APPLICATIONS OF MOLECULAR DYNAMICS SIMULATIONS IN PROTEIN X-RAY CRYSTALLOGRAPHY

by

Oleg Mikhailovskii

A Dissertation

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy

Department of Chemistry

West Lafayette, Indiana

May 2020

THE PURDUE UNIVERSITY GRADUATE SCHOOL

STATEMENT OF COMMITTEE APPROVAL

Dr. Nikolai R. Skrynnikov, Chair

Department of Chemistry

Dr. Jeffrey T. Bolin

Department of Biological Sciences

Dr. Carol B. Post

Department of Medicinal Chemistry and Molecular Pharmacology

Dr. John S. Harwood

Department of Chemistry

Approved by:

Dr. Christine Hrycyna

For my family

ACKNOWLEDGMENTS

First, I would like to thank my parents, Tatiana and Vladimir, who supported me when I changed the field of study from mathematics to biochemistry. I am glad to have them as my never-ending pillar of support. I want to thank my fiancée, Dasha, who was by my side physically when the circumstances allowed but always mentally the whole time I was in the graduate school at Purdue. Also, I deeply want to thank my brothers, Fedor and Dmitriy. When I could not go back home, they came to visit and brought a part of home to me.

Next, I wish to express my appreciation for all the help that came from my high school and university mentors, Ilya Chistyakov, Stanislav I. Kublanovsky and Sergey Kryzhevich. We parted our professional ways when I chose to continue my career in other fields rather than math, but they encouraged me and backed my decision to accept such a challenge.

I would like to thank the academic community that I met while being here at Purdue. Of course, I feel deep gratitude to my advisor, Nikolai Skrynnikov. Without him, my journey as a PhD student would not be possible. I also want to thank all of the members of his lab both former and current and, especially, Adam Groves, Sergey Izmailov, Olga Rogacheva, Tairan Yuwen and Yi Xue.

I wish to express my gratitude to my supervisors while I performed my duties as a teaching assistant. Most notably, I want distinguish Cindy Harwood, Mathew Tantama and Trevor Anderson. Also, being a PhD student, I appreciate the instructors of the courses that I took. Especially, I am glad that Nicholas Noinaj who taught me the basics of macromolecular crystallography, which included not only the mathematical principles, but also growing crystals and performing diffraction experiments.

With no doubt, I want to send a special thank you to my Advisory Committee Members, Jeffrey T. Bolin, Carol B. Post and John S. Harwood. They were always kind to me and helped me to stay on track.

I apologize to those who I do not mention here by name. But I hope they still know that I appreciate their input to where I am.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

PDB: Protein Data Bank

MD: Molecular dynamics

GPU: Graphical processing unit

XRD: X-ray diffraction

NMR: Nuclear Magnetic Resonance spectroscopy

TLS: Translation-libration-screw model parametrization

AMBER: Assisted Model Building with Energy Refinement

CHARMM: Chemistry at Harvard Macromolecular Mechanics

P19X: Modified CHARMM force field for heavy atoms crystallographic refinement

CSDX: P19X force field enhanced by Cambridge Structural Database of small molecules

CDL: Conformationally dependent library

MPD: Methyl-pentanediol

PEG: Polyethylene glycol

TAD: Torsional angles dynamics

SA: Sequential torsion angles and Cartesian dynamics

WO: Grid search optimization of the weight of crystallographic term

PME: Particle mesh Ewald scheme

ASU: Crystallographic asymmetric unit

UC: Crystallographic unit cell

DNA: Deoxyribonucleic acid

ABSTRACT

X-ray crystallography is a foundation of the modern structural biology. Thus, refinement of crystallographic structures remains an important and actively pursued area of research. We have built a software solution for refinement of crystallographic protein structures using X-ray diffraction data in conjunction with state-of-the-art MD modeling setup. This solution was implemented on the platform of Amber 16 biomolecular simulation package, making use of graphical processing unit (GPU) computing. The proposed refinement protocol consists of a short MD simulation, which represents an entire crystal unit cell containing multiple protein molecules and interstitial solvent. The simulation is guided by crystallographic restraints based on experimental structure factors, as well as conventional force-field terms. We assessed the performance of this new protocol against various refinement procedures based on the Phenix engine, which represents the current industry standard. The evaluation was conducted on a set of 84 protein structures with different realizations of initial models; the main criterion of success was free R-factor, R_{free}. Initially, we performed the re-refinement of the models deposited in the PDB bank. We found that in 58% of all cases our protocol achieved better R_free than Phenix. As a next step, we conducted the refinement on three different sets of lower-quality models that were manufactured specifically to test the competing algorithms (average C^{α} RMSD from the target structures 0.75, 0.89, and 1.02 Å). In these tests, our protocol outperformed the refinement procedures available in Phenix in up to 89% of all cases. Aside from R-factors, we also compared geometric qualities of the models as measured by MolProbity scores. It was found that our protocol led to consistently better geometries in all of the refinement comparisons.

Recently, a number of attempts have been made to fully utilize the information encoded in protein diffraction data, including diffuse scattering, which is dependent on molecular dynamics in the crystal. To understand the nature of this dependence, we have chosen three different crystalline forms of ubiquitin. By post-processing the MD data, we separated the effects from different types of motion on the diffuse scattering profiles. This analysis failed to identify any features of the diffuse scattering profiles that could be uniquely linked to certain specific motional modes (e.g. small-amplitude rocking motion of protein molecules in the crystal lattice). However, we were able to confirm the previous experimental observations, made in the

combined X-ray diffraction and NMR study, suggesting that the amount of motion in the specific crystal is reflected in the amplitude of diffuse scattering.

CHAPTER 1. INTRODUCTION

The contemporary paradigm of macromolecular biology is that structure underpins function. Thus, many biochemical studies rely on structure of biological macromolecules, such as proteins, nucleic acids, lipids and their various complexes, to address the function. Based on structural information, the researchers try to predict the system's behavior: e.g. drug efficacy, mechanisms of signal transduction, protein stability, etc.

X-ray crystallography is by far the most powerful method for protein structure determination, as indicated by the RCSB statistics. Approximately 90% of the structures deposited into Protein Data Bank (PDB) are solved using this method. Even with the emergence of 3D electron microscopy, which has undergone an exponential growth over the last decade, scientists continue to report approximately 10 times more of crystallographic structures than structures solved by other techniques. Thus, further exploration in the field of protein crystallography and improvements in both experimental practices and computational methods remain highly relevant for the future progress of structural biology. In particular, the advent of GPU-based computers opens new avenues for building highly accurate structural models.

Figure 1.1 illustrates a general pipeline used in X-ray structure determination. In this work, we focus on the *in-silico* methods associated with this technique. Specifically, we have used molecular dynamics (MD) tools to design and implement the advanced structure refinement procedure. The results of this project are covered in CHAPTER 2. In the second project we modelled the X-ray diffuse scattering effect based on the MD trajectories. These results are stated in CHAPTER 3. We summarize our observations in the CHAPTER 4.

In the current chapter, we first introduce the mathematical models used in both research projects. Next, we address the specific concepts used in the area of structure refinement. Finally, we overview the origins of diffuse scattering and its relationship with protein dynamics. While in this dissertation we focus primarily on proteins, almost everything that is discussed below can be generalized to other types of macromolecules.

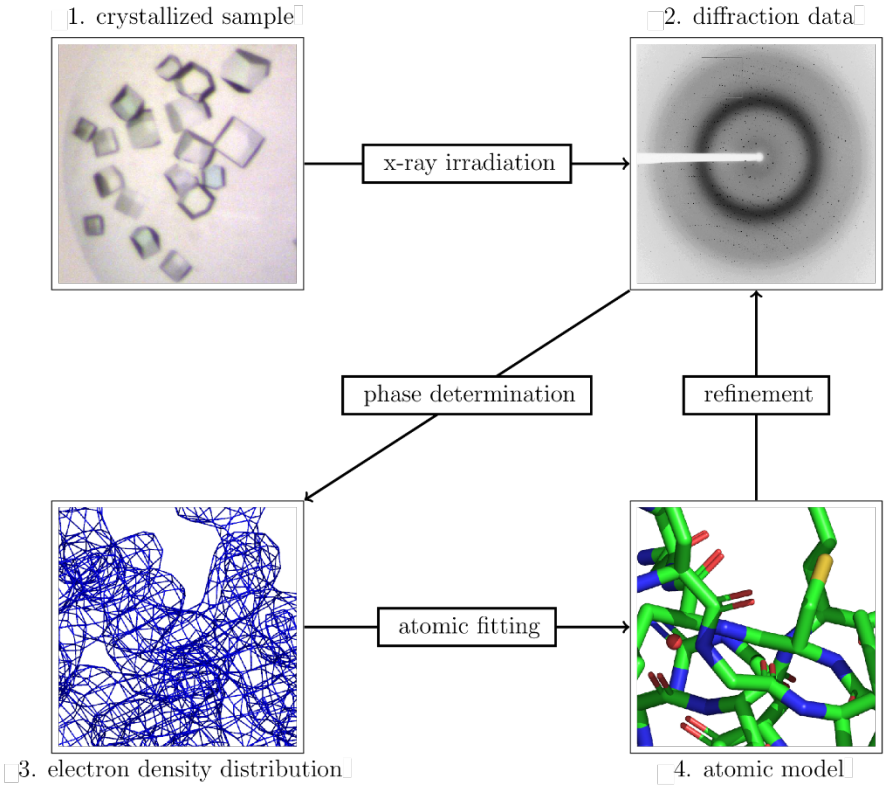

Figure 1.1. Crystallographic structure determination pipeline.

1.1 Protein X-ray crystallography basics

1.1.1 Fundamentals of X-ray diffraction

X-ray waves are scattered by electrons in a sample, hence giving rise to a multitude of secondary waves of the same wavelength from all the electrons in the sample. Therefore, the resultant wave in each given direction is a sum of the secondary waves from the electrons in the sample. These waves are much weaker than the primary ones.

In the case of crystal, the sample is built of the blocks called *unit cells*. These unit cells are repeated periodically in the three spatial dimensions. Superimposing all the secondary waves from the electrons in a whole unit cell, one can consider it as a single source of energy. Given the regularity in the structure of the crystal, it becomes obvious that in some directions the diffracted waves ‘align’ and come to a detector ‘in phase’, even though in most of the directions the waves interfere destructively. Such effect multiplies the energy of the secondary waves by the squared number of the unit cells, which makes them detectable. These waves are called *Bragg*

reflections, and the relation describing the diffracting directions is known as Bragg's law (see Figure 1.1):

$$2d \sin\theta = n\lambda,$$

where d is the spacing between diffracting planes, θ is the incident angle, λ is the wavelength of the primary wave and n is an integer.

If $\boldsymbol{\sigma}'$ and $\boldsymbol{\sigma}''$ are the unit vectors corresponding to the directions of the primary and secondary waves, the expression for a scattering vector $\boldsymbol{s}$ is as follows:

$$\boldsymbol{s} = \frac{\boldsymbol{\sigma}' - \boldsymbol{\sigma}''}{\lambda}.$$

Then, the expansion of the Bragg's law into the three-dimensional real space is provided by the Laue Equations [1]:

$$\boldsymbol{s} \cdot \boldsymbol{a} = h, \boldsymbol{s} \cdot \boldsymbol{b} = k, \boldsymbol{s} \cdot \boldsymbol{c} = l,$$

where the dot '·' is scalar product of two vectors, $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}$ are the vectors representing the periods of the crystal, and h, k, l are integers referred to as *Miller indices.* The triplet of h, k, l corresponds to a particular reflection spot on a diffracting pattern. We denote the observed intensity of the secondary wave at this spot as $I_{obs}(\boldsymbol{s})$.

In the simplest case, the intensity of N_{atoms} immobile structured atoms in a unit cell is described by the formula below:

$$I_{calc}(\boldsymbol{s}) = \sum_{k=1}^{N_{atoms}} \sum_{j=1}^{N_{atoms}} f_k(\boldsymbol{s}) f_j(\boldsymbol{s}) \cos\left[2\pi \boldsymbol{s} \cdot \left(\boldsymbol{r}_j - \boldsymbol{r}_k\right)\right], \qquad (1.1)$$

where $\boldsymbol{r}_j, \boldsymbol{r}_k$ are their positions and $f_k(\boldsymbol{s}), f_j(\boldsymbol{s})$ are spherically symmetric atomic scattering factors that are known for all chemical types of atoms [2]. This expression allows to calculate the diffraction pattern knowing the positions of the atoms. Vice versa, substituting the calculated diffraction pattern of the observed intensities, $I_{obs}(\boldsymbol{s})$, the solution of the set of equations (1.1) with respect to the atomic positions $\{\boldsymbol{r}_j\}$ is the problem the crystallographers face when they *solve* or *determine structures*.

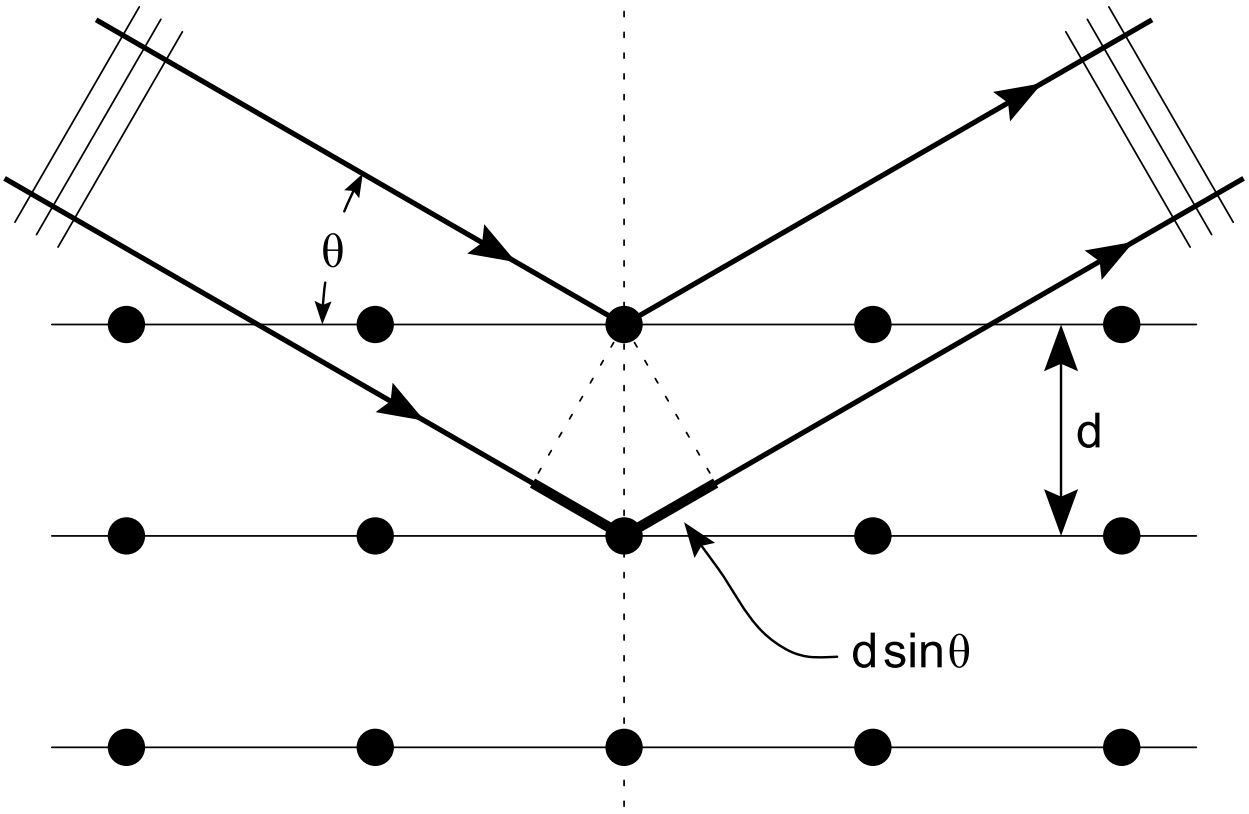

Figure 1.2. A schematic illustration of Bragg's law. Black circles represent unit cells as a single source of scattering in the lattice. The source of the incoming beam is located at the upper left corner. The beam is coming to the reflecting planes spaced by the distance d at an angle θ. The difference between the path lengths of the primary spherical waves according to the Pythagorean theorem is $2d \sin\theta$. Therefore, to produce constructive interference, the difference must be a multiply of the wavelength, λ. The more reflecting planes are in the crystal, the more pronounced is the effect of the in phase superposition. That is, the higher intensity of the diffracted beam.

1.1.2 Structure factors, reciprocal space and reflection resolution

In practice, researchers often operate with structure factors instead of the intensities. *Structure factor* as a function of a reflection $\boldsymbol{s}$ is the ratio between the secondary wave amplitudes in the same direction of the following two experiments: 1) the original crystal considered above when introducing Bragg's law (Figure 1.2) and 2) an imaginary crystal as the original one, but with single electrons at the origins of each unit cell.

As wave amplitude is a complex number, it has a magnitude and a phase. Then, the formal expression for the structure factors is as follows using the same notation as previously:

$$\boldsymbol{F}(\boldsymbol{s}) = F(\boldsymbol{s})\exp[i\phi(\boldsymbol{s})] = \sum_{j=1}^{N_{atoms}} f_j(\boldsymbol{s})\exp\left[i2\pi\boldsymbol{s}\cdot\boldsymbol{r}_j\right]. \quad (1.2)$$

Taking the magnitude of that expression and squaring it gives a simple relationship of it to intensities: $F^2(\boldsymbol{s}) = I(\boldsymbol{s})$. Thus, the problem of structure determination can be reformulated as the solution of the following system:

$$\left|\sum_{j=1}^{N_{atoms}} f_j(\boldsymbol{s})\, exp\left[i2\pi\boldsymbol{s}\cdot\boldsymbol{r}_j\right]\right| = F_{obs}(\boldsymbol{s}), \boldsymbol{s} \in S,$$

where S are the available Bragg scattering vectors. Unfortunately, the direct solution for this problem is impossible: usually there is an extremely large number of unknown parameters (atomic coordinates) and equations.

Because of the Laue equations, there is a mapping between scattering vectors $\boldsymbol{s}$ and Miller indices hkl. Because of this correspondence, we will sometimes substitute $\boldsymbol{s}$ on hkl and vice versa later in the text for convenience.

It is also convenient to introduce here the *reciprocal space basis* $\boldsymbol{a}^*, \boldsymbol{b}^*, \boldsymbol{c}^*$, for which the following conditions hold: $\boldsymbol{a}^*\cdot\boldsymbol{a} = \boldsymbol{b}^*\cdot\boldsymbol{b} = \boldsymbol{c}^*\cdot\boldsymbol{c} = 1,\ \boldsymbol{a}^*\cdot\boldsymbol{b} = \boldsymbol{a}^*\cdot\boldsymbol{c} = 0,\ \boldsymbol{b}^*\cdot\boldsymbol{a} = \boldsymbol{b}^*\cdot\boldsymbol{c} = 0,$ $\boldsymbol{c}^*\cdot\boldsymbol{a} = \boldsymbol{c}^*\cdot\boldsymbol{b} = 0$. Therefore, each scattering vector can be easily expressed in the reciprocal space basis: $\boldsymbol{s} = h\boldsymbol{a}^* + k\boldsymbol{b}^* + l\boldsymbol{c}^*$. The atomic positions in the unit cell are defined in fractional coordinates along each of the periodicity vectors $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}$ as $\boldsymbol{r} = x\boldsymbol{a} + y\boldsymbol{b} + z\boldsymbol{c}$. The scalar product needed to calculate structure factors would have the following form:

$$\boldsymbol{s} * \boldsymbol{r} = hx + ky + lz.$$

The spacing d between reflecting planes shown on Figure 1.2 is the inverse of the scattering vector length: $d(\boldsymbol{s}) = |\boldsymbol{s}|^{-1}$. This value is called the resolution of the reflection. The larger values of hkl correspond to a denser set of reflecting planes, hence, to smaller values of d and considered as reflections of higher resolution.

1.1.3 Electron density distribution calculated from structure factors

By design, the major contribution to the observed intensities on a diffraction pattern from a crystal comes from the secondary waves reflected by the electrons of that crystal. Most of these

electrons are part of atomic composition of the crystal. Thus, their distribution helps to understand where the atoms are located in a unit cell.

Due to the periodic nature of crystal, the function describing *electron density distribution* is defined as a real three-dimensional non-negative function of the fractional coordinates $\boldsymbol{r}$:

$$\boldsymbol{\rho}(\boldsymbol{r}) = \boldsymbol{\rho}(x, y, z) = \boldsymbol{\rho}\big(x + i_x, y + i_y, z + i_z\big),$$

where i_x, i_y, i_z can be arbitrary integer numbers.

By the same reason, this function can be represented as Fourier series with the complex structure factors of the crystal as coefficients summing over the Bragg reflections hkl:

$$\boldsymbol{\rho}(x, y, z) = \sum_{hkl} \boldsymbol{F}^{crystal}(hkl) \exp[-i2\pi(hx + ky + lz)]\,. \qquad (1.3)$$

As one can see from this formula, the determination of the electron density is impossible from a single X-ray diffraction experiment since only the magnitude information is available while no phase information is being collected. Thus, often the distribution is approximated using the experimental amplitudes and phases calculated from a model of the crystal.

Among the difficulties introduced by the nature of experiment one can also point out incompleteness of the set of detected reflections. Another obstacle is the uncertainty of the observed amplitudes.

1.1.4 Structure factors parameters

The considered examples of the diffraction experiments above were thought experiments in ideal conditions: ideal lattice order and immobile atoms all in the same configuration in all unit cells across the crystal. In practice, these conditions are violated and various corrections for experimental and physical nature of the crystal are needed. Further, we discuss some computational approaches on how to mitigate these problems.

Occupancy

First, let us consider the situation when some atoms in the unit cell have multiple positions due to some reason. These alternate conformations might have a different character since the X-ray experiment provides both space-wise and time-wise averaged data. The spatial averaging masks the possibility that some atoms are located at different positions across the crystal unit

cells, for example, they are in conformation A in 60% of the unit cells and in conformation B in the rest 40% of the unit cells. The time averaging leads to the interpretation that 60% of the time during data collection the atoms stay in conformation A and 40% of the data collection time the atoms are in conformation B.

In the case of biologic macromolecules, such situation might be caused by multiple possible states of the structure and/or the dynamics. Also, there might be a combination of the space and time averaging, which is hard to distinguish in the standard approach for structure determination. Nevertheless, both situations are described by the introduction of an additional fourth parameter to the atoms' coordinates – *occupancy*. Moreover, there might be more than two alternate conformers. Hence, the number of the unknown in the system of equation to determine the structure might grow even more than by 25%:

$$\boldsymbol{F}(\boldsymbol{s}) = \sum_{j=1}^{N_{a.c.}} q_j f_j(\boldsymbol{s}) \exp\left[i2\pi \boldsymbol{s} \cdot \boldsymbol{r}_j\right],$$

where the summation goes over all the atoms and their alternate conformers in the unit cell. The occupancies of the conformers are limited: $0 < q_j \leq 1$ and usually there is a single unity conformation for most of the atoms or the sum of the occupancies per atom sum up to 1. Water molecules, which are bound to protein only part of the time, might illustrate the exception to these conditions. They would have a partial occupancy that would be less than 1 since the rest of the time the position of that molecule is unknown.

The lower the partial occupancy of an atom, the harder it is to identify them because of the decreasing contribution to the structure factors. Thus, it is extremely difficult to detect more than two alternate conformers. Their identification also gets harder with the lower resolution of the structure.

Usually, the partial occupancies are present for the mentioned solvent molecules and ions. While in the macromolecules, the occupancy is typically the same for a whole group of atoms such as amino acid residue side chain or a flexible loop region.

Displacement parameters

In the previous paragraph, we discussed large-scale static and dynamic disorders related to spatial and time averaging from the experimental data. Another case of disorder occurs when

there are no multiple distinct conformations but only small-scale fluctuations around a single position. Again, both can happen across the crystal and/or over time of the data collection. These are typically modelled by Gaussian distributions.

In the simplest case of isotropic displacement, the probability distribution P_j of a shift $\boldsymbol{\Delta r}$ of the j-th atom is described by the following formula:

$$P_j(\boldsymbol{\Delta r}) \sim \exp\left[-4\pi^2 \frac{|\Delta \boldsymbol{r}|^2}{B_j}\right],$$

where the parameter is $B_j > 0$. The isotropic factor also called *isotropic B-factor* assumes that the displacement is happening in all directions with equal probability.

The isotropic model is an idealization. In practice, for example, the presence of a bond constraints the movements of the atom. If the experimental data allows to introduce more parameters per atom, a more sophisticated anisotropic modelling is adopted:

$$P_j(\boldsymbol{\Delta r}) \sim \exp\left[-\frac{1}{2}\boldsymbol{\Delta r}\Delta U^{-1}\boldsymbol{\Delta r}\right].$$

Here, ΔU is a symmetric, positive definite matrix that has six parameters describing the probability of movements along the principal axes.

The introduction of the displacement parameters into the overall picture renders the following formulae for the structure factors in isotropic and anisotropic cases, respectively:

$$\boldsymbol{F}(\boldsymbol{s}) = \sum_{j=1}^{N_{a.c.}} q_j f_j(\boldsymbol{s}) \exp\left[-\frac{1}{4}B_j|\mathbf{s}|^2\right] \exp\left[i2\pi \boldsymbol{s}\cdot \boldsymbol{r}_j\right],$$

and

$$\boldsymbol{F}(\boldsymbol{s}) = \sum_{j=1}^{N_{a.c.}} q_j f_j(\boldsymbol{s}) \exp\left[-2\pi^2 \boldsymbol{s} U_j \boldsymbol{s}\right] \exp\left[i2\pi \boldsymbol{s}\cdot \boldsymbol{r}_j\right]. \qquad (1.4)$$

Bulk solvent: exponential model

So far, we have considered the case when only structured elements are present in the unit cell. In macromolecular crystallography, 27-78% of the crystal by volume consists of solvent [3]. If one locates a water molecule or a prosthetic group or other adjunct used to crystallize the sample at a well-defined region, they can be treated using the methods described above.

Moreover, the higher the resolution, the more ordered molecules can be fitted into electron density distribution.

However, even the introduction of atomic occupancy and displacement parameters is not enough to produce a model that has a good agreement between the calculated structure factors and the experimental ones at low resolution. A large portion of the solvent cannot be modelled that way due to its dynamic nature. Time and spatial averaging of experimental data can provide only blurry and featureless electron density corresponding to the solvent, which is exactly the reason why only low resolution is getting affected. The Fourier coefficients in the formula (1.3) corresponding to high resolution are fast-oscillating components, and they are absent. Thus, the disordered solvent, also called *bulk solvent*, requires a special approach.

Babinet's principle states the diffraction patterns from a diffracting body and from a hole of the same size and shape are equal in amplitude and opposite in phase (see Figure 1.3). This statement is applicable to the structure factors of the structured part and bulk solvent of the crystal, $\boldsymbol{F}_{obs}^{structured}(\boldsymbol{s})$ and $\boldsymbol{F}_{obs}^{bulk}(\boldsymbol{s})$, respectively. The electron density calculated from low resolution data are both distributed almost equally smoothly and are complementary to each other, hence:

$$\boldsymbol{F}_{obs}^{structured}(\boldsymbol{s}) \approx -\boldsymbol{F}_{obs}^{bulk}(\boldsymbol{s}),$$

at $\boldsymbol{s}$ corresponding to low resolution reflections.

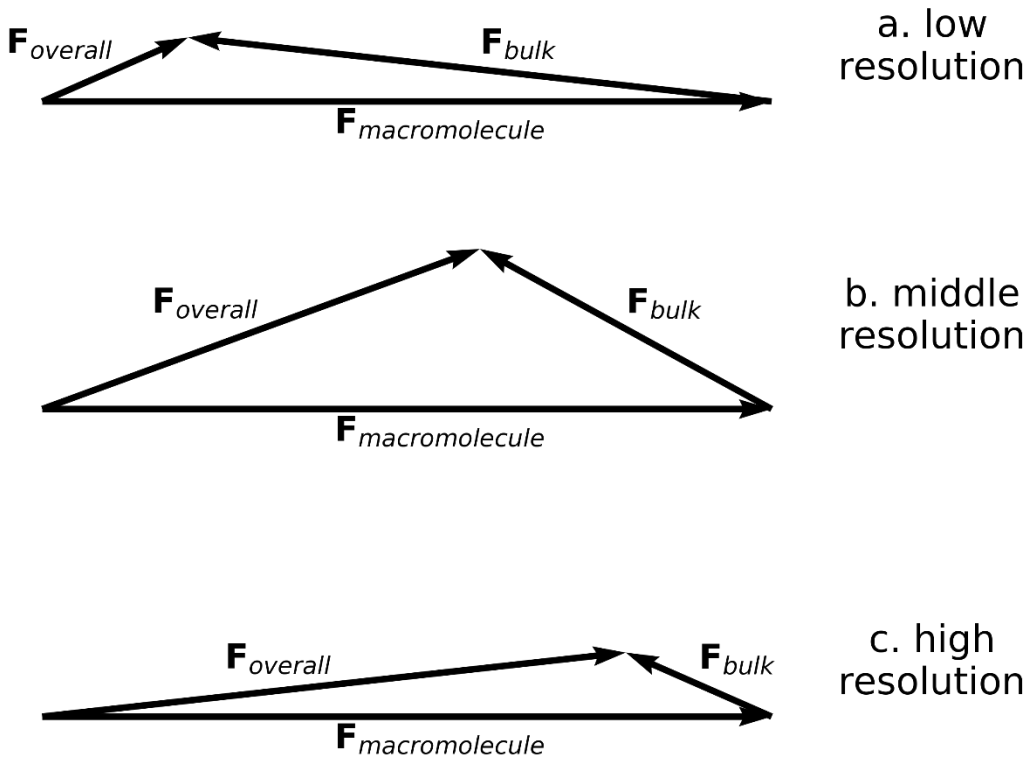

Figure 1.3. Schematic representation of Babinet's principle in macromolecular crystallography. At low resolution the electron density of a macromolecular component and the complementary bulk solvent are almost indistinguishable. Therefore, corresponding structure factors are approximately equal in amplitude but opposite in phase.

Using this observation, the simplest way to include bulk solvent into the model is to add the negative component of the structured molecules with a Gaussian weight, which would negate bulk solvent impact for high resolution reflections:

$$\boldsymbol{F}^{model}(\boldsymbol{s}) = \boldsymbol{F}^{structured}(\boldsymbol{s}) + \boldsymbol{F}^{bulk}(\boldsymbol{s}) \approx \boldsymbol{F}^{structured}(\boldsymbol{s}) \left(1 - k_{sol} \exp\left[-\frac{1}{4} B_{sol} |\boldsymbol{s}|^2\right]\right).$$

The scale coefficients k_{sol} and B_{sol} can be estimated by their physical significance or by least-squares minimization to reduce the discrepancy between the observed and the model data.

The phenomenon of almost cancelling each other's structure factor component holds only at very low resolutions [4] requiring a more rigorous treatment. Currently, there exists another approach to handle this limitation. It is probably the most popular one after the described exponential bulk solvent model and is called *flat bulk solvent model*.

Bulk solvent: flat model

In this model crystallographers consider the electron density corresponding to bulk solvent being flat. They introduce a grid of points with unity values in the unit cell. The points inside the minimal region occupied by ordered atoms are assigned to zero. Such grid is called the molecular mask, and it emulates the bulk solvent electron density distribution. The finer the grid, the more precise the mask is. Further, the Fourier transform of the mask, $\boldsymbol{F}^{mask}(\boldsymbol{s})$, provides the corresponding structure factors.

$$\boldsymbol{F}^{model}(\boldsymbol{s}) = \boldsymbol{F}^{structured}(\boldsymbol{s}) + \boldsymbol{F}^{bulk}(\boldsymbol{s}) = \boldsymbol{F}^{structured}(\boldsymbol{s}) + K_{sol}(|\boldsymbol{s}|)\boldsymbol{F}^{mask}(\boldsymbol{s}).$$

Analogous to the exponential model, one needs to introduce the scaling coefficients, $K_{sol}(|\boldsymbol{s}|)$, which correct the values for actual electron density. As previously, the choice of Gaussian scale function works well to achieve better agreement between structure factors of the model and the experimental ones [5]. When we correct for bulk solvent in CHAPTER 2 we use exactly this approach.

1.1.5 Scaling to the experimental values

Similarly to the atoms in unit cells, each building block of the crystal also vibrates, one needs the displacement parameters to model that. We have mentioned in paragraph 1.1.1 on fundamentals of X-ray diffraction, that intensities and structure factors depend on the number of unit cells in the crystal. This parameter is unknown *a priori* but needs to be considered. Therefore, the overall structure factor of the crystallographic model can be written as follows:

$$\boldsymbol{F}^{model}(\boldsymbol{s}) = k_{overall}k_{isotropic}k_{anisotropic}\left(\boldsymbol{F}^{structured}(\boldsymbol{s}) + k_{mask}\boldsymbol{F}^{mask}(\boldsymbol{s})\right).$$

Here, all the factors are dependent on reflections $\boldsymbol{s}$, except $k_{overall}$:

- $k_{mask} = k_{sol}exp\left(-\frac{B_{sol}s^2}{4}\right)$, where k_{sol}, B_{sol} are the flat bulk-solvent parameters,
- $k_{isotropic} = exp\left(-\frac{Bs^2}{4}\right)$, where B is a scalar parameter,
- $k_{anisotropic} = exp\left(-\frac{2\pi^2\boldsymbol{s}^T\boldsymbol{U}_{cryst}\boldsymbol{s}}{4}\right)$, where $\boldsymbol{U}_{cryst}$ is the overall anisotropic scale matrix,
- Denoting $\boldsymbol{F}'_{model} = k_{isotropic}k_{anisotropic}\left(\boldsymbol{F}^{structured}(\boldsymbol{s}) + k_{mask}\boldsymbol{F}^{mask}(\boldsymbol{s})\right)$, $k_{overall} = \frac{\sum_{\boldsymbol{s}} F_{obs}|\boldsymbol{F}'_{model}|}{\sum_{\boldsymbol{s}}|\boldsymbol{F}'_{model}|^2}$, where the sum is over all reflections.

1.2 Macromolecular structure refinement

1.2.1 Statement of the problem

Once macromolecular crystallographers obtained the experimental values of intensities from the diffraction pattern, they try to solve two interconnected problems: 1) determine the electron density distribution yet lacking the phase information (see paragraph 1.1.3) and 2) identify the set of atoms in that distribution, which provides a hypothesis on the initially unknown phases of structure factors. Thus, neither of these problems can be fully solved independently, and they are approached iteratively. An initial atomic model always contains a lot of errors that need to be corrected to achieve a valid atomic model.

We can divide these errors into two classes. The first one is molecular geometry errors and includes such cases as atomic clashes and impossible bond lengths or angles, etc. The other class is the poor agreement of the model with experimental data.

An accurate model is required to draw accurate structural conclusions, which affect how different features both structural and functional are interpreted. Hence, the elimination of model imperfections is an undoubtedly necessary step to perform. This is the goal of *crystallographic refinement*.

Even in its simplest form, the problem of solving the system of equations (1.1) is too complicated for modern computational hardware. In a realistic setup, the system is non-linear, it contains thousands of equations depending on thousands of parameters. Additionally, the existence and uniqueness as necessary conditions for analytical approach are not ensured. Even the model with true parameters, which are accurate from the molecular geometry point of view, might still produce not ideal conformity between calculated and observed structure factors because of measurement errors and/or approximations described in the corrections paragraph 1.1.4. As such, before the development of bulk solvent models, crystallographers usually cut the range of observed reflection lower than 6-7 Å artificially.

Therefore, researchers opt to another, more realistic task: given a set of inaccurate model parameters they try to minimize the discrepancy between the observed and calculated structure factors also adjusting the parameters. More formally, they minimize some target function, which is a combined measure of errors in the model. For example, the simplest form of such function is a least-squares target first introduced by Booth [6]:

$$T_{LS,F}(\boldsymbol{x}) = \sum_{hkl} \left(F_{hkl}^{model}(\boldsymbol{x}) - F_{hkl}^{obs}\right)^2, \tag{1.5}$$

where the summation goes over some set of reflections used in the refinement and the model structure factors depend on a set of model parameters, e.g. all atomic positions, occupancies, B-factors, etc. Sometimes structure factors are also substituted for intensities, due to the relationship discussed in Fundamentals paragraph 1.1.1:

$$T_{LS,I}(\boldsymbol{x}) = \sum_{hkl} \left(I_{hkl}^{model}(\boldsymbol{x}) - I_{hkl}^{obs}\right)^2 = \sum_{hkl} \left(\left(F_{hkl}^{model}(\boldsymbol{x})\right)^2 - \left(F_{hkl}^{obs}\right)^2\right)^2.$$

Clearly, the minimization reduces the discrepancy between the experimental structure factors and the ones calculated from the model parameters. Moreover, if after minimization such function reaches zero, and it would give an exact solution for the non-linear system (1.1). As mentioned above, the existence of such a solution is rarely the case in practice, so one might want to choose another target function. We discuss the options below in paragraph 1.2.4.

However, there is an issue with the objectivity of the target function value as a score of refinement success. It depends on the specific experiment details, such as the number of reflections used in refinement or the magnitudes scale. This, for example, makes the functions values incomparable between different structures. Therefore, another measure known as R-factors was also introduced by Booth [7] and is used nowadays:

$$R = \frac{\sum_{hkl} \left|F_{hkl}^{model}(\boldsymbol{x}) - F_{hkl}^{obs}\right|}{\sum_{hkl} F_{hkl}^{obs}} \text{ or } R = \frac{\sum_{hkl} \left|F_{hkl}^{model}(\boldsymbol{x}) - F_{hkl}^{obs}\right|}{\sum_{hkl} F_{hkl}^{obs}} * 100\%. \tag{1.6}$$

We will use the first expression further in the text when dealing with model validation. Also, besides the problems experienced when using least-squares target, the use of R-factor as a target indicates an issue of differentiability, which is needed for minimization methods described next.

1.2.2 Minimization methods

There are three main types of algorithms used for optimization problems: Zero order algorithms, First order algorithms and Second order algorithms. Zero order or pure search methods are not used in macromolecular refinement due to an enormous computational load that would be required. Thus, we focus on the latter two approaches.

We start with the Taylor series approximation of the function to minimize, $f(\boldsymbol{x})$, near a point $\boldsymbol{x}_0$, a column vector of the current set of parameters:

$$f(\boldsymbol{x}) \approx f(\boldsymbol{x}_0) + \left|\frac{\mathrm{d}f(\boldsymbol{x})}{\mathrm{d}\boldsymbol{x}}\right|^{t}_{x=x_0} (\boldsymbol{x}-\boldsymbol{x}_0) + \frac{1}{2}(\boldsymbol{x}-\boldsymbol{x}_0)^t \left|\frac{\mathrm{d}^2 f(\boldsymbol{x})}{\mathrm{d}\boldsymbol{x}^2}\right|_{x=x_0} (\boldsymbol{x}-\boldsymbol{x}_0).$$

Here we omitted the terms of the third and higher orders as it is usually done for the refinement problems.

The formula can be rewritten in terms of the difference between the argument and the point in the vicinity of which we approximate our function $f(\boldsymbol{x})$. Denoting $\Delta\boldsymbol{x} = \boldsymbol{x} - \boldsymbol{x}_0$,

$$f(\boldsymbol{x}_0 + \Delta\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{x}_0) + \left|\frac{\mathrm{d}f(\boldsymbol{x})}{\mathrm{d}\boldsymbol{x}}\right|^{t}_{x=x_0} \Delta\boldsymbol{x} + \frac{1}{2}\Delta\boldsymbol{x}^t \left|\frac{\mathrm{d}^2 f(\boldsymbol{x})}{\mathrm{d}\boldsymbol{x}^2}\right|_{x=x_0} \Delta\boldsymbol{x}.$$

Thus, taking the derivative of the function, the quadratic form with respect to the shift vector $\Delta\boldsymbol{x}$ on the right side becomes linear:

$$\left|\frac{\mathrm{d}f(\boldsymbol{x})}{\mathrm{d}\boldsymbol{x}}\right|^{t}_{x=x_0+\Delta x} = \left|\frac{\mathrm{d}f(\boldsymbol{x})}{\mathrm{d}\boldsymbol{x}}\right|^{t}_{x=x_0} + \Delta\boldsymbol{x}^t \left|\frac{\mathrm{d}^2 f(\boldsymbol{x})}{\mathrm{d}\boldsymbol{x}^2}\right|_{x=x_0}.$$

The function reaches its extremum (minimum or maximum) if the function's gradient is zero. Therefore, the following condition on the vector $\Delta\boldsymbol{x}$ gives the search direction:

$$\Delta\boldsymbol{x} = \left|\frac{\mathrm{d}^2 f(\boldsymbol{x})}{\mathrm{d}\boldsymbol{x}^2}\right|^{-1}_{x=x_0} \left|\frac{\mathrm{d}f(\boldsymbol{x})}{\mathrm{d}\boldsymbol{x}}\right|_{x=x_0}.$$

In the full-matrix method, the Hessian (the matrix of second derivatives) is calculated directly. However, typical refinement procedure would involve 10^4 parameters, so the full calculation of 10^8 elements in the matrix requires a lot of memory and computational time. Hence, in other second order methods, one may estimate only some of the elements in the Hessian.

The methods of the first order rely only on the calculation of the first derivates and assume the Hessian to be a unity matrix. Such simplification saves time, but it brings up a problem of the speed of convergence. The comparison of the methods used in structure refinement can be found on Figure 1.4 diagram.

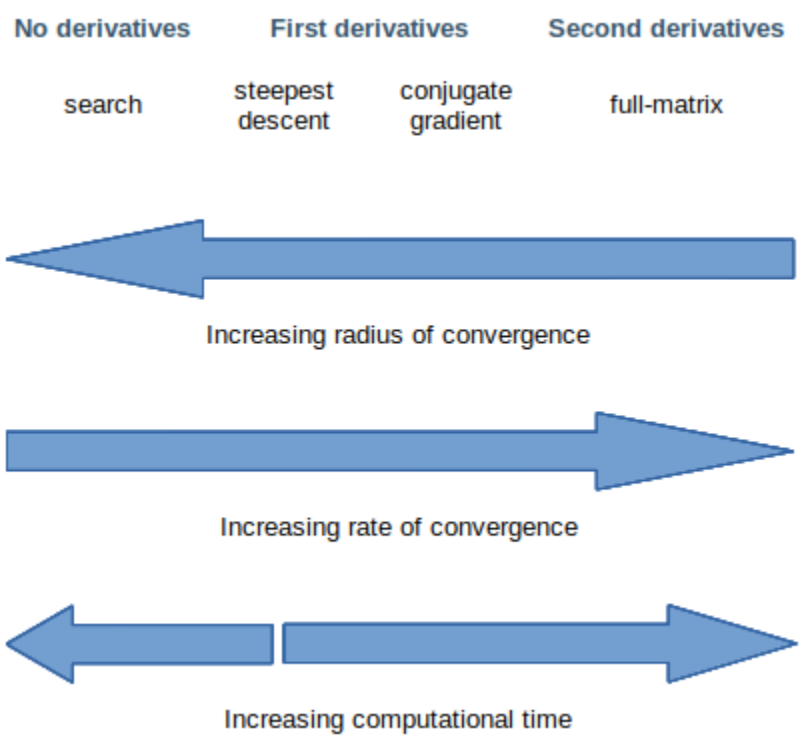

Figure 1.4. The comparison of some optimization methods on three scales. Radius of convergence indicates how close the starting model should be for a successful search. Rate of convergence indicates how fast the minimum would be found. Computational time indicates how long the procedure would take.

Being the most robust method in macromolecular structure refinement and providing an advantage of a large radius of convergence, we relied on the steepest descent algorithm in our project [8]:

$$\boldsymbol{x}^{new} = \boldsymbol{x}^{old} - \left|\frac{\mathrm{d}f(\boldsymbol{x})}{\mathrm{d}\boldsymbol{x}}\right|_{\boldsymbol{x}=\boldsymbol{x}^{old}}.$$

1.2.3 Computational load

The first refinement programs were implemented in the 1970s. At that time, the computers were weak, and a straightforward refinement cycle could take days. Moreover, one such cycle would be insufficient to achieve a reasonable model and the best agreement with experimental data.

To perform one step of minimization, one needs to compute structure factors, then one needs to at least compute gradients or even some second derivatives of the target function depending on the optimization scheme (see previous paragraph). The computational load of the first sub-step is proportional to the product of the number of parameters and the number of reflections. In the simplest scenario of steepest descent, either numeric or analytic calculation of the first derivatives roughly takes the amount of time to compute a single target value multiplied by the number of parameters. On top of that, the total time must be multiplied by the number of iterations. Therefore, some efficient computational tricks were needed to make the software usable.

