UNIVERSITY OF CALIFORNIA,
IRVINE

INSIGHTS INTO THE MODELLING OF CRYSTALLOGRAPHIC STRUCTURE AND DISORDER FROM MOLECULAR DYNAMICS SIMULATIONS OF PROTEIN CRYSTALS

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Pharmaceutical Sciences

by

David Charles Wych

Dissertation Committee:
David L. Mobley, Chair
Andrej Luptak
Thomas Poulos

2021

DEDICATION

for my mom
who is still with me
in everything

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

I would like to thank the American Crystallographic Association for allowing me to present at the Annual meeting in 2019, and their Structural Dynamics journal for publishing our work: "Liquid-like and rigid-body motions in molecular dynamics simulations of a crystalline protein"[110].

I would also like to thank the UC Lab FEE's CS3 project for providing funding for my work and introducing me to many brilliant researchers and collaborators.

I would like to thank David L. Mobley for his guidance, kindness, and patience — I could not have asked for a better advisor. He has gone out of his way to make my time in graduate school fulfilling and rewarding, and provided me with opportunities I could only have dreamed of.

I would like to thank Michael E. Wall for his mentorship, time, and willingness to put up with an endless barrage of questions and ideas. I am more than grateful to have had the opportunity to work with him and learn from him — it has been a true gift and a pleasure.

Finally, I'd like to thank Emma Jo Serbiak, Moe Abdul-Rahim, Julian Mackie, my sister Elayna Luksis, my father Guy Wych, Rachel Alvelais, Jan Whitson, Isabel Smith, Charlie Montgomery, and the rest of my friends and family for their love and support.

VITA

David Charles Wych

EDUCATION

Doctor of Philosophy in Pharmaceutical Sciences **2021**
University of California, Irvine *Irvine, CA*

Bachelor of Arts in Physics **2015**
Claremont McKenna College *Claremont, CA*

RESEARCH EXPERIENCE

Graduate Research Assistant **2016–2021**
University of California, Irvine *Irvine, California*

CNLS Virtual Summer Intern **June-August 2020**
University of California, Irvine *Irvine, California*

Graduate Student Assistant **January-August 2019**
Los Alamos National Laboratory *Los Alamos, New Mexico*

Applied Machine Learning Summer Fellow **June-August 2018**
Los Alamos National Laboratory *Los Alamos, New Mexico*

TEACHING EXPERIENCE

Instructor – Physical Biochemistry (PHRMSCI 171) **Fall 2020**
University of California, Irvine

Co-Instructor – Physical Biochemistry (PHRMSCI 171) **Fall 2019**
University of California, Irvine

Teaching Assistant – Physical Biochemistry (PHRMSCI 171) **Fall 2018**
University of California, Irvine

Teaching Assistant – General Chemistry Lab (CHEM 1LC) **Fall 2017**
University of California, Irvine

REFEREED JOURNAL PUBLICATIONS

Liquid-like and Rigid Body Motions in Molecular Dynamics Simulations of a Crystalline Protein **2019**
Structural Dynamics

REFEREED CONFERENCE PUBLICATIONS

Molecular Dynamics Simulations of Protein X-ray Crystallographic Diffuse Scattering **July 2019**
American Crystallographic Association Annual Meeting 2019

SOFTWARE

RigidBodyMotions.py `http://www.github.com/mewall/lunus/`
Code for producing translated and rotated copies of the protein, for analysis published in [111]. Available in the `lunus` *repository in the* `/scripts/command_line/` *directory.*

pdbio.py `http://www.github.com/davidwych/pdbio/`
Minimalist code for reading and writing PDB codes line-by-line, and storing and manipulating the crystal symmetry and atom information.

ABSTRACT OF THE DISSERTATION

Insights into the Modelling of Crystallographic Structure and Disorder From Molecular Dynamics Simulations of Protein Crystals

By

David Charles Wych

Doctor of Philosophy in Pharmaceutical Sciences

University of California, Irvine, 2021

David L. Mobley, Chair

For decades, X-ray crystallography has been the primary tool used to measure the structure of macromolecules at atomic resolution. As fruitful as this method has been, standard X-ray crystallographic methods often provide a model of only the average structure of macromolecules, ignoring structural heterogeneity and correlated disorder. New methods incorporating heterogeneity and correlated disorder are currently in development, but none have yet been able to overcome the "R-factor gap": the consistent and significant discrepancy between the accuracy of macromolecular crystallographic structure factor prediction and the higher accuracy achieved in small-molecule crystallography. It has been suggested that closing this gap will require more accurate models of ordered solvent and disorder. The work presented here uses molecular dynamics (MD) simulations of protein crystals to investigate models of correlated disorder put forward to predict the "diffuse scattering" (non-Bragg X-ray diffraction) which results from correlated disorder in protein crystals. We also present insights in to the modeling of conformational ensembles, ordered solvent, protonation states, and side-chain disorder gathered from crystalline MD simulations.

Chapter 1

Background on Structural and Dynamical Models of Crystalline Proteins

Proteins are the workforce of the cellular world. Virtually every task a living things is able to accomplish on the molecular and cellular level is carried out by proteins. Proteins are composed of a chain of amino acids with different structural and electro-chemical properties. The unique order in which the amino acids are chained together determines the structure that the protein will fold in to and the dynamics of the folded structure, allowing for the careful positioning of amino acids so that their structural and electro-chemical properties can be leveraged to perform specific tasks. Because of the central role that proteins play in biology — and the relationship between a protein's structure and its function — one of the most fundamental tasks in structural biology and biochemistry is the precise measurement of protein structure. While there are many methods used to measure protein structure, the most widely used method is X-ray crystallography.

X-ray crystallography takes advantage of a quirk in the diffraction of light of a particular wavelength through protein crystals: if the wavelength of the light and the spacing and orientation between crystal lattice planes is correctly tuned, the peaks in the resulting diffraction image will encode information about the arrangement of atoms within the lattice plane. The intensity of these peaks are decoded to produce a static picture of the average structure of the proteins which constitute the crystal. However, proteins are not static: they explore a complex free energy landscape, and the dynamics resulting from their traversal of this landscape can be important to their function. New techniques are being developed which aim to analyze the parts of the diffraction images *between* the peaks, which encode information about the correlated motions of the atoms [95].

To understand both the current state of structural and dynamical modelling in X-ray crystallography, and the challenges which remain, a basic understanding of the theory of X-ray scattering and crystallographic modelling will be required. An introduction to both is sketched out below, following the conventions of Meisburger and Ando (2017) [69].

1.1 Basic X-ray Scattering Theory

X-rays scattered elastically off of the electrons bound to atoms produce spherical waves of the same wavelength, radiating outward. For an elastically scattered X-ray plane wave with a particular wavelength λ, the difference in direction between the wavevector of an incident ($\mathbf{k}; |\mathbf{k}| = 2\pi/\lambda$) and scattered ($\mathbf{k}'$) wave is given by $\mathbf{q} = \mathbf{k}' - \mathbf{k}$ where $\mathbf{q}$ is called the "scattering vector", and its magnitude is given by $|\mathbf{q}| = 4\pi \sin(\theta)/\lambda$, where 2θ is the angle between the incident and scattered wavevectors.

For multiple atoms, the photons scattered elastically off of the electrons at different positions — separated by a vector $\mathbf{r}$ — will differ by a phase shift of $\Delta\phi = (\mathbf{k}' - \mathbf{k}) \cdot \mathbf{r} = \mathbf{q} \cdot \mathbf{r}$. This is

simply the difference in path length l (Figure 1.1) for the two scattered waves, converted to radians by a factor of $(2\pi/\lambda)$:

$$\Delta\phi = \mathbf{q}\cdot\mathbf{r} = |\mathbf{q}||\mathbf{r}|\cos(\theta) = \frac{4\pi\sin(\theta)}{\lambda}|\mathbf{r}|\cos(\theta)$$

$$= \frac{2\pi}{\lambda}|\mathbf{r}|2\sin(\theta)\cos(\theta) = \frac{2\pi}{\lambda}|\mathbf{r}|\sin(2\theta) = \frac{2\pi}{\lambda}l.$$

The effect of this phase shift can be expressed by way of a complex exponential $e^{i\Delta\phi} = e^{i\mathbf{q}\cdot\mathbf{r}}$, called the "phase factor", so the phase difference corresponds to turning a unit vector in the complex plane by the angle $\Delta\phi$. If $\mathbf{q}\cdot\mathbf{r}$ is such that the path length is some multiple of the wavelength, in the far-field diffraction limit the waves will constructively interfere (Figure 1.1) corresponding to a phase factor of 1 ($l = n\lambda \rightarrow e^{i\mathbf{q}\cdot\mathbf{r}} = e^{i2\pi n} = 1$).

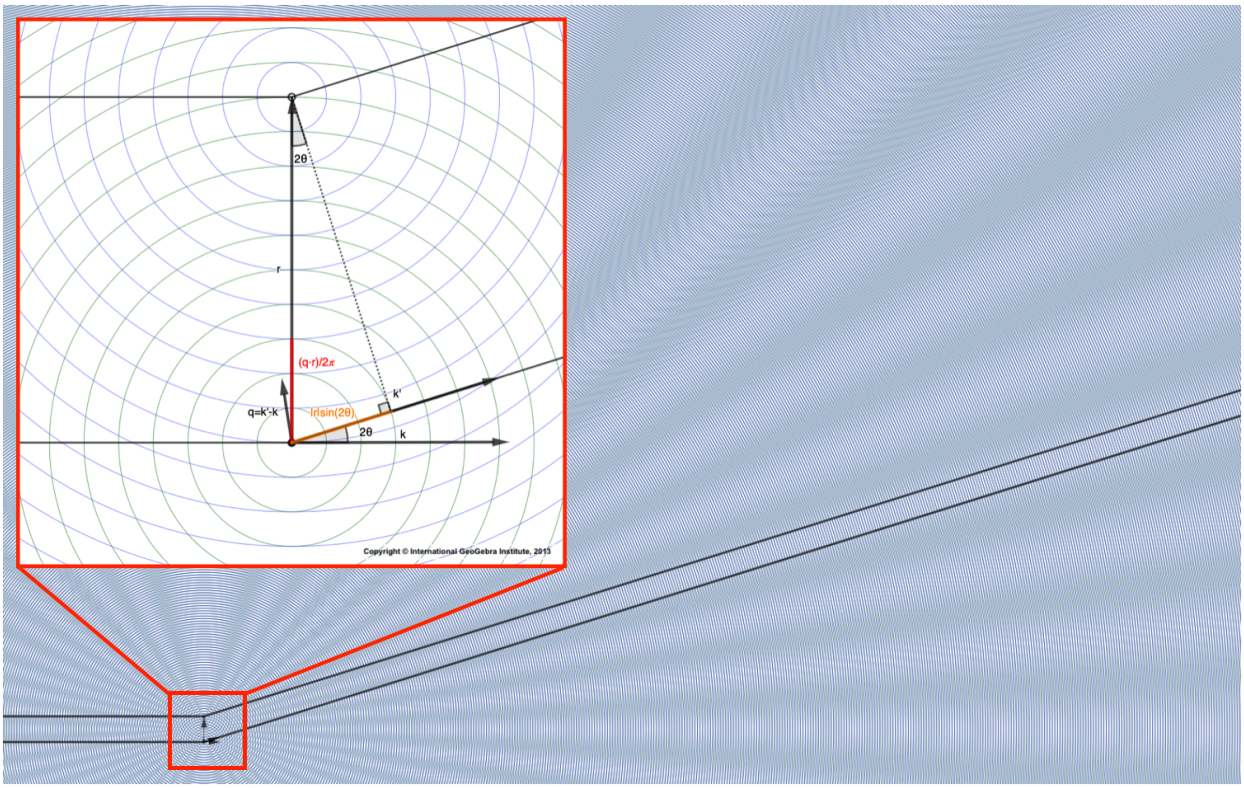
Electrons bound to atoms scatter differently from isolated electrons, depending on each atom's distribution of electron density. This difference is represented with the "atomic scattering factor," $f_n(\mathbf{q})$ for each atom n. We can calculate the scattering amplitude from an atomic electron by integrating the complex exponential across the whole atomic density, weighting by density at each point in space: $f_n(\mathbf{q}) = \int_{\text{atom}} \rho(\mathbf{q}) e^{i\mathbf{q}\cdot\mathbf{r}} d^3\mathbf{r}$.

In a system of many atoms, at positions $\mathbf{r}_n$, the scattered wave has an amplitude equal to the sum of the contributions from each atom: $F(\mathbf{q}) = \sum_n f_n(\mathbf{q}) e^{i\mathbf{q}\cdot\mathbf{r}_n}$, referred to as the "Molecular Form Factor" or "Structure Factor". We can extend the form factor to the entire electron density by switching from summation to integration over the entire scattering volume:

$$F(\mathbf{q}) = \int \rho(\mathbf{r})\ e^{i\mathbf{q}\cdot\mathbf{r}} \mathrm{d}^3\mathbf{r}.$$

The form factor is the Fourier transform of the density.

Figure 1.1: Inset: Difference in path length ($|\mathbf{r}|\sin(2\theta)$) is equal to the dot product of the scattering vector ($\mathbf{q} = \mathbf{k}' - \mathbf{k}$) with the separation vector ($\mathbf{r}$), multiplied by a conversion factor ($\lambda/2\pi$). The far-field diffraction from two scattering atoms is constructive in the far field (light blue) when the difference in path length is equal to an integer multiple of the wavelength, and destructive when it is not (dark blue).



The form factor is therefore a complex-valued function of the scattering vector. However, the intensity measured at the detector is real-valued and is proportional to the squared modulus of the form factor: $I(\mathbf{q}) \propto |F(\mathbf{q})|^2$. The squared modulus of the form factor is equal to the Fourier transform of the auto-correlation in the electron density:

$$|F(\mathbf{q})|^2 = \int (\rho \star \rho)(\mathbf{r}) e^{i\mathbf{q}\cdot\mathbf{r}} d\mathbf{r}.$$

This auto-correlation (the correlation of the density with itself at a displacement $\mathbf{r}$) is also known as the "Patterson Function" of the density, and it can be calculated directly from experiment (up to a constant of proportionality) by feeding the measured intensity through the inverse Fourier transform.

However, because the intensity measured at the detector is proportional to the squared modulus of the form factor, but not the form factor itself, we can get information from experiment about the *amplitude* of the form factor at each scattering vector, but all information about the *phases* of the form factor at each scattering vector is lost. The recovery of the phases, knowing only the amplitudes, is known as the "phase problem". The phase problem can be solved by a few different means: indirectly, by estimation with molecular replacement, or directly, in more special cases, using e.g. isomorphous replacement[31] or anomalous dispersion[34]. Once the phases are known or estimated, the form factor can be fed through the inverse Fourier transform to recover the corresponding electron density.

1.1.1 X-ray Scattering in Crystalline Systems

In a crystal, where the density is periodic in space, we can express the full density for a crystal in terms of the density with respect to the origin of each of "unit cells" (the repeating structural units which make up the crystal). That is, the position of each atom $\mathbf{r}_i$ can be re-expressed in the form $\mathbf{r}_i = \mathbf{R_n} + \mathbf{r}_j$ where $\mathbf{R_n}$ is the position of the origin of unit cell $\mathbf{n}$ and $\mathbf{r}_j$ is the position of the jth atom with respect to the unit cell origin. Because the unit cells are arranged periodically, we can express the position of each unit cell origin as $\mathbf{R_n} = n_1\mathbf{a}_1 + n_2\mathbf{a}_2 + n_3\mathbf{a}_3$, where $\mathbf{a}_1, \mathbf{a}_1$ and $\mathbf{a}_3$ are the vectors describing the direction and length of each unit cell axis, and $\mathbf{n} = [n_1, n_2, n_3]$ is a vector of integers which index the unit cells. Substituting this new formulation in to the expressions above, we can see that the form factor for the crystal is given by $F_{\text{cryst.}}(\mathbf{q}) = \sum_{\mathbf{n}} F_{\mathbf{n}}(\mathbf{q}) e^{i\mathbf{q}\cdot\mathbf{R_n}}$, where $F_{\mathbf{n}}$ is the form factor for unit cell $\mathbf{n}$, relative to its origin.

The intensity measured at the detector is now expressible in terms of the sum over the many unit cells: $I(\mathbf{q}) \propto |F_{\text{crystal}}|^2 = \sum_{\mathbf{n}} \sum_{\mathbf{m}} F_{\mathbf{n}}(\mathbf{q}) F^*_{\mathbf{m}}(\mathbf{q}) e^{i\mathbf{q}\cdot(\mathbf{R_n}-\mathbf{R_m})}$. By defining the average form factor for the N total unit cells $F_{\text{avg.}}(\mathbf{q}) = \frac{1}{N}\sum_{\mathbf{n}} F_{\mathbf{n}}(\mathbf{q}) = \int \rho_{\text{u.c.avg.}}(\mathbf{r}) e^{i\mathbf{q}\cdot\mathbf{r}} \mathrm{d}^3\mathbf{r}$, we can

separate the intensity in to two terms:

$$I(\mathbf{q}) = |F_{\text{crystal}}(\mathbf{q})|^2 = I_{\text{Bragg}}(\mathbf{q}) + I_D(\mathbf{q}),$$

where

$$I_{\text{Bragg}}(\mathbf{q}) = |F_{\text{avg.}}(\mathbf{q})|^2 \left| \sum_{\mathbf{n}} e^{i\mathbf{q}\cdot\mathbf{R_n}} \right|^2,$$

and

$$I_D(\mathbf{q}) = \sum_{\mathbf{n}} \sum_{\mathbf{m}} (F_{\mathbf{n}}(\mathbf{q}) F^*_{\mathbf{m}}(\mathbf{q}) - |F_{\text{avg.}}(\mathbf{q})|^2) e^{i\mathbf{q}\cdot(\mathbf{R_n}-\mathbf{R_m})}.$$

I_{Bragg} is the "Bragg" intensity which results from the truly periodic average electron density, and I_D is the "diffuse" intensity, which results from the parts of the density which deviate from the average, sometimes called the "variational" or "continuous" scattering.

Because vectors $\mathbf{R_n}$ form a lattice, the Bragg phase factors $e^{i\mathbf{q}\cdot\mathbf{R_n}}$ will have terms equal to 1, corresponding to in-phase scattering, when the dot product $\mathbf{q}\cdot\mathbf{R_n} = 2\pi n$ with $n \in \mathbb{Z}$. This corresponds to the scattering from *an entire lattice plane*: because the wavelength λ of the X-ray beam is well-defined, the position of the photons is uncertain, so we no longer consider scattering off of individual atoms, but rather scattering off of all atoms which intersect a particular lattice plane normal to $\mathbf{R_n}$.

We can construct a new lattice, called the "reciprocal lattice", with lattice points $\mathbf{q}$ satisfying the in-phase scattering condition above. This new lattice in reciprocal space has the principal axes $\mathbf{a}_1^*, \mathbf{a}_2^*$ and $\mathbf{a}_3^*$, where $\mathbf{a}_1^* = (2\pi/V)\ \mathbf{a}_2^* \times \mathbf{a}_3^*$, $\mathbf{a}_2^* = (2\pi/V)\ \mathbf{a}_3^* \times \mathbf{a}_1^*$, and $\mathbf{a}_3^* = (2\pi/V)\ \mathbf{a}_1^* \times \mathbf{a}_2^*$, and $V = \mathbf{a}_1 \cdot (\mathbf{a}_2 \times \mathbf{a}_3)$ is the volume of the unit cell. Each reciprocal lattice vector is in a direction normal to two of the lattice vectors and has a length of 2π divided by the spacing between the lattice planes. The nodes of reciprocal lattice are at $\mathbf{G_h} = h_1\mathbf{a}_1^* + h_2\mathbf{a}_2^* + h_3\mathbf{a}_3^*$,

and $\mathbf{h} = [h_1, h_2, h_3]$ are called the "Miller indices".

When the crystal is positioned such that $\mathbf{q}$ overlaps with a node of the reciprocal lattice ($\mathbf{q} = \mathbf{G_h}$ for some choice of h_1, h_2 and h_3), there is constructive interference:

$$\mathbf{q} \cdot \mathbf{R_n} = \mathbf{G_h} \cdot \mathbf{R_n} = 2\pi(h_1 n_1 + h_2 n_2 + h_3 n_3).$$

This is a restatement of *Bragg's Law*, which says that constructive interference occurs when waves scatter off of lattice planes separated by a distance d with an angle θ such that $2d\sin(\theta) = n\lambda$, because $\mathbf{G_h} \cdot \mathbf{R_n}$ is the path length difference converted to an angle, similar to the case sketched out above for scattering off of individual atoms. In such cases where where the scattering vector hits a reciprocal lattice node, the exponential factors will be 1, and the value of $I_{\text{Bragg}}(\mathbf{q})$ at that scattering vector $\mathbf{q}$ will be proportional to the average unit cell structure factor $|F_{\text{avg.}}(\mathbf{q})|^2$.

The Ewald Sphere Construction

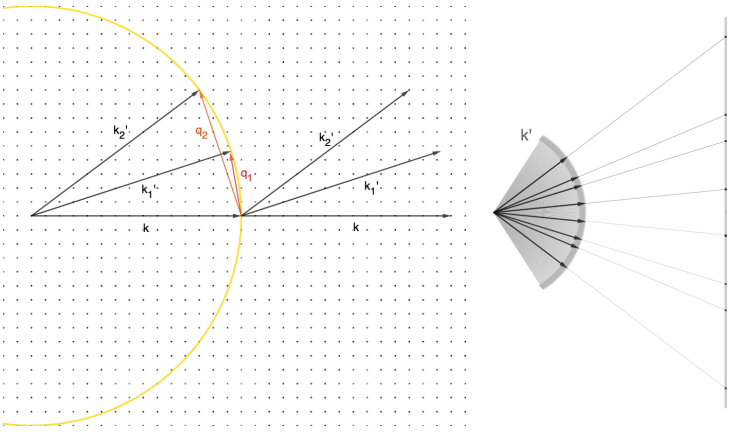


Figure 1.2: The reciprocal lattice as black dots (left) with incident wavevector $\mathbf{k}$ and scattered wavevectors $\mathbf{k}_1'$ and $\mathbf{k}_2'$ in black, scattering vector $\mathbf{q}_1$ and $\mathbf{q}_2$ in red and orange, and the Ewald sphere in yellow (left). For any scattered wavevector ($\mathbf{k}'$) whose associated scattering vectors ($\mathbf{q}$) coincide with a reciprocal lattice node — like $\mathbf{k}_2'$ in the left image — interference is constructive and a spot appears on the detector in the far field (right)

We can think of crystallographic experiments as a means of exploring reciprocal space. We direct an X-ray beam at a crystal, and record the intensity of reflections at various angles using a detector. At certain angles, θ, the scattering vector $\mathbf{q} = \mathbf{k}' - \mathbf{k}$ will overlap with a reciprocal lattice node (Figure 1.2), creating a peak in the diffraction image – these peaks

are known as "Bragg peaks" due to their connection with Bragg's Law.

The surface in reciprocal space swept out by all scattering angles θ for a given incident wave is known as the "Ewald sphere". Because the magnitude of the incident and reflected wavevectors **k** and **k'** is $2\pi/\lambda$, when we decrease the wavelength of the incident X-ray beam, we increase the magnitude of the wavevector, and the area of reciprocal space we can explore. By rotating the crystal, we change the position of the Ewald sphere with respect to reciprocal space, allowing for a new set of Bragg peaks to be resolved at various scattering angles (Figure 1.3). By recording diffraction images at many different incident angles, we can build up a full three-dimensional picture of the reciprocal lattice, which gives us information about the average form factor for the unit cell at each scattering vector.

Because the form factor is the Fourier transform of the unit cell electron density, Bragg peaks at Miller indices further from the origin correspond to "higher resolution" features in the electron density: a *longer* scattering vector along each axis in reciprocal space corresponds to a *shorter* separation distance between lattice planes from which we detect constructive interference when the scattering vector cross a reciprocal lattice node. If the phases were known, adding together three-dimensional sinusoids with frequencies determined by the scattering vector magnitudes and amplitudes given by the magnitude of the form factor at each scattering vector with the appropriate phases would reproduce the average electron density in the unit cell.

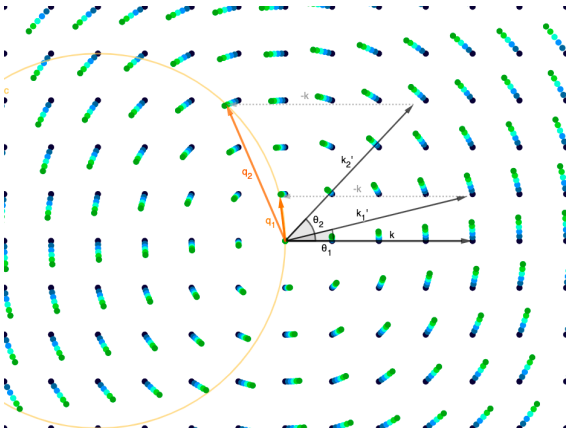


Figure 1.3: By rotating the crystal, we expose different reciprocal lattice nodes to the Ewald sphere. This happens at different rotation angles (different colored dots) and different scattering vectors $\mathbf{q} = \mathbf{k}' - \mathbf{k}$ (different scattered wavectors $\mathbf{k}_1'$ or $\mathbf{k}_2'$).

The Debye-Waller Factor and B-factors

Under the approximation that atoms exhibit small-scale harmonic disorder about their average positions, the density gets smeared out about the atom's average position, and there is a new factor introduced to the form factor, called the "Debye-Waller" factor, $T_j(\mathbf{q}) = e^{-1/2\mathbf{q}^T\mathbf{U}_j\mathbf{q}}$, where $\mathbf{u}_j = \mathbf{r}_j - \langle\mathbf{r}_j\rangle$ is the displacement of the jth atom in the unit cell from its average position, and $\mathbf{U}_j = \langle\mathbf{u}_j\mathbf{u}_j^T\rangle$ is the variance-covariance matrix of atomic displacements. This effect was originally thought to be so significant that Bragg diffraction would not be possible: Peter Debye supposedly warned Max von Laue that thermal fluctuations would preclude the possibility for constructive interference in crystal diffraction [27], but von Laue went forward with his experiments anyway. As it happens, these fluctuations simply augment the intensity of the Bragg peaks: the Debye-Waller factor introduces an additional multiplicative factor to the expression for the intensity from Bragg scattering, reducing the intensity as the displacements increase: $I_{\text{Bragg}}(\mathbf{q}) = |\sum_j f_j(q)T_j(\mathbf{q})e^{i\mathbf{q}\cdot\langle\mathbf{r}_j\rangle}|^2|\sum_{\mathbf{n}} e^{i\mathbf{q}\cdot\mathbf{R}_\mathbf{n}}|^2$, where $f_j(q)$ is the atomic form factor. Hence, the Bragg scattering contains information not only about the average position of the atoms in real space, but also information about the degree of disorder about each atom's average position.

Refinement of the Debye-Waller factor, alongside refinement of the structure factors, can thus give us an estimate of the degree of small-scale harmonic disorder for each atom, which is commonly published as a "B-factor" or "temperature factor" equal to $\frac{8}{3}\pi^2\langle u_j^2\rangle$ in the case where disorder is assumed to be isotropic, or as anisotropic displacement parameters (or "ADPs") which are the 6 unique elements of the matrix $\mathbf{U}_j$. The reduction in intensity at the Bragg peaks caused by harmonic disorder is more pronounced at larger scattering vectors, making the refinement of the Debye-Waller factors more difficult for low-resolution data.

1.2 Standard Crystallographic Methods

We can record diffraction images at various incident angles by simply rotating the crystal (using a goniometer), but the persistent bombardment of the crystal by X-ray radiation causes damage, so often measurements will be made at various locations along the crystal, continuing until the damage is severe enough to inhibit clean diffraction. Measurements at different incident angles can also be made with smaller crystals using the technique of serial crystallography, in which reflections are collected from many smaller crystals at random orientations. However, because each crystal is randomly oriented, this technique requires more data and computational resources to align and differentiate (or "merge") reflections taken at random and unknown incident angles.

By recording the intensities at these reciprocal lattice nodes, we can obtain information about the average unit cell form factor $|F_{\text{avg.}}(\mathbf{q})|^2$ at each scattering vector. The intensities at each Bragg peak are different, depending on the structure and small-scale harmonic disorder of the atoms in the unit cell, and the various planes of symmetry passing through the atoms in the unit cell. The intensity of each unique Bragg peak across all the different diffraction images is indexed and measured relative to background noise in a process known as "integration". Once again, because these intensities are proportional to the squared *amplitude* of the form factor, we lose all information about the *phases*, and we must estimate them to predict the true average unit cell electron density. There are a variety of techniques to do this, the most common of which is Molecular Replacement, but ab initio methods, isomorphous replacement[31], and anomalous scattering[34] can also be used to recover the phases.

For Molecular Replacement, crystallographers use the known structure of proteins which are similar to the protein under study ("homologous structures"), along with the crystal symmetry operations, to calculate the amplitude and phase of the structure factors for the unit cell, assuming the molecular model is correct. The structure factors from the model

are compared to the magnitudes of the structure factors measured in experiment, with the reciprocal space axes aligned by rotating and translating the Patterson function of the model until it roughly matches the experimental Patterson. By iteratively refining the modelled structure, calculating model structure factors (F_c), and comparing with observed structure factors from experiment (F_o), crystallographers can arrive at a best-guess structure for the protein. These structure factors are often deposited along with the structure files as "composite" maps ($2F_o - F_c$), which correspond roughly to the model density and have the mean subtracted so that they are centered on zero, and "difference" maps ($F_o - F_c$), which correspond to differences in density between model and experiment. Positive difference density generally corresponds to areas of the model density with too many electrons (compared to experiment) while negative difference density corresponds to areas lacking electrons, though negative difference density can be harder to interpret. Densities are often deposited in non-absolute units, so the standard is to display composite maps at an iso-surface of 1σ (one standard deviation above the mean) and the difference density at 3σ.

The agreement between the calculated and observed structure factors is captured in a statistic known as the "reliability factor" or "R-Factor", R, where

$$R = \frac{\sum_{h_1 h_2 h_3} ||F_o| - |F_c||}{\sum_{h_1 h_2 h_3} |F_c|}.$$

If the magnitude of calculated and observed structure factors matched exactly, the R-factor would be zero. As the calculated structure factors deviate from the observed structure factors, the R-factor increases. To protect against over-fitting to the data, a random selection of structure factors (often, 10% of the total number of structure factors) are left out in the calculation of the R-factor during refinement for cross-validation, with the R-factor for the remaining (or "working") structure factors referred to as the "R_{work}". After refinement of the structural model, an R-factor is calculated using the structure factors left out, known as the "R_{free}." If R_{work} and R_{free} are similar, then the model is considered valid, however,

if R_{free} is much larger than R_{work}, then the model is considered to have been overfit to the data.

The "R-factor Gap"

Despite improvements in both experimental and refinement techniques, R_{work} and R_{free} are rarely lower than 0.15 in macromolecular crystallography. Small-molecule crystallographers routinely obtain R-factors much lower than this (0.02-0.05) with discrepancies between calculated and observed structure factors largely attributable to measurement error and internal consistency of the data. James Holton and collaborators were able to show that the discrepancy between small-molecule and macromolecular crystallography is not attributable to larger error in experimental apparatuses or deficiencies in measurement — the problem lies in the models used to predict structure factors: they simulated many different forms of experimental error, and error in data collection and processing, and found that standard crystallographic methods *should* converge to small molecule precision, but do not, thanks to deficiencies in the modelling of protein and solvent density. It has since been suggested that improvements can be made by modelling the protein as structurally heterogeneous, and by more accurately modelling the solvent (beyond bulk solvent corrections and placement of full-occupancy waters into solvent peaks). [40]

1.3 Modern Methods

Various improvements on the standard crystallographic refinement procedure have been introduced in the preceding decades. A common thread between them has been the rejection of the model of a protein as a static object. Instead, these new and improved models view proteins as dynamic objects, which sample from an ensemble of structures, whether on the

scale of the entire protein, domains, smaller groups of amino acids, or individual side chains. Cutting edge methods are being developed, which expand refinement beyond the Bragg peaks, modelling correlated motions which give rise to the diffuse scattering; these methods will be discussed in the next section.

1.3.1 Anisotropic Displacement Parameters

As mentioned above, under the assumption of small-scale harmonic disorder, the Debye-Waller factor can be refined alongside the structure factors during refinement. The refinement of anisotropic displacement parameters (ADPs) requires estimating the full variance-covariance matrix of atomic displacements for each atom, which has six unique elements. If the data is low resolution, the data-to-parameter ratio can be too small, leaving the parameters susceptible to overfitting. However, crystallographers will often use techniques such as as TLS (Translation, Libration, Screw) refinement or Normal Modes or Elastic Network refinement to overcome this problem.

In TLS refinement, segments of the protein are collected in to groups, as small as a few amino acids or as large as entire domains, and the entire group is treated as a rigid body which can undergo uniform translation, screw-like rotation around an axis, and libration, which is a mixture of the two [84]. The B-factors or ADPs for each of the atoms can then be inferred relative to the rigid body motions of the group [53], lowering the number of independent parameters to refine. If successful, this technique can also provide insight into the correlated motions of large groups of atoms in the protein.

In the normal modes models, the structure of the protein is described by a set of generalized coordinates (C-α coordinates, dihedral angles, or any other set of variables which cover the essential degrees of freedom in the system) which are subject to a potential energy function that is approximately quadratic about the minimum [56]. The dynamics of the system

are then approximated by inserting the generalized coordinates and the quadratic potential into the Lagrangian and solving for the modes of the system at various frequencies. Often, the lowest frequency modes are used, and higher frequency modes are ignored, though the implied dynamics can change quite dramatically depending on which modes are included or ignored.

Elastic network (EN) models and, in particular, Gaussian network (GN) models, describe the protein as a series of atoms connected by springs, excited by the injection of thermal energy from the environment. The number of springs is determined by a cutoff, with a larger cutoff allowing for the modelling of more long-range interactions. The spring constant(s) is the parameter(s) of the model, tuned to reproduce experimental data. The modes of the system can be solved for through eigenvalue decomposition of the Kirchoff matrix for the system [81]. The EN and GN models are similar, with the main difference being that in the GN model, both changes in the relative *distance* between the nodes of the network and changes in the *direction* of the vector between the nodes are penalized by the potential, whereas in EN, only the distance is constrained.

Both models, and their relatives, have been used to great success in modelling the isotropic B-factors[30, 56, 20] and ADPs[50, 51, 80] in crystallographic refinement. In addition to modelling harmonic disorder of the individual atoms, these models can also be used to infer collective motions of the proteins' secondary structure elements and domains, as well as lattice disorder due to crystal contacts, which give rise to diffuse scattering. The application of these models to the prediction of diffuse scattering is described below.

1.3.2 Ensemble and Multi-conformer Models

It has been known for many decades now that there is heterogeneity in the structures adopted by the various proteins in crystals [88]. In particular, some side chains can be expected to

sample from a wide range of conformations, leading to broader areas of density associated with the side chains which are impossible to model with a single structure undergoing harmonic disorder. With the advent of room-temperature X-ray crystallography, the inclusion of this heterogeneity has become more necessary, as the switch from cryo temperatures to room temperature broadens the distribution of conformations from which the side chains, and the proteins as a whole, sample.

Recent work by James Fraser and Henry van den Bedem and colleagues have shown that more than 35% of side chains experience significant changes to their conformational distributions when studied at room temperature versus cryo temperatures [26]. These conformational changes can be triggered both by perturbations in temperature and mutation, and can alter contact networks with other side chains, as well as bound ligands, with broad implications for the study of allostery and drug design [96].

The modelling of structural heterogeneity has been be carried out at the level of peptide flips in the backbone, side chain reorientations, and changes in ligand binding mode, as in the qFit and qFitLigand model building programs of Fraser, van den Bedem and colleagues [49, 97]. modelling of structural homogeneity can also be carried out on the level of the entire protein, with the all-atom ensemble generated by molecular dynamics sampling of a potential landscape, as in `phenix.ensemble_refinement`. These methods have been shown to fit X-ray data better than single-structure refinements do [55, 7].

1.4 Diffuse Scattering and Models of Correlated Disorder

Above, we saw that in addition to the constructive Bragg intensity that results from periodicity in the unit cell electron density, there is another part of the intensity, the "diffuse" or "variational" scattering, which results from correlated density fluctuations about the average across the many unit cells. This part of the intensity is weaker than the Bragg peaks, and largely consists of patterns of diffuse density in the form of haloes around the Bragg peaks, or streaks and "speckles" between them, with more intricate patterns visible for crystals of sufficient quality, measured with detectors of sufficient resolution (Fig 1.4). To model this part of the density, we need to model the dynamics which would give rise to fluctuations in the density which are correlated across many unit cells.

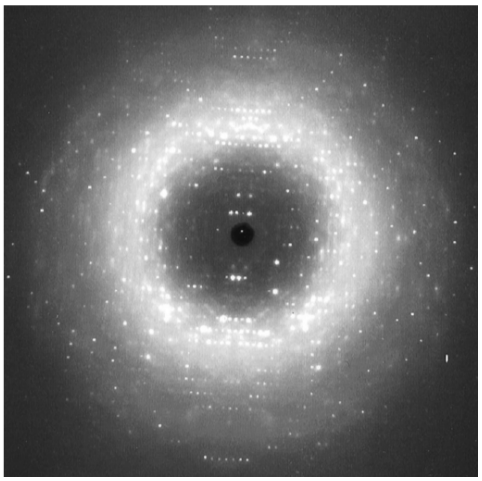


Figure 1.4: Diffraction image from Wall, Ealick, and Gruner (1997) showing point-like Bragg peaks and diffuse scattering. Streaks between Bragg peaks and speckles are visible in the solvent ring, and more cloudy, patterned features are visible at higher resolution. Copyright (1997) National Academy of Sciences.

Under the assumption that there is a distribution for the density in each unit cell, from which we can calculate an ensemble average, the diffuse intensity can be expanded as follows:

$$I_D(\mathbf{q}) = \sum_{\mathbf{n}} \sum_{\mathbf{m}} (F_{\mathbf{n}}(\mathbf{q}) F^*_{\mathbf{m}}(\mathbf{q}) - |F_{\text{avg.}}(\mathbf{q})|^2) e^{i\mathbf{q}\cdot(\mathbf{R_n}-\mathbf{R_m})}$$

$$= \sum_{\mathbf{n}} \sum_{\mathbf{m}} (\langle F_{\mathbf{n}}(\mathbf{q}) F^*_{\mathbf{m}}(\mathbf{q})\rangle - |\langle F_{\text{cell}}(\mathbf{q})\rangle|^2) e^{i\mathbf{q}\cdot(\mathbf{R_n}-\mathbf{R_m})}$$

$$= \langle |F_{\text{crvstal}}(\mathbf{q})|\rangle^2 - |\langle F_{\text{crystal}}(\mathbf{q})\rangle|^2 = \langle |F_{\text{crystal}}(\mathbf{q}) - \langle F_{\text{crystal}}(\mathbf{q})\rangle|^2\rangle$$

The equation above is known as Guinier's Equation [32], and is the basis for the calculation of diffuse scattering for every one of the models outlined below. It is plain to see in the last expression that the diffuse scattering results from deviations of the crystal density from the average – that is, from dynamics.

Under an additional assumption of harmonic deviations for each atom about their average positions, we can express the diffuse scattering as:

$$I_D(\mathbf{q}) = \sum_{\mathbf{n}j} \sum_{\mathbf{m}k} f_j(q) f_k^*(q) e^{i\mathbf{q}\cdot(\mathbf{R_n}-\mathbf{R_m})} e^{i\mathbf{q}\cdot(\langle \mathbf{r}_j\rangle - \langle \mathbf{r}_k\rangle)} T_j(\mathbf{q}) T_k(\mathbf{q}) \{e^{\mathbf{q}^T \mathbf{V}_{\mathbf{n}j\mathbf{m}k}\mathbf{q}} - 1\}$$

where $\mathbf{V}_{\mathbf{n}j\mathbf{m}k} = \langle \mathbf{u}_{\mathbf{n}j} \mathbf{u}_{\mathbf{m}k}^T \rangle$ is the displacement-covariance matrix for each pair of atoms in the crystal (atom j in unit cell $\mathbf{n}$ and atom k in unit cell $\mathbf{m}$). The exponent $\mathbf{q}^T \mathbf{V}_{njmk} \mathbf{q}$ can be thought of as the displacement covariance matrix projected on to the vector $\mathbf{q}$. If the displacement covariance for atoms j and k separated by $\langle \mathbf{r}_j \rangle - \langle \mathbf{r}_k \rangle$ is high along the same the direction as $\mathbf{q}$, the diffuse scattering at $\mathbf{q}$ will be stronger. Under Taylor approximation to first order, this expression can be rewritten as:

$$I_D(\mathbf{q}) = \sum_{\mathbf{n}j} \sum_{\mathbf{m}k} f_j(q) f_k^*(q) e^{i\mathbf{q}\cdot(\mathbf{R_n}-\mathbf{R_m})} e^{i\mathbf{q}\cdot(\langle \mathbf{r}_j\rangle - \langle \mathbf{r}_k\rangle)} T_j(\mathbf{q}) T_k(\mathbf{q}) \{\mathbf{q}^T \mathbf{V}_{\mathbf{n}j\mathbf{m}k}\mathbf{q}\}$$

So, in the same way that Bragg scattering can be leveraged to estimate the average electron density of the unit cell, the diffuse scattering could, in principle, be leveraged to estimate the displacement covariance matrix (and thus the correlated motions) for each pair of atoms in the crystal. However, in contrast to Bragg refinement, the data-to-parameter ratio in the diffuse case leaves the problem of refinement hopeless: the matrix $\mathbf{V}_{\mathbf{n}j\mathbf{m}k}$ has six independent parameters for every pair of atoms in the illuminated volume of the crystal which, even for small crystals and tight beams, is far too many parameters to refine given any reasonable amount of data collected.

