 1RNA and 1D23. Crystal structures of the RNA and DNA crystal were constructed according to space group P212121, where four symmetry-related copies of the duplex make up one unit cell. We constructed a supercell of 2×2×2 unit cells by using the PropPDB module of the Amber12 package[190], yielding an orthogonal box of size 68.22×89.22×98.22 Å for 1RNA, and 77.86×79.26×66.60 Å for 1D23. Views of the supercells are shown in Figure 5-1. For the DNA crystal simulation, Mg^{2+} ions were used as counterions: there are two $Mg(H_2O)_6^{2+}$ complexes per asymmetric unit in the 1D23 crystal structure, and 7 additional $Mg(H_2O)_6^{2+}$ ions per duplex were added to neutralize the 18 phosphate groups. The RNA supercell system was neutralized with 26 Na+ ions per duplex. All counterions and water molecules were added to the supercell using the AddToBox program in Amber. The DNA system for the solution simulation contained one DNA duplex and 5045 water molecules together with 18 Na^+ ions in a truncated octahedral box with periodic boundary conditions; the RNA simulation had 7694 waters and 26 Na^+ ions. Further simulation details are given in Table 5-1.

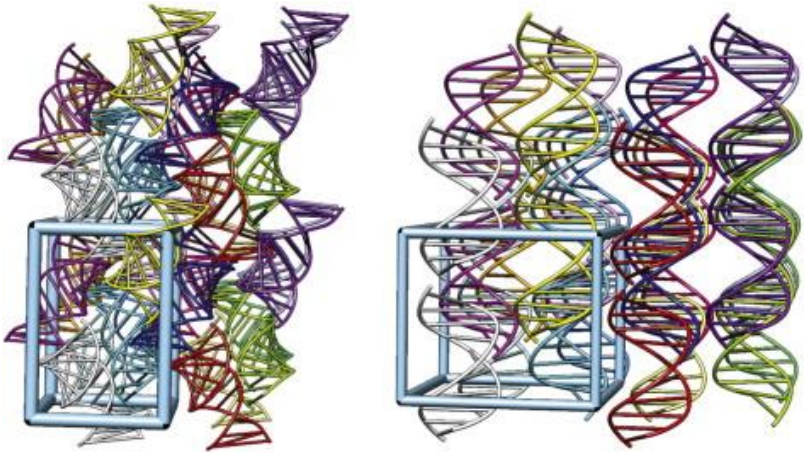

Figure 5-1. Two three dimension views of the RNA (left) and DNA (right) packing in supercells. The supercell contains 8 unit cells in a 2 × 2 × 2 arrangement; each unit cell comprises four RNA/DNA molecules. The cyan box represents one unit cell. Image generated using Chimera.

	DNA	RNA	DNA (soln)	RNA (soln)
Number of waters	8578	12 300	5045	7694
Number of ions	288	832	18	26
Total number of atoms	46 246	65 892	15 785	23 988
Mean internal pressure, atm	30	97	1.0	1.0
Density, g/cm3	1.433	1.446	1.044	1.034
Simulation speed, ns/day	16.7	21.5	24.3	26.9
Time analyzed, μs	2.0	0.5	0.5	1.0

Table 5-1. Details of the simulations. Simulation speed is measured on a single NVIDIA GTX780 card for the RNA simulations, and on a Quadro K5000 card for the DNA simulations. Ions are Mg^{2+} for the DNA crystal and Na^{+} for the other simulations. Atom counts do not include the "extra points" in the TIP4P/Ew waters.

We carried out crystal simulations using the parm99 force field[83] with bsc0 modifications[191] for both duplexes, and added the chiOL3 modifications[192] for RNA. The water model used was TIP3P[82] for DNA and TIP4P-Ew[48] for RNA; monovalent ion parameters were taken from Ref. [159] and Mg^{2+} parameters from Ref. [193]. Simulations of nucleic acids in solution show only small conformational dependencies on the water model used. As an example, we compared our TIP4P-Ew

simulation of 1RNA in solution with results from Ref. [180], where a TIP3P water model was used. Average value twist, roll, helical twist, propeller twist, slide, x-displacement and helical rise were essentially the same in the two simulations. We do not yet have similar comparisons for crystal simulations.

The minimization, equilibration, and production dynamics were performed using the pmemd module of Amber12[190], [194], [195]. The SHAKE algorithm was applied to all bonds involving hydrogen atoms. Force calculations were performed with periodic boundary conditions, a 9.0 Å cutoff on real space interactions, a homogeneity assumption for long-range Lennard–Jones contributions, and smooth particle-mesh Ewald electrostatics. An energy minimization procedure was used first to remove any bad contacts in the starting conformation. Next, the system was equilibrated at the experimental data collection temperatures of 308 K for RNA and 273 K for DNA, with successive restraints on the RNA/DNA atoms of 10.0, 1.0 and 0.1 kcal/(mol-Å^2) for a total of 40 ns. The volume was kept fixed at the experimental value, and the system pressure monitored. The number of water molecules was adjusted by trial and error to obtain a simulation with an external pressure near 1 atm. Finally, unrestrained production dynamics were propagated at a 2 fs time step for 2.0 μs for DNA and 0.5 μs for RNA. A total of 4000 equally-spaced snapshots for DNA and 2500 for RNA were saved for subsequent analysis. Some details of the simulation are collected in Table 5-1. The solution simulations followed the general procedures of the "ABC" simulations[181]–[183], [196] [27], [28], [29] and [46], and placed a single duplex in a truncated octahedron, neutralized by Na^+ ions, which were equilibrated in a similar fashion.

Results were analyzed using Amber cpptraj module, and figures were prepared by Chimera and VMD molecular visualization programs. Helical parameters throughout the trajectories were monitored with the Curves+ program.[197] The BI and BII configurations in DNA were characterized by the angles ε and ξ of the DNA backbone or by the angle difference (ε − ξ), which is − 90° for BI and + 90° for BII phosphates.[189] Interfaces were analyzed using the PISA program.[198]

5.4. Analysis of duplex structures and comparison with solution simulations

The root mean square deviations (RMSDs) from the crystal structure for all heavy atoms (of all 32 duplexes) are shown in Figure 5-2. For the curves labeled "best-fit", an optimal translation and rotation movement to fit the crystal coordinates was determined for each duplex in each snapshot; these values reach an apparent plateau of about 1.2 Å in about 200 ns of simulation. For the curves labeled "lattice", the duplexes were superimposed on the crystal configuration using only the lattice symmetry parameters, with no fitting (see Ref. [62] for a detailed explanation of these statistics). These latter values reflect both deviations of individual duplexes from the crystal configuration and the degradation of the lattice itself. The latter continues for the entire trajectory, and will be analyzed in Section 4.

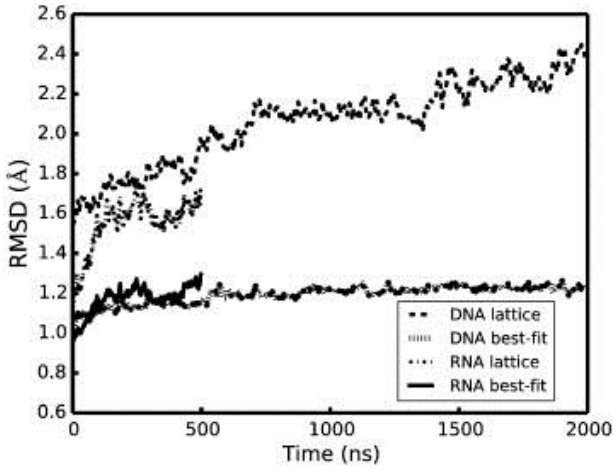

Figure 5-2. Positional RMSDs of all heavy atoms for RNA and DNA relative to the initial X-ray structure in the course of simulation.

The best-fit and solution average structures are compared to the experimental crystal configuration in Figure 5-3, and some statistics are given in Table 5-2. As expected, and as seen in earlier crystal simulations, the average structure in solution deviates much more from the crystal configuration than does the average structure from the crystal simulation. The crystal simulation

deviations are larger in the backbone for both DNA and RNA than for the bases. Some details of these differences are provided in the following sections.

Figure 5-3. Superpositions of the solution average structure (orange) and the best-fit average structure (blue) for RNA (left) and DNA (right) versus the deposited crystal structures (green).

	All heavy atoms	Backbone	Base	All heavy atoms (solution)
Best-fit (RNA)	0.78 (0.71)	0.87	0.59	1.38 (1.12)
Lattice (RNA)	0.93	1.04	0.74	~
Best-fit (DNA)	0.77 (0.64)	0.89	0.56	1.83 (1.34)
Lattice (DNA)	1.12	1.23	0.95	~

Table 5-2. Root mean square deviations (Å) from the deposited crystal structures for 1RNA and 1D23. The statistics in each box are heavy atoms, backbone atoms, base atoms RMSD of average structure against the crystal structure. The values in parentheses exclude the terminal residues.

5.4.1. Groove widths

The A- and B-form helices for RNA and DNA have very different geometries, and can be monitored as groove "widths" and "depths". In Figure 5-4, we use a simple measure groove width, based on phosphate–phosphate distances from one chain to its complementary strand. For RNA, the sequence dependence of the widths of the major groove is suppressed in solution, leading to a more

regular pattern of distances than are seen in the crystal. Figure 5-3 and Figure 5-4 show that the major groove widens slightly compared with crystal structure. This behavior of solution simulations has been discussed earlier, in connection with attempts to characterize how NMR constraints determine the details of A-form helices in RNA.[199] Furthermore, fluctuations of the major groove in the crystal simulation are much smaller than those in the solution simulation.

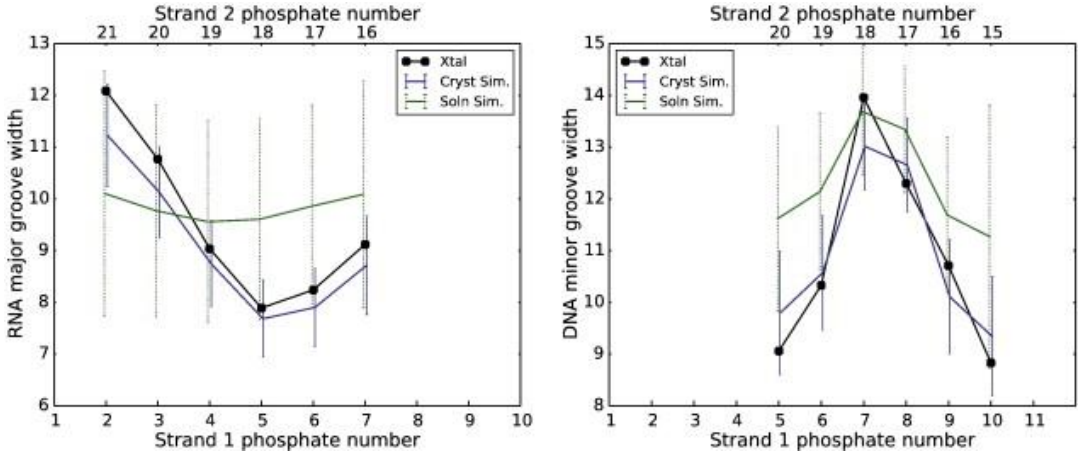

Figure 5-4. Plots of major-groove width for RNA and minor-groove width for DNA in crystal and solution simulations. Widths in Å are defined by the distance between the phosphate atoms shown at the top and bottom. Vertical bars give the standard deviations of the fluctuations seen in the simulations.

For DNA, the minor groove width is the greatest at phosphates P7/P8, not only in crystal structure but also in the crystal and solution simulations. This is related to the presence of a BII phosphate conformation at positions P7 and P17 in the crystal structure. The larger minor groove widths at positions P10 and P20 in the solution are related to the larger fluctuations in the terminal residues that are characteristic of solution simulations. In addition, the wide minor groove at P7 is consistent with the large positive roll deformation (Figure 5-7, below) which compresses the major groove and then widens the minor groove.[200] (Because of the way steps are numbered, a wide minor groove separation between P7 and P18 should be correlated with a positive roll angle at position P5.)

5.4.2. Backbone conformations

The backbone of the 1RNA structure contains two kinks (at A23/U24 and U10/A11), which divide the whole duplex into three roughly equal regions. The kinks result from changes in the α and γ backbone torsions from their typical values near 280 and 70, respectively. Table 5-3 lists the average values in both simulations, showing a reversion in the simulation back to the canonical A-type values. This reversion suggests that crystal packing forces are not sufficient (at least for the simulation analyzed here) to retain the unusual crystal configuration.

	α (cryst.)	**α (cryst. sim. ave)**	**α (solution sim. ave)**	**γ (cryst.)**	**γ (cryst. sim. ave)**	**γ (solution sim. ave)**
A11	155	279 (29.9)	281 (15.1)	178	73 (23.7)	64 (12.6)
U24	210	282 (12.9)	282 (12.4)	127	64 (9.5)	63 (9.4)

Table 5-3. Average α and γ angles (degrees) in crystal and solution simulations. The data in parentheses are standard deviations.

The DNA backbone generally has two major sub-forms, BI and BII, that can be characterized by the ϵ and ξ backbone torsion angles.[188], [189], [201] The range of ϵ and ξ torsion angles for BI structure varies from 120° to 210° and from 235° to 295°, respectively, while for BII these vary from 210° to 300° and from 150° to 210°, respectively. The difference (ε − ξ) is near + 90° for BII and near –90° for BI. In the crystal structure, P2, P7, P12, and P17 have a BII conformation, and all others have a BI conformation. Figure 5-5 shows the fraction of BII conformation in the crystal and solution simulations. Both P7 and P17 largely maintain a BII conformation, whereas P2 and P12 revert to a mixture of BI and BII conformations. The percentages of the BII conformation at P2, P7, P12 and P17 are much smaller in the solution simulation.

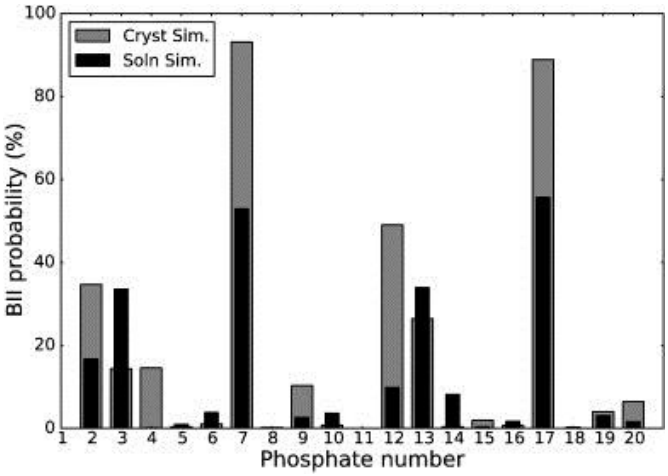

Figure 5-5. Conformational substates (BII) probability in the crystal and solution simulation along the sequence: P2 to P10 are in strand 1 and P12 to P20 are in strand 2. In the crystal structure, P2, P7, P12, and P17 have a BII conformation, and all others have a BI conformation.

5.4.3. Base pair step parameters

Nucleic acid helices are commonly characterized by six helicoidal parameters (tilt, roll, twist, shift, slide, and rise) that describe the configuration of base steps[202]; simulation results are given in Figure 5-6 and Figure 5-7. There are a number of observations that can be made. First, the shift and tilt parameters are evened out in solution for RNA; this is much less true for DNA, and is less true for the other base-step parameters. Solution and crystal simulations are in close agreement for the rise, roll and helical twist parameters, especially for DNA. The fluctuations vary little between steps in the crystal simulation, except for the more flexible shift, rise, tilt and twist of the end base pair A13·U16/A14·U15 step. Second, fluctuations in base step parameters in solution (green dashed bars) are much greater than those in the crystal simulation (blue solid bars).

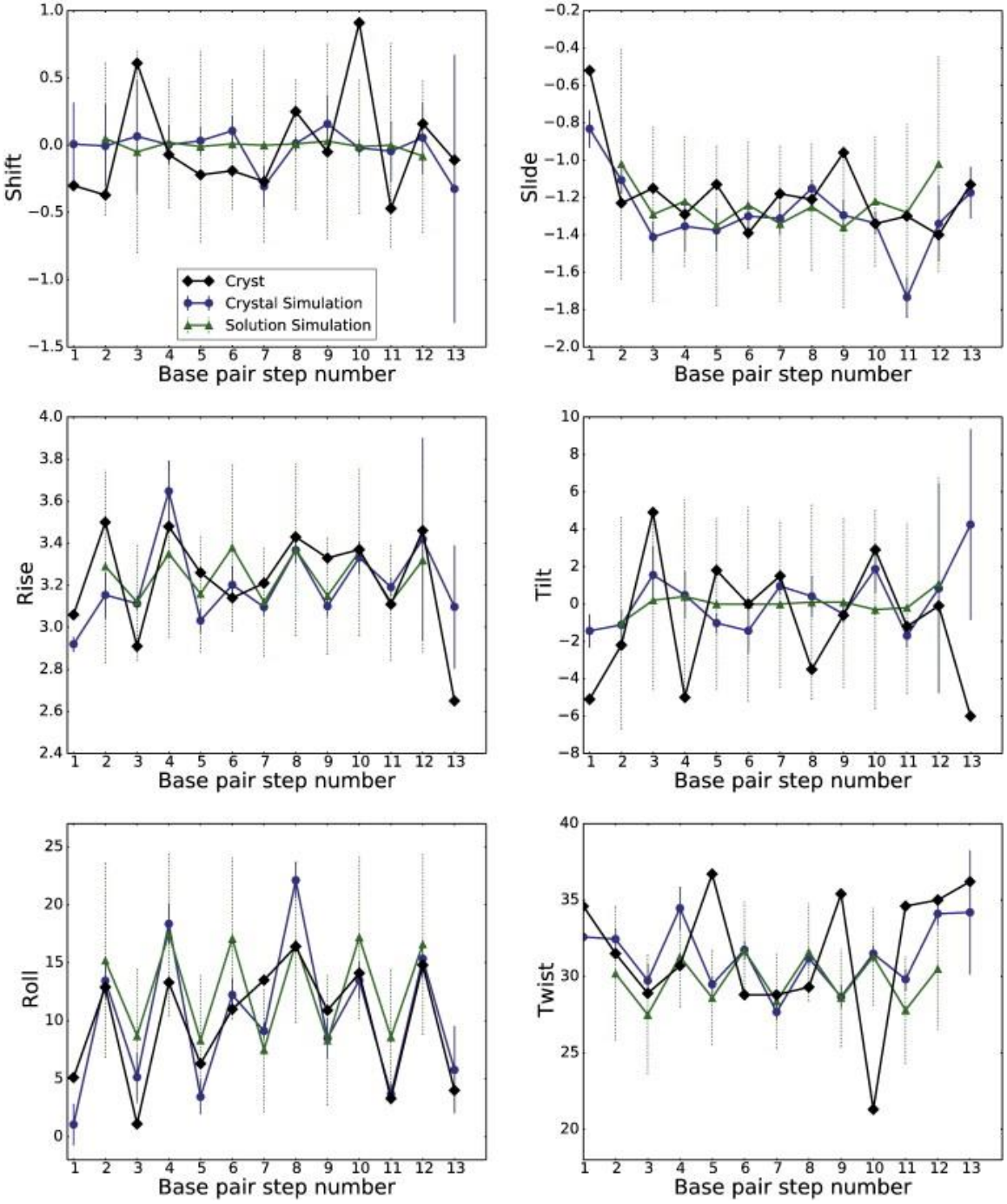

Figure 5-6. RNA base pair step parameters, translational parameters are in angstroms (Å) and rotational parameters are in degrees (°).

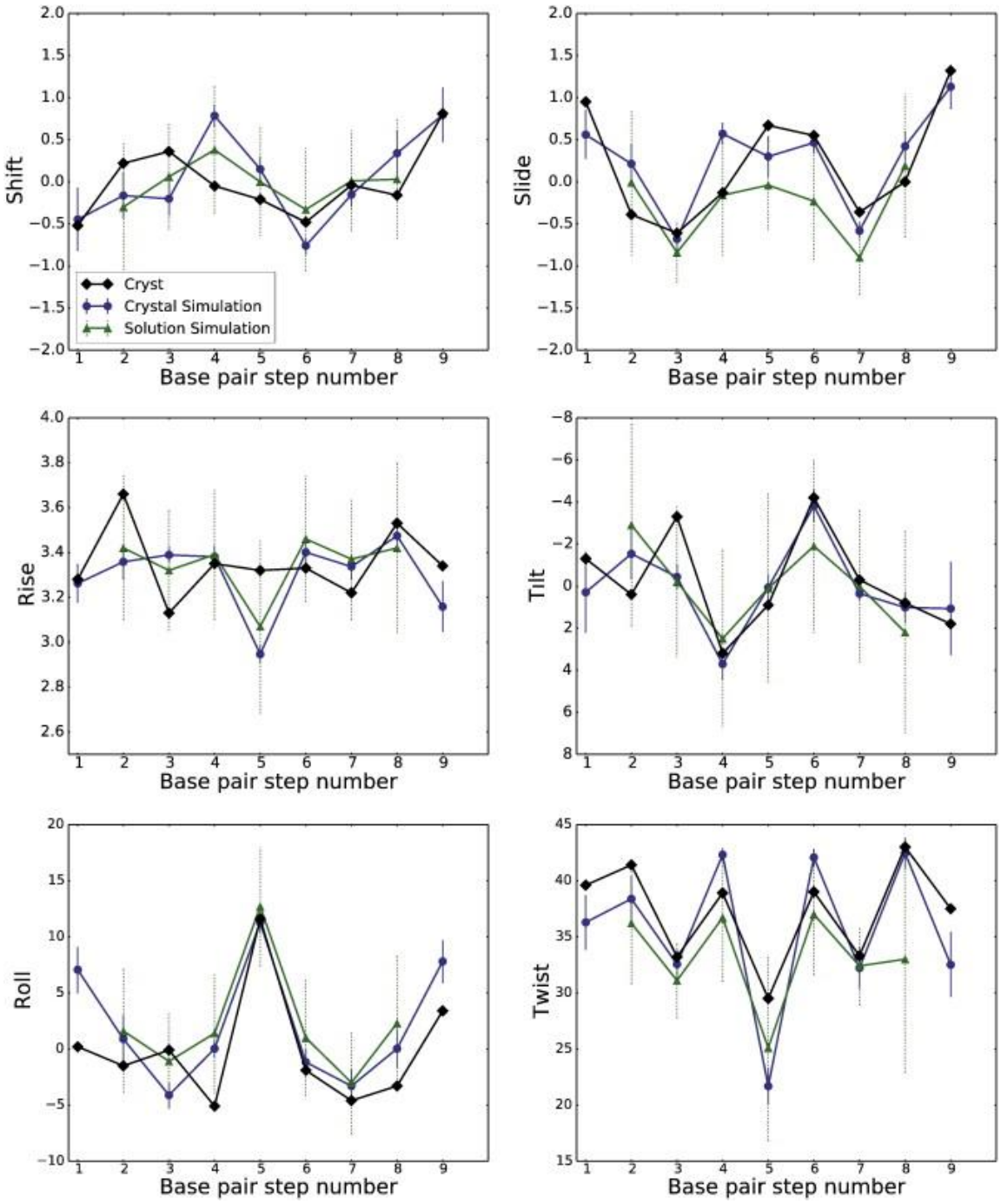

Figure 5-7. DNA base pair step parameters, translational parameters are in angstroms (Å) and rotational parameters are in degrees (°).