The two main elements that were employed are crystal symmetries and Fourier transform. First, in some crystal forms, the unit cells contain symmetries. A single block named *asymmetric unit* can be used to generate the whole unit cell provided a set of operations such as rotation, translation and screw. In that case the contents of all asymmetric units in the cell are assumed to be identical. This helps to reduce the computational time since now the summations (both the internal and external loops) in structure factors and target function formulae go over a fraction of atoms. The second trick refers to paragraph 1.1.3 on electron density distribution and the relationship between electron density and structure factors: they are Fourier transforms of each other. In 1956, Cooley and Tukey [9] followed a Sayre's suggestion [10] and proposed a machine algorithm to calculate the approximation of structure factors efficiently using fast Fourier transform. Now, instead of the number of reflections multiplied by the number of atoms complexity, the number of operations would be proportional to simply the number of atoms with a small coefficient. Similarly, one might exploit the Fourier transform to calculate the gradient.

1.2.4 Target functions and model validation

A simple test of refinement using the least-squares target (1.5) shows that for sufficiently large molecules the number of observations is usually lower than the number of parameters. This exposes the problem of over-fitting the experimental data. Another problem that arises with the least-squares target is a small radius of convergence.

For example, in a review on crystallographic refinement [11], Urzhumtsev and Lunin mimicked the refinement of a protein. They generated several models that varied in RMSD (root-mean-squared deviation) against an ideal model from 0.3 Å to 1.4 Å. A set of structure factors

was generated for the ideal model. Next, they refined the models against the set of noiseless structure factors cut at different resolution ranges of 3 Å, 2 Å and 1 Å. Even in the case of the smallest deviations from the reference model, the tests showed that the refinement exhibits over-fitting if there are not enough data points available (3003 parameters against 2290 observables). Namely, the discrepancy between the structure factors was negligible, but the refined model was wrong when compared against the reference. The second conclusion was that the poorer the initial model is, the worse the refined one is in comparison with the reference model, and the worse the model-to-data fit is as judged by R-factors. One can find more on these tests in [11].

To overcome these problems, one first needs to increase data-to-parameters ratio. The two ways to do that are imposing restraints to increase the amount of data used in refinement or decrease the number of independent parameters imposing constraints on them. In our project, we implement the approach when additional data is incorporated into refinement.

When refining macromolecular coordinates, crystallographers typically approach these issues by introduction of an additional term to the target:

$$T = T_{LS} + T_{geometry},$$

where the term $T_{geometry}$, also called force field, implements the restraints on such parameters as bond lengths and angles, dihedral and improper dihedral angles to enforce planarity, and van der Waals and electrostatic interactions. The first four components represent bonded interactions and the latter two – non-bonded interactions:

$$T_{geometry} = T_{bonds} + T_{angles} + T_{dihedral} + T_{improper} + T_{vdW} + T_{electrostatic}. \quad (1.7)$$

Unfortunately, this modification of the target term alone is not enough in practice. Other issues arise even after the additional restraints are imposed, such as a question of weighting between the crystallographic and geometry restraints. Similar to the reasons to introduce R-factors, since the observed data have arbitrary units and the absolute scale is never known, the results of refinement might vary depending on the non-physical scaling factor. Some of the other problems include the noisiness of experimental values and errors in the model, which are irremovable, such as when a part of macromolecular structure cannot be modelled. For example, the latter problem manifests itself analogously to overfitting, which is discussed in the next paragraph. We explore the target function options to mitigate the mentioned challenges in the following sub-paragraphs.

Model validation

Even if after the refinement the corresponding R-factor is low, the model can be incorrect. Since the target has a lot of local minimums, the optimization often gets trapped in them instead of the desired global minimum. This is linked to structure factors expression, which involves a lot of sines and cosines. The higher the resolution, the more such terms are present and the more rugged the profile of the target becomes. Thus, to distinguish between the true model and an incorrect one, a validation measure is needed.

Obviously, one needs to use some complementary data to those that have been used during the refinement. If the refinement is unrestrained, one can check the correctness of bond lengths and angles. In 1992, Brünger suggested to exclude a fraction of crystallographic data from the refinement procedure [12]. Typically, one would randomly and uniformly select 5-10% of the reflections and label them as a test set, while the rest would be used as regularly during the refinement and called a worked set. Now, as previously one would calculate so-called free R-factor (1.6) but the summation would go only over the test set. The discrepancy between the observed and model structure factors during the refinement might go down, but if it went a wrong way (e.g. over-restraining) or if the parameters-to-data ratio is too high (e.g. over-fitting) then the difference between the R_{free} and R_{work} factors would be enormously high.

Simple tests show that such big gaps are exactly the case in a realistic scenario when a model with irremovable errors is being refined with least-squares target even against an error-free dataset and the partial model becomes even more distorted. This makes the direct comparison of the observed data and those generated from a model ill-founded, suggesting a modification to the crystallographic term of the target.

Crystallographic term of target

Maximum likelihood methods are known to be more robust than least-squares in the case of noisy data. Maximization of the probability that the structure factors of a current model reproduce the experimental values is a common approach to solve the problems mentioned before. Such method was introduced into macromolecular crystallography by Lunin, Bricogne, Read, Pannu, Murshudov and others [13]–[21]. The maximization of the probability can be

reformulated in a more convenient form of minimization of negative logarithm [22], which now serves as a standard target, for example in Phenix refinement module [23]:

$$T_{ML} = \sum_{hkl} \Psi(\boldsymbol{F}_{model}, F_{obs}, \alpha, \beta), \text{ with}$$

$$\Psi = \begin{cases} -ln\left(\frac{2F_{obs}}{\varepsilon\beta}\right) + \frac{{F_{obs}}^2}{\varepsilon\beta} + \frac{\alpha^2|\boldsymbol{F}_{model}|^2}{\varepsilon\beta} - ln\, I_0\left(\frac{2\alpha|\boldsymbol{F}_{model}|F_{obs}}{\varepsilon\beta}\right), acentric\ reflections \\ -\frac{1}{2}ln\left(\frac{2}{\pi\varepsilon\beta}\right) + \frac{{F_{obs}}^2}{2\varepsilon\beta} + \frac{\alpha^2|\boldsymbol{F}_{model}|^2}{2\varepsilon\beta} - ln\, cosh\left(\frac{\alpha|\boldsymbol{F}_{model}|F_{obs}}{\varepsilon\beta}\right), centric\ reflections \end{cases},$$

where the coefficient ε depends on the Miller index hkl and on the space group of the crystal and is equal to the number of symmetry operations that, when applied to the vector hkl, leave it unchanged. I_0 is the zero-order modified Bessel function of the first kind. α and β are the parameters that accumulate model errors and uncertainties. We will discuss the estimation of parameters more in CHAPTER 2.

Another option is the hybrid of the maximum likelihood and least-squares [24]. One similarly performs the estimation of α and β, but instead of the original target, now the optimization of the following function is done:

$$T_{ML}^{*} = \sum_{hkl} w_{hkl}^{*}\left(F_{hkl}^{model}(\boldsymbol{x}) - F_{hkl}^{*}\right)^2,$$

where the weights, w_{hkl}^{*}, and adjusted structure factors, F_{hkl}^{*}, are expressed by the means of the experimental structure factors and the uncertainty parameters of the likelihood, α and β. The exact expressions are derived by Lunin, Afonine and Urzhumtsev [24].

Force field term of target

The general form of the force field term of target for macromolecular coordinates refinement was expressed in equation (1.7). Yet, the question of which values should be plugged-in as ideal ones into the restraints stands. We have analyzed the most popular software used for restrained structure refinement according to RCSB statistics and came up with the following list: Xplor/CNS, PROFFT/PROLSQ, SHELXL, REFMAC, Phenix and BUSTER+TNT. We distinguished two main datasets used in these and other less popular programs:

1. The current standard, values developed by Engh and Huber [25] were and are used in a number of programs such as Xplor and CNS developed by Brünger and colleagues [26],

[27], SHELX/SHELXL by Sheldrick and Schneider [28], PROFFT/PROLSQ by Konnert and Hendrickson [29], [30], BUSTER and TNT by Blanc, Bricogne and Tronrud [31], [32]. The original Xplor parameters were a modification of an all-atom CHARMM force field that did not require explicit hydrogen atoms to be used during refinement (P19X), the reasons for that will be discussed the next sub-paragraph. The Engh and Huber values were derived for each amino acid by querying appropriate chemical fragments from the Cambridge Structural Database of small molecules. The resultant dataset (CSDX) absorbed P19X. However, the accuracy of hydrogen atoms parameters was announced to be limited [33].

1.1. REFMAC5 dictionary is an extension over Engh and Huber dataset, which incorporates monomer-based approach with dynamic definition of links and modifications [34]. It was previously used in Phenix and is currently used in REFMAC [35].

2. Conformationally Dependent Library (CDL) developed by Berkholz et al. [36], which is the current trend [37] and is implemented in TNT and Phenix and has shown to achieve better R-factors [38], [39]. It also can be used in SHELXL [40]. This library was developed by the analysis of 3-residue segments from the Protein Geometry Database, which included high-resolution structures at 1.0 Å or better.

To summarize, the CSDX dataset represents single-value paradigm, which disregards environment and provides the parameters on atom-type basis, the CDL considers two neighboring residues and REFMAC5 dictionary is a mixture of these approaches.

Hydrogen atoms and solvent treatment; consequences for non-bonded interactions

As we have discussed it in the computational load paragraph 1.2.3, historically researchers needed to sacrifice some details in the model to perform refinement efficiently. Hence, to reduce the number of parameters, one would disregard hydrogens from the model. The justification for that was that the contribution from their electrons to the observed intensities is six to eight times smaller than a typical heavy atom in a protein. Thus, their location cannot be seen in the electron density distribution. If only the resolution is high enough, approaching 1.0 Å or beyond the peaks corresponding to hydrogens could be distinguished [41].

From the other point of view, hydrogens make up roughly half of all the atoms present in macromolecular structure. It has been shown that their explicit modelling improves model geometry and provides better agreement between model and experimental structure factors by reducing R-factors [42]. Therefore, the general recommendation is to use these light atoms explicitly [43].

Another consequence of the omitting hydrogens from the model manifested itself in the restraints on non-bonded interactions. First, again because of computational load and second due to electrostatic artifacts in structure determination coupled with the absence of proper solvent treatment to model such contacts. Lennard-Jones potential representing van der Waals forces was replaced by simple repulsive function, and electrostatic interactions were disregarded. The only software that kept the ability to model non-bonded forces is Xplor-NIH/CNS since it's also used to refine structures using nuclear magnetic resonance spectroscopy (NMR) data. However, this program is no longer widely used, it is recommended not to use full non-bonded potential energy in crystallographic refinement in its manual.

Table 1.1. Summary of refinement target function option in the most popular protein crystallography software. Note: Anti-bumping conditions, e.g. simplified non-bonded interactions term, are implemented in all programs except for PROFFT.

	Target function		
Program	Crystallographic term forms	Geometry term restraints	Notes
Xplor/CNS	LS, ML	all classical terms from formula (1.7)	Full non-bonded term is not recommended
PROFFT/PROLSQ	LS	bonds, torsion angles, planarity, chiral centers	No symmetry related restraints
SHELXL	LS	bonds, planarity and chiral volumes	No torsion-angle restraints or specific hydrogen-bond restraints
REFMAC	LS, ML	bonds, angles, torsion angles, planarity, chirality	
Phenix	LS, ML	bonds, angles, torsion angles, planarity, chiral volumes	
BUSTER-TNT	ML	bonds, angles, torsion angles, planarity, chiral centers	Special non-bonded "close" contacts restraints

Such simplification provides a different view on the modification of the crystallographic portion of the target function: to switch off some of its terms and/or add new ones. For example, since traditionally the refinement is done using a single asymmetric unit (see paragraph 1.2.3), it might be beneficial to use some restraints to prevent bad clashes between a model in symmetry-related units, which appears to be an expansion of the mentioned non-bonded interactions modelling. Table 1.1 summarizes the potentials used in the most popular refinement software.

1.2.5 Improving convergence of optimization: molecular dynamics and simulated annealing

So far, we have discussed the modifications of the target function which would improve the parameters-to-data ratio. The second problem brought up in the discussion of optimization against error-free data is the convergence: if a model has too many errors, the minimization cannot reach the global minimum (see previous paragraph). The first approach to tackle the problem is to improve the simplest gradient-based techniques. A comprehensive review on such enhancements, which progressively use more and more second derivatives, can be found in [8], also see paragraph 1.2.2.

Another technique to overcome this issue is called *simulated annealing*. The crystallographic target is a multi-dimensional function with lots of local minima points (see paragraph 1.2.4). Therefore, sometimes it gets trapped during optimization and fails to reach the desired global minimum (Figure 1.5). From this point of view, one can consider the target as a potential energy. In this case, one can introduce kinetic energy to let the system overcome the barrier. A molecular dynamics simulation program assigns high initial velocities to atoms to provide the system more freedom of movements rather than follow the gradients of the target. Then, it slowly reduces the introduced momentum in hope that eventually the model will fall into a global minimum.

Such an approach provides a great deal for correcting large errors and saves time of manual corrections, but the main drawback is the large amount of CPU time. To the best of our knowledge, the simulated annealing protocol is implemented only in Xplor/CNS and Phenix among the programs discussed above.

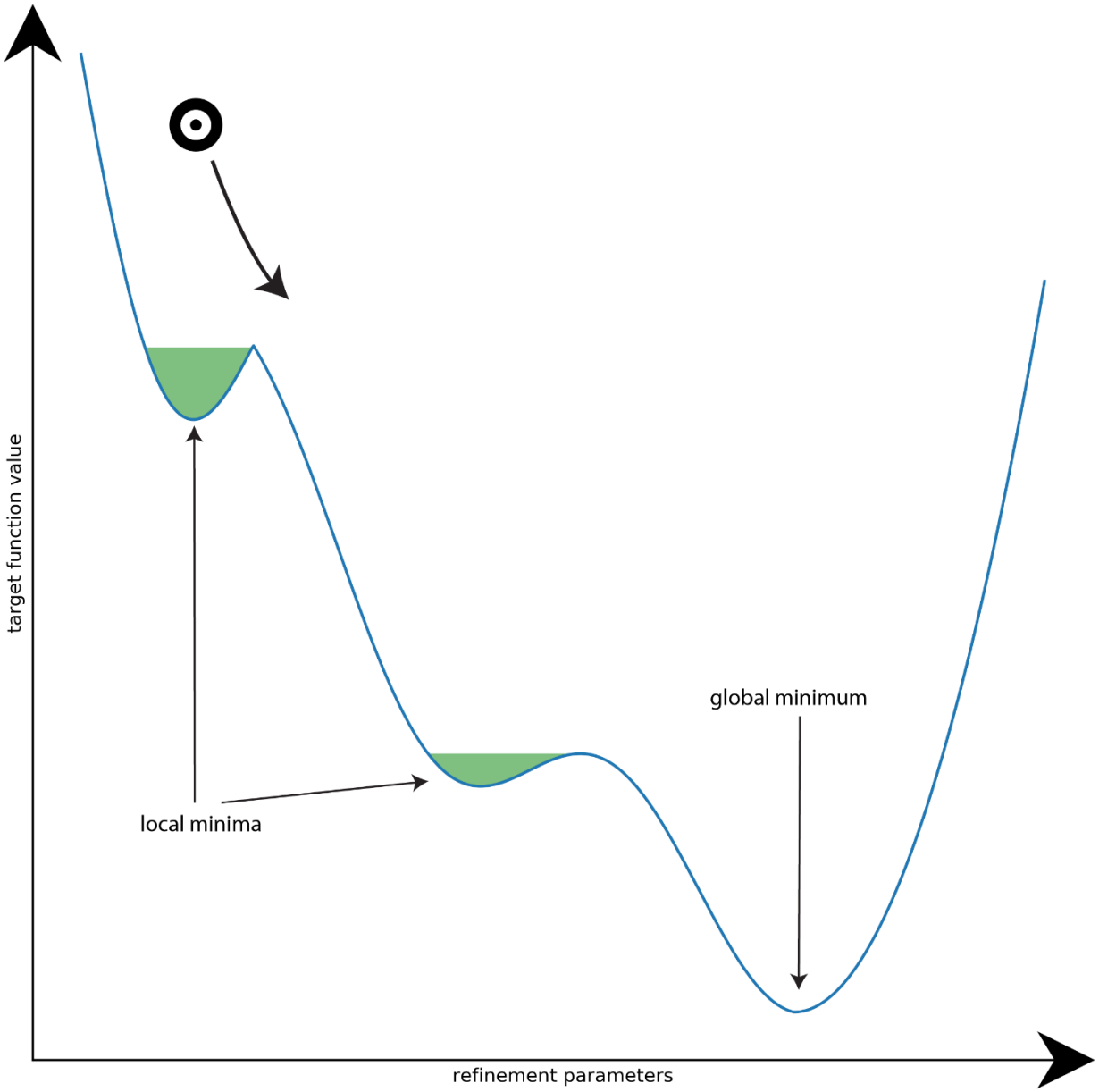

Figure 1.5. Schematic representation of simulated annealing principle. The system is represented as a ball at the upper left corner and its target value profile is projected on 1-dimensional x-axis with values on y-axis. The goal of refinement is to achieve the global minimum. In the regular optimization the system can stuck in the local minimum dips of the curve shown in green. During simulated annealing, the 'heated' system naturally overcomes these barriers.

The first attempt to introduce this idea into crystallographic refinement dates back to 1987 when Karplus, Brünger and Kuriyan published an application to crambin and α-amylase inhibitor [27]. After that, the molecular dynamics driven optimization was successfully implemented by

Fujinaga and Gros [44], [45]. They inspected several sophisticated protocols with the simulations carried after target function minimization, in vacuum and using an asymmetric unit cell:

1. conventional unrestrained MD, e.g. there is only force field term present for potential energy,
2. energy minimization (both just the pure MD potential energy and its combination with the crystallographic restraints), and
3. simulated annealing.

Here, one might ask whether the classical MD or energy minimization using the force field component alone help to achieve better models while maintaining the agreement with experimental data. Unfortunately, neither of these approaches leads to the desired result. Our preliminary tests showed that the energy minimization results in a poor agreement between the experimental data and those calculated from the final structure, even though the geometry of the model improves. Similar conclusions were drawn for unrestrained MD in the Fujinaga and Gros works.

Another implementation of MD-driven approach is a module for NAMD called xMDFF [46]: molecular dynamics flexible fitting for low-resolution X-ray crystallography, which extends the original MDFF module designed for cryo-EM refinement [47]. This program is an example of so-called real-space refinement. Instead of using either LS or ML crystallographic term, the authors added a term based on the electron density calculated from the observed data and the phases from current model (see paragraph 1.1.3). In addition, they also included the term which restrained the secondary structure. Instead of the refinement tool, this module is rather more designed for restrained MD that would fit structures into electron density and authors assess the produced models by comparing with the geometry to the reference structure. The only study using this tool for crystallographic refinement, e.g. not only the geometry but also R-factors were examined, employed the classical approach of simulated annealing [46].

1.2.6 Other advanced refinement protocols: multi-start refinement, structure-factor averaging and ensemble models

The next improvement over the simulated annealing approach is to start several refinements of the same model. Due to the random number generator used in an MD engine, some refined structures would be better than others. Thus, a more optimal model could be

selected against a single model from the classic refinement run, based on both molecular geometry and experimental data fit.

Since the observed intensities reflect only the averaged structure over time and space, one can also average structure factors of the obtained models [48]. This approach improves phases for the election density, reduces model bias and noise introduced by the deviation of a single model from a true one. Therefore, it also helps to better identify the uncertainty parameters in the ML term since they are calculated from the model structure factors.

The method of considering several models and averaging over them can be advanced even further. Instead of treating each run of refinement separately, one can start with an ensemble of structures and refine them simultaneously against the observed data. This leads to a better agreement with experimental structure factors and alleviates local errors from a single model [49]–[52]. Now, the alternate conformers can be modelled explicitly. One generates several models of the same structure and the calculated structure factors are averaged and the resultant data are refined against the observed ones. Here, it is important to notice that all the models are independent of each other from the intermolecular interactions standpoint.

Another advantage of this method is that it might provide new insights into static and dynamic disorder of various systems. For instance, it has been done for TCR–peptide–MHC interface [53], hen egg white lysozyme [54], human complement factor D [55]. Overall, the generation of ensemble representation promises to provide more adequate analysis for further investigation of the structures. Reviews on that topic can be found in [56]–[59].

The major disadvantage of the ensemble representation is the increase of the number of parameters while the amount of data is kept the same. Hence, one should be careful of the over-fitting problem discussed in paragraph1.2.4. Specifically, a large gap between R-work and R-free factors should be avoided.

It is worth pointing that the ensemble refinement does not necessarily require molecular dynamics simulation engine to be used, but rather can be done with the regular minimization technique. However, one aspect that was presented in some of the studies discussed in this paragraph but is not touched in our project does require MD consideration. Besides the ensemble representation and spatial averaging of structure factors, one can also introduce time averaging of structure factors, which would depend on a ‘memory’ parameter, e.g. how long the spatial averages should play the role. Such an approach is an attempt to fully mimic the nature of the

observed data. Another example of such method was presented by Burnley et al. [60] where they re-refined not just one system of interest but 20 different structures. It was shown that the ensemble treatment not only improves the statistics reflecting agreement to experimental data but also might reveal important functional dynamics.

1.2.7 Potential improvements in refinement

We have established above that MD-based protocols of refinement drastically improve the radius of convergence. Therefore, we will focus on such techniques. Even though it is known that simulated annealing protocols might diverge from the true models if the initial model is already close, we will also discuss other approaches, which do not necessarily involve the heating of the simulated system.

As one can see, the traditional approaches do not take into consideration the following three conditions while there are evidences that these three factors are crucial and affect the quality of the final model:

1. state-of-the-art force field / potential energy for the geometry restraints term of the target function,
2. explicit solvent, and
3. explicit representation of crystal unit cell with periodic boundary conditions.

The purpose of implementing these features is to provide a more realistic representation of structure models. Even though it has been mentioned back in 1989 that the refinement can be performed in explicit solvent and whole unit cell with periodic boundary conditions to account for intermolecular interactions and possible alternate conformers [44], it has never been done. Importantly, such an approach implies all-atom ensemble models.

All the discussed refinement techniques if they involved MD simulations of some sort were carried in vacuum using single asymmetric unit because of the issues mentioned in the paragraph 1.2.3 on computational expenses. While earlier, the use of space groups symmetries was handy to save time, today's computational capabilities allow to not assume that all molecules in the unit cell are in the same average configuration. Even in the current refinement procedures for ensemble generation, the average structure factors are adjusted against the experimental ones, but all contributing models are assumed to be independent of each other.

Traditionally, solvent and electrostatics treatment during the refinement was oversimplified even for simulated annealing schemes. Nevertheless, studies discussed below show that their reintroduction along with an all-atom model provides better structures both in terms of biological relevance and agreement to the experimental data.

The consideration of the full unit cell with explicit solvent and, especially, under crystallization conditions provides a better agreement with the X-ray observations. Thus, for example, Kuzmanic and colleagues have shown that traditional in-vacuo refinement of even high 1.0 Å resolution structures can underestimate atomic fluctuations expressed as B-factors [61]. In [62], the authors observed that atomic fluctuations computed from the simulation, which utilized crystallization conditions, closely reproduce the fluctuations derived from experimental B-factors and that the X-ray structure is preserved better in comparison with the simulation in pure water.

The only current software that supports the inclusion of explicit solvent for the force field term of the target, e.g. more precise electrostatics treatment, is previously mentioned xMDFF module of NAMD. The presence of solvent during MD simulations was shown to be beneficial for structure geometry [63]. Also, the authors of NAMD after the exploration of X-ray derived restrained MD simulations suggested that the introduction of solvent affects the quality of structure positively, especially for globular proteins which are exposed to solvent [64]. Their current recommendation is to perform the last round of refinement in an explicit solvent [65]. Yet this aspect is not fully explored in the context of refinement performance against other engines. The fitting simulation in explicit solvent produced better R-factors when compared to implicit solvent and in vacuo simulations, but those values were still worse than the deposited ones by roughly 0.05 or 5%. The caveats for this piece of software are that: 1) it relies on real-space fitting into electron density, which is strongly biased by the starting model quality, 2) it is done for a single model.

The all-atom ensemble models in a unit cell with explicit solvent also naturally raise the question of parameters for the geometry term of the target function. To the best of our knowledge, only four programs perform crystallographic refinement using other than the force fields mentioned in 1.2.4 paragraph. First, it is one of the previously discussed programs, Xplor, and its crystallographic refinement technically can plug in any custom force field. However, it is done in vacuum, hence the problems concerning non-bonded interactions term arise as discussed in the sub-paragraph 1.2.4 on hydrogens. Second, it is xMDFF of NAMD, which uses

CHARMM force field by default and can perform MD-based refinement in vacuum, implicit and explicit solvents. Third, *FFX/Force Field X* uses a Amoeba polarizable force field and performs the refinement in implicit solvent. It has been shown, that modelling electrostatics and inter-/intra-molecular contacts more precisely leads to better results both in terms of geometric qualities and the agreement with the experimental data [66]–[70]. Finally, the same idea of employing a more realistic force field underlies the project of *Rosetta-Phenix* refinement [71], which focuses on low resolution structures.

Phenix has also recently incorporated the ability to use Amber force field, which is soon to be released officially, even though the electrostatics treatment reintroduction is not clearly justified. The corresponding interactions are present in the target, but the refinement is done in vacuum, hence, one might experience the same problems as discussed earlier (see sub-paragraph 1.2.4 on the geometry term of target function) manifested by unrealistic values for non-bonded interactions terms and corrupting the model thereafter.

To summarize, our project is the first attempt to include explicit solvent, state-of-the-art physics-based force field and explicit representation of unit cell with periodic boundary conditions into the crystallographic refinement. MD simulations of crystals in conjunction with crystallographic data can provide a better insight into dynamic nature of macromolecular structure rather than a single static model, which is not necessarily a true structure due to averaging over time and space of experimental data. Another advantage of our program would be GPU-accelerated computations, which drastically reduce the time needed to refine structure, especially those which contain many atoms and observed reflections. So far, only FFX and xMDFF can perform calculations on GPU units.

The only current advancement established in refinement that we do not cover is time-averaged trailing of structure factors. In comparison with the ensemble refinement, the models from the previous steps of the refinement are introduced into the weighted average of structure factors. However, we set the plan to add it in the future. It has been implemented in several works [49], [60], [72], [73]. These researches showed that models built in such fashion can exhibit large structural mobility, which is functionally important. And even such non-explicit ensemble models are preferred over one-model structure description.

Finally, since we are interested in the refinement against crystallographic data, we do not consider protocols that generate macromolecular models de novo.

1.3 Rocking motions through the lens of diffuse scattering

1.3.1 Rocking motions in ubiquitin crystals

Usually, all macromolecules experience small rigid-body deviations from their average ‘ideal’ positions in crystal and such deviations affect the range of diffraction resolutions as pointed in several studies [74]–[79].

Independent atomic motions are modelled through the introduction of isotropic or anisotropic B-factors, as discussed in the displacement corrections sub-paragraph 1.1.4. Overall crystal motions are introduced in a similar fashion as pointed in paragraph 1.1.5 on scaling to experimental values. The motions of intermediate scale such as dynamics of a protein domain or any other group of residues or/and atoms are routinely modelled through translation-libration-screw (TLS) parametrization, see for example [80]. This technique bears the same underlying idea which we touched when talked about the reduction of the number of parameters of the model. It is done to achieve better data-to-parameters ratio during the refinement. In this case, there are 20 refinable parameters per group of atoms with presumably correlated motions [79]. Such approach is biased to the choice of atomic groups, uses only Bragg data and, therefore, does not necessarily lead to a correct model [81]–[83]. As well as B-factors and occupancy parameters, TLS has no ability to distinguish between static and dynamic disorders.

Also, since the TLS approach does not provide hints on the timescale of correlated motions, we have directly observed such rocking motions in crystals for the first time [84]. In that study, we made use of magic-angle spinning NMR spectroscopy, X-ray diffraction (XRD) and MD simulations of explicit crystal lattices to characterize the rigid-body motions of ubiquitin in different crystalline forms: MPD-ub, cubic-PEG-ub and rod-PEG-ub. Such names reflect different precipitation agents (methyl-pentanediol (MPD) and polyethylene glycol (PEG), respectively) and different symmetry relations. These crystals corresponded to previously deposited structures of 3ONS, 3N30 and 3EHV, respectively.

First, we were able to show that the local dynamics is on the ps-ns timescale and is similar between MPD-ub and cubic-PEG-ub as judged by MAS NMR relaxation rates and order parameters. These results were successfully reproduced by explicit MD model of the crystals. To evaluate the parameters obtained from NMR experiments, we produced 1-us-long all-atom MD trajectories of a 2x2 block of unit cell for MPD-ub and single unit cell for cubic-PEG-ub. These

trajectories contained 24 and 48 ubiquitin molecules, respectively.

Further, we evaluated the rigid-body motions in the three crystal forms. We found that the rocking motions are much more pronounced in cubic-PEG-ub than rod-PEG-ub, and rod-PEG-ub motions are slightly more expressed than in MPD-ub as judged by MAS NMR and MD-generated relaxation rates, order parameters. Next, the structures of cubic-PEG-ub and rod-PEG-ub were solved by conventional XRD methods. Their B-factors analysis and TLS modelling confirmed the finding that cubic-PEG-ub is the crystal with the most amount of rigid-body motions.

Concluding, it becomes clear that different motions in different crystalline forms affect structure determination. Interestingly, we found that the amplitude of rigid-body motions correlates with obtainable resolution in our crystals: rod-PEG-ub resolution is 2.2A, cubic-PEG-ub crystal resolution is 2.91A. A similar trend holds for the original structures 3ONS (MPD-ub) 1.8A, 3EHV (rod-PEG-ub) 1.81A, 3N30 (cubic-PEG-ub) 3A.

1.3.2 Diffuse scattering and Guinier formula

According to the classical macromolecular crystallographic approach, researchers use only the intensities of the Bragg peaks to determine the structure and other information is disregarded. Structure modelling using Bragg data can reveal deviations from average positions. Unfortunately, it cannot explain whether these motions are coupled.

The realistic diffraction patterns of most crystals are not clear. In addition to the diffracting reflection spots they also contain smeared background, which is present due to motions in the crystal. Figure 1.6 illustrates the case of only simple translational disorder. The diffraction of the perfect lattice produces sharp Bragg peaks while small deviations from the ideal order introduce cloud-like background.

It has long been known that proteins can preserve their functioning in crystalline form [85]–[88]. The diffuse scattering information reflects the information about disorder and motions in crystal and cannot be extracted from purely Bragg data [89]. Therefore, due to the dynamic nature of proteins, it would be useful to decode the information from diffuse scattering, which is usually omitted, in order to better model proteins' motions. In turn, that would enhance the understanding of the underlying biological process coupled with macromolecular activity.

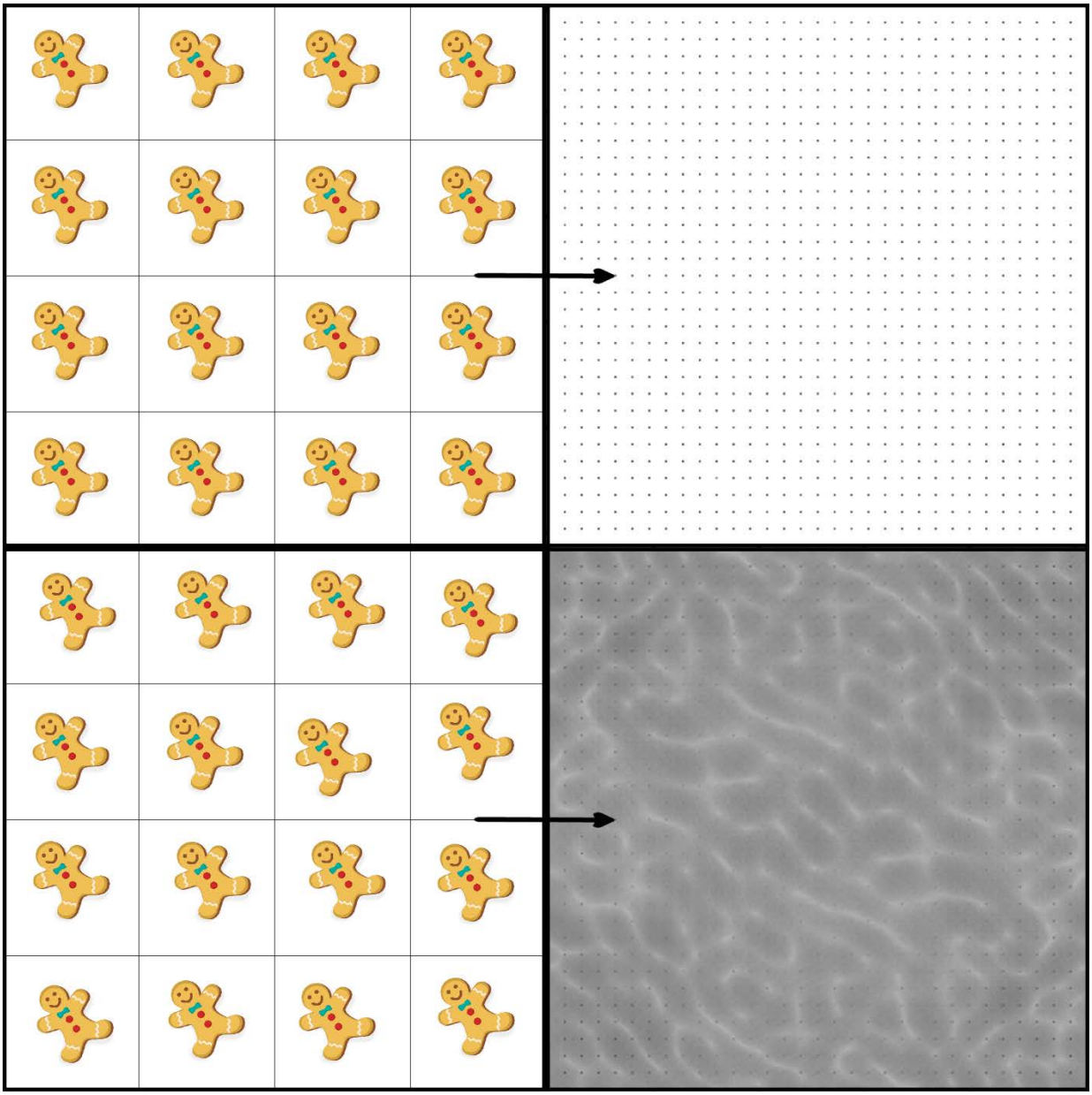

Figure 1.6. Atomic motions in crystal are the source of diffuse scattering. A specific example of how translational motions affect diffaction pattern: perfectly ordered crystal lattice produces sharp Bragg peaks, while translations from the perfect lattice result in cloud-like background.

Equation (1.2) describes the simplest case of just N static atoms in the crystal. Let us consider a more general case when the crystal consists of more than just a single unit cell and the location of the atoms across unit cells might differ. The overall scattering intensity formula looks as follows:

$$I_{total}(\boldsymbol{s}) = \sum_M \sum_N \exp\left[2\pi i\left(\boldsymbol{s}\cdot(\boldsymbol{R}_M - \boldsymbol{R}_N)\right)\right] \times \sum_j \sum_k f_j f_k \exp\left[2\pi i\left(\boldsymbol{s}\cdot(\boldsymbol{r}_{M,j} - \boldsymbol{r}_{N,k})\right)\right],$$

where the first double summation goes over the unit cells, $\boldsymbol{R}_M$ and $\boldsymbol{R}_N$ are the positions of the cells in the lattice, and the second double summation goes over the atoms in the units cells and $\boldsymbol{r}_{M,j}$ and $\boldsymbol{r}_{N,k}$ are the positions in fractional coordinates. If $N_{unit\ cells}$ is the number of unit cells in the crystal and assuming strict order of the atoms in the crystal, the expression reduces to the familiar formula:

$$I_{total}(\boldsymbol{s}) = N_{unit\ cells}^2\, F(\boldsymbol{s})\cdot\ F^*(\boldsymbol{s}),$$

where $F(\boldsymbol{s})$ is the unit cell's structure factor, and we have a perfect Bragg diffraction in this case.

Next, let us assume that each atom is displaced by a small vector $\boldsymbol{\delta}_j$ from the average position across the crystal $\langle \boldsymbol{r}_j \rangle$. The intensity can now be rewritten:

$$I_{total}(\boldsymbol{s}) = \sum_M \sum_N \exp\left[2\pi i\left(\boldsymbol{s}\cdot(\boldsymbol{R}_M - \boldsymbol{R}_N)\right)\right] \times \sum_j \sum_k \left(f_j\, f_k \exp\left[2\pi i\left(\boldsymbol{s}\cdot\left(\langle \boldsymbol{r}_j\rangle - \langle \boldsymbol{r}_k\rangle\right)\right)\right] \times \exp\left[2\pi i\left(\boldsymbol{s}\cdot\left(\boldsymbol{\delta}_{M,j} - \boldsymbol{\delta}_{N,k}\right)\right)\right]\right).$$

The variation from the average positions produces diffuse scattering and one can divide these variations in several types by the range of interactions:

1. uncorrelated random atomic motion,
2. correlated motions within unit cells,
3. correlated motions across several unit cells,
4. long-range interactions.

Assuming random uncorrelated isotropic displacement of atoms and averaging over unit cells, the formula simplifies such that the exponent, which includes atomic displacements, becomes extracted:

$$I_{total}(\boldsymbol{s}) = N_{unit\ cells}^2 \sum_j f_j^2\left(1 - \exp(-4\pi^2\langle \boldsymbol{\delta}_j^2\rangle * \boldsymbol{s}^2)\right) + \sum_M \sum_N \sum_j \sum_k \exp[2\pi i(\boldsymbol{s} * (\boldsymbol{R}_M - \boldsymbol{R}_N))] \times f_j f_k \exp[2\pi i(\boldsymbol{s}\cdot\left(\langle \boldsymbol{r}_j\rangle - \langle \boldsymbol{r}_k\rangle\right))] \times \exp[-2\pi^2\left(\langle \boldsymbol{\delta}_j^2\rangle + \langle \boldsymbol{\delta}_k^2\rangle\right) * \boldsymbol{s}^2].$$

Here we approximated the average of displacements as follows: $\langle \boldsymbol{\delta}_j - \boldsymbol{\delta}_k\rangle^2 \approx \langle \boldsymbol{\delta}_j^2\rangle - \langle \boldsymbol{\delta}_k^2\rangle$. The first term in the sum represents the diffuse scattering that is spherical and is modulated by the

atomic B-factors.

Next, if we assume atomic translational displacements are fully correlated within unit cells then the average displacement would depend only on a unit cell representative and one could write the formula where a single average displacement parameter is present:

$$I_{total}(\boldsymbol{s}) = (N_{unit\ cells}^2(1 - \exp(-4\pi^2\langle\boldsymbol{\delta}^2\rangle * \boldsymbol{s}^2)) + \sum_M \sum_N \exp[2\pi i(\boldsymbol{s} * (\boldsymbol{R}_M - \boldsymbol{R}_N))] \times exp[-4\pi^2\langle\boldsymbol{\delta}^2\rangle * \boldsymbol{s}^2]) \times F(\boldsymbol{s}) \cdot F^*(\boldsymbol{s}),$$

where $F(\boldsymbol{s})$ is the average unit cell structure factor across the crystal. The diffuse scattering that arises in this case is of type two.

To include rotational motions of molecules in unit cells, let us similarly to the previous cases combine the atomic motions into varying structure factors and averaging them:

$$\begin{aligned} I_{total}(s) &= \sum_M \sum_N \exp[2\pi i\ s * (R_M - R_N)]\, F_N(s) * F_M^*(s) \\ &= N_{unit\ cells} \sum_M \langle F_N(s) * F_M^*(s)\rangle_N \exp[2\pi i\ s * \Delta\boldsymbol{R}_M] \\ &= N_{unit\ cells} \sum_M (\langle F(s)\rangle^2 + \langle (F_N(s) - \langle F(s)\rangle)(F_M(s) - \langle F(s)\rangle)\rangle_N\ ) \\ &\times \exp[2\pi i\ s * \Delta\boldsymbol{R}_M]. \end{aligned}$$

Here, $\Delta\boldsymbol{R}_M$ are the differences between unit cell origins. Clearly, the first part of the equation is the classical Bragg scattering, while the second part containing correlations between unit cells corresponds to diffuse scattering. In our case, we are particularly interested in rigid-body motion in unit cells, so it is convenient to rewrite this formula to separate the diffuse scattering intensity explicitly:

$$I_{total}(s) = N_{unit\ cells}^2\langle F(s)\rangle^2 + I_{diff}(s),$$
$$I_{diff} = N_{unit\ cells}\langle |F_N(s) - \langle F(s)\rangle|\rangle_N.$$

This equation is also known as the Guinier equation and has been proven to be suitable for modelling motions in unit cells in the studies that we discuss below.

1.3.3 Exploration of diffuse scattering

There is a limited number of studies on diffuse scattering in protein crystals. By the point of the review by Welberry and Weber in 2016 [90] there has been published less than 30

attempts to investigate the relationships between protein dynamics in crystalline form and diffuse scattering over the past three decades and not much published after that. Those studies included investigations of tropomyosin [91]–[93], insulin [94], lysozyme in various crystalline forms [95], [96], DNAs [97], 6-phosphogluconate dehydrogenase [98], and how correlated motions of different ranges manifest themselves on diffuse scattering patterns. In these pioneering studies, key patterns of diffuse scattering were distinguished and techniques for modeling disorder such as liquid-like motions, normal modes, and rigid-body motions were formulated.

Here, we would like to focus on diffuse scattering modelling, which unfortunately remains relatively small niche with a small contributing community. Miziguchi and Kidera modelled lysozyme diffuse scattering patterns using normal mode based refinement protocol [99]. Faure et al. modelled diffuse scattering of orthorhombic lysozyme also using normal mode analysis and molecular dynamics and showed its similarity in form with the experimental data [100]. The result of modelling using MD simulations was a program called SERENA [101]. Later, they modelled diffuse scattering of tetrahedral lysozyme using isotropic translation-libration analysis and claimed close agreement to experimental data [102]. An attempt to reproduce the experimental X-ray scattering for tRNA was also made by the use of multi-cells and convolutional methods to model atomic disorder [103], [104]. In 1995, diffuse scattering was simulated for myoglobin to investigate how well MD samples conformational space [105]. Next, in a series of works, Wall and colleagues modelled calmodulin [106] and staphylococcal nuclease [107] diffuse scattering and found it to be close to the experimental one using multi-conformer refinement and liquid-like motions analysis developed in the mentioned insulin study [94]. Later, Hery et al. have had lysozyme as a test case for MD-based derivation of diffuse scattering, which well reproduced experiment and, particularly, in the context of rigid-body motions [108]. In a series of works, Meinhold and Smith studied staphylococcal nuclease X-ray scattering profiles and patterns derived from MD simulation and compared those to experimental data [109]–[111]. Riccardi et al. made another attempt to evaluate elastic network models of staphylococcal nuclease by comparison of diffuse scattering predicted by normal mode, liquid-like, and TLS models [112]. In 2014, Wall continued his study on staphylococcal nuclease by producing 1.1-us long MD trajectory of a single unit cell and 5.1-us long trajectory of 2x2x2 block of unit cells, which progressively enhanced previously developed results partially due to a more extensive conformational sampling [113], [114]. Van Benschoten compared liquid-like

motions, normal mode, and TLS models of disorders for cyclophilin A and trypsin [83]. Similar comparison was later done to cyclophilin A, a flavodoxin-like protein WrpA, alkaline phosphatase by Peck et. al [115], and, most recently, cyclophilin A and lysozyme models of disorder and the respective diffuse scattering 3D maps were analyzed by de Klijn in 2019 [116]. More detailed information on the key studies of diffuse scattering can be found in the review by Meisburger and Ando [117].

Several investigators noted that diffuse scattering can be used to verify TLS models [80], [82], [118], ensemble models [60], [113] and others such as detailed contact model [119] and, vice versa, to be used for model building [81], [120]–[122]. Summarizing the current progress in diffuse modelling, unfortunately, there still does not exist a technique that would achieve a correlation coefficient with the experimental data of more than approximately 0.70. Thus, this field needs to be explored more. As it has been mentioned in several recent reviews [90], [120], [123]–[125], taking into account modern progress in data processing and new high quality detectors, now is the time to include the information encoded in diffuse scattering into structure solution instead of omitting it as it is done conventionally.

1.3.4 Application of Guinier formula to compare diffuse scattering of ubiquitin in different crystal lattices based on MD simulation trajectory

So far, the studies which exploited MD simulations approach to investigate proteins used the Guinier formula to generate diffuse scattering profiles and maps from trajectories and compare those to the experimental ones. The obtained results only explored how bad or good the agreement between them is. Also, there has not been done any direct comparison between different crystalline forms of the same structure.