Therefore, to move forward with the task of constructing structural models which include information about correlated disorder which can be refined against the diffuse scattering, we must construct simpler models, with fewer parameters to fit.

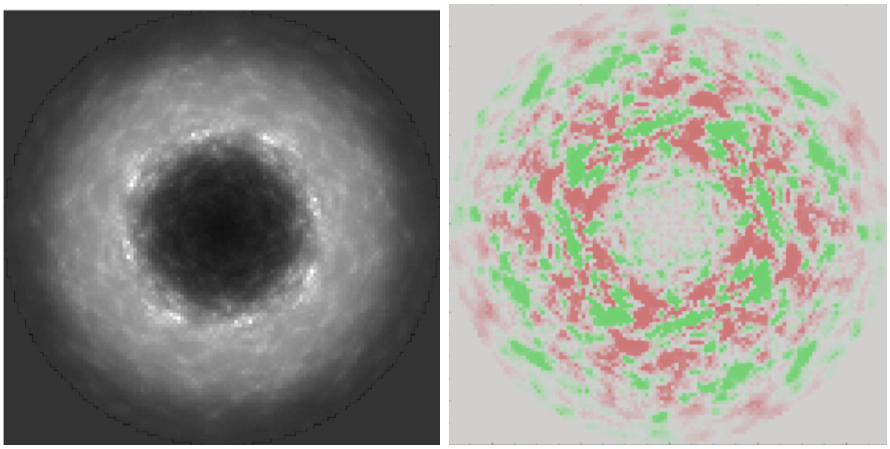


Figure 1.5: Simulated diffuse scattering data (left). The anisotropic component is calculated by subtracting away the average intensity in radial shells (right – positive: green; negative: red)

Before delving in to the specific models proposed to interpret diffuse scattering, it is important to note various general features in the diffuse scattering which will help in understanding them all. The diffuse scattering can be separated in to two main components: the isotropic and anisotropic components (Figure 1.5). The isotropic component is radially symmetric, and is mostly associated with solvent-protein interactions – this component of the diffuse scattering is often called the "solvent ring" [1].

The anisotropic component is all those features which remain once the radially symmetric component is subtracted away. This component can contain streaks, speckles, haloes around Bragg peaks, and cloudy features which span large swaths of reciprocal space. In general, features in the anisotropic component closer to Bragg peaks correspond to correlated motions which span distances of a unit cell or larger – these features are often referred to as Thermal Diffuse Scattering (TDS). Features in the diffuse scattering farther from the Bragg peaks include substantial contributions from more local correlated motions, and have proven much harder to model. These components of the diffuse scattering are what must be predicted if we

[1] Though we now know this component is produced by contributions from both the protein and solvent, in roughly equal proportion, the name "solvent ring" comes from a time when it was believed this component came only from the solvent, and is often still referred to by this name.

wish to produce models of correlated disorder which are relevant to functionally-important motions in proteins.

There are four main models of this kind which have been studied extensively: the Rigid Body Motions model (RBM), the liquid-like Motions (LLM) model, the Elastic Network/Normal Modes model (EN/NM), and the Molecular Dynamics (MD) model. We will discuss each of these models in brief below, with special attention paid to the modelling of diffuse scattering.

1.4.1 The Rigid Body Motions (RBM) Model

The Rigid Body Motions (RBM) model assumes that the protein(s) which comprise the unit cell (or an ensemble of them) undergo rigid translation and rotation about their average center of mass. Though the TLS model (a special case of RBM model, with rigid groups smaller than the protein as a whole) has been used extensively and successfully to predict B-factors and ADPs, the application of the TLS model to the prediction of diffuse scattering has been less encouraging. The application of rigid body motions models more broadly have been somewhat more successful, though the degree to which they explain the source of diffuse scattering is a topic of debate. There is also a distinction to be made between coupled and independent rigid-body motions, which will be important to keep in mind when attempting to differentiate and compare the models described below.

A special case of rigid body motions are those produced by lattice-coupled motions due to the interactions between proteins and their neighbors in the crystal lattice. This type of rigid body motions was the first to be connected explicitly with diffuse scattering, in the form of diffuse "streaks" which appear along crystal lattice planes. In the early days of the modelling of diffuse scattering, analysis was restricted to single two-dimensional diffraction images. Diffraction images collected from crystals of tropomyosin contained streaks which were interpreted as the result of transverse motions along the filament arms of tropomyosin

(like the vibrations of a string) and haloes which were interpreted as the result of coupled motions of neighboring proteins in the lattice [6]. Follow-up studies of tropomyosin, analyzing six diffraction images rather than just one, showed further evidence for long-range correlated motions due to transverse waves propagated along the filament arms [12]. Studies of diffraction from hen-egg-white lysozyme crystals showed similar streaks between Bragg peaks which were connected to rigid body molecular displacements along two perpendicular directions in the crystal lattice, connected with close-contacts between neighboring proteins along these axes (no streaks were found along the other axis, as there were no comparable close contacts along the other direction in the lattice) [22, 4]. These models were validated qualitatively, as quantitative validation was not yet possible.

Further study of the diffuse scattering from hen-egg-white lysozyme by Benoit, Faure, and Perez showed that models of individual, independent rigid-body translational and rotational motions (roughly equal in contribution) fit to the diffuse scattering could be used to calculate the root-mean-square C-α fluctuations in good agreement with the B-factors from refinement, though once again the comparison with the diffuse scattering from experiment was qualitative, and the degree of rotational and translational motion was not refined but guessed to an order of magnitude based on visual agreement with the scattering observed in diffraction images [75]. These finding were, in part, a challenge to the models of Clarage *et al.*, who had proposed a model of homogeneous disorder, the "Liquid-like-motions" (LLM) model: "The high proportion of rigid-body displacements which is deduced... strongly contradicts previous interpretations of tetragonal lysozyme diffuse-scattering data in terms of homogeneous disorder (Clarage *et al.*, 1992). To our understanding, this discrepancy mostly lies in the fact that the type of correlation accounted for by homogeneous disorder models is intrinsically limited, which thus may lead to an erroneous interpretation of the experimental diffuse scattering." [75] This would be the first in a long string of disagreements over the main source of diffuse scattering. However, the argument in favor of rigid-body motion was supported in a paper published shortly after by Smith, Genest, and Héry, which found evi-

dence for rigid body motions in groups of backbone atoms and associated rigid side-chains in a 1ns molecular dynamics simulation of orthorhombic lysozyme [37].

The RBM-LLM debate went cold for a few decades before being revived in the mid 2010s in a series of studies by Wall, van Benschoten and collaborators [94, 95]. Thanks to the developments of Wall *et al.* in the late 1990s[104], it was possible to construct full three-dimensional diffuse scattering maps which could be analyzed more finely and quantitatively. A three-dimensional diffuse scattering map was collected from diffraction by crystals of dimeric GpdQ by van Benschoten *et al.* and the anisotropic component was modelled using a TLS model, refined against the Bragg data. Firstly, refinements using different rigid groups for the TLS model (sub-domains, monomers, or the entire protein dimer), despite predicting different collective motions for the protein, did not produce significant difference in R_{work} and R_{free}, though all TLS refinements produced lower R_{free} values than non-TLS refinement. The diffuse scattering, however, was highly dependent on the TLS groups selected: the different diffuse scattering maps resulting form the different TLS rigid group selections showed little similarity with each other. It was suggested that the different molecular motions implied by the TLS models could, in principle, be (in)validated by comparison with the experimental diffuse scattering [94]. A followup study of the diffraction from cyclophilin A (CypA) and trypsin crystals by van Benschoten *et al.* compared the predictions of various TLS models ("Phenix", "TLSMD", and "whole molecule") with those from the LLM model and coarse-grained normal-modes models. None of the TLS models agreed well with the diffuse scattering data (correlation coefficients between 0.02-0.14) nor did they agree well with each other (CCs between 0.06-0.22), despite, again, yielding satisfactory and similar agreement with the Bragg data (R_{work} and R_{free} between 0.16 and 0.18). These results suggested that the protein motions which produced the diffuse scattering were correlated on distances shorter than would be suggested by the TLS models [95]. It's important to note as well, that neither of the other models tested agreed particularly well with the diffuse scattering data, with correlation coefficients for the full map between 0.4 and 0.5.

These results cast doubt on the rigid body motions model as a viable model for the diffuse scattering. However, two papers from Ayyer *et al.* in 2016 and de Klijn *et al.* in early 2019 may revive it. Ayyer *et al.* were able to show that the speckled patterns visible in X-ray Free Electron Laser (XFEL) diffraction data from crystals of photosystem II were consistent with independent rigid-body translational lattice disorder, and they were able to leverage the information in the diffuse scattering under this assumption to extend the resolution limit of the data beyond the Bragg diffraction limit, from 4.5 to 3.5 Angstroms, and also to phase the diffraction pattern directly [3].

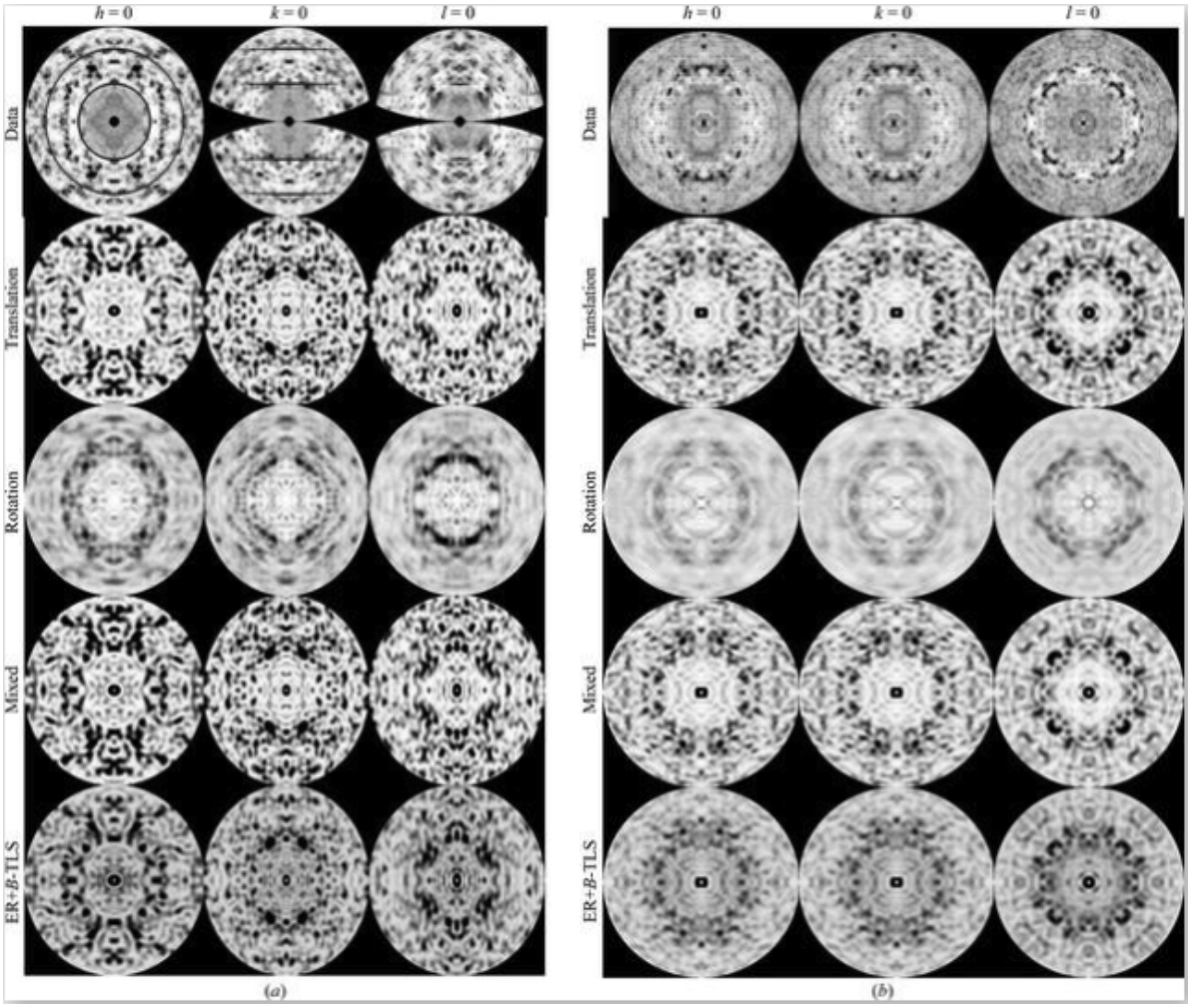


Figure 1.6: Diffuse scattering data predictions, with experimental data in the top row, and the other rows labeled by the RBM model type used to predict the diffuse data: Translation, Rotation, Mixed Translation and Rotation, and Ensemble Plus Rigid Body Motions models from rows two to five. De Klijn *et al.* (2019), *IUCrJ* (CC-BY).

De Klijn *et al.* made a stronger case for the RBM model in a paper confidently titled "Rigid-body motion is the main source of diffuse scattering in protein crystallography" [19]. They generated supercell models of CypA and hen-egg-white lysozyme assuming independent rigid-body translational and rotational motions (fit to the B-factors), as well as internal motions generated by ensemble refinement (after subtraction of TLS contributions to the B-factors) for comparison to experimental diffuse scattering. The mixed rigid-body translational and rotational motions models agreed best with the data, and the addition of internal motions did not significantly improve the agreement with the data, leading them to suggest that RBMs are the main source of diffuse scattering. However, even the best models showed relatively poor agreement with both the full diffuse scattering data (CC of 0.47 for the mixed RBM model) and the anisotropic component (CC of 0.53 for the same). Though the authors say that the diffuse scattering maps predicted by the RBM models show "remarkable resemblance" to the experimental data, the reader is invited to draw their own conclusions (Figure 1.6).

1.4.2 The Liquid-like Motions (LLM) Model

The liquid-like Motions (LLM) model starts with the assumption that the displacement of atoms from their average positions are harmonic, but it adds the additional assumptions that the displacements of the atoms are isotropic about the average position and the elements of the displacement covariance matrix $\mathbf{V}_{\mathbf{n}j\mathbf{m}k}$ decay to zero as the distance between atoms j and k ($\Delta\mathbf{r}_{\mathbf{n}j\mathbf{m}k}$) increases. The connection between this assumption and the model's name can be made by analogy to water: the correlation between the displacements of water molecules is close to unity when the molecules are adjacent to each other, and decays to zero as we consider molecules further and further apart. Under these assumptions, the formula for the first-order approximation to the diffuse scattering becomes:

$$I_{\mathrm{D,LLM}}(\mathbf{q}) = \sum_{\mathbf{n}j} \sum_{\mathbf{m}k} f_j(q) f_k^*(q) e^{i\mathbf{q}\cdot\Delta\mathbf{r}_{\mathbf{n}j\mathbf{m}k}} T_j(\mathbf{q}) T_k(\mathbf{q}) \{ e^{iq^2\sqrt{\langle u_j^2\rangle\langle u_k^2\rangle}\Gamma(\Delta\mathbf{r}_{\mathbf{n}j\mathbf{m}k})} - 1 \}$$

where $\Gamma(\Delta\mathbf{r}_{\mathbf{n}j\mathbf{m}k})$ is the correlation function describing the decay in correlation between the atomic displacements of atoms j and k as a function of their average distance apart. Here, we can see that (setting aside the Debye-Waller factors, which are refined from the Bragg scattering) we need only refine the mean squared displacements ($\langle u_j^2\rangle$, and $\langle u_k^2\rangle$) and the parameter(s) of the correlation function Γ, which is normally chosen to be a sum of exponential terms: $\Gamma(\Delta\mathbf{r}_{\mathbf{n}j\mathbf{m}k}) = \sum_p e^{-|\Delta\mathbf{r}_{\mathbf{n}j\mathbf{m}k}|/\gamma_p}$. Very often, this sum is restricted to one or two terms, leaving the relatively straightforward task of refining one or two global correlation parameters (γ_p) against the diffuse scattering (along with the mean squared displacements, $\langle u^2\rangle$).

If we assume all mean squared atomic displacements are the same, and that correlations are the same regardless of location in the crystal and only depend on the distance between atoms, the expression becomes even simpler:

$$I_{\mathrm{D,LLM}}(\mathbf{q}) = 1/(2\pi)^3 q^2 \langle u^2\rangle e^{-q^2\langle u^2\rangle} |F_{\mathrm{avg.}}(\mathbf{q})|^2 * \hat{\Gamma}(\mathbf{q})$$

where $F_{\mathrm{avg.}}$ is the structure factor of a reference crystal with all atoms at their average positions.

That is, the diffuse scattering for the liquid-like motions model is proportional to the convolution of the intensity with the Fourier transform of the correlation function (or, the product of the Patterson and the correlation function in real space, with the correlation function suppressing features of the Patterson at large separation distances). How much the Patterson is "smeared out" by the correlation function depends on the functional form and parameters of the correlation function. This expression can be easily extended to include *anisotropic* correlation by making the correlation function dependent on the *direction* of the displacement between atoms as well as their distance.

The liquid-like motions model as presented above was introduced by Caspar, Clarage and collaborators in 1988, in a study of the diffraction from crystalline insulin [9]. They found a correlation length of 6 Angstroms and an RMS displacement of about 0.4 Angstroms. The same system was studied by Caspar, Clarage and collaborators in both the tetragonal and triclinic crystalline forms, and they found the same correlation length, with slightly different RMS displacements (0.33 and 0.49 for triclinic and tetragonal, respectively)[14]. A series of studies by Faure *et al.* published in 1994 analyzed the predictions of the normal modes, molecular dynamics, and liquid-like motions models in lysozyme. Although the authors were unable to conclusively differentiate between the accuracy of the models, they calculated that the longer-distance correlations of the normal modes perturbations should produce diffuse features about 125 times more intense than those from shorter-range correlated motions, as in the LLM model [24, 25]. As with the early studies discussed in the section above on the *Rigid Body Motions Model*, these studies were performed largely with single diffraction images, analyzed qualitatively, and fit by trial and error.

In 1997, Wall, Ealick, and Gruner published a study in which they were able to collect three-dimensional diffuse scattering data for the first time, from crystals of *Staphylococcal nuclease*. Fitting quantitatively to this data, they found the best agreement with a LLM-like diffuse scattering model with a correlation length of 10 Angstroms and an RMS displacement of 0.36 Angstroms [104]. This study also introduced R-factors and correlation coefficients as measures of goodness-of-fit for diffuse scattering data.

Phillips, Clarage, and Wall performed a similar study on a calmodulin-peptide complex published in the same year[103]. In this study, they tested two different correlation functions in the LLM model: one which decays as usual ($\Gamma(\Delta r) = e^{-|\Delta r|/\gamma}$) and another which adds a factor out front ($\Gamma(\Delta r) = \frac{1}{\Delta r} e^{-|\Delta r|/\gamma}$), testing both an isotropic and anistropic version for both. The second correlation function was so chosen as in the large-$\mathbf{q}$-limit the functional form for Fourier transform of the correlation function would resemble that of a Debye solid,

and the MSD and correlation length parameters could be used to estimate values for the entropy and specific heat of the crystal. Analysis of the large-scale diffuse features using the isotropic models yielded an RMS displacement of about 0.4 Angstroms with a correlation length of 5 Angstroms for the first correlation function and 12 Angstroms for the second, but was not able to definitively say which functional form was better (the R-factor for both was about 0.4).

The streaks in the diffuse scattering were well modelled (Figure 1.7[2]) using an *anisotropic* LLM model, which uses the same correlation functions as above, but replaces the isotropic variance with a matrix corresponding to the variance in each spatial dimension. This anisotropic model yields a correlation *matrix* rather than a correlation length. This anisotropic LLM model showed highest correlation for a 0.4 Angstrom RMS displacement with a correlation length of 135 Angstroms along the end-to-end packing direction of the unit cell, and also showed weakest correlation along the direction of the lattice containing large a solvent cavity and a flexible portion of the protein.

More tests of the liquid-like motions model in macromolecular crystallography were performed in a series of studies published by Meinhold and Smith in 2005 and 2007 [67, 65]. The first study did not refine the parameters of the liquid-like motions model directly against the data, but showed evidence from molecular dynamics simulations that C-α displacement correlations decay as a function of distance, on average, for both

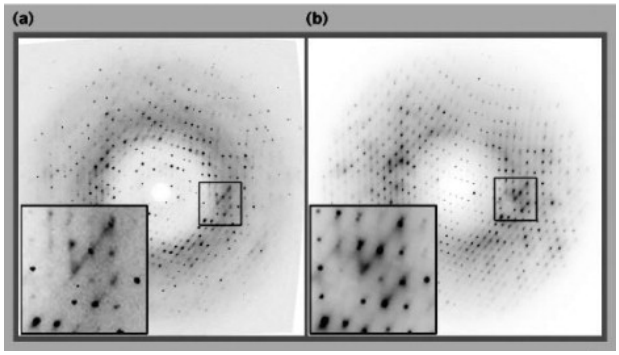


Figure 1.7: Figure from Wall, Clarage, and Philips (1997), showing an experimental diffraction image (left) and simulated diffraction data (right) using the anisotropic LLM model, showing that the streaks between Bragg peaks are well reproduced.

[2]Reprinted from *Structure*, 5 (12), Wall M., Clarage J., and Phillips G., "Motions of calmodulin characterized using both Bragg and diffuse X-ray scattering", 1599-1612, Copyright (1997), with permission from Elsevier.

intraprotein and interprotein atom pairs. The second study performed an analysis of both LLM and molecular dynamics (MD) models of diffuse scattering, and found that MD and LLM models differ in the correlation lengths predicted, and the convergence between MD and LLM-derived diffuse scattering patterns is poor. Another study by Riccardi, Phillips, and Cui in 2010 evaluated the liquid-like, rigid body, and normal modes modes of diffuse scattering against data collected from *Staph. nuclease*, and found good agreement for the liquid-like motions model with a correlation length of 10 Angstroms [82]. They also found that a TLS (rigid body) model yields very good agreement with the B-factors but suggests atomic correlations which are much too high, when compared to the other models.

More recently, van Benschoten and collaborators tested the LLM model against high quality diffuse data from CypA and Trypsin collected with a pixel-array detector, and found relatively good agreement [95]. They analyzed the model-data agreement by correlation coefficient (CC) to the anisotropic component of the diffuse scattering in spherical shells of reciprocal space, and found the best agreement for CypA with a correlation length of 7.1 Åand an RMS displacement of 0.38 Angstrom (total CC=0.58 total; max CC=0.74 in 3.67-3.28 Angstrom shell) and for Trypsin with a correlation length of 8.35 A and an RMS displacement of 0.32 Angstrom (total CC=0.44 Angstrom; max CC=0.72 in 4.53-4 Angstrom shell).

A study conducted by Arianna Peck, Frederick Poitevin, and T.J. Lane from 2018 analyzed the diffuse scattering from three different crystalline systems, and tested single-protein rigid body, elastic network, and two liquid-like motions models (one with correlations restricted to the asymmetric unit, one with correlations at arbitrary distances) [74]. Though they found that none of the models could consistently reproduce the anisotropic data with a CC with experiment greater than 0.5, the asymmetric-unit-constrained liquid-like motions model consistently out-performed the elastic network and rigid body motions models: this LLM models had correlation lengths of about 18 Angstroms, much higher than correlation lengths

of previous studies. However, when LLM-like correlations were allowed to extend beyond the asymmetric unit, the correlation coefficient with experimental data increased dramatically, to a CC of 0.67-0.71. This result suggested that correlated motions which extend across proteins are important in the modelling of diffuse scattering, a finding supported by work to be discussed in Chapter 3.

1.4.3 The Elastic Network/Normal Modes (EN/NM) Models

The Elastic Network (EN) model is an extension of the Normal Modes (NM) model's assumption of harmonic disorder about the atoms' average positions, in which the entire collection of atoms is assumed to behave as masses connected by springs. The EN model is able to reproduce the low-frequency modes present in NM model with two main advantages: (i) the NM model has 3N parameters for a protein of N atoms, whereas the number of parameters in the EN model depends on the distance cutoff used to define connections between nodes (ii) in the EN model, the minimum energy configuration can be defined with respect to the crystal structure, whereas in the NM model, the energy minimum is defined by the potential, and can be quite far from the crystal structure [82]. Thermal excitation drives the ensemble of masses and springs toward collective motions described by the modes of the system, which are eigenvectors of the covariance matrix. These modes are used as a basis set with which to construct a structural ensemble, and the dynamics implied by this ensemble are used to predict the diffuse scattering.

Though normal modes models had been used to predict B-factors and ADPs for quite some time, the first study using normal modes to predict diffuse scattering was conducted by Mizuguchi, Kidera and Go in 1994.[70]. At the time, the quality of experimental data was too poor to do any quantitative analysis of agreement, but they were able to show the sensitivities of the features in the predicted diffuse scattering to changes in the adjustable

parameters of the LLM and NM models. Faure, *et al.* were the first to directly compare the predictions of the liquid-like, molecular dynamics, and normal modes models to experimental diffuse scattering data, though only qualitatively [24, 25]. This work showed that long-distance correlations predicted by the normal modes model should produce diffuse features many times more intense than those produced from short-range LLM-like correlations.

Riccardi, Phillips and Cui published results in 2010[82] comparing the predictions of various elastic network models (one empirical, one parametrized), an LLM model, and a rigid-body TLS model. First, for elastic network models, they found that moving from isolated molecules to molecules in a lattice context significantly improved agreement with the temperature factors. They also found that both the elastic network models qualitatively agree with the liquid-like motions model when a sufficient number of modes are included, however they diverge from each other quite drastically for the same number of wavevectors. When the number of modes included is drastically reduced, both models become qualitatively similar to the TLS model.

This is to be expected, as the LLM model is a special case of the normal modes/elastic network model under additional assumptions [99], and with few modes the proteins behave more as a collection of rigid groups. This will be a common feature of diffuse scattering models: the difficultly in establishing a "best" model for the prediction of diffuse scattering comes largely from how similar the models can behave with the right choices of model assumptions and parameters.

In their own comparison study, Peck, Lane, and Poitevin refined an elastic network model against Bragg data before comparing the predicted diffuse scattering with data from three crystalline systems, finding minimal correlation (CC=0.09-0.2) [74]. They suggest that allowing the parameters to be refined against the diffuse data may have improved agreement, however, they also note that these models would probably remain poor so long as the networks were restricted to intra-protein atom pairs. These comments would prove prescient.

A elastic network model of a unit cell (rather than a single protein) computed by Wall, Fraser and Wolf for *Staphyloccocal nuclease*, refined against diffuse scattering data, was able to achieve much higher correlations with the anisotropic diffuse scattering data (CC=0.54). However, a more complex elastic network model would prove to be even more successful[106].

In 2020, Ando, Meisburger and Case published the most comprehensive and high-quality analysis of diffuse scattering and models of correlated disorder to date [68]. They collected an extremely high-quality, fine-grained three-dimensional diffuse scattering map from triclinic lysozyme at room temperature with a photon-counting pixel array detector, carefully accounting for and removing experimental noise. First, they were able to connect the power-law decay in the intensity of the haloes around Bragg peaks to long range correlations from acoustic phonon-like vibrational lattice dynamics. Moreover, they were able to construct a lattice aware elastic network model which treated proteins as rigid bodies connected to their neighbors by Gaussian and directional springs. This model was able to reproduce the anisotropic haloes around a set of 400 Bragg peaks. This lattice dynamics model was not able to accurately reproduce the experimentally refined ADPs. However, when an internal elastic network model (with rigid side chains) was coupled to this lattice aware elastic network model, the agreement with the experimentally-derived ADPs improved dramatically. This combined internal and lattice-connected elastic network model showed remarkable agreement with the experimental diffuse Patterson function, and reproducing the decay in diffuse Patterson fluctuations as a function of the scattering vector almost exactly. This combined internal and lattice-connected elastic network model outperformed either model alone, in addition to significantly outperforming a crystalline MD simulation with 343 unit cells, especially at high resolution.

This work was a significant advance. By suppressing domain-specific internal motions in their combined internal and lattice-connected elastic network model, they were able to validate (for the first time) a model of internal dynamics of a protein which could not have been

distinguished from the domain-motion-suppressed model by Bragg data alone, by comparison to the diffuse Patterson (Figure 5 in [68]). The improvement in CC with the diffuse Patterson was slight (0.01 gain in CC), but consistent across all resolution shells: the authors were careful to point out in the peer review notes that this claim (the ability to distinguish between disorder models which agree equally well with the Bragg data) has been made by other papers, but their work was the best example of such discrimination to date, as they were able to reproduce the total anisotropic scattering to truly impressive levels of accuracy.

1.4.4 The Molecular Dynamics (MD) Model

The Molecular Dynamics model starts with the atomic structure deduced from crystallography or other methods of structure determination, and assigns properties to the atoms based on a set of parameters known as the "force field." The force field parameters include the masses, charges, and van der Waals radii of the atoms, the average bond length and effective bond spring constant, bond angles for each set of three connected atoms, and the torsional parameters which control how freely a set of four atoms rotates around the central bond. These parameters are plugged into a Newtonian functional form for the potential, which is integrated in time-steps to create a stop-motion trajectory describing the dynamics of the system. Different force fields set different values for these parameters and produce different results.

The MD model of collective motions differs from the others discussed above in that it is essentially "model free": the only underlying assumptions of the model are the the force-field parameters (which are fixed) and the quasi-Newtonian framework (no chemistry or quantum mechanics). Though the other models have outperformed MD in their prediction of the anisotropic diffuse scattering, these models have the advantage of tunable parameters, whereas the "parameters" (read: the forcefield) of the MD model are fixed. The main

advantage of the MD model over the LLM, RBM, and NM models for prediction of diffuse scattering is the inclusion of solvent effects. The MD model is the only model which is able to reproduce the full diffuse scattering signal, isotropic and anisotropic. While the accuracy of the force fields has been extensively tested and improved over the preceding decades, mostly through single-molecule experiments, the accuracy of their predictions for protein-protein interactions and crystal contacts is still being tested [73, 46].

In a sense, the true power of the MD model has never been in its own predictive capacity, but in its ability to analyze and contextualize the assumptions underlying other models. From the early days of the modelling of correlated motions and diffuse scattering, MD simulations have been used to validate and contextualize the motions predicted by normal modes models [41, 43], and rigid body motions models [28, 92]. It wasn't until the work of Faure *et al.* in 1994 [24] that a 600ps MD simulation was used to predict diffuse scattering directly: this work showed that diffuse scattering predicted by NM and MD models were similar in form, however, the finer details at higher resolution were better captured by the normal modes model. Work by Clarage, Phillips and collaborators [15] showed that the deficiencies of the MD model may be due to sampling: a 1ns simulation of a myoglobin crystal was shown to insufficiently sample collection motions and correlations in atomic positions in low-frequency modes. The atomic displacement covariance matrices calculated in the first and second halves of the simulation were significantly different, indicating a lack of convergence, and the diffuse scattering calculated from the trajectory did not match experiment.

The first promising analyses of diffuse scattering using MD simulations were carried about by Meinhold and Smith, publishing three papers between 2005 and 2007 [67, 66, 65]. In all three of these works, Meinhold and Smith ran 10ns NPT simulations of a unit cell of Staphyloccocal nuclease, and calculated the diffuse scattering using Guinier's equation, for comparison with diffuse scattering data collected by Wall, Ealick and Gruner[104].

In the first of these works[67], they started by showing that the B-factors from experiment

are over-predicted by the MD in regions of high flexibility, and that rigid body translations and rotations contribute only a minority portion (about 0.25 A^2) of the total mean squared fluctuations. They also showed that the correlation in atomic displacements decays as a function of their pairwise distance apart, but the average correlation as a function of pairwise distance was different for intra- and interprotein atoms pairs. Additionally, they showed that the agreement between calculated and predicted diffuse scattering improves logarithmically as a function of simulation length. They found, similarly, that the convergence of displacement covariance matrix elements for intraprotein atom pairs converged logarithmically as a function of simulation time, though the fraction of converged matrix elements increased differently in different simulations, indicating that different simulations may get trapped in different free-energy basins.

In the second of these works[66], the isotropic profile of the total diffuse scattering calculated from the full simulation was shown to be in good agreement with the experimental profile (though missing a "shoulder", or dip in intensity, at slightly higher resolution than the peak intensity) and they were also able to show that the agreement between the calculated and experimental diffuse intensity improved logarithmically as a function of simulation length, and that the secondary structure elements contribute most significantly to the peak in the diffuse scattering profile. They were also able to use Principal Component Analysis (PCA) clustering to associated intense features in the MD-predicted diffuse scattering profile with specific intra- and interprotien motions from the trajectory, and that the first five PCA features associated with intramolecular cross-correlations contributed the most to the large-scale features in the diffuse scattering.

In the third of these works, multiple 10ns simulation were analyzed and the results from the first were reproduced, but in addition they estimated that the variance-covarance matrix of atomic displacements converges on the scale of $\sim$1 microsecond timescale, and that the correlation coefficient of pairwise atomic displacements decayed as a function of their pairwise

distance apart, for both inter-protein and intra-protein atom pairs. This is consistent with the LLM, or "isotropic correlations (IC)" model (as they call it in the paper). The values for the correlation length found by fitting an exponential to the calculated pairwise correlation with distance were consistent (10 Å) with LLM-refinement against diffuse data from previous experiments by Wall, Ealick and Gruner [104], however, when an LLM model was fit to the diffuse scattering data predicted from MD, the correlation length was higher ($\sim$14 Å).

Michael E. Wall and collaborators followed up on these studies with a 1.1 microsecond simulation of a Staph. nuclease unit cell in 2014 [105]. In this case, the correlation between the calculated and experimental diffuse scattering data was excellent across many different time scales (0.99 and above for 10, 100 and 1000 ns trajectories), and robust to both sampling timescales and changes in the forcefield. However, when the anisotropic diffuse scattering was isolated, and compared with experimental data, the correlation coefficient dropped to 0.646 for 10ns, 0.654 for 100 ns and 0.832 for 1000 ns. They also found that, were they to test the reproducibility of the C-α displacement covariance matrix elements, as in Meinhold and Smith (2005), the agreement between the matrix elements from the first and second halves of the simulation was low (CC=0.517). This suggested that the convergence of the elements of the C-α displacement covariance matrix is a poor measure of the convergence of the diffuse scattering as a whole. This work was also able to differentiate between the contributions to the diffuse scattering from the protein, the solvent, and the protein-solvent cross-term, showing that though the total intensity contains roughly equal contribution from the protein and solvent, the anisotropic component of the diffuse intensity results almost entirely from fluctuations in the structure of the protein.

Follow up work by Wall in 2018 [100] with an even longer and larger (5 microsecond, 2x2x2 unit cell) trajectory confirmed that a one microsecond trajectory is sufficiently converged to model the diffuse scattering, albeit perhaps only for this system. In fact, the agreement with the anisotropic diffuse scattering was roughly consistent across all 100 ns chunks of

the full 5 microsecond simulation, with correlation coefficients ranging from 0.62 to 0.68 – these are the highest correlation coefficients with the anisotropic component of the diffuse scattering yet achieved by MD simulations, probably owing to the supercell (vs. unit cell) simulation paradigm. As suggested by other studies, the inclusion of correlated motions across unit cell boundaries appears to be important in modelling the anisotropic component of the diffuse scattering. The qualitative agreement between the anisotropic components of the simulated and experimental diffuse scattering maps is quite good, though some large scale features are not reproduced. The similarity in the time required for convergence of the diffuse scattering between unit cell and supercell simulations suggests that the convergence time may be independent of system size above a certain number of unit cells.

The agreement between predicted and experimental B-factors was very good (CC=0.94), similar to the agreement with the B-factors from the unit cell simulation (CC=0.95), and better than the agreement found by another supercell simulation in 2016 from Case *et al.* for hen-egg-white lysozyme (CC=0.78 for all heavy atoms) [46]. In both of these supercell simulations, it was observed that the structure experiences a small but significant large-scale drift away from the crystal structure atomic positions across the full supercell. Case *et al.* posit that this drift may be due to deficiencies in the modelling of crystal contacts.

Finally, in the work by Ando, Meisburger and Case mentioned above in the section on *Elastic Network/Normal Modes (EN/NM)* models, MD simulations of hen-egg-white lysozyme were prepared with 1, 3x3x3, 5x5x5, and 11x11x13 unit cells (for 5, 5, 2, and 1 microsecond simulation time respectively). The supercell systems were able to model the anisotropic haloes around Bragg peaks (associated with phonon-like acoustic vibrations) more accurately as more unit cells were included, however it overestimated the standard deviation profile of these features in the largest supercell system, which they were able to be reproduce almost exactly with their combined internal and lattice-connected elastic network model. The MD model was also less successful at reproducing the features of the diffuse Patterson map,

suggesting that there is still ample room for improvement in the MD modelling of diffuse scattering, and MD force field treatment of large-scale atomic correlations more broadly. Unlike the elastic network models, no MD-predicted internal protein motions have been well validated against features in the diffuse scattering maps, nor the diffuse Patterson (though agreement with the Patterson can be quite good; the Patterson predicted by crystalline MD in Wall [2018] had a CC with the experimental Patterson of 0.7, whereas the combined internal and lattice-connected elastic network model had a CC of about 0.8 or more).

Chapter 2

Background on MD Simulation of Protein Crystals and modelling of Densities and Dynamics

Molecular Dynamics (MD) simulations have been used for decades as a means to probe dynamics at atomic detail and at time-scales down to the order of femtoseconds or picoseconds, both of which are not easily accessible using more standard "bench" tools in biochemistry and structural biology. The accuracy of molecular dynamics simulations depends on how well parametrized the force fields are, however, leveraging experimental measurements to improve force field parameterization has proven challenging. Many force fields (and, in particular, the "gold standard" force fields from AMBER and CHARMM) have proven useful in expanding our understanding of biochemical phenomena such as ligand-binding[16], the relationship between protein structure and function[71], and allostery[33], but they can perform quite differently (especially with regard to torsional free energy profiles[79]), and there is ample room for improvement (in many respects, but particularly with regard to protein-protein interactions and/or crowded environments[73, 76]).

Crystalline MD (abbreviated cMD below) systems are prepared and parametrized identically to standard solution single-protein systems (sMD), with a few idiosyncrasies, to be discussed below. However, production simulation and analysis of these systems requires some care, knowledge of the crystallographic context (reviewed in the previous chapter), and some bespoke software. Despite computational advances, cMD simulations can often be very expensive, with microsecond-scale "supercell" simulations (many unit cells) requiring dozens of hours of compute time on thousands of cores, even with message-passing-interface (MPI) enabled clusters with high-bandwidth, low-latency connections between nodes. Below, I present an overview of the preparation and simulation of cMD systems, along with a review of previous work.

2.1 Crystalline MD System Preparation

Molecular dynamics simulations require a structure file and a topology file as input: the structure file simply lists the identity and coordinates all of the atoms present in the system, while the topology file describes the properties of the atoms and their bonded and non-bonded interactions with each other (based on the parameters of force field). Very often the structures from crystallographic refinement will have missing atoms in regions of the density that are not well resolved, and those parts of the structure need to be built back in, either by homology modelling, or simpler forms of modelling based on common backbone torsional profiles and rotomeric states for side-chains (now easily accomplished, through programs like `phenix`[57], `coot`[23], `UCSF Chimera`[77], and others). Once one has arrived at a complete structure for the protein and any bound ligands or co-factors, one must choose a force field for parameterization, to create the topology file.

Waters present in the crystal structure are often (but not always) removed before construction of the unit cell or supercell systems, for two reasons: (i) we often wish to test

the force field with respect to the reproducibility of crystallographic waters, and (ii) minimization with crystal waters present can lead to "unsettled water" errors (waters stuck in prohibitively high-energy configurations), either due to interactions with the protein, or clashes with waters placed in by solvation. The gold standard MD force fields and TIP3P or SPC/E waters perform very well at reproducing crystal waters in the appropriate positions if crystallographic waters are removed from the start, especially when the system is restrained to the crystal structure [102].

Once the asymmetric unit structure is complete, hydrogens are added with protonation states determined by the experimental pH (with the help of a software package, such as `proPKA`[90, 72] or `pdb2pqr`[21]), and the structure is propagated to a unit cell using the appropriate symmetry operation for the system. This can be accomplished either on an atom by atom basis, using a code similar to the pseudo-code outlined below, or using the `UnitCell` program from the `ambertools` software suite [8], which reads the unit cell size and space group information directly from the `CRYST1` and `SCALE` records in the `.pdb` structure file. Once a unit cell has been constructed, copies of the unit cell can be appended along any of the unit cell faces to create a supercell of arbitrary size.