The twist of nucleic acid helices is an "emergent" property of force fields, that is difficult to associate with particular interactions or backbone torsion angle preferences. The Amber potentials

lead to slightly undertwisted helices in solution simulations.[186], [203] There are two unusual twist values (36.7 and 21.3) in the RNA duplex at steps 5 and 10 shown in Figure 5-6, which is caused by two distinguished kinks dividing the whole duplex into three parts. However, in our simulation, these two unusual twist values in kink regions become 29.5 (28.7 in solution) for step 5 and 31.6 (31.5 in solution) for step 10 in average structures. Table 5-4 shows that the average twist is larger in the crystal simulation than in solution; it seems likely that the coaxial stacking interfaces (discussed below) serve to prevent the relaxation to an undertwisted state that is typically seen in solution simulations with this force field[165], [179]–[183], [186]. Some steps in the solution simulations show evidence of bimodal distributions in the twist[204]; no such behavior is seen in the DNA simulations reported here. This may be due to sequence or crystal-packing effects, and further analysis of these points will be a subject of future study.

	RNA	**DNA**
Exp. crystal	31.7	37.3
Sim. crystal	31.4	35.7
Sim. solution	29.8	32.3

Table 5-4. The average twist (in degrees) for the RNA and DNA average structures. Twist values were calculated using the program Curves +.

5.4.4. Fluctuations about the average structure

Simulated B-factors can be calculated from an MD trajectory using (Eq. 20) and compared to crystallographic B-factors:

$$B = \frac{8\pi^2}{3}\langle \mu^2 \rangle \qquad \textbf{(Eq. 20)}$$

where μ is the root mean square fluctuation about a mean position. In Figure 5-8, the simulated and experimental B-factors for RNA and DNA are reported. For both solution and crystal simulations, the fluctuations are calculated by superimposing each duplex onto the crystal conformation using an

optimal rotation and translation movement. The phosphate groups show the highest deviations from the average structure, and the low points between these peaks represent base atoms.

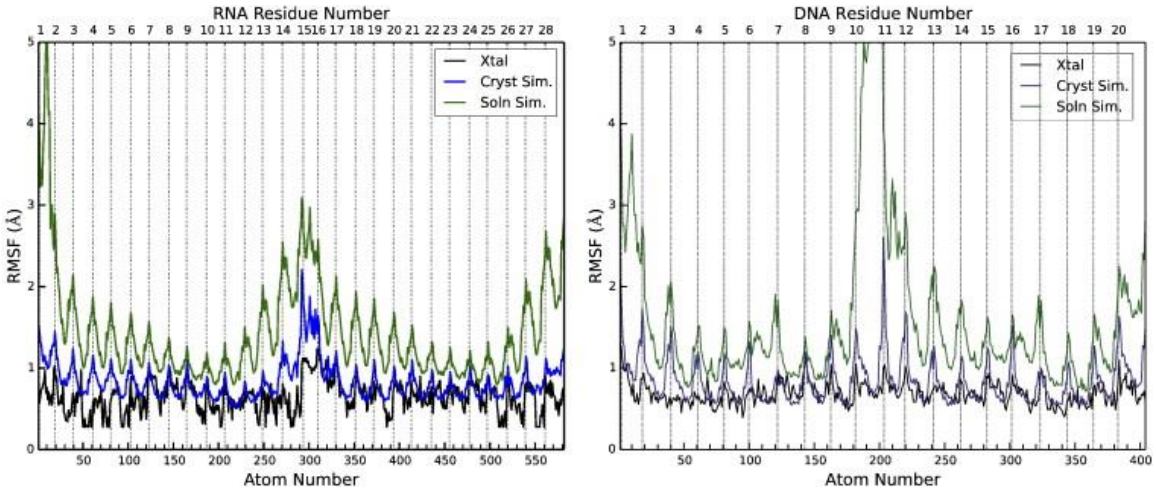

Figure 5-8. Root-mean-square fluctuations as function of heavy atom number for RNA (left) and DNA (right). The left half of each figure represents chain A, and the right half, chain B. Light vertical lines identify the phosphate atoms (excluding residues U1A15 for RNA and C1C11 for DNA). B-factors in 1RNA/1D23 (PDB entries) were converted to fluctuations using (Eq. 20).

As expected, the fluctuations in solution (green curves) are higher than those in the crystal simulation (blue curves), not only at the chain termini (where large fluctuations and fraying take place in solution), but also in the centers of the duplexes. For RNA, the fluctuations in the crystal simulation are close to experiment except near the chain termini. This is also the case for the base atoms in the DNA (i.e. at the low points between peaks in Figure 5-8), but the sugar-phosphate backbone atoms have larger fluctuations in the simulation. Note that the fluctuations reported here do not include the effects of the lattice distortion, which is fairly large for both RNA and DNA, as shown in Figure 5-2 and Figure 5-15; inclusion of these effects would result in fluctuations that are much larger than those shown in Figure 5-8 (data not shown).

5.5. Contacts and packing interactions

Two of the features that distinguish the present simulations from earlier ones are the size of the supercell used and the length of the simulations. We have 32 copies of each RNA or DNA duplex, and hence 32 copies of each of the interactions between duplexes. This allows us to analyze the nature of these interactions, and the ability of current force fields and simulation protocols to correctly describe the behavior of these relatively weak interactions.

5.5.1. Crystal interfaces for RNA

In a crystal, each nucleic acid molecule makes several contacts with its neighboring molecules. Generally, the size of the pairwise interface between two neighboring molecules can be characterized by the number of contacts or the buried interface area. PDBe PISA (Proteins, Interfaces, Structures and Assemblies) shows three interfaces with nearest neighbors for RNA, shown in Figure 5-9 and Table 5-5. Two interfaces (1 and 2 in Figure 5-9) involved in the minor grooves of the symmetry related helices face each other at their ends; hydrogen bonds stabilize both contacts. The third interface places the major grooves of two duplexes face to face with each other. A few van der Waals contacts and CH–O hydrogen bonds (among residues U1, U22 and A9, U10, U15) stabilize this interface. The total surface area buried by the three listed interfaces is 1516 Å^2 (twice the total surface area of the three interfaces listed in Table 5-5), since in the crystal each duplex participates in two independent copies of each interface, (e.g. one on the "left" and one on the "right"). The total surface area for an isolated duplex is 5134 Å2, so that 29% of the surface area is buried, and 71% is accessible to water and ions. This buried interface area is smaller than for many proteins: for example, the triclinic form of hen lysozyme (PDB code 4LZT) has 3030 Å2 (or 46%) of its surface area in contact with other proteins. The relatively small contact area for RNA (and for DNA, as discussed below) may increase the likelihood that the contacts may degrade during an imperfect simulation.

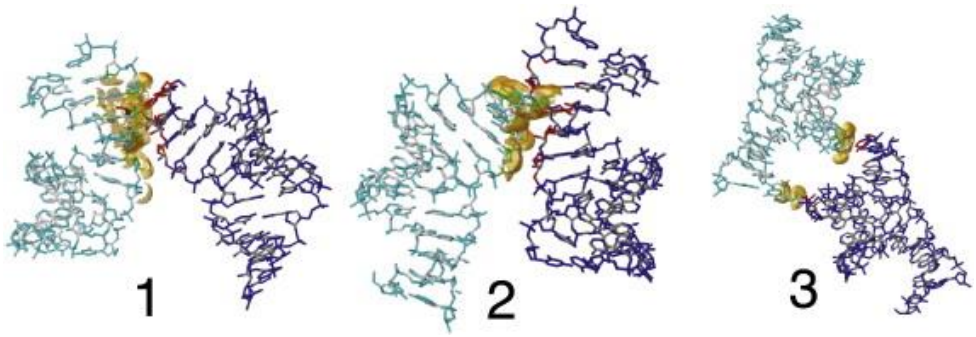

Figure 5-9. Lattice contacts, showing the orientation of one RNA asymmetric unit relative to its neighbors. The contacts are shown as defined by PISA.

Type	Interactions	Area (Å2)	Atom 1	Symmetry operator	Atom 2	Xtal. dist.	Cryst. sim. dist.
1	Minor groove–minor groove	311.5	O2/U1	−x+1, y−½, −z+½	O2'/A11	3.01	2.78 (0.20)
			O2'/A28		O2/U18	3.02	2.80 (0.19)
			O3'/A28		O2'/U18	3.00	3.10 (0.23)
			O2'/U1		OP1/U12	3.30	4.28 (0.55)
			O2'/U2		O2'/U10	2.46	4.02 (0.80)
			O2'/A3		O2'/U20	3.20	3.86 (0.86)
2	Minor groove–minor groove	304.3	O2/U15	x−½, −y+½, −z	O2'/A25	2.67	3.31 (0.96)
			O2'/A14		O2/U4	2.77	2.99 (0.39)
			O3'/A14		O2'/U4	3.01	3.69 (1.36)
3	Major groove–major groove	138.8	C6/U15	−x+½, −y+1, z−½	O2'/U22	3.17	4.44 (0.94)
			C5/U15		O3'/U22	2.80	3.75 (0.69)
			O3'/A9		O5'/U1	3.53	4.10 (1.11)
			O3'/A9		C5'/U1	3.30	4.26 (1.10)
			OP1/U10		C5'/U1	3.32	4.19 (1.21)

Table 5-5. Hydrogen bonds, van der Waals contacts and interactions between symmetry-related duplexes in RNA.The data in parentheses are standard deviations. Interface areas were computed by PISA.

The first and second interfaces for RNA involve hydrogen bond interactions, whereas the third interface for RNA and all interfaces for DNA are characterized by less-specific van der Waals contacts. Most of the contact distances become larger in the simulation for both RNA and DNA;

exceptions include most of the hydrogen bonds in RNA interfaces 1 and 2; the longer (and presumably weaker) CH–O interactions in RNA interface 3 are less-faithfully maintained; this interface also has a much smaller buried surface area in the crystal.

The averages shown in Table 5-5 hide a large amount of heterogeneity among the 32 copies of each interface present in the supercell. As an example, Figure 5-10 shows the variation of two of the hydrogen bonds in interface 2 for each copy. It is clear that these interactions are well-maintained in about 22–24 copies, and completely broken in 8–10 copies. It is likely that longer simulations would result in more copies becoming distorted, leading to progressive distortion of the lattice, as is seen more clearly in the longer DNA simulation discussed below. The average distance of the molecular centers of mass of the duplexes interacting across interface 1 increases by 0.5 Å; across interface 3 the distance decreases by about the same amount; across interface 2 it stays almost the same as in the crystal (Figure 5-11). Individual interfaces can deviate from the average by up to about 1 Å, indicative of significant heterogeneity.

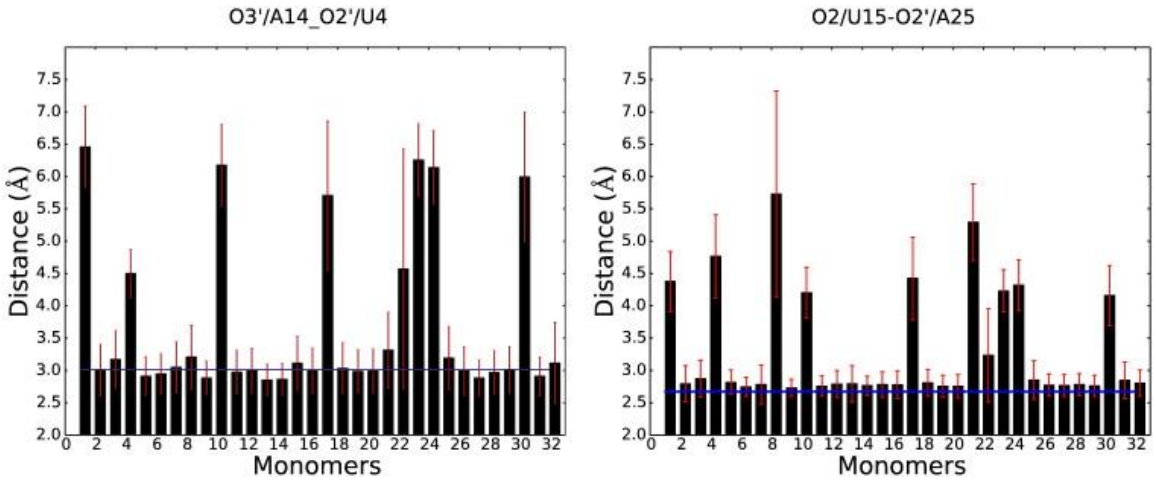

Figure 5-10. The average distances of O3'/A14–O2'/U4 and O2/U15–O2'/A25 for the 32 duplexes in RNA supercell for interface 2 (red lines are standard deviations, blue lines are the distance in crystal structure, and black bars give the average distances for every duplex).

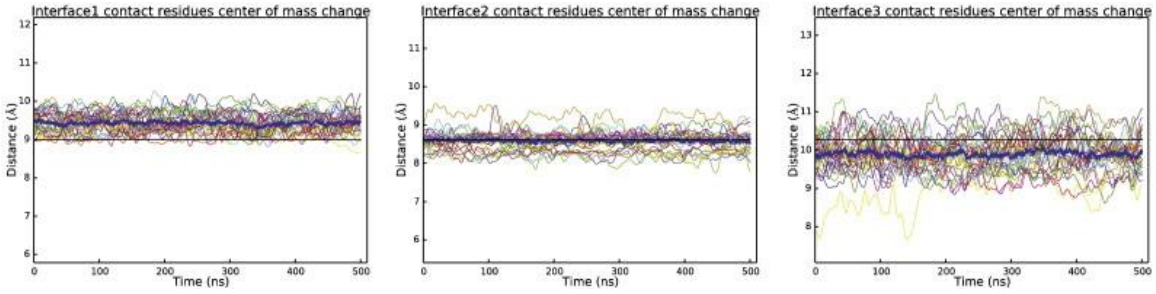

Figure 5-11. Distances of center of mass between the interface residues for RNA, for each of the 32 copies of each interface. Black lines show the value from the X-ray structure. Bold blue lines are the average distances between interface centers of mass.

5.5.2. Crystal interfaces for DNA

The five interfaces for DNA are shown in Figure 5-12 and Table 5-6. In interface 1 the decamer helices are stacked one on top of another along the z-axis leading to co-axial stacking. The second and third interfaces are along the x axis with P3, P4 facing P17, P18, and P8 and P9 facing P12 and P13, respectively. For the fourth and fifth interfaces, P6 and P16 are in close proximity to the inter-helix gaps of neighboring duplexes. The surface area buried by these contacts is 1218 Å2, which is 31% of the total surface area of an isolated duplex.

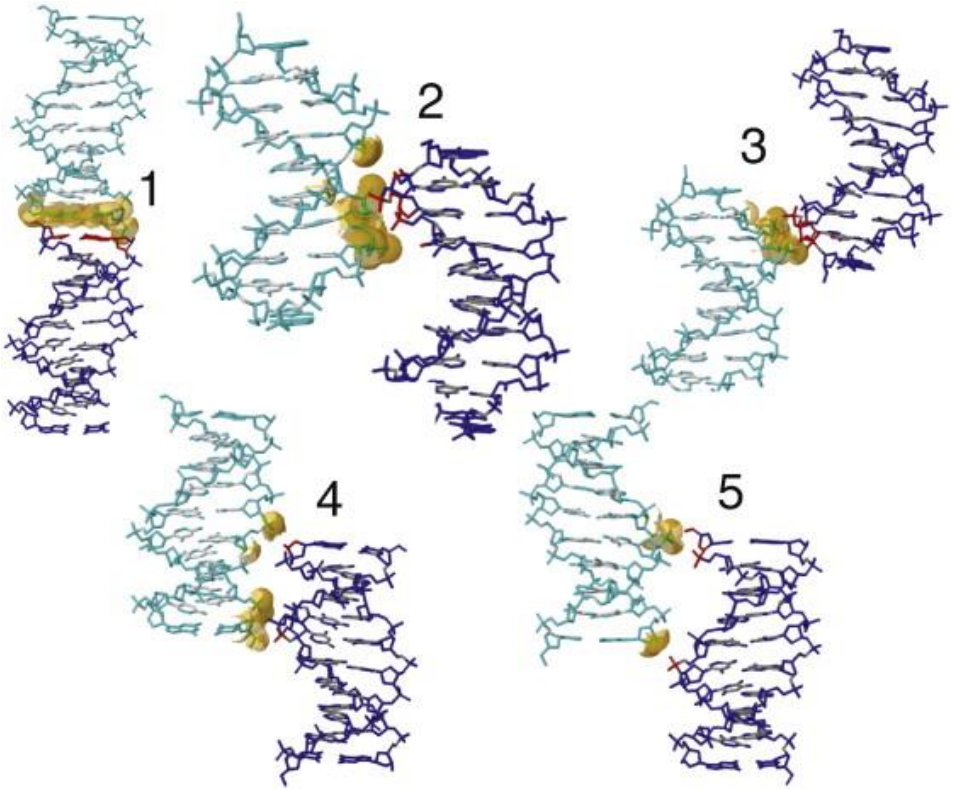

Figure 5-12. Same as Fig. 9, but for DNA.

Type	Interactions	Area (Å2)	Atom 1	Symmetry operator	Atom 2	Xtal. dist.	Cryst. sim. dist.
1	Terminal–terminal	199.9	N3/C11	x, y, z−1	N3/G20	3.49	3.51 (0.23)
			O3'/G10		O5'/C1	3.37	3.63 (0.89)
2	Backbone–minor groove	154.2	OP1/A3	x−½, −y+½, −z	O4'/A17	3.65	4.77 (1.08)
			P/A3		P/A17	6.49	7.59 (1.53)
			P/T4		P/A17	4.65	5.79 (1.93)
			OP1/T4		OP2/A17	3.88	5.83 (1.33)
			OP2/T4		O5'/A17	3.79	4.62 (0.96)
3	Backbone–backbone	102.7	OP1/A13	x−½, −y+½, −z+1	OP2/C9	3.90	5.24 (1.39)
			O3'/G12		O5'/T8	3.93	5.23 (1.45)
			P/A13		P/C9	5.15	6.65 (1.29)
			P/G12		P/T8	8.25	7.12 (1.08)
4	Minor groove–major groove	76.1	P/G6	−x+½, −y, z−½	O3'/G20	3.74	4.29 (1.05)
			OP2/G6		O3'/G20	2.68	3.37 (1.00)
			P/G6		P/G20	8.07	8.36 (1.22)

			P/G12		P/G6	6.28	7.71 (1.10)
5	Major groove–major groove	76.0	P/G16	−x+½, −y+1, z−½	O5'/C1	3.40	4.63 (1.14)
			O3'/G10		P/G16	3.93	3.96 (0.77)
			P/G16		P/G2	6.29	8.09 (1.03)
			P/G10		P/G16	8.36	7.57 (0.96)

Table 5-6. van der Waals contacts and interactions between symmetry-related helices in DNA. The data in parentheses are standard deviations.

Hydrated Mg^{2+} frequently mediates intermolecular contacts between adjacent DNA molecules. Evidence from molecular dynamics simulations by Hartmann[189] supports that Mg^{2+} bound to the DNA major groove have an effect on DNA structure and dynamics; hydrated Mg^{2+} often forms a stable intra-strand cross-link between the two purines and increases the BII population. Li et al. [205] have argued that the binding of $^{Mg2+}$ rigidifies DNA compared to Na^{+}.

Two $Mg(H_2O)_6^{2+}$ clusters are identified in the DNA deposited structure, located at G16·A17 in the minor groove and at G6·A7 in the major groove, adjacent to two of the BII sites at P7 and P17; these are shown in Figure 5-13. These sites are also highly occupied with Mg^{2+} in the simulation. Additional ion sites must be present to neutralize the total charge, but were not modeled in the deposited structure. Other relatively well-localized sites for hydrated magnesium are evident in the simulation, both in the grooves and near phosphate positions (see Figure 5-13). Almost all of the Mg^{2+} ions remain coordinated to six water molecules throughout the simulation: for 0.58% of the time, the ions have 5 waters in the first coordination shell. All of the ions have approximately the same diffusion constant (about 3 × 10− 9 cm2/s), so that the ions initially placed in crystallographic density have the same mobility as those randomly added to achieve neutrality. A fuller analysis of ion distributions probably requires additional simulations, preferably with less lattice distortion than that seen here, since magnesium ions are involved with bridging interactions between duplexes as well.