At the same time, during the investigation of rocking motions in ubiquitin, we did solve two structures. Yet, it was done traditionally using only Bragg data and the diffuse scattering data was dismissed. It is also suggested that rigid-body motions dominate the influence on diffuse scattering [116]. The analysis of different models showed that other motions also have to be taken into account to reconstruct the experimental signal [115]. Therefore, given the promising future of diffuse scattering, we wanted to estimate if we could see any evidence of different magnitudes of rocking motions, which we observed using other methods. In other words, do there exist any specific features or watermarks in diffuse scattering that could

distinguish amplitudes of rigid-body motions regardless of crystal space group?

We formulated the goal of our project to investigate how different types of protein motions influence diffuse scattering profile of ubiquitin in different crystalline forms. First, since we know that our MD trajectories closely reproduce the results from NMR and XRD experiments [84], we might also expect the diffuse scattering profiling from trajectories to be successful and reproduce the experimental profiles. Another advantage of such approach is that we could numerically characterize each type of motion based on MD trajectory and try to find a correlation with the diffuse scattering profiles.

Another positive premise of success to our study was that the decomposition of protein and solvent component contribution into diffuse scattering was done in a study on staphylococcal nuclease by Meinhold and Smith [110]. Therefore, our attempt to look at the effect of different types of motions was promising.

CHAPTER 2. MACROMOLECULAR REFINEMENT

2.1 Project product

We aimed to develop a modification of the Amber MD simulation package [126] which could be used as an X-ray crystallography refinement tool with the state-of-the-art force field. In the refinement protocol, we planned to implement explicit solvent treatment and explicit unit cell with periodic boundary conditions rather than traditional asymmetric unit refinement in vacuum. The potential benefits of such approach were discussed in the introductory CHAPTER 1.

Our tool could also potentially be employed in the rebuilding of poorly diffracting regions such as loops and tails, as well as to be used for restrained dynamics to evaluate force fields.

2.2 Summary of Amber modifications

Recent benchmarks and the existing refinement protocols tell that Amber's force field is at least one of the best force fields for the simulation of protein crystal structures [127]–[129]. Hence, one of its most recent versions recommended by the developers, ff14SB [130], was used in this project. We have selected Amber16 package as a base for the refinement software. To accomplish our goal, we have written a Fortran module for CPU-based version and a CUDA module for GPU-based version with Python interface to call auxiliary functions from The Computational Crystallography Toolbox (*cctbx*) open source library. We also changed the original files to call the additional methods from the newly written modules. Our code could be divided into two main parts: 1) the calculation of structure factors of a macromolecule and 2) the calculation of crystallographic force term from the observed and the calculated structure factors.

In the text we mainly refer to the GPU accelerated version of our code, yet all that was done can be accomplished with the CPU version but in longer time period.

2.3 Theoretical basis of the modifications

The overall potential energy in the modified version of Amber can be expressed as follows: $T = w * T_{X-ray} + T_{AMBER}$, where T_{AMBER} is the original Amber force field, T_{X-ray} is the introduced X-ray restraints term, and w is its weight. This overall energy form is essentially the

classic one used in almost all refinement procedures with two terms where the first one is based on experimental results and the second one is based on *a priori* knowledge.

For our initial tests, we used one of two most common forms for the term based on the experimental data, the least-squares target function. However, it quickly becomes clear that the least-squares function performs well only in case when the model is close to a complete one, otherwise systematic errors should be introduced [17], [18], [131]. The usage of such target function leads to a huge gap between R-free and R-work while the model is not being improved. We omit the presentation of our results with the LS target, but similar findings are summarized, for example, in the recent review by Urzhumtsev and Lunin [11].

Therefore, we mainly considered the second common target function: maximum likelihood [18], [132], [133] which is known to improve macromolecular models [134]. We use the form of negative logarithm of the maximum-likelihood function [22], which was introduced in CHAPTER 1:

$$\mathrm{T_{ML}} = \sum_{hkl} \Psi(\boldsymbol{F}_{model}, F_{obs}, \alpha, \beta), \text{ with}$$

$$\Psi = \begin{cases} -ln\left(\frac{2F_{obs}}{\varepsilon\beta}\right) + \frac{{F_{obs}}^2}{\varepsilon\beta} + \frac{\alpha^2|\boldsymbol{F}_{model}|^2}{\varepsilon\beta} - ln\, I_0\left(\frac{2\alpha|\boldsymbol{F}_{model}|F_{obs}}{\varepsilon\beta}\right), acentric\ reflections \\ -\frac{1}{2}ln\left(\frac{2}{\pi\varepsilon\beta}\right) + \frac{{F_{obs}}^2}{2\varepsilon\beta} + \frac{\alpha^2|\boldsymbol{F}_{model}|^2}{2\varepsilon\beta} - ln\, cosh\left(\frac{\alpha|\boldsymbol{F}_{model}|F_{obs}}{\varepsilon\beta}\right), centric\ reflections \end{cases}$$

To obtain the structure factors of the structured atoms in crystal and particularly of the macromolecular component, we first calculate the structure factors of the macromolecule using the direct summation formula:

$$\boldsymbol{F}(s) = \sum_{n=1}^{Natoms} q_n * f_n(s) * \exp\left(-\frac{B_n s^2}{4}\right) * \exp(2i\pi \boldsymbol{r}_n \boldsymbol{s}),$$

where $f(s) = \sum_{k=1}^{P} a_k \exp\left(-\frac{b_k s^2}{4}\right)$ – atomic scattering factor approximation (P depends on the approximation and a_k, b_k are specific for atom type, we used the *it1992* scattering table [135]), q_n – atomic occupancy, B_n – atomic isotropic B-factor, $\boldsymbol{r}_n = (x_n, y_n, z_n)$ – atomic coordinates, $s^2 = \boldsymbol{s}^T G^* \boldsymbol{s}$, $\boldsymbol{s}$ – column-vector of Miller indices, G^* – reciprocal-space metric tensor. The direct summation formula provides a more precise description of how the scattering waves from the crystal electrons affect the diffraction data instead of the Fourier method suggested by Sayre [10] and Cooley and Tukey [9].

Afterwards, provided the experimental structure factors, the scaling procedure *scaler.run* of *cctbx* library [136] is called to correct the obtained macromolecular structure factors $\boldsymbol{F}_{macromolecule}$ for the bulk solvent effect and overall, isotropic and anisotropic factors as in the formula below (see also paragraph 1.1.5):

$$\boldsymbol{F}_{model} = k_{overall}k_{isotropic}k_{anisotropic}(\boldsymbol{F}_{macromolecule} + k_{mask}\boldsymbol{F}_{bulk\ solvent}). \quad (2.1)$$

To achieve an agreement with the experimental data, one needs to minimize T_{X-ray}. We implemented the steepest descent method, which is the most robust minimization approach in structure refinement [8]. The overall force vector applied to each atom at every step of the simulation becomes $\boldsymbol{v}_{Amber} + w * \left(-\nabla T_{X-ray}\right)$ instead of the original $\boldsymbol{v}_{Amber}$ vector induced by the Amber force field. Hence, we calculate the crystallographic force term for each atom of the structure by taking a negative partial derivative of this term with respect to atomic coordinates $\boldsymbol{r}_m = (x_m, y_m, z_m)$ and add the weighted correction to the Amber force vector.

Further, we devise the derivatives for the maximum likelihood crystallographic target function since the least-squares is inappropriate for macromolecular refinement.

As we expand the refinement from a single asymmetric unit to the whole unit cell, we assume the P1 space group. Therefore, all ε from the crystallographic term of the target function are equal 1, and there are no centric reflections. As a result, all the terms in the sum have the following form:

$$\Psi = -ln\left(\frac{2F_{obs}}{\beta}\right) + \frac{{F_{obs}}^2}{\beta} + \frac{\alpha^2|\boldsymbol{F}_{model}|^2}{\beta} - ln\, I_0\left(\frac{2\alpha|\boldsymbol{F}_{model}|F_{obs}}{\beta}\right). \quad (2.2)$$

Consequently, one simplifies the partial derivatives of the target function with respect to the changes in atomic coordinates:

$-\frac{\partial T_{X-ray}}{\partial x_m} = -\sum_{hkl}\frac{\partial \Psi}{\partial|\boldsymbol{F}_{model}|}\frac{\partial|\boldsymbol{F}_{model}|}{\partial x_m}$, and $\frac{\partial \Psi}{\partial|\boldsymbol{F}_{model}|} = \frac{2\alpha^2|\boldsymbol{F}_{model}|}{\beta} - \frac{2\alpha F_{obs}}{\beta}\frac{I_1\left(\frac{2\alpha|\boldsymbol{F}_{model}|F_{obs}}{\beta}\right)}{I_0\left(\frac{2\alpha|\boldsymbol{F}_{model}|F_{obs}}{\beta}\right)}$.

Since the amplitude of a structure factor is non-negative, using $\frac{\partial(\boldsymbol{F}_{model}{}^*\boldsymbol{F}_{model})}{\partial x_m} = 2\left(Re(\boldsymbol{F}_{model}) * \frac{\partial Re(\boldsymbol{F}_{model})}{\partial x_m} + Im(\boldsymbol{F}_{model}) * \frac{\partial Im(\boldsymbol{F}_{model})}{\partial x_m}\right)$, the crystallographic force term of the *m*-th atom in the *x* dimension can be calculated as follows:

$$-\frac{\partial T_{X-ray}}{\partial x_m} = -\sum_{hkl}\left(\frac{2\alpha^2|\boldsymbol{F}_{model}|}{\beta} - \frac{2\alpha F_{obs}}{\beta}\frac{I_1\left(\frac{2\alpha|\boldsymbol{F}_{model}|F_{obs}}{\beta}\right)}{I_0\left(\frac{2\alpha|\boldsymbol{F}_{model}|F_{obs}}{\beta}\right)}\right)\frac{\partial|\boldsymbol{F}_{model}|}{\partial x_m}$$

$$= -\sum_{hkl}\frac{\left(\frac{\alpha^2|\boldsymbol{F}_{model}|}{\beta} - \frac{\alpha F_{obs}}{\beta}\frac{I_1\left(\frac{2\alpha|\boldsymbol{F}_{model}|F_{obs}}{\beta}\right)}{I_0\left(\frac{2\alpha|\boldsymbol{F}_{model}|F_{obs}}{\beta}\right)}\right)}{|\boldsymbol{F}_{model}|}\frac{\partial(\boldsymbol{F}_{model}{}^{*}\boldsymbol{F}_{model})}{\partial x_m}$$

$$= -2\sum_{hkl} w_{hkl}\frac{\left(\frac{\alpha^2|\boldsymbol{F}_{model}|}{\beta} - \frac{\alpha F_{obs}}{\beta}\frac{I_1\left(\frac{2\alpha|\boldsymbol{F}_{model}|F_{obs}}{\beta}\right)}{I_0\left(\frac{2\alpha|\boldsymbol{F}_{model}|F_{obs}}{\beta}\right)}\right)}{|\boldsymbol{F}_{model}|}\left(Re(\boldsymbol{F}_{model})\right.$$

$$\left.* \frac{\partial Re(\boldsymbol{F}_{model})}{\partial x_m} + Im(\boldsymbol{F}_{model}) * \frac{\partial Im(\boldsymbol{F}_{model})}{\partial x_m}\right).$$

We use flat bulk solvent model and we assume that $\frac{\partial \boldsymbol{F}_{bulk\ solvent}}{\partial x_i} = 0$ for further derivations. From one point of view, this shortcut is linked to non-differentiability of the solvent model and is in accordance to the widely adopted practice of most contemporary refinement suits. On the other hand, we justify that simplification due to unordered effect of the bulk solvent. It is important to notice here that this simplification makes the derivatives slightly off when compared to numerical estimations: i.e. one shifts an atom by small value Δ in one direction and to obtain the numerical derivative of the crystallographic target along that dimension with respect to atomic coordinates one uses the standard formula:

$$\frac{\partial T_{X-ray}}{\partial x_m} = \frac{T_{X-ray}(x_m + \Delta) - T_{X-ray}(x_m - \Delta)}{2\Delta}.$$

Importantly, if one keeps the maximum likelihood parameters (α and β) fixed as well as the structure factors scaling coefficients ($k_{overall}$, $k_{isotropic}$, etc.), the numerical derivatives match the semi-analytical ones.

Denoting the m-th term in the direct summation formula $q_m * f_m(s) * \exp\left(-\frac{B_m s^2}{4}\right) * \exp(2i\pi \boldsymbol{r}_m \boldsymbol{s})$ as $\boldsymbol{e}_m$:

$$-\frac{\partial T_{X-ray}}{\partial x_m}$$

$$= -2k_{overall}\sum_{hkl}\left[\frac{k_{isotropic}k_{anisotropic}\left(\frac{\alpha^2|\boldsymbol{F}_{model}|}{\beta}-\frac{\alpha F_{obs}}{\beta}\frac{I_1\left(\frac{2\alpha|\boldsymbol{F}_{model}|F_{obs}}{\beta}\right)}{I_0\left(\frac{2\alpha|\boldsymbol{F}_{model}|F_{obs}}{\beta}\right)}\right)}{|\boldsymbol{F}_{model}|}\left(Re(\boldsymbol{F}_{model})\right.\right.$$

$$\left.\left.*\frac{\partial Re(\boldsymbol{F}_{macromolecule})}{\partial x_m}+Im(\boldsymbol{F}_{model})*\frac{\partial Im(\boldsymbol{F}_{macromolecule})}{\partial x_m}\right)\right]$$

$$= -4\pi k_{overall}\sum_{hkl}\left[\frac{hk_{isotropic}k_{anisotropic}\left(\frac{\alpha^2|\boldsymbol{F}_{model}|}{\beta}-\frac{\alpha F_{obs}}{\beta}\frac{I_1\left(\frac{2\alpha|\boldsymbol{F}_{model}|F_{obs}}{\beta}\right)}{I_0\left(\frac{2\alpha|\boldsymbol{F}_{model}|F_{obs}}{\beta}\right)}\right)}{|\boldsymbol{F}_{model}|}\left(-Re(\boldsymbol{F}_{model})\right.\right.$$

$$\left.\left.*Im(\boldsymbol{e}_m)+Im(\boldsymbol{F}_{model})*Re(\boldsymbol{e}_m)\right)\right].$$

Here, again all sums are taken over the working set of Miller indices, and h is the first component of the vector of Miller indices. Similarly, one can write the crystallographic force term expression along y and z dimensions.

2.4 Methods

2.4.1 Test structures selection criterion

To reduce computational time and to avoid individual treatment of the deposited models, we scanned the PDB for the protein structures, which meet the following conditions:

1. experimental data are deposited (either structure factors or scattering intensities),
2. no twinning is present,
3. structure mass is lower than 40kDa per asymmetric unit,
4. asymmetric unit contains only protein chains without modified residues, ligands or gaps (i.e. at least one backbone heavy atom per residue must be present, sidechain atoms might be absent),
5. atomic occupancies are all equal to 1.0,
6. unit cell size is less than 200000 $Å^3$,
7. unit cell dimensions are large enough to comprise a doubled non-bonded cutoff radius of the default 8.0 Å (see [126]),
8. number of water molecules is no more than 50 per asymmetric unit.

Condition (2) would lead to a modification of the target function and its derivatives, which we do not have implemented at the time. Condition (4) is intended to avoid rebuilding missing protein parts and derivation of non-standard parameters for Amber force field. Condition (7) is coupled with the GPU code of Amber, which is currently deemed to be unsafe in situations not matching the requirement. Condition (8) is similar to the condition about gaps in proteins and was introduced to avoid the bias by structured solvent since currently there is no solution on how to treat it in the body of Amber source code. In the future, the crystallographic water or ligands might be handled by estimating electron density maps in a fashion introduced in [60].

Using the criterion above, we ended up with 84 structures of different resolution and geometric qualities from different space groups. We have assessed different refinement setups on this set or, in certain cases, on a subset of 74 structures where R_{free} value was available.

2.4.2 Preparation of input files for refinement

We removed crystallographic water molecules from the deposited pdb-files if present. Then, we rebuilt missing heavy atoms and hydrogens (see paragraph 2.4.1, condition (4)).

Further, we assigned B-factors: the values for the previously missing heavy atoms were put as of their preceding neighbors, the values for the missing hydrogens were standardly set to their bearers as done in Phenix suite. B-factors optimization usually is performed on a different step in the overall pipeline of refinement rather than the coordinate refinement. Thus, we did not pursue that goal and left the published B-factors fixed. One should notice that these procedures are similar to those suggested by Burnley et al. [60] for ensemble refinement.

We considered two major cases of initial models of structures to be refined:

1. Mimicking real life – three different initial conformation sets are prepared from the deposited one by deforming it with regular MD for 100 ps (MD1 set), 1 ns (MD2 set), 10 ns (MD3 set). The MD simulations are preceded by 20 ps period of heating of the system and followed by 10 ps cooling.
2. Improvement of the deposited model (D set).

MD1 set of models RMSD over $C\alpha$ atoms from the deposited model is 0.75 Å and MolProbity percentile is 96% on average. MD2 set structures have 0.89 Å RMSD and 98.05% MolProbity score percentile on average. MD3 set has 1.02 Å RMSD and 98.16% MolProbity score percentile. Generally, the longer the MD, the larger the RMSD was, but that was not always the case. The summary of RMSD values for the distorted structures can be found in Figure 2.1. Our focus is on the D and MD1 sets, however, MD2 and MD3 sets do provide some insights into the benefits of the radius of convergence of Amber/Amber refinement, which are discussed below.

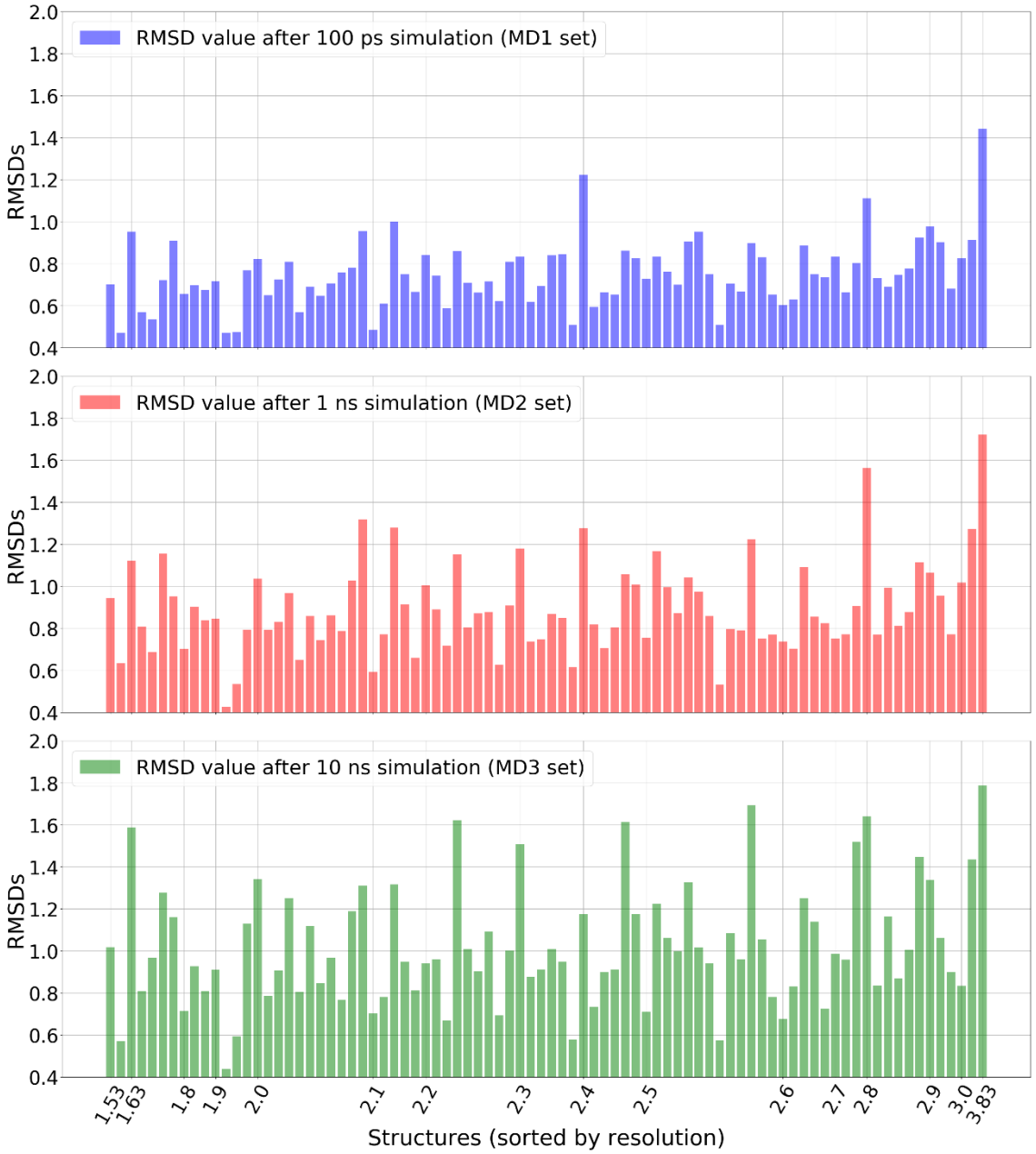

Figure 2.1. The plot represents root mean square deviations over Ca atoms of the distorted models (MD1, MD2, MD3 sets) against the deposited models (D set).

The exact parameters and flags used during regular MD simulations to obtain the distorted models and during crystallographic refinement are the following and in accordance with the recommended Amber settings [126]:

1. ntb = 1, periodic boundary conditions, constant volume,
2. ntp = 0, flag for constant pressure dynamics: no pressure scaling is applied,

3. to employ TIP3P water model:
    a. ntc = 2, flag for SHAKE algorithm: bonds involving hydrogen are constrained,
    b. ntf = 2, force evaluation method: bond interactions involving H-atoms omitted,
4. ntt = 3, Langevin dynamics,
5. gamma_ln = 2., collision frequency in ps^{-1}: small value is advantageous in terms of sampling or stability of integration,
6. cut = 8.0, non-bonded interactions cutoff radius in Angstroms,
7. dt = 0.002, time step in ps,
8. temp0 = 298.0, reference temperature in Kelvins.

During the heating the following parameters were used:

1. ntr = 1, restrain protein 10.0 kcal/mol,
2. ntt = 1, switch for temperature scaling: constant temperature, using the weak-coupling algorithm.

As for the experimental data, we used *phenix.cif2mtz* routine to unify the format of all the deposited files. By the end of the procedure we prepared the files in such a way that:

1. the intensities were converted to structure factors, if present,
2. the fraction of R_{free} factors were adjusted to 10% of all reflections,
3. structure factors were expanded from the original space group to P1 group.

Since we suggest the whole unit cell approach, we needed the expansion of structure factors (condition (3)) to maintain the data-to-parameters ratio used in refinement. The importance of this ratio is overviewed in the introductory CHAPTER 1. In an ideal case, we would use the raw data before the reduction due to symmetries in the space group, which is not affected by averaging. Such example is covered in paragraph 2.10.

2.4.3 Main Amber-based refinement protocol

For the refinement with our Amber modification, we used the following basic protocol (we call it Amber/Amber setup further in the text, see also Figure 2.2):

1. rebuild whole unit cell, add counter ions to neutralize the system, and place explicit water molecules into the voids,
2. minimize the energy of such water box over 500 steps,
3. heat the system up to the room temperature over 10,000 steps (20 ps),

4. refine the structure for 5,000 steps with evenly increasing X-ray energy term weight *w* from 0 to 1 (10 ps),
5. Gradually cool down the model over 5,000 steps with the constant unit weight (10 ps).

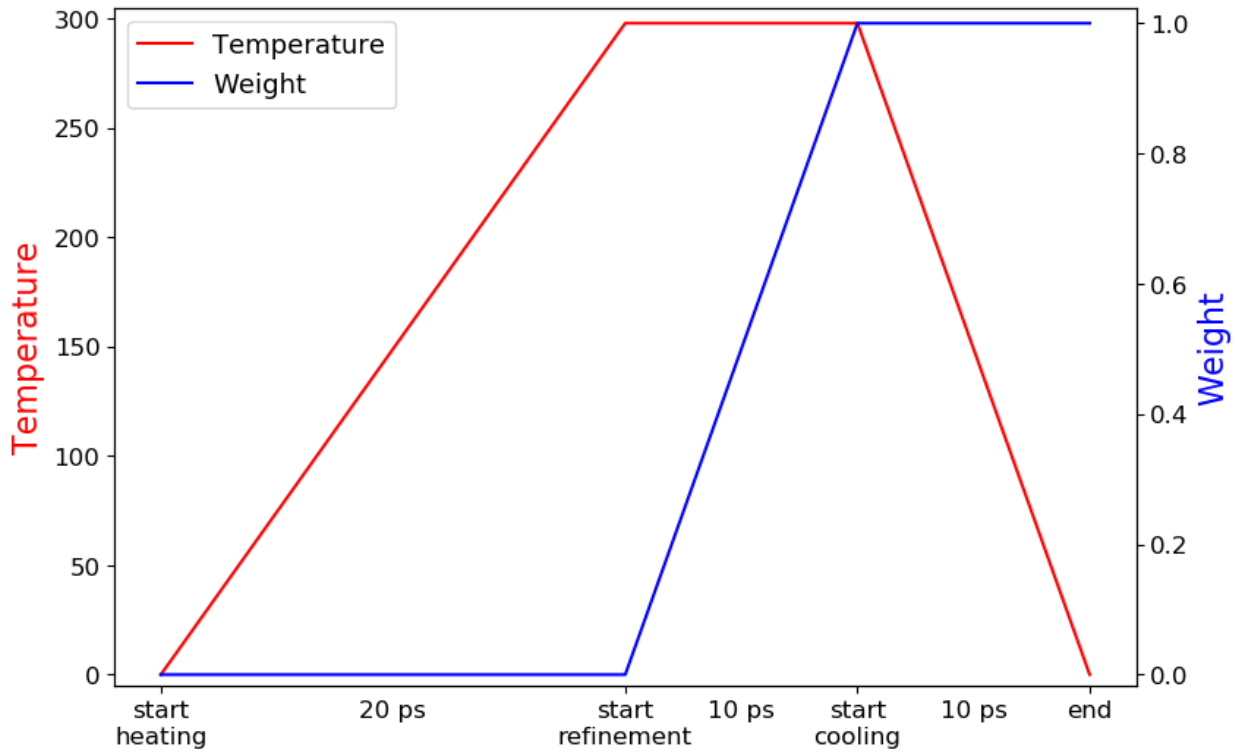

Figure 2.2. Representation of our general Amber-based refinement protocol: temperature and X-ray term weight control during refinement.

Steps (1)-(3) are preparatory while (4) and (5) perform the crystallographic refinement. As one can see from equation (2.2), the maximum likelihood target function depends on the parameters α and β that are estimated based on the model and experimental structure factors. As such, during steps (4) and (5) we updated these parameters every 100 steps along with bulk solvent mask using the aforementioned scaling *cctbx* routine and the estimator for the uncertainty parameters [137], [138]. In the follow-up paragraph, we will justify our choice of the weight.

At the end of the refinement, we remove explicit water molecules from the simulation box that we added initially. It is worth noting that we use these molecules to perform MD simulations and properly model non-bonded interactions of the force-field component of the target function. Importantly, to calculate instantaneous structure factors and the crystallographic term of the

target function we need to use bulk solvent modelling instead. Therefore, after we possess a refined protein model, we autodetect missing crystallographic water with Phenix procedure using $mF_{obs} - DF_{model}$ maps [23]. This manipulation adds bound molecules into the macromolecule component of structure factors equation (2.1), consequently, it results in a slightly different bulk-solvent mask. Such manipulation helps to achieve better agreement with the experimental structure factors (reduce R-factors) while maintaining the geometric qualities of the macromolecular model. After such the bound water addition, we also properly compare our results with the deposited data, which also have crystallographic waters. Such protocol was run two times to collect more comprehensive statistics since the results are dependent on the initial MD random seed.

Aside from the standard MD simulation parameters covered in paragraph 2.4.2 and the crystallographic term weight, there are only two tunable variables: the length of refinement and the non-bonded interactions cutoff radius. Their choice is discussed in paragraph 2.5.

2.4.4 Amber/Amber selection of crystallographic weight

The selection of the crystallographic weight, w, appears to be critical. Phenix and other popular refinement packages have a common built-in procedure to compute the X-ray term weight, which is based on gradients calculations of each term of the target function with respect to atomic parameters [18], [139]. In the case of our Amber modification, we tried the implementation of this function and modified the refinement protocol described in paragraph 2.4.3 as follows: in the beginning of stage (4) of Amber/Amber setup we calculate the weight as in Phenix, then start with the zero weight and by the end of the stage we reach the estimate. Afterwards, we re-calculate the weight again and use it during stage (5). Such choice appears to be sub-optimal and significantly depends on the quality of the initial structure.

Thus, we have implemented the approach proposed in [69] where the crystallographic term has unity weight as the *a priori* geometry term. Briefly, the original crystallographic maximum likelihood target is formulated as a conditional probability of observing the measured data (F) given an atomic model (X). Therefore, the Bayes' rule can be applied to the function that is being maximized during refinement since the measured data (F) given instead:

$$P(X|F) = P(F|X)P(X).$$

$P(X)$ is the prior probability of the atomic model and can be given as Boltzmann factor $\exp(-\frac{T_{a\ priori}(X)}{kT})$, where $T_{a\ priori}$ is the potential energy of the crystal (in our case, the Amber force field). Finally, since we minimize the negative logarithm of the probability, the overall target expands into the following formula:

$$T(X) = -\log\big(P(F|X)\big) + \frac{T_{a\ priori}(X)}{kT}.$$

In these terms, the maximum probability problem suggests that the weight of the crystallographic term of the target function should be equal to kT. At 300K it is equals to 0.6. Further, their grid search tests proved that the choice of 1.0 provides a good balance between the terms of the crystallographic component (2.2) and Amoeba force field component of the target function. Hence, we increase the weight from 0.0 to 1.0 during stage (4) and maintain constant 1.0 weight during stage (5) (Figure 2.2).

2.4.5 Phenix-based protocols

Due to the ability of Phenix to utilize Amber force field, we decided to compare our results with the results of this program. We used the default Phenix refinement schedule except the following options to achieve a one-to-one comparison between the abilities of Phenix and our Amber-based procedure:

- 5 macrocycles (we have tried 3, 5, and 7 macrocycles and the increase from 5 to 7 macrocycles did not show considerable improvement),
- individual hydrogen coordinates refinement (since the presence of explicit hydrogens is required by Amber force field, as well as it is also known to improve model geometry),
- no B-factors refinement,
- no occupancies refinement,
- maximum likelihood target function,
- turned on direct summation formula for structure factors calculation.

Further, we introduced the following variations to this basic protocol. First, simulated annealing dynamics could be performed in both Cartesian and Torsion angles spaces [140], giving 4 options: sequential torsion angles and Cartesian dynamics (full SA for short), only Cartesian dynamics, only torsion angles dynamics (TAD for short), absence of any dynamics (no

SA). Secondly, one can also employ a more sophisticated approach to crystallographic term weight instead of the default gradient-based procedure described above. Namely, grid search target function T weight optimization (WO) scheme. Finally, one can choose between two available gradients: the built-in Phenix gradient and the Amber gradient. We refer to them as Phenix/Phenix and Phenix/Amber setups, respectively. Such modifications to the basic configuration provide us a total of 16 different protocols. As with the Amber/Amber setup, crystallographic bound water molecules were added at the end of each setup.

2.4.6 Refinement results evaluation criterion

During initial experiments we considered using so called Q-score for the assessment of structure quality: a combined measure of R_{free} and MolProbity score introduced in [141] to select the best result when R values are closely distributed over the set of refined structures:

$$Q = R_{free} + c(MP^{max} - MP),$$

where MP is the MolProbity score percentile of the refined structure and MP^{max} is the maximum MolProbity percentile among all considered protocols. The weight c is the ratio between the ranges of R factors and MolProbity score percentiles:

$$c = \frac{R_{free}^{max} - F_{free}^{min}}{MP^{max} - MP^{min}}.$$

However, this measure worked inappropriately with our results due to poor R_{free} values of the failed Phenix setups, hence, giving a huge gap between the best and the worst results, thus making the best results indistinguishable. More on that issue will be discussed further. We, therefore, opted to a simple comparison of the three characteristics: R_{free} being the primary one, MolProbity score and MolProbity score percentile being the secondary. MolProbity score percentiles are calculated based on the PDB statistics of the structures which have resolution within ±0.1A margin of the evaluated structure.

2.5 Auxiliary results: Amber-based performance

2.5.1 Influence of the length of refinement MD

In our basic Amber/Amber setup we have two periods of 10 ps when the crystallographic force is being applied (see paragraph 2.4.3, steps (4) and (5)). For this paragraph we denote this protocol as “10 ps + 10 ps” and following the same pattern we named others, for example, 10 ns of step (4) and 100 ps of step (5) correspond to “10 ns + 100 ps” nomenclature. We have also tried to increase the two periods to evaluate the necessary length of the refinement. We have tested these variations on five random test structures with the initial models from the MD1 set. We found that longer refinement might sometimes improve the results both in terms of R_{free} and MolProbity, however, not significantly and not consistently. Therefore, we decided to use the shortest “10 ps + 10 ps” protocol. The full comparison is in Table 2.1 where the structures are sorted by the volume of the corresponding unit cell.

Table 2.1. Summary of Amber/Amber performance for selected structures when using different length of refinement.

						Ramachandran (%)		RMSD			
Refined strucutures	R-work	R-free	R-free – R-work	Clashscore	Poor rotamers (%)	outliers	favored	bonds	angles	Molprobity score	Molprobity Percentile
3K9P											
10 ps + 10 ps trial 1	0.227	0.289	0.062	0.12	2.32	0.19	96.44	0.0146	2.19	1.06	98.42
10 ps + 10 ps trial 2	0.227	0.288	0.062	0.23	2.53	0.19	96.82	0.0145	2.20	1.09	98.17
100 ps+100 ps trial 1	0.218	0.286	0.068	0.00	1.27	0.19	96.07	0.0135	2.08	0.85	99.57
100 ps+100 ps trial 2	0.222	0.286	0.064	0.23	1.48	0.19	97.38	0.0134	2.10	0.84	99.59
100 ps+100 ps trial 3	0.222	0.289	0.067	0.35	1.48	0.19	97.38	0.0134	2.08	0.88	99.47
100 ps+100 ps trial 4	0.226	0.290	0.064	0.12	2.11	0.19	96.44	0.0135	2.12	1.03	98.59
2 ns+100 ps trial 1	0.218	0.291	0.073	0.12	1.69	0.19	97.75	0.0136	2.06	0.77	99.80
2 ns+100 ps trial 2	0.217	0.284	0.067	0.12	1.48	0.19	97.94	0.0136	2.08	0.69	99.91
2 ns + 2 ns	0.226	0.298	0.072	0.23	1.48	0.37	97.38	0.0134	2.06	0.84	99.59
10 ns+100 ps trial 1	0.217	0.287	0.071	0.35	1.27	0.19	97.94	0.0136	2.09	0.72	99.89
10 ns+100 ps trial 2	0.216	0.286	0.070	0.23	1.05	0.19	97.19	0.0134	2.07	0.75	99.83
2J7I											
10 ps + 10 ps trial 1	0.256	0.294	0.038	0.23	2.08	0.20	98.21	0.0148	2.12	0.83	99.62
10 ps + 10 ps trial 2	0.262	0.302	0.040	0.11	3.54	0.40	98.41	0.0150	2.11	0.96	99.30
100 ps+100 ps trial 1	0.249	0.282	0.033	0.00	2.92	0.20	98.81	0.0140	2.03	0.85	99.59
100 ps+100 ps trial 2	0.250	0.284	0.034	0.00	1.46	0.20	98.61	0.0141	2.04	0.62	99.95
100 ps+100 ps trial 3	0.256	0.292	0.037	0.00	1.46	0.20	97.82	0.0143	2.04	0.67	99.87
100 ps+100 ps trial 4	0.250	0.284	0.035	0.00	1.67	0.20	98.41	0.0141	2.03	0.67	99.87

Table 2.1 continued.

2 ns+100 ps trial 1	0.238	0.274	0.036	0.00	1.88	0.20	98.02	0.0139	1.98	0.71	99.87
2 ns+100 ps trial 2	0.237	0.272	0.035	0.00	2.92	0.20	97.62	0.0140	2.00	0.93	99.41
10 ns+100 ps trial 1	0.237	0.279	0.042	0.23	1.04	0.20	98.61	0.0139	1.97	0.60	99.97
10 ns+100 ps trial 2	0.236	0.274	0.037	0.00	2.08	0.20	98.81	0.0138	1.98	0.74	99.86
4UG3											
10 ps + 10 ps trial 1	0.233	0.285	0.052	0.26	0.69	0.23	97.75	0.0148	2.03	0.65	99.94
10 ps + 10 ps trial 2	0.231	0.279	0.048	0.00	0.93	0.00	97.52	0.0148	2.04	0.60	99.97
100 ps+100 ps trial 1	0.228	0.280	0.052	0.00	1.62	0.00	97.75	0.0139	1.93	0.72	99.89
100 ps+100 ps trial 2	0.228	0.288	0.060	0.00	1.16	0.23	97.07	0.0137	1.94	0.71	99.91
100 ps+100 ps trial 3	0.228	0.285	0.057	0.00	1.39	0.23	97.97	0.0137	1.93	0.62	99.96
100 ps+100 ps trial 4	0.228	0.280	0.052	0.00	1.16	0.00	97.75	0.0139	1.94	0.60	99.97
2 ns+100 ps trial 1	0.223	0.279	0.056	0.00	0.46	0.23	97.30	0.0135	1.89	0.63	99.96
2 ns+100 ps trial 2	0.223	0.277	0.054	0.00	1.39	0.00	97.30	0.0137	1.91	0.74	99.86
10 ns+100 ps trial 1	0.227	0.278	0.051	0.00	0.93	0.23	96.85	0.0139	1.92	0.69	99.91
10 ns+100 ps trial 2	0.227	0.279	0.053	0.00	0.23	0.00	97.52	0.0138	1.92	0.60	99.97
4C0M											
10 ps + 10 ps trial 1	0.269	0.316	0.047	0.11	0.66	0.00	98.92	0.0151	2.12	0.55	99.98
10 ps + 10 ps trial 2	0.268	0.311	0.043	0.11	0.88	0.00	98.74	0.0147	2.10	0.55	99.98
100 ps+100 ps trial 1	0.266	0.316	0.049	0.23	1.32	0.00	99.64	0.0139	2.05	0.68	99.92
100 ps+100 ps trial 2	0.269	0.322	0.053	0.11	0.88	0.00	98.74	0.0139	2.04	0.55	99.98
2 ns+100 ps trial 1	0.264	0.317	0.053	0.11	0.66	0.00	99.28	0.0139	2.04	0.55	99.98
2 ns+100 ps trial 2	0.266	0.319	0.053	0.00	0.22	0.00	98.74	0.0137	2.02	0.50	100.00

Table 2.1 continued.

2 ns + 2 ns	0.271	0.321	0.050	0.11	0.22	0.00	98.38	0.0136	2.02	0.55	99.98
10 ns+100 ps trial 1	0.264	0.320	0.056	0.00	1.32	0.00	98.20	0.0140	2.04	0.59	99.97
10 ns+100 ps trial 2	0.267	0.319	0.053	0.23	1.10	0.00	98.38	0.0139	2.03	0.62	99.96
4BHC											
10 ps + 10 ps trial 1	0.202	0.257	0.055	0.00	0.58	0.31	95.86	0.0153	2.18	0.79	99.74
10 ps + 10 ps trial 2	0.198	0.246	0.048	0.10	1.16	0.15	95.71	0.0151	2.13	0.89	99.39
100 ps+100 ps trial 1	0.191	0.238	0.047	0.00	1.36	0.31	95.55	0.0143	2.05	0.91	99.31
100 ps+100 ps trial 2	0.194	0.242	0.048	0.10	1.94	0.15	96.93	0.0142	2.04	0.94	99.27
2 ns+100 ps trial 1	0.197	0.252	0.055	0.10	1.55	0.46	96.63	0.0147	2.09	0.90	99.37
2 ns+100 ps trial 2	0.195	0.250	0.055	0.00	1.55	0.15	96.78	0.0146	2.06	0.84	99.59
10 ns+100 ps trial 1	0.195	0.246	0.051	0.00	2.71	0.15	96.63	0.0141	2.03	1.05	98.48
10 ns+100 ps trial 2	0.199	0.247	0.048	0.00	2.52	0.46	96.01	0.0146	2.10	1.08	98.31

2.5.2 Non-bonded interactions cutoff: 8 Å vs 10.5 Å

Next, we tested another parameter of the refinement MD steps that can be tweaked, cutoff for non-bonded interactions. Unlike Xplor, where non-bonded interactions are truncated completely out of the cutoff radius, Amber uses a particle mesh Ewald scheme (PME) [142] since periodic boundary conditions are also employed. In brief, the non-bonded energy (both electrostatics and van der Waals terms) are calculated explicitly inside the cutoff radius, and reciprocal space is then used to calculate the energy outside the cutoff. In some MD software, such as CHARMM or CPU version of Amber, the developers allow to select different cutoffs for electrostatics and van der Waals forces when PME is in actions. However, a single value is used instead in the GPU version of Amber, which we adapted for the refinement.

In paragraph 2.4.1, we described the test structure selection criterion. However, the 84 structures are reduced to 58 if one chooses 10.5 Å non-bonded cutoff due to condition (6) in the GPU implementation so that the unit cell is sufficiently large. Table 2.2 shows the comparison of Amber/Amber setup when using 8 Å (our basic option) and 10.5 Å cutoffs with the initial models devised in paragraph 2.4.2 (D, MD1, MD2, MD3 sets).

There is no apparent correlation between the cutoff and MolProbity score. The differences in the geometric qualities of the refined structures are marginally small, less than 1% of MolProbity score percentile on average. The range of the differences across the 58 structures from all 4 initial structures sets is from -22.3% to 18.4%. As in Table 2.2, negative difference indicates the advantage of the larger non-bonded cutoff, and positive difference indicates the advantage of smaller non-bonded cutoff.

Table 2.2. Summary of Amber/Amber refinement performance depending on the non-bonded interactions cutoff radius. Average differences are calculated between the best out of two runs for each of the cutoff values. Positive values in the differences indicate the advantage of smaller 8 Å cutoff, and negative values of the differences indicate larger 10.5 Å cutoff advantage.

Initial structures set	Average difference in R_{free}	Average difference in MolProbity percentile	Number of structures with 8A cutoff best result	Number of structures with 10.5A cutoff best result
D	0.0021	-0.0354	43	15
MD1	0.0020	0.2175	40	18
MD2	-0.0014	-0.8408	23	35
MD3	-0.0039	-0.0062	23	35

The larger cutoff radius tends to improve the agreement with experimental data when the initial structures have a larger RMSD to the deposited model. It is worth to be mentioned though that the 10.5 Å cutoff also helped refinement to converge in several cases from MD3 set of initial models (with the biggest RMSD from the deposited models) where 8 Å cutoff setting failed. Therefore, the larger cutoff radius proved to be beneficial in certain cases.

Since we focus on Deposited and MD1 sets of initial structures, we selected 8 Å cutoff results to compare with the performance of Phenix-based setups.

2.6 Auxiliary results: Phenix-based refinement using single asymmetric units and whole unit cells

As we have mentioned on multiple occasions, traditional refinement is performed using a single asymmetric unit (ASU) of the unit cell. Multiple studies have shown that ensemble refinement has a number of advantages that we covered in the introduction chapter. However, in such setups structure factors of ensemble models are being averaged and refined against the experimental ones, but the multiple conformations are independent of each other. To the best of our knowledge, we performed the first ensemble refinements in explicit unit cells (UC). This way, the explicit condition on different conformers in unit cells are implied: they must physically co-exist in the crystal. Here, unlike our Amber/Amber protocol, the periodic boundary conditions cannot be applied by the design of the program.

With this idea in mind, we first have compared *phenix.refine* performance in the two cases: ASU refinement and UC refinement on the set of deposited models (D set of initial models). We have selected the best of the 16 protocols described in paragraph 2.4.5 based on R_{free} value for each of these categories. The UC approach achieved better results than the ASU approach in terms of R_{free} in 64 cases out of 84. The average difference in R_{free} between the two setups is 0.011 in favor of the UC approach. Interestingly, the geometric qualities of the refined models vary quite significantly but the average difference in MolProbity percentiles is negligibly small, -0.3%, given the range of the differences, from -50.2% to 62.7%. Figure 2.3 depicts structure-wise comparison of the results between ASU and UC refinement on the D set.