It is also possible to seed the supercell with structures from an ensemble model. Ensemble refinement software (such as `phenix.ensemble_refinement`[7]) allows one to generate a valid ensemble with a chosen number of structures. As such, if one is generating an ensemble with the explicit purpose of seeding a supercell, one should generate an ensemble with the number of structures necessary to use a different ensemble member at each position in the crystal, and send each member of the ensemble to a different position in the supercell, using the unit cell box size and space group information. However, if one is working with an ensemble of a different size, care should be taken to seed the supercell in such a way as to minimize bias (for instance, taking care not to place the same structure adjacent to itself in the supercell, which may lead to false correlations in dynamics).

PSEUSDO-CODE FOR SYMMETRY PROPAGATION

```
def UnitCell(protein, uc_id):
    '''Generates symmetry-propagated copies of of a protein in a P212121 unit cell
    Input
    -----
    protein (object) -- class containing (i) an array (atoms) of atom objects with
    coordinates x, y, and z, and (ii) crystinfo attribute containing unit cell side lengths
    a, b, and c
    uc_id (int) -- index from 0 to 3, indicating which unit cell protein to output
    Output
    ------
    protein (object)'''
    for atom in protein.atoms:
        #convert to fractional coordinates
        atom.x /= protein.crystinfo.a
        atom.y /= protein.crystinfo.b
        atom.z /= protein.crystinfo.c
        #if assymetric unit, pass
        if uc_id = 0:
            pass
        #else, use the symmetry operations, in this case P212121
        if uc_id = 1:
            atom.x =  atom.x + 0.5
            atom.y = -atom.y + 0.5
            atom.z = -atom.z
        if uc_id = 2:
            atom.x = -atom.x + 0.5
            atom.y = -atom.y
            atom.z =  atom.z + 1/2
        if uc_id = 3:
            atom.x = -atom.x
            atom.y =  atom.y + 0.5
            atom.z = -atom.z + 0.5
        else:
            print("uc_id must be integer from 0 to 3"); break
        #convert back to real coordinates
        atom.x *= protein.crystinfo.a
        atom.y *= protein.crystinfo.b
        atom.z *= protein.crystinfo.c

    return protein
```

The complete supercell structure can then be parametrized. It is recommended, if one is working with a structure with bound ligands or cofactors, to separate off each constituent in to its own structure file to be parametrized, using `tleap` or `pdb2gmx`, from the AMBER[8] and GROMACS[5] software suites, respectively. This allows one to ensure that the parametrization is successful for each individual component before combining them. In AMBER, one can use `tleap` to parametrize the constituents individually, and then `source` the appropriate `.lib` `.frcmod` and `.param` files at the start of the parametrization script for the full system, before loading in the structure file for supercell. In `GROMACS`, one can prepare each constituent with `pdb2gmx`, pointing toward an appropriate `forcefield.itp` file for each, and concatenate the `.top` files from each individual parametrization (either manually, or by sourcing each constituent's topology information as a `.tpr` file), and convert the supercell `.pdb` structure file to a `.gro` file (using, e.g., `parmed`[86]), ensuring that the `[moleculetype]` section at the end of the full system `.top` file has all of the molecules listed in the correct order and with the correct number of occurrences. As an aside, we have found the "SwissParam" server (`https://www.swissparam.ch`) especially handy for the parametrization of small molecules, as it only requires a `.mol2` file as input, and outputs a `GROMACS`-ready `.itp` file with parameters from the Merck Molecular Force Field (MMFF) and CHARMM22, which we have found to be satisfactory[113], though one can also use the Generalized Amber Force Field (`gaff2`).

The system must then be solvated. Often, these systems have a net charge, so one must replace the appropriate number of water molecules with neutralizing ions to bring the full system to net zero charge. Additionally, it is advisable, though perhaps not necessary, to reproduce the solute content of the crystallographic mother liquor as accurately as possible. Cerutti *et al.* (2008) showed that a cMD simulation of biotin-bound streptavidin with cryoprotectant, high-salt-content mother liquor accurately reproduced improved agreement with the data versus simulation in water only [11]. Whether or not the concentrations of solute in the mother liquor are preserved in the cavities of the crystal is an open question,

though certain experimental considerations are appropriate: for instance, if hanging-drop diffusion or evaporation were used to prepare the crystals, the concentration of solute in the crystal is likely to be higher than in the bulk mother liquor [10]. However, inclusion of various components of the mother liquor can be difficult to achieve if, for example, accurate and/or well-tested parameters for the components are not available. For components that can be included, one can estimate the number of molecules which need to be added (knowing the molarity) by either (i) using the reported solvent content or the Matthews coefficient[1][62] and the unit cell volume to estimate the volume of solvent in the unit cell or (ii) solvating the crystal, and using the number of waters added to estimate the volume of solvent, using the molar mass and density of water (updating this estimate if more water is added).

2.1.1 Constant pressure (NPT) versus constant volume (NVT) ensembles for solvation and equilibration

Because the system has been prepared to exactly reproduce the crystalline structure using the unit cell and space group information, special attention must be paid to the preservation of the length and aspect ratio of the box sides, as changes would alter the crystal's precise symmetry. However, solvation presents a challenge. Standard solvation methods in both `AMBER` and `GROMACS` software suites tile the supercell volume with smaller, pre-equilibrated boxes of solvent, and remove waters which clash with the atoms in the system.

Case and Cerutti [10] warned that this tiling-and-culling procedure can lead to vacuum bubbles, which could alter the volume of the system during NPT equilibration. They recommend a solvation procedure with a uniform distribution of waters placed around the protein, followed by restrained energy minimization and more weakly restrained equilibration, tun-

[1]The Matthews coefficient is the crystal volume per unit protein molecular weight, which Matthews showed has a straightforward relationship to the solvent content: $V_{\text{solvent}} = 1 - \frac{1.23}{V_M}$, where V_{solvent} is the fractional solvent volume and V_M is the Matthews coefficient in cubic angstroms per dalton

ing the number and distribution of waters placed around the protein so that the box size and aspect ratio equilibrates correctly. However, we find that one can adequately solvate the crystalline system while retaining the correct box size and aspect ratio by rigorously enforcing the box dimensions in the NVT ensemble — albeit, the NVT restriction requires additional rounds of solvation, minimization and equilibration, as the initial system is often drastically under-pressurized. Case and others have since also moved to simulating almost exclusively in NVT [45, 105, 47, 102, 100, 68], making this practice the *de facto* standard for the simulation of protein crystals.

Solvent is added, and the entire system is minimized, often using a combination of steepest-descent and conjugate gradient algorithms, with optional restraints to the crystal structure. After initial solvation and minimization, the system is equilibrated with restraints to either the minimized or crystallographic structure (we have found that this choice makes little difference if the production simulations are to be unrestrained). The system at this stage is often drastically under-pressurized in the NVT ensemble (pressure less than -1000 bar in a 2x2x2 supercell, for solvent content greater than 30%), but the system can be subjected to iterative rounds of additional solvation and equilibration, until the pressure is brought up to the neighborhood of 1 bar; in our simulations we repeat this procedure until the average pressure (plus or minus the standard error) lies within the range -100 to 100 bar.

Systems parametrized with different force fields equilibrate to the target pressure at slightly different rates, even with the same number of waters added during each round of solvation. We prepared 2x2x2 unit cell systems of *staph. nuclease* and *endoglucanase*, with the AMBER 14SB and CHARMM 27 force fields, and added an equivalent number of waters to each system for five rounds of iterative solvation and 5 nanoseconds of equilibration (pressure analyzed in the final 2 ns of equilibration). The results are system dependent, but in general, CHARMM produces higher pressures upon initial solvation, and the force fields require different numbers of waters to reach the -100 to 100 bar range. We suspected that this difference

may be due to the fact that CHARMM has Leonard-Jones parameters for polar hydrogens, whereas AMBER does not, so we also performed a test with a system parametrized with the AMBER force field, substituting CHARMM Leonard-Jones parameters for polar hydrogens. We found that this has the effect of increasing the pressure at every level of solvation, but was not sufficient to make the behavior of the two force fields equivalent. One important takeaway is that across both systems and force fields, the pressure comes up roughly linearly as a function of the number of waters added, making it quite simple to estimate the number of waters needed to achieve atmospheric pressure. Automated pipelines for the preparation of crystalline supercell systems can use this property to their advantage.

2.2 Restrained and Unrestrained Production Simulation

The heavy atoms of the protein are tied to the initial propagated crystal structure positions with strong restraints during equilibration, allowing the solvent to exit the bulk and fill in the protein's water network. Restraints are often relaxed for production, or eliminated completely, however unrestrained simulations are often found lacking [112]. Many simulations have used restraint constants on the order of 200 kJ mol^{-1} nm^{-2}, however much weaker restraints can be used, with varied effect (which we will discuss in a subsequent chapter).

Unrestrained simulations consistently show a large-scale "drift" away from the crystal structure: a deterioration of crystal lattice contacts leading to a small but noticeable reorientation of the proteins (through rigid body displacements of entire proteins or domains). Though unrestrained simulations show the unbiased dynamics implied by the force field, this drift makes accurate predictions of crystallographic data more difficult, so it is important to understand how much we can bias the forcefield toward the crystallographic structure without

suppressing the dynamics implied by the force field (this will be discussed in a later chapter as well).

In 2015, Janowski *et al.* published the results of a comprehensive test of the effectiveness of various force fields at reproducing a set of experimental crystallographic measurements, using unrestrained cMD simulations of a 2x2x3 unit-cell triclinic lysozyme system, as well as a test of the similarities and differences between sMD and cMD simulations of the same system (the sMD/cMD comparison will be discussed below) [47]. Though all force fields showed comparable performance in faithfully reproducing structural elements (particularly non-terminal alpha helices and beta sheets) the simulation using the `AMBERff 14SB` force field had the lowest backbone and heavy-atom RMSD to the experimental crystal structure, and the lowest R-work and R-free when refining the experimental model into the simulation average electron density. Refining a structural model against the structure factors calculated from best-performing simulation (as one would refine a structural model against data from diffraction images) produced R_{work} and R_{free} values on par with the experimental results. The comparison also showed that forcefields have improved over time, with `AMBER`'s `ff14SB` fore field exhibiting lower backbone and heavy-atom RMSDs to the crystal structure and higher agreement with B-factors than `ff99SB` (with the improvement attributed to side-chain torsion energy profiles)[2]. None of the simulations produced RMSD deviations from the crystal structure greater than 0.5 angstroms for the backbone and 1 angstrom for heavy atoms. This finding was in line with work published by Hu and Jiang in 2009 [42], showing that, for a four unit cell cMD tetragonal lysozyme system, the OPLS-AA and AMBER03 force fields outperformed GROMOS96: exhibited lower RMSDs to the experimental crystal structure and showed better agreement with experimental B-factors (though the difference may also have been attributable to differences in the force field *family* rather than improvements over time).

[2]Other studies have confirmed the improvement of forcefields over time, by comparison to NMR data and melting curves, though different force field families perform differently with respect to the formation and stability of various secondary structure elements[60]

The main differences between the simulations using the various force fields were found in their modelling of 3_{10} helices, with differences in structure more generally attributable mostly to differences in the modelling of hydrogen bonds. However, all forcefields produced particularly high fluctuations for solvent-exposed side chains (higher than would be expected from experimental B-factors, though refined B-factors are known to underestimate true average atomic fluctuations [54]).

Consistent with other cMD simulations (and the simulations in the work to be presented here), Janowski *et al.* saw a large-scale deterioration in the crystal lattice over the course of the unrestrained simulations [47]. This drift is not substantial enough to dissolve the crystal lattice completely, just a noticeable and significant difference in average structure. They observed that each monomer's average COM was between 0.2 and 0.5 angstroms away from the ideal crystal lattice position – increasing with simulation time, but plateauing after about a microsecond – with the amount of drift being force field dependent. Some monomers drift away progressively over time, while some drift away and return. They suggest that this drift may be due to inaccuracies in the modelling of crystal contacts: most of the hydrogen bonds across the the interfaces between unit cells were reproduced less than 50% of the time, but the data was not consistent enough to provide insight about the specific source of the deterioration (especially one that is consistent across all the tested force fields).

2.3 Ordered water in crystalline MD simulations

Large-scale structural drift in unrestrained crystalline systems presents a modest problem for prediction of atomic coordinates and fluctuations for the protein, but appears to be a much larger problem for the prediction of the positions of ordered solvent molecules, with small-scale structural drift in the backbone and side chains leading to significant knock-on effects in the modelling of ordered solvent. This is important, as interactions with ordered waters can

affect the protein structure, allosteric interactions and, particularly, the binding of ligands. Though water models (e.g. TIP3P and SPC/E) are tuned to reproduce thermodynamic properties, they may be lacking in their ability to reproduce complex solvent dynamics and interactions with protein side chains. However, the *choice* of water model may not be all that important in this regard: Hu and Jiang [42] found that, for a four unit cell cMD simulation of tetragonal lysozyme, the choice of water model affected diffusion dynamics, but had little effect on the structural and energetic properties of the system. That is to say, all water models may perform equally well, but may all be found equally lacking when it comes to fine-scale modelling of dynamics coupled to the protein. This is in line with findings from Gilson and Henriksen, whose studies of host-guest systems (small models used as test beds for MD simulation methodologies of protein-ligand binding) found that though different water models produced similar predictions for binding free energies, the predictions for enthalpies and entropies of binding could vary significantly, and the positions of waters around the host-guest systems were different, depending on the protein force field and water model used[36].

In an article published in 2018, Wall *et al.* [102] showed that a 1 microsecond 2x2x2 unit cell simulation of endoglucanase with a force constant of 209.2 kJ mol^{-1} nm^{-2} for harmonic restraints to heavy atoms was able to reproduce the crystal waters with remarkable precision and recall. The precision and recall statistic measures the number of MD-predicted waters within a certain distance of the nearest crystallographic water, providing a robust metric for the accuracy of ordered water prediction in MD studies. Ninety five percent of the top one hundred crystallographic waters (defined based on the strength of the density peak) were reproduced to within one angstrom; ninety eight percent were reproduced to within 1.4 angstroms. Elimination of these restraints reduced the precision and recall dramatically, with 40 to 50 percent of the top one hundred crystallographic waters reproduced to within one angstrom and 50 to 60 percent reproduced to within 1.4 angstroms. These results are in line with our work on *protein kinase A* (PKA), to be discussed in a later chapter, but

I will note here that ordered solvent prediction worsens slowly with reduction in the force constant for restraints, with marked reduction in precision and recall occurring between 20 and 2 $\mathrm{kJ \cdot mol^{-1} \cdot nm^{-2}}$.

These precision and recall statistics were more impressive than comparable studies. Nakasako and Higo (2002)[39], and Atlan *et al.* (2018)[1], were only able to reproduce 60-70% of crystallographic waters with similar precision. However, the simulation paradigms were quite different. Nakasako and Higo ran solution single-protein simulations for only a nanosecond, computed the density by counting water molecules in voxels (rather than calculating structure factors), and compared to cryogenic X-ray data, whereas the results of Wall and Mobley *et al.* were compared to room-temperature X-ray and neutron diffraction. Atlan et al ran four room-temperature simulations of a Yb^{3+}-substituted mannose-binding protein (MBP) unit cell, each with one of the four dimers removed – to facilitate faster sampling of protein-protein interactions and to make the solvation process easier to tune – and averaged the results, for comparison to cryogenic X-ray diffraction data. These simulations were all run with very strong (1000 $\mathrm{kJ \cdot mol^{-1} \cdot nm^{-2}}$) restraints on heavy atoms. The authors report that these strong restraints were necessary to compare the MD water density with the experimental density, and to ensure that side-chains did not change conformation, however, it seems possible that restraints this severe may *over*-bias the simulation toward the crystallographic model, preventing the MD forcefield from falling in to local minima more amenable to accurate prediction of the solvent. Thus, the weaker restraints of Wall *et al.* may have helped that study achieve better results.

All told, room temperature cMD simulations with moderate restraints, compared to room-temperature diffraction data (X-ray, or neutron, or both) appear to provide the best model for ordered water in crystalline protein systems. However, even in cases where the cMD is able to reproduce crystallographic waters with near-ideal precision and recall, R-factors for MD models refined against experimental data remain greater than $\sim$0.13, suggesting once

again that static models, even those with complex modelling of ordered solvent, are not sufficient to fully reproduce even just the Bragg diffraction data (n.b. this R-factor would be considered fairly good for macromolecular crystallography, but small-molecule crystallographers routinely obtain R-factors much lower than this, suggesting that macromolecular crystallographic models need improvement — see the section on the R-factor gap from the previous chapter).

It seems likely, based on improvements moving from single-structure to structurally heterogenous or ensemble-based models [49, 97, 55, 7], that incorporating ordered solvent into models of heterogeneity/disorder will be advantageous. MD models are uniquely suited to this task, as they allow for prediction of not just the Bragg data, but of the full diffuse scattering map, a large part of which results from correlated disorder in the solvent and from solvent-protein interactions. However, even the best MD models show room for improvement: Wall *et al.* [105] showed that the isotropic component of diffuse scattering comes mostly from solvent and protein-solvent interactions and the cMD-predicted isotropic diffuse scattering they calculated exhibited discrepancies in the 0.3-0.6 inverse-angstrom range, which were reproduced in later simulations of the same system [111], suggesting that MD force fields are inadequately modelling correlated disorder in the solvent and from solvent-protein interactions – though there are myriad potential explanations for this discrepancy.

2.4 Single-protein versus crystalline MD

Many studies have shown that cMD simulations provide more accurate models of protein crystallographic data than sMD simulations. Stocker *et al.* studied this question explicitly [91], presenting the results of 2 ns simultions of sMD, single-unit-cell, and double-unit-cell cMD systems. They found that, though the results were similar for both, the cMD simulation reproduces data from both X-ray (B-factor) and NMR data slightly better than the sMD

simulations. Another study by van Walser *et al.* in 2002 [107] showed that agreement with the experimental structure improved as they moved from sMD simulations to 1 unit cell and 2 unit cell cMD simulations. Even more unit cells provide even better agreement and greater detail, as shown in the work by Case, Ando and Meisburger mentioned in the previous chapter: in this case a 7x7x11 cMD model of triclinic lysozyme was necessary to reproduce haloes in diffraction data with reciprocal space sampling of the same resolution as the diffraction data. However, this model was not as successful at modelling either the haloes or the diffuse Patterson as was a lattice-connected full-protein elastic network (EN) model. Taken together, these results suggested that increasing the size of cMD systems is not the best avenue for improvement – there does appear to be a law of diminishing returns. What's more, the fact that the lattice-connected full-protein EN model performed better than a simpler rigid body EN model suggests that dynamics resulting from the coupling of internal protein motions to the lattice make up a significant contribution to the dynamics captured in the diffuse scattering – MD force fields, in their current state, seem to struggle with modelling this exact category of dynamics and interactions.

Janowski *et al.* (2015)[47] conducted a study of the similarities and differences between sMD and cMD simulations of triclinic lysozyme, as mentioned above, and found that the unrestrained supercell system takes much longer to reach equilibrium than the solution single-protein system (roughly an order of magnitude longer)[3]. They point to the constrained crystalline environment, making solvent rearrangement more difficult, in explaining these differences. However, one can also imagine that rearrangements of (particularly solvent exposed) groups of side chains or secondary structure elements that would be easily accessible in the free energy landscape of the sMD simulation are hindered or blocked completely by interactions with neighboring protein residues.

[3]n.b. Cerutti and Case, in their 2018 review, point out that cMD systems "equilibrate" or "converge" on very different time scales depending on the system, and the property being measured or predicted – e.g. backbone position, B-factors, or diffuse scattering[10]

Janowski *et al.* also calculated the instantaneous RMSD to the crystal structure in two different ways: (i) "best fit" RMSD, in which each monomer's trajectory was aligned to the crystal structure in each frame, measuring only fluctuations due to internal dynamics, and (ii) "lattice" RMSD, in which each monomer was mapped back to the unit cell using the crystal symmetry operations in each frame, measuring RMSD due to fluctuations from both internal and lattice dynamics. The "lattice" method produced higher RMSDs than the "best fit" method for the cMD simulation, however, the "best fit" method for the cMD simulation produced *lower* RMSDs to the crystal structure than the solution sMD simulation, suggesting that the incorporation of lattice interactions improves agreement with the crystal structure model. The instantaneous RMSD of the sMD model also fluctuates more wildly than the cMD model. Additionally, if one computes the average protein structure from the cMD ensemble, and the average protein structure from the sMD simulation, the RMSD to the crystal structure for the latter is higher. This difference is reflective of an under-reported benefit of cMD simulations over sMD simulations: though the computational resources required to run cMD simulations are higher than those required for sMD simulations, the explicit modelling of the crystalline context leads to better agreement with the experimental data which is difficult to achieve without it. Validating force fields against solution-state simulations may lead force field developers astray, or require additional solution-state simulation corrections which are not otherwise necessary, were the researchers to attempt cMD instead.

However, when Janowski and Case *et al.* considered dynamics on the scale of individual side chains, the cMD and sMD simulations behaved more similarly: the χ_1 angle distributions (capturing rotations about the bond connecting side chains to the backbone) from sMD and cMD simulations were about the same. Where there were differences in side chain disorder, they were often found in charged or polar side chains exposed to the solvent: charged or polar interactions with neighboring proteins in the cMD simulation stabilized these residues, leading to higher disorder in the sMD simulation.

These results suggest that dynamics on the level of individual side chains, particularly those in the core of the protein, may be equally well modelled in sMD simulations as in cMD simulations. However, if one wishes to validate the dynamics of larger-scale structural features, or allostery, against crystallographic data (TLS/EN/ensemble models of Bragg data, or diffuse scattering data) single protein simulations may not be sufficient.

2.5 Calculating densities and diffuse scattering from MD trajectories

At various points above, I've mentioned direct comparisons between the densities from experiment and densities calculated from MD trajectories. Knowing the atomic form factors, it is trivial to calculate the densities from MD trajectories by inputting the coordinates and atomic form factors for each atom in the system in to the equation for the structure factor, though the software would be tedious to construct from scratch. Once the structure factors for each frame of the simulation have been calculated, the diffuse scattering can be computed with Guinier's equation. Calculation of these densities from MD trajectories is made easier through the computational crystallography toolbox (`cctbx`), an open source software suite for crystallographic data processing and analysis, which has tools for calculating structure factors, taking coordinate files as input. The most advanced version of this software comes in the form of `xtraj.py`, written by Michel E. Wall, and available in the repository for `lunus` (`https://www.github.com/mewall/lunus`), a software suite for the calculation of diffuse scattering. `xtraj.py` takes in a `GROMACS .xtc` trajectory file as input, and outputs the amplitudes and phases of the structure factors from the simulation, as well as the diffuse scattering, in an `.mtz` file format. It uses `cctbx`, the open source component of the `phenix` project for crystallographic analysis[58]. It can be run on a single core, multiple cores, or multiple nodes of a high performance computing cluster, taking advantage of the embarrass-

ingly parallel nature of calculating average structure factors to allow for efficient calculation of densities even for large supercell cMD systems with up to millions of atoms simulated for microseconds or longer. Additionally, this software can output the "F000" structure factor (the total number of electrons in the density), which allows for conversion of densities to the absolute scale (electrons per cubic angstrom), which can be useful, if not necessary, when comparing densities from different simulations.

Calculating densities in this way unlocks another powerful capability unique to cMD simulations: the separability of electron densities in to their individual components. In standard crystallographic experiments it is possible, in principle, to discriminate solvent density from protein density in high-resolution data through careful and precise modelling, but it is difficult and perhaps specious for lower resolution data. In both cases, validating that a volume of density is in fact, say, water density (rather than protein, solute, or ion density) requires a separate experimental observable (such as neutron diffraction or solid-state-NMR). For cMD density calculations, this discrimination is utterly trivial. One simply feeds in coordinates from only those component of the system from which one wishes to calculate the density. This allows densities to be calculated separately from the protein, water, ions, ligands, solute molecules — whatever one likes — which can later be superimposed on top of the crystallographic structure or the experimental density. Additionally, this software allows for the density from the entire supercell to be folded back on to the asymmetric unit, using the unit cell space group symmetry operations, averaging over not only the trajectory, but over the ensemble present in each frame.

What's more, this software is not limited to cMD simulations: it can be easily applied to solution single-protein MD simulations as well. In this case, the entire periodic bounding box of the simulation is treated as a P1 unit cell (a strange crystal, to be sure — it is difficult to image a crystal maintaining its periodicity with a large volume of solvent separating the adjacent proteins). One must be careful to rotationally and translationally align the system

so that, in each frame, the protein is superimposed on top of the crystal structure. However, densities calculated from these simulations can be compared to experimental densities in the same way as outlined above, taking advantage of the same benefit of separability into components. Preliminary tests of this capability for single protein simulations show that they are surprisingly capable at reproducing ordered water density from even short trajectories. This capability is powerful, considering the computational cost and complexity of tracking solvent density through kinematic grid-based methods. We suspect that this method of modelling ordered solvent density will be useful to many molecular dynamics practitioners, and expect to publish results from a study of this kind in the near future.

Chapter 3

Models of Diffuse Scattering Investigated Through Crystalline MD Simulations

3.1 Introduction

In Chapter 1, Sections 1 though 3, we reviewed standard crystallographic methods, which work exclusively with Bragg diffraction. The Bragg data result from average features in the density, consistent across all unit cells. It is possible to extract information about disorder from this average picture, through refinement of the Debye-Waller factor (or "B-factor"), which is related to small-scale atomic displacement and attenuates the intensity of the Bragg peaks. Anisotropic displacement parameters (ADPs) can be refined using the Translation, Libration, Screw (TLS) model, NM/EN models, ensemble models, or more fine-grained models of structural heterogeneity, providing a more detailed picture of protein structure and structural variation. However, different models of disorder or structural heterogeneity can

yield equivalent agreement with the Bragg data. Additional experimental observables are required to differentiate and (in)validate them.

In Chapter 1, Section 4, we introduced diffuse scattering, which results from dynamics, and has been tested as a means of differentiating and/or (in)validating models of correlated disorder in protein crystals. van Benschoten *et al.* (2015 and 2016)[94, 95] found that agreement with the anisotropic diffuse scattering data from CypA and trypsin was insufficient to distinguish between different Translation Libration Screw (TLS) models which agreed equally well with the Bragg data, however they were able to show that the liquid-like motions (LLM) and elastic network (EN) models provided better agreement than any of the TLS models. Peck *et al.* did similar comparisons, using agreement with the anisotropic diffuse scattering data from CypA, WrpA and alkaline phosphatase to show that the LLM model (particularly, one which allows for correlations between neighboring molecules) shows significantly higher agreement than rigid body translation or rotation models, elastic network models, or multiconformer models[74]. de Klijn *et al.*, similarly used agreement with the anisotropic diffuse scattering from CypA and hen egg-white lysozyme to compare models of correlated disorder; however, they came to a completely different conclusion: because a mixed rigid-body translation and rotation model showed roughly equivalent agreement to that of the same model with internal motions included (generated by ensemble Bragg refinement), they posited that rigid body motions are the main source of diffuse scattering[19]. Ayyer *et al.* identified diffuse scattering features in diffraction data from photosystem II consistent with rigid-body translational disorder, and were able to leverage the information contained in the diffuse scattering to extend the resolution of the diffraction beyond the Bragg limit, and perform model-free phasing of the structure factors[3]. Finally, Ando, Meisburger and Case found remarkable agreement with diffuse scattering data from triclininc lysozyme by using a combined internal and lattice-connected elastic network (EN) model — better than the agreement found using a rigid-body lattice-connected EN model[68]. To summarize, some systems and methods of analysis supporting the LLM model (which is a special case of an

EN network model[99]), and others supporting the rigid body motions model. In any case, the field seems intent on determining the "main source" of diffuse scattering.

In Chapter 1, Section 4.4, and Chapter 2, we discussed the crystalline Molecular Dynamics (cMD) model. The cMD model is the only model capable of predicting the full diffuse scattering pattern (isotropic *and* anisotropic), thanks to explicit modelling of the solvent, but it has not demonstrated better agreement with the anisotropic diffuse scattering than the LLM or EN models have. This may be due to the fact that the LLM and EN models have tune-able parameters, while the cMD model does not. The cMD model is "model free" in some sense: the parameters are fixed by the choice of force field (though one can, and some have[52], tuned the force field parameters on the fly).

For cMD simulations, force fields have been shown to have improved over time[47], with more modern force fields providing better accuracy with respect to a wide array of structural and dynamical phenomena, including agreement with crystallographic observables such as the instantaneous and average RMSD to the crystal structure and B-factors. For solution-state simulations, the same finding holds, with force fields improving over time with respect to the modelling of solvation thermodynamic measurements[60, 29]. These findings inspire confidence: after selecting a (modern, well-tested) force field, we let the model "take the wheel," so to speak, and see where it takes us. We trust, but verify, checking agreement along some measures to give us confidence in the models predictions about others. Here, I'll present a study along these very lines.

3.2 System Setup

We simulated a 2x2x2 unit cell cMD model of *Staphylococal nuclease* using the `AMBER14SB` and `CHARMM27` force fields. The crystal structure model for *Staph. nuclease* was missing

five residues on the N-terminus and eight residues on the C-terminus, which were modelled back in based on extension of existing secondary structure. The bound thymidine-3'-5'-bisphosphate (pdTp) molecule was parametrized with the SwissParam server [113] in the both simulations. The structure was propagated to the supercell in the manner described in the previous chapter, and hydrogen atoms were added using `pdb2gmx` with protonation states assigned automatically, assuming a pH of 7. Both systems were solvated with TIP3P waters using `gmx solvate`, and neutralized with chloride ions, using `gmx genion -neutral`, before minimization using the steepest descent algorithm.

As previously described, the cMD system pressure was large and negative upon equilibration after initial solvation: -1439±39 bar and -1795±252 bar for `AMBER` and `CHARMM` respectively (as reported by `GROMACS`'s `gmx energy`). The system was subjected to iterative rounds of solvation and equilibration at a temperature of 298 K, with heavy atoms strongly restrained (1000 $\text{kJ} \cdot \text{mol}^{-1} \cdot \text{nm}^{-2}$), for a duration of 100 ps for `CHARMM` and 5 ns for `AMBER` – thus the higher standard error `CHARMM above` – until the average pressure, plus or minus standard error, was within the range of -100 to 100 bar. The two systems required 17,557 and 17,138 waters, for `AMBER` and `CHARMM` respectively. The systems were both subjected to 100 ns production-equilibration with heavy atoms restrained more weakly (200 $\text{kJ} \cdot \text{mol}^{-1} \cdot \text{nm}^{-2}$) before restraints were released for 600 ns unrestrained production. Equilibration and production simulations were carried out using the leapfrog algorithm (`integrator = md` in `GROMACS`'s .mdp specification format), neighborhood searching using Verlet cutoffs, with a cutoff for the short-range neighbor list of 1.5 nm (`rlist = 1.5`) and updating every 10 frames (`niter = 10`). All bonds were constrained with the LINCS algorithm (`constraints = all-bonds`, `constraint-algorithm = lincs`). Both simulations used a timestep of 2 femtoseconds, with coordinate and data output every 2 picoseconds.

The potential energy of the system was monitored to get a sense of the convergence time for the system. The potential energy drifted for about 100 ns out of the total 600 ns, and

remained relatively stable thereafter. However, to ensure analysis was not affected by the drift in the system, analysis was carried out ignoring the first 200 ns (thus, keeping only 400 ns of trajectory for analysis). Over the course of the analyzed trajectory, the heavy-atom RMSDs of both systems to the initial crystal structure supercell increased from 2.7 to 3.1 Å and 2.6 to 2.9 Å for `AMBER` and `CHARMM` respectively. These RMSDs may seem high given the RMSDs reported in the previous chapter for similar systems, however our analysis was carried out by translationally and rotationally aligning the entire supercell, whereas previously reported RMSDs were calculated by either mapping proteins back to the asymmetric unit or unit cell before translational and rotational alignment; translationally and rotationally aligning the full supercell can lead to larger artifacts (as will be discussed in the section below regarding *Analysis of Covariance of Atom Pairs*). Additionally, the N- and C-termini of the protein, and the loop from residues 40 to 60, are all very labile. Considering that each protein started from the same initial conformation, the flexibility in these regions alone can increase RMSD across the ensemble, compared to other systems, such as lysozyme. Promisingly, the B-factors predicted by both systems were similar to each other, similar to previous simulations of the same system, and similar to experimentally refined B-factors (see Figure 7 from [100]).

We then moved to analyzing the diffuse scattering predicted by the final 10 ns of the trajectory, using the software mentioned in Chapter 2, Section 5.

3.3 Diffuse Scattering Prediction

We computed the three-dimensional diffuse scattering maps from the analyzed portion of the cMD trajectories, and compared the results to the diffuse intensity from experiment. The Pearson Correlation Coefficient (abbreviated CC hereafter) between the full diffuse maps calculated from both trajectories and full experimental diffuse map was greater than 0.9 in

both cases. The isotropic diffuse intensity (the average intensity in radial shells of reciprocal space) was subtracted from both cMD diffuse scattering maps and the experimental map, leaving the anisotropic component. The CC between simulation and experiment for the ansitropic map was 0.58 for the `AMBER` cMD simulation and 0.63 for the `CHARMM` simulation (Figure 3.1). These values were lower than the 0.68 CC previously reported for a similar simulation of *Staph. Nuclease* [100], which used the `CHARMM27` force field (and which was re-calculated against the same experimental map compared with the two simulations presented in this work, reaffirming the result). The discrepancy may be due to a variety of factors including, but not limited to, the LINCS constraints applied to all bonds here (whereas LINCS constraints were used only on hydrogen bonds in the previous simulation) and the length of the trajectory analyzed (previously 5 microseconds, versus the 0.6 microseconds analyzed here).

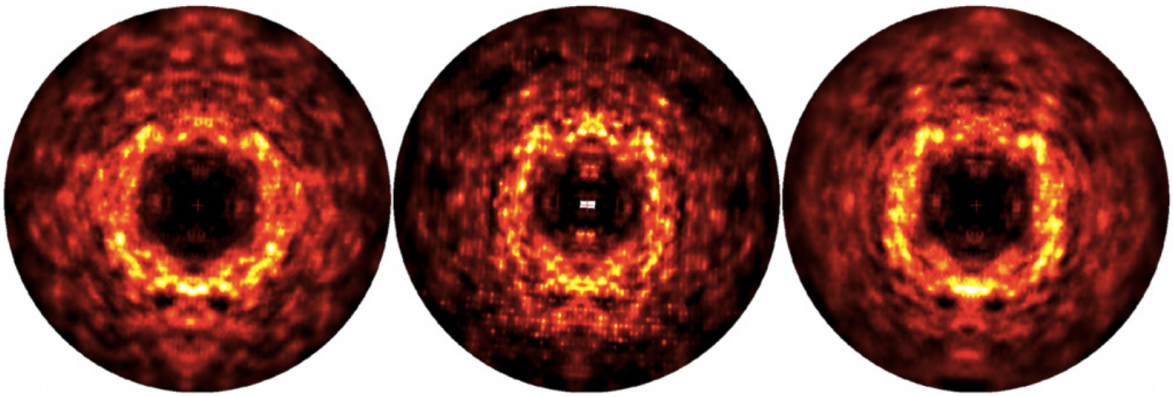


Figure 3.1: Anisotropic diffuse scattering predicted by the 200-400 ns unrestrained production segment of the `AMBER` simulation (left), from experiment (center), and predicted by the 200-400 ns unrestrained production segment of the `CHARMM` simulation (right), with sampling at each Miller index, out to a resolution of 1.8 Å. Features highlighted by the subtracting minimum intensity value in constant resolution shells of reciprocal space, prior to visualization in ADXV[2].

One hint that the discrepancy may be due to the length of the trajectory came from analysis of the accumulation of the diffuse intensity from smaller 100 ns segments of the full 400 ns trajectory. The CC between modelled and experimental structure factors were between 0.53 and 0.58 for all individual 100 ns segments using either force field. If the diffuse intensity from the 100 ns segments was accumulated "incoherently", by simply averaging the intensity

overall segments, the final intensities exhibited CCs of 0.58 and 0.63 for AMBER and CHARMM respectively (the values reported above). However, when the diffuse intensity from the 100 ns segments was accumulated "coherently", by averaging the the complex structure factors, the agreement was worse (0.54 and 0.58 for AMBER and CHARMM respectively). Indeed, the gains in agreement for the incoherently accumulated data appear as though they may continue to increase, were the simulation extended beyond the 400 ns analyzed in our study, perhaps attaining the correlation seen in previous studies of the same system, with diffuse scattering predictions accumulated from a much longer trajectory (Figure 3.2).

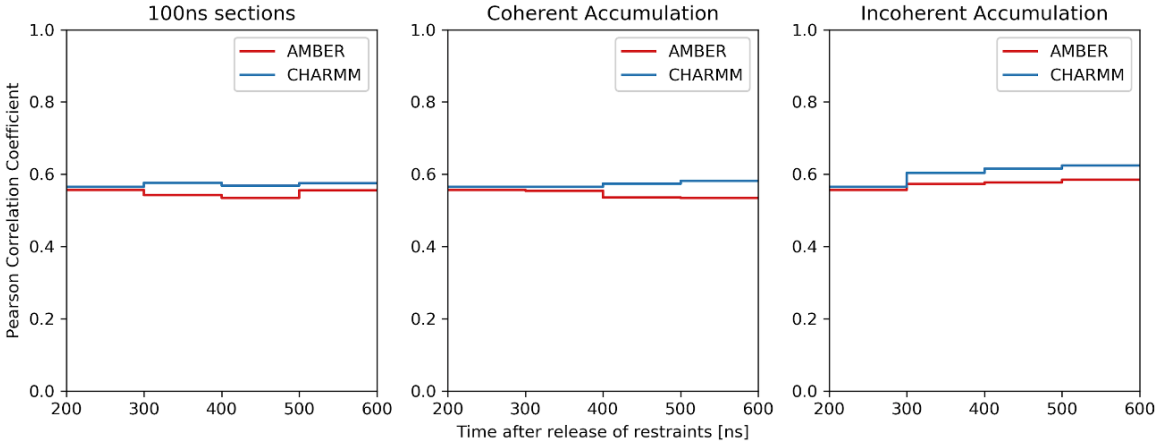


Figure 3.2: CC between experimental and simulated anisotropic diffuse scattering from 100 ns segments of 200-600ns cMD unrestrained production trajectories using AMBER (red) and CHARMM (blue): calculated in isolation (left), accumulated "coherently", by averaging complex structure factors (center), and by accumulated "incoherently", by simply averaging the intensities themselves (right).

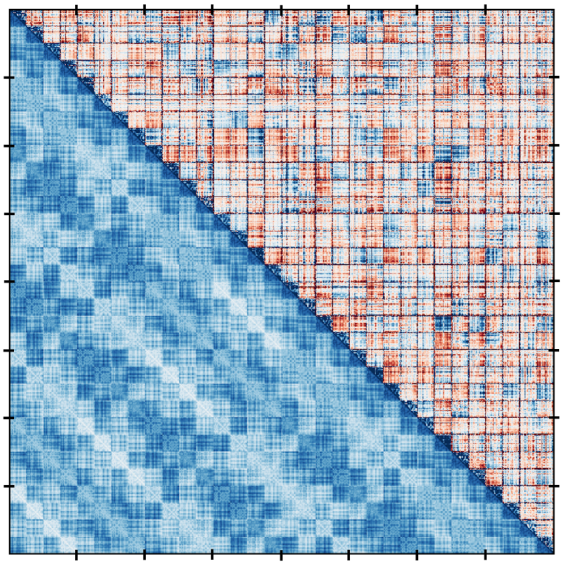
3.4 Analysis of the Covariance of Atom Pairs

In Chapter 1, Section 4.2, we introduced the Liquid-like Motions (LLM) model, which is based around the assumption that the elements of the variance covariance matrix of atom pairs decay exponentially with the distance between the atoms. Therefore, if we analyze the covariance matrix for atom pairs in our cMD system, we can analyze the degree to which the assumptions of the liquid-like motions model are supported by the the dynamics implied by

the force field. In total, our system contains on the order of 100,000 non-solvent atoms so the variance covariance matrix for all atom pairs would contain on the order of 60,000,000,000 unique elements – far too many elements to compute from a trajectory efficiently. So, here, we restricted our analysis to the variance covariance matrix for backbone carbon (C-α) atoms.

There are 149 C-α atoms in every 32 proteins in our supercell, and each pair of atoms has its own 3x3 displacement covariance matrix, leading to to a full C-α variance covariance matrix of size 14,304x14,304. However, we need a single covariance for each atom pair to analyze the assumptions of the isotropic LLM model, so, we took the trace of each 3x3 displacement covariance matrix: this is a generalization of the variance for vector-valued random variables, with negative trace when the displacements of atoms vary in opposite directions[1]. The resulting matrix, after taking the trace of each 3x3 sub-matrix, is of size 4768x4768, and is square-symmetric: the diagonal elements correspond to the the mean squared displacement for each C-α atom, and the off diagonal elements correspond to the vector-valued-"variance" (or "covariance", as I will refer to it from now on) for the atom pairs indexed by row and column. To find the relationship between the atom pair distance and the covariance of the atoms pairs, we computed

Figure 3.3: Covariance and distance matrices from cMD simulation of *Staph. nuclease*. Upper triangular elements: trace of atom pair displacement-covariance matrices, from -0.2 Å^2 (red) to 0 (white) to 0.2 Å^2 (blue). Lower triangular elements: average atom pair distance from 0 Å(dark blue) to 92 Å(white). Ticks separate unit cells. Sub-matrices along the diagonal are intraprotein distances and covariances. Elements can be matched up by mirror symmetry through the diagonal.