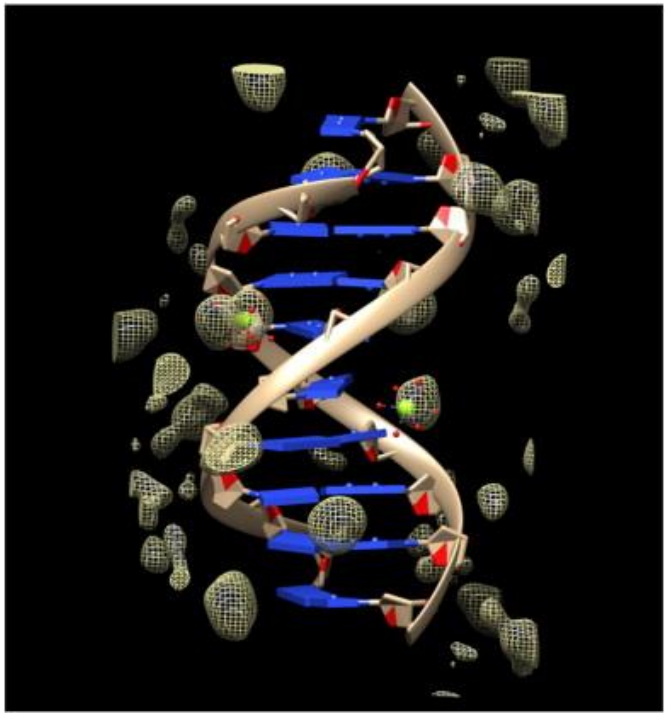

Figure 5-13. Cartoon view of the 1D23 crystal structure, showing the two identified Mg^{2+} ions as green spheres with attached water molecules (smaller red spheres). The mesh structure shows the Mg^{2+} distribution from the simulation.

The analysis of the deformations in the distances between the centers of mass between the interface residues (Figure 5-14) implicates all interfaces in crystal lattice deformations. The coaxial stacking (interface 1) shows small fluctuations and a compression by about 0.5 Å, whereas the other interfaces increase in average length by up to 0.5 Å, and individual copies exhibit much larger variations from the mean. These results are consistent with the larger average values and fluctuations for individual contacts that are shown in Table 5-6. Deviations from the mean behavior are much greater for DNA than for RNA (compare Figure 5-11 and Figure 5-14).

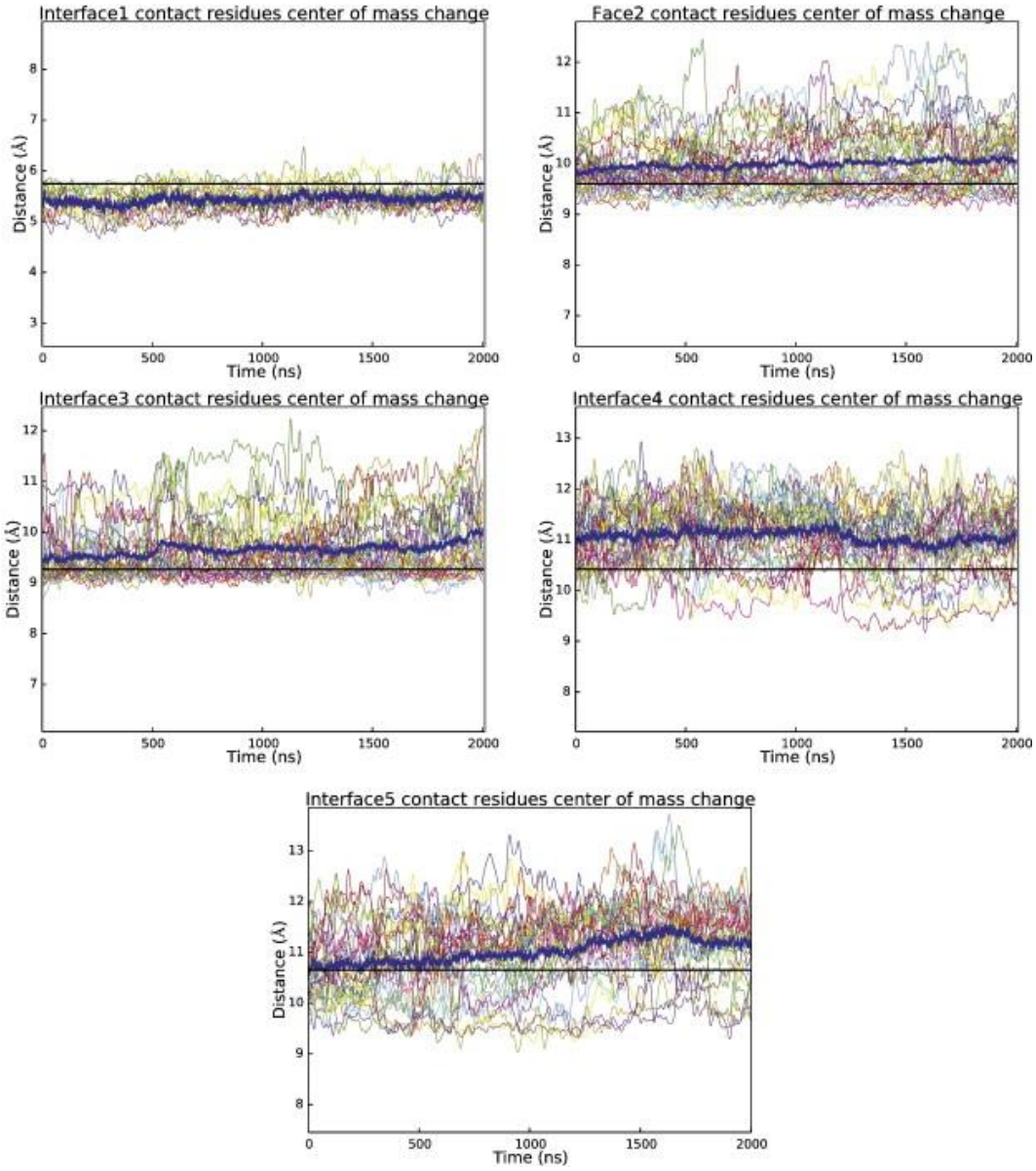

Figure 5-14. Distances between the centers of mass between the interface residues for DNA. Black lines show the value from the X-ray structure. Bold blue lines are the average distances between interface centers of mass.

The displacements of the centers of mass from their ideal positions (averaged over time) for each duplex in the supercell are shown in Figure 5-15. The average distance of each point from its ideal position is 0.37 Å for RNA and 1.77 Å for DNA. (For comparison, in a recent simulation of lysozyme, the mean deviations of the proteins from their ideal lattice positions was 0.27 Å.) Such

large translations of some duplexes in the supercell correspond to a significant loss of crystal symmetry, which can also be seen in Figure 5-2.

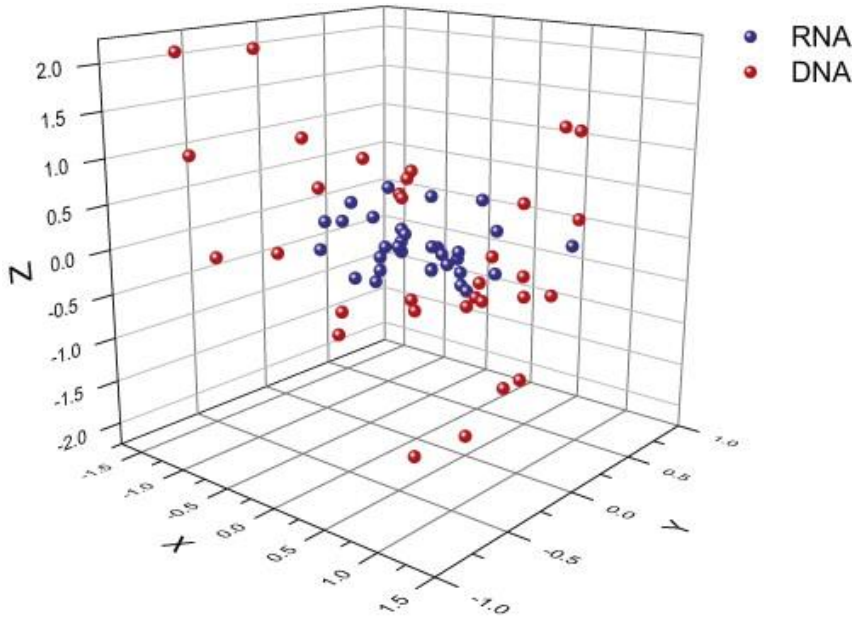

Figure 5-15. The center of mass position for each of the 32 duplexes, where the origin represents their position in an ideal lattice.

5.6. Conclusions

We have performed crystal simulations of 1D23 and 1RNA in the "supercell" with 32 copies of each duplex, as well as solution simulation of isolated duplexes. As has been seen in earlier studies, the average duplex structures from crystal simulation match the experimental structure much more closely than the average structures from solution simulations with the same force field. There is a general tendency for the solution simulation to flatten the variations in helicoidal parameters seen in the crystal (and, presumably, stabilized by crystal packing interactions.) Fluctuations about the average structure in the crystal simulation are in good agreement with refined B-factors for base atoms, but are somewhat higher for the sugar-phosphate backbone atoms; fluctuations in solution are considerably higher than in the crystal lattice.

On the other hand, the contacts that stabilize the crystal lattice are not extensive in these crystals, especially for DNA, and our simulations show a progressive deformation of the lattice, such that the duplexes move away from the "proper" positions in the supercell by 1 to 2 Å. This is more true for DNA than for RNA, perhaps by virtue of having fewer hydrogen bonds between duplexes to stabilize the lattice. Another factor may simply be the shorter simulation time for RNA. The results of the present work represent the largest all-atom crystal "supercell" simulations of nucleic acid to date, and provide a clearer understanding of the structural and dynamical parameters for RNA and DNA crystals. They also offer insights into a new challenge for force-field simulations in crystals and non-crystalline applications with weak interactions between chains. Future studies are planned to investigate dependencies on force fields, water models, and methods of representing ion interactions.

5.7. Acknowledgements

This work was supported by NIH grant GM103297. We thank Tom Cheatham for the useful discussions.

Section IV. Improved crystallographic methods through crystal molecular dynamics

Chapter 6. *Improved ligand geometries in crystallographic refinement using AFITT in Phenix*

6.1. Abstract

Modern crystal refinement programs rely on geometry restraints to overcome the challenge of a low data to parameter ratio. While the classical Engh & Huber restraints work well for standard residues, the chemical complexity of ligands and small molecules presents a particular challenge. Most current approaches either limit ligand restraints to those that can be readily described in the Crystallographic Information File format, thus sacrificing chemical flexibility and energetic accuracy, or they employ protocols that lengthen refinement times and hinder automated refinement workflows. We present the results of combining AFITT and the Phenix software suite, which together generate more chemically accurate models for small molecules. A Phenix-AFITT refinement uses a full molecular mechanics force field for ligands during refinement. It is fully integrated with a standard refinement protocol and requires practically no additional steps from the user, thus making it ideal for high throughput workflows. Phenix-AFITT refinements also handle multiple ligands in a single model, alternate conformations and covalently bound ligands. Refinements using AFITT significantly reduce ligand energies and lead to improved geometries without detriment to R-free factors.

6.2. Introduction

Structural knowledge is fundamental to our understanding of biomolecular function and related drug discovery efforts. X-ray crystallography remains the pre-eminent method for obtaining detailed structural information about molecules. Continued advances in data collection and processing, model building and structure refinement have gone a long way toward making crystallography a semi-automated, reliable technique for high-throughput structural biology. In the course of

crystallographic structure solution, the process of refinement is used to optimize the atomic coordinates against the experimental data. However, because of the low data to parameter ratio in a typical experiment, additional *a priori* knowledge must be introduced into the optimization algorithm to make it tractable. Usually, this additional knowledge is in the form of stereochemical restraints for bond lengths, angles, steric exclusions as well as additional restraints for dihedrals, chirality and other geometry restraints.

For standard biomolecular residues (proteins and nucleic acids), most modern refinement programs base these restraints on the so-called Engh & Huber restraints developed in 1991[11] from a survey of small molecule crystal structures and with later corrections added in 2002[12]. Engh & Huber restraints function reasonably well for standard residues, but even in this case deficiencies have been exemplified [206]–[209]. On the other hand modeling of small molecular ligands presents a particular challenge due to their more complex chemistry, conformations and energetics. Thus small molecules and ligands are not accurately modeled by the standard set of restraints [210], [211]. In fact, recent studies suggest that as many as 60% of the structures deposited in the Protein Data Bank (PDB, Berman et al., 2000) may contain questionable ligand structures [212].

Significant effort has been placed into developing tools for accurate representation of ligand restraints in crystallographic models. Some [208], [213]–[217] employ sophisticated approaches to derive the same type of stereochemical restraints as those used by the Engh & Huber set for standard residues. In other words, these restraints must conform to the standard Crystallographic Information File (CIF) restraint dictionary format provided to the restraint program[218], [219]. Other approaches [220]–[223] focus on more accurate ligand representation through the use of more elaborate protocols, force fields or quantum methods.

The former approaches suffer from an inherently inadequate representation of the ligand by attempting to force the restraints into the standard restraint CIF format that insufficiently models or wholly ignores energetic effects such as electrostatics and dispersion forces. The latter approaches

are often complicated to use, requiring multiple additional steps from the user, and are therefore difficult to integrate into an automated workflow and the speed requirements of many modern day high throughput labs such as those involved in pharmaceutical drug discovery.

Here we present a more accurate but efficient structural modelling of small molecules in the refinement process using the combined power of two crystallographic applications. Phenix [90] is the widely popular suite of software for integrated crystallography that includes the phenix.refine [91] application for refinement; AFITT ([217]; http://www.eyesopen.com/afitt) is OpenEye's package for automated ligand placement in crystal density. AFITT models ligand stereochemistry with the well-regarded Merck Molecular Mechanics Force Field (MMFF, [224]–[227]) but until now was forced to write the restraints to the standard CIF format restraint files, to the detriment of the improved ligand modelling by MMFF. By seamlessly integrating AFITT with Phenix, the user gains the powerful advantage of a full molecular mechanics representation of the ligand while being able to maintain the same efficient refinement workflow. Furthermore, alternating steps of standard macromolecular refinement followed by highly accurate ligand refinement is no longer necessary as both sets of restraints are applied simultaneously.

Here, we provide a comparison of refinements on a test set of 265 ligands. We compare refinement using AFITT-derived CIF restraint dictionaries in refinement and obtaining the ligand geometry gradients in refinement directly from AFITT. Thus our comparison does not depend on the possible differences between force fields used by various ligand modelling tools, but hinges only on the improvement gained by simultaneously representing the ligand with the full molecular mechanics force field during the course of refinement. We see that Phenix-AFITT refinements yield improved, lower-energy small molecule structures while ensuring the same degree of agreement with experimental data as obtained with the refinement packages most widely in use today. Thus, a Phenix-AFITT refinement provides the user with a fully integrated ligand refinement that ensures accurate modelling of ligand chemistry. The implementation of AFITT in Phenix is versatile, easy to use and powerful. Refinements can include different types of ligands and multiple instances of each

ligand type. Support for ligands with full or partial alternate conformations is fully integrated as is refinement of ligands covalently bound to the macromolecule.

6.3. Methods

phenix.refine [91] optimizes a crystal structure via a series of repeated cycles. During each cycle a series of parameters of the user's choosing are optimized. These usually include the atomic coordinates and the isotropic atomic displacement parameters but can also include, for example, Translation-Libration-Screw parameters, bulk solvent scaling and anisotropic atomic displacement parameters. Each optimization is conducted by minimizing a residual function of the model against the experimental data using a maximum likelihood approach.

AFITT [217] is package developed by OpenEye Scientific Software for small molecule real-space fitting in biomolecular crystallography. It uses a combination of an electron density shape matching algorithm and a molecular mechanics force field to fit small ligands into experimental density while maintaining accurate chemical geometry. AFITT uses an "adiabatic" method to find the best relative weight between these two components. It can be run without a solvent model or using either the Sheffield [228] or the Poisson-Boltzmann scheme to model solvation effects. AFITT uses the Merck Molecular Mechanics Force Field (MMFF94)[224]–[227], [229]. This force field was designed to reproduce *ab initio* accuracy in a broad range of chemical functionality and has been shown to produce satisfactory results with small molecules typically encountered in biomolecular crystallography[222], [230], [231].

In the case of reciprocal space atomic coordinate refinement, the Phenix refinement target function has the form

$$E_{Phenix} = w \cdot E_{x-ray} + E_{geometry} \qquad \textbf{(Eq. 21)}$$

where E_{x-ray} is the residual of the structure factors, $E_{geometry}$ is the residual due to the Engh & Huber restraints [11], [12] and *w* is the weighting factor. If a ligand is present, the $E_{geometry}$ term can further be divided:

$$E_{Phenix} = w \cdot E_{x-ray} + E_{protein} + E_{ligand\_non-bonded} + E_{ligand\_bonded} \quad \textbf{(Eq. 22)}$$

where E_{ligand_bonded} represents the so-called bonded terms in the geometry restraints that include bonds, angles and torsion angles. $E_{ligand\_non-bonded}$ represents the non-bonded terms that in the case of Engh & Huber are the atomic steric overlap restraints. During a Phenix-AFITT refinement the last term in the equation above is replaced by a residual calculated by AFITT:

$$E_{Phenix-AFITT} = w \cdot E_{x-ray} + E_{protein} + E^{Phenix}_{ligand\_non-bonded} + E^{AFITT}_{ligand\_bonded} \quad \textbf{(Eq. 23)}$$

The implementation in Phenix combines the *phenix.refine* refinement scheme and optimization algorithm while using AFITT to obtain the E_{ligand_bonded} part of the residual. A Phenix-AFITT refinement is invoked with:

```
phenix.refine mymodel.pdb mymodel.mtz myligand.cif \
  use_afitt=True afitt.ligand_names=BCL
```

This implementation automatically searches for all instances of the ligand specified by the user (Bacteriochlorophyll A in the example above) and uses AFITT to calculate the geometry gradients for those instances. More than one type of ligand can be included. The required restraints CIF dictionary file specifies the bond/angle topology for AFITT, but not the actual restraint force constants, which are calculated internally by AFITT. This implementation also accounts for alternate conformations on ligand atoms and ligands covalently bound to the macromolecule. A heuristically determined weight of 10 is placed on the $E^{AFITT}_{ligand\_bonded}$ term by default, but the user also has the option of modifying the weight. Additionally, there is a simple command line tool to quickly obtain MMFF ligand energies from a given PDB coordinate file.

6.4. Results and discussion

Testing of the implementation of AFITT in Phenix was performed using a set of 189 protein PDB structures taken from the Iridium dataset[232]. This set contains a curated and chemically varied set of protein-ligand structures. Some structures contained multiple small molecules resulting in 304 small molecule instances in total. Because many of the structures lacked reflection test sets (to calculate R-free[97]) a new test set of reflections was assigned and each model was refined using *phenix.refine* with the default strategy for 10 macrocycles to remove "memory" of the original test set. Next each structure was refined for a further 5 macrocycles using the default strategy and with the ligand modeled either using an AFITT-derived CIF dictionary file (hereafter referred to as "AFITT-CIF" refinement) or with the new Phenix-AFITT implementation. Using this strategy, as opposed to comparing results with another non-AFITT ligand restraint tool, allowed for rigorous testing of the benefits of implementing ligand restraints in the way advocated here; that is, by applying a full molecular mechanics treatment of the ligand but in a manner fully integrated with the refinement optimizer.

Phenix-AFITT refinement produces significantly lower ligand energies, both compared to the original deposited coordinates and compared to refinement using an MMFF-based CIF restraints file (Figure 6-1). For our test set the mean post-refinement ligand energy was 402.09 kJ/mol for the set of deposited structures, 350.81 kJ/mol for the AFITT-CIF refined structures and 260.54 kJ/mol for the Phenix-AFITT refined structures. This signifies an average reduction of ligand conformational energies of 34% versus the deposited conformation in the PDB and 22% versus refinement using MMFF derived CIF restraint dictionaries. The right-hand panel of Figure 6-1 shows scatterplots of the Phenix-AFITT refined ligand energies versus either the deposited PDB or AFITT-CIF refined energies on a per-model basis. It is apparent that Phenix-AFITT refinement results in significant reduction of ligand conformational energy compared to both other datasets. Furthermore, the greatest energy reduction using AFITT in Phenix (distance below the identity line) tends to occur in

the ligands with highest starting energies. In other words, the higher the ligand conformational energy the more a Phenix-AFITT refinement is able to reduce it.

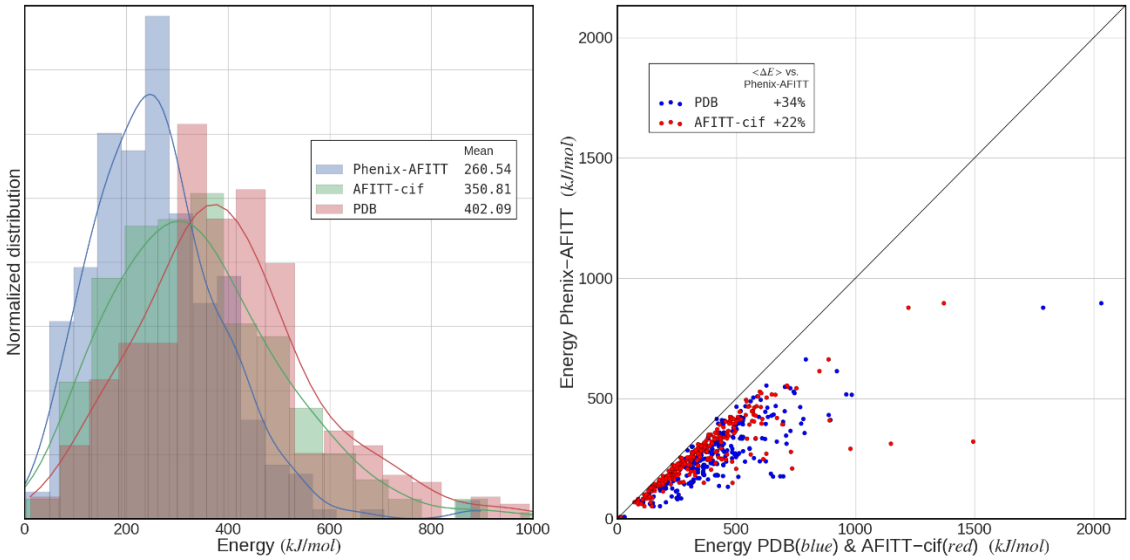

Figure 6-1. Ligand conformational energies from PDB-deposited models, AFITT-CIF refinement and Phenix-AFITT refinement. Left panel displays three histograms with kernel density estimates (KDE) of the distributions for the full set of test ligand energies. Means of each set of ligand conformation energies are shown in the legend. Right panel displays a scatterplot comparing the conformation energy of each ligand obtained from a Phenix-AFITT refinement against either the deposited PDB model (blue dots) or the models after refinement with an MMFF-derived CIF dictionary file (red dots). The mean percent reduction in energy from using Phenix-AFITT protocol is 34% vs. the PDB conformations and 22% vs. AFITT-CIF.