However, the longer the deposited model undergoes the MD to become the initial one for refinement, the better become both the R_{free} factors and MolProbity score percentiles in ASU-refined model rather than in UC-refined models. On average the difference values between the

UC approach and ASU approach are: 0.006 and -4.00%, 0.000 and -12.21%, -0.016 and -15.37% for MD1, MD2, MD3 sets, respectively.

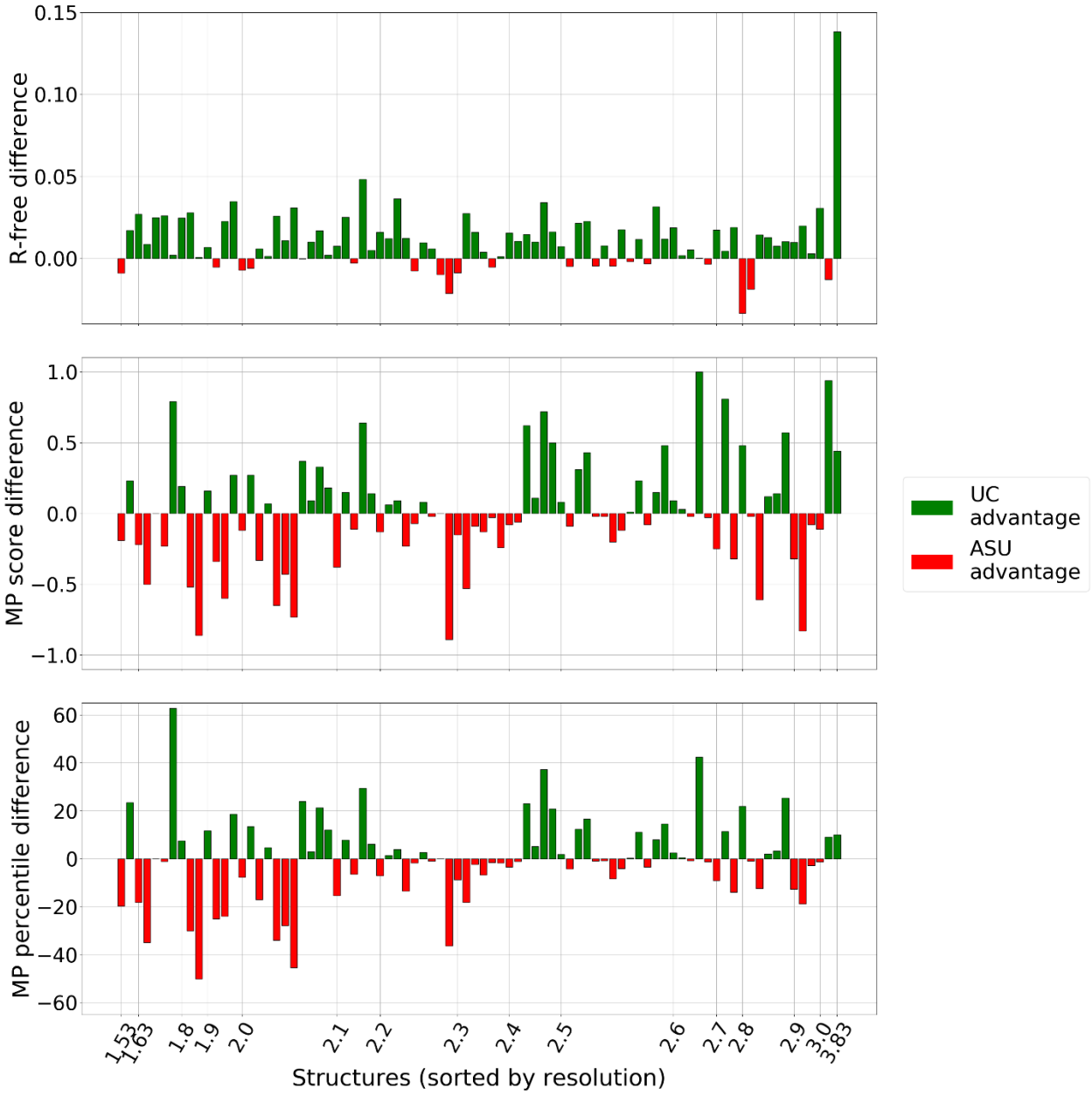

Figure 2.3. Summary of *phenix.refine* performance using single asymmetric unit (ASU) and whole unit cell (UC) with the deposited models as initial ones. Green bars indicate the advantage of UC approach, red bars indicate the advantage of ASU approach.

2.7 Main Results: Comparison of Amber-based and Phenix-based protocols

Here we proceed with the main results of our investigations: the comparison of our Amber/Amber setup to the classic scheme of ASU Phenix refinement and the proposed UC Phenix refinement. After the comparison, we highlight some advantages of our method. Next, we showcase an application of Amber/Amber setup to the real problem: we were able to obtain experimental data from our collaborators before the reduction due to space group symmetry to refine the structure of GRB2 adaptor protein with MPD co-crystallization factor. Finally, we conclude with a brief overview of the web-service based on our Amber refinement module.

2.7.1 Example of refinement comparison, the case of 3K9P

To show how we compare Amber-based and Phenix-based refinement, we selected the structure with PDB code 3K9P. This is a structure of ubiquitin-conjucating enzyme and ubiquitin complex. The lengths of proteins are 217 and 79 residues, respectively. The space group is P 1 2_1 1, the resolution is 2.8 Å and the reported R_{free} is 0.296. We chose the initial model from the MD1 set and first compare with the ASU approach of Phenix to demonstrate how we compared our Amber-based refinement with Phenix-based setups.

We ran Amber/Amber protocol twice with the only difference in the initial random seed and the 16 Phenix-based protocols. The best run was selected in each category. The following Table 2.3 summarizes the results. The selected cells represent the best protocols. Thus, for example, the best Phenix-based protocol turned out to be the one using Amber14 force field with torsion angles dynamics and the standard gradient-based weight for the crystallographic terms.

We should mention that the best Phenix-based protocol varies from one structure to another. For example, one can find that the protocol with CDL geometry restraints without any simulated annealing worked the best for the N-terminal SH3 domain of GRB2 (see paragraph 2.7.6).

Using such a scheme for the comparison, we evaluated our Amber/Amber setups on the four sets of initial models (D, MD1, MD2, MD3) across the 84 structures.

Table 2.3. Comparison of the refinement results of the corrupted initial model for 3K9P structure. The best results in each category are highlighted with boxes. P/A – Phenix with Amber14 force field, P/P – Phenix with Phenix force field, SA – simulated annealing, TAD – torsional angle dynamics, Cartesian – Cartesian dynamics, WO – weight optimization. The best Amber/Amber run is the second trial. Phenix-based protocol with Amber14 force field with torsional angles dynamics without weight optimization is the best.

	R-work	R-free	R-free - R-work	Clash score	Poor rotamers (%)	Ramachandran outliers	Ramachandran favored	Molprobity score	Molprobity percentile
MD1 initial structure	0.419	0.422	0.280	0.35	0.63	0.37	96.25	0.88	99.47
Amber results									
Run 1	0.221	0.282	0.061	0.58	2.11	0.19	96.63	1.16	97.42
Run 2	0.227	0.275	0.048	0.23	3.80	0.37	97.19	1.18	97.10
Phenix results									
P/P, SA, WO	0.240	0.326	0.085	15.04	0.00	3.75	83.52	2.37	57.92
P/P, no SA, WO	0.293	0.357	0.064	9.49	0.00	1.87	87.64	2.11	69.03
P/P, no SA, no WO	0.265	0.347	0.082	20.59	0.42	0.37	89.14	2.38	57.41
P/P, SA, no WO	0.227	0.331	0.104	20.82	0.42	5.24	78.65	2.57	48.24
P/A, SA, WO	0.393	0.502	0.109	4.16	0.00	7.12	76.78	1.97	74.41
P/A, no SA, WO	0.286	0.332	0.046	0.69	0.00	0.00	96.63	0.94	99.27
P/A, no SA, no WO	0.236	0.331	0.095	4.86	0.00	1.12	91.76	1.75	82.12
P/A, SA, no WO	0.292	0.417	0.125	14.11	0.00	6.74	79.40	2.40	56.56
P/P, Cartesian, no WO	0.220	0.326	0.107	24.76	0.00	5.24	79.40	2.63	45.40
P/P, Cartesian, WO	0.243	0.337	0.094	17.12	0.42	4.87	81.65	2.45	53.97
P/P, TAD, no WO	0.242	0.340	0.098	39.10	0.00	2.62	82.77	2.77	38.31
P/P, TAD, WO	0.260	0.337	0.077	31.47	1.27	3.75	82.02	2.77	38.31
P/A, Cartesian, no WO	0.285	0.409	0.124	15.97	0.42	8.24	74.53	2.51	51.03
P/A, Cartesian, WO	0.298	0.389	0.091	0.69	0.42	5.24	85.77	1.37	93.05
P/A, TAD, no WO	0.218	0.310	0.092	5.78	0.00	1.50	92.88	1.77	81.43
P/A, TAD, WO	0.263	0.323	0.060	4.63	0.00	0.75	92.88	1.69	84.23

2.7.2 Comparison across the whole test set: ASU case

First, we compared how well Amber-based and Phenix-based setups improve the deposited structure if it is used as the initial model. Figure 2.4 depicts the relative comparison of the resulting values. Clearly, Amber/Amber setup outperformed all the 16 Phenix-based protocols in cases 65 out of 84. The average improvement of R_{free} factor and MolProbity score percentile among the structures is 0.0181 and 9.225%, respectively.

Next, we compared the results of the best Amber-based and Phenix-based setups with the distorted initial model MD1 based on R_{free} value in a similar fashion (Figure 2.5). We remind that the average RMSD over $C\alpha$ atoms from the deposited model was 0.75 Å and the average MolProbity percentile was 96%. The results appeared to be even more impressive as compared with the deposited models set. Amber/Amber protocol produced better R_{free} factors than Phenix-based refinement for 75 out of the 84 structures with the average improvement of 0.0215. The average improvement in terms of geometry quality is 18.555% MolProbity score percentiles.

Following up, we tested the refinement on the MD2 and MD3 sets (0.89 Å and 1.02 Å average RMSDs, 98.05% and 98.16% average MolProbity score percentiles, respectively). We achieved the results analogous to the earlier outcomes (see Figure 2.6 and Figure 2.7). Using MD2 starting model, R_{free} produced by Amber/Amber setup was better in 68 cases out of 84 with the average improvement of 0.0174 and MolProbity score percentile improvement was 15.569%. It should be mentioned that the poorer the starting model (i.e. case of MD3), the worse the performance refinement for both Amber/Amber and Phenix setups. This could be observed from the absolute values of R-factors. The comparison between the setups in this case should be taken with a grain of salt: Amber/Amber setup performs better than *phenix.refine* in 52 cases out of 84 and the R_{free} improvement is 0.0192 and the MolProbity percentile improvement is 16.480%. The figures with absolute values can be found in the Appendix.

The direct comparison of Amber/Amber setup with non-bonded cutoff radius of 10.5 Å has shown no significant improvement when compared to *phenix.refine* as well.

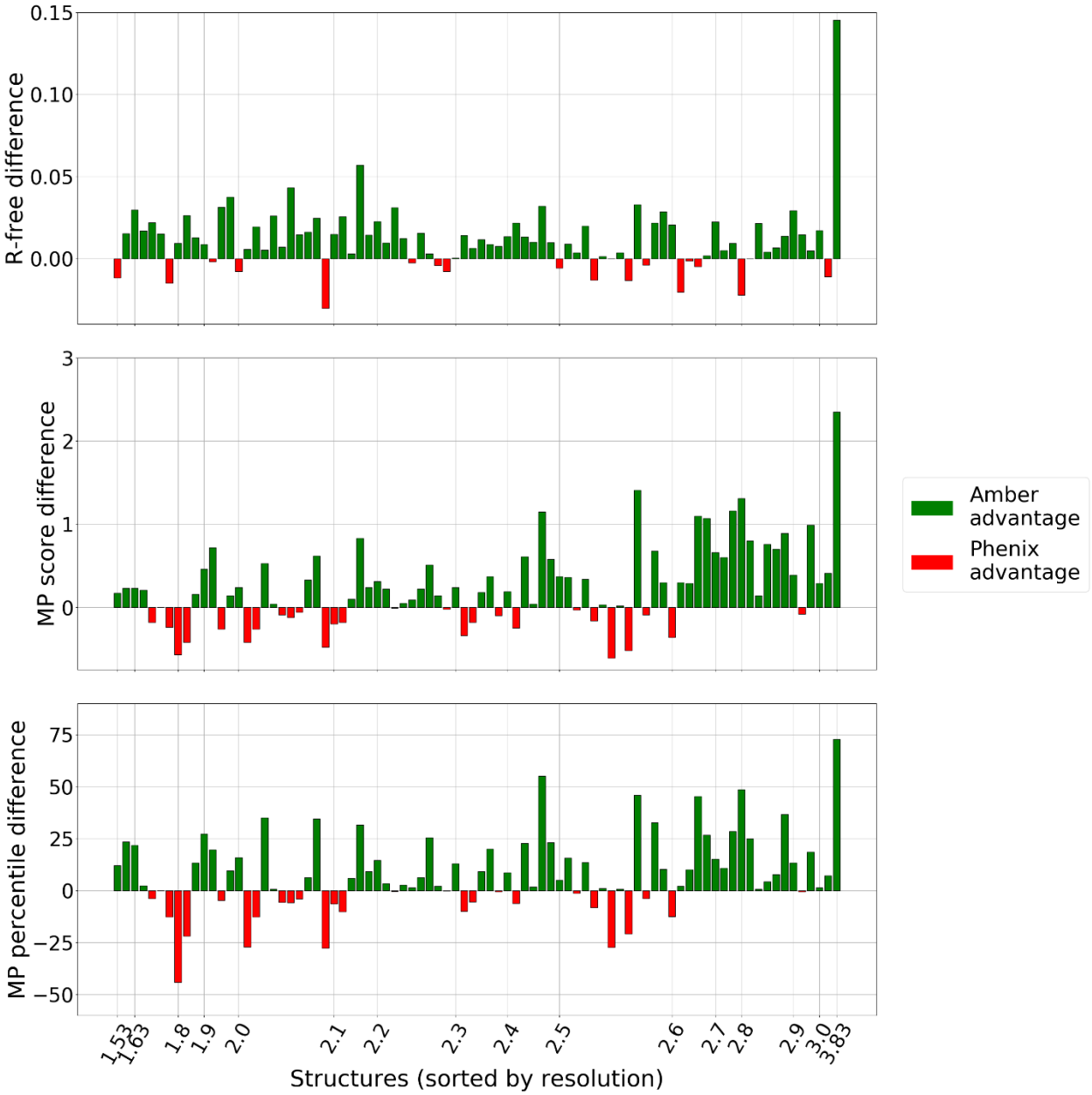

Figure 2.4. The plots show the difference between R_{free} factors, MolProbity scores and MolProbity percentiles of the re-refined deposited models. Green bars represent the superiority of our Amber-based setup. Red bars represent the superiority of Phenix-based ASU setups.

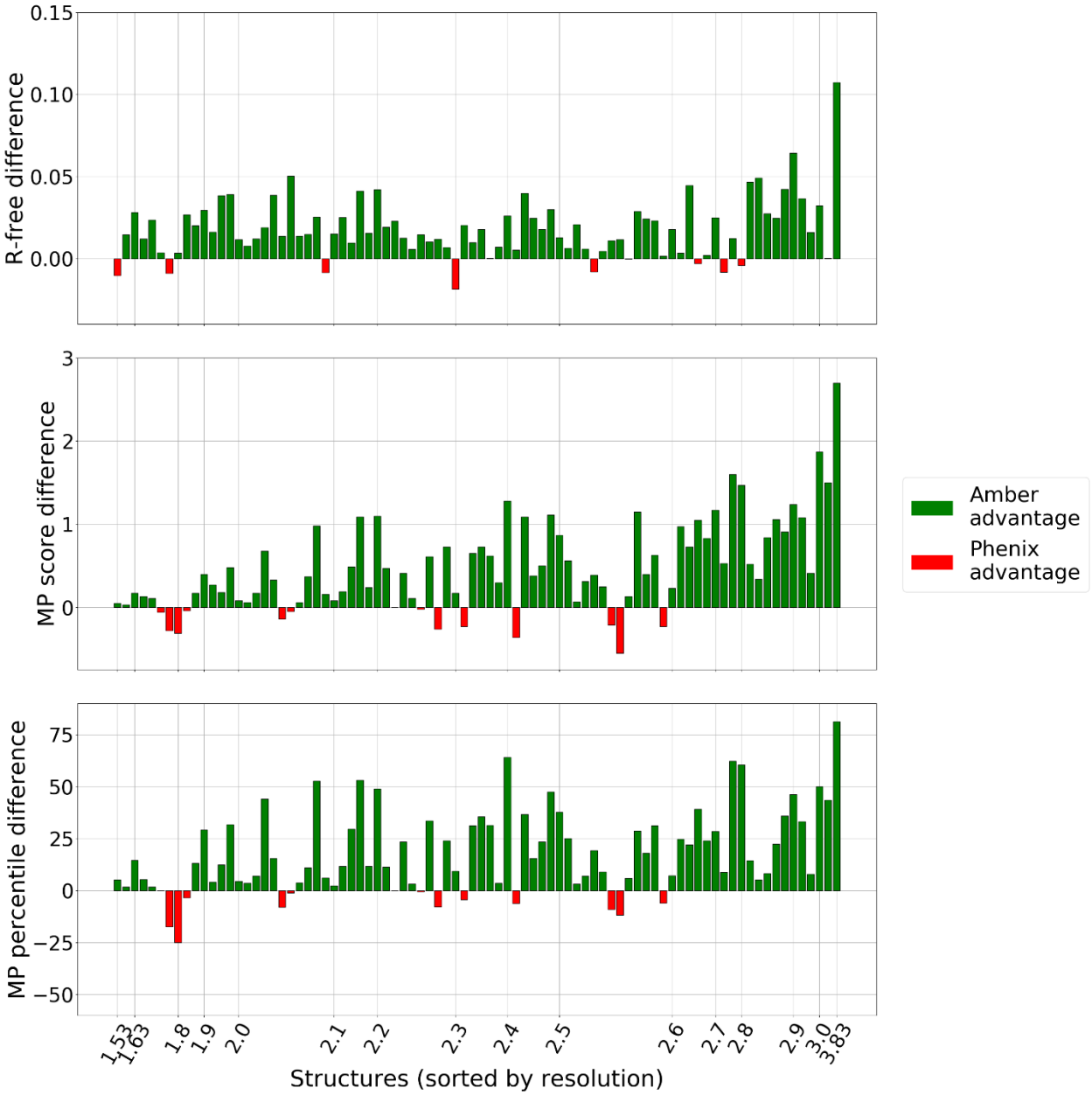

Figure 2.5. The plots show the difference between R_{free} factors, MolProbity scores and MolProbity percentiles of the refined MD1 models. Green bars represent the superiority of our Amber-based setup. Red bars represent the superiority of Phenix-based ASU setups.

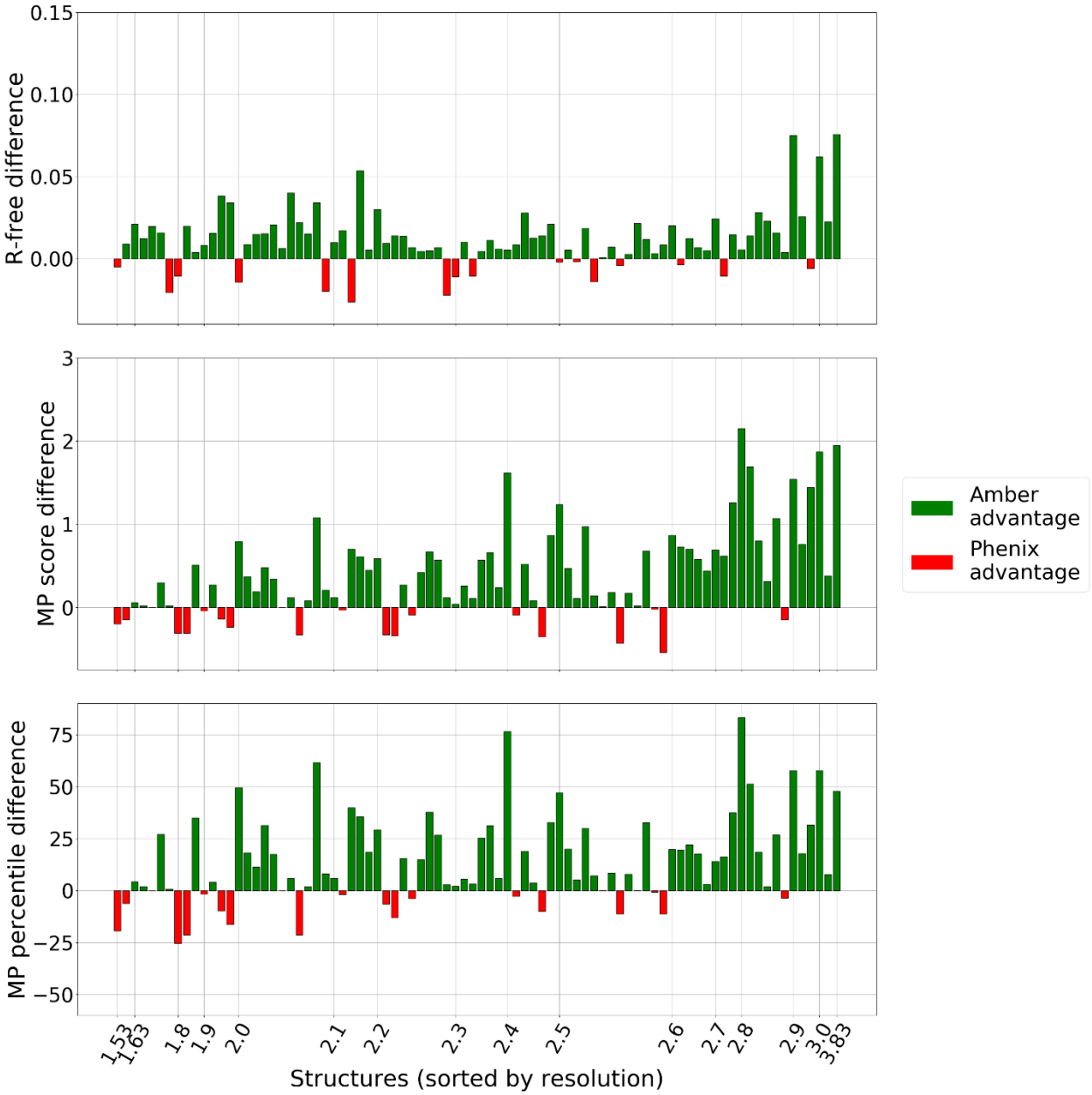

Figure 2.6. The plots show the difference between R_{free} factors, MolProbity scores and MolProbity percentiles of the refined MD2 models. Green bars represent the superiority of our Amber-based setup. Red bars represent the superiority of Phenix-based ASU setups.

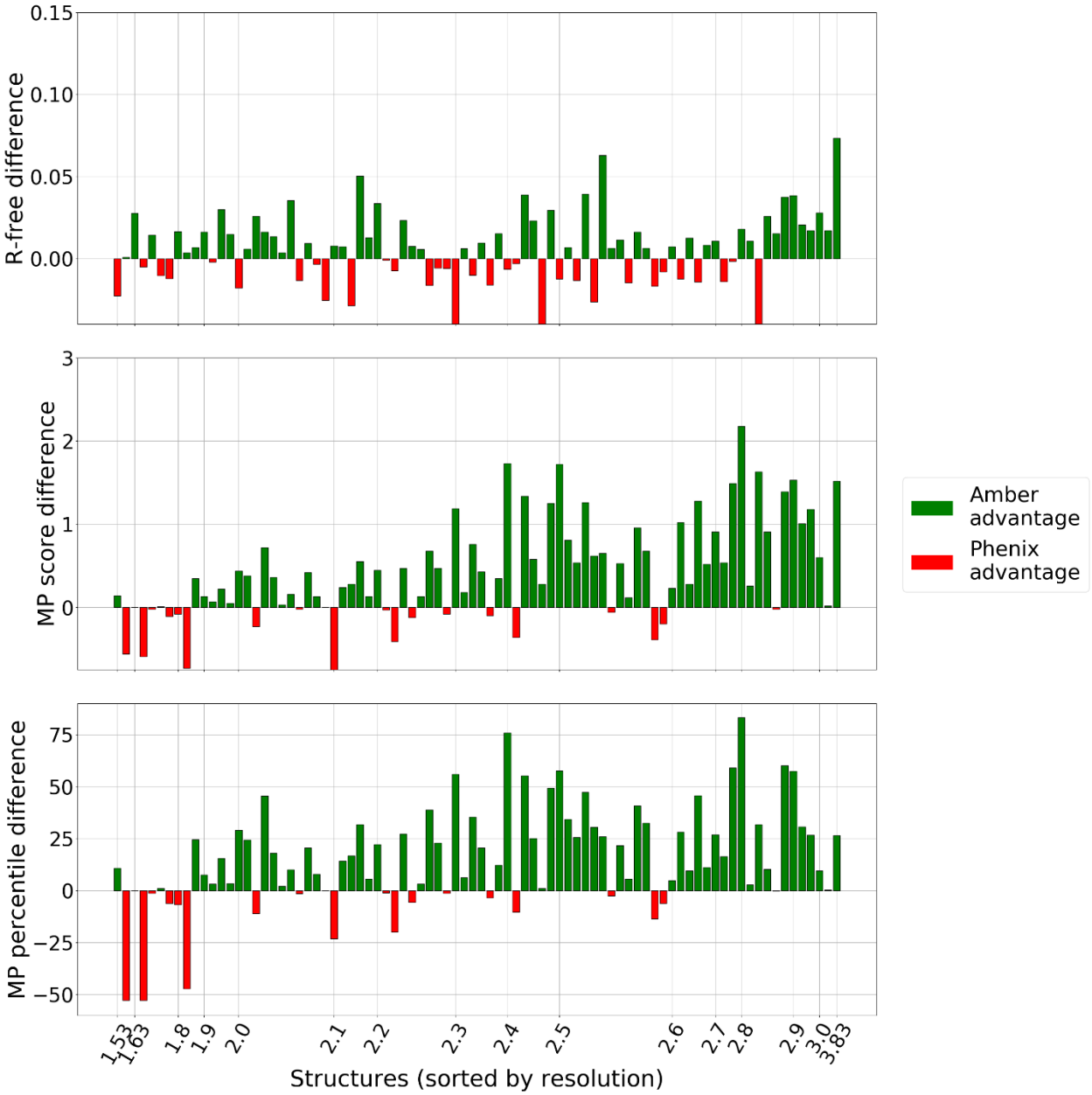

Figure 2.7. The plots show the difference between R_{free} factors, MolProbity scores and MolProbity percentiles of the refined MD3 models. Green bars represent the superiority of our Amber-based setup. Red bars represent the superiority of Phenix-based ASU setups.

Overall, as one can see from the figures and the statistics above, Amber-based refinement leads to consistently better R_{free} factors and MolProbity scores even with the simplest approach versus Phenix-based refinement.

At last, we assessed the performance of refinement protocols against the PDB structures. We compared the re-refined models (the results from the D set of initial models) and the newly refined models (MD1, MD2, MD3 sets) versus the deposited models. Table 2.4 shows this comparison between Amber/Amber, Phenix-based refinement protocols and the Protein Database deposited structures, which had R_{free} factors available (74 out of 84). Clearly, Amber/Amber setup provides the best result in more than 50% of the cases from Deposited, MD1 and MD2 initial models: even starting with a poor model, the refined structure is frequently better than the originally published one.

Table 2.4. Comparison of Amber-based and Phenix-based ASU refinement with the PDB deposited data.

Initial model	Amber produces the best model	Phenix produces the best model	PDB deposition is the best model
Deposited	43	12	19
MD1 (0.75 Å rmsd)	47	4	23
MD2 (0.89 Å rmsd)	39	5	30
MD3 (1.02 Å rmsd)	24	11	39

2.7.3 Comparison across the whole test set: UC case

We have also refined the 84 structures with Phenix using the entire unit cell approach and compared the results to the Amber/Amber results in the fashion introduced in paragraphs 2.7.1 and 2.7.2. Unlike the ASU case, the Amber-based refinement performance is not as prominent but still very competitive.

On the D set of initial structures, our protocol gave better results than Phenix in 47 cases out of 84 with an average improvement of 0.0078 R_{free} units and 14.220% MolProbity score percentiles (Figure 2.8). The comparison on MD1 set is again more striking: 73 out of 84 structures benefit from Amber/Amber refinement rather than *phenix.refine* with improvements to R_{free} and MolProbity score percentiles of 0.0147 and 24.040%, respectively (Figure 2.9).

Afterwards, we decided to run additional tests. We accumulated 5 trials of the best Phenix-based setup and matched them with 5 runs of Amber-based setup for both D and MD1 sets. We also tried to increase the length of refinement in selected cases. As we have mentioned in paragraph 2.4.6 we used Q-scores in the initial tests, so we have also tried to extend the length of Amber-based procedure 10 times with some of the structures from the MD1 set where our Amber/Amber setup was outperformed by the Q-score. Instead of 10 ps intervals during the refinement, we used 100 ps since sometimes it leads to better outcomes (see Table 2.1). Despite these add-ons, the pattern in the comparison remained the same.

The statistics for the MD2 set of initial models is as follows: 67 out of the 84 structures are refined better with Amber/Amber setup with average improvements in R_{free} and MolProbity score percentiles of 0.0165 and 27.865% (Figure 2.10). The Amber-based protocol on the MD3 set outperforms Phenix-based refinement in 61 out of the 84 cases (Figure 2.11). The advantage is 0.0266 in R_{free} and 34.329% in MolProbity percentiles.

Amber/Amber protocol looks more modest in comparison with the Phenix-based UC refinement in terms of the benefits in R_{free} than in the ASU case (paragraph 2.7.2). Despite that, the geometric qualities of the Amber-refined models markedly profit from this approach as measured against *phenix.refine* models in the UC case.

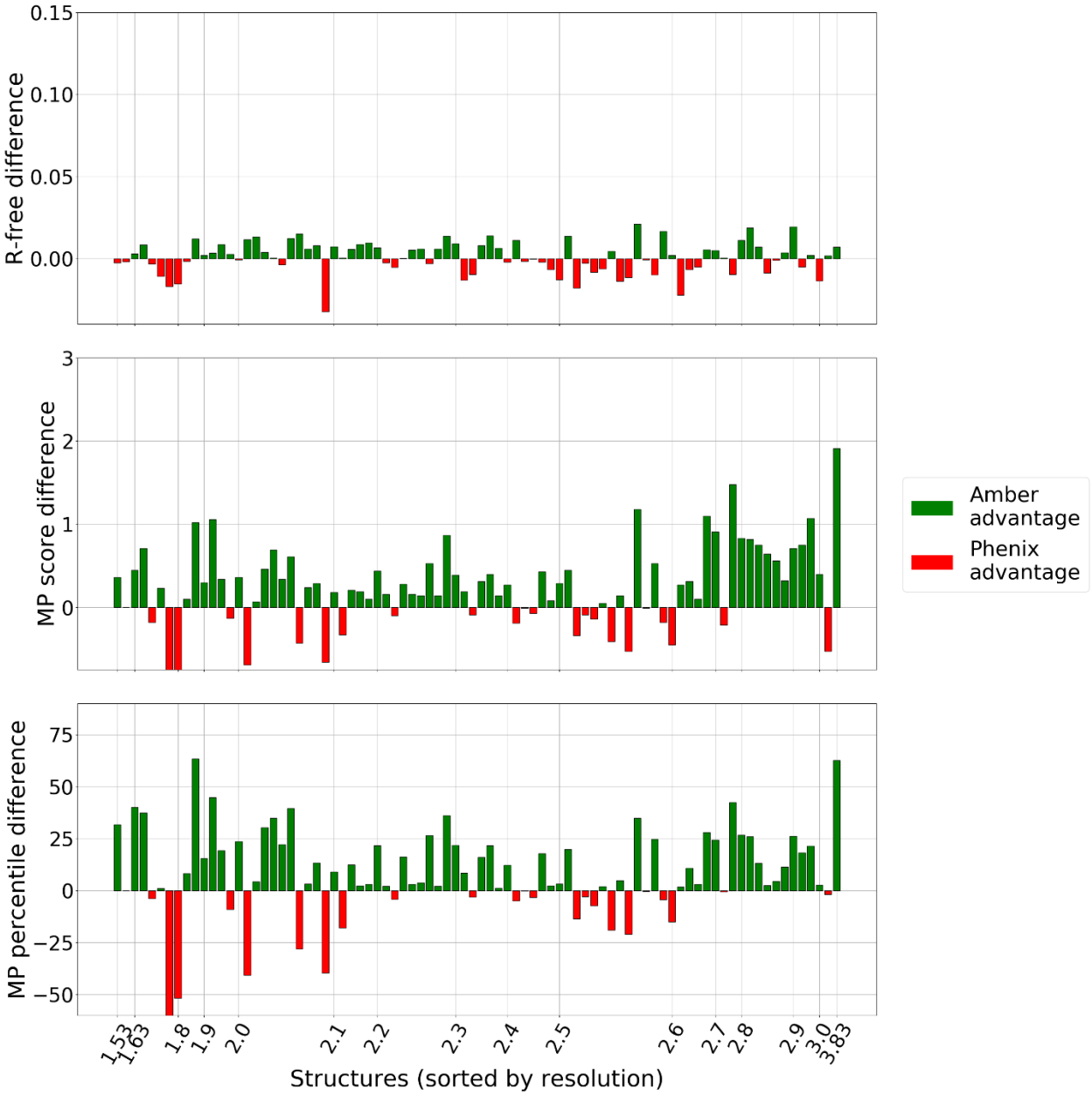

Figure 2.8. The plots show the difference between R_{free} factors, MolProbity scores and MolProbity percentiles of the re-refined deposited models. Green bars represent the superiority of our Amber-based setup. Red bars represent the superiority of Phenix-based UC setups.

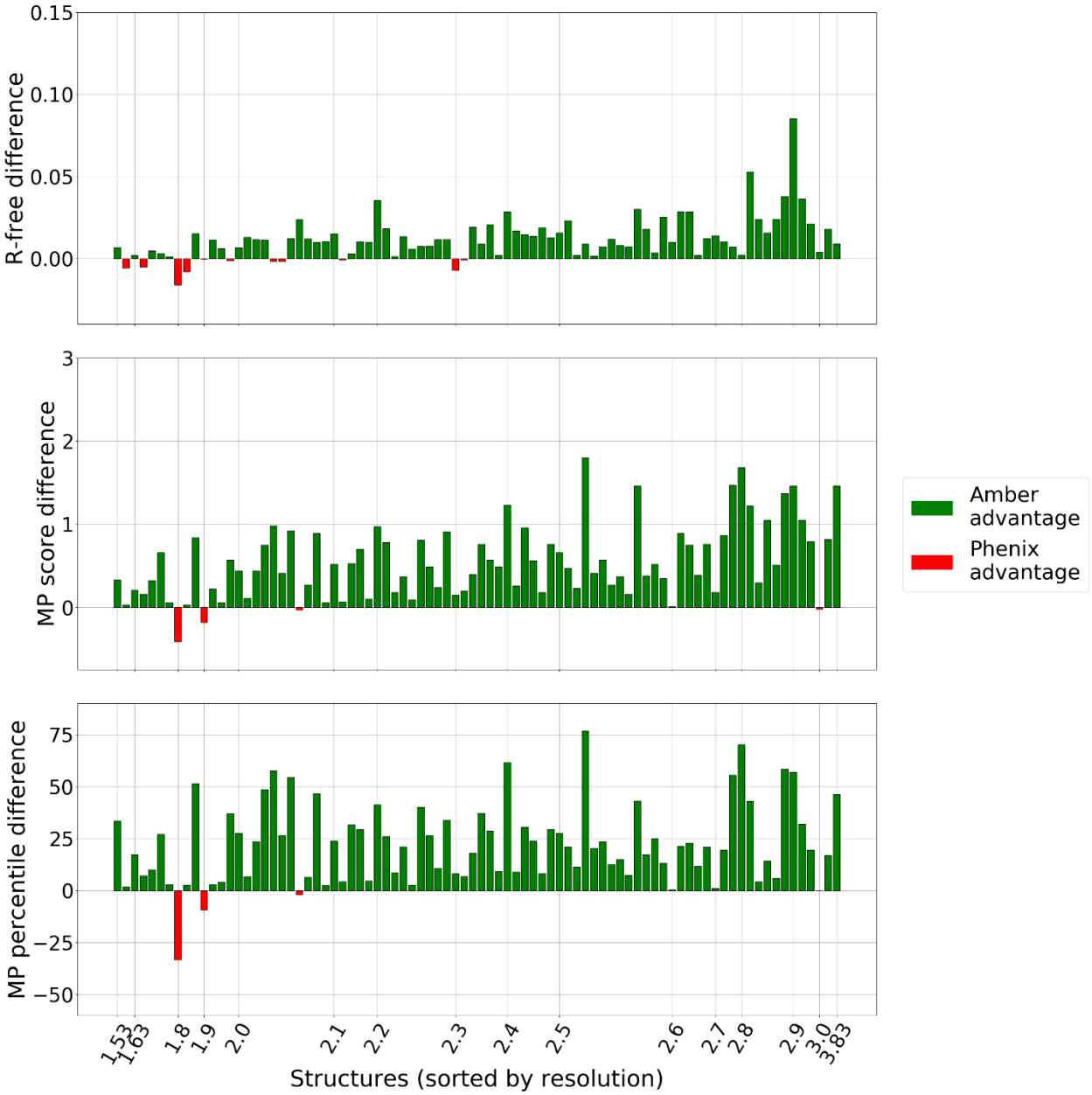

Figure 2.9. The plots show the difference between R_{free} factors, MolProbity scores and MolProbity percentiles of the refined MD1 models. Green bars represent the superiority of our Amber-based setup. Red bars represent the superiority of Phenix-based UC setups.

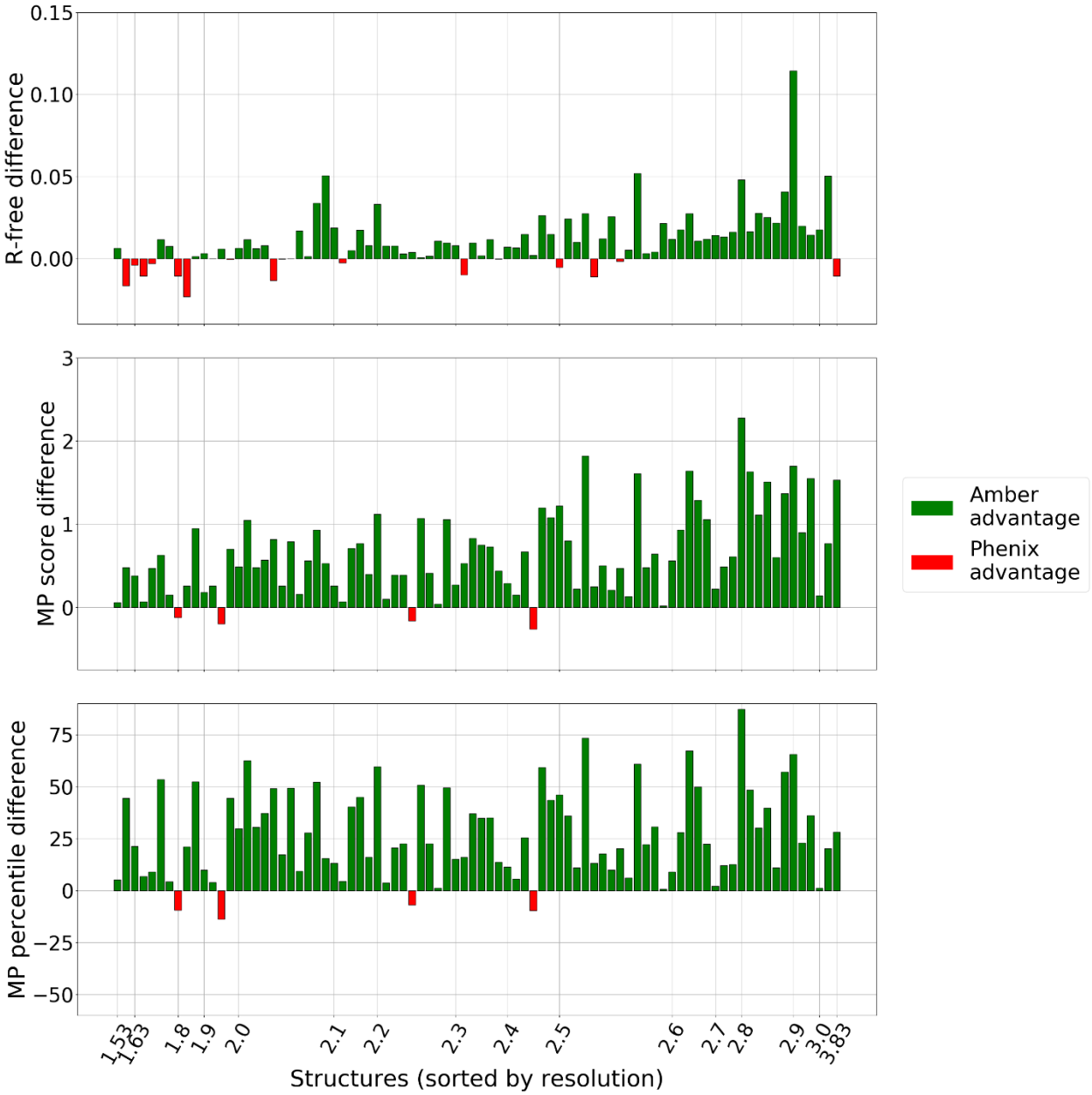

Figure 2.10. The plots show the difference between R_{free} factors, MolProbity scores and MolProbity percentiles of the refined MD2 models. Green bars represent the superiority of our Amber-based setup. Red bars represent the superiority of Phenix-based UC setups.

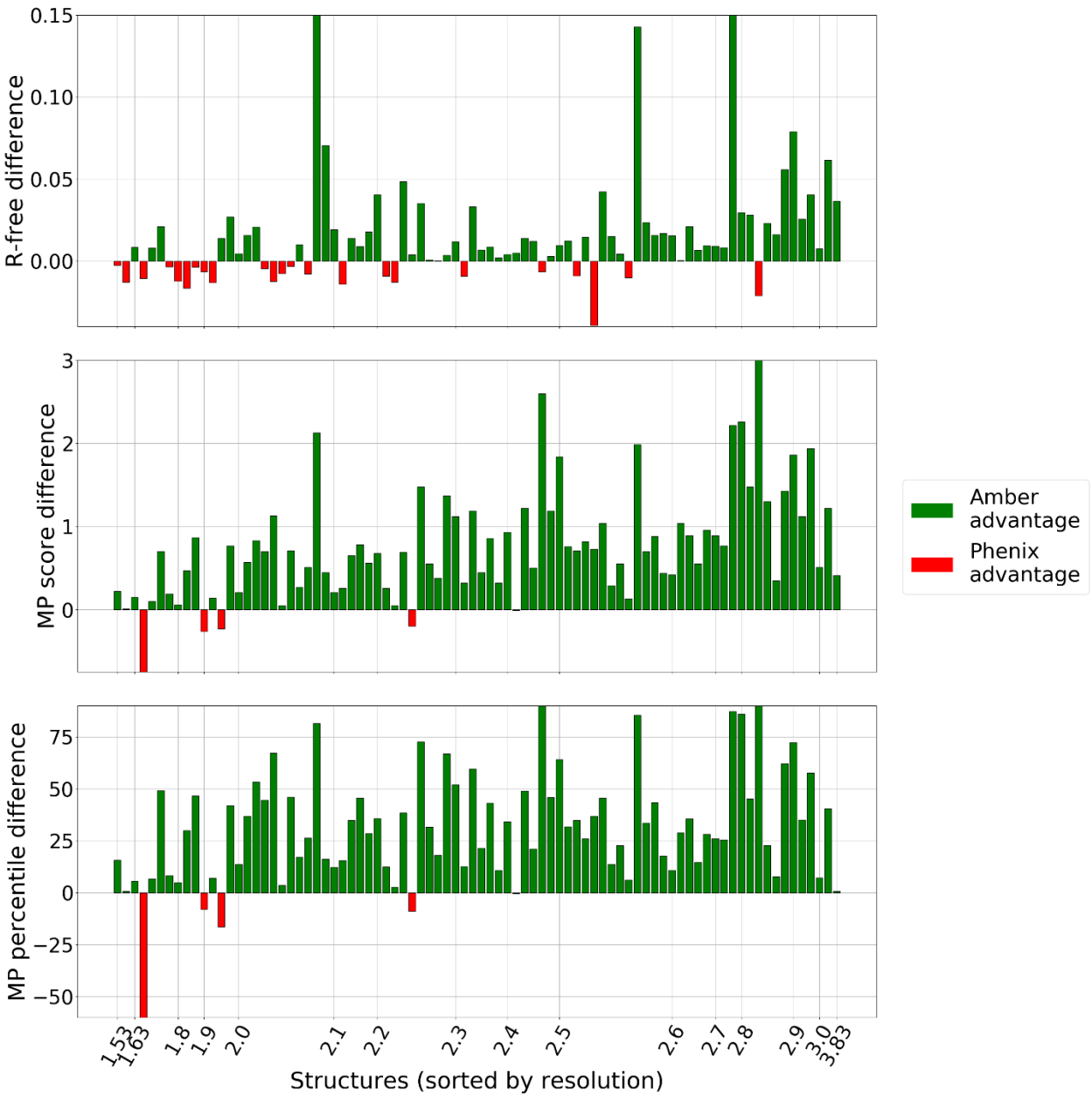

Figure 2.11. The plots show the difference between R_{free} factors, MolProbity scores and MolProbity percentiles of the refined MD3 models. Green bars represent the superiority of our Amber-based setup. Red bars represent the superiority of Phenix-based UC setups.