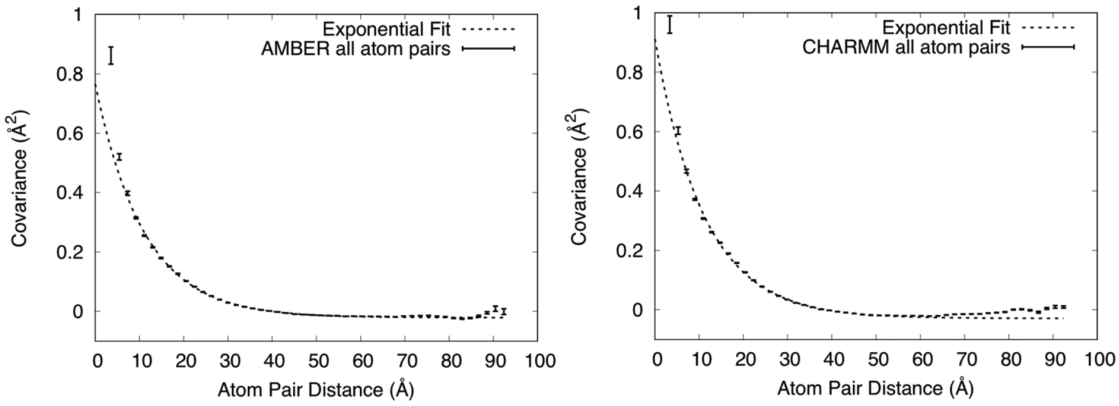


[1]We can define a variance for matrices: $\mathrm{Var_M}(\mathbf{X}) = \mathrm{E}\left[\|\mathbf{X} - \mathrm{E}[\mathbf{X}]\|_2^2\right] = \mathrm{E}[(\mathbf{X} - \mathrm{E}[\mathbf{X}]) \cdot (\mathbf{X} - \mathrm{E}[\mathbf{X}])]$ $= \mathrm{E}\left[\sum_{i=1}^{n}(X_i - \mathrm{E}[X_i])^2\right] = \sum_{i=1}^{n}\mathrm{E}\left[(X_i - \mathrm{E}[X_i])^2\right] = \sum_{i=1}^{n}\mathrm{Var}(X_i) = \sum_{i=1}^{n} C_{ii}$, where $\mathbf{C}$ is the covariance matrix, and C_{ii} are the diagonal elements. The last expression is the trace: $\mathrm{Tr}(\mathbf{C})$, thus: $\mathrm{Var_M}(\mathbf{X}) = \mathrm{Tr}(\mathbf{C})$.

a matrix of the average distance between C-α atoms from the trajectory, and compared the upper-triangular elements of each matrix, element-wise (Figure 3.3). We divided the distance between atom pairs in to to 50 bins, and found the mean and standard error of the covariance within each bin. For each simulation, for the smallest-distance bin (<5 Å) the covariance between atoms pairs was inordinately high (20 times higher then the next distance bin), most likely due to high displacement covariance between bonded atoms.

3.4.1 All atom pairs

Figure 3.4: Average atom pair covariance plus or minus standard error versus atom pair distance for *all* atom pairs in cMD simulation of *Staph. nuclease* using the `AMBER` (left) and `CHARMM` (right) force fields. Dashed lines show exponential fit to $C(r) = ae^{-r/\gamma} + b$, with γ =11.0 Åfor `AMBER` and γ =11.1 Åfor `CHARMM`



For both simulations, the relationship between covariance and atom pair separation decreases exponentially beyond 5 Å, crossing in to the negative at about 36-40 Å, and rising back above zero at about 90 and 80 Å for `AMBER` and `CHARMM`, respectively. Both plots were fit to an exponential function of the form $C(r) = ae^{-r/\gamma}+b$ where r is the separation distance between atom pairs, with the constant b added to account for the crossover in to negative covariance. The covariance versus distance data from the `AMBER` simulation yielded a correlation length of γ =11.0 $\pm$ 0.1 Å and a MSD of a = 0.79 $\pm$ 0.01 Å^2 and an offset constant of b = -0.022 $\pm$ 0.001 Å^2; the `CHARMM` simulation yielded a correlation length of γ =11.1 $\pm$ 0.2 Å and a MSD of a = 0.94 $\pm$ 0.02 Å^2 and an offset constant of b = -0.029 $\pm$ 0.001 Å^2 (Figure 3.4).

The dip in to negative covariance is most likely a result of the translational and rotational alignment before calculation of the covariance matrix: rotation about a center of mass produces negative covariance for the atoms on opposite sides of the center of rotation; the shortest supercell side length is 48.5 Å, and longest unit cell diagonal is about 94 Å, so rotation would produce negative covariances for atom pairs separated by 25-94 Å, with the negative covariances due to rotation increasing with distance (atom pairs further away from the center of rotation move more relative to each other).

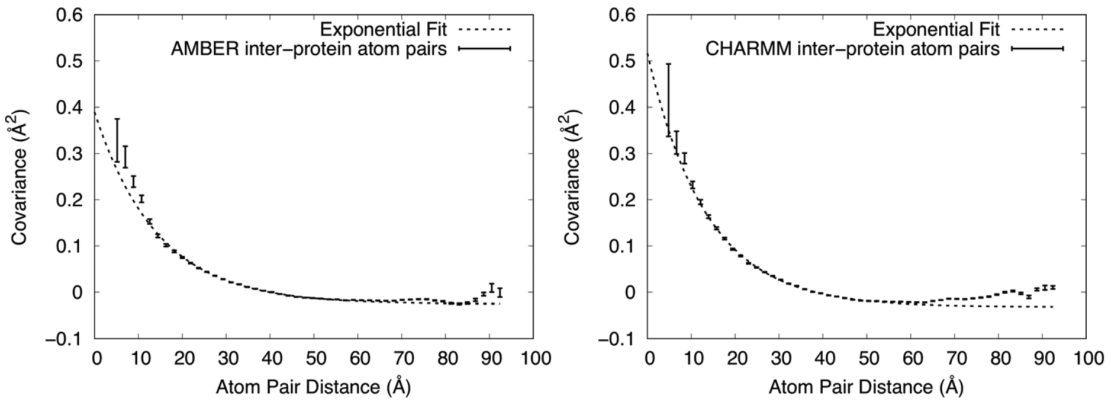
Here, we see that the liquid-like motions (LLM) model is well supported by the relationship between covariance and distance, for all pairs of C-α atoms: on average the covariance decreases exponentially as a function of the atom pair separation distance. The relationship between covariance and distance seems to be independent of the choice of force field, with the correlation length implied by both simulations equal, within uncertainty bounds provided by the fitting procedure (fit using the nonlinear least-squares Levenberg-Marquardt algorithm in `gnuplot`'s `fit` module[108]).

Next, we separated the covariance and distance matrix elements in to intraprotein atom pairs (atom pairs within proteins) and interprotein atom pairs (atom pairs across proteins), and performed a similar analysis as above for both.

3.4.2 Interprotein atom pairs

Interprotein atom pairs exhibited much the same relationship as all atom pairs for the mean covariance versus separation distance. For interprotein atom pairs, the covariance decays exponential with the separation distance with a correlation length γ = 14.3 ± 0.4 Å, MSD a = 0.42 ± 0.02 Å^2, and offset b = -0.025 ± 0.001 Å^2 for `AMBER` and γ = 13.4 ± 0.3 Å, MSD a = 0.55 ± 0.02 Å^2, and offset b = -0.032 ± 0.001 Å^2 for `CHARMM`. The correlation coefficient values for intraprotein atom pairs are higher than those for all atom pairs, for

Figure 3.5: Average atom pair covariance plus or minus standard error versus atom pair distance for *inter-protein* atom pairs in cMD simulation of *Staph. nuclease* using the `AMBER` (left) and `CHARMM` (right) force fields. Dashed lines show exponential fit to $C(r) = ae^{-r/\gamma} + b$, with γ =14.3 ± 0.4 Å for `AMBER` and γ =13.4 ± 0.3 Å for `CHARMM`



both simulations, implying correlations extend to a longer length scale. That said, there are fewer interprotein atom pairs than intraprotein atom pairs at separation distances less than about 12 Å (Figure 3.6), so this increase in correlation length may simply be a result of partitioning the data rather than a real effect. The implied MSDs are also lower for intraprotein atom pairs than all atoms pairs, perhaps for similar reasons.

3.4.3 Intraprotein atom pairs

We then moved on to analyzing the average C-α covariance plus or minus standard error versus separation distance for intraprotein atom pairs (atom pairs within proteins). Unlike the all-atom and interprotein atom pairs, the relationship between covariance and separation distance for intra protein atom pairs appeared to be roughly linear – clearly an exponential fit would be unsuccessful (again, the covariance for small-separation-distance (< 5Å) atoms was much higher than the covariance at other distances, and that data point is again disregarded in the figure below). For both simulations, the covariance between intraprotein C-α atoms starts positive and decreases roughly linearly before it crosses in to the negative at about 37 Å (Figure 3.7; note the change in axis limits for both the distance and covariance axes.)

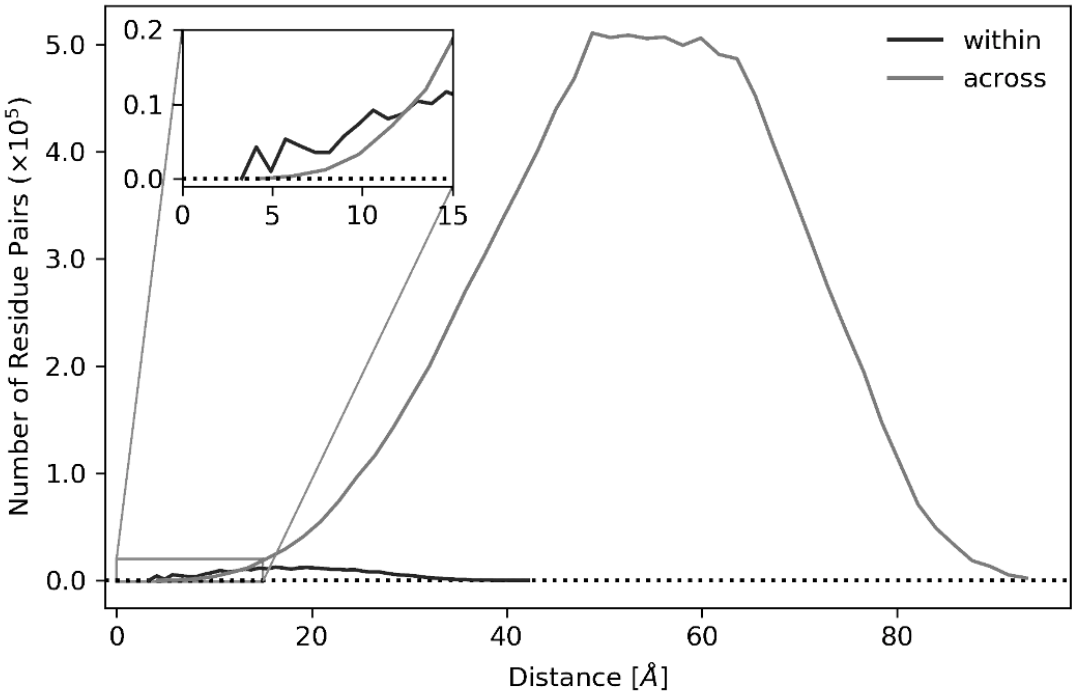


Figure 3.6: Count of the of number residue pairs as a function of average C-α separation distance for intraprotein ("within"; black) and interprotein ("across", grey) residue pairs.

A reasonable explanation for why covariance would fall off linearly and dip in to the negative so concertedly at larger atom pair separation distances can be arrived at by way of a rigid-body motions model: a protein rotating about about its center of mass will produce negative covariance for the atom pairs furthest away from each other (think of the earth: as it rotates, people in the same city will move in the same direction, but people on opposite sides of the globe from each other will be moving in the opposite direction relative to each other), and atom pairs equidistant from the center of rotation will experience equal displacements, increasing linearly as a function of the distance from the center (for small angle rotations).

To test whether a rigid body motions (RBM) model could explain this data, I wrote a piece of software (`RigidBodyMotions.py`) that would take the crystal structure as input, and generate an ensemble by shifting the coordinates through rotational or translational transformations. Rotational transformations are generated by sampling three Euler angles, each from a normal distribution with mean zero and a given standard deviation, and trans-

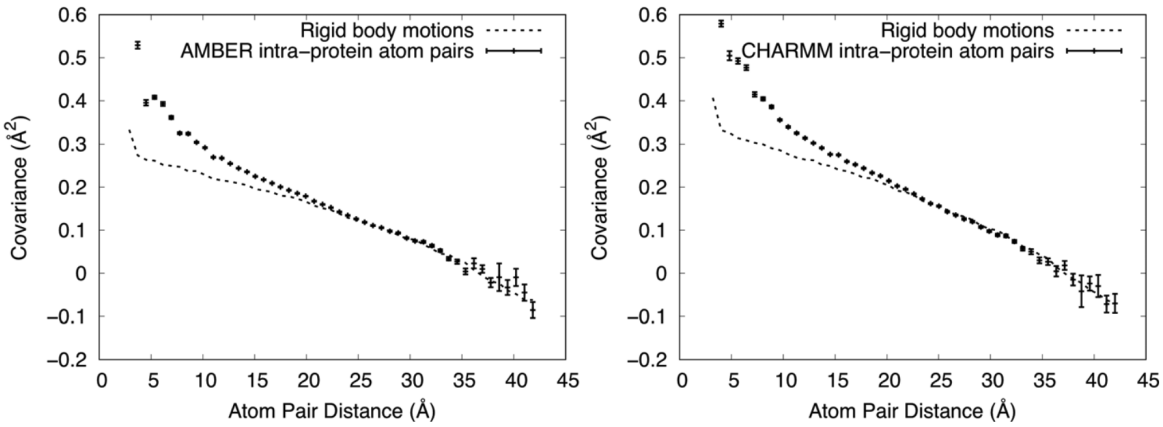


Figure 3.7: Average atom pair covariance plus or minus standard error versus atom pair distance for *intraprotein* atom pairs in cMD simulation of *Staph. nuclease* using the `AMBER` (left) and `CHARMM` (right) force fields. C-α atom pair covariance versus distance predicted by a single-protein rigid body translation and rotation model shown in the dashed line.

lations are generated by sampling a random three-dimensional vector with mean zero and a given (Normally distributed) standard deviation, consistent across each three axes – the code outputs an arbitrary number of samples, with rotational transformations performed before translational ones.

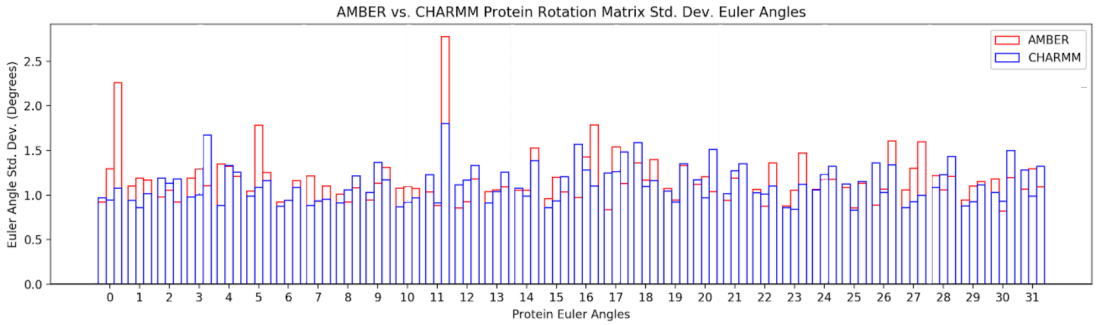


Figure 3.8: Standard deviation of the three Euler angles for the 32 proteins (indexed along the x-axis) from the 200-600ns segment of the cMD simulations of *Staph. nuclease* using the `AMBER` (red) and `CHARMM` (blue) force fields.

Initial estimates for the standard deviations were chosen based on similarly to those measured in the cMD simulation (for the Euler angles: about 1 degree, see Figure 3.8; for translations: measured on average to be $\sim$0.25 Å). The best-fit values of the standard deviations for Euler angles and translations were arrived at by visual inspection, based on agreement with the covariance versus distance data (shown in the dashed lines above in Figure 3.7). For `AMBER`

the best-fit rigid body motions model sampled Euler angles with a standard deviation of 0.95 degrees and sampled translations translations with standard deviation 0.24 Å; for CHARMM the best-fit rigid body motions model had a standard deviation of 1.05 degrees for the Euler angles sampled, and 0.27 Å for the translations. The error bars for covariance versus distance for the RBM model are too small to be visible.

The agreement of the rigid body motions model with the covariance versus distance data from both the AMBER and CHARMM simulations is excellent for atom pair separation distances above 20 Å(Figure 3.7). For distances lower than 20 Å, the discrepancies between the RBM model and cMD values decreases with distance. We calculated the residual between the cMD and RBM-modelled covariance versus distance, and found the relationship below (Figure 3.9).

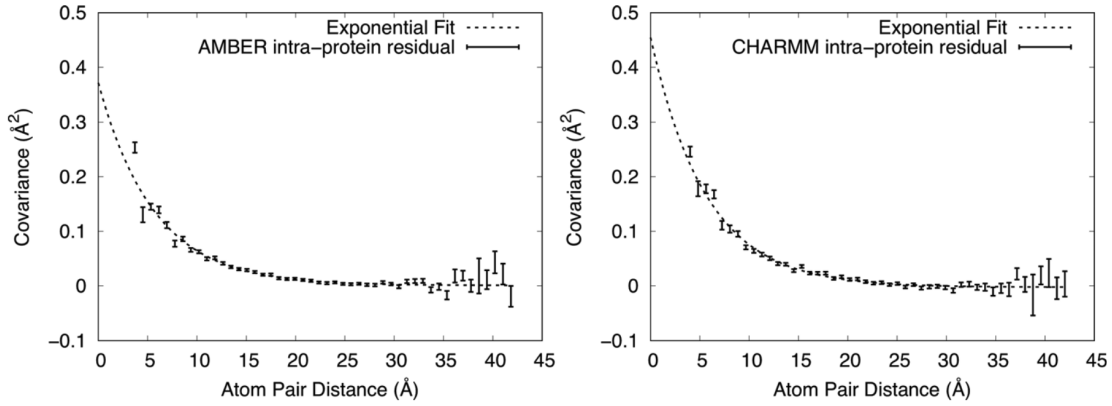


Figure 3.9: Residual between the covariance verses distance from cMD simulations of *Staph. nuclease* using the AMBER (left) and CHARMM (right) force fields, and a RBM model for each. Fit to an exponential function of the form $C(r) = ae^{-r/\gamma} + b$ shown in dashed lines.

The residual between the covariance from the cMD simulations and those predicted by the RBM model decays exponentially as a function of distance for both the AMBER and CHARMM data. As with before, the residual was fit with an exponential of the form $C(r) = ae^{-r/\gamma} + b$. The best fit exponential function had a correlation length of $\gamma = 5.7 \pm 0.2$ Å, an MSD of $a = 0.37 \pm 0.02$ Å^2 and an offset of $b = 0$ Å^2 for AMBER; for CHARMM, the best fit exponential function had a correlation length of $\gamma = 5.7 \pm 0.3$ Å, an MSD of $a = 0.46 \pm 0.03$ Å^2 and an offset of $b = 0$ Å^2.

These results show that, for C-α atoms pairs within proteins, the data from cMD simulations are well modelled by a combination of LLM-like and RBM-like covariance. The residual covariance correlation length for intraprotein atom pairs is much shorter than that of all atom pairs or interprotein atom pairs, suggesting that, once rigid body covariances are accounted for, LLM-like correlations extend over much shorter distances for individual proteins than they do over the entire system. However, beyond 12 Å the number of intraprotein C-α atom pairs is dwarfed by the number of interprotein atom pairs, and beyond about 40 Å, there are *only* interprotein atom pairs, so LLM-like exponential decay in covariance dominates the overall picture.

3.5 Discussion

3.5.1 Insights into Models of Correlated Disorder

To further understand these findings, we fit a LLM model to the diffuse scattering data from experiment (Figure 3.10). This LLM model had a correlation length of $\gamma = 6.5$ Å and an RMS displacement of $\sigma = 0.41$ Å, with a Pearson correlation coefficient (CC) to the anisotropic diffuse scattering data of 0.73 (the CC to the total diffuse scattering data is poor, but this is consistent with other LLM models, and other phenomenological models – such as the RBM and EN models – none of which model the solvent). We can convert this RMS displacement to a MSD by squaring and multiplying by three (the MSD is isotropic): $3 \times (0.41)^2 = 0.5043$ Å^2. This is similar to both the MSD suggested by analysis of the covariance of interprotein atom pairs (0.42 for `AMBER`; 0.55 for `CHARMM`) and the MSD suggested by intraprotein atom pair covariance (0.37 for `AMBER`; 0.46 for `CHARMM`), but not the MSD suggested by the covariance of all atom pairs (0.79 for `AMBER`; 0.94 for `CHARMM`). The correlation length for the LLM model refined against experimental data was 6.5 Å. This correlation length is closer to the value

for intraprotein atom pairs (5.7 Å) than it is to the correlation length of the all atom pairs (11 Å) and interprotein atom pairs (14 Å).

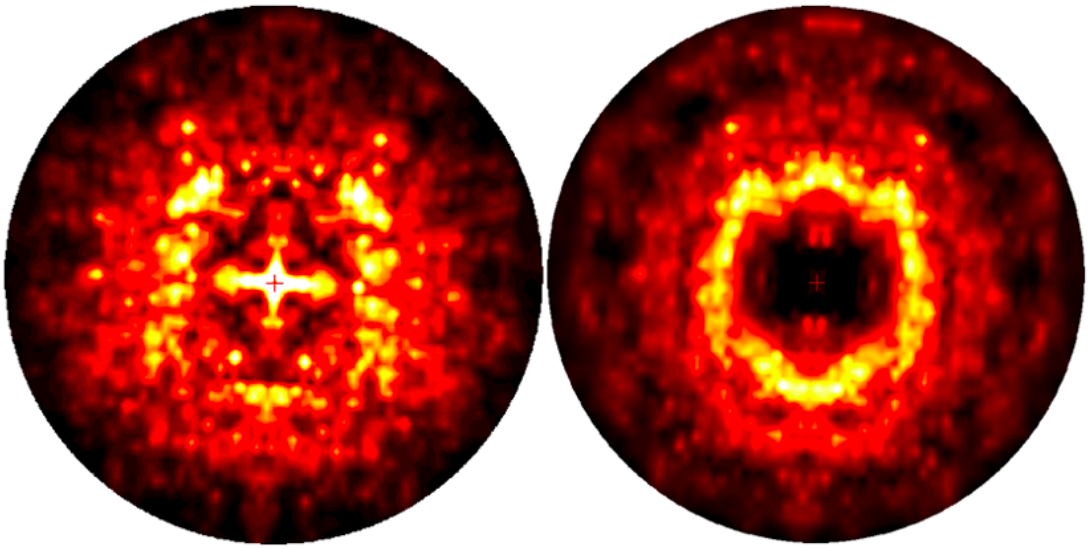


Figure 3.10: Comparison of diffuse scattering from experiment (right) and LLM model refined against the diffuse scattering (left). Features at low resolution dominate the visual comparison, but don't contribute strongly to the quantitative comparison: the Pearson correlation coefficient between the experimental and LLM-predicted anisotropic diffuse scattering is 0.73.

We also fit a simple rigid body translation model with a single displacement parameter σ to the diffuse scattering data, which yielded a refined displacement parameter of $\sigma = 0.40$ Å, and had a correlation coefficient to the anisotropic diffuse scattering data of 0.56. This value for the displacement parameter is very similar to that of the LLM model fit to the data, however the correlation coefficient is much lower. These results contrast with those of de Klijn *et al.* (2019) [19], who found that (for both CypA and Lysozyme) a good fit to the data comes in the form of a rigid body translation model, with only minimal improvements when combined with rigid-body rotational motions and/or ensemble modelling (they did not fit a LLM model to their data). In contrast to their findings, our results support the notion that rigid body motions are *not* the main source of diffuse scattering.

The correlation length and MSD from the experimentally-fit LLM model is closest to the parameters suggested by the fitting of the covariance versus distance from intraprotein atom

pairs. This diffuse scattering data is coarsely sampled (data at Miller indices) compared to the diffuse scattering from from CypA and WrpA crystals from from Peck *et al.*, which over-sampled Miller indices by a factor of three in each lattice direction[74]. The LLM models they refined against diffuse scattering data had correlation lengths of **18 Å**, and MSDs between 0.5 and 1.0 $Å^2$. These correlation lengths were larger than previous studies as well (5 Å, from calmodulin[103], 7.1 Å and 8.35 Å for CypA and trypsin, respectively[95]), however, the diffuse data in these studies were also sampled more coarsely. Our results may provide some context to the discrepancy, together with some insights from thermal diffuse scattering theory[44].

The fineness or coarseness of the sampling of diffuse scattering data may preferentially select for different types of models of disorder, and the length scales of correlations implied by them. When motions are coupled across unit cell boundaries, the diffuse intensity resulting from these motions appears closer to the Bragg peaks[74], sometimes directly on top of them. Thermal diffuse scattering theory suggests that the intensity resulting from thermal motions has local minima at Bragg peaks, which decay with distance from the Bragg peaks: motions coupled across unit cells produce features which decay quickly with the distance from the Bragg peak (as in the power law fit from Ando, Meisburger and Case (2020)[68]), whereas motions coupled on shorter length scales produce diffuse features further from the Bragg peaks, sometimes spreading across large regions of reciprocal space. Because of the coarse sampling in our study, there is a larger proportion of data points far from the Bragg peaks, whereas more finely sampled data would have a larger proportion of data closer to the Bragg peaks. For future studies (particularly those with finely sampled data) it will be important to test for differences in model fit when refined against down-sampled data. There are only a handful of well-isolated and processed diffuse scattering data sets at high resolution, so it is important to study the effect of down-sampling on model fit, so researchers working with lower resolution data have proper context with which to ground their findings.

3.5.2 Force Field Differences

The cMD simulations using the AMBER and CHARMM force fields produced very similar results in their predictions of the covariance of atom pairs. The correlation lengths implied by the exponential fit to the covariance and distance data for the simulations using both forcefields were equal to within uncertainty for all atom pairs and intraprotein atom pairs, but were not equal to within standard error for the interprotein atom pairs; the MSDs implied by the same analysis were not consistent to within uncertainty across forcefields for any set of atom pairs. The measured standard deviation of Euler angles for the rotations of individual proteins were similar between forcefields, as were the standard deviations of Euler angles for rotation and translational perturbations implied by the fitting of the covariance of intraprotein atom pairs with a rigid body model.

There were some other notable differences between the AMBER and CHARMM simulations: for all atom pairs and interprotein atoms pairs, the atom pair covariance from the AMBER simulation stays negative to a separation of 90 Å, while that of the CHARMM comes out of the negative at about 80 Å. The short distance covariance behavior (out of range in the y-axis in the plots above) was also different between the simulations: for interprotein atom pairs, the covariance between atom pairs remains exponential at short distance for the CHARMM simulation, but dips in to the negative for the AMBER simulation for the bin from 0 to 3.31 Å.

Finally, and perhaps most importantly, the diffuse scattering predicted by the simulation using the CHARMM force field had a higher correlation to the experimental data than the simulation using the AMBER force field[2]. The origin of these differences is not immediately clear. However, a few observations point toward the modelling of crystal contacts (in line

[2]I'll note here, though it applies to comparisons with the LLM model as well, that correlation is not a linear descriptor: a correlation of 0.25 is not half as good as a correlation of 0.5. In fact, multivariate Gaussian distributions with correlation 0.25 and 0.5 between the random variables look quite similar. So, small increases in correlation may not be all that meaningful. The field requires new metrics to better quantify the agreement with experimental data, e.g. some form of the mutual information. However, as things stand, the Pearson correlation coefficient is the standard for the field, so it is what we use here.

with the discussion from Case *et al.* from [10]): there are only few interprotein atom pairs at short distances (Figure 3.6), but there is a large difference between the two simulations in the covariance of interprotein atom pairs at short distances. The measured standard deviation of rotational Euler angles in the `AMBER` simulation is slightly higher than the same from the `CHARMM` simulation for almost all proteins. These results together suggest that the individual proteins in the `AMBER` simulation experience slightly higher rotational displacements from the crystal structure position, and that there is a significant difference in correlated dynamics for the closest interprotein atoms pairs (which are likely to be crystal contact residues).

3.5.3 Additional Notes and Caveats

As mentioned in the beginning of this chapter, analysis of the covariance of all atom pairs in the simulation would have been too computationally expensive to compute, so we restricted our analysis to the covariance of C-α atom pairs. For all C-α covariance data except the interprotein atoms pairs from the `AMBER` simulation, the covariance for short-separation-distance (< 5 Å) atoms pairs is much higher than the covariance for all other atom pairs at higher separation distances. This finding is not too surprising, as C-α atom pairs at short distances are often connected by bonds (and thus, their displacements would be highly correlated) whereas residues with C-α atoms separated by larger distances may only be able to interact with each other through longer range forces (charged or van der Waals interactions), if they interact directly at all. Additionally, both cMD simulations used the linear constraints solver (LINCS) to constrain the distances between all bonded atoms, which increases the covariance between bonded atoms even further. Additionally, if the C-α/backbone atoms are generally more rigid than side-chain atoms, restricting our analysis to C-α atom pairs may bias the analysis toward a rigid body motions model.

The exponential fits for all atom pairs and interprotein atoms pairs have a negative offset

($b < 0$), meaning the correlation between atom pairs does not converge to zero as the separation distance increases. Due to periodic boundary conditions, the longest possible distance between atom pairs is half the longest box side length, or one lattice vector, along each crystal axis. This offset may be the consequence of the translational and rotational alignment of the supercell performed before analysis (as suggested in [67]), or it may be due to low-frequency phonon-like crystal vibrations (which have been observed to contribute to diffuse scattering by Polikanov and Moore[78]), or it may be some other real effect. As these correlations are long-distance, however, if it were a real effect, the diffuse scattering it contributes to would be focused very close to the Bragg peaks, and might be filtered out along the with Bragg scattering, if filtering techniques are not sufficiently careful.

In 2015, Meinhold and Smith performed a similar analysis [65] on the same system, *Staph. nuclease*. They analyzed the *correlation* matrix, rather than the covariance matrix (where the former is the latter, with elements divided by the geometric mean of the variances for both atoms). They found that the correlation for intraprotein and interprotein atom pairs both decrease exponentially with separation distance, with a correlation length of 11.0 Å and 11-18 Å, respectively. The exponential fit for their intraprotein atom pairs was quite good, in disagreement with our findings here. However, they only simulated a single unit cell (whereas we simulated a 2x2x2 unit cell supercell) for 10 ns (whereas we simulated for 400 ns), and they performed their simulations at constant pressure, allowing the periodic box sides to fluctuate in length (whereas we simulated in the NVT ensemble, strictly enforcing the crystal symmetry). They do not report the RMSD of their simulated protein coordinates to the crystal structure, and though they found that the R-factor to the experimental data improves with simulation time, they also estimated in an earlier paper that the covariance matrix would take on the order of 1 microsecond to converge [67].

With regard to the diffuse scattering, we found that the agreement with the diffuse scattering data is dependent on the method used to accumulate data from smaller segments of

the trajectory (computational time allocation restraints on clusters makes it more efficient to simulate longer trajectories in smaller chunks, which can be easily extended at a later time). In particular, though Guinier's equation (strictly applied) dictates that "coherent" accumulation (in which the complex structure factors are averaged when adding new data) should be used, we found that better agreement is found by "incoherently" accumulating data (simply averaging the diffuse intensity from the segments). Although the difference in correlation is small, it may be a sign of a real effect: it may be the case that a real crystal is better described as an average over many independent domains than a single crystal undergoing dynamics as a single coherent system over long time scales. Atomic force microscopy experiments have detected a relatively high concentration of crystal defects, in studies of macromolecular crystals[61], in support of this suggestion (and the "mosaic block" model used widely in crystallography[89]).

Finally, though the LLM model outperformed the cMD model with respect to the correlation to anisotropic diffuse data, it is worth remembering that the LLM model is not capable of predicting the *full* diffuse scattering data, which the cMD system models quite well (CC $>$ 0.9). Additionally, the LLM model is fit to the data, whereas the cMD model is "model free" in that the parameters are fixed after the force field is chosen. The cMD model may improve if its parameters were allowed to be refined against the data, however, though it has been attempted before, as things stand currently, this task is expected to be incredibly computationally inefficient for a 2x2x2 supercell. Improvement of forcefield parameters is more likely to be feasible through analysis of the Bragg data (density and difference maps) than the diffuse scattering – a project we have been working toward, and which we will discuss in the next chapter.

Chapter 4

Lessons in the modelling of Protein Structure and Disorder and Ordered Water from Crystalline MD Density Analysis

4.1 Introduction

In the previous Chapter, we discussed some insights on the modelling of correlated dynamics and diffuse scattering gained by crystalline Molecular Dynamics (cMD) modelling of a 2x2x2 unit cell system of *Staphylococcal nuclease*. We saw that cMD models can perform very well at modelling the total diffuse scattering signal, perform reasonably well at modelling the anisotropic diffuse scattering, and provide insights on the character of correlated disorder for atom pairs both within and across proteins. However, there is ample room for improvement. For example, a simple, one parameter, liquid-like motions model produces predictions for

the anisotropic component of the diffuse scattering which are better correlated with the experimental data than those from the cMD model (however, the LLM model has parameters fitted to the data, while the parameters of the cMD model are fixed). Improvement of the cMD model will be dependent on improvements in force field parameterization, or on changes to the simulation regime (such as the inclusion of polarizability or quantum mechanics).

Molecular dynamics force field parameters can be improved by a variety of means. Amino acid and water model parameters have been validated or found lacking based on agreement with measured solvation- or hydration-related thermodynamic quantities[38, 87, 29], with water models being particularly important in determining the agreement with measured quantities. However, the accuracy of the force fields in these studies may be system-dependent, with force fields demonstrating agreement in certain domains (e.g. fully solvated molecules) untested in other domains (e.g. coordinated water network interactions, membrane-bound proteins, binding pockets, etc.).

A systematic study of protein force fields showed that modern force-fields are generally quite accurate in reproducing NMR data and melting curves[60]. This study also showed that force fields have improved over time, but force field families differ in their modelling of the formation and stability of secondary structure elements. A systematic cMD study mentioned in a previous chapter has also shown that force fields have improved over time, with respect to a completely different set of experimental observables (instantaneous and average RMSDs to the crystal structure and agreement with experimentally-refined B-factors)[47].

However, accurate prediction of more specific, complex phenomena such as the free energy of protein-protein association [73], and the structural dynamics of intrinsically disordered proteins [83] requires researchers to produce ad-hoc variants of the gold standard force fields. cMD simulations provide a unique model system in which to (in)validate force field parameters at a much finer level of detail, for direct comparison to experimental data. cMD simulations can also provide information about structural dynamics of side chains and al-

lostery, and coordinated water networks, allowing for the careful study of both amino acid and water model parameterization effects in the same system.

Here, we present preliminary results of a cMD study of a 2x2x2 unit cell cMD model of the catalytic domain of *protein kinase A* (PKA). PKA is a cyclic-AMP (cAMP)-dependent kinase present in the cells of the liver, kidney, nervous system, cardiovascular system, and muscular system. In cells, inactive PKA exists as a tetrameric holoenzyme — a complex of two heterodimers both containing a catalytic subunit (or "C subunit") and a regulatory subunit (the "R subunit"). The R subunit binds the C subunit at the active site cleft, blocking access to the active site. Cyclic AMP binds to the R subunit, causing the R subunit to dislodge from the cleft, exposing the active site for phosphorylation of other peptides. Once phosphorylation takes place, any number of phosphodiesterases convert the cAMP bound to the R subunit to AMP, restoring the R subunit to its original configuration. In the nervous system, an endogenous inhibitor, PKI, modulates the function of the C subunit in place of the R subunit. PKI binds potently and specifically to the free C subunit and helps in its transportation from the nucleus. PKI is involved in morphogenesis, neuronal synaptic activity and gene expression[18]. The bound peptide in the structure used in our study is part of the PKI molecule.

Three motifs are important to the activation of PKA: (i) the glycine rich loop (this region is often called the "P loop" in GTPases, however our collaborators refer to it as the "G loop", so we will use that nomenclature here) which packs over the active site, and helps to keep the ATP molecule in place, and stabilize the transition state in catalysis; (ii) the YRD motif, with the "D", or aspartic acid, residue being the catalytic base for the hydrolysis of ATP to ADP; (iii) the DFG motif, a regulatory loop which moves in and out from the active site as a consequence of inhibitor binding[13].

PKA is, in general, modulated by phosphorylation of the active site loop, but is constitutively active coming out of the ribosome, regulated by the presence of the R subunit. The "R-spine"

and "C-spine" are a series of stacked hydrophobic residues which run from the N-terminus to C-terminus, along the regulatory and catalytic subunits, respectively, whose conformations (stacked or dispersed) are commonly used as a sign of activity (tightly stacked spines are associated with activity)[93].

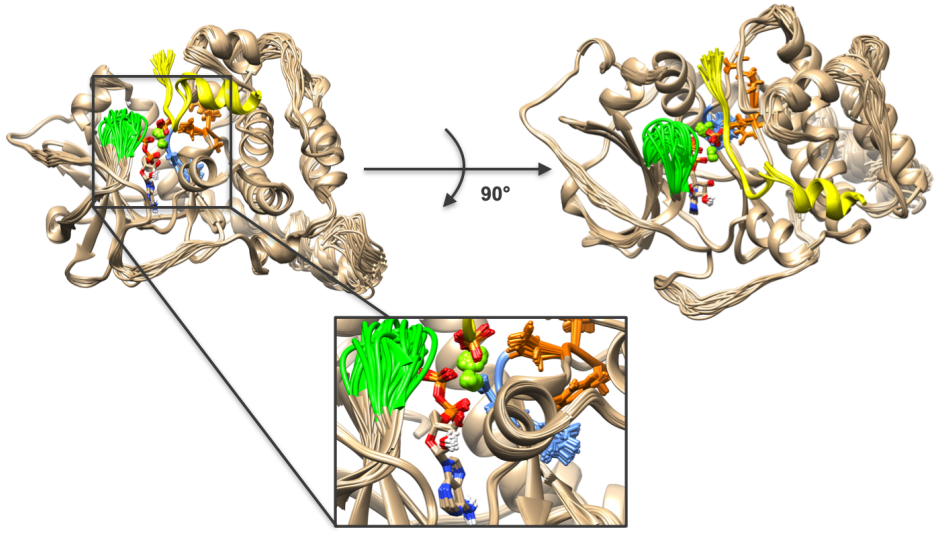


Figure 4.1: A PKA ensemble (refined by collaborators) with bound peptide in yellow, G loop in green, YRD motif in orange, and DFG motif in blue. The inlay shows active site with bound ADP (bottom), magnesium ions and phosphate molecule (top): aspartic acid (Asp; D) 166 at the end of the YRD motif shown coordinated with phosphate and magnesium ions; Aspartic acid 184 from DFG motif obscured by magnesium ions. Molecular graphics produced with UCSF Chimera[77], developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco, with support from NIH P41-GM103311.

PKA exhibits substrate binding cooperativity, with the binding of ATP leading to increased propensity for association with peptide substrate. It has been suggested that ordered water plays an important role in the allosteric pathway responsible for this effect. Two hydration sites seem to play an important role: one forming a hydrogen bond bridge between the DFG and YRD motifs, and another smoothing out the free energy landscape between the active and inactive conformations of the G loop[85]. It has also been proposed that ordered waters play in important role (along with the bound magnesium ions) in the destabilization of the transition state in catalysis, through QM/MM and MD simulations of the catalytic subunit[13].

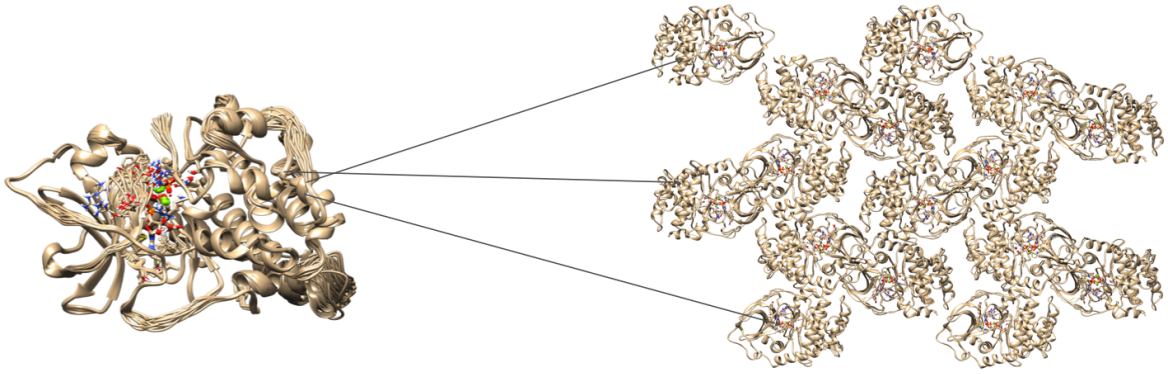


Figure 4.2: Each member of the experimentally-refined ensemble is propagated to a different place in the supercell, using the side lengths of the unit cell and the space group symmetry operations. Each frame of the simulation can be reverse-propagated, using the reverse operations, to create a dynamically changing representation of the ensemble from the supercell model.