To further validate refinement quality, the Mogul software [233] was used to assess the post-refinement ligand geometries. Mogul is a knowledge based library of accurate small molecule geometries that can be used to assess the quality of small molecule conformations against structures in the Cambridge Structural Database (CSD). Mogul has been shown to accurately evaluate both experimentally and computationally derived geometries. The results (Figure 6-2) show that both the Phenix-AFITT refined and the AFITT-CIF refined geometries are significantly better than those found in the PDB and that Phenix-AFITT geometries are better than AFITT-CIF. The respective

mean RMSD scores for bonds were 0.036 for the deposited PDB set, 0.019 for AFITT-CIF and 0.018 for the Phenix-AFITT protocol. For angles the respective mean RMSD was 3.07 for the deposited PDB set, 2.76 for AFITT-CIF and 2.06 using AFITT in Phenix. This indicates that Phenix-AFITT refinements not only reduce conformational energies according to the MMFF force field but that the conformations are actually more accurate both compared to existing accurate structures and to what one would expect to derive from quantum calculations.

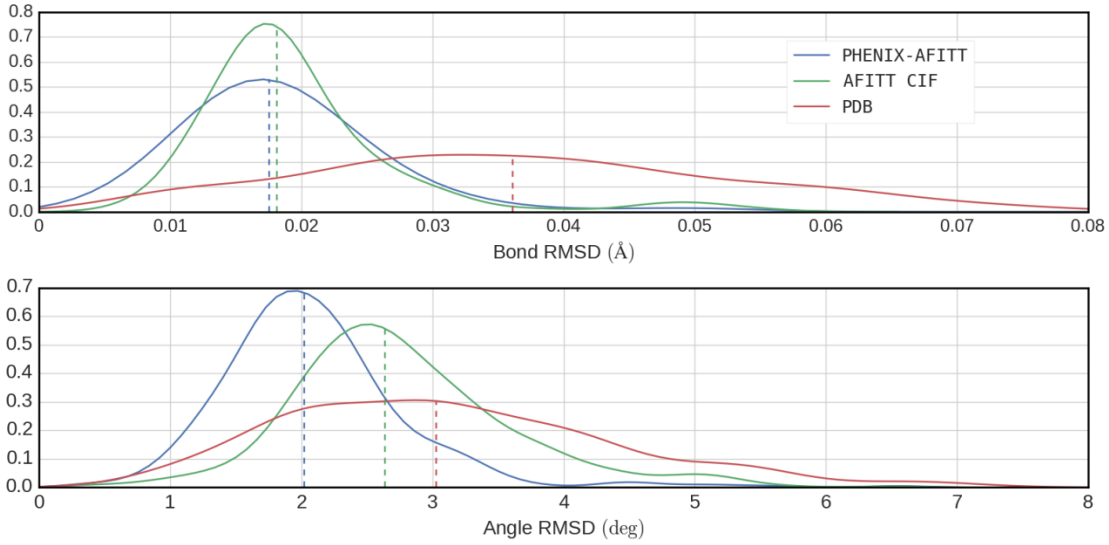

Figure 6-2. Mogul validation of the PDB deposited, AFITT-CIF refined and PHENIX-AFITT refined ligand conformations. Top panel shows bond RMSD distribution and bottom panel shows angle RMSD distribution. RMSD is relative to the Mogul library of "ideal" bonds and angles. Dashed vertical lines indicate median of each distribution.

While significantly reducing ligand energies, Phenix-AFITT refinements did not result in poorer agreement with experimental data. The mean R-free of the structures refined with AFITT-CIF was 0.231 while for those refined with the Phenix-AFITT protocol the mean R-free was 0.232 (Figure 6-3). A pairwise comparison of the Phenix-AFITT and AFITT-CIF refinements yields a mean difference in R-free between the two methods of 0.0012 which is statistically insignificant. Thus on average AFITT-CIF results in slightly lower R-free, but this difference is very small (0.0012 on average) and within the margin of error. We found only one case with an R-free difference greater

than 0.01. This was the case of PDB entry 1CTR which resulted in a ΔR-free of 0.025. However, the starting model for 1ctr is problematic in itself with 32% of rotamer outliers, a clash score of 50 and a MolProbity [234] score of 4.07. Thus one could expect to see large R-factor fluctuations upon refinement in this case.

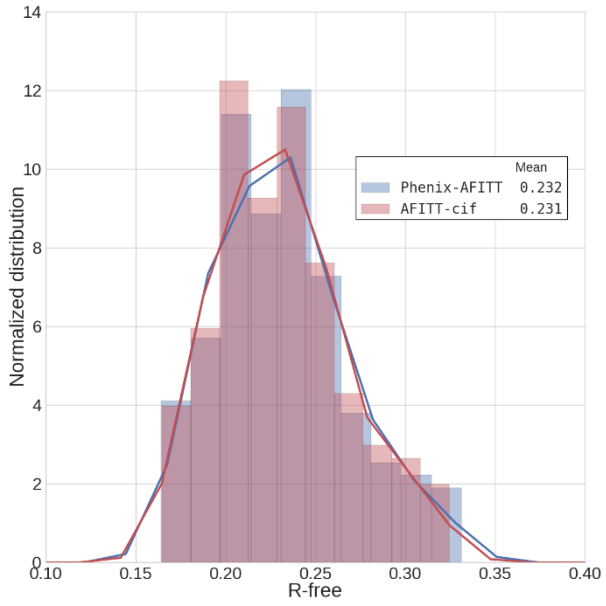

Figure 6-3. R-free distributions after refining the test set using either AFITT-CIF or the Phenix-AFITT protocol. Means of each distribution are shown in the legend.

Figure 6-4 shows a more detailed comparison of eight randomly selected structures from the test set with a total of 10 ligands. As can be seen, the Phenix-AFITT refinement leads to significantly lower energies in all cases. In some cases (e.g. second ACD instance in PDB entry 1CVU) AFITT-CIF restraints lead to ligand energies that are much higher than even the deposited coordinates, while using AFITT directly in Phenix obtains lower energies. At the same time, a comparison of R-free shows that the fit to experimental data remains essentially the same between the two refinements. The structure with the highest R-free difference in the entire data set, PDB entry 1CTR, is included in the panel. Phenix-AFITT can also handle ligands with alternate conformations. Figure 6-5 shows a similar comparison of energies and R-free factors for five PDB structures containing multiple ligands with alternate conformations. The conclusions are similar.

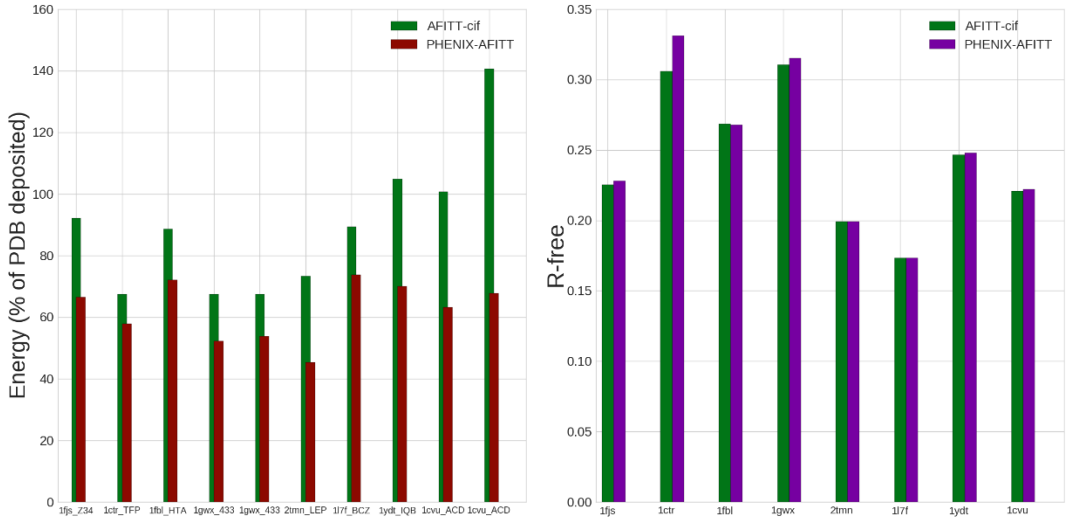

Figure 6-4. Comparison of 8 randomly selected PDB structures. Left-hand panel shows energies obtained with AFITT-CIF and Phenix-AFITT refined ligand restraints as a percentage of the deposited ligand energy. Labels provide the PDB code followed by the ligand's 3-letter code. Some PDB structures have more than one instance of a ligand. Right-hand panel shows the R-free obtained after refinement with E-H or AFITT geometry restraints on the ligands.

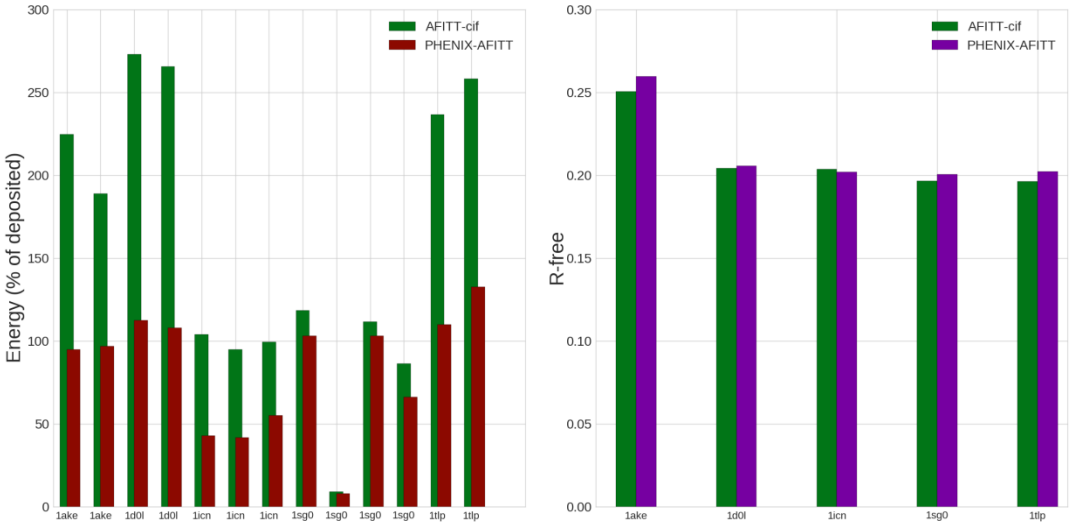

Figure 6-5. Comparison of 5 PDB structures containing ligand instances with alternate conformations. All labels and statistics as in Figure 6-4.

Structure refinement with the Phenix-AFITT protocol is somewhat slower than refinement with a previously prepared CIF dictionary file. Figure 6-6 presents a histogram of runtime differences between AFITT-CIF and a Phenix-AFITT refinement as a percentage of the AFITT-CIF runtime. In general, a Phenix-AFITT refinement is slower by an average of 16% compared to the same refinement using E-H restraints. In addition there were five structures with refinement times more than twice as slow as with the traditional algorithm: 1q41:2235%, 1sq5:1141%, 1q1g:440%, 1hq2:219%, 1dd7:110%. These five outliers have been omitted from the plot.

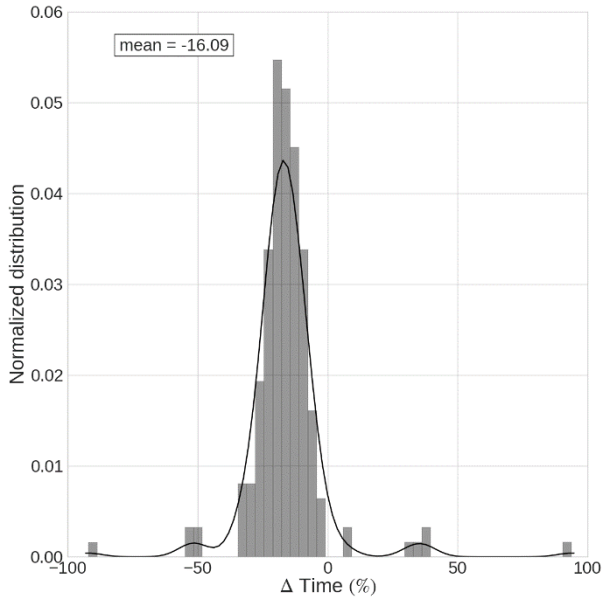

Figure 6-6. Difference in run time between traditional E-H and Phenix-AFITT refinement as a percentage of E-H refinement run time. Positive numbers indicate that the Phenix-AFITT refinement is faster; negative numbers that Phenix E-H is faster. Five outliers (1q41, 1sq5, 1q1g, 1hq2, 1dd7) have been omitted from the plot.

6.5. Discussion

The Phenix-AFITT protocol is a new tool that improves on the limited set of geometry restraints typically used in modern refinement programs. By implementing an interface to AFITT into Phenix refinement, a more accurate set of chemical restraints is made available that leads to a significant reduction in ligand conformational energies and an improvement in ligand geometries. This is

accomplished without detriment to the model's fit to experimental data and with only a modest increase in refinement time. AFITT is fully integrated with *phenix.refine*, is easy to use and automatically handles multiple ligands, alternate conformations and covalent linkage. A user's guide is available on-line in the Phenix documentation under the "Structure Refinement and Restraint Generation" heading.

The Phenix-AFITT protocol not only improves on the deposited PDB ligand geometries but also on those obtained with refinements using a CIF-format restraints file derived from the same MMFF force field that AFITT uses. This is noteworthy because it underscores that improved refinement results are not solely the result of using a better force field but also of how that force field is implemented within the refinement target function. Most target functions in use today only allow for representation of bond, angle, simplified dihedral and atomic overlap penalty terms. Thus a crystal refinement restraints representation of a ligand's geometry parameters necessarily constrains the force field into a more primitive representation. For example, non-bonded interactions (electrostatics and van der Waals forces) are no longer accurately represented in the restraints format. The fact that the Mogul validation results indicate a greater improvement between AFITT-CIF and Phenix-AFITT refinements in angle than in bond geometries, is consistent with this interpretation: angles should be more susceptible to differences in other force field terms than bonds. Unfortunately the CIF-like restraints dictionaries are in widespread use today because they represent the same geometry restraints function as has been found to function well with protein residues. As a move is made to more accurate refinement, it can be hoped that refinement target functions will more often be implemented according to the paradigm presented here so as to more accurately represent the complex conformational space of small molecules and ligands.

Chapter 7. *Implementing molecular dynamics for improved crystallographic model refinement with Phenix and Amber*

7.1. Abstract

The refinement of accurate biomolecular crystallographic models relies on a correct set of geometry restraints to supplement the low data to parameter ratio inherent in the experiment. We present a new tool that integrates Phenix crystallographic refinement with the Amber package for molecular dynamics and thus makes available the full all-atom Amber force field for precise modelling of biomolecular chemistry. The advantages of Amber force field include a carefully derived set of torsion angle potentials, an extensive and flexible set of atom types, Lennard-Jones treatment of non-bonded interactions and a fully accurate treatment of crystalline electrostatics. The new combined method, Phenix-Amber, improves model quality with over 87% of structures displaying a better *MolProbity* score as compared to traditional refinement. Electrostatics are more accurately modeled and there is a 7% increase in the number of hydrogen bonds retained through refinement. Additionally overfitting is reduced with 93% of models having a smaller R_{work}-R_{free} gap. We further see that improvements are greatest at lower resolutions and poorer starting models. An efficient implementation and a user-friendly tool for preparing input files offer an excellent tool for improving the quality and accuracy of refined models. Furthermore, the flexible implementation will simplify the future development of more advanced applications such as Amber-based ensemble refinement, quantum mechanical representation of active sites and accurate molecular dynamics simulated annealing.

7.2. Introduction

Accurate structural knowledge lies at the heart of our understanding biomolecular function, interactions, mechanisms of enzyme activity as well as cellular processes and pathological states. With over 100,000 structures in the Protein Data Bank[16] solved via x-ray diffraction methods, crystallography is the eminent method for determining biomolecular structure. Crystal structure

refinement is a computational technique that plays a key role in post-experiment data interpretation. Refinement of atomic coordinates entails solving an optimization problem to minimize the residual difference between the experimental and model structure factor amplitudes.[89], [95], [235] However, due to inherent experimental limitations and low data to parameter ratio, the employment of an additional set of restraints, commonly referred to as geometry or steric restraints, is key to successful and reasonable structural refinement.[236] These restraints, which can be thought of as a prior in the Bayesian sense, provide additional observations to the optimization target and reduce the danger of overfitting. Their use leads to higher quality, chemically accurate models.

Most current refinement programs[91], [237]–[239] employ a set of geometry restraints first proposed by Engh & Huber in 1991 and later augmented and improved in 2001[11], [12]. This set of restraints is based on a survey of accurate high-resolution small molecule crystal structures from the Cambridge Structural Database and includes restraints on interatomic bond lengths, angles and torsion angles as well as parameters to prevent steric overlap between atoms and enforce proper chirality and planarity. The Engh & Huber restraints function reasonably well but a number of limitations have been identified over the years. Some of these limitations include: a lack of adjustability to changes in local conformation, the low number of parameters, the targets that are a result of averaging, bias arising from sampling only high-resolution crystal structures, inaccurate dihedral restraints, ignorance of electrostatic and quantum dispersive interaction and consequent lack of accounting for hydrogen bonding cooperativity. [206]–[209], [240]

An alternative approach is the use of geometry restraints based on all-atom force fields used for molecular dynamics studies. This is not a novel idea. In fact, some of the earliest implementation of refinement programs employed molecular dynamics force fields.[124], [235], [241] However, at the time restraints derived from coordinates of ideal fragments[242], [243] were found to provide better refinement results. Limitations of molecular dynamics-based restraints insufficient flexibility stemming from few atom types and overstabilization of electrostatic effects that limited conformational sampling. Since then, however, the methods of molecular dynamics and

corresponding force fields have seen significant development and improvement. Current force fields contain more atom types and are easily expandable. They are parametrized against accurate quantum mechanical calculation results not feasible just a few years ago as well as more accurate experimental results. Significant methodological advances, such as the development of Particle Mesh Ewald[44], [45] for accurate calculation of crystalline electrostatics and improved temperature and pressure control algorithms have greatly increased accuracy. Modern force fields have been shown to agree well with experimental data[66], [68], [69], [244], [245], including crystal diffraction data[62]–[64], [77], [78].

We have undertaken to implement the use of the Amber ff14SB force field as an alternate set of geometry restraints to the Engh & Huber set. Here we present an integration of the *Phenix* software package for crystallographic refinement[90] and the *Amber* software package[26] for molecular dynamics. We present results of extensive refinements of over 6000 structures and compare them to traditional refinement, both in terms of model quality, chemical accuracy and agreement with experimental data. We also describe the implementation and discuss future directions.

7.3. Model quality

To compare the refinement using Amber against traditional refinement using Engh & Huber restraints, we ran *phenix.refine* on a set of 6084 randomly selected structures from the Protein Data Bank. Structures ranged from 0.9 Å to 4.3 Å with the majority of structures (96%) falling in the 1.0-3.0 Å range. Coordinate files were obtained directly from the PDB and inputs prepared via the automated *phenix.AmberPrep* program (see section 4 below for details). Each model was then subjected to 10 macrocyles of refinement using the default *Phenix* strategy (bulk solvent model and anisotropic scaling factor, reciprocal space coordinate refinement and isotropic or anisotropic B-factor refinement) using either Engh & Huber or Amber geometry restraints. Quality of the resulting models was assessed via *MolProbity*[234] and *cpptraj*[110] (available in *AmberTools*).

A summary comparison of these refinements is presented in Table 7-1. We found that Phenix-Amber consistently produced higher quality models. 88% of the Phenix-Amber refinements exhibited improved (lower) *MolProbity* scores, 90% contained fewer clashes between atoms and 91% had fewer backbone residue Ramachandran outliers. Box plot distributions of some indicators of model quality are shown in Figure 7-1 (see Supplementary Data for all figures). Mean values of model quality descriptors are better using Phenix-Amber and appear to have a more narrow distribution around those means than those obtained with Engh & Huber. Phenix-Amber outliers are also less extreme. Phenix-Amber refinement improves the overall *MolProbity* score, clash score, number of Ramachandran outliers, number of residues in favored Ramachandran space and the number of rotamer outliers.

	No. of models better with Amber	**Percent of models better with Amber**	**Mean improve. with Amber**	**Max. improve. with Amber**	**Min. improve. with Amber**
Clash score	5449	90%	4.52	154	-34
No. of H-bonds	5646	93%	15.3	377	-20
MolProbity score	5331	88%	0.46	3.5	-1.79
Ramachandran outliers	5508	91%	0.44	25.7	-13.0
Ramachandran favored	4795	79%	1.33	42.6	-33.3
Rotamer outliers	4396	73%	0.84	75.0	-15.4
C-beta deviations	2225	37%	-2.05	62	-183
R_{work}	55	1%	-.0162	0.126	0.182
R_{free}	2026	33%	-0.0045	0.304	-0.237
R_{diff} (R_{free} - R_{work})	5644	93%	0.0118	0.178	0.159

Table 7-1: Summary of improvement obtained when refining the set of 6084 models with Amber as compared to using traditional Engh & Huber restraints.