Finally, we compiled a table similar to Table 2.4 to estimate the re-refinement abilities of the Amber/Amber setup Phenix-based UC protocols. As seen from Table 2.5, our newly developed refinement module for Amber still produces the best result in roughly half of the 74 PDB entries which had R_{free} available.

Table 2.5. Comparison of Amber-based and Phenix-based UC refinement with the PDB deposited data.

Initial model	Amber produces the best model	Phenix produces the best model	PDB deposition is the best model
Deposited	36	22	16
MD1 (0.75 Å rmsd)	43	8	23
MD2 (0.89 Å rmsd)	31	10	33
MD3 (1.02 Å rmsd)	21	13	40

2.7.4 Conformational diversity example: 3ZQ7

To illustrate what the whole unit cell Amber-based refinement can achieve, we selected the 3ZQ7 crystal from P 4_3 2_1 2 space group. This is a randomly selected structure that had a high symmetry space group among the 84 test structures. Each asymmetric unit contains a 102 residue long chain of DNA-binding domain of response regulator from E. coli. The reported structure has 2.52 Å resolution with R_{free} value of 0.283. Below we describe our results of refinement from the MD1 set of initial models.

First, we proceeded with DSSP assignment of the secondary structure [143], [144]. Then, the refined macromolecules were superimposed based on the secondary structure Cα atoms. This way, our approach produced an ensemble of 8 asymmetric units. Next, we calculated the average structure and RMSDs to each of the models to color the cartoon representation of the ensemble on Figure 2.12. From the figure one can see that variations in positions of back bone residues and side chain atoms reach up to 2.5 Å (residues at the bottom of the figure) and 7 Å (residues at the top of the figure), respectively. To conclude, our approach indeed might provide another perspective on conformational diversity aside from fixed alternate conformers as other ensemble models.

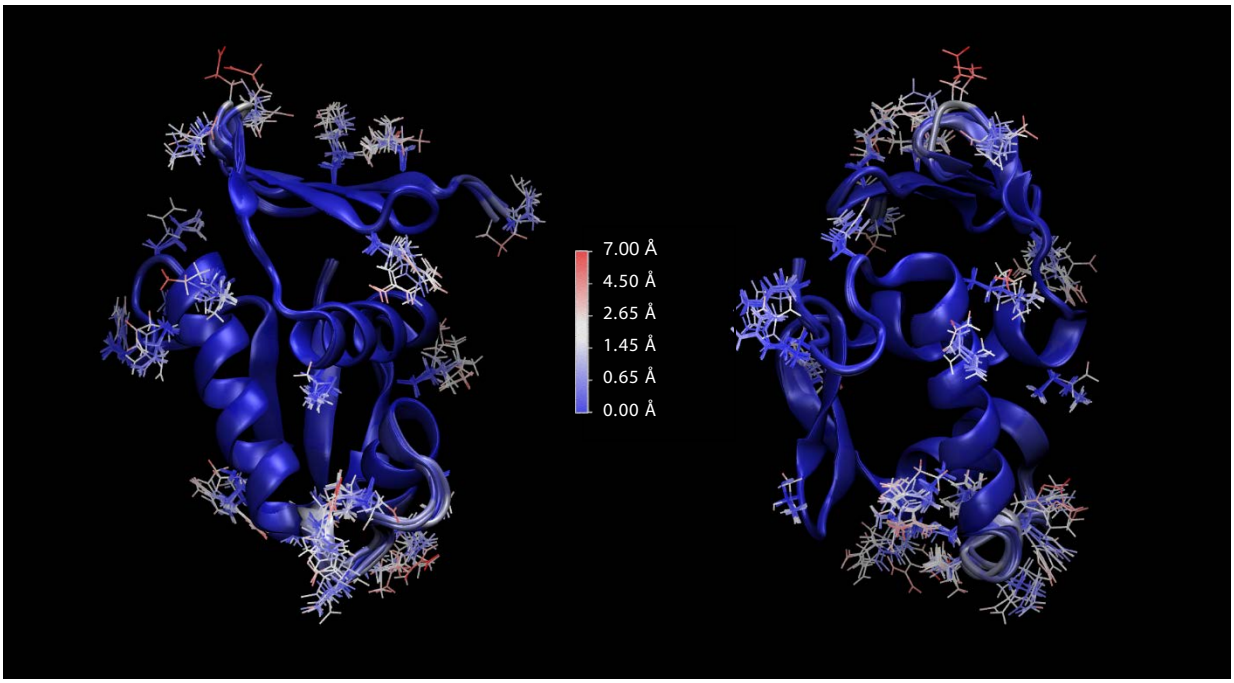

Figure 2.12. 8 superimposed asymmetric units of 3ZQ7. Red color corresponds to higher conformational variability, higher RMSD to the average structure. Blue color corresponds to lower conformational variability, lower RMSD to the average structure.

2.7.5 Natural representation of alternate conformers example: 3C57

The observations in this paragraph were guided by the idea of the explicit conformational diversity application to distinguish states of alternate conformers in macromolecules. Using the same criteria as we employed to choose the test structures in paragraph 2.4.1, except eliminating the restriction (5) on non-unity occupancies, we have selected the 3C57 PDB structure. This was the structure with the smallest size of unit cell to showcase how our approach can benefit structures with alternate conformers. The structure has P 2_1 2_1 2_1 space group providing 4 asymmetric units and has one homodimer per asymmetric unit. The dimer consists of DNA-binding transcriptional activator DevR. The structure has 1.7 Å resolution and reported 0.206 R_{free} value. Each monomer is 95 amino acids long and has five alternate conformers.

In this case we increased the time of refinement from the total of 20 ps to 4 ns: 2 ns constant temperature MD with increasing crystallographic weight and 2 ns of cooling MD with a constant crystallographic weight. This was done to increase the chances of observing the transition between alternate conformers. We selected the first alternate conformer of the two-

component transcriptional regulatory protein as the initial model and refined the ensemble using the extended Amber/Amber setup. The secondary structure remained almost intact except termini after our Amber-based re-refinement: average RMSD against the deposited model is 0.1A over the four asymmetric units.

Figure 2.13 depicts a cartoon representation along with some of sidechains which had alternate conformers by the end of re-refinement. The initial position of the sidechains is represented in blue, the alternate reported conformer is in orange, and the refined models of the ensemble are in green. Thus, for example, one of the 4 copies from the ensembles' M194 residue of the protein's chain B flipped the side chain to the alternate reported conformation (panels A and B) after the refinement. All the representatives of the chain B L160 swapped their conformation into the second reported possible position (panels C and D). Some other residues, which originally had alternate conformers like chain B L165, appeared to be near the initial conformation. This suggests that the MD approach if given enough time might help to determine different occupational states.

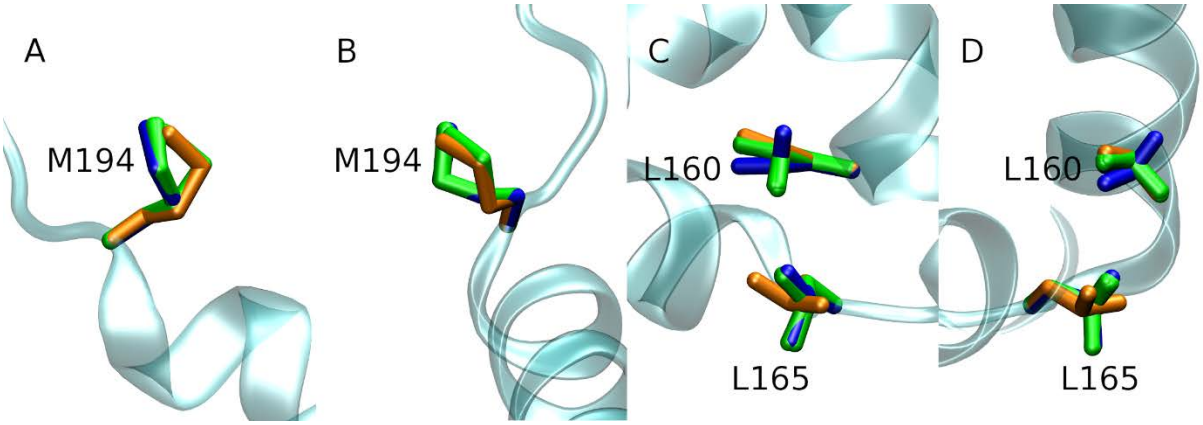

Figure 2.13. Projections of side chains with alternate conformers of 3C57. The first reported alternate conformation is represented in blue. The second reported alternate conformation is represented in orange. The refined multiple conformers are represented in green. Panels (A) and (B) show different projections of chain B M194 residue. Panels (C) and (D) show different projections of chain B L160 and L165 residues.

2.7.6 True real-life example: N-terminal SH3 domain of GRB2 adaptor protein

As a part of a collaboration with the I. Bezprozvanny laboratory, S. Korban kindly provided us with the experimental data and a model of the N-terminal SH3 domain of GRB2 protein in apo form (PDB ID 6SDF). The model was co-crystallized with the MPD, thus, before

proceeding with the Amber/Amber refinement we determined the Amber force field parameters for this agent using Gaussian package [145] and prepared the structure according to our standard procedure (see paragraph 2.4.2). Therefore, along with the protein component of the crystal, we also refined the position of the co-factor.

The space group for this crystal was determined to be R 3. However, since we were able to obtain raw data in P 1 space group at 2.5A resolution, we used them instead for the refinement.

We ran the same 16 Phenix-based protocols and two trials of Amber-based setup and compared the achieved results similar to what we did in the bulk tests (paragraph 2.7.1). Below follows Table 2.6 with this comparison and the results achieved by the Bezprozvanniy lab where they used REFMAC5 and Phenix.

Our independent to the collaborators' attempt to achieve the best possible model resulted in comparable R_{free} factors from Amber-based and Phenix-based ASU protocols: 0.2150 versus 0.2082, respectively. However, the geometric qualities of the Amber-refined model are almost perfect, unlike in the case of Phenix-based ASU setups: 99.253% against 73.126% MolProbity score percentile, respectively. Interestingly, Phenix-based UC refinement results did not follow the trends notes in paragraph 2.6 and produced worse R_{free} factors and better MolProbity score percentile than in ASU case: 0.2211 and 89. 421%, respectively. Also, the best Phenix-based protocols for ASU and UC cases are different.

Another feature in this example is the incorporation of a co-factor into the refinement protocol. In a similar fashion one can derive non-standard protein residue parameters for Amber force field. If one has no access to the Gaussian software, there is a general Amber built-in script that can be used for this purpose: antechamber [146], [147]. This highlights the possibility of the extension for our protocol (paragraph 2.4.3) to more cases, which we restricted while selecting the test structures (paragraph 2.4.1). The MPD molecules fit into the electron density as well in Phenix-based UC refinement: 28 out of 45 molecules have real space correlation of more than 0.8 in our refined model and 27 out of 24 molecules for Phenix. This also supports their inclusion into Amber/Amber refinement.

Table 2.6. Summary of N-terminal SH3 domain of GRB2 protein refinement. The best results in each category are highlighted with boxes. P/A – Phenix with Amber14 force field, P/P – Phenix with Phenix force field, SA – simulated annealing, TAD – torsional angle dynamics, Cartesian – Cartesian dynamics, WO – weight optimization. The best Amber/Amber run is the first trial. The best Phenix-based ASU protocol is the one with CDLv1.2 restraints without any simulated annealing and weight optimization. The best Phenix-based UC protocol is the one with CDLv1.2 restraints without any simulated annealing and with weight optimization.

	R-work	R-free	R-free - R-work	Clash score	Poor rotamers (%)	Ramachandran outliers	Ramachandran favored	Molprobity score	Molprobity percentile
REFMAC5	0.160	0.210	0.050	12.43	1.00	0.00	99.13	1.61	80.78
Amber results									
Run 1	0.186	0.216	0.029	0.61	1.78	0.00	98.36	0.89	98.98
Run 2	0.187	0.215	0.028	0.50	1.67	0.00	98.16	0.84	99.25
Phenix results: ASU case									
P/P, SA, WO	0.190	0.218	0.028	9.94	0.00	0.00	97.39	1.64	79.50
P/P, no SA, WO	0.194	0.211	0.016	4.97	0.00	0.00	99.13	1.26	92.82
P/P, no SA, no WO	0.182	0.212	0.030	10.43	0.00	0.00	99.13	1.54	83.67
P/P, SA, no WO	0.183	0.213	0.030	12.42	0.00	0.00	99.13	1.61	80.78
P/A, SA, WO	0.249	0.271	0.023	7.95	0.00	2.65	93.81	1.85	70.22
P/A, no SA, WO	0.212	0.214	0.002	1.99	0.00	0.00	97.39	1.09	96.88
P/A, no SA, no WO	0.182	0.209	0.027	6.46	0.00	0.00	100.00	1.36	89.79
P/A, SA, no WO	0.216	0.258	0.042	5.46	0.00	5.22	89.57	1.86	69.69
P/P, Cartesian, no WO	0.185	0.216	0.031	14.41	0.00	0.00	98.26	1.67	78.25
P/P, Cartesian, WO	0.203	0.228	0.026	12.42	0.00	0.87	98.26	1.61	80.78
P/P, TAD, no WO	0.180	0.208	0.028	19.87	0.00	0.00	98.26	1.79	73.13
P/P, TAD, WO	0.198	0.216	0.018	9.44	0.00	0.00	98.26	1.50	85.19
P/A, Cartesian, no WO	0.236	0.287	0.051	13.41	0.00	3.48	89.57	2.20	53.52
P/A, Cartesian, WO	0.258	0.285	0.028	4.97	0.00	0.00	93.91	1.67	78.25
P/A, TAD, no WO	0.197	0.234	0.037	7.45	0.00	0.00	96.52	1.64	79.50
P/A, TAD, WO	0.221	0.235	0.014	1.49	0.00	0.00	97.39	1.01	98.04

Table 2.6 continued.

Phenix results: UC case									
P/P, SA, WO	0.180	0.245	0.065	21.83	0.00	2.51	85.41	2.48	38.21
P/P, no SA, WO	0.200	0.221	0.021	6.80	0.00	0.00	98.55	1.37	89.42
P/P, no SA, no WO	0.182	0.224	0.042	13.15	0.00	0.00	96.81	1.82	71.54
P/P, SA, no WO	0.165	0.241	0.076	27.91	0.00	3.86	83.96	2.61	31.53
P/A, SA, WO	0.233	0.290	0.057	20.29	0.11	5.51	85.02	2.46	39.35
P/A, no SA, WO	0.206	0.234	0.027	3.48	0.00	0.00	98.55	1.14	95.79
P/A, no SA, no WO	0.186	0.228	0.042	9.01	0.00	0.00	98.45	1.48	85.90
P/A, SA, no WO	0.209	0.280	0.070	16.75	0.00	6.47	82.80	2.42	41.69
P/P, Cartesian, no WO	0.163	0.240	0.077	30.07	0.11	3.96	83.38	2.65	29.53
P/P, Cartesian, WO	0.183	0.247	0.063	21.73	0.00	2.80	84.93	2.49	37.58
P/P, TAD, no WO	0.165	0.231	0.066	37.20	0.22	1.74	88.99	2.63	30.66
P/P, TAD, WO	0.182	0.228	0.046	25.31	0.00	1.26	91.50	2.40	42.69
P/A, Cartesian, no WO	0.181	0.260	0.079	22.43	0.22	6.31	82.33	2.55	34.54
P/A, Cartesian, WO	0.230	0.272	0.042	12.44	0.00	6.67	83.29	2.30	48.27
P/A, TAD, no WO	0.178	0.232	0.054	11.33	0.00	1.06	93.82	1.98	64.13
P/A, TAD, WO	0.201	0.235	0.033	8.51	0.00	1.16	94.88	1.81	72.14

2.7.7 Performance timing

Another important aspect of structure determination is the time needed to obtain a refined structure. Amber/Amber setup requires a significantly smaller time than it is used to unravel the best protocol of Phenix for a structure. For example, we present the statistics and timing for the discussed 3K9P structure in Table 2.7. Cleary, the benefit of our refinement module in speed is at least 4.7 times.

One should also mention that the best Phenix-based protocol varies from structure to structure. Thereby, the particular advantage of our method is the absence of necessity to fine-tune variable parameters and, consequently, the less amount of time needed to find the best model without manual intervention.

Table 2.7. 3K9P R-factors, MolProbity, and timing statistics.

3K9P	R_{free}	RMSD $C\alpha$
MD1	0.42	0.73 Å
	Best R_{free}	MolProbity percentile
Amber-based refinement (2 runs, 2.4 hrs)	0.27	97 %
Phenix-based ASU refinement (16 runs, 11.4 hrs)	0.31	81 %
Phenix-based UC refinement (16 runs, 16.7 hrs)	0.32	53 %

As seen from the formula for structure factors of the whole model (2.1), one needs to consider bulk solvent contribution and further scale the data. At the moment, we use the implementation of such procedure from the *cctbx* library [136]. *cctbx* interface and Amber code are based on different programming languages and require frequent data passage back and forth. This passage is coupled to data structures reorganization. Therefore, such a bridge comes with significant computational expenses. Nevertheless, our GPU code is ~10x faster than the CPU code as calculated on the test cases. Even though it is significantly slowed down by that piece of CPU calculations. We expect that the translation of the scaling procedures on GPU and more optimization of the currently existing X-ray related GPU code will drastically improve the current performance. The work in this direction is currently going with Amber developers.

2.7.8 Web server

The final product of the project is our refinement server, which produces consistently better models in comparison with Phenix-based approach for further investigations of proteins via completely automated pipeline without the need to fine-tune various refinement schedules.

On top of our refinement module we have built a web server currently located at http://purcell.chem.purdue.edu:8000/refinements using Django framework with Celery task scheduler and custom in-house Python scripts to deliver the service. The server provides an opportunity to refine macromolecular structures in the PDB format against structure factors anonymously. Alternatively, one can register and keep track of the refinement jobs as well.

A huge benefit of this server is that the user does not need to worry about the execution of the process and receives a notification through an e-mail once the structure of interest is refined. The availability of GPU-accelerated computational power on users' side is also unnecessary, and one can track the progress from a handheld device such as smartphone or tablet. A sample job upload page of a registered user is presented in Figure 2.14.

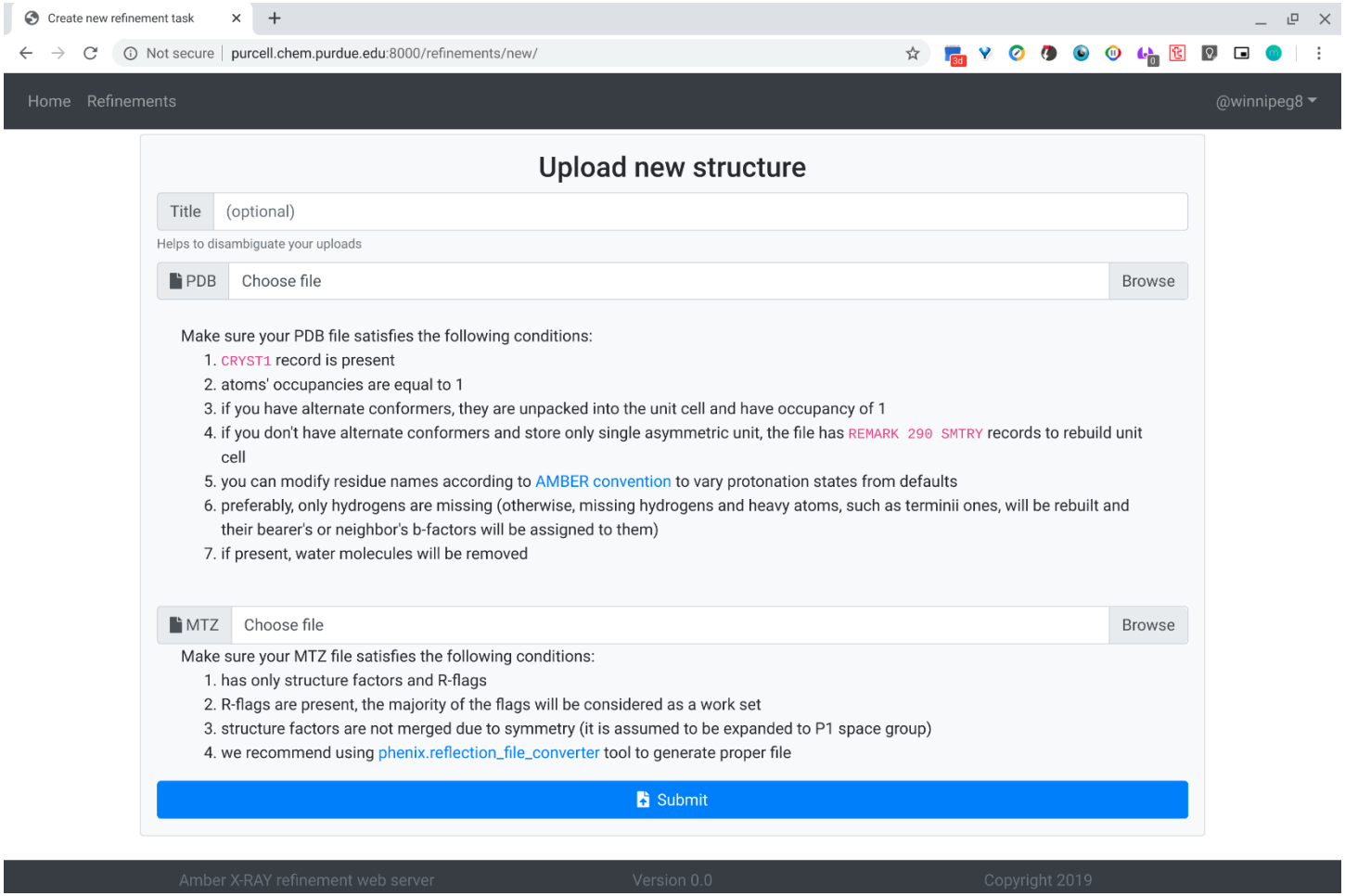

Figure 2.14. Job upload page of a registered user on the Amber-assisted X-ray refinement web-server.

CHAPTER 3. DIFFUSE SCATTERING

3.1 Diffuse scattering profiling

To compare molecular dynamics of 3ONS, 3N30 and 3EHV crystal structures and, particularly, to investigate rigid-body motions in these trajectories, we simulated radially averaged diffuse scattering profiles. The diffuse intensities were calculated according to the Guinier equation, which was introduced in CHAPTER 1:

$$I_{diff}(hkl) = < |\boldsymbol{F}_n(hkl)|^2 >_n - |< \boldsymbol{F}_n(hkl) >_n|^2,$$

where hkl are Miller indices, $I_{diff}(hkl)$ is the corresponding diffuse intensity, $\boldsymbol{F}_n(hkl)$ is the structure factor of the whole frame for the corresponding Miller indices. To calculate structure factors, we use the direct summation formula (1.4). Since we know the precise location of each atom at each given moment in our simulations, we do not need the corrections covered in paragraph 1.1.4. Hence, we use unity occupancies and zero B-factors for all atoms, including solvent molecules where specified and the direct summation formula simplifies to the following:

$$\boldsymbol{F}(\boldsymbol{s}) = \sum_{j=1}^{N_{atoms}} f_j(\boldsymbol{s}) \exp\left[i2\pi \boldsymbol{s} \cdot \boldsymbol{r}_j\right]. \quad (3.1)$$

We have three different crystal simulations with different unit cell dimensions. Therefore, the reciprocal space coordinates, e.g. Miller indices, do not have a direct relationship between them. However, we need to use some invariant to compare the intensities. Thus, instead of the intensities versus Miller indices dependence, we chose to compare the intensities versus the resolutions corresponding to respective Miller indices. These corresponding scattering resolutions are calculated from the Miller indices based on unit cell dimensions, providing a unit cell independent measure.

The deposited observed data corresponding to our simulated crystals had from 96.5% to 99% of all possible reflections. However, since our goal is to simulate the hypothetical intensities for new experiments, we generated 100% complete sets of Miller indices to be used in our modelling. The minimum cutoff value was set to 1.8 Å as the best resolution of the three crystals. The maximum cutoff value for the resolutions was set to 30 Å since the reflections become extremely rare above this resolution.

Clearly, the diffuse scattering intensities produce overwhelmingly crowded plots. Hence, the intensities were radially averaged. Following the same argument as for the lower resolution cutoff, it is more rational to consider inverse resolution scale to increase intensity values distribution at lower resolution (30 Å) and reduce this density at upper resolution (1.8 Å). The corresponding interval of direct resolutions translates to (0.033 $Å^{-1}$, 0.555 $Å^{-1}$) on inverse resolution scale. The inverse resolution dimension was dissected into 50 bins, and the average I_{diff} value was computed for each of these bins. Similar techniques are employed in a number of studies mentioned in the introduction [82], [89], [99], [110], [112], [123]. We call such averaged curves as diffuse scattering profiles further in the text.

Finally, the averaging in the Guinier formula was performed over 2000 frames (uniform sampling of 2 μs long trajectories). In all the results, except where we predict the experimental profile, we omitted: 1) hydrogen atoms to accelerate the calculations unlike CHAPTER 2, and 2) ions and solvent. The solvent effects are discussed in the following paragraph and paragraph 3.3.

3.2 Methods

3.2.1 Trajectories preparation

Table 3.1. Summary of crystal simulations setups.

	3EHV	3ONS	3N30
Unit cell dimensions (Å)	45.823, 52.630, 96.402	49.204, 49.204, 62.986	106.61, 106.61, 106.61
Unit cell angles	90, 90, 90	90, 90, 120	90, 90, 90
Water residues	6198	8772	23419
Chlorine atoms	48	192	0
Protein heavy atoms	14448	14448	28896
Total heavy atoms	20694	23412	52315

Over the course of the crystal simulations, the molecules undergo a slow drift across periodic boundaries. This is a harmless effect for the refinement in general, but it requires a correction for the purpose of diffuse scattering simulations. Clearly, such overall drift in coordinates introduces the same phase shift in all structure factors (see formula (3.1)). Therefore,

the first component of Guinier's formula, $< |\boldsymbol{F}_n(hkl)|^2 >_n$, remains unaffected, but the second component, $|< \boldsymbol{F}_n(hkl) >_n|^2$, becomes severely distorted.

Thus, as preliminary actions, 1) using GROMACS utility *trjconv* we have eliminated jumps of protein chains occurred due to the periodic boundary conditions with *–pbc nojump* subcommand, 2) performing subcommand *–fit translation* on "protein-H" group (i.e. protein component) we eliminated the drift. A similar approach is suggested by Wall [114].

To check that this correction procedure also correctly addresses the drift of water (not only protein component) we have performed several tests. First, we traced the centers of mass translations of solvent and protein lattice separately and ensured that they follow similar paths for each of our trajectories: 3ONS, 3N30 and 3EHV crystal MD simulations. Figure 3.1 depicts such paths for the 3N30 crystal during the first 100 ns of the simulation. We estimated the speed of the changes in the difference between the coordinates of the corresponding centers of mass. This value does not exceed 1 Å over the sampled 2000 frames for all three trajectories: 0.90 Å, 0.54 Å, 0.99 Å for 3N30, 3EHV, 3ONS, respectively. Taking the large dimensions of unit cells into account, we conclude that water drifts together with the protein as one may expect.

Second, we compared the diffusion and the overall drift of solvent molecules. Based on the previous test, the water drift and the crystal lattice drift are tightly coupled. Hence, we estimated the diffusion coefficients of the lattice center of mass as a measure of water drift using CPPTRAJ. Next, we measured the true diffusion coefficient of the solvent by similar means. As one can see from Table 3.2, it turns out that the water diffusion is much faster than its drift.

Finally, we recorded a control simulation consisting entirely of water and verified that its diffuse scattering response is unaffected by the drift correction treatment. Obviously, the diffuse scattering intensities are slightly different for the original and the corrected trajectories (top panel of Figure 3.2). However, the radially averaged profiles are practically identical (bottom panel of Figure 3.2), giving the maximum difference of 0.0018% along the range of intensities between them. These three points validate our drift correction strategy. Therefore, the effect of the water drift on diffuse scattering can be safely neglected.

Table 3.2. Diffusion coefficients as estimated by CPPTRAJ analysis of the three original trajectories.

	3ONS	3N30	3EHV
Protein lattice	5.09	0.28	1.94
Solvent diffusion	300.89	582.43	168.77

3.2.2 Separation of motions

Along with the original trajectories we have generated several pseudo-trajectories for each of the crystals corresponding to the three kinds of motions:

- internal motions - molecules from MD frames are superimposed onto molecules in the crystallographic structure,
- rotational motions - 1UBQ molecules are superimposed onto molecules from MD frames and then translated to their positions in the crystal lattice according to the crystallographic structure (using center-of-mass coordinates),
- translational motions - 1UBQ molecules are superimposed onto molecules in the crystallographic structure and then translated to their positions in the crystal lattice according to the MD data (using center-of-mass coordinates).

All superpositions and translations above are based on Cα atoms within the secondary structure of ubiquitin. We have used a 1UBQ crystallographic structure to isolate rotational and translational motions for the unbiased comparison between different structures. In this way, we have estimated the impact of different kinds of motions on the diffuse scattering profiles based on the above pseudo-trajectories.

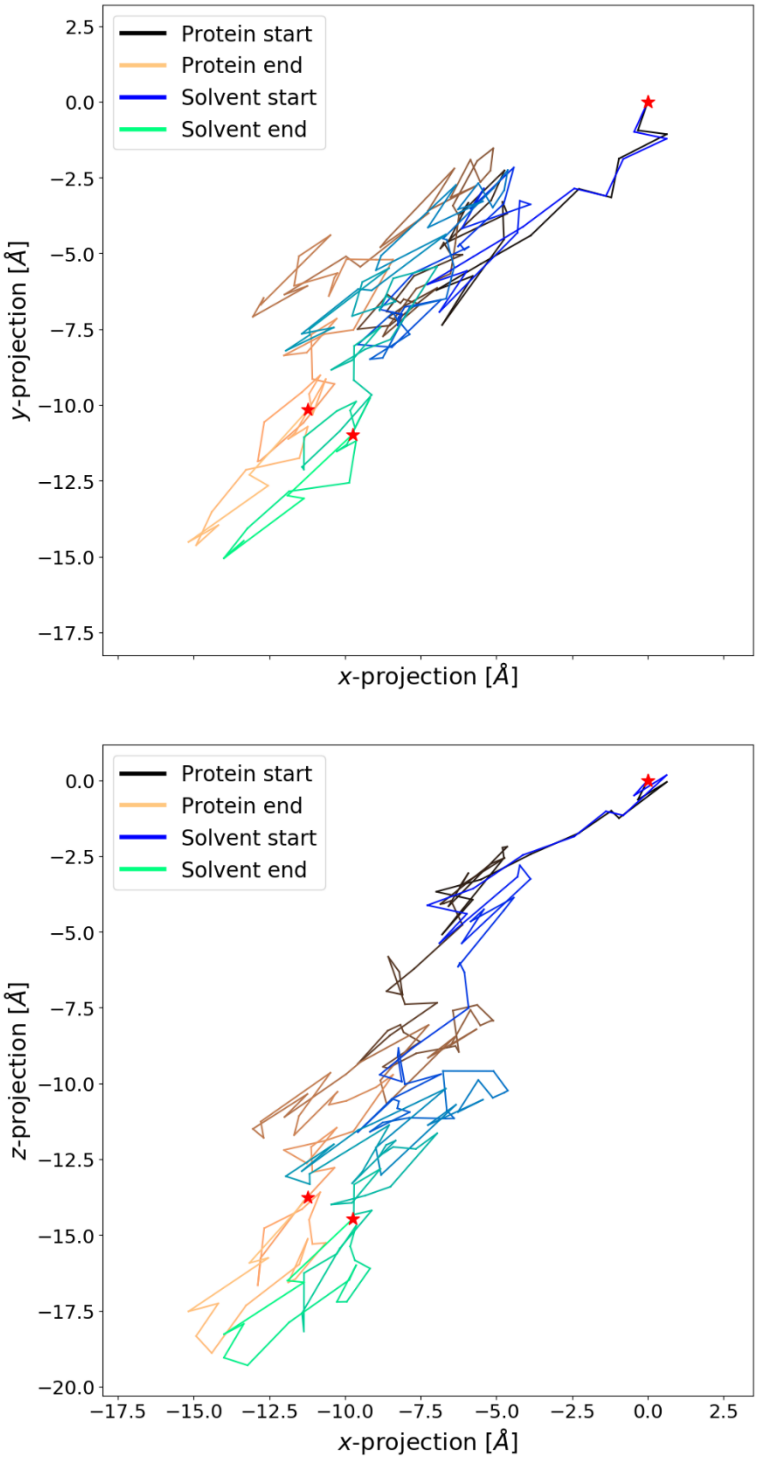

Figure 3.1. Drift correction strategy validation. Projection of the protein's and solvent's centers of mass paths from the 3N30 2 μs trajectory. The first 100 ns are shown.

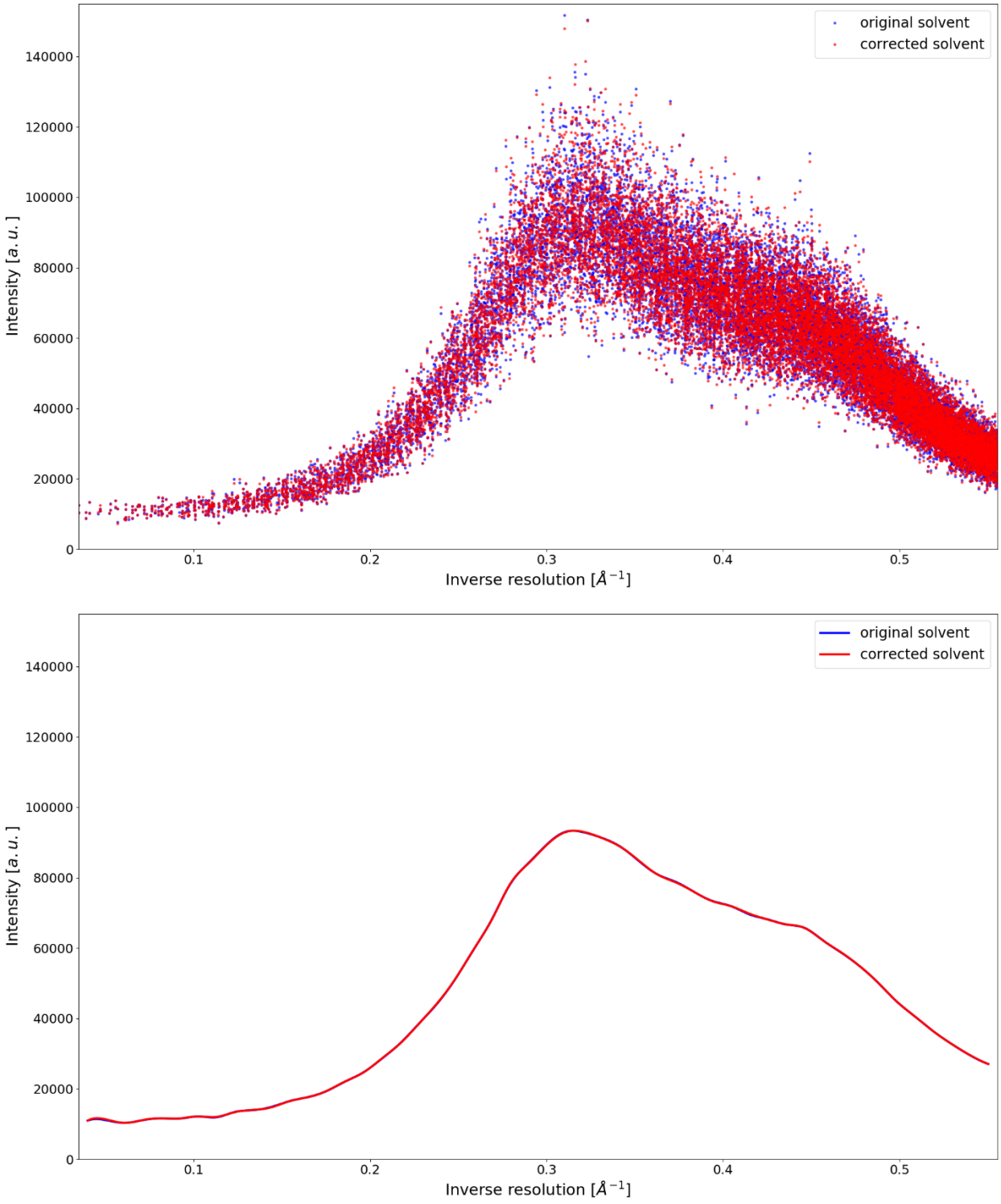

Figure 3.2. Diffuse scattering intensities of the control pure water simulations. Top panel shows all intensities sorted by inverse resolution. Bottom panel shows the radially averaged intensities.

3.2.3 Profile's independence on the unit cell dimension and scaling

Even though we have established the strategy for the invariant comparison of intensities across different crystals (see paragraph 3.1), one can also note that unit cell dimensions do not have a significant effect on the simulated diffuse scattering profiles. Here, we mean that we can change the unit cell dimensions for the calculations in the Guinier formula, i.e. for the calculations of structure factors. Yet the crystal MD simulations are still performed using the original values. Figure 3.3 depicts the profiles calculated from 3ONS crystal using different cell dimensions in the Guinier formula. It is clear that the profile based on the smaller 3ONS unit cell would exhibit larger fluctuations over the inverse resolutions range, which is attributed to a less dense distribution of reflections in the bins. Thus, the maximum difference between the averaged intensities of the profiles calculated with the two sets of dimensions comprised 2.89%.

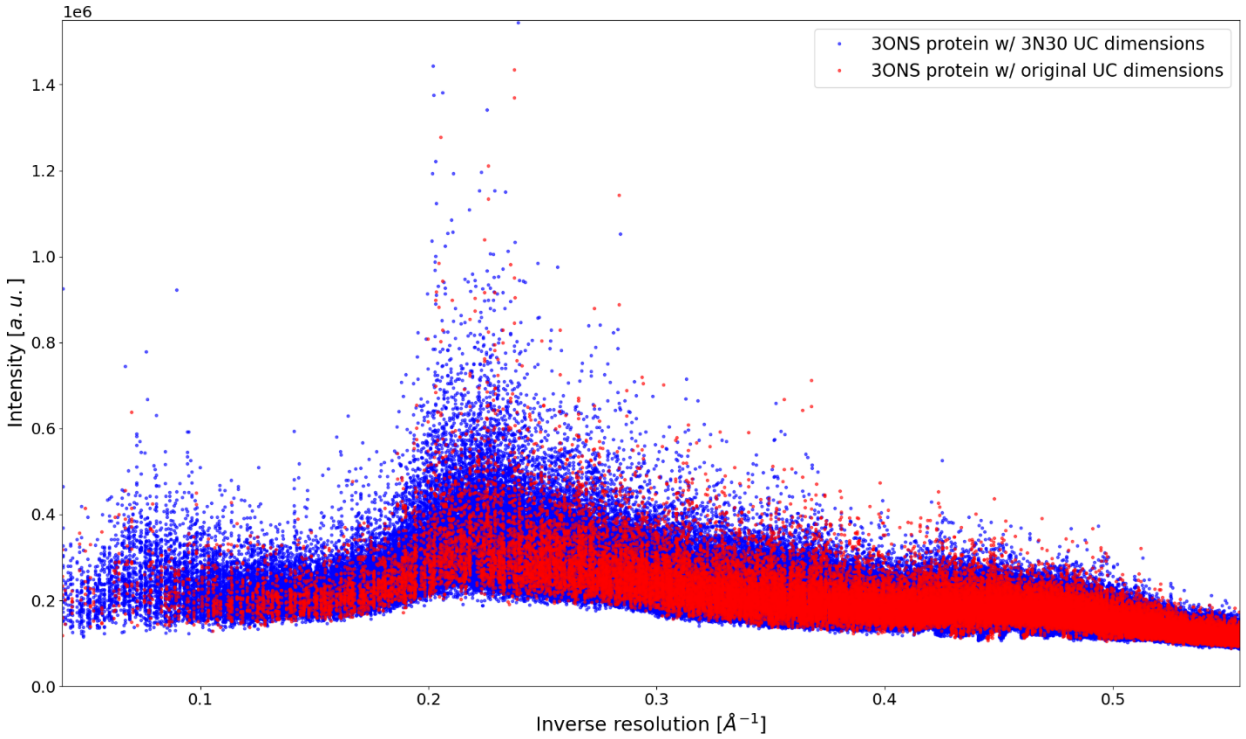

Figure 3.3. Diffuse scattering profiles of 3ONS crystal trajectory using the actual crystal unit cell parameters and the unit cell parameters of 3N30, alternatively.

Therefore, we used the 3N30 unit cell dimensions for the purposes of comparison between structures, unless otherwise specified. These dimensions are the largest among the crystals, hence, they provide the smoothest curves given the same resolution range.

To eliminate the differences in the number of scatterers, e.g. crystal size, the structure factors and intensities need to be normalized. Let us first estimate a structure factor for a frame of a trajectory. If all n atoms in the frame are of the same type with approximately equal scattering factors, our problem is analogous to that of the displacement of a particle due to Brownian motion in two dimensions, where n is the number of equal steps.

In other words, by looking at the direct summation formula one can see that $\boldsymbol{F}(hkl)$ is a sum of n exponents with quasi-random phases, and the length of the resulting vector is proportional to $\sqrt{n}$ (see Figure 3.4). Hence, the following relationship holds for any structure factor: $\boldsymbol{F}(hkl) \sim \sqrt{n}$, and consequently, $I_{diff}(hkl) \sim n$. Therefore, the intensities calculated from 3EHV and 3ONS pseudo-trajectories from paragraph 3.3 are scaled by a factor of 2 since they have two times less atoms than 3N30 (see Table 3.1). It is worth noting that only heavy atoms of proteins are considered in this estimation as well as in structure factors calculations (see paragraph 3.1).