4.2 System Setup

4.2.1 Ensemble Model

In this work, we studied only the catalytic subunit of PKA, with a peptide bound to the active site cleft, which mimics the interaction with the R subunit. Our 2x2x2 unit cell system (space group P212121, four proteins per unit cell, for thirty two proteins total) was seeded with a thirty two member ensemble model of PKA refined by collaborators against 2.4 Å X-ray diffraction data (Figure 4.2). The core of the protein was refined against the experimental structure factors, and the N- and C-termini were added after the fact, and energy minimized, to produce a full structural model, which was fed in to `phenix.ensemble_refinement`[7] to produce the ensemble model. Each ensemble member has a bound peptide whose coordinates are also refined as part of the ensemble. Each member of the ensemble has a bound adenosine diphosphate (ADP) molecule, phosphate molecule ($PO_4H_2^{2-}$ – with protonation state

determined by the pH, in the range from 5.8 to 6.7), and two magnesium ions, all placed in identical positions in the active site, while the ensemble protein atoms have different coordinates.

After removing crystallographic waters, each member of the ensemble was sent to a different place in the supercell using my `python` software module `pdbio.py` (a custom, minimalist software program design to read, manipulate, and write PDB files) and propagation code similar to the pseudocode outlined in Chapter 2 (`propagate.py`, in the Appendix). The same code was used in reverse (`reverse_propagate.py`, also in the Appendix), for analysis, sending each protein in the supercell back to the asymmetric unit. These two scripts together allowed us to measure structural and dynamical changes to the protein from an *ensemble* perspective (Figures 4.2 and 4.3).

The protein and magnesium atoms were parametrized using the `AMBER14SB` force field, using `tleap` from the `ambertools`[8] software suite (version 20.9), using the `phosaa10` parameter set for the phosphorylated serine (SEP) and tyrosine (TPO) residues. The bound ADP molecule was parametrized based on the work of Meagher *et al.* (2003)[64]. The geometry of the phosphate molecule was determined by `mp2/aug-cc-pvdz` QM geometry optimization, before being parametrized with the `gaff2` force field.

In the first round of simulations, all histidines were doubly protonated (residue name HIP), whereas in all later simulations, the protonation states were assigned based on data from neutron diffraction, collected by collaborators. Histidines 87, 158, and 260 remain doubly protonated (HIP), whereas residues 62, 131, 142, and 294 are singly protonated on the epsilon nitrogen (HIE), and histidine 68 is protonated on the delta nitrogen (HID). A later section will discuss the differences in conformation observed in the cMD systems when switching from double protonated to singly protonated forms, but it is sufficient to say for now that in some cases, a switch in protonation state in the cMD simulation can (in)validate a protonation state model when compared to experimental density.

The fully parametrized supercell structure was solvated using GROMACS's [5] `solvate` program, made neutral in charge by the substitution of water molecules with the appropriate number of chloride ions using GROMACS's `genion` with the flag `-neutral`, and additional water molecules were substituted with choride and magnesium ions sufficient to mimic the crystallization buffer $MgCl_2$ concentration of 0.05 M. The TIP3P parameter set was used for all simulation waters[48].

In a procedure similar to that outlined in the previous chapter, the system was subjected to iterative rounds of solvation, energy minimization (using the steepest-descent algorithm), and NVT equilibration (2 ns total, with harmonic restraints to the crystal structure positions with a force constant of 1000 $kJ \cdot mol^{-1} \cdot nm^{-2}$), until the average pressure, plus or minus standard error, was brought within the range from -100 to 100 bar (as close to 1 bar as possible). The system was simulated for 10 ns with weaker restraints (to be discussed below), to ensure that the system was fully equilibrated under the new restraint regime, before starting 100 ns of restrained "production" simulation.

For the initial system, in which histidines were all modelled as doubly protonated, the force constant used for harmonic restraints was 200 $kJ \cdot mol^{-1} \cdot nm^{-2}$. For the other systems, with histidine protonation states assigned based on the data from neutron diffraction, four separate simulations were prepared, with the only difference between them being the strength of the harmonic restraints: the four simulations had force constants of 200, 20, 2, and 0.2 $kJ \cdot mol^{-1} \cdot nm^{-2}$, respectively. These simulations were meant to test the effect of different heavy-atom harmonic restraint strengths on the structure, dynamics and ordered solvent modelling in cMD systems (Figure 4.3).

All simulations used a time step of 2 femtoseconds, with coordinate output every 2 picoseconds. Other specifications for the simulation parameters are detailed in the GROMACS `.mdp` files in the Appendix. For all analysis below, the final 10 ns were used for density-based analysis, and final frames were used for coordinate-based analysis.

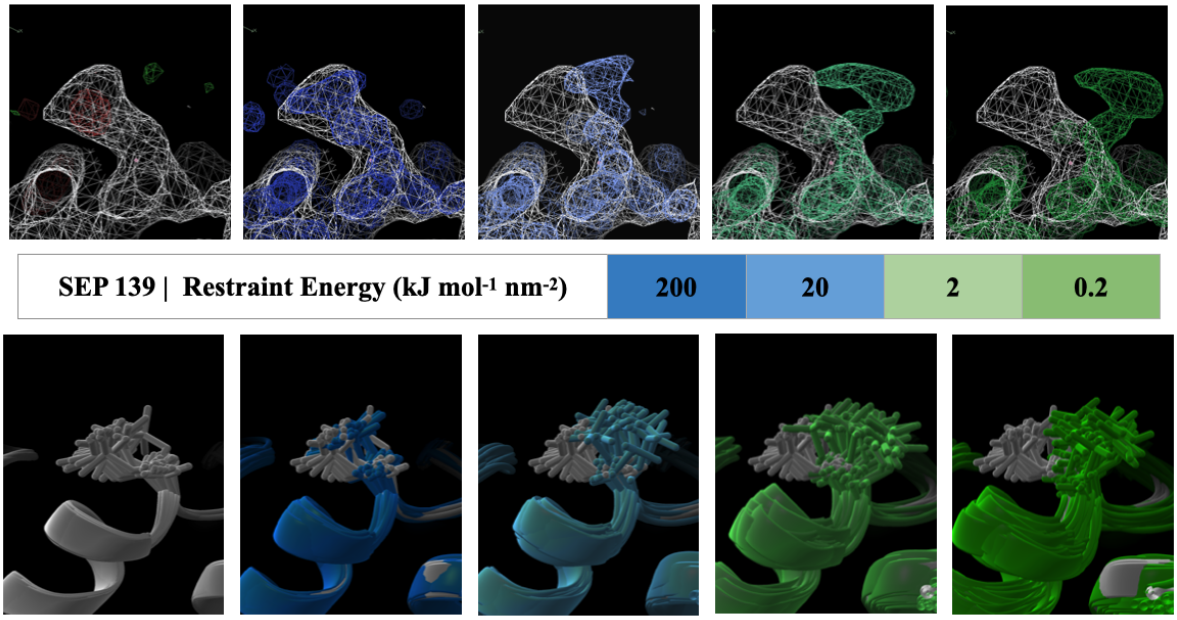


Figure 4.3: Phosphorylated serine (SEP), residue 139, from *protein kinase A*, from cMD simulations with various force constants for harmonic restraints. Top: average densities calculated using `xtraj.py`. Bottom: ensembles from the final frame of each simulation generated using `reverse_propagate.py`. Starting ensemble and density in white for both top and bottom. SEP 139 shows a concerted change in ensemble conformation as the restraints are relaxed.

4.3 Density-based Analysis

Densities were calculated from the final 10 ns of the production trajectories, down-sampling to frames every 10 ps. Each frame was aligned to the initial propagated-crystal-structure supercell using translational fitting (`gmx trjconv -fit translation`), with molecules kept whole by mapping coordinates back across the box using periodic boundary conditions (`gmx trjconv -pbc mol`), before feeding the trajectory and initial propagated-crystal-structure supercell in to `xtraj.py` to produce structure factor and intensity `.mtz` files (`fcalc.mtz` and `icalc.mtz`, respectively).

Densities were calculated separately for the full system, water molecules, magnesium atom, and calcium atoms, for all trajectories. The `fft` method from `CCP4` (Collaborative Computational Project Number 4, version 7.1 [109]) was used (along with the electron count, or "F_{000}", and volume information outputted by `xtraj.py`) to calculate densities on an abso-

lute scale (electrons per cubic angstrom): this allows us to make more precise comparisons between simulations, which would not be possible using relative density cutoffs (such as the standard 3σ cutoff for visualizing solvent density). These files were converted to more standard, diffraction-like intensity `.mtz` (`Iobs` and `SIGIobs` columns of a `.ccp4` map) files using the `sftools` method from `CCP4`.

The calculation of separate densities for the full system, water molecules, magnesium ions, and calcium ions allowed us to discriminate between protein and solvent density (which is not possible using standard crystallographic techniques, except for proteins with little structural flexibility or heterogeneity at high resolution). The `peakmax` method from `CCP4` was was used to generate a `.pdb` file with water molecules at the positions determined by peaks in the solvent map. This allowed us to refine the protein and solvent coordinates separately, combining them for a full structural model. To help in validating solvent density from the cMD simulations we also constructed "Polder maps", which are a special case of OMIT map constructed by excluding certain solvent atoms from refinement and eliminating bulk solvent corrections[59], to identify weak solvent density that is supported by experiment.

4.4 Discussion

4.4.1 B-factor Analysis

The initial experimentally-refined structural model used to seed the ensemble refinement was used a target for B-factor analysis. B-factors were calculated from the cMD simulations by refining a single structure against the intensities calculated from the cMD trajectories, using `phenix.refine`. Figure 4.4 shows the results for all heavy atoms. Unsurprisingly, for the most strongly restrained system, the B-factors are drastically underestimated. However, as the restraints are relaxed, the B-factors approach the experimental values more and more

closely, with some B-factors beginning to be overestimated. Figure 4.5 shows a comparison of the cMD and experimental B-factors for each restraint force constant simulation, with experimental B-factors on the abscissa and cMD B-factors on the ordinate.

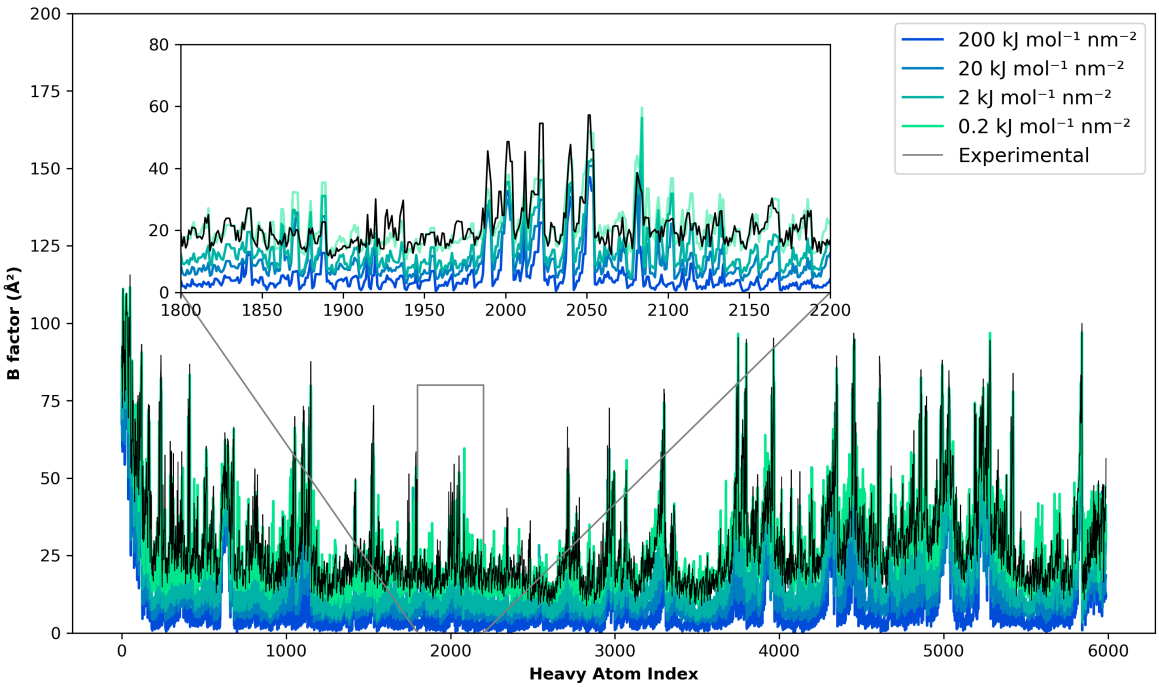


Figure 4.4: B-factors from refinement of crystal structure model in to density calculated from the final 10ns of cMD simulations with harmonic restraint energy force constant of 200 (blue), 20 (light blue), 2 (turquoise), and 0.2 (green) $\text{kJ} \cdot \text{mol}^{-1} \cdot \text{nm}^{-2}$. B-factors refined against experimental data in black. Inset showing B-factors for cMD-refined heavy atoms from index 1800-2200: the overall pattern of B-factors is reproduced even at high restraint energy, however the predictions of the B-factor values themselves improves as restraint force constants are diminished.

The Pearson correlation coefficient between the experimentally-refined and cMD-density-refined B-factors is high (CC=0.94) and is constant across all simulations, suggesting that the overall pattern for the B-factors is largely preserved, even at high restraint energy (see the legends in in Figure 4.5). However, as the restraints are relaxed, the root mean squared error (RMSE) between the experimental and cMD-density-refined B-factors decreases, from 18.8 Å^2 for the 200 $\text{kJ} \cdot \text{mol}^{-1} \cdot \text{nm}^{-2}$ restraint force constant simulation to 5.6 Å^2 for the 0.2 $\text{kJ} \cdot \text{mol}^{-1} \cdot \text{nm}^{-2}$ restraint force constant simulation. This indicates that the *values* for the B-factors predicted by the simulation get more accurate as the restraints are relaxed.

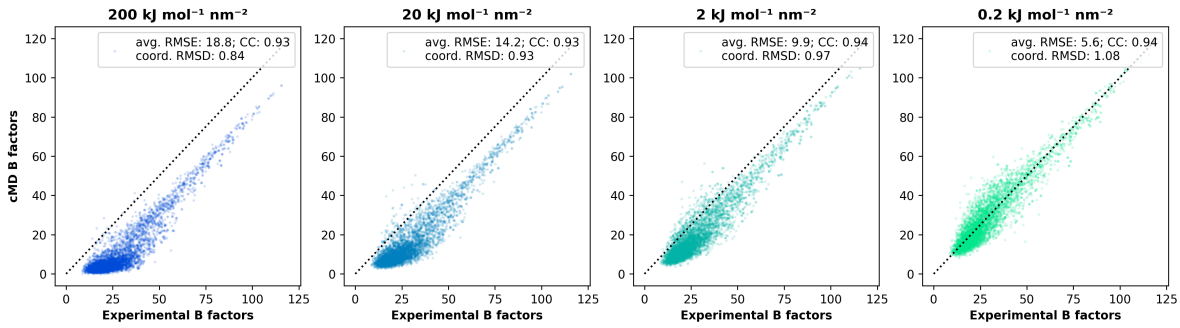


Figure 4.5: Comparison between experimental (abscissa) and cMD (ordinate) B-factors for simulations with restraint energy force constant of 200 (left), 20 (center-left), 2 (center-right), and 0.2 (right) $\mathrm{kJ \cdot mol^{-1} \cdot nm^{-2}}$. Correlation coefficient between experimentally-refined and cMD-density-refined B-factors is 0.93-0.94 for all simulations, whereas the RMSE decreases from 18.8 Å to 5.6 Å going from the 200 to 0.2 $\mathrm{kJ \cdot mol^{-1} \cdot nm^{-2}}$ restraint force constant simulations.

The heavy-atom RMSD between the average coordinates of the cMD-density-refined structure and the experimental crystal structure increases, from 0.84 Å for the 200 $\mathrm{kJ \cdot mol^{-1} \cdot nm^{-2}}$ restraint force constant simulation to 1.08 Å for the 0.2 $\mathrm{kJ \cdot mol^{-1} \cdot nm^{-2}}$ restraint force constant simulation (Figure 4.5). All together, these results suggests that although the cMD simulations may be converging to a slightly different structure than the structure refined against experimental data (as evidences by the RMSD), the cMD simulations are faithfully reproducing the overall *character* of small-scale harmonic disorder for the heavy atoms, even at high restraint energies (with high B-factor atoms from experimental refinement corresponding to relatively high B-factor atoms in the cMD-density refined structure, even at high restraint energy, and *vice versa* for low B-factor atoms) though as the restraints are relaxed, the predicted *values* for the B-factors improve.

4.4.2 Structural Deviation

Next, we compared the atomic coordinates of the models refined against the structure factors from the cMD trajectories to the atomic coordinates from the model refined against experimental data. Do to this, we calculated a histrogram of the heavy atom displacements between the experimentally-refined and cMD-density-refined structures. Unsurprisingly, the

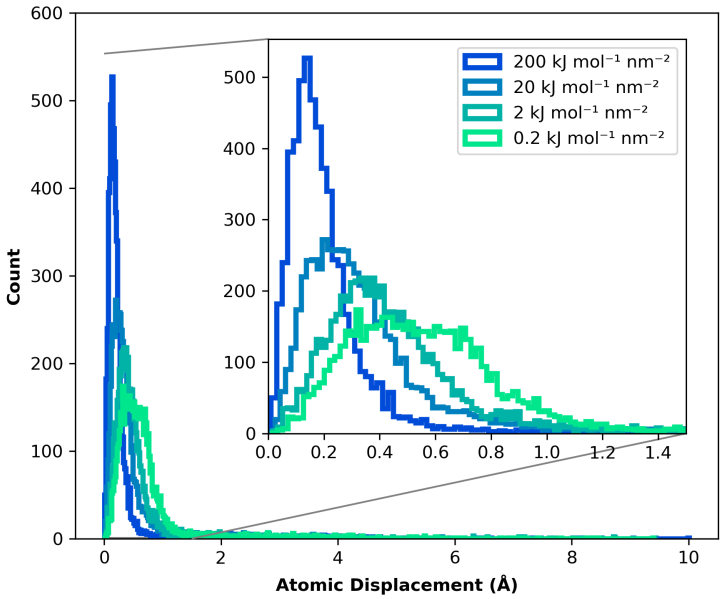


Figure 4.6: Histogram of heavy-atom atomic displacements compared to the experimentally-refined crystal structure, for the models refined against the structure factors from the 200 (blue), 20 (light blue), 2 (turquoise), and 0.2 (green) $\mathrm{kJ \cdot mol^{-1} \cdot nm^{-2}}$ restraint force constant cMD simulations.

coordinates from the model refined against the density from the most strongly restrained simulation exhibited the lowest atomic displacements to the crystal structure coordinates – the strong restraints are keeping the displacements low. As the restraint force constants are reduced, the distribution of atomic displacement shifts toward having a larger mean and a broader width (Figure 4.6). The distributions of atomic displacements appear surprisingly well-converged and smoothly varied with the relaxation of the restraints, considering that side chain atoms make up most of the atoms in the system and their displacements at low restraint energies are influenced not only by thermal excitation but by complex interactions with neighboring side chains and the solvent.

To better understand the change in the distribution of atomic displacements with the relaxation of restraints, we converted the histrograms of atomic displacements to probability density functions, considering the atomic displacement not as an experimental observable, but as a random variable sampled by the simulation. The probability density functions were fit to a Weibull distribution $\left(\frac{k}{\lambda}\right)\left(\frac{x}{\lambda}\right)e^{-(x/\lambda)^k}$ with shape parameter k and scale parameter λ. The fit was optimized using the `optimize.curve_fit` function from the open source scientific Python package `scipy`[98], which uses the Levenberg-Marquardt algorithm (Figure

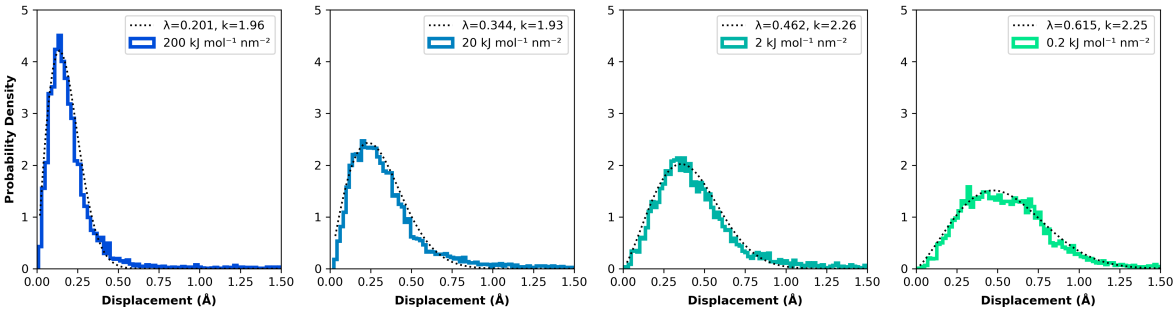


Figure 4.7: Histrogram of heavy atom atomic displacements to the experimentally-refined crystal structure, for the models refined against the structure factors from the 200 (blue), 20 (light blue), 2 (turquoise), and 0.2 (green) kJ $\cdot$ mol^{-1} $\cdot$ nm^{-2} restraint force constant cMD simulations, converted to probability density functions. Fits to Weibull distribution $\left(\frac{k}{\lambda}\right)\left(\frac{x}{\lambda}\right)e^{-(x/\lambda)^k}$ in dotted lines. Shape and scale parameters for each fit, k and λ respectively, presented in the labels.

4.7). The fits are quite good, with the variance of the parameters on the order of 10^{-6} and 10^{-4}, for λ and k respectively. The distributions were not well fit by Maxwell(-Boltzmann) or Rayleigh distributions.

The shape parameter, k, is similar for the 200 and 20 kJ $\cdot$ mol^{-1} $\cdot$ nm^{-2} distributions ($k =$ 1.93 - 1.96), and similar for the 2 and 0.2 kJ $\cdot$ mol^{-1} $\cdot$ nm^{-2} distributions ($k =$ 2.25 - 2.26), but they are not similar to each other, within uncertainty. The scale parameter, λ falls off linearly as a function of the logarithm of the restraint force constant (decreasing the force constant by an order of magnitude increases the scale parameter by a constant value, about 0.14).

In statistics, the Weibull distribution is often used to model failure rates — where the random variable is the time to failure — as it is the limiting distribution of the minimum of a large number of i.i.d. random variables that are at least zero[17]. In that context, a value of $k > 1$ indicates that the failure rate increases over time, and the failure rate itself is proportional to time to the power of $k - 1$. In this case the random variable is the displacement of the cMD-density-refined heavy atoms to those from the experimentally-refined crystal structure. However, a connection between the two can be made in the sense that the heavy atoms are exploring a harmonically-restrained space around the crystal structure positions with

simulation time, and the "failure" rate can be interpreted as the rate at which heavy atoms significantly deviate from the positions they are meant to occupy in the crystal structure. In this sense, the change in the shape parameter between the 200/20 and 2/0.2 $\mathrm{kJ \cdot mol^{-1} \cdot nm^{-2}}$ simulations may indicate that at some restraint force constant in between, the force field begins to influence the positions of the atoms more than the restraints do, and the "failure rate" converges to a new value. However, this analysis and interpretation is speculative at best.

Histidine Protonation State Discrimination

In addition to being used as a target for refinement, the structure factors calculated from the cMD trajectories can be analyzed on their own, as Fourier-transformed densities, for comparison to experimental composite ($2F_O - F_C$) and difference ($F_O - F_C$) maps. A nice example of this can be seen in the comparisons between the densities calculated from the initial cMD simulation, with all histidines doubly protonated, and the otherwise equivalent simulation with histidines protonated based on data from neutron diffraction.

Most of the histidines which changed protonation state, from doubly-protonated (HIP) to singly-protonated on either the delta (HID) or epsilon (HIE) nitrogen, did not change conformation, with little to no change evident in the density for the side chains calculated from the analyzed segment of the cMD trajectories. However, two residues did change conformation drastically after the protonation state was changed, both of which went from doubly protonated to protonated on the epsilon nitrogen: histidines 62 and 294. After this conformational change, we find that histidine 62 in the doubly protonated state is rotated almost ninety degrees relative to the density from experiment, whereas in the epsilon-protonated state the densities are almost equivalent. The effect of the protonation state on histidine 294 is even more drastic: in the doubly protonated state, the aromatic ring rotates and moves several angstroms away from the crystal structure position, opening up space for a

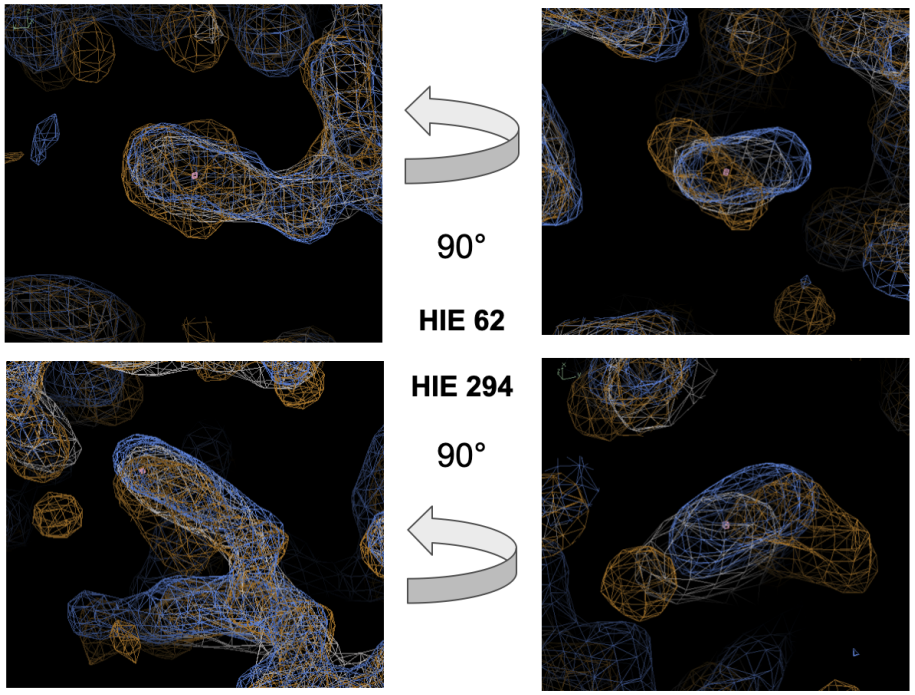


Figure 4.8: Comparison between experimental $2F_O - F_C$ map displayed as a 1 σ isosurface (white), the density calculated from the original cMD simulation with all histidines double protonated (orange), and the density calculated from the otherwise equivalent cMD simulation with protonation state determined by neutron diffraction (blue), both displayed as a 1 σ isosurface. Epsilon-nitrogen-protonated histidine (HIE) 62 above, HIE 294 below.

coordinated water; while the epsilon-protonated density is not perfectly aligned with the experimental density, the agreement is much improved (Figure 4.8).

In these cases, then, we find that density-based analysis from cMD simulations can (in certain cases) discriminate between correct and incorrect protonation states for histidines. Although many of the histidines did change conformation (in a way which was consistent across the ensemble) when changed from doubly-protonated to singly-protonated, when there were large conformational changes from this transition, they brought the cMD-predicted density in to closer agreement with the experimental density. There were no observed cases for this system in which the cMD-predicted density substantially disagreed with the experimental density in the simulation where the protonation states were correctly assigned.

4.4.3 Ordered Water Analysis

We've touched on another application of density-based cMD-trajectory analysis in a previous chapter: the prediction of ordered waters from crystallographic models. Here, we analyzed the effect of the strength of heavy-atom harmonic restraints on the precision and recall of crystallographic waters in cMD simulations. The precision and recall measures the number of cMD-density-refined waters which are within a certain distance of a crystallographic water. In this case, we measure the fraction of experimentally-refined crystallographic waters which have a cMD-density-refined water within a certain distance. The faster the curve slopes upward toward one, the more accurate the cMD model (in the limit, 100% of crystallographic waters have a cMD-predicted water in the exact same place — a distance of zero).

In this work, we converted the cMD-calculated structure factors to intensities to be used as as targets for refinement, and used `phenix.refine`, with the single-conformation experimental structure as input, to refine both the protein and solvent coordinates. This gives us a set of cMD-density-refined waters we can compare to those refined against experimental data. This allows us to test the degree to which the structure factors implied by cMD simulations recover the solvent density, and thus the crystallographic waters, from experimental refinement, at each restraint energy. There were 454 total waters refined against the structure factors from the 200 $\mathrm{kJ \cdot mol^{-1} \cdot nm^{-2}}$ restraint force constant simulation, 425 for the 20 $\mathrm{kJ \cdot mol^{-1} \cdot nm^{-2}}$ restraint force constant simulation, 346 for the 2 $\mathrm{kJ \cdot mol^{-1} \cdot nm^{-2}}$ restraint force constant simulation, and 296 for the 0.2 $\mathrm{kJ \cdot mol^{-1} \cdot nm^{-2}}$ restraint force constant simulation. That is, the solvent density contains fewer solvent peaks which can be identified by `phenix` as spots in which to place crystallographic waters.

Next, we tested the precision and recall of waters refined against the cMD data with respect to the positions of experimentally-refined crystallographic waters (Figure 4.9). The 200 $\mathrm{kJ \cdot mol^{-1} \cdot nm^{-2}}$ restraint force constant simulation recovers nearly all (137 out of 144; 95%)

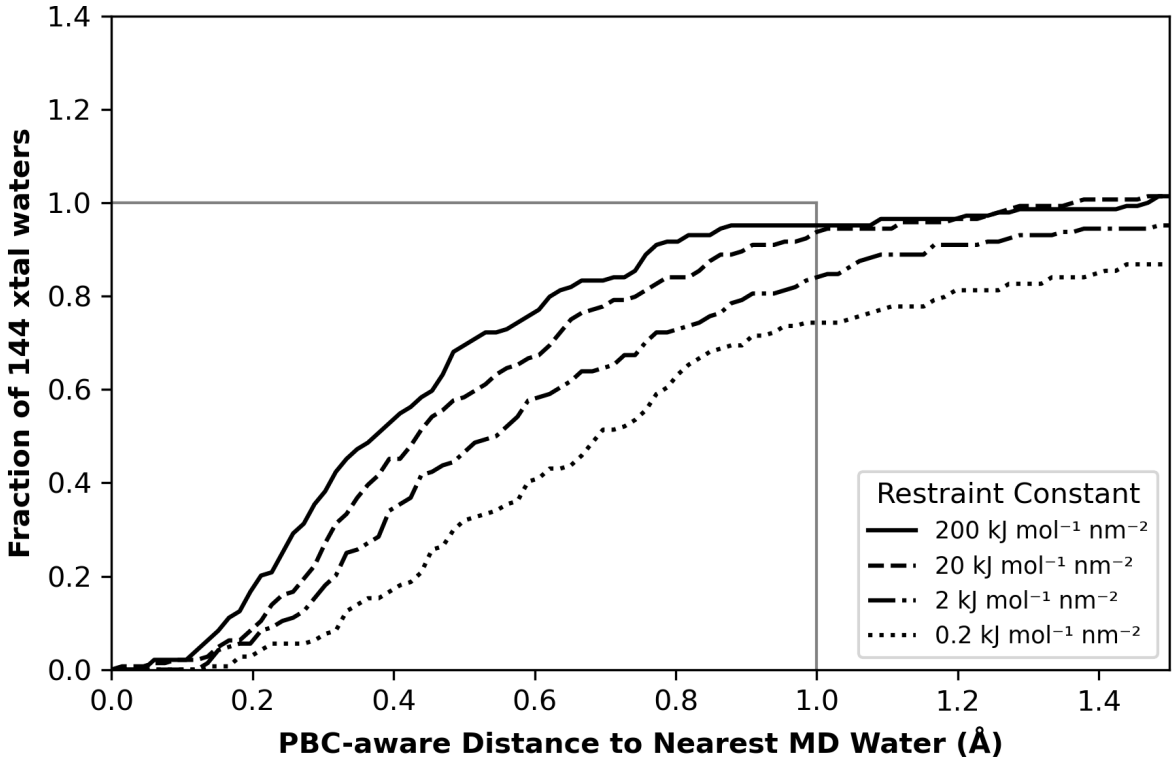


Figure 4.9: Precision and recall statistics for waters refined against the structure factors from cMD simulations with restraints with force constants 200 (solid), 20 (dashed), 2 (dash-dotted), and 0.2 (dotted) $\text{kJ} \cdot \text{mol}^{-1} \cdot \text{nm}^{-2}$. 200 and 20 $\text{kJ} \cdot \text{mol}^{-1} \cdot \text{nm}^{-2}$ restraint force constant simulations recover the same percentage of crystallographic waters to within 1 Å, while the precision and recall drops off for the 2 and 0.2 $\text{kJ} \cdot \text{mol}^{-1} \cdot \text{nm}^{-2}$ restraint constant simulations.

of all crystallographic waters to within one Angstrom. The results are similar for the 20 $\text{kJ} \cdot \text{mol}^{-1} \cdot \text{nm}^{-2}$ restraint force constant simulation, with 94% of crystallographic waters recovered to within one Angstrom. However, there is a steep drop off after this, with the 2 and 0.2 $\text{kJ} \cdot \text{mol}^{-1} \cdot \text{nm}^{-2}$ restraint force constant simulations recovering 84% and 74% of crystallographic waters within one Angstrom.

These results are in line with the results of the previous section. There is a change in behavior for the system between 20 and 2 $\text{kJ} \cdot \text{mol}^{-1} \cdot \text{nm}^{-2}$ — in the previous case, with respect to the shape parameter describing the distribution of atomic displacements from the crystal structure; in this case with respect to the number and position of waters refined against the cMD data. There may be a connection between these two findings: as the force field takes

over in influence from the restraints to the crystal structure heavy atoms, the structural and electrostatic interactions responsible for holding ordered waters in place may be corrupted.

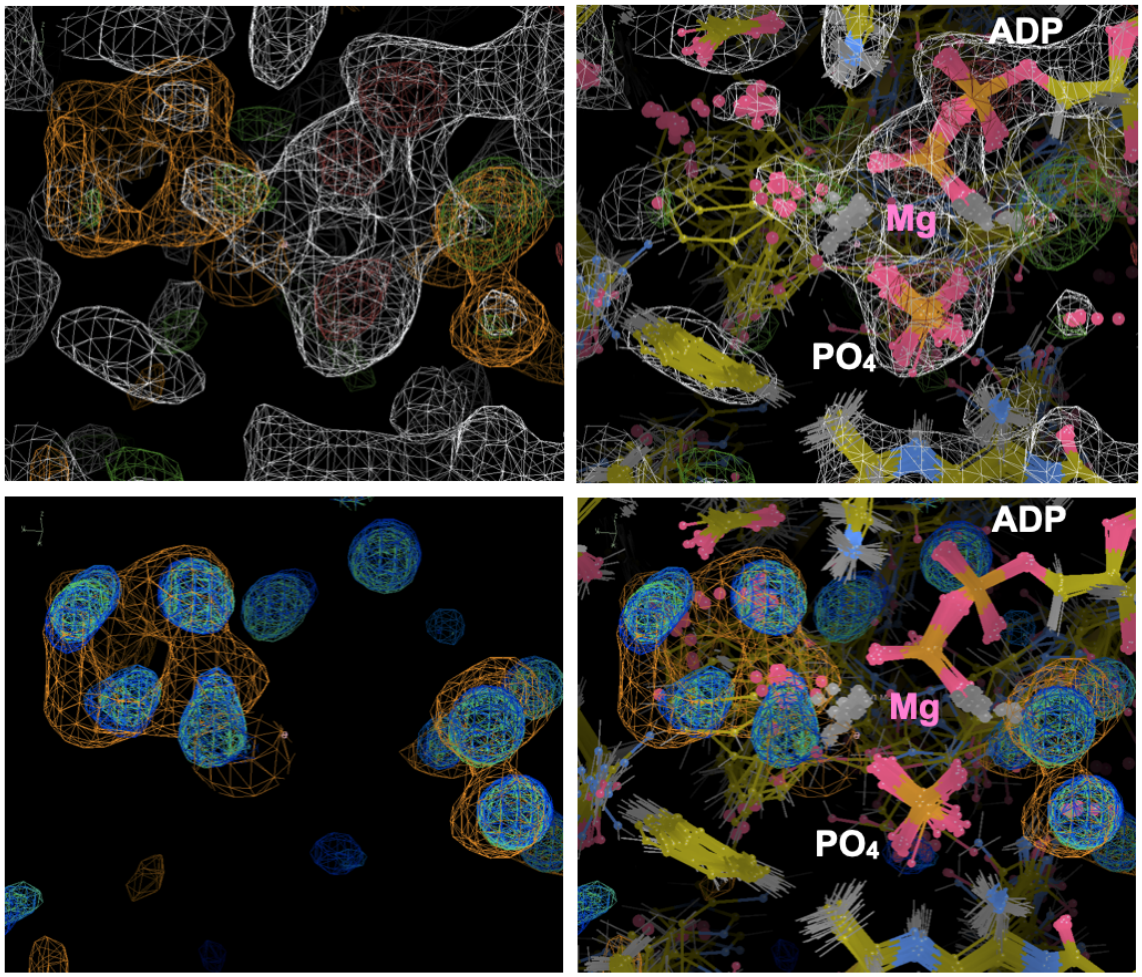


Figure 4.10: Composite and difference maps from experiment, Polder map excluding active site waters, and solvent density from cMD simulatios. Top left: experimentally-refined composite map (displayed as a 1σ isosurface, white) and difference map (displayed as 3σ, positive in green, negative in red), and Polder map (displayed as a 3σ isosurface, orange); top right; experimentally-refined composite and difference maps, with experimentally-refined ensemble protein (orange bonds) and water (red) atoms; bottom left: cMD solvent density (isosurface at absolute density of 11 e/Å^3; densities from the highest to lowest restraint force constant simulations displayed in dark blue to light green), and Polder map (orange); bottom right: ensemble protein and solvent coordinates, with cMD solvent densities and Polder map.

We can also identify areas of cMD-predicted solvent density around the active site to determine how well the force field performs at reproducing catalytically-relevant water positions. The experimentally-refined ensemble model has waters placed in to solvent peaks in the composite map, however, we also constructed Polder maps by removing the solvent molecules

around the active site, and computing an OMIT map without bulk solvent corrections: this map identifies areas of weak solvent density which are still consistent with the data from experiment (but may have been un-modelled, perhaps by over-fitting, for the full structural model). Comparisons between the experimental composite and difference maps, cMD solvent maps from each different restraint force constant simulation, and the Polder map, are shown below for two interesting areas around the active site (Figures 4.10 and 4.11).