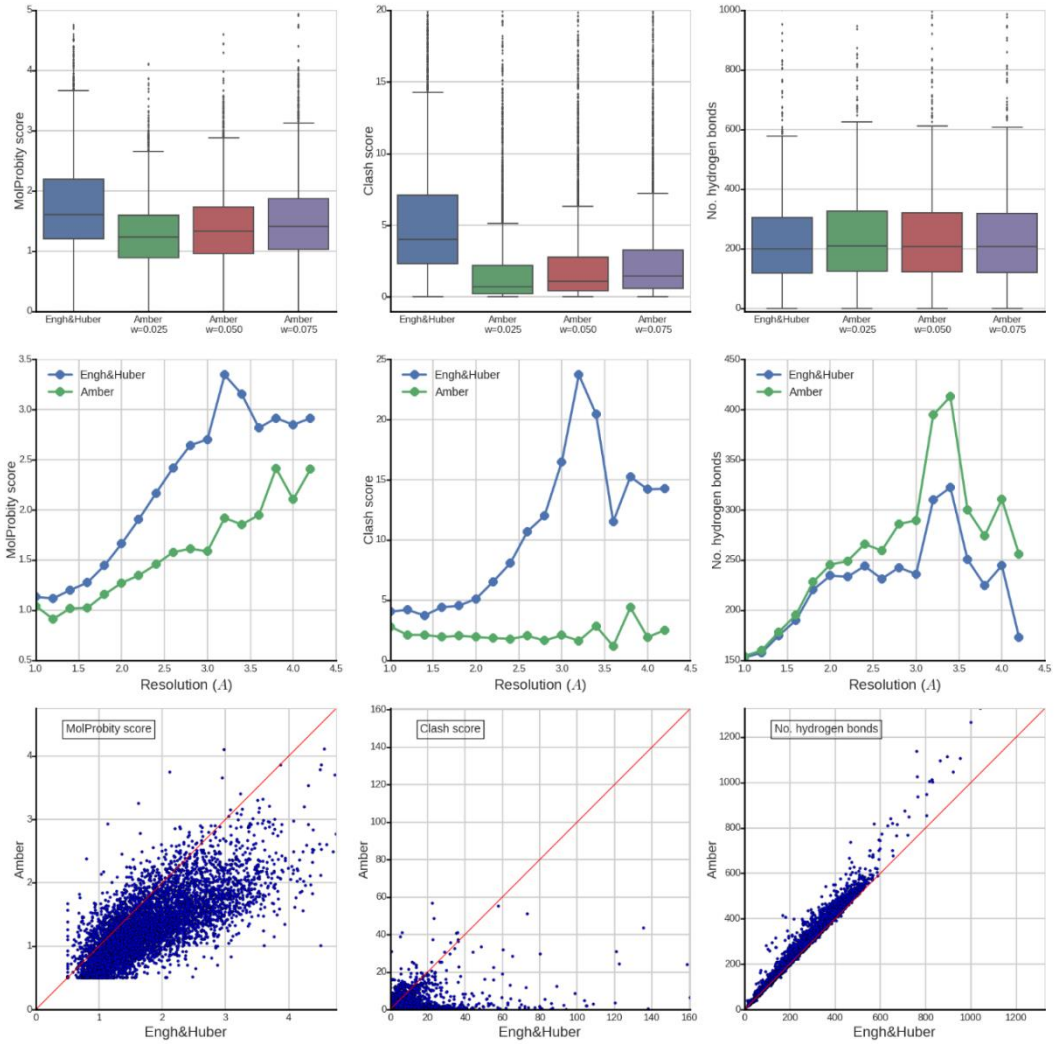

Figure 7-1. Phenix-Amber refinement improves model quality. Row I: MolProbity score, clash score and number of hydrogen bonds in final refined asymmetric unit obtained with Engh & Huber refinement and Amber refinement. Box plots are over entire set of 6084 structures. Amber refinements are shown using a wxc_scale weight of 0.025, 0.050 and 0.075. Row II: MolProbity score, clash score and number of hydrogen bonds mean values per resolution bin for Engh & Huber (blue) and Amber (green; wxc_scale weight 0.025). Row III: Scatterplots of MolProbity score, clash score and number of hydrogen bonds for 6084 structures obtained via Engh & Huber refinement (horizontal axis) and Amber refinement(vertical axis).

Refinement with Amber incorporates explicit restraints based on electrostatic forces and a Lennard-Jones potential to model quantum repulsion and dispersion forces. Consequently we find improved modelling of electrostatics using Phenix-Amber with 93% of models displaying more hydrogen bonds than traditional refinement. On average, we find 7.0% more hydrogen bonds when refining with Amber. Though generally small, these slight changes can be meaningful especially when interpreting interaction distances at active sites. For example, Figure 7-2 shows a typical nucleotide base-pair hydrogen bond interaction. The distortion in the Engh & Huber refined structure is not excessive, but the lack of explicit electrostatic restraints coupled with slight ambiguity in the electron density (2.7 Å resolution) allows for a slight rotation of the guanine residue resulting in a break of the hydrogen bond. In contrast, Amber is able to maintain all three base-pair bonds correctly.

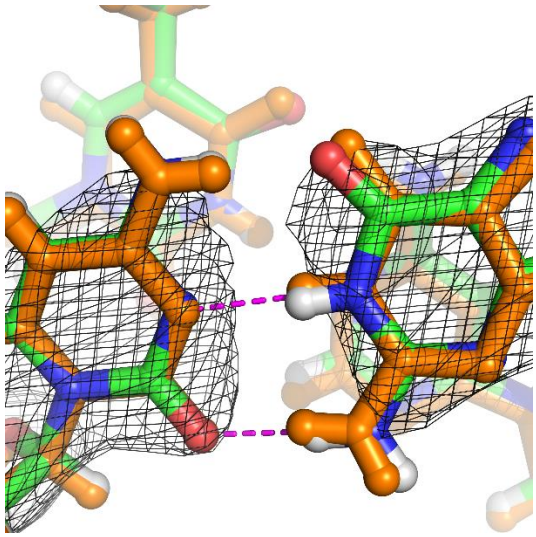

Figure 7-2. Phenix-Amber refinement improves modelling of electrostatics. Example of hydrogen bonding in a guanine-cytosine base pair from PDB:424d refined via Phenix-Amer refinement(color) and Engh & Huber refinement (orange). Purple lines display hydrogen bonds formed in the Engh & Huber refined model. One of the three base-pair hydrogen bonds is broken when refining with Engh & Huber restraints but maintained with Amber.

Phenix-Amber refinement improves on traditional refinement across all resolution bins, but the improvement is especially strong for lower resolution structures where electron density is less precise

and geometry restraints play a more important role. Figure 7-1, second row compares mean *MolProbity* scores, clash scores and number of hydrogen bonds between Engh & Huber refinement and Amber refinement for various resolution bins (see Supplementary Data for other properties). In all cases improvement using Phenix-Amber increases with worsening resolution. For example the difference in mean *MolProbity* score between Phenix-Amber and traditional refinement is about 0.02 at high resolution and increases to more than 1.0 at 3.2 Å resolution. Clash scores are particularly striking: for refinement using Engh & Huber restraints clash scores steadily increase as resolution drops resulting in some very high numbers of clashes. On the other hand, the mean clash score with Amber restraints appears to be independent of resolution and remains consistent at about 3 clashes per 100 atoms. This indicates that Amber is particularly effective at maintaining high model quality once electron density based enforcement of quality is removed. Here it should be noted, that our test set of 6084 structures only included 209 structures at less than 3.0Å resolution (3.5% of total test set). Thus the statistics below that cut-off can be treated as less reliable.

Next we ask whether Amber-based improvement is dependent on the quality of the starting model. The results presented so far were based on refinements starting from PDB deposited models which are usually of high quality. Scatter plots of model quality parameters (Figure 7-1, row III; see Supplementary Data for additional plots) indicate that improvement with Amber tends to increase as the quality of the starting model deteriorates. For example, in the case of *MolProbity* scores, the greatest differences between Amber and traditional refinement are observed for scores greater than 3. A similar observation holds for clash scores and hydrogen bonds. To probe further, we carried out refinements on poor (low quality) starting models. We selected nine diverse structures of varying resolution from the test set of 6084 and ran rudimentary molecular dynamics on each structure using the *phenix.dynamics* program for 100, 200, 500, 2000 and 5000 steps resulting in structures with an atomic root mean square deviation (RMSD) of about 0.3, 0.5, 1.0, 2.0 and 5.0 Å each from the deposited model. We next ran 10 macrocycles of refinement with either Engh & Huber or Amber force field restraints on each model with the same strategy as described above. Figure 7-3 shows

MolProbity, clash score and hydrogen bond number results for one of the structures, PDB:3h70 (see Supplementary Data for all results). In all cases Phenix-Amber outperforms traditional refinement across the entire spectrum of structural distortion, but the degree of improvement with Phenix-Amber rises with the amount of distortion. Again this is especially striking in the case of clash scores which increase dramatically with traditional refinement after a distortion of about 1.0Å RMSD but remain at about the same constant level with Phenix-Amber. Thus Amber based refinement appears to perform even better on poorer models than on high-quality ones.

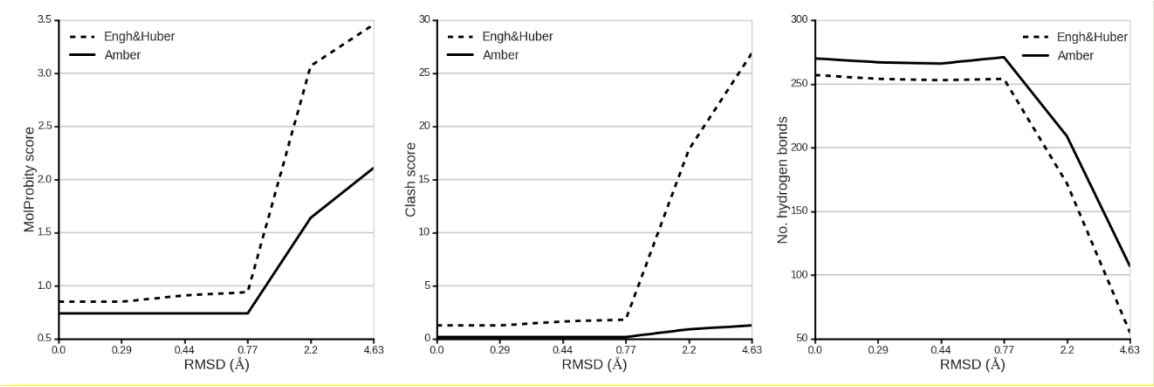

Figure 7-3. Phenix-Amber refinement improves model quality more with poorer starting models. Molprobity score, clash score and number of hydrogen bonds per asymmetric unit for a PDB:3h70 refined using Engh & Huber (dotted line) and Amber (solid line) restraints. The deposited model was first "shaken" with phenix.dynamics to obtained various degrees of RMSD divergence from the starting model.

7.4. Agreement with experimental data

While providing high quality, chemically sensible structures, refinement results must also be consistent with experimental data and avoid overfitting. Usually, the R_{free} factor[109] is used to assess the goodness of fit to experimental data. However R_{gap}, i.e. the difference between R_{free} and R_{work}, is also important: too high an R_{gap} is indicative of model overfitting to the working set of structure factor amplitudes. Figure 7-4 displays the R_{work}, R_{free} and R_{gap} obtained for the test set of 6084 refinements. R_{work} is higher by a value of about 1-2% for Amber-Phenix refinement across all

resolution ranges. On the other hand R_{free} is approximately the same for both Phenix-Amber and traditional refinement. Thus, the agreement of the refined models with experimental data is roughly equal using either set of geometry restraints, except at high resolution (above 1.5Å) where Engh & Huber improves on Amber by about 1-2%. At high resolution, restraints imposed by the experimental electron density become more important and this indicates the need for the user to decrease the relative weight between the x-ray and the geometry restraints in the target function.

More importantly, however, R_{gap} is consistently lower for Amber refinement. The amount of improvement increases with decreasing resolution, from about 0.3% (absolute) at 1.0Å resolution to about 2.5% on average at 3.0Å. This indicates that there is less overfitting when refining with Amber. Thus, Phenix-Amber refinement results in quantitatively similar agreement with experimental data as traditional refinement but with less overfitting to the working data set. Especially in the mid to low resolution range where the electron density is more ambiguous and overfitting more likely to lead to pernicious model effects, it is gratifying to see the large decrease in R_{gap} between traditional and Phenix-Amber refinement.

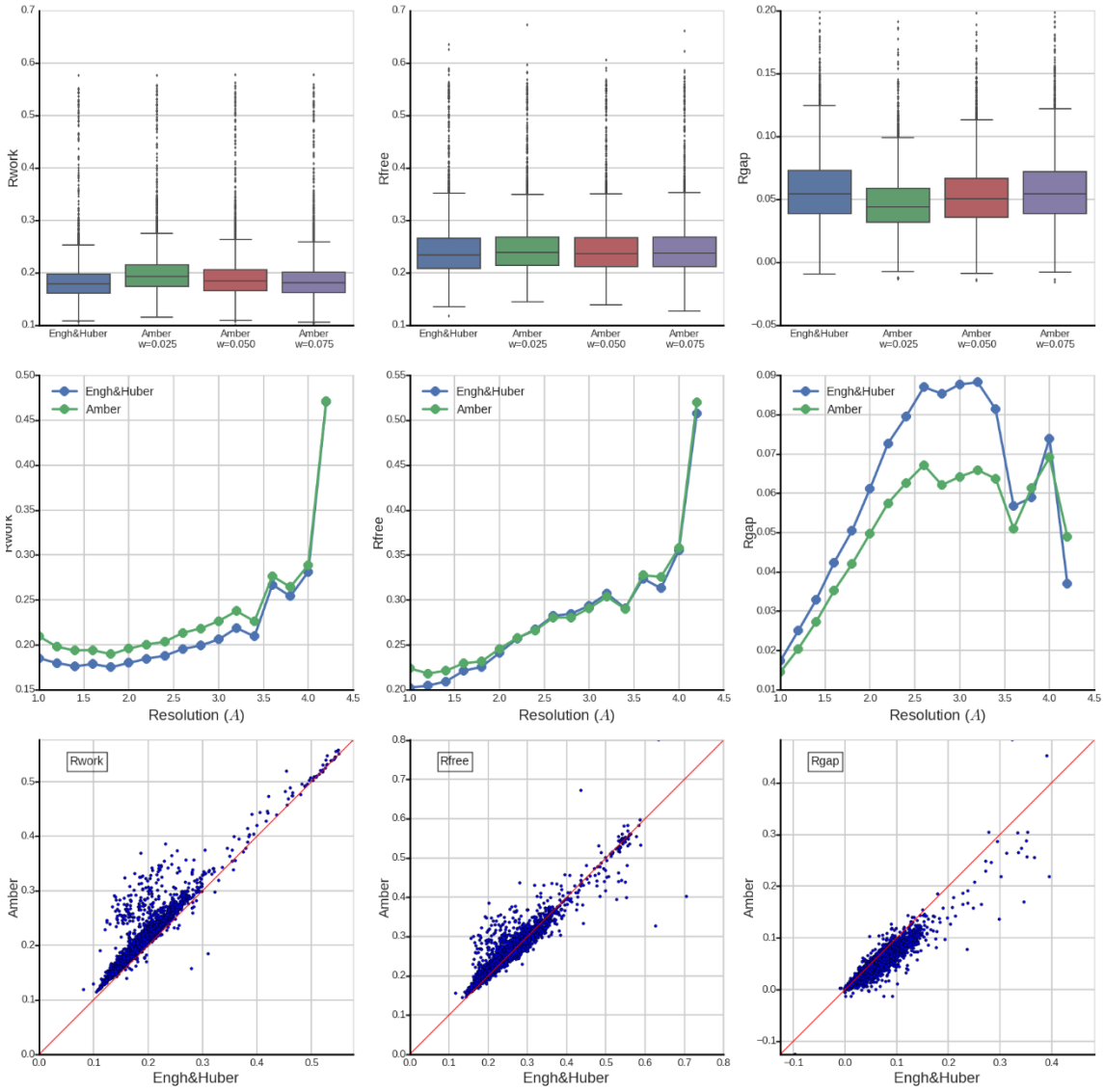

Figure 7-4. Phenix-Amber reduces model overfitting to working set of reflections. Row I: Box plot distributions of R_{work}, R_{free}, and R_{gap} over the 6084 models refined using Engh & Huber restraints and Amber restraints. Amber distributions are shown for three wxc_scale weights (0.025, 0.050 and 0.075). Row II: Mean values of R_{work}, R_{free}, and R_{gap} separated by resolution bin for Engh & Huber (blue) and Amber (green) refinement. Row III: Scatter plots of R_{work}, R_{free}, and R_{gap} obtained by Engh & Huber refinement (horizontal axis) and Amber refinement (vertical axis).

7.5. Implementation details

The target function optimized in *phenix.refine*'s reciprocal space atomic coordinate refinement stage is of the general form

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

where all the terms are functions of the atomic coordinates, T_{xyz} is the target residual to be minimized, T_{exp} is a residual between the observed and model structure factors and quantifies agreement with experimental data, $T_{xyz\_restraints}$ is the residual of agreement with the geometry restraints and w is a scale factor that modulates the relative weight between the experimental and the geometry restraint terms. In traditional refinement $T_{xyz\_restraints}$ is calculated using the set of Engh & Huber restraints:

$$T_{xyz} = w * T_{exp} + T_{Engh\&Huber}$$

To implement Phenix-Amber we substitute this term with the potential energy calculated using the Amber force field:

$$T_{xyz} = w * T_{exp} + E_{AmberFF}$$

were the Amber term is intentionally represented now by an E to emphasize that we directly incorporate the full potential energy function calculated in Amber using the ff14SB force field. Internally this is implemented using a shared library implementation and an API we wrote for AmberTool's *sander* and *mdgx* molecular dynamics programs. The interface to Phenix is achieved via Python's C API and Boost.Python instructions such that all calculations are performed internally with no external system calls from Phenix to Amber. Thus the two programs are fully and seamlessly integrated and memory efficiency is optimal.

For easy and automated running of Phenix-Amber refinement a helper program called *phenix.AmberPrep* assists users in preparing the necessary input files. *phenix.AmberPrep* requires a single pdb file and automatically produces three input files (PDB format coordinates, Amber format topology, Amber restart format coordinates). Hydrogens are automatically added and if there are any

unusual ligands, parameters are determined automatically. In addition residue names and numbering is changed to fit Amber and Phenix requirements, di-sulfide bridges are recognized and geometry minimization is optionally performed to ensure that bad clashes from addition of hydrogens are not present.

At present, *phenix.AmberPrep* prepares files for refinement using Amber's ff14SB force field. However, it can also easily be modified to use any force field currently available in the Amber release and updated as improved force fields are developed. Crystallographic model structures typically do not include all of the crystal liquor present in the experimental crystal and this is usually represented by a bulk solvent model. However, because of this concern arises about the electrostatic charges on charged residues refined using Amber. These charges are ordinarily screened to some degree by the bulk solvent but this screening effect is missing if the entire solvent is not explicitly modeled. Therefore we have also created a modified FF14SB force field referred to as *redq* with reduced charge values on the charged residues. We tested *redq* on several structures but little change in the obtained results (see Supplementary Data). Some refinements are slightly better with FF14SB and some are slightly better with redq, but most produce very similar results. Presumably the additional restraints imposed by the electron density are enough to offset the effect of unscreened electrostatic charges. Nevertheless, we have included the option for the user to select the reduced charge force field if one wishes.

The refinement target function includes a weight parameter *wxc_scale* that scales the relative contribution of the x-ray and geometry restraints term. This parameter depends strongly on the set of restraints used. To find the correct weight to use in Phenix-Amber refinement we performed extensive refinements on a set of 100 structures randomly selected from the PDB with various weight settings ranging from 0.006 to 0.5 (Figure 7-5 and Supplementary data.) as well as probing other scaling algorithms. We concluded that the optimal setting of *wxc_scale* for Phenix-Amber refinement ranges from 0.02 to 0.08. The lower end of this range results in optimal model quality with slightly worse R_{free} (the importance of the Amber geometry restraints target is scaled up) while

the upper end of that range improves experimental agreement at the cost of a slight decrease in model quality (the importance of x-ray restraints is scaled up). We then ran the full set of 6084 test structures using weight settings of 0.025, 0.050 and 0.075 and found that indeed a weight of 0.025 produce the best quality models, while weights of 0.050 and 0.075 resulted in models whose quality was slightly worse (but still better than with Engh & Huber refinement) but slightly better R_{free}. Thus, we recommend using *wxc_scale=0.025* in the *phenix.refine* input parameters by default, but the user may set the weight manually according to the tradeoff they wish to make between experimental agreement and model quality.

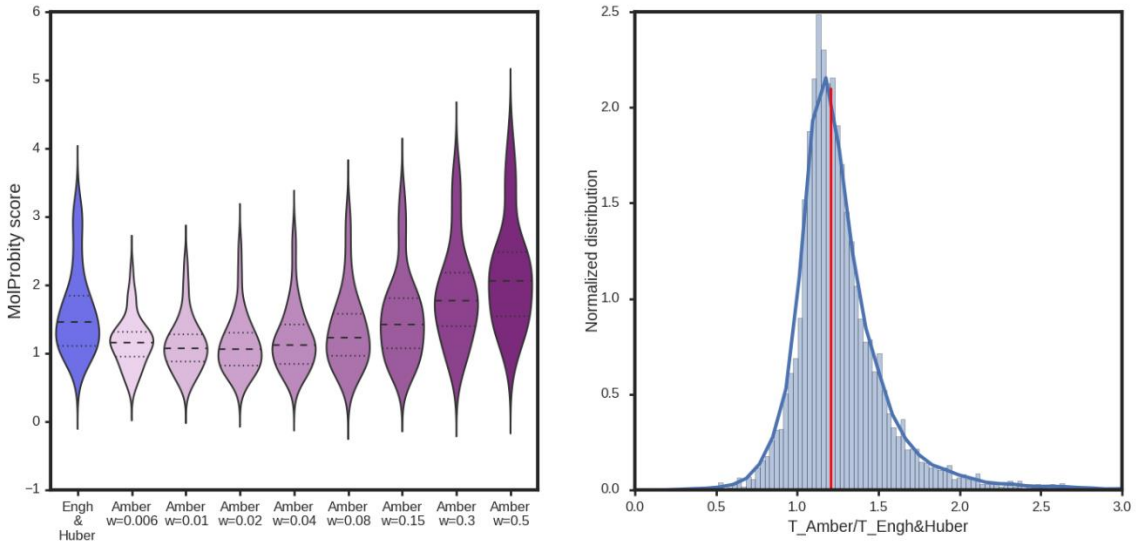

Figure 7-5: Left hand: violin plot distributions of the MolProbity score obtained via Engh & Huber refinement (blue) and various wxc_scale weights of Amber refinement (purple). Optimal results in terms of model quality are obtained with a weight of 0.2. Slightly better results in terms of R_{free} are obtained with a weight of 0.08 while still providing an improvement in model quality over traditional refinement. Right hand: Historgram of the ratio of Amber refinment time to Engh & Huber refinement time. Median (1.23) is shown by the red line.