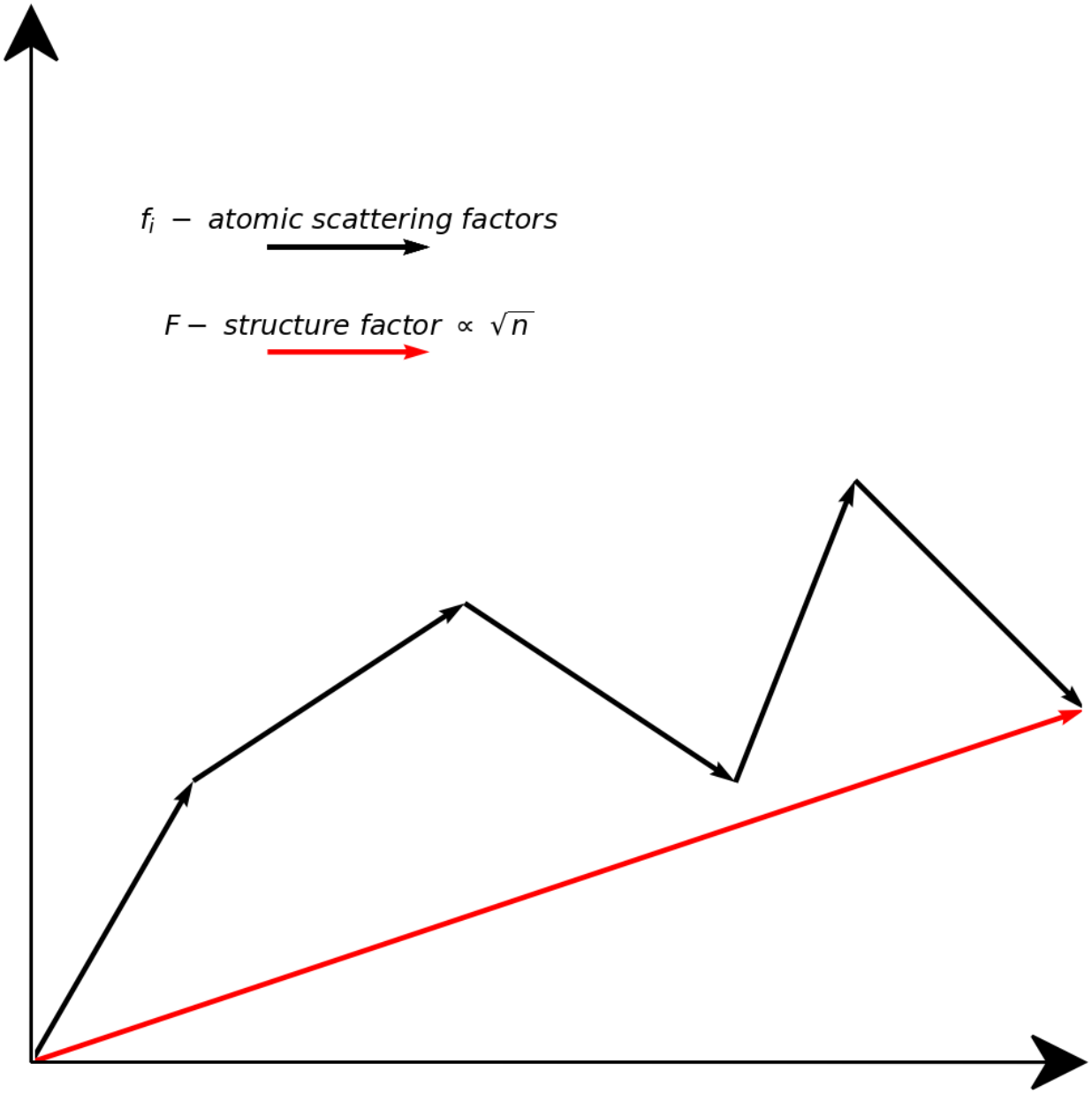

Figure 3.4. Schematic representation of the total structure factor for a given reflection as a sum of the contributing atomic scattering factors.

3.3 Results

3.3.1 Pseudo-trajectories profiles comparison

First, we analyzed the influence of each type of motion on the diffuse scattering profile for the three crystals (see Figure 3.5). Clearly, all three crystals exhibit a similar pattern in the profiles when the curve corresponding to internal motions is disregarded. The dominant input into the overall profile is the intensities generated by the rotational (i.e. rocking) motions. This input is followed by the intensities generated by translational motions.

The internal motions influence the diffuse scattering of 3EHV and 3ONS in a similar fashion at lower resolutions (>10 Å, or <0.2 $Å^{-1}$). However, the higher the resolution, the less pronounced is their relative effect in 3ONS compared to 3EHV. Unlike the first two crystals, 3N30 diffuse scattering intensities are affected by the internal protein motions the least.

Next, we compared the profiles for each type of motions between the crystals. As expected, the internal motions across the MD simulations produced very similar diffuse intensities response (see Figure 3.6, top left panel). The profiles generated using the pseudo-trajectories that represent rotational and translational motions appeared to be similar. The 3EHV diffuse scattering intensities associated with these motions are the smallest. The 3ONS crystal intensities are affected slightly more than in the case of 3EHV, while the 3N30 intensities are the biggest. These observations are logically concluded in the same pattern of the overall intensities (Figure 3.6, bottom right panel).

The magnitudes of the intensities of rotational motions pseudo-trajectories support the results of [84] where we estimated the amplitude of rocking motions in 3N30 to be larger than it is in 3ONS and 3EHV.

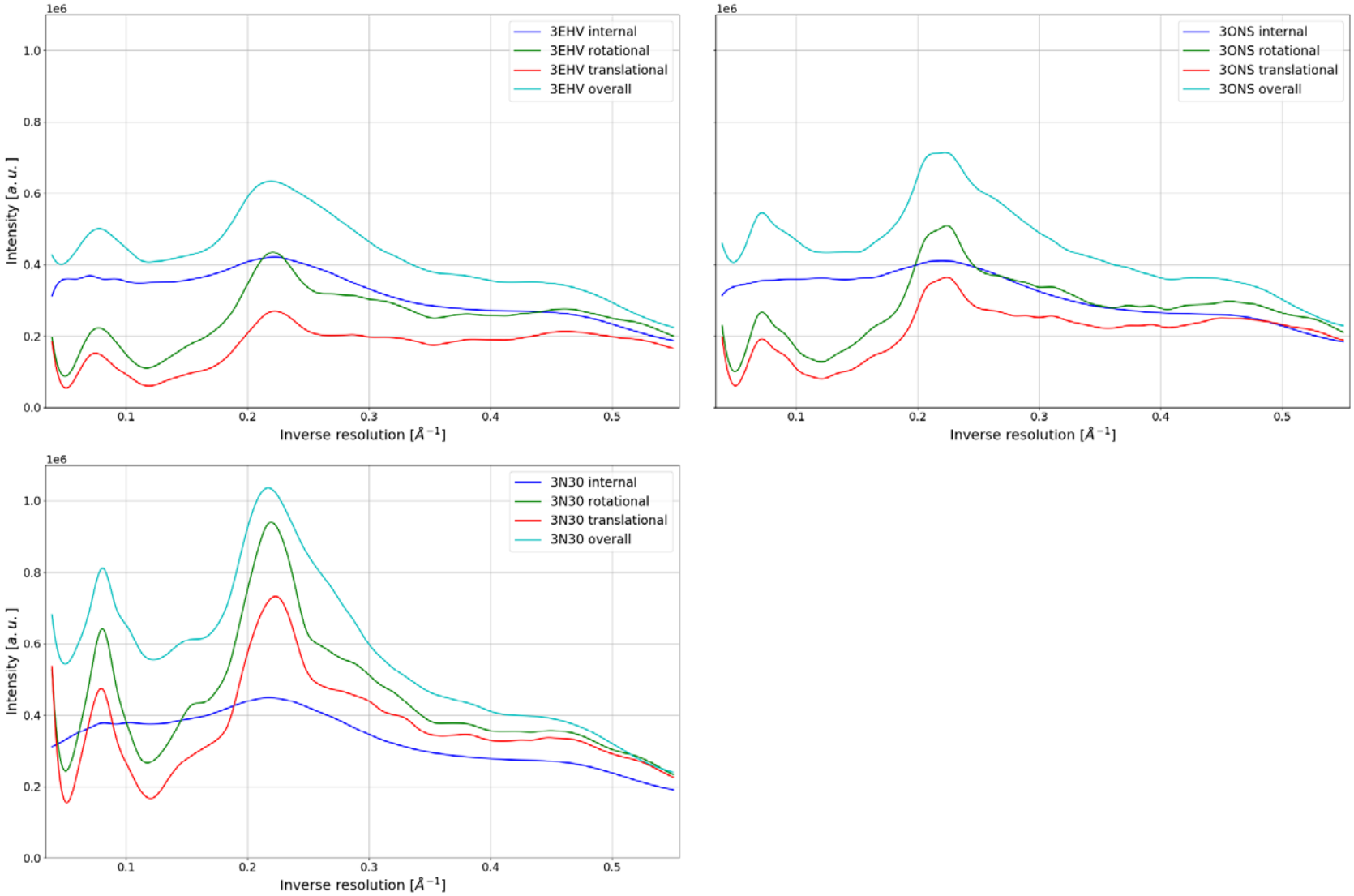

Figure 3.5. Comparison of diffuse scattering profiles for each pseudo-trajectory by ubiquitin crystals. The solvent component for the overall profile is disregarded.

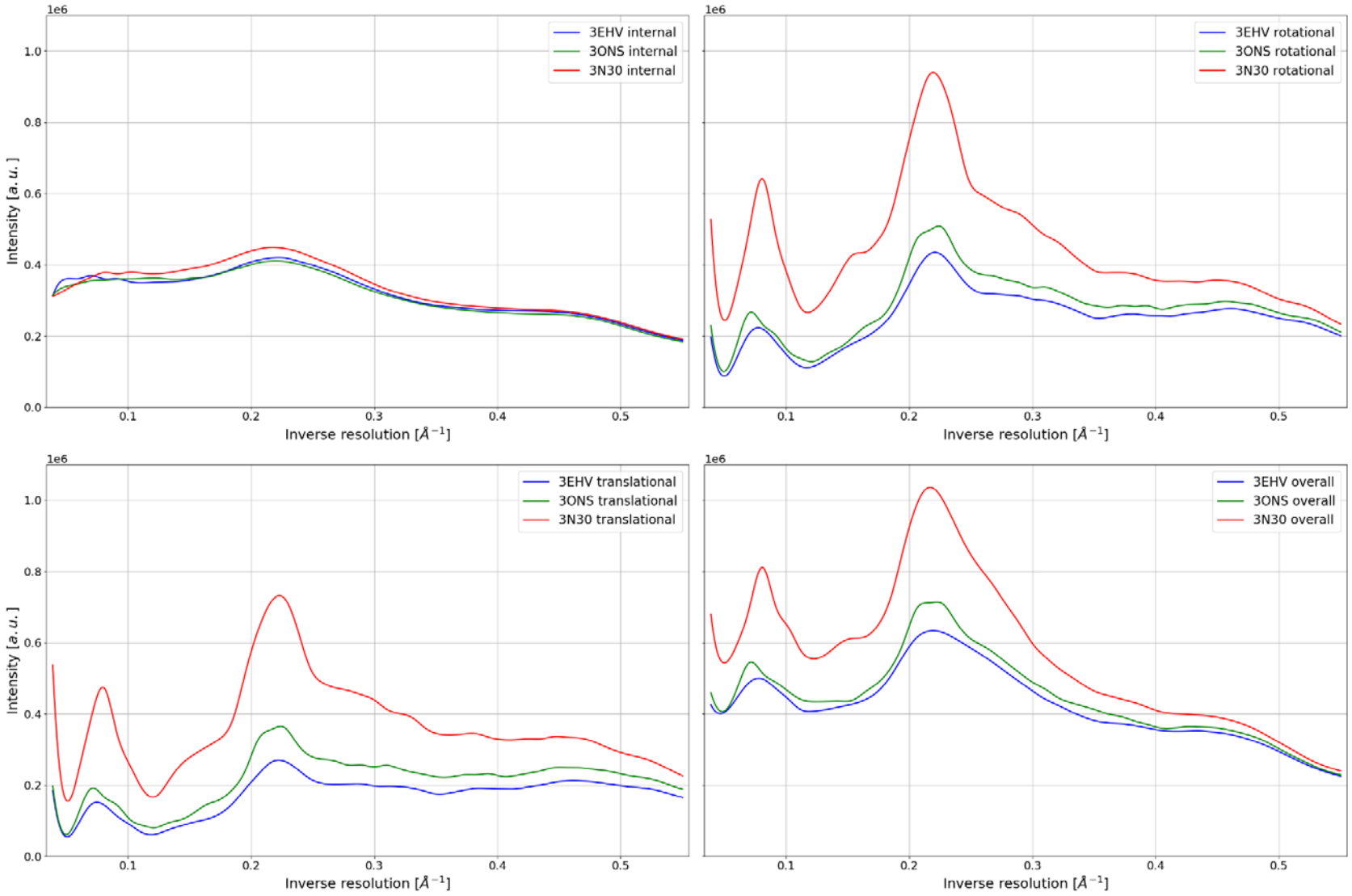

Figure 3.6. Comparison of diffuse scattering profiles for each pseudo-trajectory by the type of motion. The solvent component for the overall profile is disregarded.

3.3.2 Experimental data simulation

To predict the experimental diffuse scattering profiles, solvent and ions have been taken into account as well as the accurate MD unit cell dimensions for each individual crystal. Solvent molecules of trajectories after elimination of crystal drifting were put back into the original simulation box using GROMACS *trjtool -pbc atom* command for 3EHV and 3N30 trajectories. In the case of 3ONS simulations with non-orthorhombic unit crystal cell, the GROMACS output had corrupted unit cell dimensions and, therefore, we had to use the VMD *pbc wrap* command after copying the correct dimensions from the original trajectory.

Analogously to the previous paragraph, we first analyzed the results for each of crystals. In Figure 3.7 below, we do not apply any normalization and again omit hydrogen atoms. The solvent contribution to the diffuse scattering profile is specifically interesting for the resolutions of less than 3.7 Å (or more than 0.37 $Å^{-1}$). It is the least prominent relative to the protein part for the 3EHV structure. While the most significant influence of the solvent part is present for the 3ONS crystal.

Next, we compare the predicted diffuse scattering profiles, where the normalization is done according to the number of heavy atoms in each simulation (see Figure 3.8). The profiles' curves appear to be very similar for all three crystals. 3N30 crystal diffuse scattering profile shows the largest magnitudes. However, it is not clear whether this is simply the effect of scaling or motions. Moreover, since the experimental data are in arbitrary units, it can be scaled to any desired magnitude. Therefore, we conclude that we were unable to identify any profile's feature which would point to the motions that gave rise to it. Hence, it is impossible to compare the amount of rocking motions from the diffuse scattering profiles not quantitatively nor qualitatively.

There are no distinctive features of the diffuse scattering profile indicative of rocking. If this is so, it may not be worthwhile to pursue the experimental study of diffuse scattering.

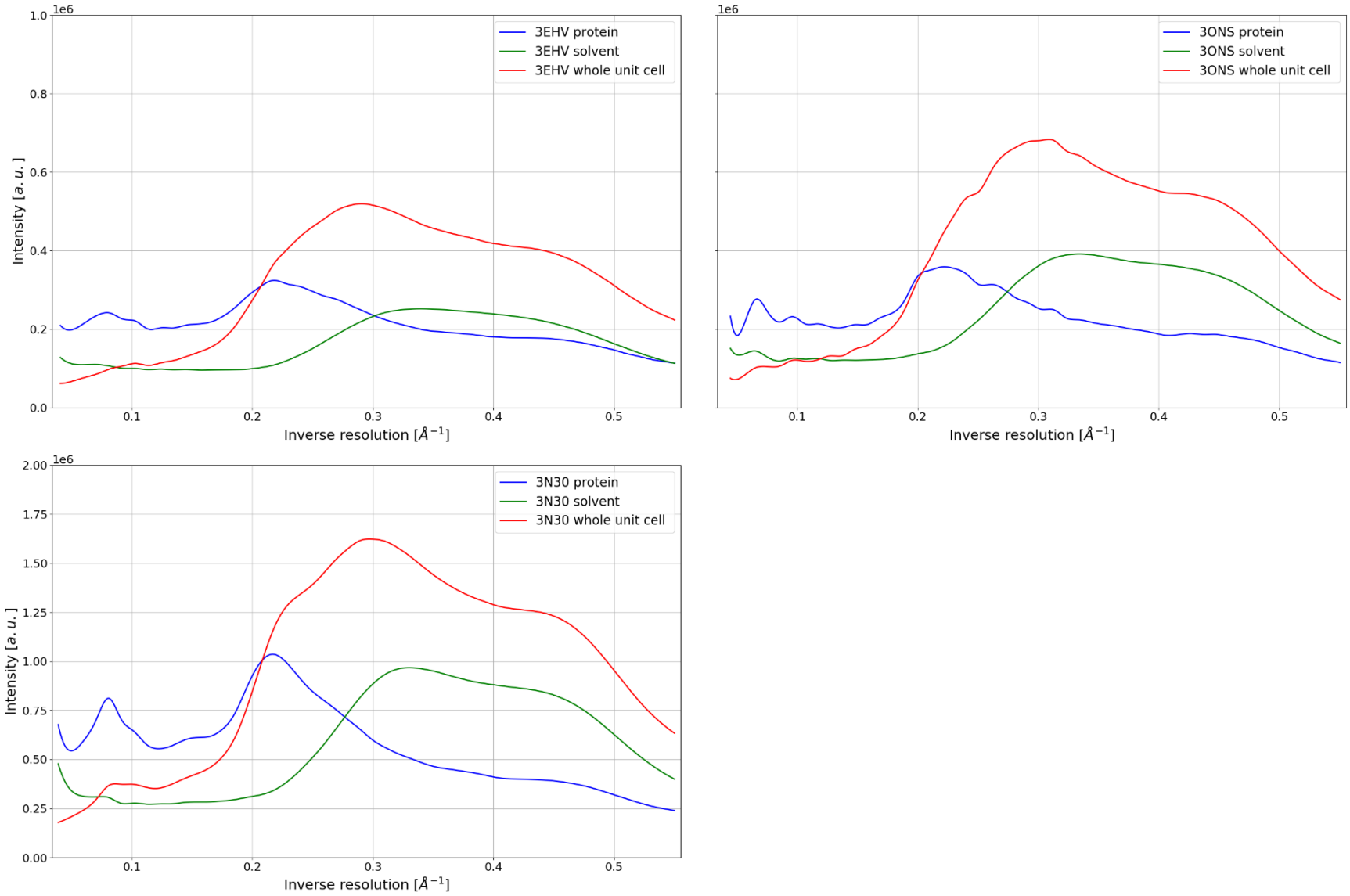

Figure 3.7. Diffuse scattering profiles of protein only, solvent only, and whole unit cell contents of the simulated crystals.

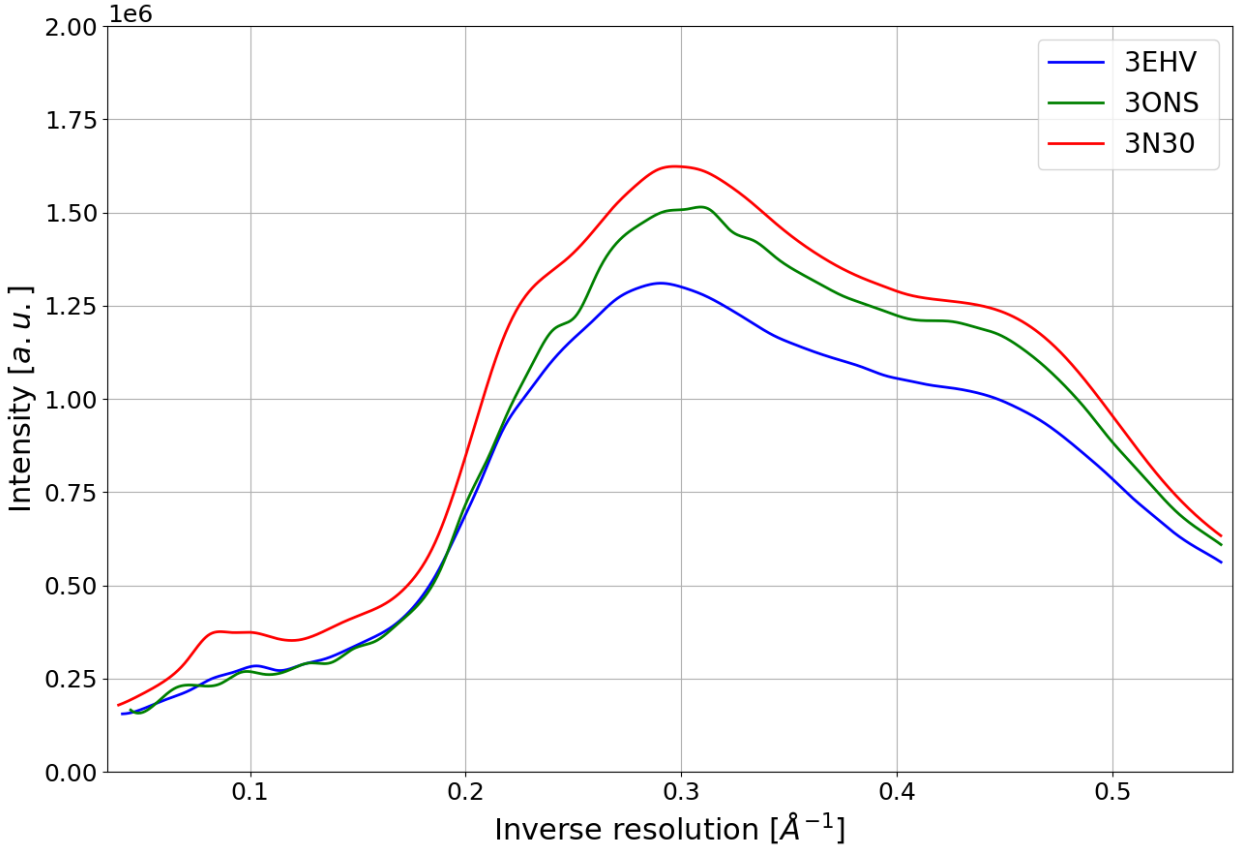

Figure 3.8. Comparison of the complete simulated diffuse scattering profiles from different crystals. Scaled by the total number of heavy atoms in the protein content.

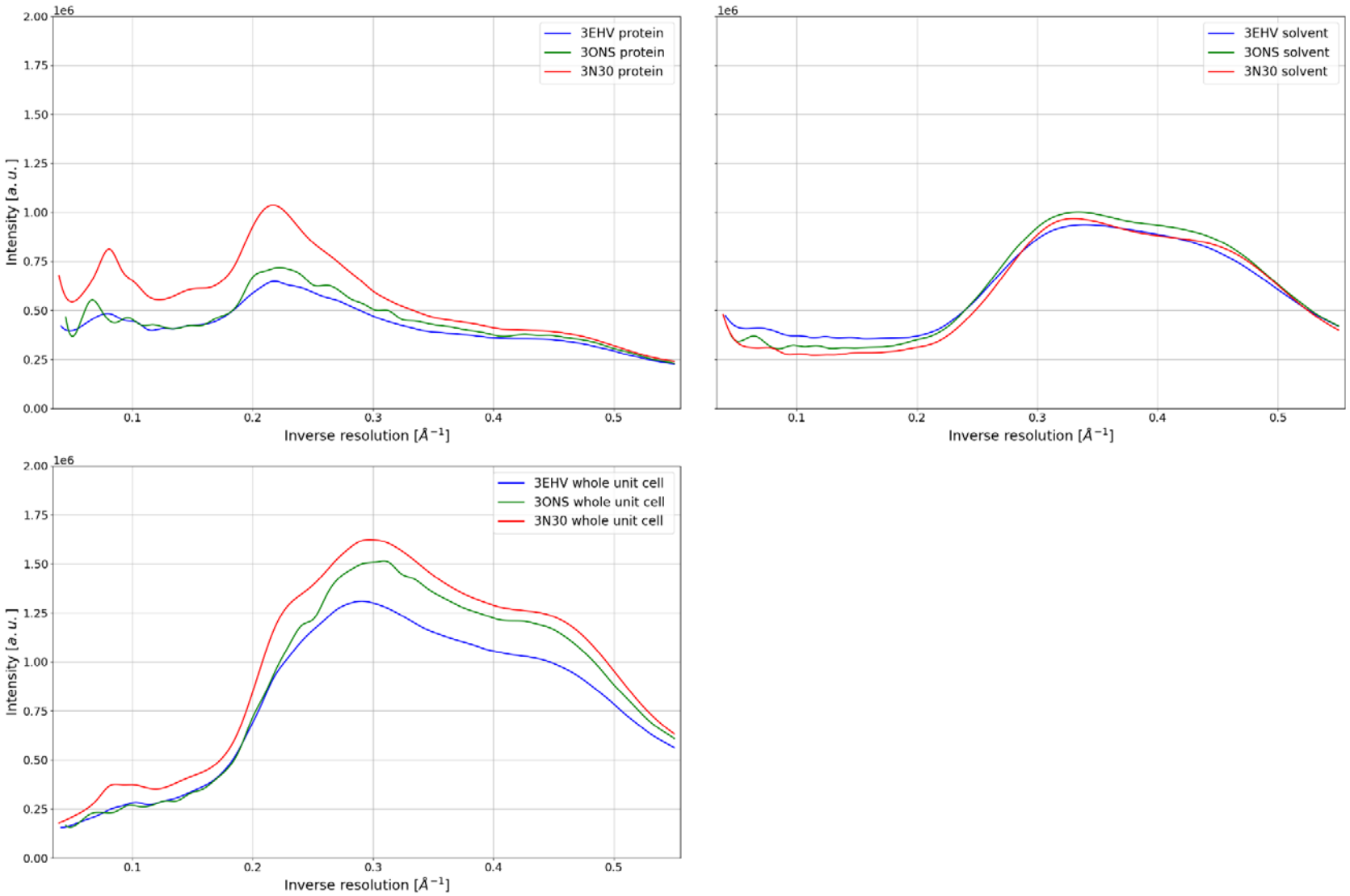

Figure 3.9. Diffuse scattering profiles of protein only, solvent only, and whole unit cell contents. Each panel is normalized based on the number of heavy atoms in the corresponding pseudo-trajectory.

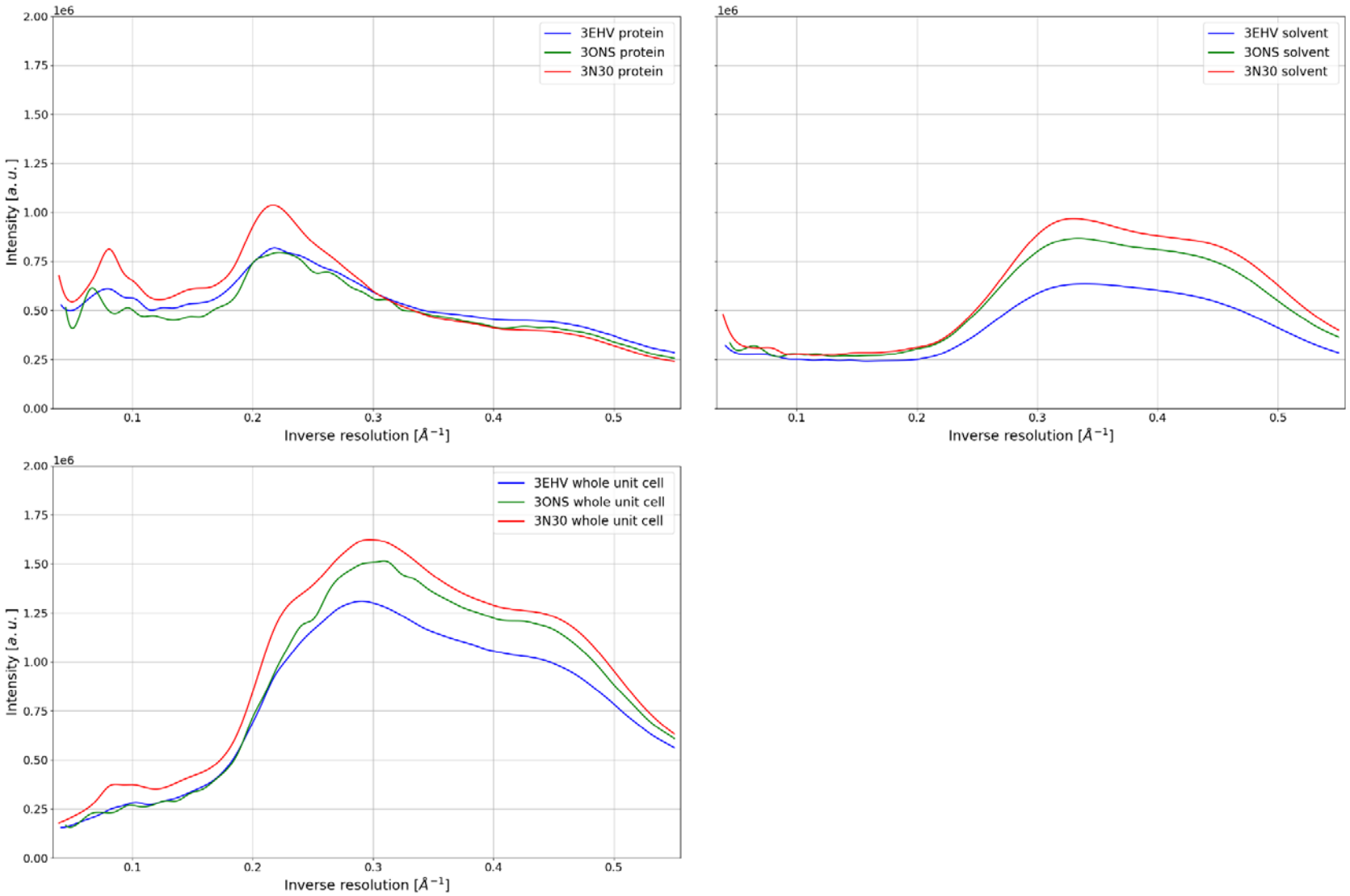

Figure 3.10. Diffuse scattering profiles of protein only, solvent only, and whole unit cell contents. Each panel is normalized based on the total number of heavy atoms in the whole unit cell simulation.

3.3.3 Solvent contribution and Babinet's principle

Interestingly, the whole unit cell diffuse scattering profiles at low resolution lie below the profiles from proteins and solvent (Figure 3.7). This implies that there is a cancelation of signal from protein and solvent at low resolution. Such effect is described in details by Podjarny, A. D. and Urzhumtsev, A. G.[4] and it originates from the structure factors of protein and solvent. As we use the formula based on structure factors to calculate the diffuse scattering, the phenomenon also manifests itself in the case of intensities. The idea has been covered in the introduction paragraph 1.1.4. Briefly, in the range of low resolutions the structure factors of protein and solvent regions are almost identical in magnitude and opposite in phase, since we do not distinguish between them, and the content of the cell is considered homogeneous. "As the resolution increases, density fluctuations appear inside these regions and the anticorrelation between the corresponding structure factors disappear"[4].

Another interesting point is that the solvent contribution is different among the three crystals relative to the protein counterpart as judged by the diffuse scattering intensity. However, it is almost the same in all three simulations if normalization is based on the number of heavy atoms in the corresponding pseudo-trajectory (see Table 3.1 for numbers). Since Cl^- ions have roughly twice more electrons than a typical heavy atom in our simulations, they contribute twice more to the intensities. Hence, we account for that by doubling the number of atoms corresponding to the anions. Thus, one can conclude that the dynamics of solvent is similar in the three simulations (see top right panel of Figure 3.9).

In the case of no normalization (Figure 3.7), we established the relative impact of the solvent profile on the protein profile to be much higher in 3ONS than it is in case of 3EHV and 3N30. One can also normalize the profiles based on the total number of heavy atoms in the original simulation (see Table 3.1, here we again count chlorine atoms twice when necessary). This time, the 3N30 solvent diffuse scattering profile has the biggest magnitude and the 3EHV solvent curve again has the lowest magnitude. The resultant curves are represented on Figure 3.10.

The dominance in the whole unit cell profiles of 3ONS over 3EHV might be explained by the solvent content – 51.93% over 39.97% by volume and 38.79% and 30.34% by the number of heavy atom scatterers. The similar explanation might be applied to the dominance of 3N30 over

3ONS – 56.12% vs. 51.93% by volume, and 44.77% vs. 38.79% by the number of heavy atom scatterers. The effect of self-cancellation takes place here as well.

MD simulations suggest that contributions from “disordered” solvent molecules appear to play an important role in more slowly varying parts of diffuse scattering intensities [110]. Hence, combining the outcomes of this and the previous paragraph, we conclude that the rotational motions input into diffuse scattering is hidden under the scaling issue and solvent contribution.

3.3.4 Patterson maps

At the American Crystallography Association meeting in 2019, I had a conversation about our results with Michael Wall. He also recommended to analyze Patterson maps generated from the obtained diffuse intensities. Further, we introduce the definition of the maps and report the results.

The relationship between Patterson maps and intensities is fundamentally the same as between electron density and structure factors (see paragraph 1.1.3). Patterson function is the Fourier transform of the intensities:

$$P(u,v,w) = \sum_{hkl} |\boldsymbol{F}_{hkl}|^2 \exp[-2\pi i(hu + kv + lw)],$$

while electron density distribution is the Fourier transform of structure factors. Similarly, the Patterson function is defined in the real space with the same periodic conditions as the crystal unit cell.

Such maps are used to identify the positions of heavy atoms. The peaks’ positions in the Patterson map correspond to interatomic distance vectors. The magnitudes of the peaks are proportional to the product of the respective atomic numbers. Since the vector corresponding to i-th and j-th atoms implies the existence of the oppositely directed vectors, the function is centrosymmetric (see, for example, Figure 3.11, panel A).

To analyze the effect of rigid-body motions, we extracted single chain trajectories from the pseudo-trajectories representing the rotational motions in our three crystals. Next, using small artificial unit cell dimensions, we generated the diffuse scattering intensities for the three single chains. Finally, we plugged these intensities into the Paterson function to generate the maps. Figure 3.11 shows the sections of the maps on xy-plane at zero z value.

The results turned out to be interesting and support our previous observations as in paragraph 3.3.1. As can be seen from Figure 3.6, 3EHV and 3ONS rotational motions in crystals are smaller than in 3N30 since the peaks are more pronounced and well defined. It is also supported by the values of the Patterson functions (see Table 3.3).

Unfortunately, the interpretation of Patterson maps in a direct comparison of the magnitudes of motions between the whole crystals having different unit cells is much more complicated. This is due to the same reason mentioned in paragraph 3.1: these maps depend on the unit cell dimensions. Thus, one will also see intermolecular distance vectors along with intramolecular atomic distance vectors. Therefore, one would need to use the native unit cell parameters for the diffuse scattering intensities. Moreover, the magnitude of the vectors corresponding to intermolecular atomic distances would significantly depend on the configuration of the asymmetric units in the crystal. Since the goal of the project was to answer the question whether one could see the difference in the scattering intensities, we decided not to proceed forward with this task.

Table 3.3. Summary of Patterson functions values obtained for single chain rotational motions in crystals 3EHV, 3ONS and 3N30.

	Minimum value (a.u.)	Maximum value (a.u.)	Mean value (a.u.)
3EHV	155.083	-20.820	-1.913e-11
3ONS	-22.804	140.607	1.317e-11
3N30	-11.185	94.982	3.990e-11

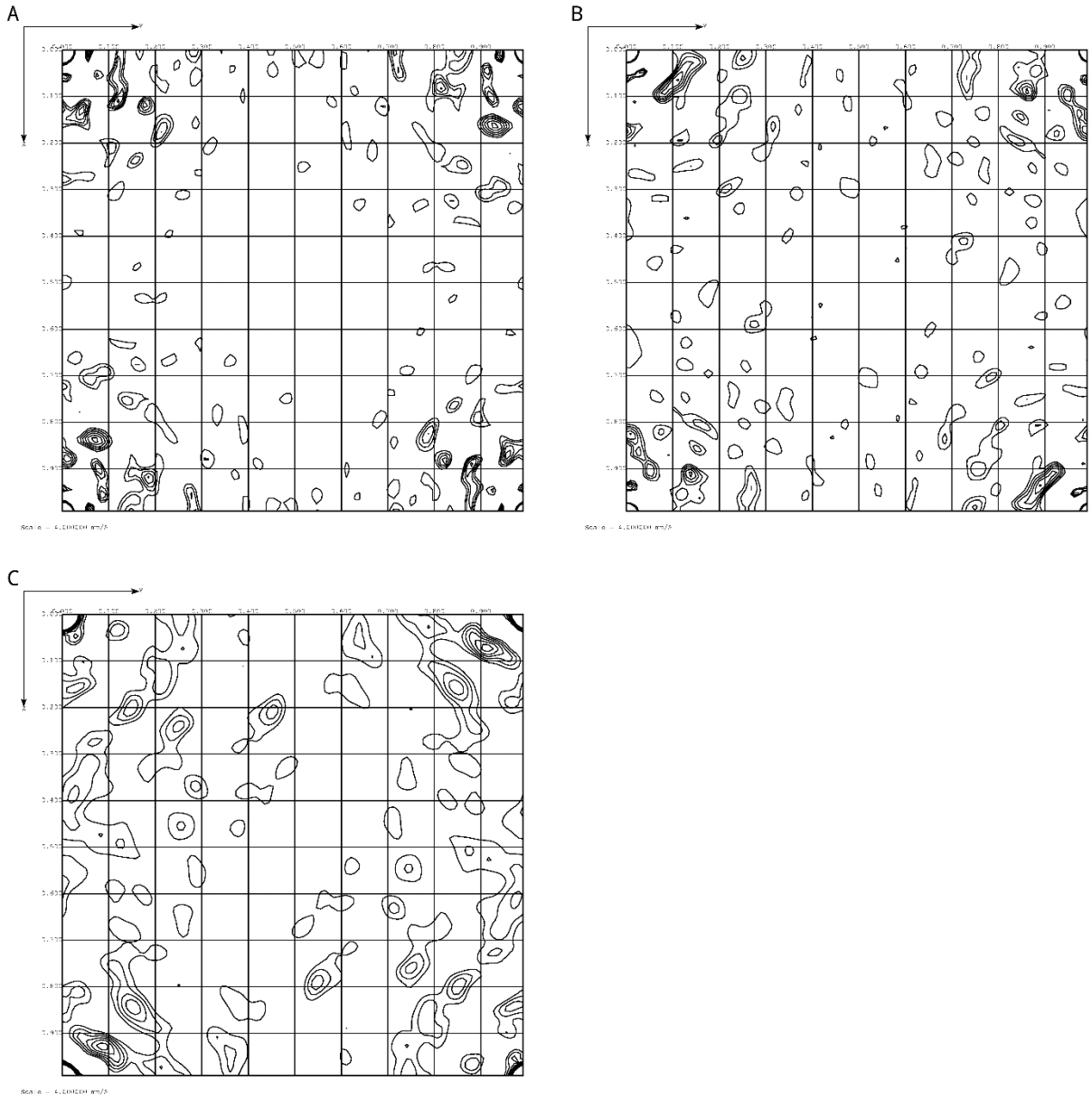

Figure 3.11. Representation of xy-plane sections of Patterson maps at z value of zero. Panels A, B and C correspond to the data obtained for single chain rotational motions in crystals 3EHV, 3ONS and 3N30, respectively. Panels dimensions are in fractional coordinated.

CHAPTER 4. DISCUSSION AND FUTURE DIRECTIONS

4.1 Macromolecular refinement

4.1.1 Discussion

To the best of our knowledge, we have created the first refinement software that operates with entire unit cells and employs periodic boundary conditions. Generally, ensemble models suggest that asymmetric units are independent during refinement. In Phenix tests of the whole unit cell approach, all the asymmetric units had to co-exist in the unit cell (paragraph 2.7.3). Unlike our Amber modification, Phenix cannot utilize the periodic boundary conditions. The results of the Phenix UC approach showed better agreement with experimental data than the traditional ASU approach, but mixed results in terms of geometric qualities of the re-refined models (see paragraph 2.6). Our Amber/Amber setup outperforms both ASU and UC Phenix setups in terms of geometry qualities on average, which we expected from the state-of-the-art force field. Even though Phenix UC approach showed similar R_{free} results as Amber/Amber protocol on the re-refinement tests, our results are quite striking on the simulated initial models (MD1, MD2 sets) and the deposited models against the ASU approach (see paragraphs 2.7.2 and 2.7.3) in terms of both R_{free} and MolProbity measures. This showcases the larger radius of convergence of our method against both the ASU and UC Phenix approaches. In part, we attribute that not only to the better force field but also to the different method of minimization of the target function (see paragraphs 1.2.2 and 2.3). However, if the starting model is way too poor (MD3 set), the comparison becomes meaningless since no protocols can refine structures well.

Moreover, our software is GPU-accelerated. As we noticed in paragraph 1.2.7, only FFX and xMDFF package are able to perform the refinement on graphics processors for now. They are one of the few to employ different from the widely used force fields as well (see paragraph 1.2.4). However, there are some major differences with our approach. First, it is impossible to plug in an explicit solvent into the FFX package. Second, FFX does the refinement using a single asymmetric unit and not an ensemble. The latter implies that the treatment of the implicit solvent to model non-bonded interactions is also limited and does not account for crystal packing. Yet, the polarizable force field also yields better geometry than those obtained by the classical force

fields. Third, xMDFF employs the real space crystallography term and constraints on the secondary structure of macromolecules. Therefore, our package is unique of its kind.

Paragraph 2.7.6 shows that, in principle, the inclusion of ligands is possible, however, it requires some additional procedures to derive the Amber force field parameters. At this point, we have no automation of the files' preparation such that they would be compatible with both Amber/Amber and Phenix/Amber setups due to different standards. Thus, we encourage manual intervention in this task. However, with the creation of a unified format that would not be an issue. The same is applicable to non-standard residues.

In case of already good starting model, one of the inevitable downsides of Amber/Amber setup is the initial increase of R-factors and RMSD against the target or deposited structure on the stages of minimization and heating (see paragraph 2.4.3), and the longer those stages are the worse the R-factors statistics become. Elimination of the minimization stage resulted in explosion of simulation setups. Elimination of the heating stage practically moved it to the first stage of refinement and R-factors still increased in the beginning of this stage. We attribute these issues to the correction of poor geometry features of the starting models, such as clash score and Ramachandran outliers.

Currently, there is a technical limitation for our method that we do not employ modelling of crystallographic water molecules during the refinement and focus on macromolecules only. The use of a high-resolution data implies that the geometry restraints become less critical, and the restraints based on experimental data are mainly important, since the ratio of observables to parameters increases. Therefore, our current approach is particularly valuable for lower resolution structures where the influence of high-quality force field and ensemble representation is especially significant, and the presence of explicit solvent is limited [148].

To summarize, the implementation of the proposed enhancements increases the geometric quality of the outcome in comparison with the performance of Phenix package [149], which core is The Computational Crystallography Toolbox (*cctbx*) library [150] also used by CCP4 suite [151]. Even the usage of Phenix in conjunction with an advanced Amber force field does not affect geometry as much. On top of this engine, we built a web server, which not only can be used even by a non-specialist from a handheld device but also delivers significant time savings.

4.1.2 Future directions

As we mentioned previously, our Amber modification currently refines only coordinates of proteins. However, in the refined models we do have B-factors, which researcher would also want to refine in principle. We have tried several schedules that included the B-factors refinement. For example, we added the B-factors refinement between stages (4) and (5) and after stage (5) (see paragraph 2.4.3). In this scenario, the agreement with the experimental data at the end of stage (5) was poorer than it was at the beginning. This is due to the maximum likelihood parameters estimation and somewhat similar to the increase of R-factors during the heating stage. The relative weight of the crystallographic terms becomes smaller than it was at the end of stage (4) and the structure is released for dynamics again. Therefore, more testing needs to be done

A potential amplification of the current state of Amber/Amber setup is accounting for diffraction data twinning. Since not all crystals are perfect and there are intergrown ones, the addition of it would potentially broadly expand the range of applications.

The implementation of time-averaged crystallographic restraints would also give our refinement protocol another boost [60]. Since it is known that such treatment not only improves R_{free} factors but also gives insights into the dynamics of proteins, such tool would be especially useful for GPU-based runs.

Another important application of Amber/Amber protocol would be alternate conformers optimization, especially those of the backbone residues, since they are hard to identify [152]. As our setup provides not only an ensemble model but also is supplied GPU-acceleration, the performance is drastically sped up in comparison with the current software. In paragraph 2.7.5, we presented a proof-of-concept. However, this is not a routine job yet.

In the same manner one would also try to model missing loop. Currently, all available software solutions design missing protein elements with either *ab initio* or template-based approaches. In its turn, Amber module can make use of experimental data. We did try to rebuild missing tail of the 3ONS ubiquitin structure, but our various protocol to treat the B-factors of the missing region failed so far by producing poor R_{free}. The solution of this problem is closely related to the problems with B-factors optimization and alternate conformers optimization.

Another interconnected issue is the crystallographic water detection. As stated in paragraph 2.4.3, we currently run a Phenix routine at the end of the coordinates' refinement. This routine picks up the bound water molecules by electron density map calculation and as it is done in the

cctbx library. However, we might implement such a procedure into the body of Amber to account for the bound molecules during the coordinates' refinement.

Finally, since diffuse scattering affects the precision of measured intensities, it would be useful to test the joint refinement against both Bragg and diffuse scattering data. That would be a blend of the two major concepts discussed in this dissertation. With the time-averaged crystallographic restraints, one could potentially try to estimate the atomic displacement parameters (B-factors). The refined model would not only agree with the experimental data but also have naturally derived B-factors.

At this point, we are working on the testing of our X-ray refinement related code as well as scaling procedures of the official codebase of Amber package. The latter makes the code completely independent of the *cctbx* library and speeds up the performance. The core of the code will be officially released in Amber 20, while the rest that is currently being tested will be rolled out through one of the updates.

4.2 Diffuse scattering

4.2.1 Discussion

Just like in the X-ray, NMR and straight MD experiments showed, by decomposition of pseudo-trajectories we can see the correlation between the magnitude of rotational motions and the corresponding diffuse scattering intensities amplitudes. However, there are several studies which indicate that one or the other type of motion is critical for diffuse scattering, for example, internal motions dominate in staphylococcal nuclease case [114], and rigid-body motions are claimed to be the main source of the diffuse scattering in cyclophilin A and lysozyme cases [116]. In the absence of solvent, we can see that the rotational motions dominate in ubiquitin crystals, but the amplitudes of the intensities from the rotational motions are of similar order to the ones from the translational motions. Moreover, the presence of the solvent in the experiment totally smears distinguishable differences between the different crystals of ubiquitin. Unfortunately, we conclude that even though one can incorporate the diffuse data into refinement. It seems virtually impossible to compare the motions solely based on that data.