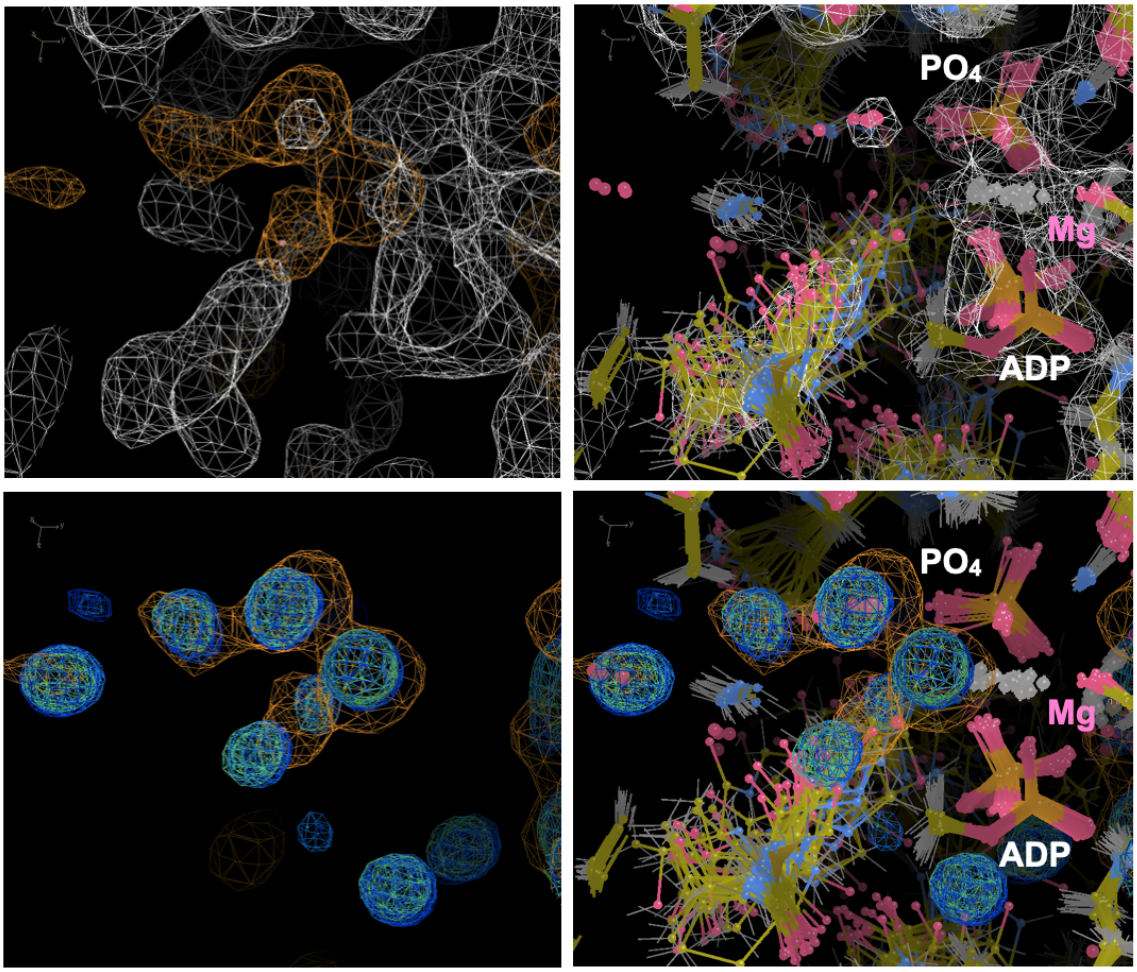


Figure 4.11: Composite and difference maps from experiment, Polder map excluding active site waters, and solvent density from cMD simulatios. Top left: experimentally-refined composite map (displayed as a 1σ isosurface, white) and difference map (displayed as 3σ, positive in green, negative in red), and Polder map (displayed as a 3σ isosurface, orange); top right; experimentally-refined composite and difference maps, with experimentally-refined ensemble protein (orange bonds) and water (red) atoms; bottom left: cMD solvent density (isosurface at absolute density of 11 e/Å^3; densities from the highest to lowest restraint force constant simulations displayed in dark blue to light green), and Polder map (orange); bottom right: ensemble protein and solvent coordinates, with cMD solvent densities and Polder map.

There are a few things to note in general for the two figures above: (i) all of the crystallo-

graphic waters around the active site are reproduced as solvent peaks in the cMD simulations, for all restraint force constants, (ii) many of the cMD-predicted solvent density peaks are supported by the Polder map, even if they do not have associated crystallographic waters. Both magnesium ions in the active site have *very* closely associated waters (within one Å), both of which are supported by the Polder density, but only one of which had crystallographic waters modelled in by the ensemble refinement (the modelled in water is to the left of the magnesium in the top right and bottom right images in Figure 4.10; the unmodelled water is to the left of the magnesium ion in the top right and bottom right images in 4.11). Both of these areas of solvent density in the cMD simulations coordinated with the magnesium ions are associated with other waters, organized in tight hydrogen bonding networks: both of these hydrogen bonding networks are supported by the Polder map, but the ensemble refinement misses most of the waters in these networks. There is very little difference density in these regions as well, suggesting that the crystallographic model does not predict a lack or excess of electrons in these regions.

4.4.4 Side Chain Disorder Analysis

Next, we moved from analyzing the disorder captured by the B-factors present in the models refined against the data from cMD simulations, to analyzing the more complex forms of disorder in the simulation ensembles. In the experimentally-refined ensemble, some side chains occupy roughly the same conformation in every representative structure, but many side chains are modelled with a broad range of conformations across the ensemble in areas where the experimental density is poorly resolved. In fact, this the *purpose* of ensemble modelling: to take advantage of the fact that areas of low-resolution density imply structural heterogeneity or disorder in order to produce an ensemble of structures, all of which are equally compatible with experimental data. In the simulations, this disorder is also modelled, in this case by the force field. However, there are some striking differences between the

structural ensemble generated by refinement and the structural ensemble generated by the force field.

In Figure 4.12, I present a few different examples of the types of order and disorder present in the experimentally-refined ensembles, and the reverse-propagated ensembles from the final frame of the 200 $\mathrm{kJ \cdot mol^{-1} \cdot nm^{-2}}$ restraint force constant simulation, along with the density calculated from the final 10 ns of the trajectory. Both the experimental composite $(2F_O - F_C)$ map and the cMD density are displayed as 1σ isosurfaces (for lysine 217, the chloride density is displayed as a 12σ isosurface).

For lysine 105, the experimental composite map is well-resolved only for the backbone, and the ensemble refinement produces significant structural heterogeneity for the side chain; the cMD ensemble has significant structural heterogeneity for the side chain as well, and the only resolved density is along the backbone, matching the experimental composite map. In contrast, for lysine 189, the experimental composite map is well-resolved along the backbone and the entirety of the side chain (this side chain is involved in the positioning of the ADP molecule in preparation for catalysis, so its conformation is stabilized by interactions with the ADP in the active site); the cMD density and ensemble are similarly well-resolved and structurally homogeneous. Lysine 217 presents an interesting case in which the experimental composite map is resolved only along the backbone, but the refined ensemble is relatively structurally homogeneous, with the side chain modelled in to density far from the backbone; however, even in the most strongly-restrained simulation, the cMD ensemble converges to a different configuration than the experimentally-refined ensemble, and the cMD model suggests that the density far from the backbone is chloride density (this is supported by the experimental difference map, which displays positive difference density, even with the side chain atoms modelled in to the off-backbone density). The cMD model of lysine 217 has about the same structural homogeneity as the experimentally-refined ensemble, but the position of the side chain is different.

Because both the experimentally-refined ensemble and reverse-propagated cMD ensembles have many different members, it is hard to characterize differences between the two ensembles with a single measure (like the RMSD), particularly for disordered side chains. Lysine 105, for instance, has significant structural heterogeneity in the side chain in both the experimentally refined ensemble and the cMD ensemble, so each atom in the cMD ensemble is likely quite far from its starting position in the experimentally refined ensemble, after 100ns of simulation. Even on average, the heavy atoms in this side chain would have a high RMSD to their starting positions. However, the same is true for lysine 217: all of the atoms in the side chain from the cMD reverse-propagated ensemble are far from their starting positions in the experimentally-refined ensemble, and thus would have a high RMSD. However, this high RMSD between the cMD and experimentally-refined ensemble models is somewhat more "meaningful" than in the case of lysine 105: it represents a significant difference in the modelling of the side chain position and structural homo-/heterogeneity between the experimentally-refined and cMD side chain ensembles.

Using lysines 105 and 217 as characteristic examples, we can derive a reasonable test for significant differences between the experimentally-refined and cMD modelled ensembles. For a side chain with significant structural heterogeneity, if the position and disorder of the side chain are modelled equivalently by the experimental refinement and the cMD model, each side chain in the ensemble will wander from its initial position, leading to a high ensemble-averaged RMSDs from the crystal structure positions, however, if the disorder in the cMD ensemble is similar to that of the experimentally refined ensemble, the atoms will explore the same space, leading to similar standard deviation of atomic positions (SDAPs) between the atoms of the experimentally-refined and cMD ensembles. This would be indicative of model agreement: though the RMSD is high, the cMD model is building in the correct amount of structural heterogeneity for the side chain, measured by high SDAP. (For a side chain with little structural heterogeneity and disorder, both the RMSD between the cMD and experimentally refined ensemble atoms and SDAP of the side chain atoms in each ensemble

will be low — this would also be indicative of model agreement). However, if a side chain has low structural heterogeneity in the experimentally-refined ensemble and low structural heterogeneity in the cMD ensemble, but the side chain has a different position in the cMD ensemble, there will be a high RMSD between the experimentally-refined and cMD ensemble atoms, but a *small and similar* SDAP for the atoms from the two ensembles. In some cases the SDAP for the atoms in the cMD ensemble may be *smaller* than those from the experimentally-refined ensemble, if the cMD model predicts a more certain position for the side chains in the ensemble. All of this is to say: a combination of the RMSD and SDAP can be used to identify significant differences between the cMD and experimentally-refined ensembles.

Another case similar to that of lysine 217 mentioned above, is phosphorylated serine (SEP) 139, (Figure 4.3): as the restraint force constants are lowered, the cMD model converges on a new conformational ensemble. In this case, the ensemble-average SDAPs for the side chain atoms stay relatively constant (there is not a lot of structural heterogeneity for the side chains, across the ensemble, at any restraint strength) while the RMSDs to the crystal structure positions increase (the conformational ensemble moves to a completely different position). This is the another prototypical case of meaningful difference between the cMD and experimentally-refined ensemble models, but this is by far the most extreme case we observed in our cMD model: the force field forces the side chain in to a significantly different position from the experimentally-refined ensemble, and the deviation gets more substantial as the restraints are relaxed.

Backbone, Side Chain, and Residual RMSD and SDAP

We reverse-propagated the proteins from the final frame of each of the cMD simulation, so all of the different ensemble members were superimposed in the asymmetric unit, as was the case for the original experimentally-refined ensemble. We then matched each protein from

the reverse-propagated supercell with the experimentally-refined ensemble member which provided its starting structure. From here, we calculated the RMSD to the crystal structure position in the following way: for each ensemble member, and for each heavy atom in the protein, we calculated the average RMSD over all the heavy atoms in the residue, then averaged these residue-averaged RMSDs over the entire ensemble. This gives us an ensemble average RMSD for each residue in the protein, for each restraint force constant simulation.

We then performed a similar calculation for the SDAP: for each heavy atom in each ensemble (the experimentally-refined ensemble, and the reverse-propagated ensemble from the final frame of each restraint force constant simulation), we calculated the three-dimensional ensemble-average position, then calculated the (euclidean) displacement from this average position for each heavy atom, summed the squared deviations, averaged over all the equivalent atoms in the ensemble, and took the square root of this average variance to find the standard deviation. We then calculated the average SDAP for each residue in a similar way as above (averaging the SDAPs over all heavy atoms in each residue).

If we plot the RMSD and SDAP simultaneously (with rising SDAP going up from the x-axis, and rising RMSD going down from the x-axis) we can simultaneously observe the change in ensemble-and-residue-averaged RMSD and SDAP for each residue in the protein (for the experimentally-refined ensemble there is SDAP only, RMSD is relative *to* the experimentally-refined-ensemble) as the restraint strength is diminished. The results for all side chains are shown in Figure 4.13.

Areas of high initial (experimentally-refined-ensemble) SDAP correspond to regions of high structural heterogeneity (as in the G loop, residues 50-56 in the left inset of Figures 4.13, 4.14, and 4.15, gray scatter points from the experimentally-refined ensemble, light to dark green scatter points from the strongest to weakest restraint force constant simulations). Not surprisingly, regions of high structural heterogeneity also show large changes in RMSD as restraint force constants are decreased (as restraints relax, residues are allowed more flexibility,

wandering farther from their initial positions, leading to higher RMSD on average). However, some residues with low structural heterogeneity (low SDAP) in the experimentally-refined ensemble experience large increases in both SDAP and RMSD as the restraints are relaxed (right inset in all figures): these are residues for which the MD predicts larger amounts of structural heterogeneity than is predicted by the experimental ensemble refinement.

The order or disorder represented by the side chain RMSDs and SDAPs in the figure above are not representative of the behavior of the side chains *alone*. Side chains are so named because they are *chained* to the *side* of the backbone. If the backbone exhibits large RMSD and/or SDAP, this disorder will be present in the RMSD and SDAP of the side chain atoms as well (and vice versa if it does not). So, to isolate the behavior of the side chain on its own, we constructed a similar plot to the one above, considering only the backbone atoms, shown in Figure 4.14.

Here, we can see that many of the areas of high and/or increasing side chain RMSD and SDAP from Figure 4.13 are primarily the result of backbone motions rather than the motions of the side-chain itself. For instance, the G loop is modelled with significant structural heterogeneity in the experimentally-refined ensemble (Figure 4.14). Interestingly, in this region, the backbone SDAP *decreases* relative to the experimentally-refined ensemble for the the 200 and 20 $\mathrm{kJ \cdot mol^{-1} \cdot nm^{-2}}$ (in some cases decreasing relative to the experimentally-refined ensemble in *all* simulations, regardless of restraint strength). Thus, the cMD model is suggesting that the experimentally-refined ensemble is modelling in *too much* backbone structural heterogeneity in this region.

To isolate the RMSD and SDAP of the side-chain alone, we subtracted the backbone RMSD and c-α SDAP from those of the side chains, leaving the "residual" RMSD and SDAP, shown in Figure 4.15. We return to identifying the type of disorder present in the side chains, by examining the relationship between residual SDAP and RMSD for each residue. The residual RMSD and SDAP are both very small and remain close to zero for all restraint

force constants for most residues. Compare this to the plot for side chains (Figure 4.13) and backbone (Figure 4.14) in which, for most residues, the RMSD and SDAP increase as the restraint force constants decrease. This finding is indicative of the fact that, for the most part, the disorder present in the side chains is a knock-on effect of disorder from the backbone, with the side chain contributing very little to the ensemble disorder captured by the residual RMSD and SDAP. It is also indicative of the fact that the cMD model largely agrees with the experimentally-refined ensemble on the position and structural heterogeneity for most of the side chains.

If the residual RMSD and SDAP both increase as the restraint force constants decrease (e.g. in phenylalanine 281 or leucine 284 in the right inset of figure 4.15) the deviation from the crystal structure ensemble is matched by an increase in atomic fluctuations about the mean positions as restraints are relaxed. This is the type of disorder we mentioned above as indicative of agreement between the ensemble models: though the RMSD between the cMD and experimentally-refined ensemble models is high, if the SDAP for both is also high, the two models are in agreement. Because an increase in SDAP moves upward, and an increase in RMSD moves downward, an increase in both RMSD and SDAP with a decrease in the restraint force constant is visually represented by scatter points which mirror each other in their movement away from the x-axis in the plots. We say that these residues "pass the mirror test".

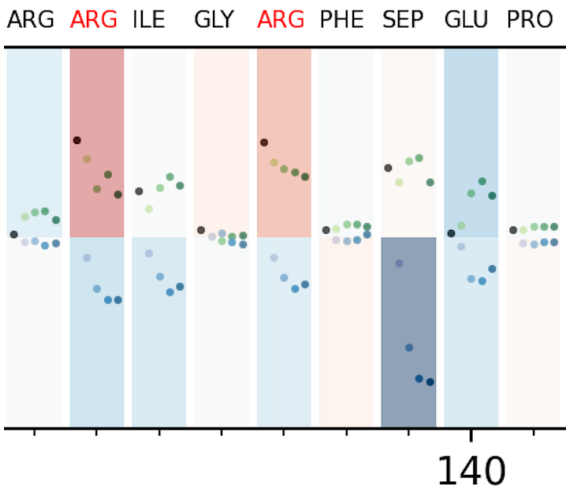


Figure 4.16: Residual SDAP (above x-axis, increasing upward) and residual RMSD (below the x-axis, increasing downward) for residues 132-141. Phosphorylated serine (SEP) 139 shows roughly constant residual SDAP, with an associated large increase in RMSD as restraints are relaxed (light to dark green above, light to dark blue below).

However, if a residue strays from its position in the experimentally-refined ensemble (RMSD

increases), but the residue is exhibiting a constant or decreased amount of structural heterogeneity (converging on a new position for the side chain or a more certain position across the ensemble in the cMD model) the SDAP will either remain constant or *decrease* relative to the SDAP of the experimentally-refined ensemble. Because an increase in RMSD and constant or decreasing SDAP move in different directions with respect to the x-axis (one away from it, one towards it), we say that these residue "fail the mirror test". We used the failed mirror test as a heuristic by which to identify residues for which the cMD model deviates from the structural ensemble modelled by experimental refinement.

As an example: consider again phosphorylated serine (SEP) residue 139, visualized in both density and coordinate views in Figure 4.3. This residue provides a clear example of the ways in which the relationship between residual RMSD and SDAP can capture meaningful differences in between the cMD and experimentally-refined ensemble models. The residual SDAP for the residue stays roughly constant for all cMD simulations at the various restraint strengths (indicating a roughly constant amount of structural heterogeneity) but the RMSD increases (indicating that the residue is moving concertedly away from its position in the experimentally-refined ensemble) (Figure 4.16). In this case, the residue is not highlighted in red in the residual RMSD/SDAP because its SDAP does not *decrease*, it only remains constant – residues highlighted in red correspond to the most extreme failures of the mirror test, with both large increases in RMSD and large *decreases* in SDAP.

There are complexities to the interpretation of the failed mirror test that warrant discussion. Lysines 105, 189, and 217 were highlighted in Figure 4.12, showing three different types of disorder. Lysine 105 has significant structural heterogeneity for the side chain in both the experimentally-refined ensemble and the cMD ensembles; lysine 189 has very little structural heterogeneity in both the experimentally-refined and cMD ensembles; lysine 217 has significant structural heterogeneity in the experimentally-refined ensemble, with a number of the side chains modelled in to off-backbone density, but the cMD model has relatively less struc-

tural heterogenity, modelling the off-backbone density as chloride or solvent density, and modelling most of the side chain in a similar conformation. As expected the residual RMSD and SDAP for lysine 189 are both very low, and remain low at all restraint energies. Also, lysine 217 is identified by the mirror test as a residue for which the experimentally-refined and cMD models diverge (Figure 4.17).

However, lysine 105 is also identified by the failed mirror test as a residue for which the experimentally-refined and cMD models diverge (Figure 4.17). In this case though, the decrease in residual SDAP comes mostly from the fact that the disorder in the side chain stays roughly constant while the disorder in the *backbone* increases (Figure 4.18). This is representative of a major difference between the cMD and experimentally-refined ensemble models: the experimentally-refined ensemble model refines the B-factors of the atoms along with the coordinates. The B-factors smooth out the density, representing small scale harmonic disorder about the atoms' average positions. The cMD model builds in this disorder *explicitly*. So, though the densities appear roughly equivalent between the two models, the coordinates represented by the density are quite different. It's important to note that the cMD model predicts disorder in the backbone which is not (entirely) small scale and harmonic: there are some ensemble members with backbone conformations that are quite different from the rest (lower left image in Figure 4.18).

In general, most of the residues identified by the failed mirror test happen to be (i) residues with significant structural heterogeneity modelled by the experimental refinement, and (ii) on the surface of the protein (Figure 4.23). The failed mirror test does appear to be useful in identifying residues for which the experimentally-refined and cMD ensembles diverge. This does not mean that the failed mirror test (and the residual RMSD and SDAP in general) cannot be used as a measure with which to identify side chains which are significantly differently modelled between the experimentally-refined and cMD ensemble models. Below are four residues, identified using the failed mirror test — glutamine 39 (Figure 4.19), arginines

134 and 137 (Figures 4.20 and 4.21), and lysine 177 (Figure 4.22) — which, despite exhibiting the same increase in backbone disorder mentioned above, also exhibit noticeably different structural ensembles for the side chains that those from the experimentally-refined ensemble.

Glutamine 39 is an interesting case in which there is a straight-forward interaction with a residue on a neighboring protein (Lysine 83, another residue which fails the mirror test), which draws the end of the side chain away from its experimentally-refined ensemble position. The backbone is more disordered than the experimentally-refined ensemble, but the side chain also adopts a noticeably different conformation (with the amino group pulled off to the left in the bottom right image of Figure 4.19). These ensemble changes are represented in the density as well (bottom left image).

Arginines 134 and 137 both deviate from their experimentally-refined ensembles in the cMD model. Arginine 134 adopts two major conformations across the ensemble in the 0.2 $kJmol^{-1}nm^{-2}$ restraints force constant simulation, with some side chains jutting off to toward an adjacent phenylalanine (to the left), and others bending toward a symmetry related copy of serine 212 (to the right), whereas the experimentally-refined ensemble is more scattered. The difference between the experimentally-refined and cMD ensemble models of arginine 137 is more subtle: the experimental model has a few side chains jutting off in to solvent density, while the cMD model has more structural homogeneity, despite the increased backbone disorder.

Lysine 177 has such a significant change from structural heterogeneity to homogeneity that it shows up in the density from the cMD simulation. In the experimentally-refined ensemble model, the side chain is modelled with significant structural heterogeneity, but in the cMD model, even with significant disorder in the backbone, the side chain adopts such a concerted conformation across the ensemble that there is resolved density along the entire side chain, even at the lowest restraint energy. The amino group of the lysine is pulled toward the backbone oxygen of arginine 308 (top left of the top right image in Figure 4.22).

As we've seen from examining each of these individual cases, the "mirror test" is an imperfect, but still useful, quantitative heuristic for identifying side chain ensembles which the cMD and experimental-refinement model differently. However, it is important to ask how well the cMD density agrees with the density from experiment, even in cases where the side-chain ensembles are modelled differently. It may be the case, for example, that the cMD converges on a different structural ensemble for the side chain, but the density implied by the new structural ensemble agrees with the experimental structure factors just as well.

To test this, we defined a mask region around each of the side chain ensembles using the `sfall` method from `CCP4`, and analyzed the correlation between the experimental density (converting the experimental structure factors to a density using `phenix.mtx2map`) and the cMD-calculated density (using the same method) in the region defined by the mask, for each residue, using the `overlapmap` method from `CCP4`. This gives us a real space correlation coefficient (RSCC) between the experimental and cMD-predicted density for each residue, for each restraint force constant cMD simulation. The results are presented in Figure 4.24.

There is an overall increase in correlation with the experimental map as the restraint force constants decrease, with an average increase in RSCC across all residues of 0.028 going from 200 to 20 $\text{kJmol}^{-1}\text{kJ}^{-2}$, 0.048 going from 200 to 2 $\text{kJmol}^{-1}\text{kJ}^{-2}$, and 0.047 going from 200 to 0.2 $\text{kJmol}^{-1}\text{kJ}^{-2}$. Thus, on average, as the restraints are relaxed, the side-chain density predicted by the cMD model agrees better with the experimental density, with the increase in correlation maxing out somewhere between 20 and 2 $\text{kJmol}^{-1}\text{kJ}^{-2}$. In fact, even though the "failed mirror test" identifies residues with significant differences between the experimentally-refined ensemble and the cMD predicted ensemble, the correlation increases for these residues as well, on average, with a RSCC increase of 0.03 going from 200 to 20 $\text{kJmol}^{-1}\text{kJ}^{-2}$, 0.05 going from 200 to 2 $\text{kJmol}^{-1}\text{kJ}^{-2}$, and 0.03 going from 200 to 0.2 $\text{kJmol}^{-1}\text{kJ}^{-2}$. Overall, the cMD density appears to agree more with the density from experiment as the restraints are relaxed. This is true even for the residues which fail the mirror test — that is for residues

with side chain ensembles that are differently modelled by the cMD simulation than those modelled the experimental refinement.

4.4.5 Conclusions

In this chapter, we've laid out a systematic study of *protein kinase A* (PKA) using a crystalline molecular dynamics simulation. This cMD system is novel in that it each of the proteins in the supercell start from a different position pulled from an experimentally-refined ensemble model. We also studied the effect of restraint strength on the modelling of the density and the dynamics of the cMD ensemble.

We found that the overall *pattern* of the heavy-atom B-factors is modelled well, even in relatively strongly restrained simulations, while the relaxation of the restraints brings the *magnitude* of the B-factors closer to those from experiment. There may be a way to systematically scale the predicted B-factors up based on the restraint strength, to identify atoms or regions which are being modelled as overly or insufficiently disordered even at strong restraint strength, by comparison with the experimental B-factors.

We also found that cMD simulations of PKA are able to reproduce the crystallographic waters from experimental refinement with very good precision and recall — similar to that of previous studies on a system of enduglucanase. Unlike previous studies, we also showed that the restraint force constant can be lowered by an order of magnitude without losing signnificant accuracy in the prediction of crystallographic waters. However, somewhere between a restraint force constant of 20 and 2 $\mathrm{kJ \cdot mol^{-1} \cdot nm^{-2}}$ the accuracy of the cMD model with respect to the prediction of crystallographic waters begins to plummet (from 95% of waters recovered to within an Angstrom to 85% or below). Further work will need to be done to isolate the source of this decline in accuracy, but the displacement of atoms from their crystal positions may to play a strong roll (as may an increase in unrestrained thermal

disorder). We were also able to identify potentially catalytically relevant water networks around the active site of PKA which were not modelled in by ensemble refinement, and were not associated with difference density (but were supported by a Polder map).

Finally, we were able to identify a heuristic with which to identify side chain ensembles which are differently modelled by the cMD than the are by the experimental refinement. The RMSD and SDAP of heavy atoms (in partcular, the residual, or side chain minus backbone RMSD and SDAP) are simple measures of disorder which manage to to capture information about the differences between two complicated ensembles. These measures can be used to easily identify residues in the cMD simulation whose positions are significantly different from those in the refined ensemble structural model, even if both ensembles exhibit significant structural heterogeneity, allowing cMD modelers to identify areas were the ensemble model may be building in side chains incorrectly (for instance, building the side chain in to density that is more likely solvent).

We hope this work serves as example of the many and varied ways in which cMD models can be useful in the modelling of crystallographic structure and disorder, particularly for structural ensembles. cMD models can be used to study crystal systems from a coordinate-based (ensemble) perspective and a density-based perspective. The density based perspective is particularly powerful: the cMD model can be used to characterize areas of density as either protein, solvent, or ion (or in some cases, a combination of multiple), a capability which is not possible with standard crystallographic methods. This capability can be leveraged to test hypotheses about the relationship between ordered water and side chains or active site molecules which might not be possible otherwise, while maintaining the ability to compare results with high quality crystallographic data. We hope others may find these techniques useful.

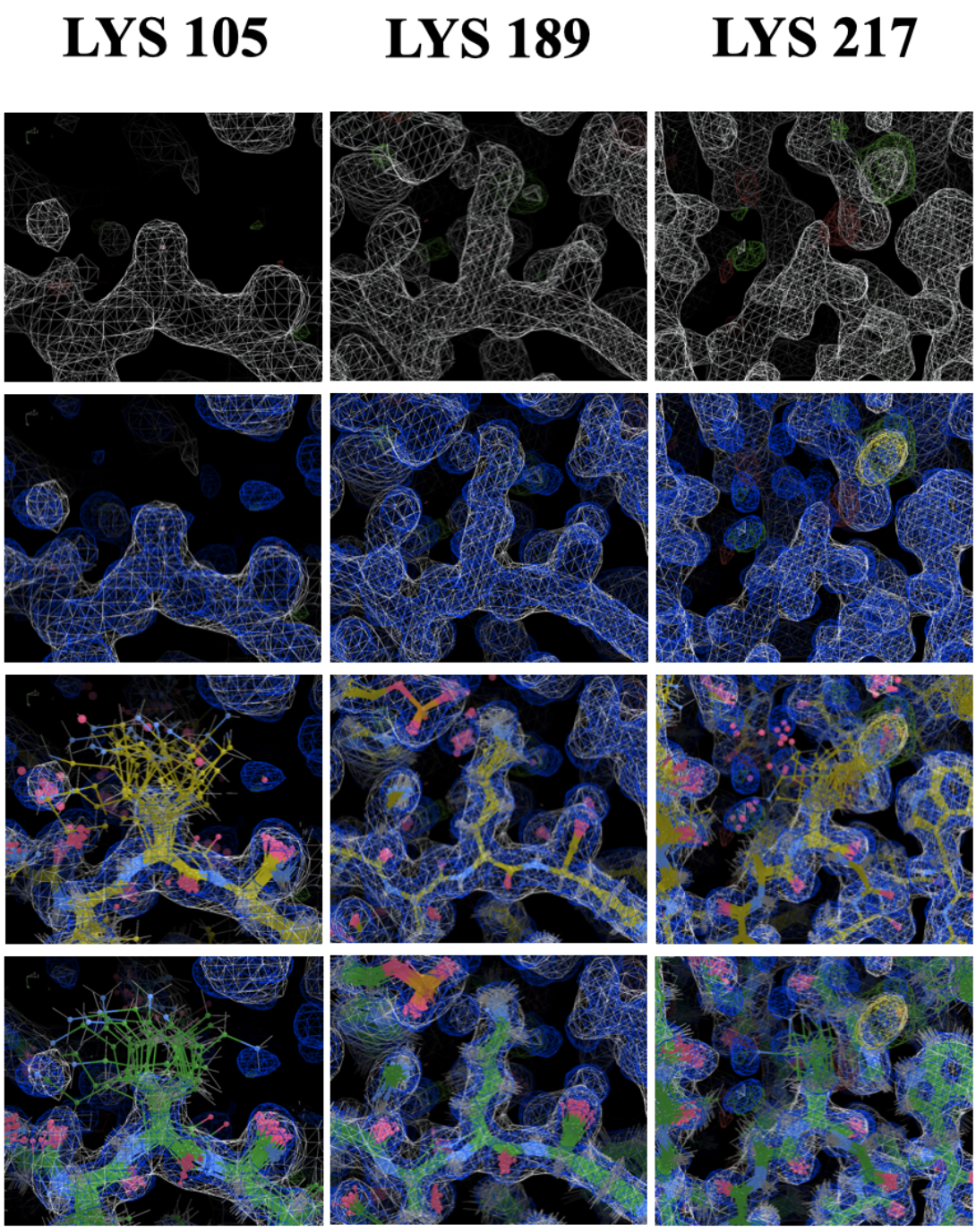


Figure 4.12: First row: experimental composite map ($2F_O - F_C$; white) and difference map ($F_O - F_C$; positive in green, negative in red) displayed as a 1σ and 3σ isosurface, respectively. Second row: exp. composite map (white) and cMD-calculated density (blue) from the final 10ns of the 200 $\mathrm{kJmol^{-1}kJ^{-2}}$ restraint force constant simulation, displayed as a 1 σ isosurface. Third row: exp. composite map (white) and cMD density (blue) with experimentally-refine ensemble model (yellow bonds). Bottom row: exp. composite map (white) and cMD density (blue) with reverse-propagated ensemble from the final frame of the cMD trajectory (green bonds). Side chain of lysine 105 is structurally heterogeneous in both ensembles; side chain of lysine 189 is structurally homogeneous in both models; side chain of lysine 217 is relatively structurally homogeneous in both models, but the models place the side chain in different positions.

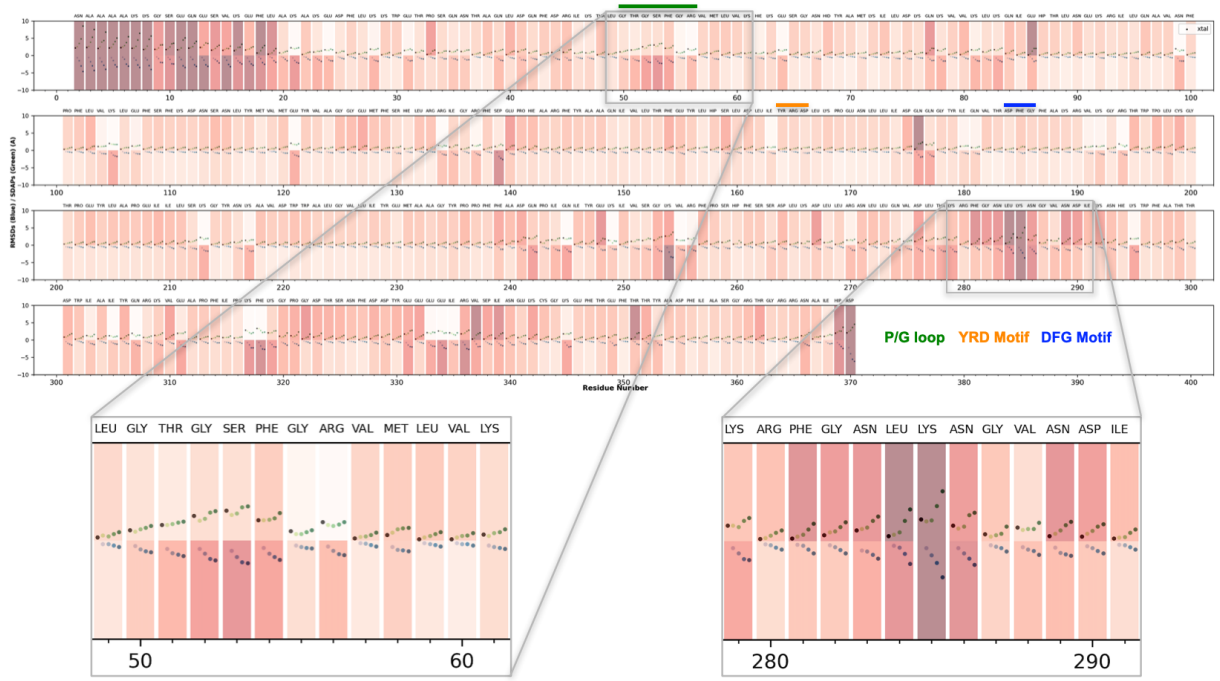


Figure 4.13: Above the x-axis: **side chain** heavy-atom ensemble-and-residue-averaged SDAP for the experimentally-refined ensemble (grey) and for the 200 $kJ \cdot mol^{-1} \cdot nm^{-2}$ (light grey-green), 20 $kJ \cdot mol^{-1} \cdot nm^{-2}$ (lime green), 2 $kJ \cdot mol^{-1} \cdot nm^{-2}$ (medium green), and 0.2 $kJ \cdot mol^{-1} \cdot nm^{-2}$ (forest green) restraint force constant cMD simulation final frame ensembles. Below the x-axis: **side chain** heavy-atom ensemble-and-residue-averaged RMSD for the cMD simulation final frame ensemble, from light to dark blue, in the same order as above. Both are colored to indicate the small (white) to large (dark red) changes in RMSD and SDAP. There is high RMSD and SDAP (and large increases in both, as restraints are relaxed) in flexibile regions, such as the G loop (green bar).

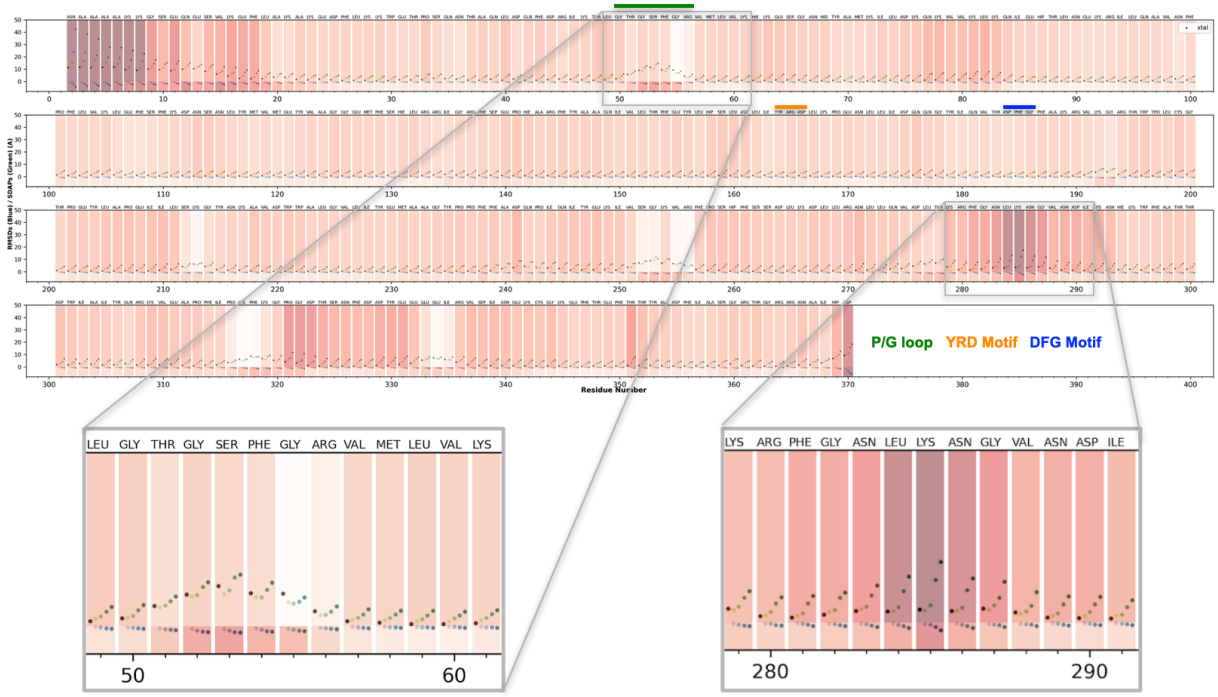


Figure 4.14: Above the x-axis: **backbone** heavy-atom ensemble-and-residue-averaged SDAP for the experimentally-refined ensemble (grey) and for the 200 $kJ \cdot mol^{-1} \cdot nm^{-2}$ (light grey-green), 20 $kJ \cdot mol^{-1} \cdot nm^{-2}$ (lime green), 2 $kJ \cdot mol^{-1} \cdot nm^{-2}$ (medium green), and 0.2 $kJ \cdot mol^{-1} \cdot nm^{-2}$ (forest green) restraint force constant cMD simulation final frame ensembles. Below the x-axis: **backbone** heavy-atom ensemble-and-residue-averaged RMSD for the cMD simulation final frame ensemble, from light to dark blue, in the same order as above, with increasing RMSD going down from the x-axis. Both are colored to indicate the small (white) to large (dark red) changes in RMSD and SDAP. There is high RMSD and SDAP (and large increases in both, as restraints are relaxed) in flexibile regions, such as the G loop (green bar).

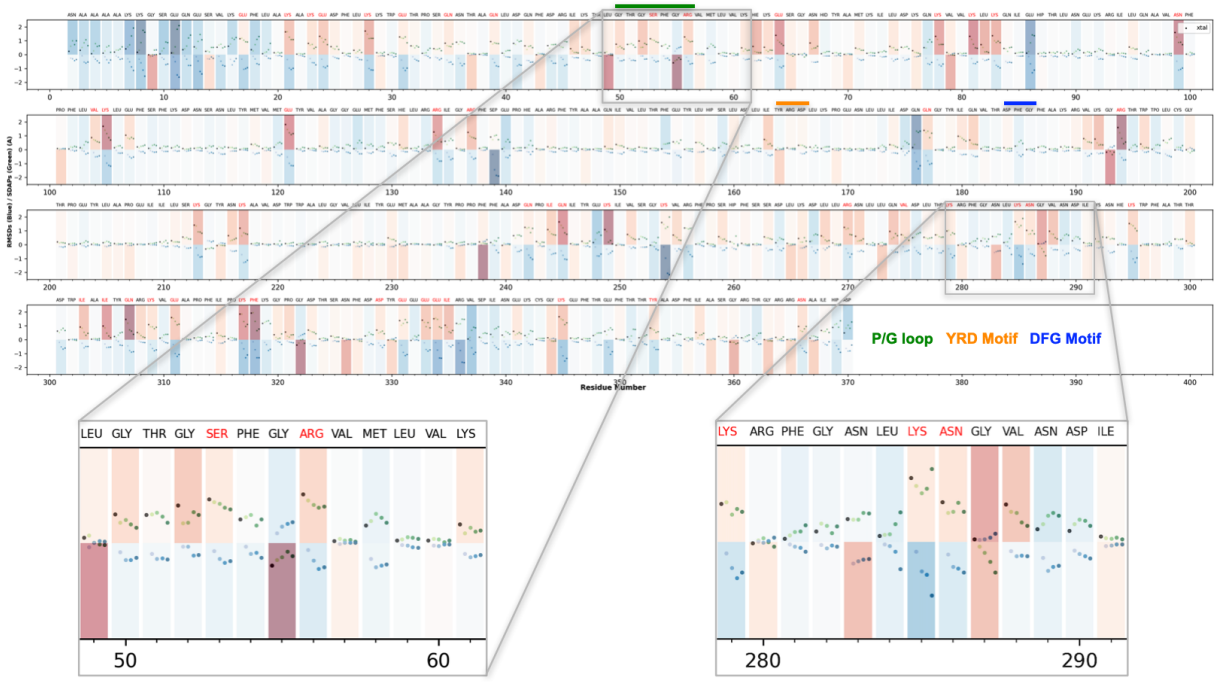


Figure 4.15: Above the x-axis: **residual** (side chain minus backbone) heavy-atom ensemble-and-residue-averaged SDAP for the experimentally-refined ensemble (grey) and for the 200 kJ $\cdot$ mol^{-1} $\cdot$ nm^{-2} (light grey-green), 20 kJ $\cdot$ mol^{-1} $\cdot$ nm^{-2} (lime green), 2 kJ $\cdot$ mol^{-1} $\cdot$ nm^{-2} (medium green), and 0.2 kJ $\cdot$ mol^{-1} $\cdot$ nm^{-2} (forest green) restraint force constant cMD simulation final frame ensembles. Below the x-axis: **backbone** heavy-atom ensemble-and-residue-averaged RMSD for the cMD simulation final frame ensemble, from light to dark blue, in the same order as above, with increasing RMSD going down from the x-axis. Both are colored to indicate the negative (dark red), zero (white), and positive (dark blue) changes in RMSD and SDAP. "Failed mirror test" residues will have negative (red) change in SDAP and positive (blue) change in RMSD – residues with both large negative SDAP change and large positive RMSD change are highlighted with red labels.