We find that Phenix-Amber refinement requires on average 27% more compute time than traditional refinement (Figure 7-5). This is not surprising as the Amber force field calculation includes additional computation of electrostatics and van der Waals forces. It should also be

mentioned that in the previously mentioned set of 100 structures, we also compared Phenix-Amber refinement results to Phenix refinement using Engh & Huber restraints with the weight optimization option enabled. Weight optimization refinement is significantly slower than normal refinement because at each stage several parallel refinements are run with various weights to find the optimal setting. We find that Phenix-Amber without optimization outperforms traditional refinement even with the optimization setting enabled (Figure 7-5 and Supplementary Data). In other words, when a user would be willing to invest time in weight optimized refinement, Phenix-Amber refinement provides a significantly faster alternative that usually yields better results.

Phenix-Amber refinement does not currently support alternate conformations or partial occupancies and gaps in the structure are not allowed (i.e. all atoms must be modeled explicitly), however these features are currently being developed and should be made available in upcoming releases. Phenix-Amber is released as part of the Phenix software package but requires that the user also have the open source AmberTools software package installed.

7.6. Discussion

We have presented refinement results obtained by integrating Phenix with the Amber software package for molecular dynamics. Our refinements of over 6000 crystal structures show that refinement using Amber's all atom molecular mechanics force field consistently outperforms Engh & Huber restraint refinement. An overwhelming majority of Amber-refined models display significantly improved model quality. At the same time the R_{free}-R_{work} gap is greatly reduced but without increasing R_{free} which indicates similar agreement to experimental data but with less overfitting. Because Phenix-Amber is easy to run and automated set-up is facilitated by the current implementation, it is to be desired that refinement with a full molecular mechanics potential energy function become a standard tool in the crystallographer's arsenal.

Furthermore, Phenix-Amber obtains relatively better results when starting with poorer models and when working with low resolution data. This is especially important, as the experimental data in a

low resolution structure often does not provide sufficient unequivocal information to properly enforce a chemically sensible structure. The need for a more accurate set of restraints is greater in this case and indeed we see that the results obtained with Phenix-Amber outperform traditional refinement even more at low resolution and with poor starting models.

The idea of using a molecular mechanics force field for geometry restraints in crystallographic refinement is not new[124], [235] but was abandoned at one point in favor of restraints based on ideal fragment geometries due to limitations of force fields available at the time.[11] However, molecular mechanics force fields have progressed enormously since then [29], [30], [33], [150], [157], and it is heartening to see this reflected in the refinement results we present here. From a theoretical standpoint molecular mechanics is better suited for refinement restraints because it includes key information that is missing from a statistical set of restraints, most importantly electrostatics and a van der Waals potential that accounts for repulsive/dispersive quantum forces. Current force fields have expanded number of atoms types that surpass the flexibility of restraint sets used in traditional refinement. Improved charge derivation schemes[155], [246], [247] have been created as well as more accurate methods for calculating crystalline lattice electrostatics[45]. Torsion potentials are finely tuned to ever more sophisticated levels of quantum theory and experimental results. Recent work has shown that the inclusion of electrostatics in crystallographic refinement improves results[73], [248] and a number of studies have shown that state of the art molecular dynamics does well with crystals.[62]–[64], [77], [130] Thus the integration of molecular dynamics restraints in crystallographic refinement appears as a natural development of the field.

Perhaps even more importantly, the integration of Amber's molecular dynamics engine Phenix software for crystallography paves the way to the development of more sophisticated applications of great promise. For example, ensemble refinement can now be run using a proper molecular dynamics force field, thus avoiding calculations that lead to poor quality structures in the ensemble. Similarly simulated annealing can now be run with an improved physics based potential. Amber is developing the RISM method for calculation of bulk solvent distribution around molecules that should lead to

improved solvent modeling in refinement. All of these methods stand to significantly contribute to future advances in macromolecular crystallography as an important transition is made from single static structure dominated view of macromolecular crystals to a dynamics and ensemble focus that is more apt to reveal functional relationships in crystals.[52], [59], [249]

7.7. Acknowledgements

We thank Nathaniel Echols, Pavel Afonine, Thomas Terwilliger and Randy Read for stimulating and helpful discussion and ideas.

Bibliography

[1] B. Rupp, *Biomolecular Crystallography*. New York: Garland Science, Taylor & Francis Group, 2010.

[2] D. Blow, *Outline of Crystallography for Biologists*. New York: Oxford University Press, 2002.

[3] C. Hammond, *The Basics of Crystallography and Diffraction*. New York: Oxford University Press, 1997.

[4] A. McPherson, "A brief history of protein crystal growth," *J. Cryst. Growth*, vol. 110, no. 1–2, pp. 1–10, Mar. 1991.

[5] P. Ewald, *Fifty Years of X-Ray Diffraction*. Springer, 1972.

[6] M. Eckert, "Max von Laue and the discovery of X-ray diffraction in 1912," *Ann. Phys.*, vol. 524, no. 5, pp. A83–A85, May 2012.

[7] W. Bragg, "The Specular Reflection of X-rays.," *Nature*, vol. 90, no. 2250, pp. 410–410, Dec. 1912.

[8] W. L. Bragg, "The Structure of Some Crystals as Indicated by Their Diffraction of X-rays," *Proc. R. Soc. A Math. Phys. Eng. Sci.*, vol. 89, no. 610, pp. 248–277, Sep. 1913.

[9] J. C. Kendrew, G. Bodo, H. M. Dintzis, R. G. Parrish, H. Wyckoff, and D. C. Phillips, "A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis," *Nature*, vol. 181, no. 4610, pp. 662–666, Mar. 1958.

[10] E. N. Maslen, A. G. Fox, and M. A. O'Keefe, *International Tables for Crystallography, Volumne C: Mathematical, Physical and Chemical Tables*, 2nd ed. Dordrecht: Kluwer Academic Publishers, 1999.

[11] R. A. Engh and R. Huber, "Accurate bond and angle parameters for X-ray protein structure refinement," *Acta Crystallogr., Sect. A*, vol. 47, no. 4, pp. 392–400, Jul. 1991.

[12] R. A. Engh and R. Huber, "Structure quality and target parameters," in *International Tables for Crystallography. Volume F: Crystallography of Biological Macromolecules*, M. G. Rossman and E. Arnold, Eds. Dordrecht: Kluwer, 2001, pp. 382–392.

[13] R. J. Read, "Improved Fourier coefficients for maps using phases from partial structures with errors," *Acta Crystallogr. Sect. A Found. Crystallogr.*, vol. 42, no. 3, pp. 140–149, May 1986.

[14] D. W. J. Cruickshank, "The determination of the anisotropic thermal motion of atoms in crystals," *Acta Crystallogr.*, vol. 9, no. 9, pp. 747–753, Sep. 1956.

[15] B. T. M. Willis and A. W. Pryor, *Thermal Vibrations in Crystallography*. Cambridge, United Kingdom: Cambridge University Press, 1975.

[16] 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.," *Nucleic Acids Res.*, vol. 28, no. 1, pp. 235–42, Jan. 2000.

[17] M. Allen and D. Tildesley, *Computer Simulations of Liquids*. Oxford University Press, 1989.

[18] J. Haile, *Molecular Dynamics Simulation: Elementary Methods*. New York: John Wiley & Sons, Inc., 1997.

[19] A. Leach, *Molecular Modelling: Principles and Applications*. Prentice Hall, 2001.

[20] T. Schlick, *Molecular Modeling and Simulation*. New York: Springer, 2002.

[21] C. Cramer, *Essentials of Computational Chemistry*. John Wiley & Sons, Inc., 2004.

[22] D. Frenkel and B. Smit, *Understanding Molecular Simulation: from algorithms to applications*. San Diego: Academic Press, 2001.

[23] B. J. Alder and T. E. Wainwright, "Studies in Molecular Dynamics. I. General Method," *J. Chem. Phys.*, vol. 31, no. 2, p. 459, Aug. 1959.

[24] A. Rahman, "Correlations in the Motion of Atoms in Liquid Argon," *Phys. Rev.*, vol. 136, no. 2A, pp. A405–A411, Oct. 1964.

[25] J. A. McCammon, B. R. Gelin, and M. Karplus, "Dynamics of folded proteins," *Nature*, vol. 267, no. 5612, pp. 585–590, Jun. 1977.

[26] D. A. Case, V. Babin, J. T. Berryman, R. M. Betz, Q. Cai, D. S. Cerutti, T. E. Cheatham, III, T. A. Darden, R. E. Duke, H. Gohlke, A. W. Goetz, S. Gusarov, N. Homeyer, P. Janowski, J. Kaus, I. Kolossváry, A. Kovalenko, T. S. Lee, S. LeGrand, T. Luchko, R. Luo, B. Madej, K. M. Merz, F. Paesani, D. R. Roe, A. Roitberg, C. Sagui, R. Salomon-Ferrer, G. Seabra, C. L. Simmerling, W. Smith, J. Swails, R. C. Walker, J. Wang, R. M. Wolf, X. Wu, and P. A. Kollman, "AMBER 14." University of California, San Francisco, 2014.

[27] A.-P. Hynninen and M. F. Crowley, "New faster CHARMM molecular dynamics engine.," *J. Comput. Chem.*, vol. 35, no. 5, pp. 406–13, Feb. 2014.

[28] J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kalé, and K. Schulten, "Scalable molecular dynamics with NAMD.," *J. Comput. Chem.*, vol. 26, no. 16, pp. 1781–802, Dec. 2005.

[29] B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl, "GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation," *J. Chem. Theory Comput.*, vol. 4, no. 3, pp. 435–447, Mar. 2008.

[30] V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, and C. Simmerling, "Comparison of multiple Amber force fields and development of improved protein backbone parameters.," *Proteins*, vol. 65, no. 3, pp. 712–725, 2006.

[31] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman, "A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules," *J. Am. Chem. Soc.*, vol. 117, no. 19, pp. 5179–5197, May 1995.

[32] D. S. Cerutti, W. C. Swope, J. E. Rice, and D. A. Case, "ff14ipq: a self-consistent force field for condensed-phase simulations of proteins.," *J. Chem. Theory Comput.*, vol. 10, no. 10, pp. 4515–4534, Oct. 2014.

[33] K. Vanommeslaeghe and A. D. MacKerell, "CHARMM additive and polarizable force fields for biophysics and computer-aided drug design.," *Biochim. Biophys. Acta*, vol. 1850, no. 5, pp. 861–871, May 2015.

[34] J. W. Ponder, C. Wu, P. Ren, V. S. Pande, J. D. Chodera, M. J. Schnieders, I. Haque, D. L. Mobley, D. S. Lambrecht, R. A. DiStasio, M. Head-Gordon, G. N. I. Clark, M. E. Johnson, and T. Head-Gordon, "Current status of the AMOEBA polarizable force field.," *J. Phys. Chem. B*, vol. 114, no. 8, pp. 2549–64, Mar. 2010.

[35] P. Ren, C. Wu, and J. W. Ponder, "Polarizable Atomic Multipole-based Molecular Mechanics for Organic Molecules.," *J. Chem. Theory Comput.*, vol. 7, no. 10, pp. 3143–3161, Oct. 2011.

[36] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak, "Molecular dynamics with coupling to an external bath," *J. Chem. Phys.*, vol. 81, no. 8, p. 3684, Oct. 1984.

[37] J. A. Izaguirre, D. P. Catarello, J. M. Wozniak, and R. D. Skeel, "Langevin stabilization of molecular dynamics," *J. Chem. Phys.*, vol. 114, no. 5, p. 2090, Feb. 2001.

[38] S. Miyamoto and P. A. Kollman, "Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models," *J. Comput. Chem.*, vol. 13, no. 8, pp. 952–962, 1992.

[39] J. Ryckaert, G. Ciccotti, and H. Berendsen, "Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes," *J. Comput. Phys.*, vol. 23, no. 3, pp. 327–341, Mar. 1977.

[40] H. C. Andersen, "Rattle: A 'velocity' version of the shake algorithm for molecular dynamics calculations," *J. Comput. Phys.*, vol. 52, no. 1, pp. 24–34, Oct. 1983.

[41] D. E. Shaw, "166 Millisecond-long molecular dynamics simulations of proteins on a special-purpose machine," *J. Biomol. Struct. Dyn.*, vol. 31, no. sup1, pp. 108–108, Jan. 2013.

[42] K. J. Bowers, R. O. Dror, and D. E. Shaw, "The midpoint method for parallelization of particle simulations.," *J. Chem. Phys.*, vol. 124, no. 18, p. 184109, May 2006.

[43] K. J. Bowers, R. O. Dror, and D. E. Shaw, "Zonal methods for the parallel execution of range-limited N-body simulations," *J. Comput. Phys.*, vol. 221, no. 1, pp. 303–329, Jan. 2007.

[44] D. M. York, T. A. Darden, and L. G. Pedersen, "The effect of long-range electrostatic interactions in simulations of macromolecular crystals: A comparison of the Ewald and truncated list methods.," *J. Chem. Phys.*, vol. 99, no. 10, pp. 8345–8348, 1993.

[45] T. Darden, D. M. York, and L. Pedersen, "Particle mesh Ewald: An N*log(N) method for Ewald sums in large systems," *J. Chem. Phys.*, vol. 98, no. 12, p. 10089, Jun. 1993.

[46] H. J. C. Berendsen, J. R. Grigera, and T. P. Straatsma, "The missing term in effective pair potentials," *J. Phys. Chem.*, vol. 91, no. 24, pp. 6269–6271, Nov. 1987.

[47] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, "Comparison of simple potential functions for simulating liquid water," *J. Chem. Phys.*, vol. 79, no. 2, p. 926, Jul. 1983.

[48] H. W. Horn, W. C. Swope, J. W. Pitera, J. D. Madura, T. J. Dick, G. L. Hura, and T. Head-Gordon, "Development of an improved four-site water model for biomolecular simulations: TIP4P-Ew.," *J. Chem. Phys.*, vol. 120, no. 20, pp. 9665–78, May 2004.

[49] D. Borek, W. Minor, and Z. Otwinowski, "Measurement errors and their consequences in protein crystallography," *Acta Crystallogr., Sect. D*, vol. 59, no. 11, pp. 2031–2038, Oct. 2003.

[50] J. M. Holton and K. A. Frankel, "The minimum crystal size needed for a complete diffraction data set.," *Acta Crystallogr., Sect. D*, vol. 66, no. Pt 4, pp. 393–408, Apr. 2010.

[51] J. M. Holton, "A beginner's guide to radiation damage.," *J. Synchrotron Radiat.*, vol. 16, no. Pt 2, pp. 133–42, Mar. 2009.

[52] N. Furnham, T. L. Blundell, M. A. DePristo, and T. C. Terwilliger, "Is one solution good enough?," *Nat. Struct. Mol. Biol.*, vol. 13, no. 3, pp. 184–5, Mar. 2006.

[53] B. T. Burnley, P. V Afonine, P. D. Adams, and P. Gros, "Modelling dynamics in protein crystal structures by ensemble refinement," *Elife*, vol. 1, p. e00311, Dec. 2012.

[54] E. J. Levin, D. A. Kondrashov, G. E. Wesenberg, and G. N. Phillips, "Ensemble refinement of protein crystal structures: validation and application," *Structure*, vol. 15, no. 9, pp. 1040–1052, 2007.

[55] J. S. Fraser, H. van den Bedem, A. J. Samelson, P. T. Lang, J. M. Holton, N. Echols, and T. Alber, "Accessing protein conformational ensembles using room-temperature X-ray crystallography.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 108, no. 39, pp. 16247–52, Sep. 2011.

[56] H. van den Bedem, G. Bhabha, K. Yang, P. E. Wright, and J. S. Fraser, "Automated identification of functional dynamic contact networks from X-ray crystallography.," *Nat. Methods*, vol. 10, no. 9, pp. 896–902, Sep. 2013.

[57] P. T. Lang, H.-L. Ng, J. S. Fraser, J. E. Corn, N. Echols, M. Sales, J. M. Holton, and T. Alber, "Automated electron-density sampling reveals widespread conformational polymorphism in proteins.," *Protein Sci.*, vol. 19, pp. 1420–31, Jul. 2010.

[58] J. S. Fraser and C. J. Jackson, "Mining electron density for functionally relevant protein polysterism in crystal structures.," *Cell. Mol. Life Sci.*, vol. 68, no. 11, pp. 1829–41, Jun. 2011.

[59] M. E. Wall, P. D. Adams, J. S. Fraser, and N. K. Sauter, "Diffuse X-ray scattering to model protein motions.," *Structure*, vol. 22, no. 2, pp. 182–4, Feb. 2014.

[60] S. Héry, D. Genest, and J. C. Smith, "X-ray diffuse scattering and rigid-body motion in crystalline lysozyme probed by molecular dynamics simulation.," *J. Mol. Biol.*, vol. 279, no. 1, pp. 303–19, May 1998.

[61] M. E. Wall, A. H. Van Benschoten, N. K. Sauter, P. D. Adams, J. S. Fraser, and T. C. Terwilliger, "Conformational dynamics of a crystalline protein from microsecond-scale molecular dynamics simulations and diffuse X-ray scattering.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 111, no. 50, pp. 17887–92, Dec. 2014.

[62] P. A. Janowski, D. S. Cerutti, J. M. Holton, and D. A. Case, "Peptide crystal simulations reveal hidden dynamics," *J. Am. Chem. Soc.*, vol. 135, no. 21, pp. 7938–7948, 2013.

[63] P. A. Janowski, C. Liu, J. Deckman, and D. A. Case, "Molecular dynamics simulation of triclinic lysozyme in a crystal lattice.," *Protein Sci.*, May 2015.

[64] C. Liu, P. A. Janowski, and D. A. Case, "All-atom crystal simulations of DNA and RNA duplexes.," *Biochim. Biophys. Acta*, vol. 1850, no. 5, pp. 1059–71, May 2015.

[65] D. Kruschel and B. Zagrovic, "Conformational averaging in structural biology: issues, challenges and computational solutions.," *Mol. Biosyst.*, vol. 5, no. 12, pp. 1606–16, Dec. 2009.

[66] B. Zagrovic, Z. Gattin, J. K.-C. Lau, M. Huber, and W. F. van Gunsteren, "Structure and dynamics of two beta-peptides in solution from molecular dynamics simulations validated against experiment.," *Eur. Biophys. J.*, vol. 37, pp. 903–12, 2008.

[67] W. F. van Gunsteren and A. E. Mark, "Validation of molecular dynamics simulation," *J. Chem. Phys.*, vol. 108, p. 6109, 1998.

[68] S. A. Showalter and R. Brüschweiler, "Validation of Molecular Dynamics Simulations of Biomolecules Using NMR Spin Relaxation as Benchmarks: Application to the AMBER99SB Force Field," *J. Chem. Theory Comput.*, vol. 3, pp. 961–975, 2007.

[69] C. Grindon, S. Harris, T. Evans, K. Novik, P. Coveney, and C. Laughton, "Large-scale molecular dynamics simulation of DNA: implementation and validation of the AMBER98 force field in LAMMPS.," *Philos. Trans. R. Soc. London. Ser. A, Math. Phys. Eng. Sci.*, vol. 362, no. 1820, pp. 1373–86, Jul. 2004.

[70] P. L. Freddolino and K. Schulten, "Common structural transitions in explicit-solvent simulations of villin headpiece folding.," *Biophys. J.*, vol. 97, no. 8, pp. 2338–47, Oct. 2009.

[71] A. Nicholls, D. L. Mobley, J. P. Guthrie, J. D. Chodera, C. I. Bayly, M. D. Cooper, and V. S. Pande, "Predicting Small-Molecule Solvation Free Energies: An Informal Blind Test for Computational Chemistry," *J. Med. Chem.*, vol. 51, no. 4, pp. 769–779, Feb. 2008.

[72] L. Meinhold and J. C. Smith, "Protein dynamics from X-ray crystallography: anisotropic, global motion in diffuse scattering patterns.," *Proteins*, vol. 66, no. 4, pp. 941–53, Mar. 2007.

[73] M. J. Schnieders, T. D. Fenn, and V. S. Pande, "Polarizable Atomic Multipole X-Ray Refinement: Particle Mesh Ewald Electrostatics for Macromolecular Crystals," *J. Chem. Theory Comput.*, p. 110309124812033, Mar. 2011.

[74] M. J. Schnieders, T. D. Fenn, V. S. Pande, and A. T. Brunger, "Polarizable atomic multipole X-ray refinement: application to peptide crystals.," *Acta Crystallogr., Sect. D*, vol. 65, no. Pt 9, pp. 952–65, Sep. 2009.

[75] M. D. Tyka, D. A. Keedy, I. André, F. Dimaio, Y. Song, D. C. Richardson, J. S. Richardson, and D. Baker, "Alternate states of proteins revealed by detailed energy landscape mapping.," *J. Mol. Biol.*, vol. 405, no. 2, pp. 607–18, Jan. 2011.

[76] D. S. Cerutti, P. L. Freddolino, R. E. Duke, and D. A. Case, "Simulations of a protein crystal with a high resolution X-ray structure: evaluation of force fields and water models.," *J. Phys. Chem. B*, vol. 114, no. 40, pp. 12811–24, Oct. 2010.

[77] D. S. Cerutti, I. Le Trong, R. E. Stenkamp, and T. P. Lybrand, "Dynamics of the streptavidin-biotin complex in solution and in its crystal lattice: distinct behavior revealed by molecular simulations.," *J. Phys. Chem. B*, vol. 113, no. 19, pp. 6971–85, May 2009.

[78] D. S. Cerutti, I. Le Trong, R. E. Stenkamp, and T. P. Lybrand, "Simulations of a protein crystal: explicit treatment of crystallization conditions links theory and experiment in the streptavidin-biotin complex.," *Biochemistry*, vol. 47, no. 46, pp. 12065–77, Nov. 2008.

[79] S. Aravinda, N. Shamala, C. Das, A. Sriranjini, I. L. Karle, and P. Balaram, "Aromatic-aromatic interactions in crystal structures of helical peptide scaffolds containing projecting phenylalanine residues," *J. Am. Chem. Soc.*, vol. 125, no. 18, pp. 5308–15, 2003.