4.2.2 Future directions

Interestingly, the profile shapes that we obtained for ubiquitin are quite similar to already known profiles for other proteins such as lysozyme (see Figure 4.1), or staphylococcal nuclease [110]. It would be useful to compare the diffuse scattering between different protein crystals to find what exactly is the source of different shape features. Another direction would be to produce crystal simulations of different proteins with different magnitudes of rotational motions belonging to the same space group. That way, we could compare not only the resolution against intensity profiles, but Miller indices against intensities, which is a three-dimensional map.

Alternatively, we could proceed with the ubiquitin in different crystalline forms and to produce supercell simulations. This would help to collect in-between Bragg peaks intensities and sample the reciprocal space more finely. Such approach was implemented by Wall to examine the diffuse scattering of the mentioned staphylococcal nuclease [114].

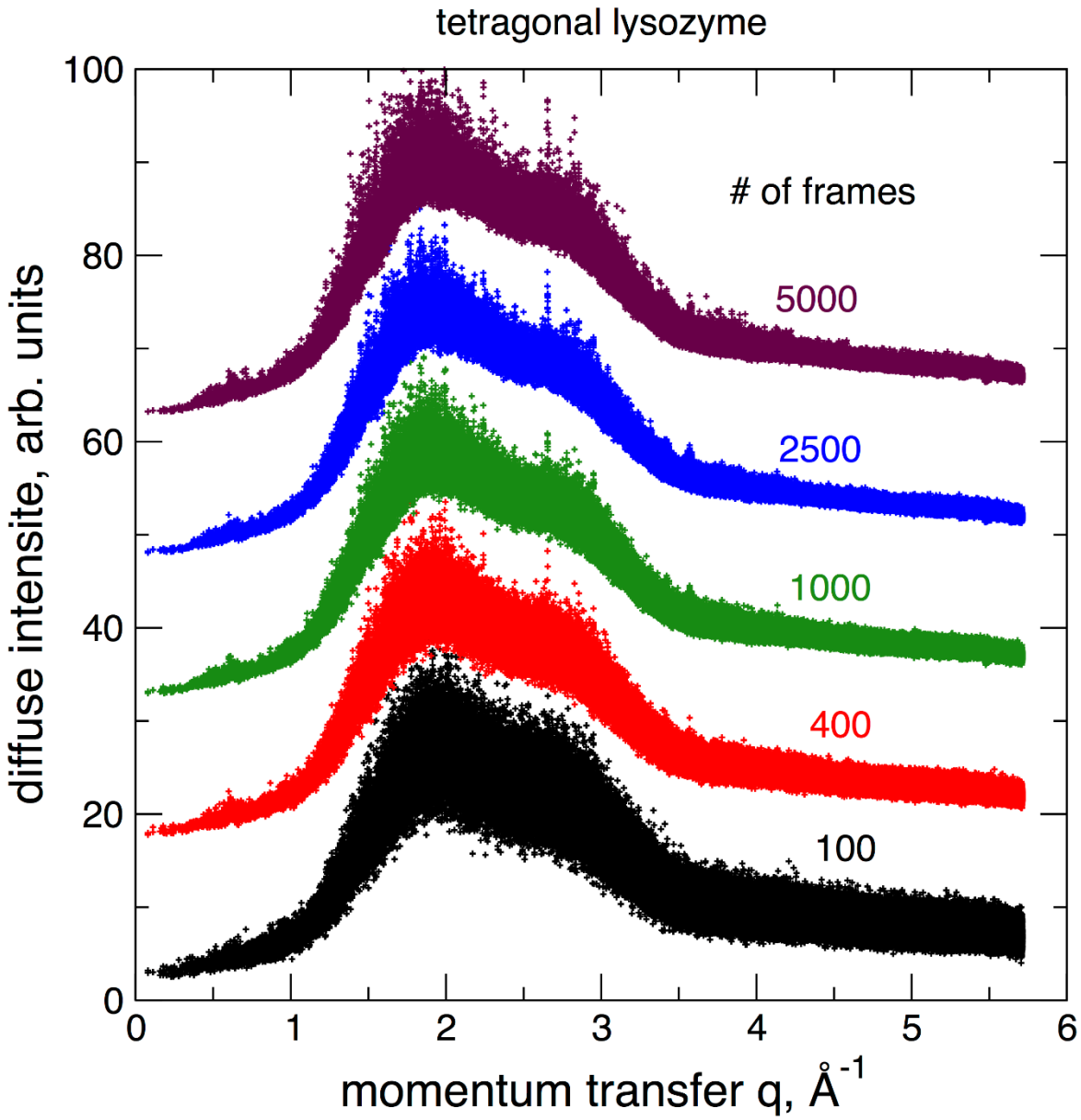

Figure 4.1. Diffuse intensities of tetragonal lysozyme depending the trajectory sampling frequency. X-axis is in momentum transfer units: $q = 2\pi|\boldsymbol{s}|$. 3.5 $Å^{-1}$ value corresponds to 30 Å resolution cutoff. Courtesy of D. A. Case.

REFERENCES

[1] P. P. Ewald, Ed., *Fifty Years of X-Ray Diffraction*. Boston, MA: Springer US, 1962.

[2] C. Colliex *et al.*, “Electron diffraction,” in *International Tables for Crystallography*, vol. C, Chester, England: International Union of Crystallography, 2006, pp. 259–429.

[3] C. X. Weichenberger, P. V. Afonine, K. Kantardjieff, and B. Rupp, “The solvent component of macromolecular crystals,” *Acta Crystallogr. Sect. D Biol. Crystallogr.*, vol. 71, pp. 1023–1038, 2015.

[4] A. D. Podjarny and A. G. Urzhumtsev, “Low-resolution phasing,” *Methods Enzymol.*, vol. 276, no. 1956, pp. 641–658, 1997.

[5] J.-S. Jiang and A. T. Brünger, “Protein Hydration Observed by X-ray Diffraction,” *J. Mol. Biol.*, vol. 243, no. 1, pp. 100–115, Oct. 1994.

[6] A. D. Booth, “Application of the method of steepest descents to X-ray structure analysis [15],” *Nature*, vol. 160, no. 4058, p. 196, Aug. 1947.

[7] A. D. Booth, “An expression for following the process of refinement in X-ray structure analysis using Fourier series,” *London, Edinburgh, Dublin Philos. Mag. J. Sci.*, vol. 36, no. 260, pp. 609–615, Sep. 1945.

[8] D. E. Tronrud, “Introduction to Macromolecular Refinement,” in *Macromolecular Crystallography Protocols, Volume 2*, vol. 364, New Jersey: Humana Press, 2007, pp. 231–254.

[9] J. W. Cooley and J. W. Tukey, “An algorithm for the machine calculation of complex Fourier series,” *Math. Comput.*, vol. 19, no. 90, pp. 297–297, May 1965.

[10] D. Sayre, “The calculation of structure factors by Fourier summation,” *Acta Crystallogr.*, vol. 4, no. 4, pp. 362–367, Jul. 1951.

[11] A. G. Urzhumtsev and V. Y. Lunin, “Introduction to crystallographic refinement of macromolecular atomic models,” *Crystallogr. Rev.*, vol. 25, no. 3, pp. 164–262, Jul. 2019.

[12] A. T. Brünger, “Free R-Value - A Novel Statistical Quantity for Assessing the Accuracy of Crystal-Structures,” *Nature*, vol. 355, no. 6359, pp. 472–475, 1992.

[13] V. Y. Lunin and A. G. Urzhumtsev, “Improvement of protein phases by coarse model modification,” *Acta Crystallogr. Sect. A Found. Crystallogr.*, vol. 40, no. 3, pp. 269–277, May 1984.

[14] G. Bricogne, “Maximum entropy and the foundations of direct methods,” *Acta Crystallogr. Sect. A Found. Crystallogr.*, vol. 40, no. 4, pp. 410–445, Jul. 1984.

[15] G. Bricogne, “A Bayesian statistical theory of the phase problem. I. A multichannel maximum-entropy formalism for constructing generalized joint probability distributions of structure factors,” *Acta Crystallogr. Sect. A Found. Crystallogr.*, vol. 44, no. 4, pp. 517–545, Jul. 1988.

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

[17] N. S. Pannu and R. J. Read, “Improved structure refinement through maximum likelihood,” *Acta Crystallogr. Sect. A Found. Crystallogr.*, vol. 52, no. 5, pp. 659–668, 1996.

[18] P. D. Adams, N. S. Pannu, R. J. Read, and A. T. Brünger, “Cross-validated maximum likelihood enhances crystallographic simulated annealing refinement,” *Proc. Natl. Acad. Sci.*, vol. 94, no. 10, pp. 5018–5023, May 1997.

[19] G. Murshudov, E. Dodson, and A. Vagin, “Application of maximum likelihood methods for macromolecular refinement,” in *Proceedings of the CCP4 study weekend: macromolecular refinement*, 1996.

[20] G. N. Murshudov, A. A. Vagin, and E. J. Dodson, “Refinement of macromolecular structures by the maximum-likelihood method,” *Acta Crystallogr. Sect. D Biol. Crystallogr.*, vol. 53, no. 3, pp. 240–255, 1997.

[21] N. S. Pannu, G. N. Murshudov, E. J. Dodson, and R. J. Read, “Incorporation of Prior Phase Information Strengthens Maximum-Likelihood Structure Refinement,” *Acta Crystallogr. Sect. D Biol. Crystallogr.*, vol. 54, no. 6, pp. 1285–1294, Nov. 1998.

[22] P. V. Afonine, R. W. Grosse-Kunstleve, and P. D. Adams, “A robust bulk-solvent correction and anisotropic scaling procedure,” *Acta Crystallogr. Sect. D Biol. Crystallogr.*, vol. 61, no. 7, pp. 850–855, Jul. 2005.

[23] P. V. Afonine *et al.*, “Towards automated crystallographic structure refinement with phenix.refine,” *Acta Crystallogr. Sect. D Biol. Crystallogr.*, vol. 68, no. 4, pp. 352–367, Apr. 2012.

[24] V. Y. Lunin, P. V. Afonine, and A. G. Urzhumtsev, “Likelihood-based refinement. I. Irremovable model errors,” *Acta Crystallogr. Sect. A Found. Crystallogr.*, vol. 58, no. 3, pp. 270–282, May 2002.

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

[26] A. T. Brünger *et al.*, “Crystallography & NMR system: A new software suite for macromolecular structure determination,” *Acta Crystallogr. Sect. D Biol. Crystallogr.*, vol. 54, no. 5, pp. 905–921, 1998.

[27] A. T. Brünger, J. Kuriyan, and M. Karplus, “Crystallographic R Factor Refinement by Molecular Dynamics,” *Science (80-.).*, vol. 235, no. 47897, pp. 458–60, 1987.

[28] G. M. Sheldrick and T. R. Schneider, “SHELXL: High-resolution refinement,” in *Methods in Enzymology*, vol. 277, 1997, pp. 319–343.

[29] J. H. Konnert, “A restrained-parameter structure-factor least-squares refinement procedure for large asymmetric units,” *Acta Crystallogr. Sect. A*, vol. 32, no. 4, pp. 614–617, 1976.

[30] W. A. Hendrickson, “Stereochemically restrained refinement of macromolecular structures,” *Methods Enzymol.*, vol. 115, pp. 252–270, Jan. 1985.

[31] E. Blanc, P. Roversi, C. Vonrhein, C. Flensburg, S. M. Lea, and G. Bricogne, “Refinement of severely incomplete structures with maximum likelihood in BUSTER-TNT,” *Acta Crystallogr. Sect. D Biol. Crystallogr.*, vol. 60, no. 12 I, pp. 2210–2221, Dec. 2004.

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

[33] R. A. Engh and R. Huber, “Structure quality and target parameters,” in *International Tables for Crystallography*, Chester, England: International Union of Crystallography, 2006, pp. 382–392.

[34] A. A. Vagin *et al.*, “REFMAC 5 dictionary: organization of prior chemical knowledge and guidelines for its use,” *Acta Crystallogr. Sect. D Biol. Crystallogr.*, vol. 60, no. 12, pp. 2184–2195, Dec. 2004.

[35] G. N. Murshudov *et al.*, “REFMAC 5 for the refinement of macromolecular crystal structures,” *Acta Crystallogr. Sect. D Biol. Crystallogr.*, vol. 67, no. 4, pp. 355–367, Apr. 2011.

[36] D. S. Berkholz, M. V. Shapovalov, R. L. Dunbrack, and P. A. Karplus, “Conformation Dependence of Backbone Geometry in Proteins,” *Structure*, vol. 17, no. 10, pp. 1316–1325, Oct. 2009.

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

[38] D. E. Tronrud, D. S. Berkholz, and P. A. Karplus, “Using a conformation-dependent stereochemical library improves crystallographic refinement of proteins,” *Acta Crystallogr. Sect. D Biol. Crystallogr.*, vol. 66, no. 7, pp. 834–842, Jul. 2010.

[39] N. W. Moriarty, D. E. Tronrud, P. D. Adams, and P. A. Karplus, “A new default restraint library for the protein backbone in Phenix: A conformation-dependent geometry goes mainstream,” *Acta Crystallogr. Sect. D Struct. Biol.*, vol. 72, no. 1, pp. 176–179, Jan. 2016.

[40] D. E. Tronrud and P. A. Karplus, “A conformation-dependent stereochemical library improves crystallographic refinement even at atomic resolution,” *Acta Crystallogr. Sect. D Biol. Crystallogr.*, vol. 67, no. 8, pp. 699–706, Aug. 2011.

[41] T. Petrova and A. Podjarny, “Protein crystallography at subatomic resolution,” *Reports Prog. Phys.*, vol. 67, no. 9, pp. 1565–1605, Sep. 2004.

[42] P. V. Afonine and P. D. Adams, “On the contribution of hydrogen atoms to X-ray scattering,” *Comput. Crystallogr. Newsl.*, vol. 3, pp. 18–21, 2012.

[43] Z. Dauter, G. N. Murshudov, and K. S. Wilson, “Refinement at atomic resolution,” in *International Tables for Crystallography*, Chester, England: International Union of Crystallography, 2006, pp. 393–402.

[44] M. Fujinaga, P. Gros, and W. F. van Gunsteren, “Testing the method of crystallographic refinement using molecular dynamics,” *J. Appl. Crystallogr.*, vol. 22, no. 1, pp. 1–8, Feb. 1989.

[45] P. Gros, M. Fujinaga, B. W. Dijkstra, K. H. Kalk, and W. G. J. Hol, “Crystallographic refinement by incorporation of molecular dynamics: thermostable serine protease thermitase complexed with eglin c,” *Acta Crystallogr. Sect. B Struct. Sci.*, vol. 45, no. 5, pp. 488–499, Oct. 1989.

[46] R. McGreevy *et al.*, “xMDFF: molecular dynamics flexible fitting of low-resolution X-ray structures,” *Acta Crystallogr. Sect. D Biol. Crystallogr.*, vol. 70, no. 9, pp. 2344–2355, Sep. 2014.

[47] L. G. Trabuco, E. Villa, E. Schreiner, C. B. Harrison, and K. Schulten, “Molecular dynamics flexible fitting: A practical guide to combine cryo-electron microscopy and X-ray crystallography,” *Methods*, vol. 49, no. 2, pp. 174–180, Oct. 2009.

[48] L. M. Rice, Y. Shamoo, and A. T. Brünger, “Phase Improvement by Multi-Start Simulated Annealing Refinement and Structure-Factor Averaging,” *J. Appl. Crystallogr.*, vol. 31, no. 5, pp. 798–805, 1998.

[49] P. Gros, W. F. van Gunsteren, and W. G. J. Hol, “Inclusion of thermal motion in crystallographic structures by restrained molecular dynamics,” *Science (80-.).*, vol. 249, no. 4973, pp. 1149–1152, Sep. 1990.

[50] J. Kuriyan, K. Ösapay, S. K. Burley, A. T. Brünger, W. A. Hendrickson, and M. Karplus, “Exploration of disorder in protein structures by X-ray restrained molecular dynamics,” *Proteins Struct. Funct. Bioinforma.*, vol. 10, no. 4, pp. 340–358, 1991.

[51] F. T. Burling and A. T. Brünger, “Thermal Motion and Conformational Disorder in Protein Crystal Structures: Comparison of Multi-Conformer and Time-Averaging Models,” *Isr. J. Chem.*, vol. 34, no. 2, pp. 165–175, 1994.

[52] F. T. Burling, W. I. Weis, K. M. Flaherty, and A. T. Brünger, “Direct observation of protein solvation and discrete disorder with experimental crystallographic phases,” *Science (80-.).*, vol. 271, no. 5245, pp. 72–77, 1996.

[53] J. Fodor, B. T. Riley, N. A. Borg, and A. M. Buckle, “Previously Hidden Dynamics at the TCR–Peptide–MHC Interface Revealed,” *J. Immunol.*, vol. 200, no. 12, pp. 4134–4145, Jun. 2018.

[54] J. E. Kohn, P. V. Afonine, J. Z. Ruscio, P. D. Adams, and T. Head-Gordon, “Evidence of Functional Protein Dynamics from X-Ray Crystallographic Ensembles,” *PLoS Comput. Biol.*, vol. 6, no. 8, p. e1000911, Aug. 2010.

[55] F. Forneris, B. T. Burnley, and P. Gros, “Ensemble refinement shows conformational flexibility in crystal structures of human complement factor D,” *Acta Crystallogr. Sect. D Biol. Crystallogr.*, vol. 70, no. 3, pp. 733–743, Mar. 2014.

[56] N. Furnham, T. L. Blundell, M. A. DePristo, and T. C. Terwilliger, “Is one solution good enough?,” *Nat. Struct. Mol. Biol.*, vol. 13, no. 3, pp. 184–185, Mar. 2006.

[57] T. C. Terwilliger *et al.*, “Interpretation of ensembles created by multiple iterative rebuilding of macromolecular models,” *Acta Crystallogr. Sect. D Biol. Crystallogr.*, vol. 63, no. 5, pp. 597–610, May 2007.

[58] E. J. Levin, D. A. Kondrashov, G. E. Wesenberg, and G. N. Phillips, “Ensemble Refinement of Protein Crystal Structures: Validation and Application,” *Structure*, vol. 15, no. 9, pp. 1040–1052, Sep. 2007.

[59] R. A. Woldeyes, D. A. Sivak, and J. S. Fraser, “E pluribus unum, no more: From one crystal, many conformations,” *Curr. Opin. Struct. Biol.*, vol. 28, no. 1, pp. 56–62, Oct. 2014.

[60] B. T. Burnley, P. V. Afonine, P. D. Adams, and P. Gros, “Modelling dynamics in protein crystal structures by ensemble refinement,” *Elife*, vol. 1, no. 2, p. e00311, 2012.

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

[62] D. S. Cerutti, I. Le Trong, R. E. Stenkamp, and T. P. Lybrand, “Simulations of a protein crystal: Explicit treatment of crystallization conditions links theory and experiment in the streptavidin-biotin complex,” *Biochemistry*, vol. 47, no. 46, pp. 12065–12077, 2008.

[63] B. Xia, V. Tsui, D. A. Case, H. J. Dyson, and P. E. Wright, “Comparison of protein solution structures refined by molecular dynamics simulation in vacuum, with a generalized Born model, and with explicit water,” *J. Biomol. NMR*, vol. 22, no. 4, pp. 317–331, 2002.

[64] D. E. Tanner, K.-Y. Chan, J. C. Phillips, and K. Schulten, “Parallel Generalized Born Implicit Solvent Calculations with NAMD,” *J. Chem. Theory Comput.*, vol. 7, no. 11, pp. 3635–3642, Nov. 2011.

[65] Y. Qi, J. Lee, A. Singharoy, R. McGreevy, K. Schulten, and W. Im, “CHARMM-GUI MDFF/xMDFF Utilizer for Molecular Dynamics Flexible Fitting Simulations in Various Environments,” *J. Phys. Chem. B*, vol. 121, no. 15, pp. 3718–3723, Apr. 2017.

[66] L. Moulinier, D. A. Case, and T. Simonson, “Reintroducing electrostatics into protein X-ray structure refinement: Bulk solvent treated as a dielectric continuum,” *Acta Crystallogr. - Sect. D Biol. Crystallogr.*, vol. 59, no. 12, pp. 2094–2103, 2003.

[67] T. D. Fenn *et al.*, “Reintroducing Electrostatics into Macromolecular Crystallographic Refinement: Application to Neutron Crystallography and DNA Hydration,” *Structure*, vol. 19, no. 4, pp. 523–533, Apr. 2011.

[68] M. J. Schnieders, T. D. Fenn, and V. S. Pande, “Polarizable atomic multipole X-ray refinement: Particle mesh ewald electrostatics for macromolecular crystals,” *J. Chem. Theory Comput.*, vol. 7, no. 4, pp. 1141–1156, 2011.

[69] T. D. Fenn and M. J. Schnieders, “Polarizable atomic multipole X-ray refinement: Weighting schemes for macromolecular diffraction,” *Acta Crystallogr. Sect. D Biol. Crystallogr.*, vol. 67, no. 11, pp. 957–965, 2011.

[70] T. D. Fenn, M. J. Schnieders, A. T. Brünger, and V. S. Pande, “Polarizable Atomic Multipole X-Ray Refinement: Hydration Geometry and Application to Macromolecules,” *Biophys. J.*, vol. 98, no. 12, pp. 2984–2992, Jun. 2010.

[71] F. Dimaio, N. Echols, J. J. Headd, T. C. Terwilliger, P. D. Adams, and D. Baker, “Improved low-resolution crystallographic refinement with Phenix and Rosetta,” *Nat. Methods*, vol. 10, no. 11, pp. 1102–1106, 2013.

[72] J. B. Clarage and G. N. Phillips, “Cross-validation tests of time-averaged molecular dynamics refinements for determination of protein structures by X-ray crystallography,” *Acta Crystallogr. Sect. D Biol. Crystallogr.*, vol. 50, no. 1, pp. 24–36, Jan. 1994.

[73] T. D. Romo, J. B. Clarage, D. C. Sorensen, and G. N. Phillips, “Automatic identification of discrete substates in proteins: Singular value decomposition analysis of time-averaged crystallographic refinements,” *Proteins Struct. Funct. Genet.*, vol. 22, no. 4, pp. 311–321, Aug. 1995.

[74] R. Soheilifard, D. E. Makarov, and G. J. Rodin, “Critical evaluation of simple network models of protein dynamics and their comparison with crystallographic B-factors,” *Phys. Biol.*, vol. 5, no. 2, p. 026008, Jun. 2008.

[75] D. W. Li and R. Brüschweiler, “All-atom contact model for understanding protein dynamics from crystallographic B-factors,” *Biophys. J.*, vol. 96, no. 8, pp. 3074–3081, Apr. 2009.

[76] G. Song and R. L. Jernigan, “vGNM: A Better Model for Understanding the Dynamics of Proteins in Crystals,” *J. Mol. Biol.*, vol. 369, no. 3, pp. 880–893, Jun. 2007.

[77] B. Stec, Zhou Rongsheng, and M. M. Teeter, “Full-matrix refinement of the protein crambin at 0.83 Angstrom and 130 K,” *Acta Crystallogr. - Sect. D Biol. Crystallogr.*, vol. 51, no. 5, pp. 663–681, Sep. 1995.

[78] F. G. Parak, “Physical aspects of protein dynamics,” *Reports Prog. Phys.*, vol. 66, no. 2, pp. 103–129, Feb. 2003.

[79] V. Schomaker and K. N. Trueblood, “On the rigid-body motion of molecules in crystals,” *Acta Crystallogr. Sect. B Struct. Crystallogr. Cryst. Chem.*, vol. 24, no. 1, pp. 63–76, Jan. 1968.

[80] C. Chaudhry, A. L. Horwich, A. T. Brünger, and P. D. Adams, “Exploring the structural dynamics of the E. coli chaperonin GroEL using translation-libration-screw crystallographic refinement of intermediate states,” *J. Mol. Biol.*, vol. 342, no. 1, pp. 229–245, 2004.

[81] P. B. Moore, “On the Relationship between Diffraction Patterns and Motions in Macromolecular Crystals,” *Structure*, vol. 17, no. 10, pp. 1307–1315, 2009.

[82] A. H. Van Benschoten *et al.*, “Predicting X-ray diffuse scattering from translation–libration–screw structural ensembles,” *Acta Crystallogr. Sect. D Biol. Crystallogr.*, vol. 71, no. 8, pp. 1657–1667, Aug. 2015.

[83] A. H. Van Benschoten *et al.*, “Measuring and modeling diffuse scattering in protein X-ray crystallography,” *Proc. Natl. Acad. Sci.*, vol. 113, no. 15, pp. 4069–4074, 2016.

[84] P. Ma *et al.*, “Observing the overall rocking motion of a protein in a crystal,” *Nat. Commun.*, vol. 6, no. 1, p. 8361, Dec. 2015.

[85] M. S. Doscher and F. M. Richards, “The Activity of an Enzyme in the Crystalline State: Ribonuclease {S},” *J. Biol. Chem.*, vol. 238, no. 7, pp. 2399–2406, 1963.

[86] S. Yeremenko, I. H. M. van Stokkum, K. Moffat, and K. J. Hellingwerf, “Influence of the Crystalline State on Photoinduced Dynamics of Photoactive Yellow Protein Studied by Ultraviolet-Visible Transient Absorption Spectroscopy,” *Biophys. J.*, vol. 90, no. 11, pp. 4224–4235, Jun. 2006.

[87] M. B. C. Moncrief, R. P. Hausinger, L. G. Hom, E. Jabri, and P. Andrew Karplus, “Urease activity in the crystalline state,” *Protein Sci.*, vol. 4, no. 10, pp. 2234–2236, Oct. 1995.

[88] M. Schmidt and D. K. Saldin, “Enzyme transient state kinetics in crystal and solution from the perspective of a time-resolved crystallographer,” *Struct. Dyn.*, vol. 1, no. 2, p. 024701, Mar. 2014.

[89] I. D. Glover, G. W. Harris, J. R. Helliwell, and D. S. Moss, “The variety of X-ray diffuse scattering from macromolecular crystals and its respective components,” *Acta Crystallogr. Sect. B Struct. Sci.*, vol. 47, no. 6, pp. 960–968, Dec. 1991.

[90] T. R. Welberry and T. Weber, “One hundred years of diffuse scattering,” *Crystallogr. Rev.*, vol. 22, no. 1, pp. 2–78, Jan. 2016.

[91] D. Boylan and G. N. Phillips, “Motions of Tropomyosin: Characterization of Anisotropic Motions and Coupled Displacements in Crystals,” *Biophys. J.*, vol. 49, no. 1, pp. 76–78, Jan. 1986.

[92] S. Chacko and G. N. Phillips, “Diffuse x-ray scattering from tropomyosin crystals,” *Biophys. J.*, vol. 61, no. 5, pp. 1256–1266, May 1992.

[93] G. N. Phillips, J. P. Fillers, and C. Cohen, “Motions of tropomyosin. Crystal as metaphor,” *Biophys. J.*, vol. 32, no. 1, pp. 485–502, Oct. 1980.

[94] D. L. D. Caspar, J. Clarage, D. M. Salunke, and M. Clarage, “Liquid-like movements in crystalline insulin,” *Nature*, vol. 332, no. 6165, pp. 659–662, Apr. 1988.

[95] J. P. B. J. Doucet, “Molecular dynamics studies by analysis of the X-ray diffuse scattering from lysozyme crystals,” *Nature*, vol. 325. p. 643, 1987.

[96] J. B. Clarage, M. S. Clarage, W. C. Phillips, R. M. Sweet, and D. L. D. Caspar, “Correlations of atomic movements in lysozyme crystals,” *Proteins Struct. Funct. Genet.*, vol. 12, no. 2, pp. 145–157, Feb. 1992.

[97] J. Doucet, J. P. Benoit, W. B. T. Cruse, T. Prange, and O. Kennard, “Coexistence of A- and B-form DNA in a single crystal lattice,” *Nature*, vol. 337, no. 6203, pp. 190–192, Jan. 1989.

[98] J. R. HELLIWELL, I. D. GLOVER, A. JONES, E. PANTOS, and D. S. MOSS, “Protein dynamics: use of computer graphics and protein crystal diffuse scattering recorded with synchrotron X-radiation,” *Biochem. Soc. Trans.*, vol. 14, no. 3, pp. 653–655, Jun. 1986.

[99] K. Mizuguchi, A. Kidera, and N. Gō, “Collective motions in proteins investigated by X-ray diffuse scattering,” *Proteins Struct. Funct. Genet.*, vol. 18, no. 1, pp. 34–48, Jan. 1994.

[100] P. Faure, A. Micu, D. Pérahia, J. Doucet, J. C. Smith, and J. P. Benoit, “Correlated intramolecular motions and diffuse x–ray scattering in lysozyme,” *Nat. Struct. Mol. Biol.*, vol. 1, no. 2, pp. 124–128, Feb. 1994.

[101] A. M. Micu and J. C. Smith, “SERENA: a program for calculating x-ray diffuse scattering intensities from molecular dynamics trajectories,” *Comput. Phys. Commun.*, vol. 91, no. 1–3, pp. 331–338, 1995.

[102] J. Pérez, P. Faure, and J.-P. Benoit, “Molecular Rigid-Body Displacements in a Tetragonal Lysozyme Crystal Confirmed by X-ray Diffuse Scattering,” *Acta Crystallogr. Sect. D Biol. Crystallogr.*, vol. 52, no. 4, pp. 722–729, Jul. 1996.

[103] A. R. Kolatkar, J. B. Clarage, and G. N. Phillips Jnr, “Analysis of diffuse scattering from yeast initiator tRNA crystals,” *Acta Crystallogr. Sect. D Biol. Crystallogr.*, vol. 50, no. 2, pp. 210–218, Mar. 1994.

[104] J. B. Clarage and G. N. Phillips, “Analysis of diffuse scattering and relation to molecular motion,” *Methods Enzymol.*, vol. 277, pp. 407–432, 1997.

[105] J. B. Clarage, T. Romo, B. K. Andrews, B. M. Pettitt, and G. N. Phillips, “A sampling problem in molecular dynamics simulations of macromolecules.,” *Proc. Natl. Acad. Sci.*, vol. 92, no. 8, pp. 3288–3292, Apr. 1995.

[106] M. E. Wall, J. B. Clarage, and G. N. Phillips, “Motions of calmodulin characterized using both Bragg and diffuse X-ray scattering.,” *Structure*, vol. 5, no. 12, pp. 1599–1612, 1997.

[107] M. E. Wall, S. E. Ealick, and S. M. Gruner, “Three-dimensional diffuse x-ray scattering from crystals of Staphylococcal nuclease,” *Proc. Natl. Acad. Sci.*, vol. 94, no. 12, pp. 6180–6184, Jun. 1997.

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

[109] L. Meinhold and J. C. Smith, “Fluctuations and Correlations in Crystalline Protein Dynamics: A Simulation Analysis of Staphylococcal Nuclease,” *Biophys. J.*, vol. 88, no. 4, pp. 2554–2563, Apr. 2005.

[110] L. Meinhold and J. C. Smith, “Correlated dynamics determining X-ray diffuse scattering from a crystalline protein revealed by molecular dynamics simulation,” *Phys. Rev. Lett.*, vol. 95, no. 21, pp. 1–4, 2005.

[111] L. Meinhold and J. C. Smith, “Protein dynamics from X-ray crystallography: Anisotropic, global motion in diffuse scattering patterns,” *Proteins Struct. Funct. Genet.*, vol. 66, no. 4, pp. 941–953, 2007.

[112] D. Riccardi, Q. Cui, and G. N. Phillips, “Evaluating Elastic Network Models of Crystalline Biological Molecules with Temperature Factors, Correlated Motions, and Diffuse X-Ray Scattering,” *Biophys. J.*, vol. 99, no. 8, pp. 2616–2625, Oct. 2010.

[113] M. E. Wall, A. H. Van Benschoten, N. K. Sauter, P. D. Adams, J. S. Fraser, and T. C. Terwilliger, “Conformational dynamics of a crystalline protein from microsecond-scale molecular dynamics simulations and diffuse X-ray scattering,” *Proc. Natl. Acad. Sci.*, vol. 111, no. 50, pp. 17887–17892, Dec. 2014.

[114] M. E. Wall, “Internal protein motions in molecular-dynamics simulations of Bragg and diffuse X-ray scattering,” *IUCrJ*, vol. 5, no. 2, pp. 172–181, Mar. 2018.

[115] A. Peck, F. Poitevin, and T. J. Lane, “Intermolecular correlations are necessary to explain diffuse scattering from protein crystals,” *IUCrJ*, vol. 5, pp. 211–222, 2018.

[116] T. de Klijn, A. M. M. Schreurs, and L. M. J. Kroon-Batenburg, “Rigid-body motion is the main source of diffuse scattering in protein crystallography,” *IUCrJ*, vol. 6, no. 2, pp. 277–289, Mar. 2019.

[117] S. P. Meisburger, W. C. Thomas, M. B. Watkins, and N. Ando, “X-ray Scattering Studies of Protein Structural Dynamics,” *Chem. Rev.*, vol. 117, no. 12, pp. 7615–7672, Jun. 2017.

[118] A. Urzhumtsev, P. V. Afonine, A. H. Van Benschoten, J. S. Fraser, and P. D. Adams, “From deep TLS validation to ensembles of atomic models built from elemental motions,” *Acta Crystallogr. Sect. D Biol. Crystallogr.*, vol. 71, no. 8, pp. 1668–1683, Aug. 2015.

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

[120] M. E. Wall, P. D. Adams, J. S. Fraser, and N. K. Sauter, “Diffuse x-ray scattering to model protein motions,” *Structure*, vol. 22, no. 2, pp. 182–184, 2014.

[121] M. A. Wilson, “Visualizing networks of mobility in proteins,” *Nat. Methods*, vol. 10, no. 9, pp. 835–837, 2013.

[122] K. Ayyer *et al.*, “Macromolecular diffractive imaging using imperfect crystals,” *Nature*, vol. 530, no. 7589, pp. 202–206, Feb. 2016.

[123] S. P. Meisburger and N. Ando, “Correlated motions from crystallography beyond diffraction,” *Acc. Chem. Res.*, vol. 50, no. 3, pp. 580–583, 2017.

[124] M. E. Wall, A. M. Wolff, and J. S. Fraser, “Bringing diffuse X-ray scattering into focus,” *Curr. Opin. Struct. Biol.*, vol. 50, no. Ccd, pp. 109–116, 2018.

[125] D. S. Cerutti and D. A. Case, “Molecular dynamics simulations of macromolecular crystals,” *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, vol. 9, no. 4, pp. 1–14, Jul. 2019.

[126] D. A. Case, “Amber 2017 Reference Manual,” *Amber*, 2017.

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

[128] J. Li *et al.*, “Molecular dynamics simulations of a new branched antimicrobial peptide: A comparison of force fields,” *J. Chem. Phys.*, vol. 137, no. 21, 2012.

[129] F. Martin-Garcia, E. Papaleo, P. Gomez-Puertas, W. Boomsma, and K. Lindorff-Larsen, “Comparing molecular dynamics force fields in the essential subspace,” *PLoS One*, vol. 10, no. 3, pp. 1–16, 2015.

[130] P. A. Janowski, C. Liu, J. Deckman, and D. A. Case, “Molecular dynamics simulation of triclinic lysozyme in a crystal lattice,” *Protein Sci.*, vol. 25, no. 1, pp. 87–102, 2016.

[131] A. M. Silva and M. G. Rossmann, “The refinement of southern bean mosaic virus in reciprocal space,” *Acta Crystallogr. Sect. B Struct. Sci.*, vol. 41, no. 2, pp. 147–157, Apr. 1985.

[132] A. T. Brünger and P. D. Adams, “Molecular dynamics applied to X-ray structure refinement,” *Acc. Chem. Res.*, vol. 35, no. 6, pp. 404–412, 2002.

[133] A. T. Brünger and P. D. Adams, “1.7 Refinement of X-ray Crystal Structures,” in *Comprehensive Biophysics*, Elsevier, 2012, pp. 105–115.

[134] R. P. Joosten, T. Womack, G. Vriend, and G. Bricogne, “Re-refinement from deposited X-ray data can deliver improved models for most PDB entries,” *Acta Crystallogr. Sect. D Biol. Crystallogr.*, vol. 65, no. 2, pp. 176–185, Feb. 2009.

[135] P. J. Brown, A. G. Fox, E. N. Maslen, M. A. O’Keefe, and B. T. M. Willis, “Intensity of diffracted intensities,” in *International Tables for Crystallography*, vol. C, E. Prince, Ed. Chester, England: International Union of Crystallography, 2006, pp. 554–595.

[136] P. V. Afonine, R. W. Grosse-Kunstleve, P. D. Adams, and A. G. Urzhumtsev, “Bulk-solvent and overall scaling revisited: faster calculations, improved results,” *Acta Crystallogr. Sect. D Biol. Crystallogr.*, vol. 69, no. 4, pp. 625–634, Apr. 2013.

[137] V. Y. Lunin and T. P. Skovoroda, “R -free likelihood-based estimates of errors for phases calculated from atomic models,” *Acta Crystallogr. Sect. A Found. Crystallogr.*, vol. 51, no. 6, pp. 880–887, Nov. 1995.

[138] P. V. Afonine, V. Y. Lunin, and A. G. Urzhumtsev, “MLMF: Least-squares approximation of likelihood based refinement criteria,” *J. Appl. Crystallogr.*, vol. 36, no. 1, pp. 158–159, 2003.

[139] N. W. Moriarty, “Computational Crystallography Newsletter,” vol. 2, 2011.

[140] R. W. Grosse-Kunstleve, N. W. Moriarty, and P. D. Adams, “Torsion angle refinement and dynamics as a tool to aid crystallographic structure determination,” *Proc. ASME Des. Eng. Tech. Conf.*, vol. 4, no. PARTS A, B AND C, pp. 1477–1485, 2009.

[141] O. Kovalevskiy, R. A. Nicholls, and G. N. Murshudov, “Automated refinement of macromolecular structures at low resolution using prior information,” *Acta Crystallogr. Sect. D Struct. Biol.*, vol. 72, no. 10, pp. 1149–1161, Oct. 2016.

[142] R. Salomon-Ferrer, A. W. Go, D. Poole, S. Le Grand, and R. C. Walker, “Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald.”

[143] W. Kabsch and C. Sander, “Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features,” *Biopolymers*, vol. 22, no. 12, pp. 2577–2637, Dec. 1983.

[144] W. G. Touw *et al.*, “A series of PDB-related databanks for everyday needs,” *Nucleic Acids Res.*, vol. 43, no. D1, pp. D364–D368, Jan. 2015.

[145] M. J. Frisch *et al.*, “Gaussian 16,” *Gaussian, Inc., Wallingford CT,*. 2016.

[146] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case, “Development and testing of a general amber force field,” *J. Comput. Chem.*, vol. 25, no. 9, pp. 1157–1174, Jul. 2004.

[147] J. Wang, W. Wang, P. A. Kollman, and D. A. Case, “Automatic atom type and bond type perception in molecular mechanical calculations,” *J. Mol. Graph. Model.*, vol. 25, no. 2, pp. 247–260, Oct. 2006.

[148] A. T. Brünger, P. D. Adams, and L. M. Rice, “Enhanced macromolecular refinement by simulated annealing,” in *International Tables for Crystallography*, 2012, pp. 466–473.

[149] P. D. Adams *et al.*, “PHENIX: A comprehensive Python-based system for macromolecular structure solution,” *Acta Crystallogr. Sect. D Biol. Crystallogr.*, vol. 66, no. 2, pp. 213–221, 2010.

[150] R. W. Grosse-Kunstleve, N. K. Sauter, N. W. Moriarty, and P. D. Adams, “The Computational Crystallography Toolbox: Crystallographic algorithms in a reusable software framework,” *J. Appl. Crystallogr.*, vol. 35, no. 1, pp. 126–136, Feb. 2002.

[151] M. D. Winn *et al.*, “Overview of the CCP4 suite and current developments,” *Acta Crystallogr. Sect. D Biol. Crystallogr.*, vol. 67, no. 4, pp. 235–242, 2011.

[152] D. A. Keedy, J. S. Fraser, and H. van den Bedem, “Exposing Hidden Alternative Backbone Conformations in X-ray Crystallography Using qFit,” *PLOS Comput. Biol.*, vol. 11, no. 10, p. e1004507, Oct. 2015.

VITA

Oleg Mikhailovskii was born in Angarsk, Irkutskaya oblast, near lake Baikal in Russia. At the age of six, his family moved to Voronezh, Russia, where he entered elementary school. At the age of twelve, he moved to Saint Petersburg, Russia, where he continued to go to middle school. At high school level, he entered Laboratory of Continuous Mathematical Education.

After the high school graduation, he got accepted to the Mathematics and Mechanics faculty of the Saint Petersburg State University. There, he received an equivalent of Master's Degree of Science in Mathematics with the specialization at differential equations and dynamics systems. Later, he remained interested in science and entered PULSe program at Purdue University in August 2014. After one year of rotations, he joined the Skrynnikov group. Currently, he is a PhD candidate in the Department of Chemistry.

APPENDIX. ABSOLUTE SCALE VALUES OF REFINEMENTS

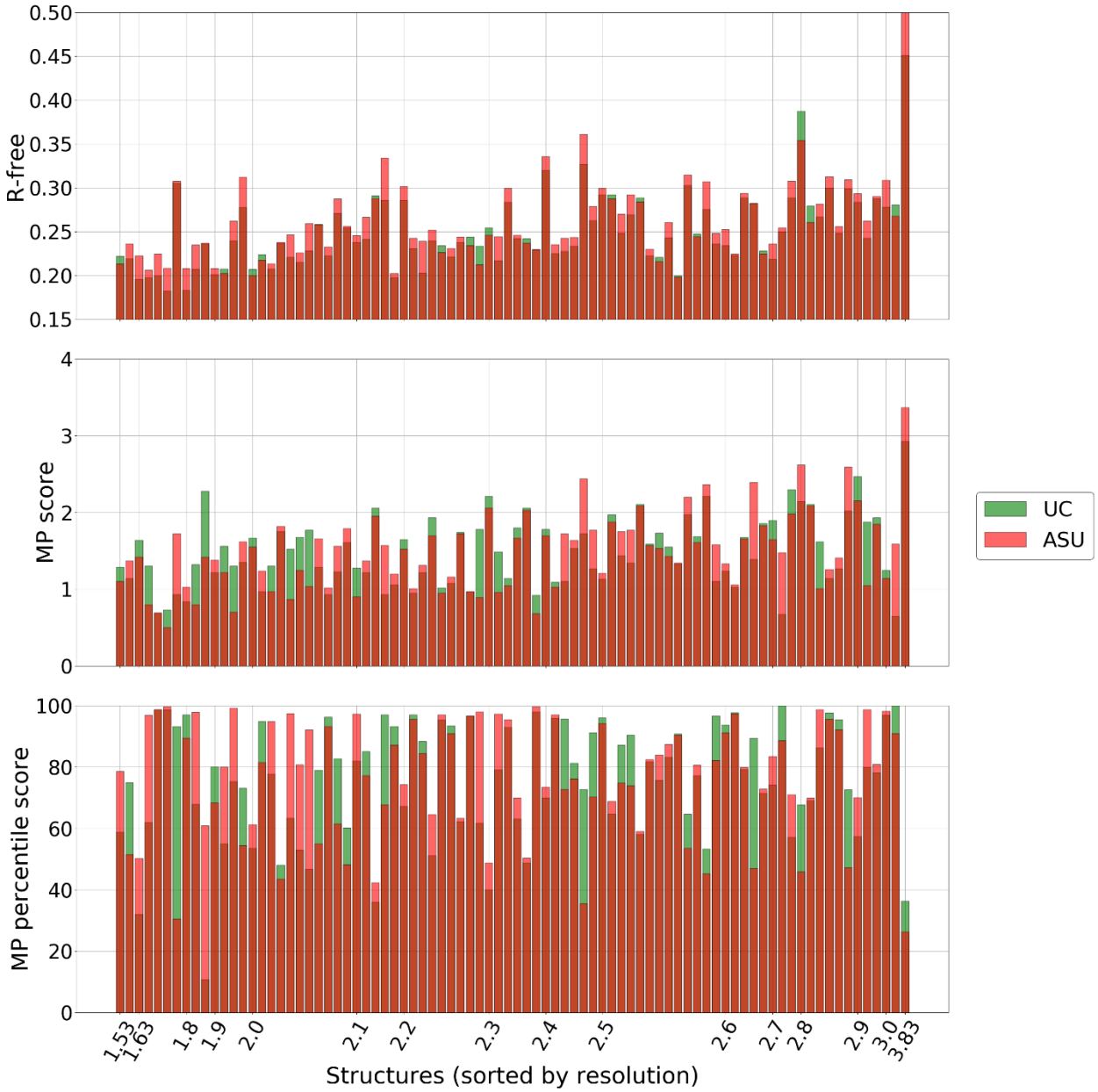

Figure 4.2. *Phenix.refine* performance using UC and ASU approaches on the deposited models. Green bars indicate the results of UC approach, red bars indicate the results of ASU approach.