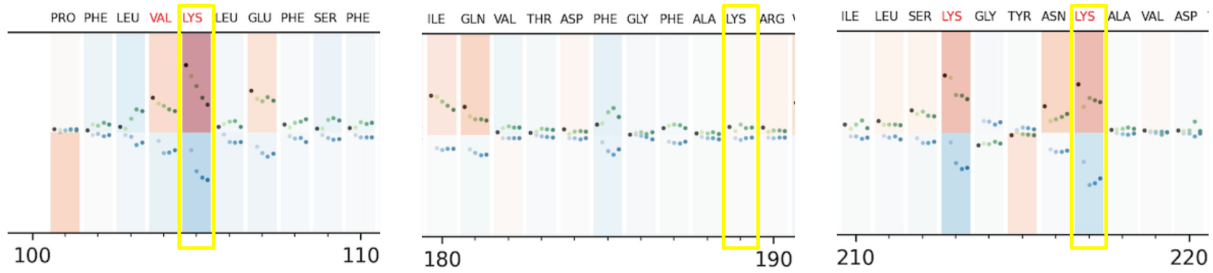


Figure 4.17: Residual SDAP (above the x-axis, increasing upward) and RMSD (below the x-axis, increasing downward) for residues 101-110, 180-190, and 210-220, with lysines 105, 189, and 217 highlighted with yellow bounding boxes. Residues which significantly fail the mirror test (large increase in RMSD, large decrease in SDAP) have their residue names highlighted in red.

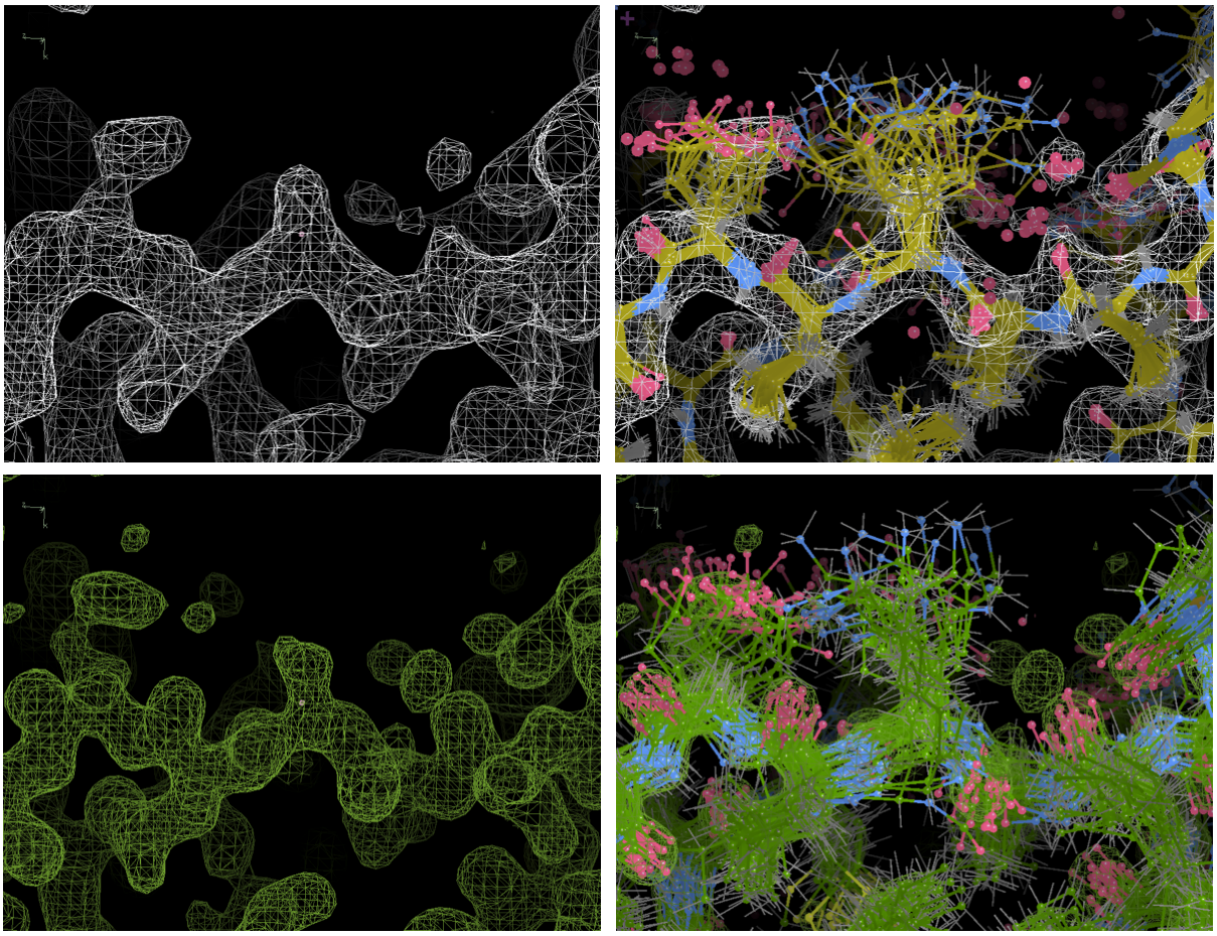


Figure 4.18: Lysine 105 — top left: experimental composite ($2F_O - F_C$)) map displayed as a 1σ isosurface; Top right: same as top left, withe experimentally-refined ensemble superimposed; bottom left: cMD density from the 0.2 kJmol^{-1}nm^{-2} restraint force constant simulation, displayed as a 1σ isosurface; bottom right: same as bottom left, with the final-frame reverse-propagated cMD ensemble from the 0.2 kJmol^{-1}nm^{-2} restraint force constant simulation superimposed.

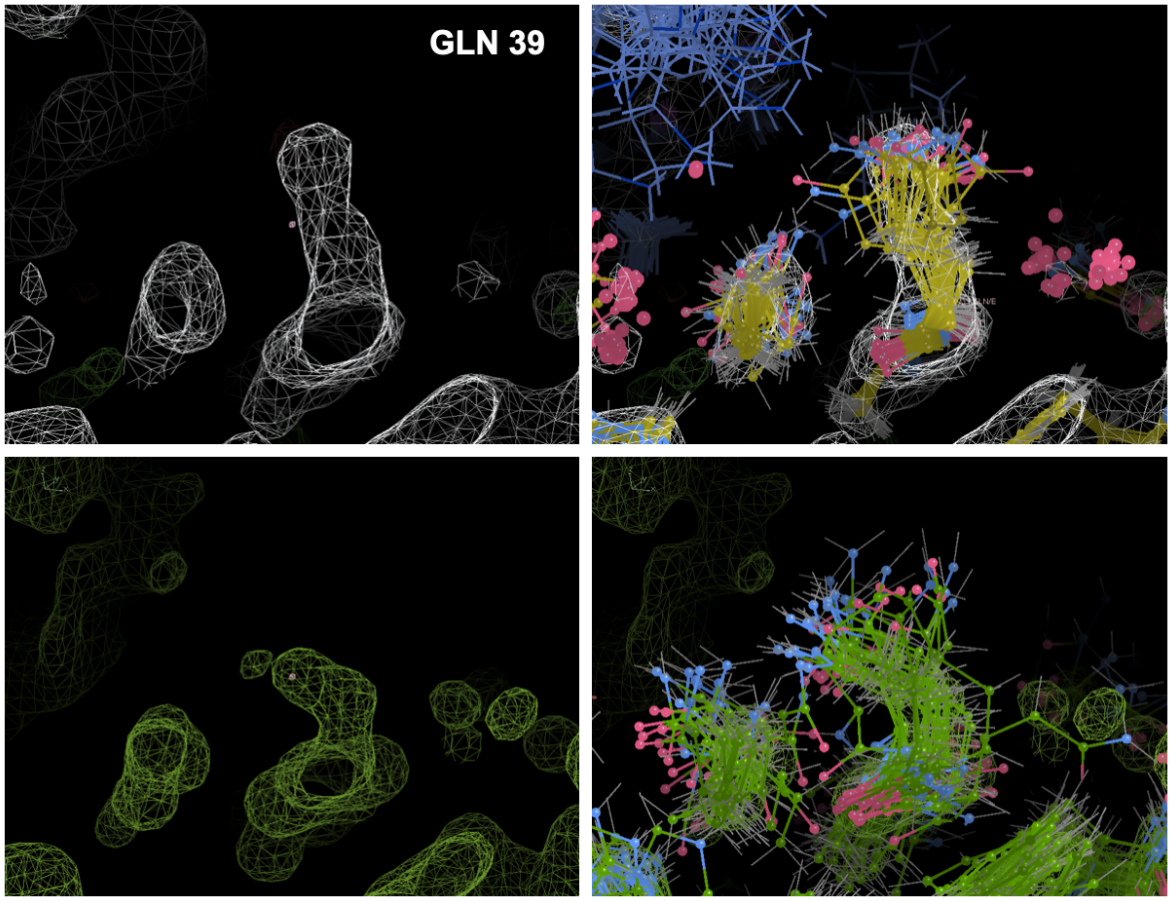


Figure 4.19: Glutamine 39 — top left: experimental composite ($2F_O - F_C$) and difference ($F_O - F_C$) map displayed as a 1σ and 3σ isosurface, respectively; Top right: same as top left, with the experimentally-refined ensemble superimposed, and symmetry-related copy of lysine 83 in blue; bottom left: cMD density from the 0.2 $\text{kJmol}^{-1}\text{nm}^{-2}$ restraint force constant simulation, displayed as a 1σ isosurface; bottom right: same as bottom left, with the final-frame reverse-propagated cMD ensemble from the 0.2 $\text{kJmol}^{-1}\text{nm}^{-2}$ restraint force constant simulation superimposed.

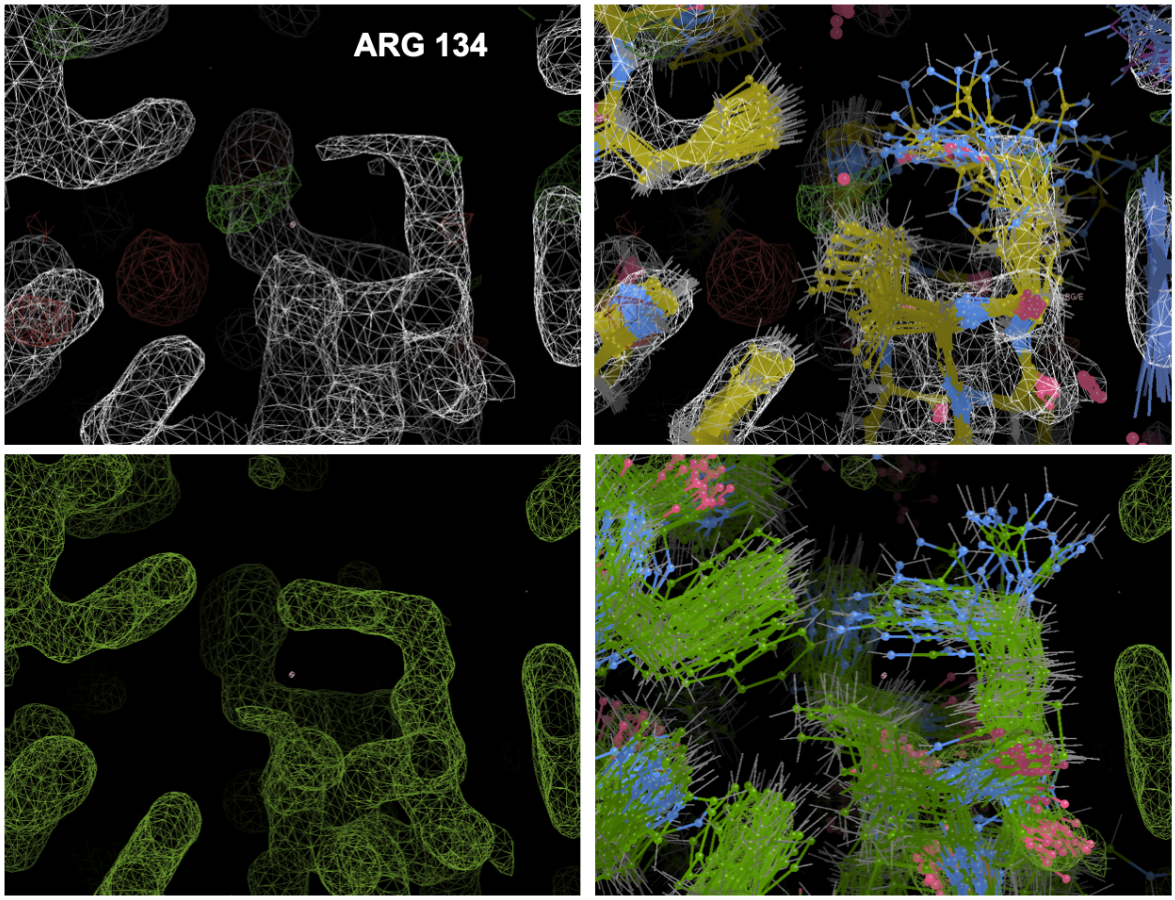


Figure 4.20: Arginine 134 — top left: experimental composite ($2F_O - F_C$) and difference ($F_O - F_C$) map displayed as a 1σ and 3σ isosurface, respectively; Top right: same as top left, with the experimentally-refined ensemble superimposed; bottom left: cMD density from the 0.2 $\mathrm{kJmol^{-1}nm^{-2}}$ restraint force constant simulation, displayed as a 1σ isosurface; bottom right: same as bottom left, with the final-frame reverse-propagated cMD ensemble from the 0.2 $\mathrm{kJmol^{-1}nm^{-2}}$ restraint force constant simulation superimposed.

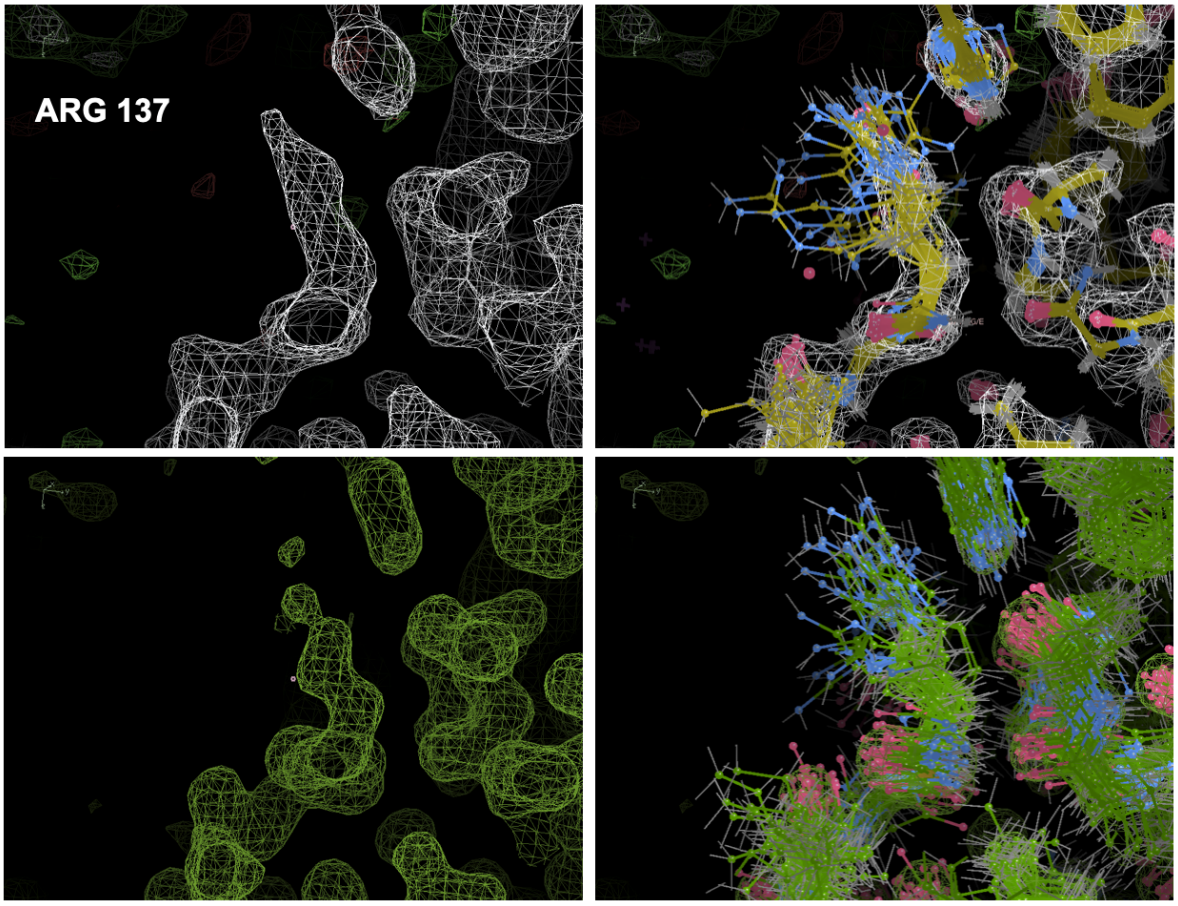


Figure 4.21: Arginine 137 — top left: experimental composite ($2F_O - F_C$) and difference ($F_O - F_C$) map displayed as a 1σ and 3σ isosurface, respectively; Top right: same as top left, with the experimentally-refined ensemble superimposed; bottom left: cMD density from the 0.2 $\mathrm{kJmol^{-1}nm^{-2}}$ restraint force constant simulation, displayed as a 1σ isosurface; bottom right: same as bottom left, with the final-frame reverse-propagated cMD ensemble from the 0.2 $\mathrm{kJmol^{-1}nm^{-2}}$ restraint force constant simulation superimposed.

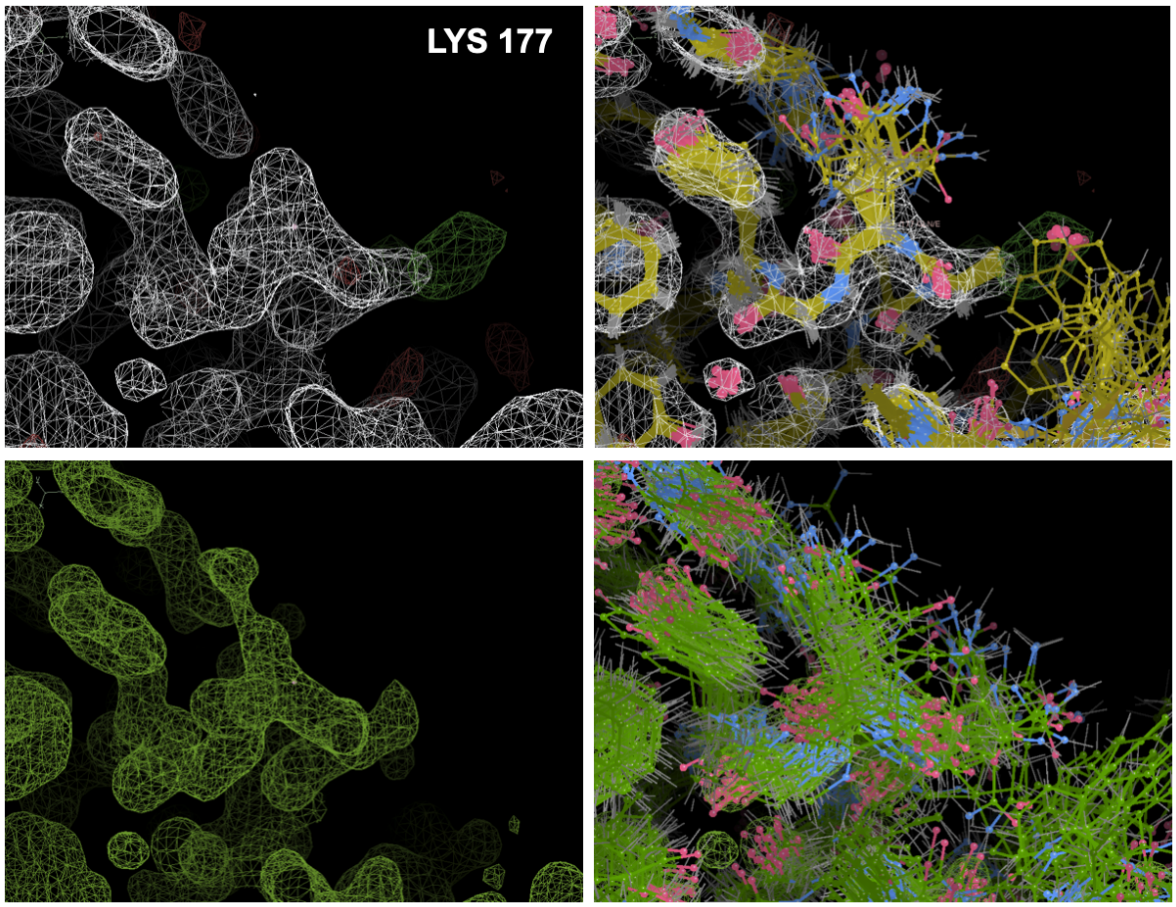


Figure 4.22: Lysine 177 — top left: experimental composite ($2F_O - F_C$) and difference ($F_O - F_C$) map displayed as a 1σ and 3σ isosurface, respectively; Top right: same as top left, with the experimentally-refined ensemble superimposed; bottom left: cMD density from the 0.2 $\mathrm{kJmol^{-1}nm^{-2}}$ restraint force constant simulation, displayed as a 1σ isosurface; bottom right: same as bottom left, with the final-frame reverse-propagated cMD ensemble from the 0.2 $\mathrm{kJmol^{-1}nm^{-2}}$ restraint force constant simulation superimposed.

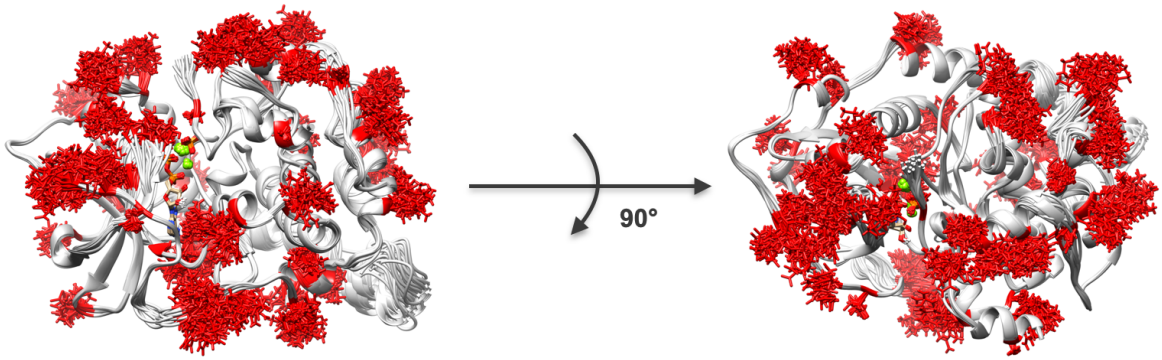


Figure 4.23: Experimentally-refined ensemble of PKA, with residues which fail the RMSD/SDAP "mirror" test highlighted in red.

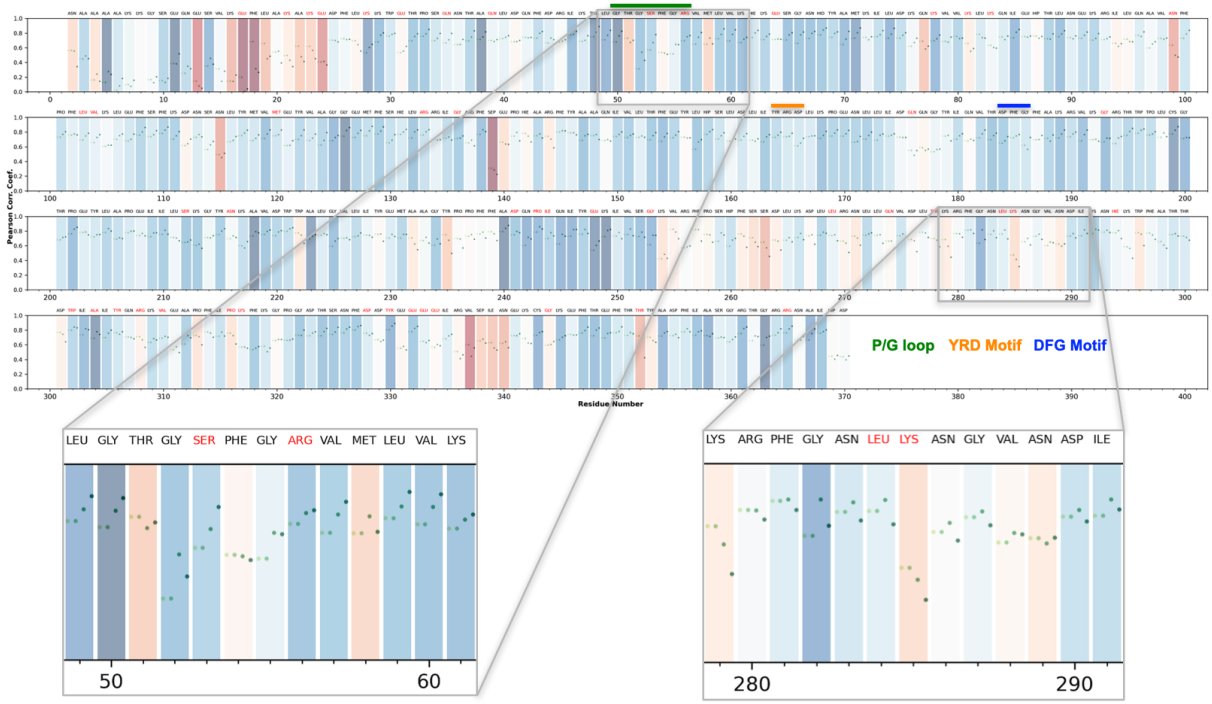


Figure 4.24: Pearson correlation coefficient (PCC) between the experimental and cMD-predicted density in a mask defined around the side chain ensemble, for each residue, for cMD simulations with restraint force constants of 200 (grey-green), 20 (grey green), 2 (medium green) and 0.2 (dark green) kJ $\cdot$ mol^{-1} $\cdot$ nm^{-2}. Colored backgrounds indicate trend in PCC with a net average decrease in correlation with relaxation of restraints colored in dark red, net zero trend in PCC in white, and net positive trend in PCC in dark blue.

Chapter 5

Future Directions in the Modelling of Crystallographic Structure and Dynamics

5.1 Diffuse Scattering

In the past decade or two, there has been renewed interest in the modelling of diffuse scattering in X-ray crystallography, as a method for probing correlated dynamics in protein crystals. Steady improvements in the modelling of conformational heterogeneity against Bragg data and correlated dynamics through diffuse scattering have been made possible thanks to better experimental methods (room temperature and time-resolved experiments), better modelling paradigms (ensemble and multiconformer refinement), better detectors (pixel array detectors, with high resolution, low point spread functions, and high signal to noise), and brighter, more energetic light sources, with the ability to produce femtosecond pulses (X-ray Free Electron Lasers, or XFELs)[106].

Improvements in the modelling of diffuse scattering have been slowed by the relative scarcity of high-quality data sets. Although the diffuse scattering accounts for about 50% of all scattered photons, the anisotropic component can be quite weak (about 10 times weaker than the isotropic), so to discriminate the signal from noise, care has to be taken to account for and either experimentally eliminate or (post-measurement) subtract away sources of background scattering from, e.g., the loop used to mount the crystal, the liquid the crystal is suspended in, and the scattering from the air. The work of Ayyer *et al.* (2016)[3] and Ando *et al.* (2020)[68] are the finest examples to date of careful data collection and processing, in addition to being some of the most exciting and groundbreaking work in this space. Ayyer *et al.* showed that the diffuse scattering (in their case, modelled as resulting from independent rigid body translational displacements) could be used to extend the resolution of the data set, and allow for model-free phasing (though Wall, Wolff, and Fraser present questions and concerns about this work in their perspective paper[106], in the section on "Phasing and resolution extension").

Ando *et al.* presented the first case in which the diffuse scattering could be used to distinguish between two roughly[1] equivalent EN models of internal motions which were both refined against the ADPs and the diffuse scattering and were equivalent in their correlation with the B-factors. In addition, they were able to show that a significant part of the agreement with the diffuse data comes from lattice-coupled motions, with only a comparatively small increase in correlation coming from the addition of internal motions. This type of finding would not possible with coarse-grained data, or data that is not carefully processed, as the lattice-coupled motions were refined against halos around the Bragg peaks, which require fine-grained data to analyze the power law dependence as a function of the distance from the center of the Bragg peak. They also took care to collect diffraction images from the crystal and the plastic capillary in which the crystal was housed, by simply translating the

[1]There was a small difference in the number of parameters between the rigid-body lattice-connected and internal-plus-lattice-connect elastic network models.

spindle arm at each angle at which the data was collected. They showed that background scattering has many independent features that vary as the spindle arm is rotated, so the rigorous collection of background scattering data for each reflection is important to achieve similar quality in the Bragg and diffuse data. Beam line scientists would do well to design their experiments, and process their data with the same level of care and sophistication. Additional tools will be required to apply the same techniques to crystals with higher levels of mosaicity, or more complex unit cells (the unit cell of triclinic lysozyme is P1, with one protein per unit cell, and their particular data set showed very small amounts of mosiacity compared to standard protein crystals).

Open source methods have been developed to aid in the collection of more high quality data sets: the `lunus` and `sematura` (`https://github.com/fraser-lab/diffuse_scattering`) software suites provide end-to-end methods for the isolation, processing, and analysis of diffuse scattering data from diffraction images (capable of processing thousands of images in minutes, on a small computing cluster). Even with a small number of suitable data sets to work with, there have been notable successes in extracting information from the diffuse scattering using a variety of models.

The liquid-like motions (LLM) model has been the most widely tested of these models, and (with the exception of the Ando *et al.* paper mentioned above) has reliably exhibited the highest correlations with the experimental anisotropic diffuse scattering data compared with other models. Its simplicity is both a strength and a weakness: it can be easily refined against the data with very little required in the way of computational resources, but it has not yet been leveraged to provide information about dynamics which are biologically or functionally illuminating. This may change in the near future: it is possible to reconstruct the variance-covariance matrix of atomic displacements using the RMS displacement and correlation length obtained through refinement (or correlation *lengths* if one is using the anisotropic LLM model), which can then be used to calculate the Hessian matrix and thus

the normal modes and eigenvalues for the dynamics of the protein implied by the model. New methods have very recently been developed[101] which move beyond assigning assigning the same covariance matrix of atomic displacements to every atom, and instead utilize the full Bragg scattering to capture information about the displacements of each atom independently (from the B-factors). Whether or not combined B-factor and LLM model is able to (a) better predict the anisotropic diffuse scattering, and/or (b) provide useful information about internal motions in proteins to the degree that more complicated elastic network models have is an open question. However, were both true, this updated LLM model would allow for simultaneous refinement of a structural and dynamical model against the Bragg and diffuse data with modest computational resources.

5.2 Crystalline Molecular Dynamics Simulations

5.2.1 Structural Modelling

The agreement between crystalline MD (cMD) models and diffuse scattering data is, in general, very good for the full diffuse scattering signal (isotropic and anisotropic), with correlation coefficients between modelled and experimental data greater than 0.9 (see [105], [68], and our work, [110]). It is often left out of discussions of the merits of various models that the cMD model is still, to this day, the only widely-studied model capable of predicting both the isotropic and anisotropic components of the diffuse scattering, as it is the only model which incorporates interactions between the protein and the solvent. However, this remark is left out for an understandable reason: the anisotropic component of the diffuse scattering comes almost entirely from the correlated dynamics of the protein alone, so this is the data we're interested in if we're concerned primarily with the modelling of protein dynamics. The other models (LLM, RBM, EN/NM) benefit from the ability to tune their

parameters against experimental data (either Bragg or diffuse), while the MD model does not.

The unrestrained cMD model does not appear to be an exceptionally good model of the average structure of protein crystals (Bragg data) at the moment. Ando *et al.* [68] showed that, for cMD simulations from 1 to over 300 unit cells, the correlation coefficient between the experimental and predicted Bragg intensities decreases dramatically with diffraction resolution (and doesn't improve all that much with the addition of many unit cells) indicating that unrestrained cMD systems are poor models for average structure at high resolution. This is unsurprising, given the prevalence of structural "drift" observed in many different unrestrained cMD simulations (remarked upon in Chapter 2). We found similar results: for *Staph. nuclease*, the heavy-atom RMSD to the crystal structure supercell was 2.5-3 Å, with the full supercell translationally fit to the starting structure. For PKA, we found that protein structures refined against the cMD-simulated structure factors increased in RMSD to the experimentally-refined structure, with a heavy atom RMSD increasing from 0.84 to 1.08 Å as restraint constants are decreased from 200 to 0.2 $\mathrm{kJ \cdot mol^{-1} \cdot nm^{-2}}$. However, the modelling of the B-factors by unrestrained simulations appears to be quite accurate (e.g. see [46], [100], and the previous chapter), and agreement with other dynamical data (such as chemical shifts from solid state NMR data) can be quite good as well (as remarked on in Chapter 2).

It appears, therefore, that unrestrained MD models are well tuned to reproduce the small scale harmonic motions that side chains and larger structural features undergo, due to thermal motion and short-range interactions with nearby residues, but the force field leads the structure as a whole to converge to an inaccurate conformation. The most promising course for force field improvement seems to be the tuning of parameters for protein-protein and protein-solvent interactions, as the force fields' reproduction of secondary structure elements and non-solvent-exposed side chains positions (for standard residues) appears to be fairly

accurate. In a previous chapter, we discussed how important the choice of water model can be in the prediction of solvation thermodynamics, and some attempts at developing a variant of the AMBER force field specifically designed for protein-protein association. Polarizable force fields or quantum mechanical treatment may be useful in the modelling of these specific interactions as well, though there is quite a bit of work to be done before either will be up to the task of simulating an entire protein (let alone many unit cells of them) for, say, tens to hundreds of nanoseconds (the time scales required to predict many biologically important motions). Quantum mechanical treatment of isolated portions of a cMD simulation may be possible in the near future, but only on short time-scales. However, simulations of this kind at smaller time scales may help to better understand what is currently missing from the gold-standard MD force fields.

In the meantime, *restrained* cMD simulations provide useful information about protein crystalline systems, while staying true to the experimentally-refined crystal structure. We showed in the previous chapter that, while the B-factors of a strongly-restrained cMD simulation may be severely underestimated due to the restriction of motion provided by the restraints, the overall *pattern* in the B-factors is well modelled at all restraint constants, with the Pearson correlation coefficient to the experimental B-factors staying very high (CC$\approx$0.94) and constant for simulations with restraint force constants spanning four orders of magnitude. What's more, strongly restrained cMD simulations do a remarkably good job at reproducing the number and position of experimentally-refined crystallographic waters. In fact, the restraints do not have to be *that* strong: the restraint force constant can be an order of magnitude weaker than in previous studies of the same kind (on a different system[102]), and the precision and recall of crystallographic waters to within an angstrom is about the same. In special cases, the differences between restrained cMD systems and the crystal structure (or experimentally-refined ensemble) can be large enough that errors in protonation state or obvious false parametrizations for non-standard residues could be observed even in the most strongly restrained simulations.

cMD simulations may also be useful for producing or improving models of structural heterogeneiety. As we showed in the previous chapter, the residues with the largest changes in the conformational ensemble from the experimentally-refined conformational ensemble were mostly solvent-exposed or crystal-contact residues. These residues were shown, in some cases, to have direct interactions with side chains from neighboring proteins in the crystal lattice. The MD simulation used to produce the experimentally-refined ensemble uses bulk solvent modelling (no explicit waters), and does not include the crystal context. Ensembles generated for refinement against experimental data could, therefore, be improve by moving from a sMD to cMD model (even a unit cell simulation would probably be sufficient for these purposes).

5.2.2 Correlated disorder modelling

Even if there are simpler parametric models (such as the LLM or EN/NM model) that fit the anisotropic diffuse scattering better than the cMD model does (assuming the results of Ando *et al.* are reproduced in other systems), it will still be of interest to both the crystallography community and the MD modelling community to work to improve the agreement of the cMD model with crystallographic data. cMD simulations are unique in that they provide a way to directly compare to experimental data that is (in some cases) incredibly precise and allows for analysis of the effect of small changes in the model (protonation state changes, different ligands, or side chain mutations) or the parameters of the force field on the (ensemble) structure and dynamics, on a scale from the entire protein to the scale of a particular side chain. In many cases, protein crystal are highly solvated (with solvent content greater than forty or fifty percent by volume) so it is not unreasonable to expect that improving agreement with crystallographic data with respect to protein-solvent interactions in the crystal context will benefit modelling in the solution-state context as well. The same is true for protein-protein interactions: indeed cMD simulations provide the ideal context in which to study

protein-protein interactions; the systems is a large, dense collection of closely interacting proteins.

5.3 Conclusion

I believe the work presented in this dissertation shows that cMD simulations can shed light on a wide variety of complex phenomena in crystallographic modelling of structure and dynamics. cMD models have been used to provide context to long-standing debates about the nature and source of diffuse scattering, and suggest future directions in the modelling of correlated disorder.

In Chapter 3, we discussed work which showed that a LLM-like model provides an excellent fit to the global C-α pair displacements, but a combination of RBM- and LLM-like motions were required to fit the data describing the displacement of intra-protein C-α atom pairs. We also suggested that different models of disorder may capture features sampled in different areas of reciprocal space, suggesting that sampling further from the Bragg peaks will preferentially select models with correlations on shorter length scales, and sampling closer to the Bragg peaks will select for models with longer range correlated motions. These findings were, in some small way, validated by the work of Ando *et al.* who found that a rigid body elastic network model fits the haloes around Bragg peaks, but a robust fit to the full anisotropic diffuse scattering data, with more data further from the Bragg peaks, required the addition of an elastic network model for internal protein motions.

In Chapter 4, we showed that cMD simulations can also be used to improve our understanding of ensemble and/or multi-conformer modelling, side chain disorder, and ordered water networks across a protein (and in particular, around the active site) to help improve crystallographic models, and potentially expand our understanding of the structure and function of

biologically-important proteins. Computational resources will continue to become cheaper and faster, and simulation software will continue to become more flexible across broader ranges of computing systems (from laptops to high-performance exa-scale computing clusters) bringing down the computational cost and complexity of simulating large, complex system at the atomic scale.

The work presented here is but a fraction at the work still left to be done. Crystallography will continue to be a powerful tool in the hands of scientists all over the world. It may become more powerful still: some of the work presented here suggests that careful crystallographic experiments should be able to provide models not just of the average structure of proteins but also their correlated dynamics (a dream in the field for quite some time). However, as things stand, the "R-factor gap" remains unexplained: the source of the difference in accuracy between small-molecule and macromolecular crystallographic models has not been identified. Crystalline molecular dynamics simulations of both Bragg and diffuse diffraction data may prove indispensable in the search for the source of this discrepancy, as it seems likely that the modelling of structural heterogeneity, correlated dynamics, and ordered water networks play a central role. cMD simulations are the finest tool currently available for the study of both, using the same model.

Bibliography

[1] Irem Altan, Diana Fusco, Pavel V Afonine, and Patrick Charbonneau. Learning about biomolecular solvation from water in protein crystals. *The Journal of Physical Chemistry B*, 122(9):2475–2486, 2018.

[2] A. Arvai. Adxv—a program to display x-ray diffraction images, 2012.

[3] Kartik Ayyer, Oleksandr M Yefanov, Dominik Oberthür, Shatabdi Roy-Chowdhury, Lorenzo Galli, Valerio Mariani, Shibom Basu, Jesse Coe, Chelsie E Conrad, Raimund Fromme, et al. Macromolecular diffractive imaging using imperfect crystals. *Nature*, 530(7589):202–206, 2016.

[4] Jean-Pierre Benoit and Jean Doucet. Diffuse scattering in protein crystallography. *Quarterly reviews of biophysics*, 28(2):131–169, 1995.

[5] Herman JC Berendsen, David van der Spoel, and Rudi van Drunen. Gromacs: a message-passing parallel molecular dynamics implementation. *Computer physics communications*, 91(1-3):43–56, 1995.

[6] Dimitri Boylan and George N Phillips Jr. Motions of tropomyosin: characterization of anisotropic motions and coupled displacements in crystals. *Biophysical journal*, 49(1):76, 1986.

[7] B Tom Burnley, Pavel V Afonine, Paul D Adams, and Piet Gros. Modelling dynamics in protein crystal structures by ensemble refinement. *Elife*, 1:e00311, 2012.

[8] D.A. Case, H.M. Aktulga, K. Belfon, I.Y. Ben-Shalom, S.R. Brozell, D.S. Cerutti, T.E. Cheatham, III, V.W.D. Cruzeiro, T.A. Darden, R.E. Duke, G. Giambasu, M.K. Gilson, H. Gohlke, A.W. Goetz, R. Harris, S. Izadi, S.A. Izmailov, C. Jin, K. Kasavajhala, M.C. Kaymak, E. King, A. Kovalenko, T. Kurtzman, T.S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, M. Machado, V. Man, M. Manathunga, K.M. Merz, Y. Miao, O. Mikhailovskii, G. Monard, H. Nguyen, K.A. O'Hearn, A. Onufriev, F. Pan, S. Pantano, R. Qi, A. Rahnamoun, D.R. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, J. Shen, C.L. Simmerling, N.R. Skrynnikov, J. Smith, J. Swails, R.C. Walker, J. Wang, H. Wei, R.M. Wolf, X. Wu, Y. Xue, D.M. York, S. Zhao, , and P.A. Kollman. Amber 2021. 2021.