[80] D. A. Case, T. A. Darden, T. E. I. Cheatham, C. L. Simmerling, J. Wang, R. E. Duke, R. Luo, R. C. Walker, W. Zhang, K. M. Merz, B. Roberts, S. Hayik, A. Roitberg, G. Seabra, J. Swails, A. W. Götz, I. Kolossváry, K. F. Wong, F. Paesani, J. Vanicek, R. M. Wolf, J. Liu, X. Wu, S. R. Brozell, T. Steinbrecher, H. Gohlke, Q. Cai, X. Ye, M.-J. Hsieh, G. Cui, D. R. Roe, D. H. Mathews, M. G. Seetin, R. Salomon-Ferrer, C. Sagui, V. Babin, T. Luchko, S. Gusarov, A. Kovalenko, and P. A. Kollman, "AMBER 12," *Univ. California, San Fr.*, 2012.

[81] J. M. Word, S. C. Lovell, J. S. Richardson, and D. C. Richardson, "Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation.," *J. Mol. Biol.*, vol. 285, no. 4, pp. 1735–47, Jan. 1999.

[82] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, "Comparison of simple potential functions for simulating liquid water," *J. Chem. Phys.*, vol. 79, no. 2, p. 926, Jul. 1983.

[83] J. Wang, P. Cieplak, and P. A. Kollman, "How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules?," *J. Comput. Chem.*, vol. 21, no. 12, pp. 1049–1074, Sep. 2000.

[84] D. S. Cerutti, R. Duke, P. L. Freddolino, H. Fan, and T. P. Lybrand, "Vulnerability in Popular Molecular Dynamics Packages Concerning Langevin and Andersen Dynamics.," *J. Chem. Theory Comput.*, vol. 4, no. 10, pp. 1669–1680, Oct. 2008.

[85] U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee, and L. G. Pedersen, "A smooth particle mesh Ewald method," *J. Chem. Phys.*, vol. 103, no. 19, pp. 8577–8593, 1995.

[86] W. Kabsch, "A discussion of the solution for the best rotation to relate two sets of vectors," *Acta Crystallogr., Sect. A*, vol. 34, no. 5, pp. 922–923, Sep. 1976.

[87] W. Kabsch, "A discussion of the solution for the best rotation to relate two sets of vectors," *Acta Crystallogr., Sect. A*, vol. 34, no. 5, pp. 827–828, Sep. 1978.

[88] W. Kabsch and C. Sander, "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features.," *Biopolymers*, vol. 22, no. 12, pp. 2577–637, Dec. 1983.

[89] G. N. Murshudov, A. A. Vagin, and E. J. Dodson, "Refinement of macromolecular structures by the maximum-likelihood method.," *Acta Crystallogr., Sect. D*, vol. 53, pp. 240–55, May 1997.

[90] P. D. Adams, P. V Afonine, G. Bunkóczi, V. B. Chen, I. W. Davis, N. Echols, J. J. Headd, L.-W. Hung, G. J. Kapral, R. W. Grosse-Kunstleve, A. J. McCoy, N. W. Moriarty, R. Oeffner, R. J. Read, D. C. Richardson, J. S. Richardson, T. C. Terwilliger, and P. H. Zwart, "PHENIX: a comprehensive Python-based system for macromolecular structure solution.," *Acta Crystallogr., Sect. D*, vol. 66, no. 2, pp. 213–21, Feb. 2010.

[91] P. V Afonine, R. W. Grosse-Kunstleve, N. Echols, J. J. Headd, N. W. Moriarty, M. Mustyakimov, T. C. Terwilliger, A. Urzhumtsev, P. H. Zwart, and P. D. Adams, "Towards automated crystallographic structure refinement with phenix.refine.," *Acta Crystallogr., Sect. D*, vol. 68, no. 4, pp. 352–67, Apr. 2012.

[92] W. Humphrey, "VMD: Visual molecular dynamics," *J. Mol. Graph.*, vol. 14, no. 1, pp. 33–38, Feb. 1996.

[93] S. McNicholas, E. Potterton, K. S. Wilson, and M. E. M. Noble, "Presenting your structures: the CCP4mg molecular-graphics software.," *Acta Crystallogr., Sect. D*, vol. 67, pp. 386–94, Apr. 2011.

[94] M. D. Winn, C. C. Ballard, K. D. Cowtan, E. J. Dodson, P. Emsley, P. R. Evans, R. M. Keegan, E. B. Krissinel, A. G. W. Leslie, A. McCoy, S. J. McNicholas, G. N. Murshudov, N. S. Pannu, E. A. Potterton, H. R. Powell, R. J. Read, A. Vagin, and K. S. Wilson, "Overview of the CCP4 suite and current developments.," *Acta Crystallogr., Sect. D*, vol. 67, pp. 235–42, Apr. 2011.

[95] R. C. Agarwal, "A new least-squares refinement technique based on the fast Fourier transform algorithm," *Acta Crystallogr., Sect. A*, vol. 34, pp. 791–809, Sep. 1978.

[96] P. L. Howell and G. D. Smith, "Identification of heavy-atom derivatives by normal probability methods," *J. Appl. Crystallogr.*, vol. 25, no. 1, pp. 81–86, Feb. 1992.

[97] A. T. Brünger, "Assessment of phase accuracy by cross validation: the free R value. Methods and applications.," *Acta Crystallogr., Sect. D*, vol. 49, pp. 24–36, Jan. 1993.

[98] G. J. Kleywegt, "Separating model optimization and model validation in statistical cross-validation as applied to crystallography," *Acta Crystallogr., Sect. D*, vol. 63, no. 9, pp. 939–940, Sep. 2007.

[99] L. A. Baez and P. Clancy, "Existence of a density maximum in extended simple point charge water," *J. Chem. Phys.*, vol. 101, no. 11, p. 9837, Dec. 1994.

[100] A. D. Scouras and V. Daggett, "The Dynameomics rotamer library: amino acid side chain conformations and dynamics from comprehensive molecular dynamics simulations in water.," *Protein Sci.*, vol. 20, no. 2, pp. 341–52, Feb. 2011.

[101] K. Modig, E. Liepinsh, G. Otting, and B. Halle, "Dynamics of protein and peptide hydration.," *J. Am. Chem. Soc.*, vol. 126, no. 1, pp. 102–14, Jan. 2004.

[102] B. Halle, "Cross-relaxation between macromolecular and solvent spins: The role of long-range dipole couplings," *J. Chem. Phys.*, vol. 119, no. 23, p. 12372, Dec. 2003.

[103] A. Lesage, L. Emsley, F. Penin, and A. Böckmann, "Investigation of dipolar-mediated water-protein interactions in microcrystalline Crh by solid-state NMR spectroscopy.," *J. Am. Chem. Soc.*, vol. 128, no. 25, pp. 8246–55, Jun. 2006.

[104] K. A. Dill, "Dominant forces in protein folding," *Biochemistry*, vol. 29, no. 31, pp. 7133–7155, Aug. 1990.

[105] O. F. Lange, D. van der Spoel, and B. L. de Groot, "Scrutinizing molecular mechanics force fields on the submicrosecond timescale with NMR data.," *Biophys. J.*, vol. 99, no. 2, pp. 647–55, Jul. 2010.

[106] T. Kortemme, A. V. Morozov, and D. Baker, "An Orientation-dependent Hydrogen Bonding Potential Improves Prediction of Specificity and Structure for Proteins and Protein–Protein Complexes," *J. Mol. Biol.*, vol. 326, no. 4, pp. 1239–1259, Feb. 2003.

[107] P. Ren and J. W. Ponder, "Polarizable Atomic Multipole Water Model for Molecular Mechanics Simulation," *J. Phys. Chem. B*, vol. 107, no. 24, pp. 5933–5947, Jun. 2003.

[108] A. Kuzmanic, N. S. Pannu, and B. Zagrovic, "X-ray refinement significantly underestimates the level of microscopic heterogeneity in biomolecular crystals.," *Nat. Commun.*, vol. 5, p. 3220, Jan. 2014.

[109] A. T. Brünger, "Free R value: Cross-validation in crystallography," *Methods Enzymol.*, vol. 277, pp. 366–396, 1997.

[110] D. R. Roe and T. E. Cheatham, "PTRAJ and CPPTRAJ: software for processing and analysis of molecular dynamics trajectory data," *J. Chem. Theory Comput.*, vol. 9, no. 7, pp. 3084–3095, Jul. 2013.

[111] 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.*, vol. 281, no. 18, pp. 4046–60, Sep. 2014.

[112] K. Henzler-Wildman and D. Kern, "Dynamic personalities of proteins.," *Nature*, vol. 450, no. 7172, pp. 964–72, Dec. 2007.

[113] D. Kern and E. R. Zuiderweg, "The role of dynamics in allosteric regulation," *Curr. Opin. Struct. Biol.*, vol. 13, no. 6, pp. 748–757, Dec. 2003.

[114] C.-J. Tsai, A. Del Sol, and R. Nussinov, "Protein allostery, signal transmission and dynamics: a classification scheme of allosteric mechanisms.," *Mol. Biosyst.*, vol. 5, no. 3, pp. 207–16, Mar. 2009.

[115] K. K. Frederick, M. S. Marlow, K. G. Valentine, and A. J. Wand, "Conformational entropy in molecular recognition by proteins.," *Nature*, vol. 448, no. 7151, pp. 325–9, Jul. 2007.

[116] A. Schmidt and V. S. Lamzin, "Extraction of functional motion in trypsin crystal structures.," *Acta Crystallogr., Sect. D*, vol. 61, no. 8, pp. 1132–9, Aug. 2005.

[117] J. E. Kohn, P. V Afonine, J. Z. Ruscio, P. D. Adams, and T. Head-Gordon, "Evidence of functional protein dynamics from X-ray crystallographic ensembles.," *PLoS Comput. Biol.*, vol. 6, no. 8, p. e1000911, Jan. 2010.

[118] B. Halle, "Flexibility and packing in proteins," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 99, no. 3, pp. 1274–1279, Jan. 2002.

[119] M. A. DePristo, P. I. de Bakker, and T. L. Blundell, "Heterogeneity and inaccuracy in protein structures solved by X-ray crystallography," *Structure*, vol. 12, no. 5, pp. 831–838, 2004.

[120] H. van den Bedem, A. Dhanik, J. C. Latombe, and A. M. Deacon, "Modeling discrete heterogeneity in X-ray diffraction data by fitting multi-conformers.," *Acta Crystallogr., Sect. D*, vol. 65, no. 10, pp. 1107–17, Oct. 2009.

[121] 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. U. S. A.*, vol. 111, no. 1, pp. 237–42, Jan. 2014.

[122] Z. Ren, P. W. Y. Chan, K. Moffat, E. F. Pai, W. E. Royer, V. Šrajer, and X. Yang, "Resolution of structural heterogeneity in dynamic crystallography.," *Acta Crystallogr., Sect. D*, vol. 69, no. 6, pp. 946–59, Jun. 2013.

[123] H. Heldenbrand, P. A. Janowski, G. Giambaşu, T. J. Giese, J. E. Wedekind, and D. M. York, "Evidence for the role of active site residues in the hairpin ribozyme from molecular simulations along the reaction path.," *J. Am. Chem. Soc.*, vol. 136, no. 22, pp. 7789–92, Jun. 2014.

[124] A. T. Brünger, J. Kuriyan, and M. Karplus, "Crystallographic R factor refinement by molecular dynamics.," *Science*, vol. 235, no. 4787, pp. 458–60, Jan. 1987.

[125] J. Hafner and W. Zheng, "All-atom modeling of anisotropic atomic fluctuations in protein crystal structures.," *J. Chem. Phys.*, vol. 135, no. 14, p. 144114, Oct. 2011.

[126] F. Avbelj, J. Moult, D. H. Kitson, M. N. G. James, and A. T. Hagler, "Molecular dynamics study of the structure and dynamics of a protein molecule in a crystalline ionic environment, Streptomyces griseus protease A," *Biochemistry*, vol. 29, no. 37, pp. 8658–8676, Sep. 1990.

[127] D. H. Kitson and A. T. Hagler, "Theoretical studies of the structure and molecular dynamics of a peptide crystal," *Biochemistry*, vol. 27, no. 14, pp. 5246–5257, Jul. 1988.

[128] J. Lautz, H. Kessler, R. Kaptein, and W. F. van Gunsteren, "Molecular dynamics simulations of cyclosporin A: the crystal structure and dynamic modelling of a structure in apolar solution based on NMR data.," *J. Comput. Aided. Mol. Des.*, vol. 1, no. 3, pp. 219–41, Oct. 1987.

[129] D. Vitkup, D. Ringe, M. Karplus, and G. A. Petsko, "Why protein R-factors are so large: a self-consistent analysis.," *Proteins*, vol. 46, no. 4, pp. 345–54, Mar. 2002.

[130] D. M. York, A. Wlodawer, L. G. Pedersen, and T. A. Darden, "Atomic-level accuracy in simulations of large protein crystals.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 91, no. 18, pp. 8715–8718, Aug. 1994.

[131] C. Brinkmann, M. S. Weiss, and E. Weckert, "The structure of the hexagonal crystal form of hen egg-white lysozyme.," *Acta Crystallogr., Sect. D*, vol. 62, no. Pt 4, pp. 349–55, Apr. 2006.

[132] J. Held and S. van Smaalen, "The active site of hen egg-white lysozyme: flexibility and chemical bonding," *Acta Crystallogr., Sect. D*, vol. 70, no. 4, pp. 1136–1146, Mar. 2014.

[133] K. Harata and T. Akiba, "Structural phase transition of monoclinic crystals of hen egg-white lysozyme.," *Acta Crystallogr., Sect. D*, vol. 62, no. Pt 4, pp. 375–82, Apr. 2006.

[134] P. J. Artymiuk, C. C. Blake, D. E. Grace, S. J. Oatley, D. C. Phillips, and M. J. Sternberg, "Crystallographic studies of the dynamic properties of lysozyme," *Nature*, vol. 280, no. 5723, pp. 563–568, Aug. 1979.

[135] K. Hinsen, "Structural flexibility in proteins: impact of the crystal environment.," *Bioinformatics*, vol. 24, no. 4, pp. 521–8, Feb. 2008.

[136] S. Hery, D. Genest, and J. C. Smith, "Fluctuation and correlation in crystalline lysozyme," *J. Chem. Inf. Model.*, vol. 37, no. 6, pp. 1011–1017, Nov. 1997.

[137] U. Stocker, K. Spiegel, and W. F. van Gunsteren, "On the similarity of properties in solution or in the crystalline state: A molecular dynamics study of hen lysozyme," *J. Biomol. NMR*, vol. 18, no. 1, pp. 1–12, 2000.

[138] Z. Hu and J. Jiang, "Assessment of biomolecular force fields for molecular dynamics simulations in a protein crystal.," *J. Comput. Chem.*, vol. 31, no. 2, pp. 371–80, Jan. 2010.

[139] Z. Hu and J. Jiang, "Molecular dynamics simulations for water and ions in protein crystals.," *Langmuir*, vol. 24, no. 8, pp. 4215–23, Apr. 2008.

[140] M. A. Walsh, T. R. Schneider, L. C. Sieker, Z. Dauter, V. S. Lamzin, and K. S. Wilson, "Refinement of triclinic hen egg-white lysozyme at atomic resolution," *Acta Crystallogr., Sect. D*, vol. 54, no. 4, pp. 522–546, Jul. 1998.

[141] B. Zagrovic and V. S. Pande, "How does averaging affect protein structure comparison on the ensemble level?," *Biophys. J.*, vol. 87, no. 4, pp. 2240–6, Oct. 2004.

[142] P. Emsley and K. Cowtan, "Coot: model-building tools for molecular graphics.," *Acta Crystallogr., Sect. D*, vol. 60, no. 12, pp. 2126–32, Dec. 2004.

[143] L. Zhou and Q. Liu, "Aligning experimental and theoretical anisotropic B-factors: water models, normal-mode analysis methods, and metrics.," *J. Phys. Chem. B*, vol. 118, no. 15, pp. 4069–79, Apr. 2014.

[144] S. Kundu, J. S. Melton, D. C. Sorensen, and G. N. Phillips, "Dynamics of proteins in crystals: comparison of experiment with simple models.," *Biophys. J.*, vol. 83, no. 2, pp. 723–32, Aug. 2002.

[145] T. Z. Sen, Y. Feng, J. V Garcia, A. Kloczkowski, and R. L. Jernigan, "The extent of cooperativity of protein motions observed with elastic network models is similar for atomic and coarser-grained models.," *J. Chem. Theory Comput.*, vol. 2, no. 3, pp. 696–704, Jan. 2006.

[146] D. Riccardi, Q. Cui, and G. N. Phillips, "Application of elastic network models to proteins in the crystalline state.," *Biophys. J.*, vol. 96, no. 2, pp. 464–75, Jan. 2009.

[147] L. Yang, G. Song, and R. L. Jernigan, "Comparisons of experimental and computed protein anisotropic temperature factors.," *Proteins*, vol. 76, no. 1, pp. 164–75, Jul. 2009.

[148] L.-W. Yang, E. Eyal, C. Chennubhotla, J. Jee, A. M. Gronenborn, and I. Bahar, "Insights into equilibrium dynamics of proteins from comparison of NMR and X-ray data with computational predictions.," *Structure*, vol. 15, no. 6, pp. 741–9, Jun. 2007.

[149] D. A. Kondrashov, A. W. Van Wynsberghe, R. M. Bannen, Q. Cui, and G. N. Phillips, "Protein structural variation in computational models and crystallographic data," *Structure*, vol. 15, no. 2, pp. 169–177, Feb. 2007.

[150] M. Rueda, C. Ferrer-Costa, T. Meyer, A. Pérez, J. Camps, A. Hospital, J. L. Gelpí, and M. Orozco, "A consensus view of protein dynamics.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 104, no. 3, pp. 796–801, Jan. 2007.

[151] J. Kuriyan, G. A. Petsko, R. M. Levy, and M. Karplus, "Effect of anisotropy and anharmonicity on protein crystallographic refinement," *J. Mol. Biol.*, vol. 190, no. 2, pp. 227–254, 1986.

[152] K. Harata and T. Akiba, "Phase transition of triclinic hen egg-white lysozyme crystal associated with sodium binding.," *Acta Crystallogr., Sect. D*, vol. 60, no. Pt 4, pp. 630–7, Apr. 2004.

[153] W. R. P. Scott, P. H. Hünenberger, I. G. Tironi, A. E. Mark, S. R. Billeter, J. Fennen, A. E. Torda, T. Huber, P. Krüger, and W. F. van Gunsteren, "The GROMOS biomolecular simulation program package," *J. Phys. Chem. A*, vol. 103, no. 19, pp. 3596–3607, May 1999.

[154] Y. Xue and N. R. Skrynnikov, "Ensemble MD simulations restrained via crystallographic data: accurate structure leads to accurate dynamics.," *Protein Sci.*, vol. 23, no. 4, pp. 488–507, Apr. 2014.

[155] D. S. Cerutti, J. E. Rice, W. C. Swope, and D. A. Case, "Derivation of fixed partial charges for amino acids accommodating a specific water model and implicit polarization.," *J. Phys. Chem. B*, vol. 117, no. 8, pp. 2328–38, Feb. 2013.

[156] "www.pharmacy.manchester.ac.uk/bryce/amber/." .

[157] R. B. Best, X. Zhu, J. Shim, P. E. M. Lopes, J. Mittal, M. Feig, and A. D. Mackerell, "Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone φ, ψ and side-chain χ(1) and χ(2) dihedral angles.," *J. Chem. Theory Comput.*, vol. 8, no. 9, pp. 3257–3273, Sep. 2012.

[158] H. W. Horn, W. C. Swope, and J. W. Pitera, "Characterization of the TIP4P-Ew water model: vapor pressure and boiling point.," *J. Chem. Phys.*, vol. 123, no. 19, p. 194504, Nov. 2005.

[159] I. S. Joung and T. E. Cheatham, "Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations.," *J. Phys. Chem. B*, vol. 112, no. 30, pp. 9020–41, Jul. 2008.

[160] D. Frishman and P. Argos, "Knowledge-based protein secondary structure assignment.," *Proteins*, vol. 23, no. 4, pp. 566–79, Dec. 1995.

[161] L. L. C. Schrödinger, "The PyMOL Molecular Graphics System, Version~1.6.0." Aug-2010.

[162] J. D. Hunter, "Matplotlib: A 2D Graphics Environment," *Comput. Sci. Eng.*, vol. 9, no. 3, pp. 90–95, 2007.

[163] H. KIM, B. MHIN, C. YOON, C. WANG, and K. KIM, "A THEORETICAL-STUDY OF A Z-DNA CRYSTAL - STRUCTURE OF COUNTERIONS, WATER AND DNA-MOLECULES," *Bull. KOREAN Chem. Soc.*, vol. 12, no. 2, pp. 214–219, Apr. 1991.

[164] A. D. MacKerell, J. Wiorkiewicz-Kuczera, and M. Karplus, "An all-atom empirical energy function for the simulation of nucleic acids," *J. Am. Chem. Soc.*, vol. 117, no. 48, pp. 11946–11975, Dec. 1995.

[165] T. E. Cheatham and D. A. Case, "Twenty-five years of nucleic acid simulations.," *Biopolymers*, vol. 99, no. 12, pp. 969–77, Dec. 2013.

[166] T. E. I. Cheatham, J. L. Miller, T. Fox, T. A. Darden, and P. A. Kollman, "Molecular Dynamics Simulations on Solvated Biomolecular Systems: The Particle Mesh Ewald Method Leads to Stable Trajectories of DNA, RNA, and Proteins," *J. Am. Chem. Soc.*, vol. 117, no. 14, pp. 4193–4194, Apr. 1995.

[167] D. M. York, W. Yang, H. Lee, T. Darden, and L. G. Pedersen, "Toward the Accurate Modeling of DNA: The Importance of Long-Range Electrostatics," *J. Am. Chem. Soc.*, vol. 117, no. 17, pp. 5001–5002, May 1995.

[168] H. Lee, T. Darden, and L. Pedersen, "Accurate crystal molecular dynamics simulations using particle-mesh-Ewald: RNA dinucleotides — ApU and GpC," *Chem. Phys. Lett.*, vol. 243, no. 3–4, pp. 229–235, Sep. 1995.

[169] V. Babin, J. Baucom, T. A. Darden, and C. Sagui, "Molecular dynamics simulations of polarizable DNA in crystal environment," *Int. J. Quantum Chem.*, vol. 106, no. 15, pp. 3260–3269, 2006.

[170] V. Babin, J. Baucom, T. A. Darden, and C. Sagui, "Molecular dynamics simulations of DNA with polarizable force fields: convergence of an ideal B-DNA structure to the crystallographic structure.," *J. Phys. Chem. B*, vol. 110, no. 23, pp. 11571–81, Jun. 2006.