[9] DLD Caspar, J Clarage, DM Salunke, and M Clarage. Liquid-like movements in crystalline insulin. *Nature*, 332(6165):659–662, 1988.

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

[11] David S Cerutti, Isolde Le Trong, Ronald E Stenkamp, and Terry P Lybrand. Simulations of a protein crystal: Explicit treatment of crystallization conditions links theory and experiment in the streptavidin- biotin complex. *Biochemistry*, 47(46):12065–12077, 2008.

[12] Susan Chacko and GN Phillips Jr. Diffuse x-ray scattering from tropomyosin crystals. *Biophysical journal*, 61(5):1256–1266, 1992.

[13] Yuhui Cheng, Yingkai Zhang, and J Andrew McCammon. How does the camp-dependent protein kinase catalyze the phosphorylation reaction: an ab initio qm/mm study. *Journal of the American Chemical Society*, 127(5):1553–1562, 2005.

[14] James B Clarage, Michael S Clarage, Walter C Phillips, Robert M Sweet, and Donald LD Caspar. Correlations of atomic movements in lysozyme crystals. *Proteins: Structure, Function, and Bioinformatics*, 12(2):145–157, 1992.

[15] James B Clarage, Tod Romo, B Kim Andrews, B Montgomery Pettitt, and George N Phillips. A sampling problem in molecular dynamics simulations of macromolecules. *Proceedings of the National Academy of Sciences*, 92(8):3288–3292, 1995.

[16] Zoe Cournia, Bryce K Allen, Thijs Beuming, David A Pearlman, Brian K Radak, and Woody Sherman. Rigorous free energy simulations in virtual screening. *Journal of Chemical Information and Modeling*, 60(9):4153–4169, 2020.

[17] Gavin E Crooks. Field guide to continuous probability distributions. *Berkeley Institute for Theoretical Sciences, Berkeley*, 2019.

[18] George D Dalton and William L Dewey. Protein kinase inhibitor peptide (pki): a family of endogenous neuropeptides that modulate neuronal camp-dependent protein kinase function. *Neuropeptides*, 40(1):23–34, 2006.

[19] T De Klijn, AMM Schreurs, and LMJ Kroon-Batenburg. Rigid-body motion is the main source of diffuse scattering in protein crystallography. *IUCrJ*, 6(2):277–289, 2019.

[20] R Diamond. On the use of normal modes in thermal parameter refinement: theory and application to the bovine pancreatic trypsin inhibitor. *Acta Crystallographica Section A: Foundations of Crystallography*, 46(6):425–435, 1990.

[21] Todd J Dolinsky, Jens E Nielsen, J Andrew McCammon, and Nathan A Baker. Pdb2pqr: an automated pipeline for the setup of poisson–boltzmann electrostatics calculations. *Nucleic acids research*, 32(suppl_2):W665–W667, 2004.

[22] J Doucet and JP Benoit. Molecular dynamics studied by analysis of the x-ray diffuse scattering from lysozyme crystals. *Nature*, 325(6105):643–646, 1987.

[23] Paul Emsley, Bernhard Lohkamp, William G. Scott, and Kevin Cowtan. Features and development of coot. *Acta Crystallographica Section D - Biological Crystallography*, 66:486–501, 2010.

[24] Ph Faure, A Micu, D Perahia, J Doucet, Jeremy C Smith, and JP Benoit. Correlated intramolecular motions and diffuse x–ray scattering in lysozyme. *Nature structural biology*, 1(2):124–128, 1994.

[25] Ph Faure, J Pérez, J Doucet, and JP Benoit. X-ray diffuse scattering and molecular dynamics in proteins. *Le Journal de Physique IV*, 4(C9):C9–293, 1994.

[26] James S Fraser, Henry van den Bedem, Avi J Samelson, P Therese Lang, James M Holton, Nathaniel Echols, and Tom Alber. Accessing protein conformational ensembles using room-temperature x-ray crystallography. *Proceedings of the National Academy of Sciences*, 108(39):16247–16252, 2011.

[27] Hans Frauenfelder. The debye-waller factor: From villain to hero in protein crystallography. *International journal of quantum chemistry*, 35(6):711–715, 1989.

[28] Sylvie Furois-Corbin, Jeremy C Smith, and Gerald R Kneller. Picosecond timescale rigid-helix and side-chain motions in deoxymyoglobin. *Proteins: Structure, Function, and Bioinformatics*, 16(2):141–154, 1993.

[29] Sadra Kashef Ol Gheta, Shuzhe Wang, William E Acree Jr, and Philippe Hunenberger. Evaluation of nine condensed-phase force fields of the gromos, charmm, opls, amber, and openff families against experimental cross-solvation free energies. *Physical Chemistry Chemical Physics*, 2021.

[30] Nobuhiro Go, Tosiyuki Noguti, and Testuo Nishikawa. Dynamics of a small globular protein in terms of low-frequency vibrational modes. *Proceedings of the National Academy of Sciences*, 80(12):3696–3700, 1983.

[31] DW Green, Vernon Martin Ingram, and Max Ferdinand Perutz. The structure of haemoglobin-iv. sign determination by the isomorphous replacement method. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 225(1162):287–307, 1954.

[32] A Guinier. X-ray diffraction in crystals, imperfect crystals and amorphous bodies. *San FranciscoL WH Freeman and Co*, 1963.

[33] Jingjing Guo and Huan-Xiang Zhou. Protein allostery and conformational dynamics. *Chemical reviews*, 116(11):6503–6515, 2016.

[34] J Mitchell Guss, Ethan A Merritt, R Paul Phizackerley, Britt Hedman, Mitsuo Murata, Keith O Hodgson, and Hans C Freeman. Phase determination by multiple-wavelength x-ray diffraction: crystal structure of a basic" blue" copper protein from cucumbers. *Science*, 241(4867):806–811, 1988.

[35] Charles R Harris, K Jarrod Millman, Stéfan J van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J Smith, et al. Array programming with numpy. *Nature*, 585(7825):357–362, 2020.

[36] Niel M Henriksen and Michael K Gilson. Evaluating force field performance in thermodynamic calculations of cyclodextrin host–guest binding: Water models, partial charges, and host force field parameters. *Journal of chemical theory and computation*, 13(9):4253–4269, 2017.

[37] Stéphanie Héry, Daniel Genest, and Jeremy C Smith. X-ray diffuse scattering and rigid-body motion in crystalline lysozyme probed by molecular dynamics simulation. *Journal of molecular biology*, 279(1):303–319, 1998.

[38] Berk Hess and Nico FA van der Vegt. Hydration thermodynamic properties of amino acid analogues: a systematic comparison of biomolecular force fields and water models. *The journal of physical chemistry B*, 110(35):17616–17626, 2006.

[39] Junichi Higo and Masayoshi Nakasako. Hydration structure of human lysozyme investigated by molecular dynamics simulation and cryogenic x-ray crystal structure analyses: On the correlation between crystal water sites, solvent density, and solvent dipole. *Journal of computational chemistry*, 23(14):1323–1336, 2002.

[40] James M Holton, Scott Classen, Kenneth A Frankel, and John A Tainer. The r-factor gap in macromolecular crystallography: an untapped potential for insights on accurate structures. *The FEBS journal*, 281(18):4046–4060, 2014.

[41] Tokio Horiuchi and Nobuhiro Gō. Projection of monte carlo and molecular dynamics trajectories onto the normal mode axes: human lysozyme. *Proteins: Structure, Function, and Bioinformatics*, 10(2):106–116, 1991.

[42] Zhongqiao Hu and Jianwen Jiang. Assessment of biomolecular force fields for molecular dynamics simulations in a protein crystal. *Journal of computational chemistry*, 31(2):371–380, 2010.

[43] Toshiko Ichiye and Martin Karplus. Collective motions in proteins: a covariance analysis of atomic fluctuations in molecular dynamics and normal mode simulations. *Proteins: Structure, Function, and Bioinformatics*, 11(3):205–217, 1991.

[44] R.W. James. *The Optical Principles of the Diffraction of X-rays*. G. Bell and Sons, London, 1994.

[45] P.A. Janowski, D.S. Cerutti, J. Holton, and D.A. Case. Peptide crystal simulations reveal hidden dynamics. *J. Am. Chem. Soc.*, 135:7938–7948, 2013.

[46] Pawel A Janowski, Chunmei Liu, Jason Deckman, and David A Case. Molecular dynamics simulation of triclinic lysozyme in a crystal lattice. *Protein Science*, 25(1):87–102, 2016.

[47] Pawel A Janowski, Chunmei Liu, Jason Deckman, and David A Case. Molecular dynamics simulation of triclinic lysozyme in a crystal lattice. *Protein Science*, 25(1):87–102, 2016.

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

[49] Daniel A Keedy, James S Fraser, and Henry Van Den Bedem. Exposing hidden alternative backbone conformations in x-ray crystallography using qfit. *PLoS Comput Biol*, 11(10):e1004507, 2015.

[50] Akinori Kidera and Nobuhiro Go. Refinement of protein dynamic structure: normal mode refinement. *Proceedings of the National Academy of Sciences*, 87(10):3718–3722, 1990.

[51] Akinori Kidera, Masaaki Matsushima, and Nobuhiro Gō. Dynamic structure of human lysozyme derived from x-ray crystallography: normal mode refinement. *Biophysical chemistry*, 50(1-2):25–31, 1994.

[52] Elmar Krieger, Tom Darden, Sander B Nabuurs, Alexei Finkelstein, and Gert Vriend. Making optimal use of empirical energy functions: force-field parameterization in crystal space. *Proteins: Structure, Function, and Bioinformatics*, 57(4):678–683, 2004.

[53] John Kuriyan and William I Weis. Rigid protein motion as a model for crystallographic temperature factors. *Proceedings of the National Academy of Sciences*, 88(7):2773–2777, 1991.

[54] Antonija Kuzmanic, Navraj S Pannu, and Bojan Zagrovic. X-ray refinement significantly underestimates the level of microscopic heterogeneity in biomolecular crystals. *Nature communications*, 5(1):1–10, 2014.

[55] Elena J Levin, Dmitry A Kondrashov, Gary E Wesenberg, and George N Phillips Jr. Ensemble refinement of protein crystal structures: validation and application. *Structure*, 15(9):1040–1052, 2007.

[56] Michael Levitt, Christian Sander, and Peter S Stern. Protein normal-mode dynamics: trypsin inhibitor, crambin, ribonuclease and lysozyme. *Journal of molecular biology*, 181(3):423–447, 1985.

[57] Dorothee Liebschner, Pavel V Afonine, Matthew L Baker, Gábor Bunkóczi, Vincent B Chen, Tristan I Croll, Bradley Hintze, L-W Hung, Swati Jain, Airlie J McCoy, et al. Macromolecular structure determination using x-rays, neutrons and electrons: recent developments in phenix. *Acta Crystallographica Section D: Structural Biology*, 75(10):861–877, 2019.

[58] Dorothee Liebschner, Pavel V Afonine, Matthew L Baker, Gábor Bunkóczi, Vincent B Chen, Tristan I Croll, Bradley Hintze, L-W Hung, Swati Jain, Airlie J McCoy, et al.

Macromolecular structure determination using x-rays, neutrons and electrons: recent developments in phenix. *Acta Crystallographica Section D: Structural Biology*, 75(10):861–877, 2019.

[59] Dorothee Liebschner, Pavel V Afonine, Nigel W Moriarty, Billy K Poon, Oleg V Sobolev, Thomas C Terwilliger, and Paul D Adams. Polder maps: improving omit maps by excluding bulk solvent. *Acta Crystallographica Section D: Structural Biology*, 73(2):148–157, 2017.

[60] Kresten Lindorff-Larsen, Paul Maragakis, Stefano Piana, Michael P Eastwood, Ron O Dror, and David E Shaw. Systematic validation of protein force fields against experimental data. *PloS one*, 7(2):e32131, 2012.

[61] AJ Malkin, Yu G Kuznetsov, and A McPherson. Defect structure of macromolecular crystals. *Journal of Structural Biology*, 117(2):124–137, 1996.

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

[63] Robert T. McGibbon, Kyle A. Beauchamp, Matthew P. Harrigan, Christoph Klein, Jason M. Swails, Carlos X. Hernández, Christian R. Schwantes, Lee-Ping Wang, Thomas J. Lane, and Vijay S. Pande. Mdtraj: A modern open library for the analysis of molecular dynamics trajectories. *Biophysical Journal*, 109(8):1528 – 1532, 2015.

[64] Kristin L Meagher, Luke T Redman, and Heather A Carlson. Development of polyphosphate parameters for use with the amber force field. *Journal of computational chemistry*, 24(9):1016–1025, 2003.

[65] Lars Meinhold, Franci Merzel, and Jeremy C Smith. Lattice dynamics of a protein crystal. *Physical review letters*, 99(13):138101, 2007.

[66] Lars Meinhold and Jeremy C Smith. Correlated dynamics determining x-ray diffuse scattering from a crystalline protein revealed by molecular dynamics simulation. *Physical review letters*, 95(21):218103, 2005.

[67] Lars Meinhold and Jeremy C Smith. Fluctuations and correlations in crystalline protein dynamics: a simulation analysis of staphylococcal nuclease. *Biophysical journal*, 88(4):2554–2563, 2005.

[68] Steve P Meisburger, David A Case, and Nozomi Ando. Diffuse x-ray scattering from correlated motions in a protein crystal. *Nature communications*, 11(1):1–13, 2020.

[69] Steve P Meisburger, William C Thomas, Maxwell B Watkins, and Nozomi Ando. X-ray scattering studies of protein structural dynamics. *Chemical reviews*, 117(12):7615–7672, 2017.

[70] Kenji Mizuguchi, Akinori Kidera, and Nobuhiro Gō. Collective motions in proteins investigated by x-ray diffuse scattering. *Proteins: Structure, Function, and Bioinformatics*, 18(1):34–48, 1994.

[71] Pramod C Nair and John O Miners. Molecular dynamics simulations: from structure function relationships to drug discovery. *In silico pharmacology*, 2(1):1–4, 2014.

[72] Mats HM Olsson, Chresten R Søndergaard, Michal Rostkowski, and Jan H Jensen. Propka3: consistent treatment of internal and surface residues in empirical p k a predictions. *Journal of chemical theory and computation*, 7(2):525–537, 2011.

[73] Albert C. Pan, Daniel Jacobson, Konstantin Yatsenko, Duluxan Sritharan, Thomas M. Weinreich, and David E. Shaw. Atomic-level characterization of protein–protein association. *Proceedings of the National Academy of Sciences*, 116(10):4244–4249, 2019.

[74] Ariana Peck, Frédéric Poitevin, and Thomas J Lane. Intermolecular correlations are necessary to explain diffuse scattering from protein crystals. *IUCrJ*, 5(2):211–222, 2018.

[75] J Pérez, Ph Faure, and J-P Benoit. Molecular rigid-body displacements in a tetragonal lysozyme crystal confirmed by x-ray diffuse scattering. *Acta Crystallographica Section D: Biological Crystallography*, 52(4):722–729, 1996.

[76] Drazen Petrov and Bojan Zagrovic. Are current atomistic force fields accurate enough to study proteins in crowded environments? *PLoS Comput Biol*, 10(5):e1003638, 2014.

[77] Eric F Pettersen, Thomas D Goddard, Conrad C Huang, Gregory S Couch, Daniel M Greenblatt, Elaine C Meng, and Thomas E Ferrin. Ucsf chimera—a visualization system for exploratory research and analysis. *Journal of computational chemistry*, 25(13):1605–1612, 2004.

[78] Yury S Polikanov and Peter B Moore. Acoustic vibrations contribute to the diffuse scatter produced by ribosome crystals. *Acta Crystallographica Section D: Biological Crystallography*, 71(10):2021–2031, 2015.

[79] Jay W Ponder and David A Case. Force fields for protein simulations. *Advances in protein chemistry*, 66:27–85, 2003.

[80] Billy K Poon, Xiaorui Chen, Mingyang Lu, Nand K Vyas, Florante A Quiocho, Qinghua Wang, and Jianpeng Ma. Normal mode refinement of anisotropic thermal parameters for a supramolecular complex at 3.42-å crystallographic resolution. *Proceedings of the National Academy of Sciences*, 104(19):7869–7874, 2007.

[81] AJ Rader, Chakra Chennubhotla, Lee-Wei Yang, Ivet Bahar, and Q Cui. The gaussian network model: Theory and applications. *Normal mode analysis: Theory and applications to biological and chemical systems*, 9:41–64, 2006.

[82] Demian Riccardi, Qiang Cui, and George N Phillips Jr. Evaluating elastic network models of crystalline biological molecules with temperature factors, correlated motions, and diffuse x-ray scattering. *Biophysical journal*, 99(8):2616–2625, 2010.

[83] Paul Robustelli, Stefano Piana, and David E Shaw. Developing a molecular dynamics force field for both folded and disordered protein states. *Proceedings of the National Academy of Sciences*, 115(21):E4758–E4766, 2018.

[84] Verner Schomaker and KN Trueblood. On the rigid-body motion of molecules in crystals. *Acta Crystallographica Section B: Structural Crystallography and Crystal Chemistry*, 24(1):63–76, 1968.

[85] Piotr Setny and Marta D Wiśniewska. Water-mediated conformational preselection mechanism in substrate binding cooperativity to protein kinase a. *Proceedings of the National Academy of Sciences*, 115(15):3852–3857, 2018.

[86] Michael R Shirts, Christoph Klein, Jason M Swails, Jian Yin, Michael K Gilson, David L Mobley, David A Case, and Ellen D Zhong. Lessons learned from comparing molecular dynamics engines on the sampl5 dataset. *Journal of computer-aided molecular design*, 31(1):147–161, 2017.

[87] Michael R Shirts and Vijay S Pande. Solvation free energies of amino acid side chain analogs for common molecular mechanics water models. *The Journal of chemical physics*, 122(13):134508, 2005.

[88] Janet L Smith, Wayne A Hendrickson, Richard B Honzatko, and Steven Sheriff. Structural heterogeneity in protein crystals. *Biochemistry*, 25(18):5018–5027, 1986.

[89] EH Snell, S Weisgerber, JR Helliwell, E Weckert, K Hölzer, and K Schroer. Improvements in lysozyme protein crystal perfection through microgravity growth. *Acta Crystallographica Section D: Biological Crystallography*, 51(6):1099–1102, 1995.

[90] Chresten R Søndergaard, Mats HM Olsson, Michał Rostkowski, and Jan H Jensen. Improved treatment of ligands and coupling effects in empirical calculation and rationalization of p k a values. *Journal of chemical theory and computation*, 7(7):2284–2295, 2011.

[91] U Stocker, K Spiegel, and WF van Gunsteren. On the similarity of properties in solution or in the crystalline state: a molecular dynamics study of hen lysozyme. *Journal of biomolecular NMR*, 18(1):1–12, 2000.

[92] John J Tanner, Paul E Smith, and Kurt L Krause. Molecular dynamics simulations and rigid body (tls) analysis of aspartate carbamoyltransferase: evidence for an uncoupled r state. *Protein Science*, 2(6):927–935, 1993.

[93] Susan S Taylor and Alexandr P Kornev. Protein kinases: evolution of dynamic regulatory proteins. *Trends in biochemical sciences*, 36(2):65–77, 2011.

[94] Andrew H Van Benschoten, Pavel V Afonine, Thomas C Terwilliger, Michael E Wall, Colin J Jackson, Nicholas K Sauter, Paul D Adams, Alexandre Urzhumtsev, and James S Fraser. Predicting x-ray diffuse scattering from translation–libration–screw structural ensembles. *Acta Crystallographica Section D: Biological Crystallography*, 71(8):1657–1667, 2015.

[95] Andrew H Van Benschoten, Lin Liu, Ana Gonzalez, Aaron S Brewster, Nicholas K Sauter, James S Fraser, and Michael E Wall. Measuring and modeling diffuse scattering in protein x-ray crystallography. *Proceedings of the National Academy of Sciences*, 113(15):4069–4074, 2016.

[96] Henry Van Den Bedem, Gira Bhabha, Kun Yang, Peter E Wright, and James S Fraser. Automated identification of functional dynamic contact networks from x-ray crystallography. *Nature methods*, 10(9):896–902, 2013.

[97] Gydo Cp Van Zundert, Brandi M Hudson, Saulo HP de Oliveira, Daniel A Keedy, Rasmus Fonseca, Amelie Heliou, Pooja Suresh, Kenneth Borrelli, Tyler Day, James S Fraser, et al. qfit-ligand reveals widespread conformational heterogeneity of drug-like molecules in x-ray electron density maps. *Journal of medicinal chemistry*, 61(24):11183–11198, 2018.

[98] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

[99] Michael E Wall. *Diffuse features in X-ray diffraction from protein crystals*. PhD thesis, Princeton University Princeton, NJ, 1996.

[100] Michael E Wall. Internal protein motions in molecular-dynamics simulations of bragg and diffuse x-ray scattering. *IUCrJ*, 5(2):172–181, 2018.

[101] Michael E. Wall. Liquid-like motions model of diffuse scattering with individual atomic b factors. *LA-UR-2122113*, 2021.

[102] Michael E Wall, Gaetano Calabró, Christopher I Bayly, David L Mobley, and Gregory L Warren. Biomolecular solvation structure revealed by molecular dynamics simulations. *Journal of the American Chemical Society*, 141(11):4711–4720, 2019.

[103] Michael E Wall, James B Clarage, and George N Phillips Jr. Motions of calmodulin characterized using both bragg and diffuse x-ray scattering. *Structure*, 5(12):1599–1612, 1997.

[104] Michael E Wall, Steven E Ealick, and Sol M Gruner. Three-dimensional diffuse x-ray scattering from crystals of staphylococcal nuclease. *Proceedings of the National Academy of Sciences*, 94(12):6180–6184, 1997.

[105] Michael E Wall, Andrew H Van Benschoten, Nicholas K Sauter, Paul D Adams, James S Fraser, and Thomas C Terwilliger. Conformational dynamics of a crystalline

protein from microsecond-scale molecular dynamics simulations and diffuse x-ray scattering. *Proceedings of the National Academy of Sciences*, 111(50):17887–17892, 2014.

[106] Michael E Wall, Alexander M Wolff, and James S Fraser. Bringing diffuse x-ray scattering into focus. *Current opinion in structural biology*, 50:109–116, 2018.

[107] Regula Walser, Philippe H Hünenberger, and Wilfred F van Gunsteren. Molecular dynamics simulations of a double unit cell in a protein crystal: volume relaxation at constant pressure and correlation of motions between the two unit cells. *Proteins: Structure, Function, and Bioinformatics*, 48(2):327–340, 2002.

[108] Thomas Williams, Colin Kelley, and many others. Gnuplot 4.6: an interactive plotting program. `http://gnuplot.sourceforge.net/`, April 2013.

[109] Martyn D Winn, Charles C Ballard, Kevin D Cowtan, Eleanor J Dodson, Paul Emsley, Phil R Evans, Ronan M Keegan, Eugene B Krissinel, Andrew GW Leslie, Airlie McCoy, et al. Overview of the ccp4 suite and current developments. *Acta Crystallographica Section D: Biological Crystallography*, 67(4):235–242, 2011.

[110] David C Wych, James S Fraser, David L Mobley, and Michael E Wall. Liquid-like and rigid-body motions in molecular-dynamics simulations of a crystalline protein. *Structural Dynamics*, 6(6):064704, 2019.

[111] David C Wych, James S Fraser, David L Mobley, and Michael E Wall. Liquid-like and rigid-body motions in molecular-dynamics simulations of a crystalline protein. *Structural Dynamics*, 6(6):064704, 2019.

[112] Yi Xue and Nikolai R Skrynnikov. Ensemble md simulations restrained via crystallographic data: accurate structure leads to accurate dynamics. *Protein Science*, 23(4):488–507, 2014.

[113] V. Zoete, M. A. Cuendet, A. Grosdidier, and O. Michielin. Swissparam, a fast force field generation tool for small organic molecules. *J. Comput. Chem*, 32(11):2359–2368, 2011.

Appendix A

Preparatory and analysis code, and production simulation .mdp paramters

`propagate.py`: an example of a propagation script, using `mdtraj`[63] and `numpy`[35] which takes in an ensemble model `.pdb` file as input, and outputs a full supercell `.pdb` file.

```
import mdtraj as md
import numpy as np

# LOAD IN ENSEMBLE
ens = md.load_pdb("ensemble.pdb")

#UNIT CELL LENGTHS (a, b, c)
UC_L = ens.unitcell_lengths[0]

# TRANSFORM TO FRACTIONAL COORDS
for i in range(32):
    ens.xyz[i] = np.apply_along_axis(lambda x: np.divide(x, UC_L), 1, ens.xyz[i])

def UC(coords, which):
    '''Takes in protein coordinates in array form and outputs coordinates
    transformed by P212121 symmetry operations
    '''
    if which == "first":
```

```
        for i in range(len(coords)):
            coords[i] = np.array([-coords[i][0]+0.5, -coords[i][1], coords[i][2]+0.5])

    if which == "second":
        for i in range(len(coords)):
            coords[i] = np.array([coords[i][0]+0.5, -coords[i][1]+0.5, -coords[i][2]])

    if which == "third":
        for i in range(len(coords)):
            coords[i] = np.array([-coords[i][0], coords[i][1]+0.5, -coords[i][2]+0.5])

    return coords

def TRN(coords, axis):
    '''Translates coordinates by a full unit cell along the selected direction
    '''
    if axis == "x":
        return np.apply_along_axis(lambda x: np.add(np.array([1.0, 0.0, 0.0]), x), 1, coords
    )

    elif axis == "y":
        return np.apply_along_axis(lambda x: np.add(np.array([0.0, 1.0, 0.0]), x), 1, coords
    )

    elif axis == "z":
        return np.apply_along_axis(lambda x: np.add(np.array([0.0, 0.0, 1.0]), x), 1, coords
    )

# OG UC
# DO NOTHING FOR THE ORIGINAL PROTEIN (ens.xyz[0])
ens.xyz[1]  = UC(ens.xyz[1], "first")
ens.xyz[2]  = UC(ens.xyz[2], "second")
ens.xyz[3]  = UC(ens.xyz[3], "third")
# UC +X
ens.xyz[4]  = TRN(   ens.xyz[4],            "x")
ens.xyz[5]  = TRN(UC(ens.xyz[5], "first"),  "x")
ens.xyz[6]  = TRN(UC(ens.xyz[6], "second"), "x")
ens.xyz[7]  = TRN(UC(ens.xyz[7], "third"),  "x")
#UC +Y
ens.xyz[8]  = TRN(   ens.xyz[8],             "y")
ens.xyz[9]  = TRN(UC(ens.xyz[9],  "first"),  "y")
```

```
ens.xyz[10] = TRN(UC(ens.xyz[10], "second"), "y")
ens.xyz[11] = TRN(UC(ens.xyz[11], "third"),  "y")
# UC +Z
ens.xyz[12] = TRN(   ens.xyz[12],            "z")
ens.xyz[13] = TRN(UC(ens.xyz[13], "first"),  "z")
ens.xyz[14] = TRN(UC(ens.xyz[14], "second"), "z")
ens.xyz[15] = TRN(UC(ens.xyz[15], "third"),  "z")
# UC +X+Y
ens.xyz[16] = TRN(TRN(   ens.xyz[16],            "x"), "y")
ens.xyz[17] = TRN(TRN(UC(ens.xyz[17], "first"),  "x"), "y")
ens.xyz[18] = TRN(TRN(UC(ens.xyz[18], "second"), "x"), "y")
ens.xyz[19] = TRN(TRN(UC(ens.xyz[19], "third"),  "x"), "y")
# UC +X+Z
ens.xyz[20] = TRN(TRN(   ens.xyz[20],            "x"), "z")
ens.xyz[21] = TRN(TRN(UC(ens.xyz[21], "first"),  "x"), "z")
ens.xyz[22] = TRN(TRN(UC(ens.xyz[22], "second"), "x"), "z")
ens.xyz[23] = TRN(TRN(UC(ens.xyz[23], "third"),  "x"), "z")
# UC +Y+Z
ens.xyz[24] = TRN(TRN(   ens.xyz[24],            "y"), "z")
ens.xyz[25] = TRN(TRN(UC(ens.xyz[25], "first"),  "y"), "z")
ens.xyz[26] = TRN(TRN(UC(ens.xyz[26], "second"), "y"), "z")
ens.xyz[27] = TRN(TRN(UC(ens.xyz[27], "third"),  "y"), "z")
# UC +X+Y+Z
ens.xyz[28] = TRN(TRN(TRN(   ens.xyz[28],            "x"), "y"), "z")
ens.xyz[29] = TRN(TRN(TRN(UC(ens.xyz[29], "first"),  "x"), "y"), "z")
ens.xyz[30] = TRN(TRN(TRN(UC(ens.xyz[30], "second"), "x"), "y"), "z")
ens.xyz[31] = TRN(TRN(TRN(UC(ens.xyz[31], "third"),  "x"), "y"), "z")

# TRANSFORM BACK TO REAL COORDINATES
for j in range(32):
    ens.xyz[j] = np.apply_along_axis(lambda x: np.multiply(x, UC_L), 1, ens.xyz[j])

# SAVE OUT THE FULL SUPERCELL
ens.save("supercell.pdb")
```

`reverse_propagate.py`: an example of a script which sends supercell proteins back to the position of the asymmetric unit, by reversing the unit cell translation and symmetry propagation operations. This script uses the helper module `pdbio.py`, a minimalist script written by me to facilitate the reading, manipulation, and writing of atom information from `.pdb` files. The link to the code is provided in the *Vita*.

```
from pdbio import *
import numpy as np
from copy import deepcopy
import sys

# GET THE FILENAME
filename = sys.argv[1]

# LOAD IN ENSEMBLE
full = PDBFile(filename)

# GET UNIT CELL SIDE LENGTHS FROM SUPERCELL BOX LENGTHS
full.crystinfo.a = full.crystinfo.a/2
full.crystinfo.b = full.crystinfo.b/2
full.crystinfo.c = full.crystinfo.c/2

# TRANSFORM TO FRACTIONAL COORDS
for atom in full.contents:
    if not isinstance(atom, str):
        atom.x /= full.crystinfo.a
        atom.y /= full.crystinfo.b
        atom.z /= full.crystinfo.c

# ATOMS PER PROTEIN
app = int(len([atom for atom in full.contents if not isinstance(atom, str)])/32)

# INDIVIDUAL PROTEINS
prots = [PDBFile(ifilename=None, crystinfo=full.crystinfo,
                 contents=full.contents[i*app:(i+1)*app]) for i in range(32)]

def UC(_prot, which):
    '''Takes in a protein PDBFile object, and outputs the same object, with
    coordinates mapped back to the assymmetric unit, based on their placement
    in the unit cell
```

```
    '''
    prot = deepcopy(_prot)
    if which == "first":
        for atom in prot.contents:
            if not isinstance(atom, str):
                coord = [atom.x, atom.y, atom.z]
                coord = [-(coord[0]-0.5), -coord[1], coord[2]-0.5]
                atom.x = coord[0]; atom.y = coord[1]; atom.z = coord[2]

    if which == "second":
        for atom in prot.contents:
            if not isinstance(atom, str):
                coord = [atom.x, atom.y, atom.z]
                coord = [coord[0]-0.5, -(coord[1]-0.5), -coord[2]]
                atom.x = coord[0]; atom.y = coord[1]; atom.z = coord[2]

    if which == "third":
        for atom in prot.contents:
            if not isinstance(atom, str):
                coord = [atom.x, atom.y, atom.z]
                coord = [-coord[0], coord[1]-0.5, -(coord[2]-0.5)]
                atom.x = coord[0]; atom.y = coord[1]; atom.z = coord[2]

    return prot

def TRN(_prot, axis):
    '''Takes in a protein PDBFile object, and outputs the same object,
    translated by a unit cell length in the selected axis direction
    '''
    prot = deepcopy(_prot)
    if axis == "x":
        for atom in prot.contents:
            if not isinstance(atom, str):
                coord = [atom.x, atom.y, atom.z]
                coord = [coord[0]-1.0, coord[1], coord[2]]
                atom.x = coord[0]; atom.y = coord[1]; atom.z = coord[2]

    if axis == "y":
        for atom in prot.contents:
            if not isinstance(atom, str):
                coord = [atom.x, atom.y, atom.z]
```

```
                coord = [coord[0], coord[1]-1.0, coord[2]]
                atom.x = coord[0]; atom.y = coord[1]; atom.z = coord[2]

    if axis == "z":
        for atom in prot.contents:
            if not isinstance(atom, str):
                coord = [atom.x, atom.y, atom.z]
                coord = [coord[0], coord[1], coord[2]-1.0]
                atom.x = coord[0]; atom.y = coord[1]; atom.z = coord[2]

    return prot

# OG UC
# DO NOTHING FOR THE FIRST PROTEIN
# (IT IS ALREADY ON THE ASYMMETRIC UNIT)
prots[1] = UC(prots[1], "first")
prots[2] = UC(prots[2], "second")
prots[3] = UC(prots[3], "third")
# UC X
prots[4]  =    TRN(prots[4], "x")
prots[5]  = UC(TRN(prots[5], "x"), "first")
prots[6]  = UC(TRN(prots[6], "x"), "second")
prots[7]  = UC(TRN(prots[7], "x"), "third")
#UC Y
prots[8]  =    TRN(prots[8],  "y")
prots[9]  = UC(TRN(prots[9],  "y"), "first")
prots[10] = UC(TRN(prots[10], "y"), "second")
prots[11] = UC(TRN(prots[11], "y"), "third")
# UC Z
prots[12] =    TRN(prots[12], "z")
prots[13] = UC(TRN(prots[13], "z"), "first")
prots[14] = UC(TRN(prots[14], "z"), "second")
prots[15] = UC(TRN(prots[15], "z"), "third")
# UC XY
prots[16] =    TRN(TRN(prots[16], "x"), "y")
prots[17] = UC(TRN(TRN(prots[17], "x"), "y"), "first")
prots[18] = UC(TRN(TRN(prots[18], "x"), "y"), "second")
prots[19] = UC(TRN(TRN(prots[19], "x"), "y"), "third")
# UC XZ
prots[20] =    TRN(TRN(prots[20], "x"), "z")
prots[21] = UC(TRN(TRN(prots[21], "x"), "z"), "first")
```

```
prots[22] = UC(TRN(TRN(prots[22], "x"), "z"), "second")
prots[23] = UC(TRN(TRN(prots[23], "x"), "z"), "third")
# UC YZ
prots[24] =    TRN(TRN(prots[24], "y"), "z")
prots[25] = UC(TRN(TRN(prots[25], "y"), "z"), "first")
prots[26] = UC(TRN(TRN(prots[26], "y"), "z"), "second")
prots[27] = UC(TRN(TRN(prots[27], "y"), "z"), "third")
# UC XYZ
prots[28] =    TRN(TRN(TRN(prots[28], "x"), "y"), "z")
prots[29] = UC(TRN(TRN(TRN(prots[29], "x"), "y"), "z"), "first")
prots[30] = UC(TRN(TRN(TRN(prots[30], "x"), "y"), "z"), "second")
prots[31] = UC(TRN(TRN(TRN(prots[31], "x"), "y"), "z"), "third")

alph = "abcdefghijklmnopqrstuvwxyzABCDEF"
# CHANGE CHAIN NAMES
for i in range(len(prots)):
    for atom in prots[i].contents:
        if not isinstance(atom, str):
            atom.chainid = alph[i]

#TRANSFORM BACK
for prot in prots:
    prot.renumber_residues()
    _idx = 1
    for atom in prot.contents:
        if not isinstance(atom, str):
            atom.x *= prot.crystinfo.a
            atom.y *= prot.crystinfo.b
            atom.z *= prot.crystinfo.c
            atom.index = _idx
            _idx += 1

def multiconf():
    '''Output multi-conformer .pdb file
    '''
    new_sys = prots[0]
    for atom in new_sys.contents:
        if not isinstance(atom, str):
            _chainid = atom.chainid
            atom.chainid = " "
            atom.resname = _chainid + atom.resname
```

```
158
159     for i in range(31):
160         #prots[i+1].make_model(i+2)
161         for atom in prots[i+1].contents:
162             if not isinstance(atom, str):
163                 _chainid = atom.chainid
164                 atom.chainid = " "
165                 atom.resname = _chainid + atom.resname
166
167         new_sys = new_sys.combine(prots[i+1])
168
169     multiconf = PDBFile(ifilename=None, crystinfo=new_sys.crystinfo, contents=[])
170     for i in range(370):
171         multiconf.combine(PDBFile(ifilename=None, crystinfo=new_sys.crystinfo,
172                                   contents=[el for el in new_sys.contents if not isinstance(
    el, str) and el.resid == i+1]))
173
174     for atom in multiconf.contents:
175         if not isinstance(atom, str):
176             atom.chainid = "A"
177
178     multiconf.write("{}_multiconf.pdb".format(filename[:-4]))
179
180 def multimodel():
181     '''Output multi-model pdb file
182     '''
183     new_sys = prots[0].make_model(1)
184
185     for i in range(31):
186         prots[i+1].make_model(i+2)
187         new_sys = new_sys.combine(prots[i+1])
188
189     new_sys.write("{}_multimodel.pdb".format(filename[:-4]))
190
191 multiconf()
192 multimodel()
```

`prod.mdp`: example GROMACS[5] `.mdp` (molecular dynamics parameters) file for a restrained 100ns simulation, using the leap-frog stochastic dynamics integrator, with 2 femtosecond times steps, coordinate output every 2 picoseconds, simulated annealing to bring the system up to temperature (300 Kelvin) for 450 picoseconds, and LINCS constraints on hydrogen bonds.

```
; VARIOUS PREPROCESSING OPTIONS
title                    = PKA_prod
; Turning on position restraints
define                   = -DPOSRES_WEAKEST
; RUN CONTROL PARAMETERS
integrator               = sd
; Start time and timestep in ps
tinit                    = 0
dt                       = 0.002                    ; 2 fs
nsteps                   = 50000000                 ; 100 ns
; mode for center of mass motion removal
comm-mode                = Linear
; number of steps for center of mass motion removal
nstcomm                  = 1000                     ; 2 ps
; Output Control Paramteres
; Output frequency for coords (x), i
; velocities (v) and forces (f)
nstxout                  = 0                        ; none
nstvout                  = 0                        ; none
nstfout                  = 0                        ; none
; Checkpointing if the MD stop, 2 or 3 times
nstcheckpoint            = 50000                    ; 1000 checkpoints
; Output frequency to log file and energy file     ; ... 1/(100 ps)
nstlog                   = 5000                     ; 10 ps
nstenergy                = 5000                     ; 10 ps
; Output frequency and precision for xtc file
nstxout-compressed       = 1000                     ; 2 ps
compressed-x-precision   = 1000
energygrps               = System
; cutoff scheme
cutoff-scheme            = Verlet
; nblist update frequency
nstlist                  = 10
Verlet-buffer-tolerance  = 0.005
```

```
; ns algorithm (simple or grid)
ns-type                  = grid
; Periodic boundary conditions: xyz (default), no (vacuum)
; or full (infinite systems only)
pbc                      = xyz
; nblist cut-off
rlist                    = 1.0
; OPTIONS FOR ELECTROSTATICS/VDW/PME
; Method for doing electrostatics
coulombtype              = pme
rcoulomb-switch          = 0
rcoulomb                 = 1.0
; Dielectric constant (DC) for cut-off or DC of reaction field
; Method for doing Van der Waals
vdw-type                 = Cut-off
; cut-off lengths
rvdw                     = 1.0
; Ewald parameters
fourier-spacing          = 0.12
pme-order                = 4
ewald-rtol               = 1e-5
ewald-rtol-lj            = 1e-3
; Dispersion Correction
DispCorr                 = EnerPres
; OPTIONS FOR WEAK COUPLING ALGORITHMS
; Temperature coupling
tcoupl                   = no
; Groups to couple separately
tc-grps                  = System
; Time constant (ps) and reference temperature (K)
tau_t                    = 2.0
ref_t                    = 298
; Random seed for v-scale algorithm
ld_seed                   = 1191993
; SIMULATED ANNEALING
; type of annealing
annealing                = yes
; number of annealing reference/control points
annealing-npoints        = 10
; times for the annealing reference/control points
annealing-time           = 0 50 100 150 200 250 300 350 400 450
```

```
;reference temperatures
annealing-temp           = 30 60 90 120 150 180 210 240 270 300
; GENERATE VELOCITIES FOR STARTUP RUN
gen_vel                  = yes
gen_temp                 = 300
gen_seed                 = 1993
; OPTIONS FOR BONDS
constraints              = h-bonds
; Type of constraint algorithm
constraint-algorithm     = lincs
; Do not constrain the start configuration
continuation             = no
; Highest order in the expansion of the constraint coupling matrix
lincs-order              = 4
lincs-iter               = 1
; Number of iterations in the final step of LINCS. 1 is fine for
; normal simulations, but use 2 to conserve energy in NVE runs.
; For energy minimization with constraints it should be 4 to 8.
; Lincs will write a warning to the stderr if in one step a bond
; rotates over more degrees than
lincs-warnangle          = 90
```