[171] Z. Gong, Y. Xiao, and Y. Xiao, "RNA stability under different combinations of amber force fields and solvation models.," *J. Biomol. Struct. Dyn.*, vol. 28, no. 3, pp. 431–41, Dec. 2010.

[172] V. Babin, D. Wang, R. B. Rose, and C. Sagui, "Binding polymorphism in the DNA bound state of the Pdx1 homeodomain.," *PLoS Comput. Biol.*, vol. 9, no. 8, p. e1003160, Jan. 2013.

[173] A. Kuzmanic and B. Zagrovic, "Dependence of protein crystal stability on residue charge states and ion content of crystal solvent.," *Biophys. J.*, vol. 106, no. 3, pp. 677–86, Feb. 2014.

[174] L. S. Ahlstrom and O. Miyashita, "Packing interface energetics in different crystal forms of the λ Cro dimer.," *Proteins*, vol. 82, no. 7, pp. 1128–41, Jul. 2014.

[175] Z. Hu, J. Jiang, and S. I. Sandler, "Water in hydrated orthorhombic lysozyme crystal: Insight from atomistic simulations.," *J. Chem. Phys.*, vol. 129, no. 7, p. 075105, Aug. 2008.

[176] K. Grzeskowiak, K. Yanagi, G. G. Privé, and R. E. Dickerson, "The structure of B-helical C-G-A-T-C-G-A-T-C-G and comparison with C-C-A-A-C-G-T-T-G-G - The effect of base pair reversals," *J. Biol. Chem.*, vol. 266, no. 14, pp. 8861–8883, 1991.

[177] A. C. Dock-Bregeon, B. Chevrier, A. Podjarny, D. Moras, J. S. deBear, G. R. Gough, P. T. Gilham, and J. E. Johnson, "High resolution structure of the RNA duplex [U(U-A)6A]2.," *Nature*, vol. 335, no. 6188, pp. 375–8, Sep. 1988.

[178] A. C. Dock-Bregeon, B. Chevrier, A. Podjarny, J. Johnson, J. S. de Bear, G. R. Gough, P. T. Gilham, and D. Moras, "Crystallographic structure of an RNA helix: [U(UA)6A]2," *J. Mol. Biol.*, vol. 209, no. 3, pp. 459–474, Oct. 1989.

[179] I. Beššeová, P. Banáš, P. Kührová, P. Košinová, M. Otyepka, and J. Šponer, "Simulations of A-RNA duplexes. The effect of sequence, solute force field, water model, and salt concentration.," *J. Phys. Chem. B*, vol. 116, no. 33, pp. 9899–916, Aug. 2012.

[180] I. Besseová, M. Otyepka, K. Réblová, and J. Sponer, "Dependence of A-RNA simulations on the choice of the force field and salt strength.," *Phys. Chem. Chem. Phys.*, vol. 11, no. 45, pp. 10701–11, Dec. 2009.

[181] S. B. Dixit, D. L. Beveridge, D. A. Case, T. E. Cheatham, E. Giudice, F. Lankas, R. Lavery, J. H. Maddocks, R. Osman, H. Sklenar, K. M. Thayer, and P. Varnai, "Molecular dynamics simulations of the 136 unique tetranucleotide sequences of DNA oligonucleotides. II: sequence context effects on the dynamical structures of the 10 unique dinucleotide steps.," *Biophys. J.*, vol. 89, no. 6, pp. 3721–40, Dec. 2005.

[182] R. Lavery, K. Zakrzewska, D. Beveridge, T. C. Bishop, D. A. Case, T. Cheatham, S. Dixit, B. Jayaram, F. Lankas, C. Laughton, J. H. Maddocks, A. Michon, R. Osman, M. Orozco, A. Perez, T. Singh, N. Spackova, and J. Sponer, "A systematic molecular dynamics study of nearest-neighbor effects on base pair and base pair step conformations and fluctuations in B-DNA.," *Nucleic Acids Res.*, vol. 38, no. 1, pp. 299–313, Jan. 2010.

[183] D. L. Beveridge, T. E. Cheatham, and M. Mezei, "The ABCs of molecular dynamics simulations on B-DNA, circa 2012," *J. Biosci.*, vol. 37, no. 3, pp. 379–397, Jun. 2012.

[184] M. Pasi, J. H. Maddocks, D. Beveridge, T. C. Bishop, D. A. Case, T. Cheatham, P. D. Dans, B. Jayaram, F. Lankas, C. Laughton, J. Mitchell, R. Osman, M. Orozco, A. Pérez, D. Petkevičiūtė, N. Spackova, J. Sponer, K. Zakrzewska, and R. Lavery, "μABC: a systematic microsecond molecular dynamics study of tetranucleotide sequence effects in B-DNA.," *Nucleic Acids Res.*, vol. 42, no. 19, pp. 12272–83, Oct. 2014.

[185] A. Pérez, F. J. Luque, and M. Orozco, "Dynamics of B-DNA on the microsecond time scale.," *J. Am. Chem. Soc.*, vol. 129, no. 47, pp. 14739–45, Nov. 2007.

[186] A. Pérez, F. J. Luque, and M. Orozco, "Frontiers in molecular dynamics simulations of DNA.," *Acc. Chem. Res.*, vol. 45, no. 2, pp. 196–205, Feb. 2012.

[187] S. Teletchea, B. Hartmann, and J. Kozelka, "Discrimination between BI and BII conformational substates of B-DNA based on sugar-base interproton distances.," *J. Biomol. Struct. Dyn.*, vol. 21, no. 4, pp. 489–94, Feb. 2004.

[188] B. Heddi, N. Foloppe, N. Bouchemal, E. Hantz, and B. Hartmann, "Quantification of DNA BI/BII backbone states in solution. Implications for DNA overall structure and recognition.," *J. Am. Chem. Soc.*, vol. 128, no. 28, pp. 9170–7, Jul. 2006.

[189] M. Guéroult, O. Boittin, O. Mauffret, C. Etchebest, and B. Hartmann, "Mg2+ in the major groove modulates B-DNA structure and dynamics.," *PLoS One*, vol. 7, no. 7, p. e41704, Jan. 2012.

[190] D. A. Case, T. E. Cheatham, T. Darden, H. Gohlke, R. Luo, K. M. Merz, A. Onufriev, C. Simmerling, B. Wang, and R. J. Woods, "The Amber biomolecular simulation programs.," *J. Comput. Chem.*, vol. 26, no. 16, pp. 1668–88, Dec. 2005.

[191] A. Pérez, I. Marchán, D. Svozil, J. Sponer, T. E. Cheatham, C. A. Laughton, and M. Orozco, "Refinement of the AMBER force field for nucleic acids: improving the description of alpha/gamma conformers.," *Biophys. J.*, vol. 92, no. 11, pp. 3817–29, Jun. 2007.

[192] M. Zgarbová, M. Otyepka, J. Sponer, A. Mládek, P. Banáš, T. E. Cheatham, and P. Jurečka, "Refinement of the Cornell et al. Nucleic Acids Force Field Based on Reference Quantum Chemical Calculations of Glycosidic Torsion Profiles.," *J. Chem. Theory Comput.*, vol. 7, no. 9, pp. 2886–2902, Sep. 2011.

[193] J. Åqvist, "Ion-water interaction potentials derived from free energy perturbation simulations," *J. Phys. Chem.*, vol. 94, no. 21, pp. 8021–8024, 1990.

[194] R. Salomon-Ferrer, D. A. Case, and R. C. Walker, "An overview of the Amber biomolecular simulation package," *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, vol. 3, no. 2, pp. 198–210, Mar. 2013.

[195] R. Salomon-Ferrer, A. W. Götz, D. Poole, S. Le Grand, and R. C. Walker, "Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald," *J. Chem. Theory Comput.*, vol. 9, no. 9, pp. 3878–3888, Sep. 2013.

[196] D. L. Beveridge, G. Barreiro, K. S. Byun, D. A. Case, T. E. Cheatham, S. B. Dixit, E. Giudice, F. Lankas, R. Lavery, J. H. Maddocks, R. Osman, E. Seibert, H. Sklenar, G. Stoll, K. M. Thayer, P. Varnai, and M. A. Young, "Molecular dynamics simulations of the 136 unique tetranucleotide sequences of DNA oligonucleotides. I. Research design and results on d(CpG) steps.," *Biophys. J.*, vol. 87, no. 6, pp. 3799–813, Dec. 2004.

[197] R. Lavery, M. Moakher, J. H. Maddocks, D. Petkeviciute, and K. Zakrzewska, "Conformational analysis of nucleic acids revisited: Curves+.," *Nucleic Acids Res.*, vol. 37, no. 17, pp. 5917–29, Sep. 2009.

[198] E. Krissinel and K. Henrick, "Inference of Macromolecular Assemblies from Crystalline State," *J. Mol. Biol.*, vol. 372, no. 3, pp. 774–797, 2007.

[199] B. S. Tolbert, Y. Miyazaki, S. Barton, B. Kinde, P. Starck, R. Singh, A. Bax, D. A. Case, and M. F. Summers, "Major groove width variations in RNA structures determined by NMR and impact of 13C residual chemical shift anisotropy and 1H-13C residual dipolar coupling on refinement.," *J. Biomol. NMR*, vol. 47, no. 3, pp. 205–19, Jul. 2010.

[200] D. R. Mack, T. K. Chiu, and R. E. Dickerson, "Intrinsic bending and deformability at the T-A step of CCTTTAAAGG: a comparative analysis of T-A and A-T steps within A-tracts.," *J. Mol. Biol.*, vol. 312, no. 5, pp. 1037–49, Oct. 2001.

[201] M. Poncin, B. Hartmann, and R. Lavery, "Conformational sub-states in B-DNA," *J. Mol. Biol.*, vol. 226, no. 3, pp. 775–794, Aug. 1992.

[202] R. E. Dickerson, "Definitions and nomenclature of nucleic acid structure parameters.," *J. Biomol. Struct. Dyn.*, vol. 6, no. 4, pp. 627–34, Feb. 1989.

[203] I. Faustino, A. Pérez, and M. Orozco, "Toward a consensus view of duplex RNA flexibility.," *Biophys. J.*, vol. 99, no. 6, pp. 1876–85, Sep. 2010.

[204] P. D. Dans, A. Pérez, I. Faustino, R. Lavery, and M. Orozco, "Exploring polymorphisms in B-DNA helical conformations.," *Nucleic Acids Res.*, vol. 40, no. 21, pp. 10668–78, Nov. 2012.

[205] W. Li, L. Nordenskiöld, and Y. Mu, "Sequence-specific Mg2+-DNA interactions: a molecular dynamics simulation study.," *J. Phys. Chem. B*, vol. 115, no. 49, pp. 14713–20, Dec. 2011.

[206] J. P. Priestle, "Improved dihedral-angle restraints for protein structure refinement," *J. Appl. Crystallogr.*, vol. 36, no. 1, pp. 34–42, Jan. 2003.

[207] W. G. Touw and G. Vriend, "On the complexity of Engh and Huber refinement restraints: the angle τ as example.," *Acta Crystallogr., Sect. D*, vol. 66, no. 12, pp. 1341–50, Dec. 2010.

[208] A. M. Davis, S. J. Teague, and G. J. Kleywegt, "Application and limitations of X-ray crystallographic data in structure-based ligand and drug design.," *Angew. Chem. Int. Ed. Engl.*, vol. 42, no. 24, pp. 2718–36, Jun. 2003.

[209] N. W. Moriarty, D. E. Tronrud, P. D. Adams, and P. A. Karplus, "Conformation-dependent backbone geometry restraints set a new standard for protein crystallographic refinement," *FEBS J.*, vol. 281, no. 18, pp. 4061–4071, 2014.

[210] G. J. Kleywegt, K. Henrick, E. J. Dodson, and D. M. F. van Aalten, "Pound-Wise but Penny-Foolish," *Structure*, vol. 11, no. 9, pp. 1051–1059, Sep. 2003.

[211] G. J. Kleywegt and T. A. Jones, "Databases in Protein Crystallography," *Acta Crystallogr., Sect. D*, vol. 54, no. 6, pp. 1119–1131, Nov. 1998.

[212] J. Liebeschuetz, J. Hennemann, T. Olsson, and C. R. Groom, "The good, the bad and the twisted: a survey of ligand geometry in protein crystal structures.," *J. Comput. Aided. Mol. Des.*, vol. 26, no. 2, pp. 169–83, Feb. 2012.

[213] N. W. Moriarty, R. W. Grosse-Kunstleve, and P. D. Adams, "electronic Ligand Builder and Optimization Workbench (eLBOW): a tool for ligand coordinate and restraint generation.," *Acta Crystallogr., Sect. D*, vol. 65, no. Pt 10, pp. 1074–80, Oct. 2009.

[214] O. S. . Smart, T. O. . Womack, A. . Sharff, C. . Flensburg, P. . Keller, W. . Paciorek, C. . Vonrhein, and G. Bricogne, "Grade, version 1.1.1." http://www.globalphasing.com, 2011.

[215] A. A. Lebedev, P. Young, M. N. Isupov, O. V Moroz, A. A. Vagin, and G. N. Murshudov, "JLigand: a graphical tool for the CCP4 template-restraint library.," *Acta Crystallogr., Sect. D*, vol. 68, no. Pt 4, pp. 431–40, May 2012.

[216] A. W. Schüttelkopf and D. M. F. van Aalten, "PRODRG: a tool for high-throughput crystallography of protein-ligand complexes.," *Acta Crystallogr., Sect. D*, vol. 60, no. Pt 8, pp. 1355–63, Aug. 2004.

[217] S. Wlodek, A. G. Skillman, and A. Nicholls, "Automated ligand placement and refinement with a combined force field and shape potential.," *Acta Crystallogr., Sect. D*, vol. 62, no. Pt 7, pp. 741–9, Jul. 2006.

[218] I. D. Brown and B. McMahon, "CIF: the computer language of crystallography," *Acta Crystallogr., Sect. B*, vol. 58, no. 3, pp. 317–324, May 2002.

[219] S. R. Hall, F. H. Allen, and I. D. Brown, "The crystallographic information file (CIF): a new standard archive file for crystallography," *Acta Crystallogr., Sect. A*, vol. 47, no. 6, pp. 655–685, Nov. 1991.

[220] O. Y. Borbulevych, J. A. Plumley, R. I. Martin, K. M. Merz, and L. M. Westerhoff, "Accurate macromolecular crystallographic refinement: incorporation of the linear scaling, semiempirical quantum-mechanics program DivCon into the PHENIX refinement package.," *Acta Crystallogr., Sect. D*, vol. 70, no. Pt 5, pp. 1233–47, May 2014.

[221] O. Y. Borbulevych, N. W. Moriarty, P. D. Adams, and L. M. Westerhoff, "Quantum Mechanics-based Refinement in Phenix/DivCon," *Comput. Crystallogr. Newsl.*, vol. 5, pp. 26–30, 2014.

[222] Z. Fu, X. Li, and K. M. Merz, "Accurate assessment of the strain energy in a protein-bound drug using QM/MM X-ray refinement and converged quantum chemistry.," *J. Comput. Chem.*, vol. 32, no. 12, pp. 2587–97, Sep. 2011.

[223] N. Yu, H. P. Yennawar, and K. M. Merz, "Refinement of protein crystal structures using energy restraints derived from linear-scaling quantum mechanics.," *Acta Crystallogr., Sect. D*, vol. 61, no. Pt 3, pp. 322–32, Mar. 2005.

[224] T. A. Halgren, "Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94," *J. Comput. Chem.*, vol. 17, no. 5–6, pp. 490–519, Apr. 1996.

[225] T. A. Halgren, "Merck molecular force field. V. Extension of MMFF94 using experimental data, additional computational data, and empirical rules," *J. Comput. Chem.*, vol. 17, no. 5–6, pp. 616–641, Apr. 1996.

[226] T. A. Halgren and R. B. Nachbar, "Merck molecular force field. IV. conformational energies and geometries for MMFF94," *J. Comput. Chem.*, vol. 17, no. 5–6, pp. 587–615, Apr. 1996.

[227] T. A. Halgren, "Merck molecular force field. III. Molecular geometries and vibrational frequencies for MMFF94," *J. Comput. Chem.*, vol. 17, no. 5–6, pp. 553–586, Apr. 1996.

[228] J. A. Grant, B. T. Pickup, M. J. Sykes, C. A. Kitchen, and A. Nicholls, "A simple formula for dielectric polarisation energies: The Sheffield Solvation Model," *Chem. Phys. Lett.*, vol. 441, no. 1–3, pp. 163–166, Jun. 2007.

[229] T. A. Halgren, "Merck molecular force field. II. MMFF94 van der Waals and electrostatic parameters for intermolecular interactions," *J. Comput. Chem.*, vol. 17, no. 5–6, pp. 520–552, Apr. 1996.

[230] T. A. Halgren, "MMFF VII. Characterization of MMFF94, MMFF94s, and other widely available force fields for conformational energies and for intermolecular-interaction energies and geometries," *J. Comput. Chem.*, vol. 20, no. 7, pp. 730–748, May 1999.

[231] K. Gundertofte, T. Liljefors, P. Norrby, and I. Pettersson, "A comparison of conformational energies calculated by several molecular mechanics methods," *J. Comput. Chem.*, vol. 17, no. 4, pp. 429–449, Mar. 1996.

[232] G. L. Warren, T. D. Do, B. P. Kelley, A. Nicholls, and S. D. Warren, "Essential considerations for using protein-ligand structures in drug discovery.," *Drug Discov. Today*, vol. 17, no. 23–24, pp. 1270–81, Dec. 2012.

[233] I. J. Bruno, J. C. Cole, M. Kessler, J. Luo, W. D. S. Motherwell, L. H. Purkis, B. R. Smith, R. Taylor, R. I. Cooper, S. E. Harris, and A. G. Orpen, "Retrieval of crystallographically-derived molecular geometry information.," *J. Chem. Inf. Comput. Sci.*, vol. 44, no. 6, pp. 2133–44, Jan. 2004.

[234] V. B. Chen, W. B. Arendall, J. J. Headd, D. A. Keedy, R. M. Immormino, G. J. Kapral, L. W. Murray, J. S. Richardson, and D. C. Richardson, "MolProbity: all-atom structure validation for macromolecular crystallography.," *Acta Crystallogr., Sect. D*, vol. 66, no. 1, pp. 12–21, Jan. 2010.

[235] A. Jack and M. Levitt, "Refinement of large structures by simultaneous minimization of energy and *R* factor," *Acta Crystallogr., Sect. A*, vol. 34, no. 6, pp. 931–935, Nov. 1978.

[236] J. Waser, "Least-squares refinement with subsidiary conditions," *Acta Crystallogr.*, vol. 16, no. 11, pp. 1091–1094, Nov. 1963.

[237] G. N. Murshudov, P. Skubák, A. A. Lebedev, N. S. Pannu, R. A. Steiner, R. A. Nicholls, M. D. Winn, F. Long, and A. A. Vagin, "REFMAC5 for the refinement of macromolecular crystal structures.," *Acta Crystallogr., Sect. D*, vol. 67, no. 4, pp. 355–67, Apr. 2011.

[238] G. M. Sheldrick, "A short history of *SHELX*," *Acta Crystallogr., Sect. A*, vol. 64, no. 1, pp. 112–122, Jan. 2008.

[239] G. Bricogne, E. Blanc, M. Brandl, C. Flensburg, P. Keller, W. Paciorek, P. Roversi, A. Sharff, O. S. Smart, C. Vonrhein, and T. O. Womack, "BUSTER." Global Phasing Ltd., Cambridge, United Kingdom, 2011.

[240] D. E. Tronrud, D. S. Berkholz, and P. A. Karplus, "Using a conformation-dependent stereochemical library improves crystallographic refinement of proteins.," *Acta Crystallogr., Sect. D*, vol. 66, no. Pt 7, pp. 834–42, Jul. 2010.

[241] A. T. Brünger, M. Karplus, and G. A. Petsko, "Crystallographic refinement by simulated annealing: application to crambin," *Acta Crystallogr., Sect. A*, vol. 45, no. 1, pp. 50–61, Jan. 1989.

[242] D. E. Tronrud, L. F. Ten Eyck, and B. W. Matthews, "An efficient general-purpose least-squares refinement program for macromolecular structures," *Acta Crystallogr., Sect. A*, vol. 43, no. 4, pp. 489–501, Jul. 1987.

[243] W. A. Hendrickson and J. H. Konnert, "Incorporation of stereochemical information into crystallographic refinement," in *Computing in Crystallography*, R. Diamond, S. Ramaseshan, and K. Venkatesan, Eds. Bangalore: Indian Academy of Sciences, 1980, pp. 13.01–13.26.

[244] W. F. van Gunsteren, J. Dolenc, and A. E. Mark, "Molecular simulation as an aid to experimentalists.," *Curr. Opin. Struct. Biol.*, vol. 18, no. 2, pp. 149–53, Apr. 2008.

[245] G. R. Bowman, V. A. Voelz, and V. S. Pande, "Atomistic folding simulations of the five-helix bundle protein λ(6−85).," *J. Am. Chem. Soc.*, vol. 133, no. 4, pp. 664–7, Feb. 2011.

[246] C. I. Bayly, P. Cieplak, W. Cornell, and P. A. Kollman, "A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model," *J. Phys. Chem.*, vol. 97, no. 40, pp. 10269–10280, Oct. 1993.

[247] F.-Y. Dupradeau, A. Pigache, T. Zaffran, C. Savineau, R. Lelong, N. Grivel, D. Lelong, W. Rosanski, and P. Cieplak, "The R.E.D. tools: advances in RESP and ESP charge derivation

and force field library building.," *Phys. Chem. Chem. Phys.*, vol. 12, no. 28, pp. 7821–39, Jul. 2010.

[248] M. J. Schnieders, T. D. Fenn, V. S. Pande, and A. T. Brünger, "Polarizable atomic multipole X-ray refinement: application to peptide crystals.," *Acta Crystallogr., Sect. D*, vol. 65, pp. 952–65, Sep. 2009.

[249] H. van den Bedem and J. S. Fraser, "Integrative, dynamic structural biology at atomic resolution—it's about time," *Nat. Methods*, vol. 12, no. 4, pp. 307–318, Mar. 2015.