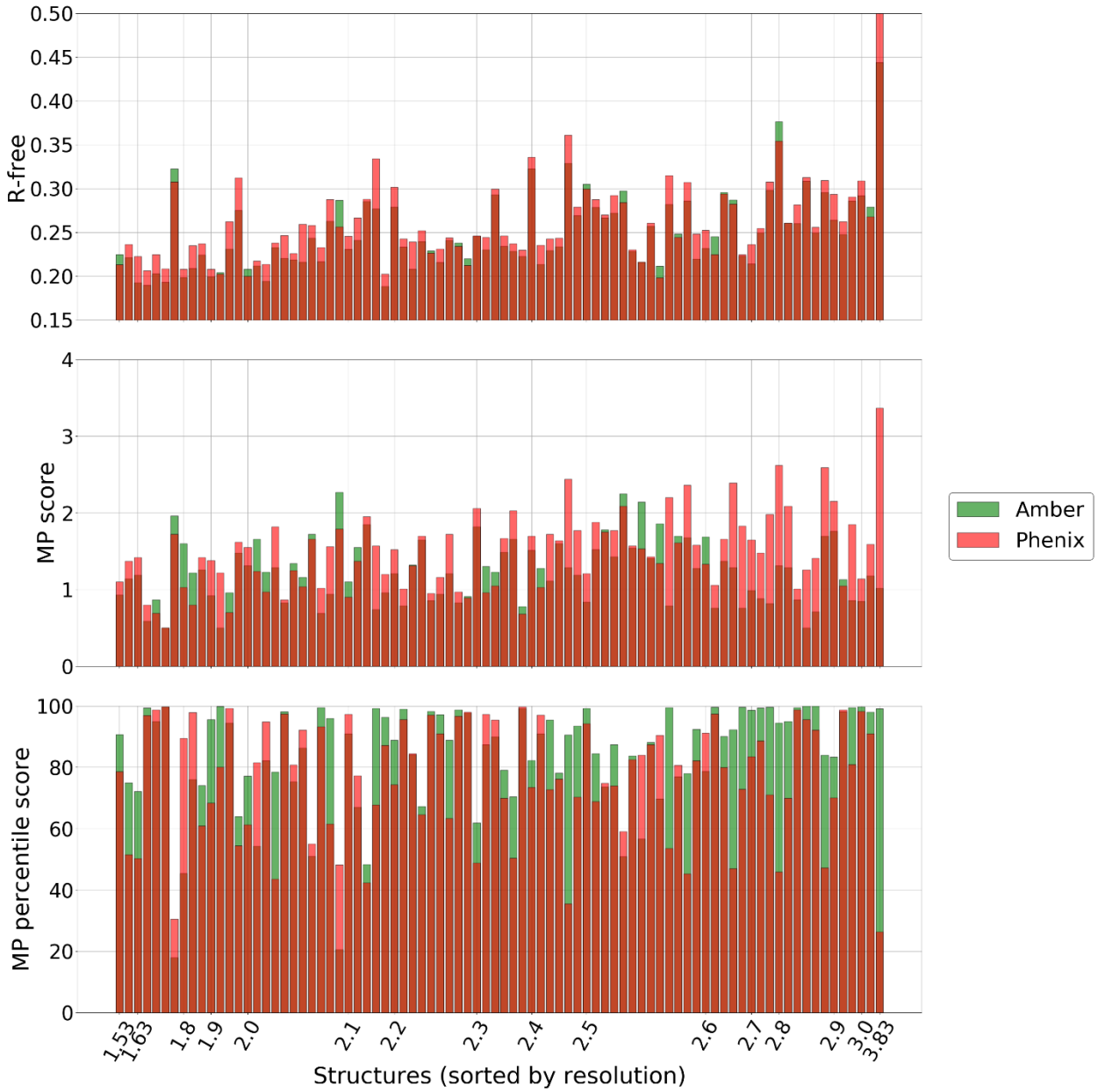

Figure 4.3. The plots show the absolute values of R_{free} factors, MolProbity scores and MolProbity percentiles of the re-refined deposited models. Green bars represent the results of our Amber-based setup. Red bars represent the results of Phenix-based ASU setups.

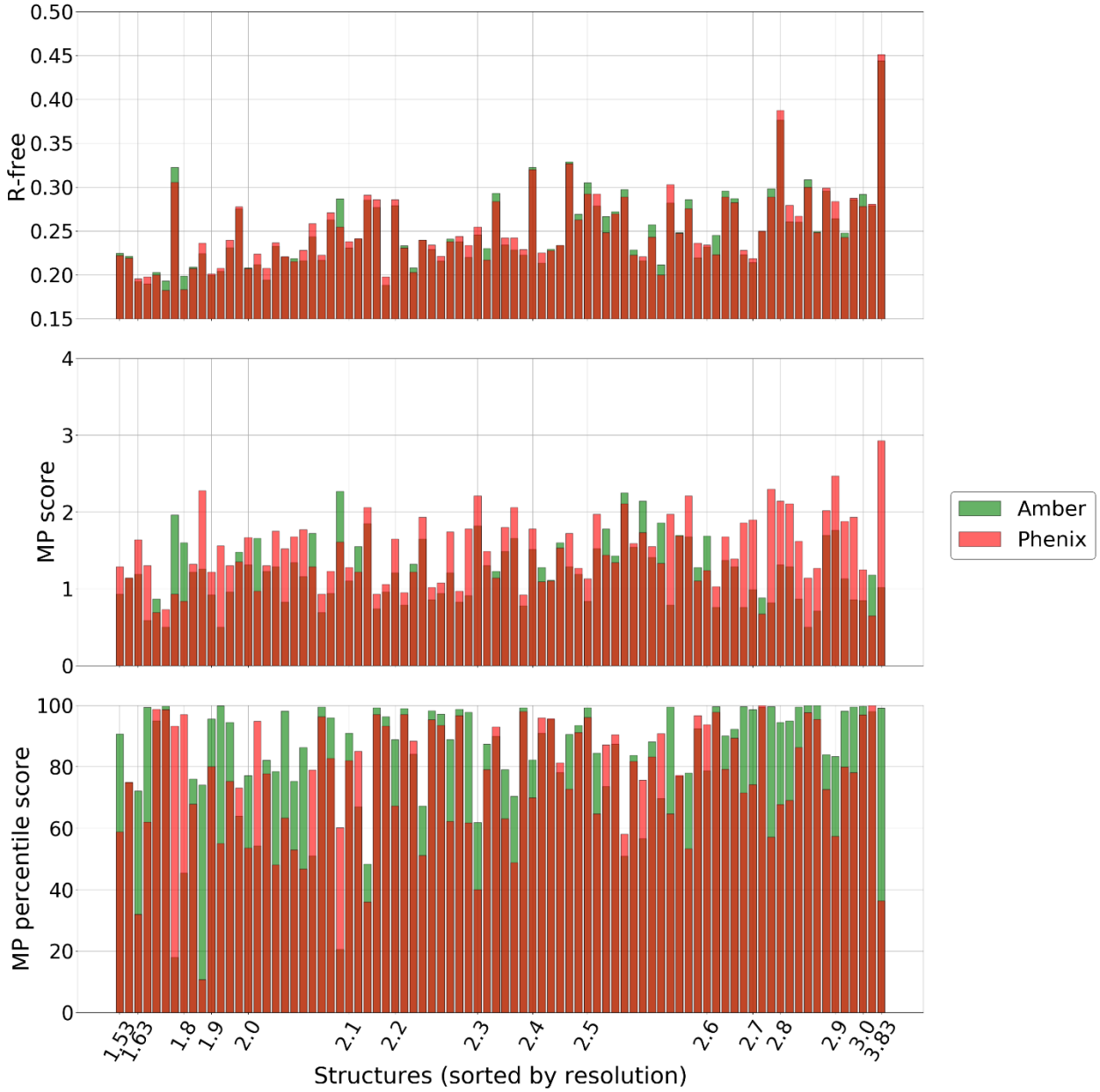

Figure 4.4. The plots show the absolute values of R_{free} factors, MolProbity scores and MolProbity percentiles of the re-refined deposited models. Green bars represent the results of our Amber-based setup. Red bars represent the results of Phenix-based UC setups.

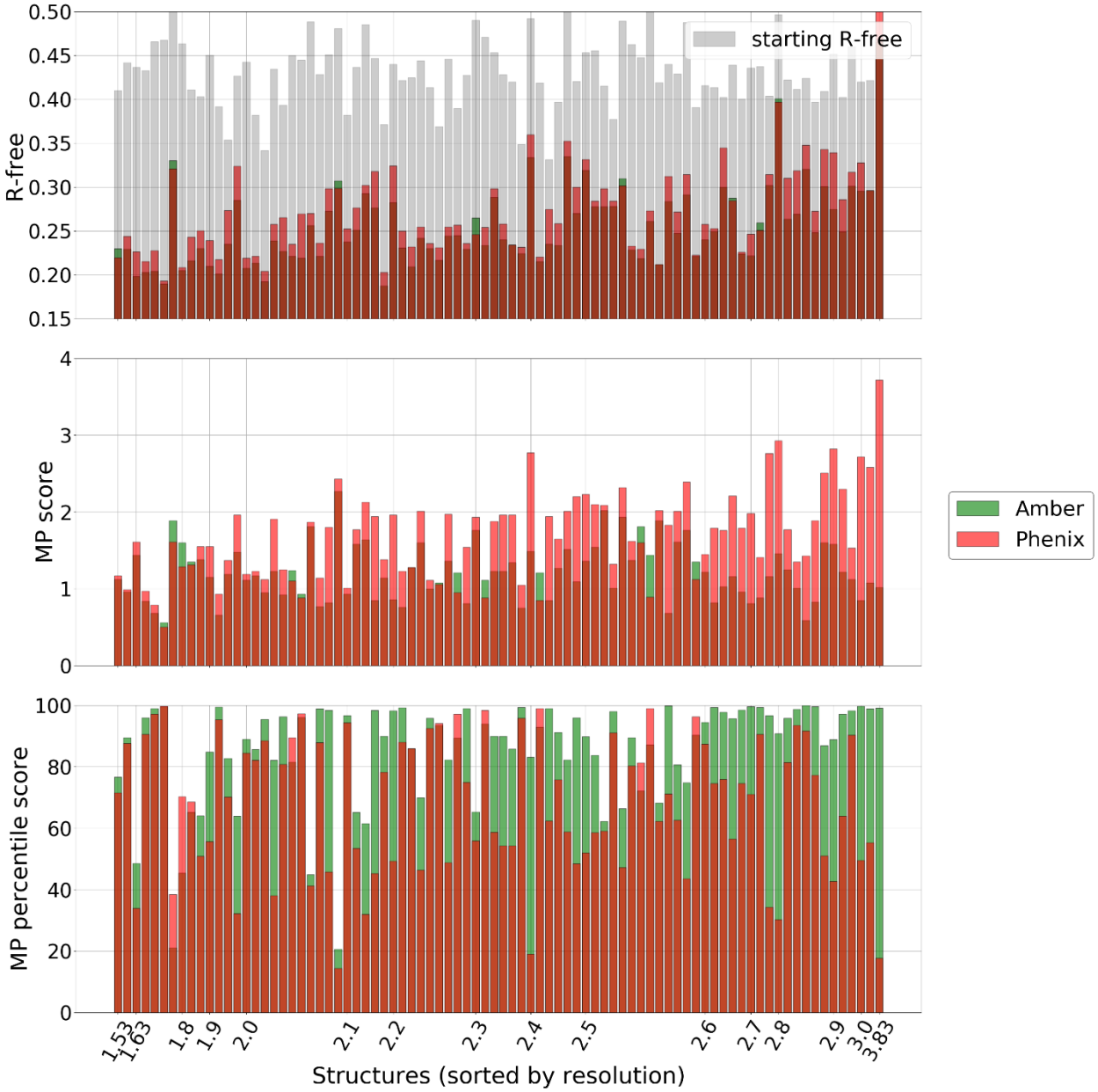

Figure 4.5. The plots show the absolute values of R_{free} factors, MolProbity scores and MolProbity percentiles of the refined MD1 models. Green bars represent the results of our Amber-based setup. Red bars represent the results of Phenix-based ASU setups.

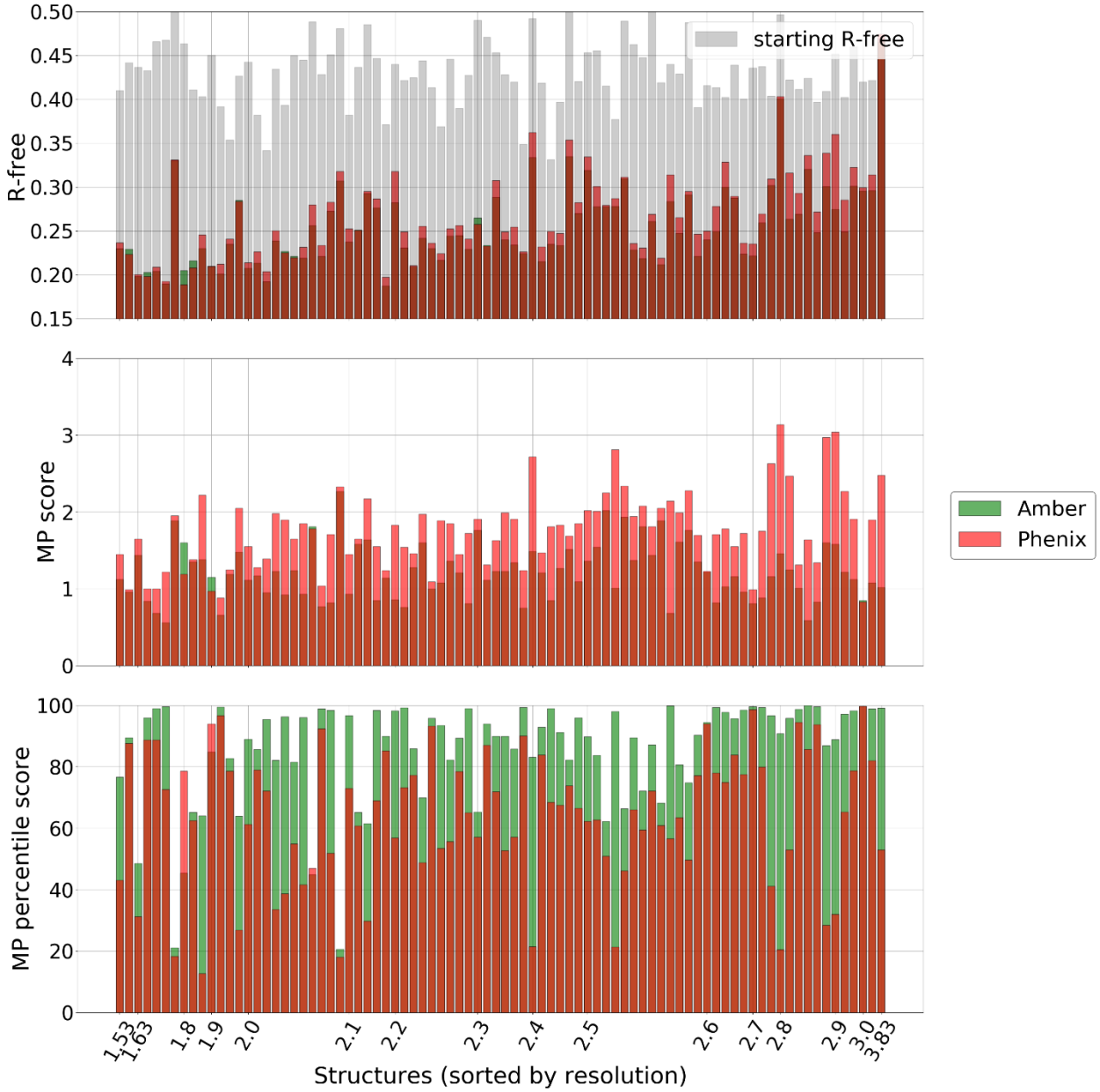

Figure 4.6. The plots show the absolute values of R_{free} factors, MolProbity scores and MolProbity percentiles of the refined MD1 models. Green bars represent the results of our Amber-based setup. Red bars represent the results of Phenix-based UC setups.

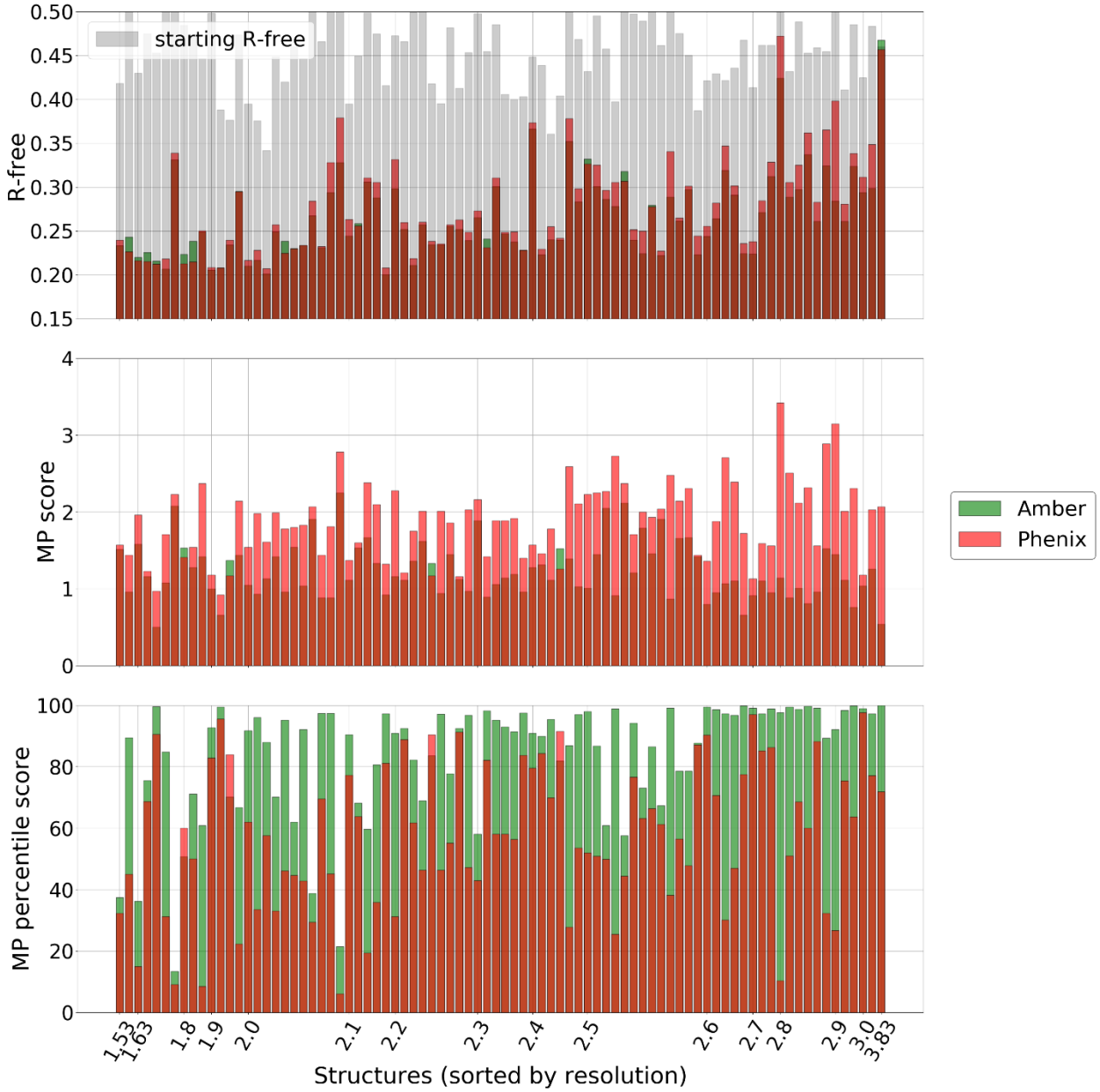

Figure 4.7. The plots show the absolute values of R_{free} factors, MolProbity scores and MolProbity percentiles of the refined MD2 models. Green bars represent the results of our Amber-based setup. Red bars represent the results of Phenix-based ASU setups.

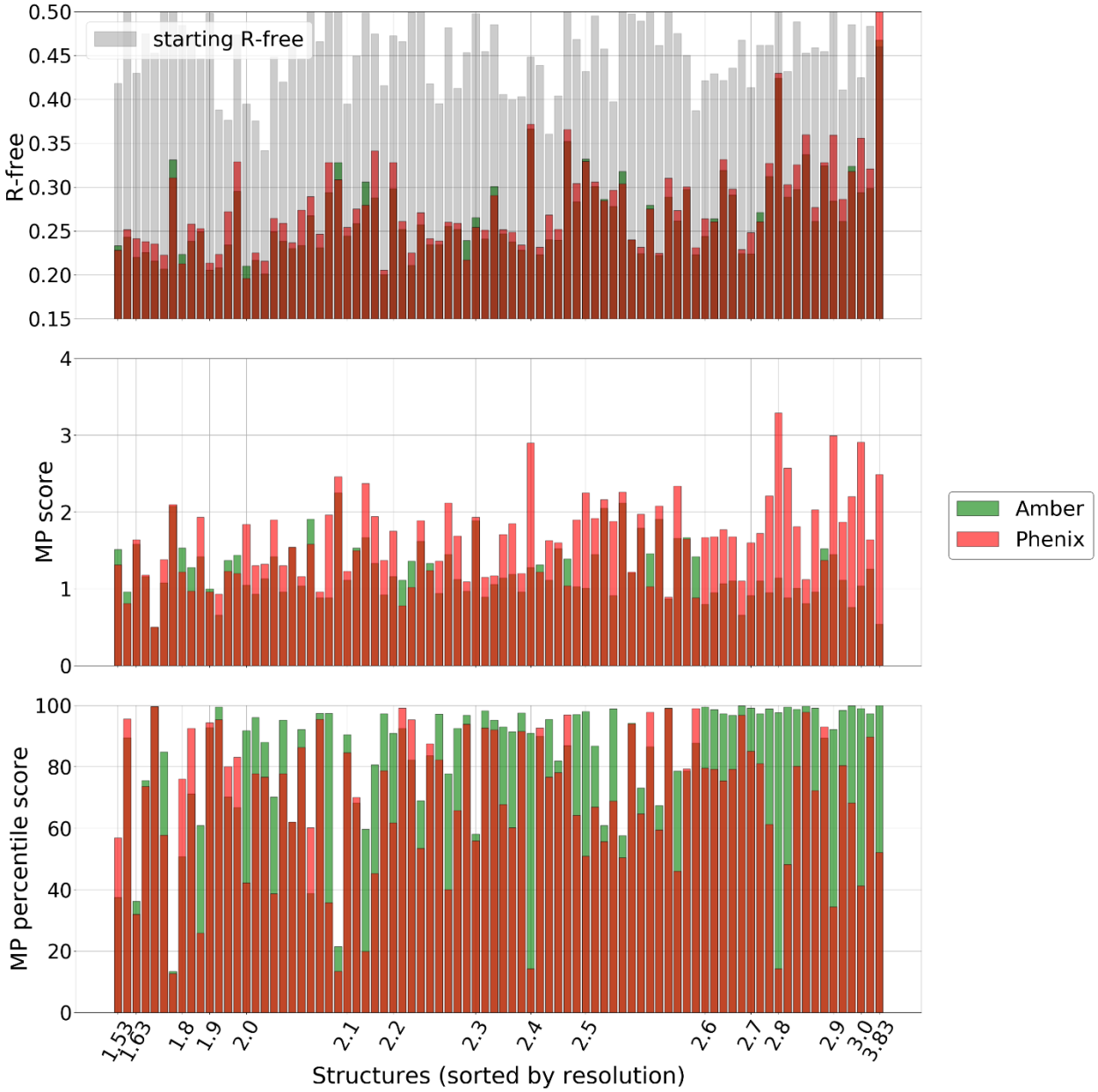

Figure 4.8. The plots show the absolute values of R_{free} factors, MolProbity scores and MolProbity percentiles of the refined MD2 models. Green bars represent the results of our Amber-based setup. Red bars represent the results of Phenix-based UC setups.

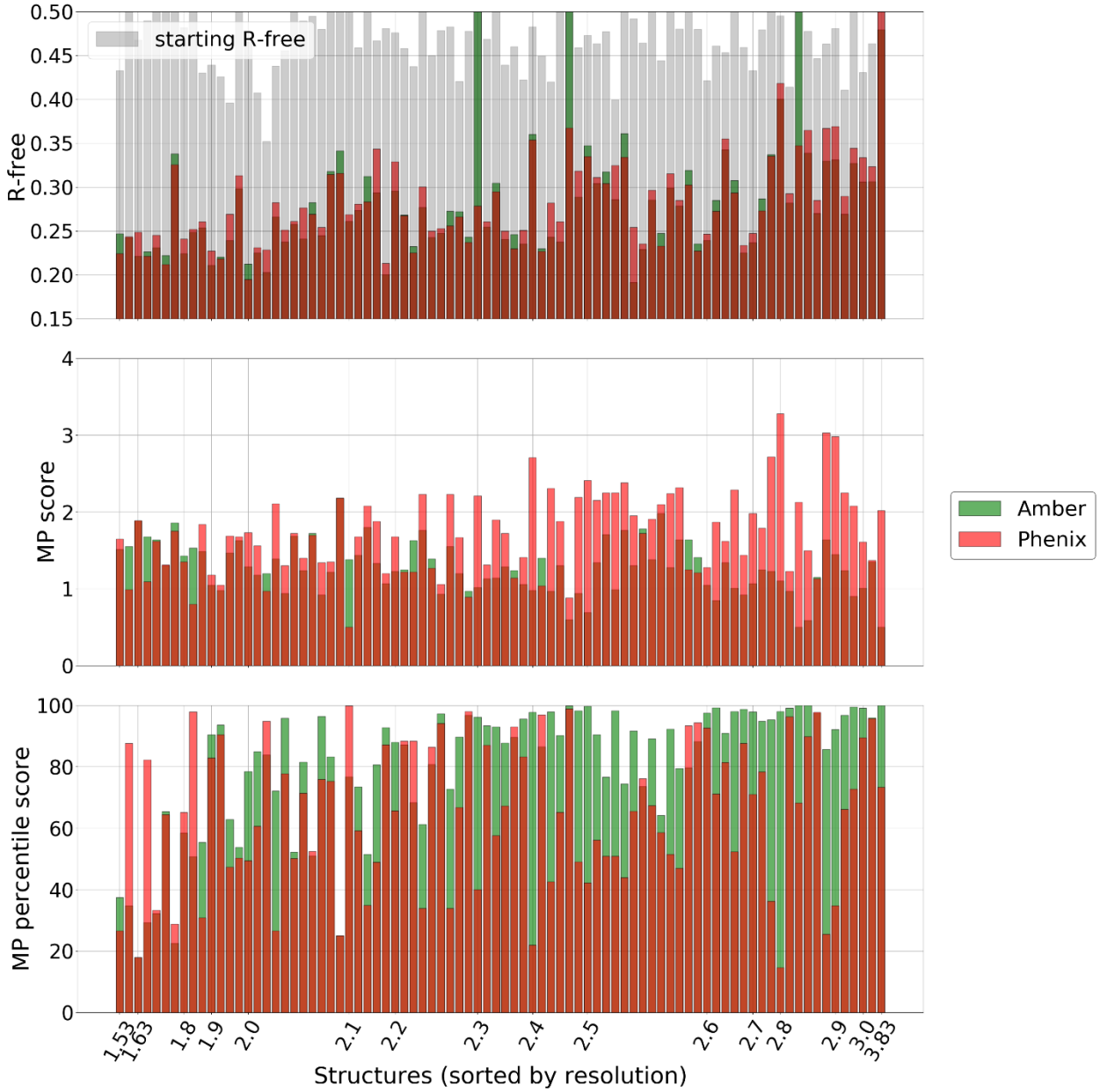

Figure 4.9. The plots show the absolute values of R_{free} factors, MolProbity scores and MolProbity percentiles of the refined MD3 models. Green bars represent the results of our Amber-based setup. Red bars represent the results of Phenix-based ASU setups.

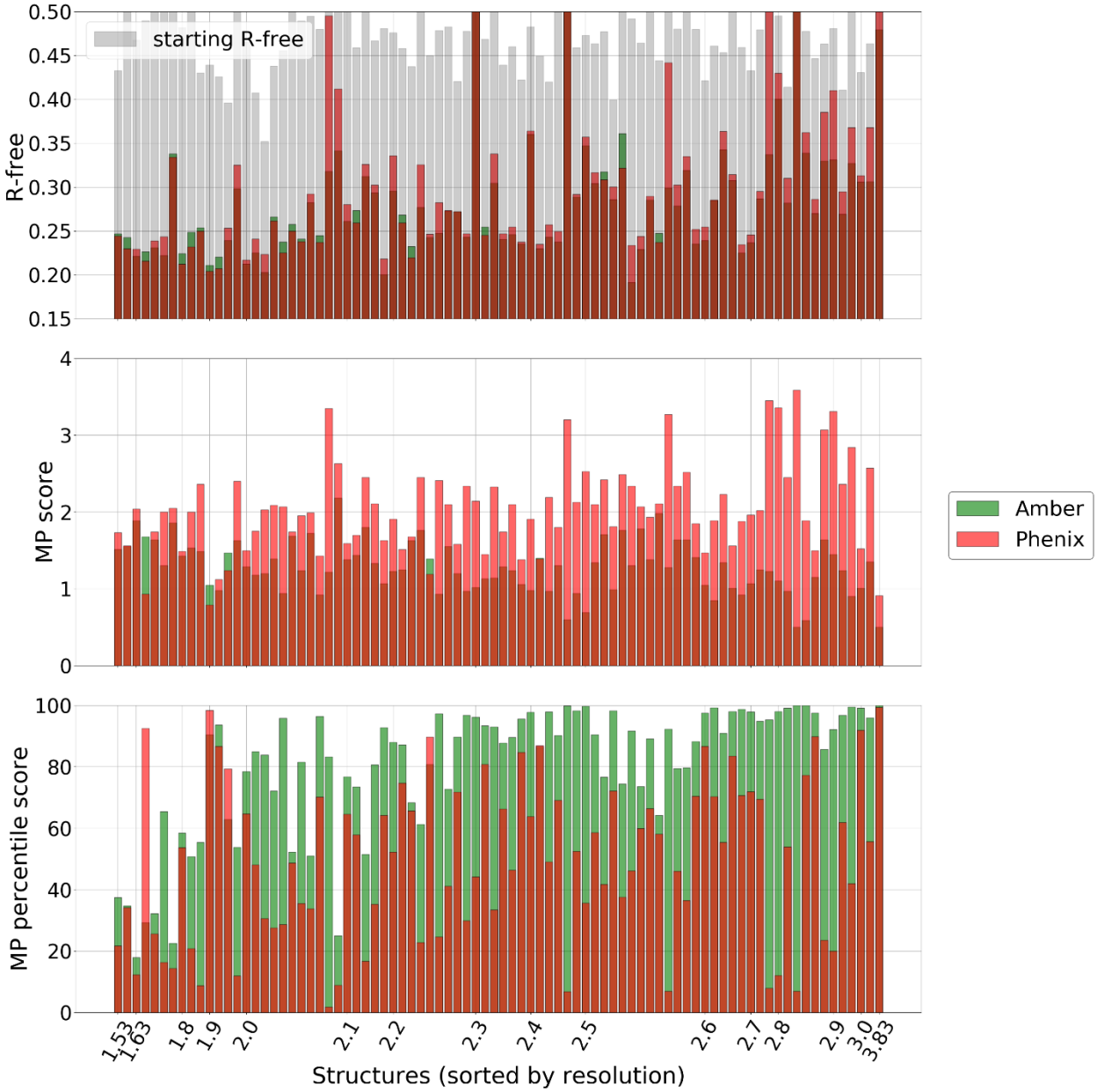

Figure 4.10. The plots show the absolute values of R_{free} factors, MolProbity scores and MolProbity percentiles of the refined MD3 models. Green bars represent the results of our Amber-based setup. Red bars represent the results of Phenix-based UC setups.

PUBLICATION

nature COMMUNICATIONS

ARTICLE

Received 3 Mar 2015 | Accepted 13 Aug 2015 | Published 5 Oct 2015

DOI: 10.1038/ncomms9361 OPEN

Observing the overall rocking motion of a protein in a crystal

Peixiang Ma[1,2,3,*,†], Yi Xue[4,*], Nicolas Coquelle[1,2,3], Jens D. Haller[1,2,3], Tairan Yuwen[4], Isabel Ayala[1,2,3], Oleg Mikhailovskii[4], Dieter Willbold[2,5,6], Jacques-Philippe Colletier[1,2,3], Nikolai R. Skrynnikov[4,7] & Paul Schanda[1,2,3]

The large majority of three-dimensional structures of biological macromolecules have been determined by X-ray diffraction of crystalline samples. High-resolution structure determination crucially depends on the homogeneity of the protein crystal. Overall 'rocking' motion of molecules in the crystal is expected to influence diffraction quality, and such motion may therefore affect the process of solving crystal structures. Yet, so far overall molecular motion has not directly been observed in protein crystals, and the timescale of such dynamics remains unclear. Here we use solid-state NMR, X-ray diffraction methods and μs-long molecular dynamics simulations to directly characterize the rigid-body motion of a protein in different crystal forms. For ubiquitin crystals investigated in this study we determine the range of possible correlation times of rocking motion, 0.1–100 μs. The amplitude of rocking varies from one crystal form to another and is correlated with the resolution obtainable in X-ray diffraction experiments.

[1] Université Grenoble Alpes, IBS, F-38044 Grenoble, France. [2] CEA, Institut de Biologie Structurale, F-38044 Grenoble, France. [3] CNRS, Institut de Biologie Structurale, F-38044 Grenoble, France. [4] Department of Chemistry, Purdue University, West Lafayette, Indiana 47907, USA. [5] Institut für Physikalische Biologie, Heinrich-Heine-Universität Düsseldorf, 40225 Düsseldorf, Germany. [6] ICS-6: Structural Biochemistry, Forschungszentrum Jülich, 52425 Jülich, Germany. [7] Laboratory of Biomolecular NMR, St. Petersburg State University, St. Petersburg 199034, Russia. * These authors contributed equally to this work. † Present address: Shanghai Institute for Advanced Immunochemical Studies (SIAIS), ShanghaiTech University, Shanghai 201210, China. Correspondence and requests for materials should be addressed to P.S. (email: paul.schanda@ibs.fr) or to N.R.S. (email: nikolai@purdue.edu) or to J.-P.C. (email: colletier@ibs.fr).

X-ray crystallography is the quintessential method for macromolecular structure determination. The method provides atomic coordinates along with atomic displacement parameters, which are generally expressed as B-factors and reflect the coordinate uncertainty around the mean positions. The coordinate precision in X-ray structures is limited by several factors, including model errors and invalid restraints[1]. The precision is also adversely affected by protein dynamics and static disorder, which together contribute to the 'blurring' of electron density maps. Motion has therefore long been treated as a nuisance limiting the effective resolution at which a crystallographic structure can be solved. Recent methodological advances have shown, however, that useful dynamical information can be extracted from X-ray diffraction (XRD) data[2–10], provided that high-resolution structural information is available. Several investigators pointed out the importance of rigid-body motions, which limit the achievable resolution in XRD experiments[4–9].

Overall motion is routinely modelled from XRD data using translation-libration-screw (TLS) analyses. However, refined TLS parameters offer only a simplified view of rotational and translational dynamics in the crystal lattice, meaning that some ambiguity remains regarding the physical nature of the modelled motion. Furthermore, diffraction data cannot provide insights into the timescale of motions, making it difficult to distinguish between static disorder and molecular motions. In other words, it is not possible to ascertain that the dynamics modeled from XRD data accurately reflect the overall motion of the molecules in the crystal.

Magic-angle spinning (MAS) NMR spectroscopy provides atomic-level-resolution access to crystalline proteins. MAS NMR is complementary to XRD in the sense that it can provide atom-specific insights into reorientational motions at a large number of sites. A number of NMR observables, in particular relaxation rate constants and dipolar couplings, probe exclusively the angular motion as sensed at each individual site while being unaffected by static disorder. Furthermore, NMR measurements can provide direct access to the timescale at which dynamics occur. It has been hypothesized before that rocking motion in crystals might be observable through spin relaxation parameters in MAS NMR[11], yet no experimental evidence has to date been produced. Rotational diffusion and its effects have been investigated for membrane proteins embedded in lipid bilayers[12–15], but reorientational fluctuations in protein crystals remain largely unexplored.

Here we report on the combined use of MAS NMR, XRD and microsecond-long molecular dynamics (MD) simulations of explicit crystal lattices to characterize the overall rocking motion and the local internal dynamics of the protein ubiquitin in three different crystal forms. Our results provide direct insight into the amplitudes and timescales of rocking motion in the three crystals. They illuminate the possibly general relationship that exists between crystalline rocking motions and the experimental resolution achieved in XRD and MAS NMR experiments.

Results

MAS NMR and XRD of three different ubiquitin crystals. Disentangling overall rigid-body motion (herein referred to as 'rocking' motion) from internal dynamics is a challenge, regardless of whether XRD or MAS NMR is used as an experimental tool. This is because both types of motion contribute to the dynamics-related observables, that is, to B-factors in XRD and to relaxation and dipolar-coupling parameters in MAS NMR. In the present study, these complications were circumvented by using different crystal forms of the same protein, allowing us to assume that the internal dynamics are similar—an assumption that we verify below—and thus to focus on differences in overall motion of the protein in the crystal lattices.

We prepared three different crystal forms of the 8-kDa globular protein ubiquitin. These crystals are henceforth referred to as MPD-ub, cubic-PEG-ub and rod-PEG-ub, reflecting the different precipitation agents (methyl-pentanediol (MPD) and polyethylene glycol (PEG), respectively) and the morphology of the crystals. Structures for the three crystal forms have been solved before and correspond to Protein Data Bank entries 3ONS (ref. 16), 3N30 (ref. 17) and 3EHV (ref. 18), respectively. To ensure that our crystals were consistent with the previously reported structures, XRD data were collected on the three crystals. For the two types of PEG crystals, we collected diffraction data at 100 K and solved the structures by molecular replacement, confirming the identity to the two already reported sets of coordinates. Our MPD-ub crystals appeared too thin for conventional structure determination when crystallized under the conditions that yield high-quality MAS NMR spectra. Nevertheless, a powder pattern obtained by rotating a scoop of MPD-ub crystals into the X-ray beam yielded a distribution of Bragg peaks similar to that calculated from the previously deposited structure (see Methods section). Thus, our crystals display the same space group as crystals previously obtained in the same crystallization conditions.

We used MAS NMR to further study the three crystal forms and obtain information about their dynamics. Figure 1 shows MAS NMR ^{1}H–^{15}N correlation spectra recorded on the three crystal forms. A first interesting observation concerns the number of peaks found in the three spectra. In MPD-ub, which has been extensively characterized before[19–21], one set of

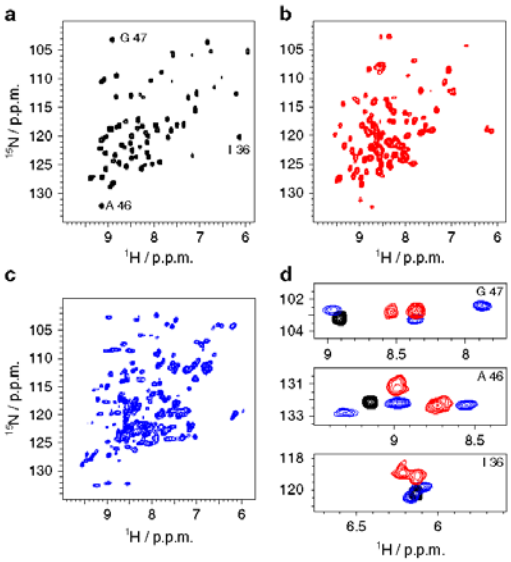

Figure 1 | High-resolution solid-state NMR spectra of three different crystal forms of ubiquitin. ^{1}H–^{15}N NMR spectra of MPD-ub, cubic-PEG-ub and rod-PEG-ub are shown in **a**–**c**, respectively. (**d**) Three regions of the spectra with well-isolated peaks, showing the different peak multiplicity observed in the different crystals (the residue numbers are indicated in each subpanel). A set of assigned HN and NCA spectra as well as methyl H-C spectra are shown as Supplementary Figs 1, 2 and 3, respectively.

well-resolved ^{1}H–^{15}N cross-peaks is observed. In cubic-PEG-ub many residues give rise to two peaks, as exemplified in Fig. 1d. In rod-PEG-ub we find—for several instances of well-isolated regions of the spectrum—three peaks per residue. This peak multiplicity is in good agreement with the number of non-equivalent molecules in the asymmetric unit of the crystals, i.e. one (MPD-ub), two (cubic-PEG-ub) and three (rod-PEG-ub), respectively. Of note, similar peak duplication has been reported previously in NMR spectra of ubiquitin crystals (prepared under slightly different conditions and resulting in different NMR spectra) and polymorphs of GB1 crystals[22–25]. We obtained residue-specific assignments of a majority of HN resonances in cubic-PEG-ub, using a set of ^{1}H- and ^{13}C-detected three-dimensional correlation spectra (assignments are reported in Supplementary Table 1). Owing to the higher spectral complexity arising from the three non-equivalent molecules, we did not assign the spectra of rod-PEG-ub.

Internal dynamics in different crystals from MAS NMR and MD. We conducted $^{1}H^{N}$-detected ssNMR experiments on highly deuterated protein samples to study dynamics in MPD-ub and cubic-PEG-ub. In what follows, we rely on three different experimental observables that concurrently probe a wide range of timescales at each amide site in the protein and are informative of both amplitudes and timescales of the dynamics. The first parameter, ^{1}H–^{15}N dipolar-coupling derived squared order parameter S^2, report on the amplitude of motion of HN bond vectors. The value of S^2 can range from 1 for a completely rigid bond to 0 for fully dynamically disordered peptide planes. The dipolar-coupling derived order parameters reflect the net effect from all reorientational motions occurring on timescales shorter than about 100 μs. The second parameter, the ^{15}N R_1 spin relaxation rate constant, is sensitive to both the amplitude and the timescale of ^{1}H–^{15}N bond vector motions. This relaxation parameter is particularly sensitive to dynamics on timescales from tens of picoseconds to ~100 nanoseconds (Supplementary Fig. 4). The third parameter, the ^{15}N $R_{1\rho}$ spin relaxation rate constant, is also sensitive to both the amplitude and timescale of the motion, but mainly to slower motion, occurring on the ns–μs timescale (see Supplementary Fig. 5 and discussion below). Analysing these three experimental observables therefore provides good insight into motional properties of individual protein residues over a wide range of timescales.

Figure 2a–d shows a comparison of site-specific amide ^{15}N R_1 rate constants and NH order parameters in MPD-ub and cubic-PEG-ub, obtained at 300 K sample temperature. These data reveal that the local dynamics in the two crystal forms are generally similar, with few differences. Overall, residues located in secondary structure elements have high order parameters S^2 and low R_1 relaxation rate constants, indicating that these residues are motionally restricted in both crystal forms. Previous studies of MPD-ub showed that low-amplitude motions in the secondary-structure elements occur primarily on the picosecond timescale[20]. Certain details of local dynamics are reproduced in both crystals. For example, an alternating pattern of low/high motional amplitudes in strand β2 is observed in both MPD-ub and cubic-PEG-ub (residues T12–V17, dashed outline in Fig. 2). This pattern arises from alternation of amides which are hydrogen bonded or otherwise exposed to solvent[26]. Similarities between the two crystals are also found in several loop regions, such as the α1–β3 loop and the β3–β4 loop, which show similarly increased flexibility (as reflected in the increased R_1 and decreased S^2 values). Yet, distinct differences in dynamic behaviour are observed at certain sites, as evident from Fig. 2a,b. For example, high R_1, low S^2 and high $R_{1\rho}$ (see further below, Fig. 3) values in the β1–β2 loop in MPD-ub are indicative of extensive ns-timescale motion. In contrast, this loop appears rigid in cubic-PEG-ub, displaying similar dynamics to residues in the secondary-structure regions. Another prominent example is residue Q62 located in the α2–β5 loop, which displays significant flexibility in cubic-PEG-ub but seems relatively stiff in MPD-ub. It is also worth noting that the order parameters in MPD-ub are overall slightly higher than in cubic-PEG-ub. When applying an overall scaling factor of 1.04 to the S^2 values from cubic-PEG-ub, the agreement with MPD-ub data is significantly improved (see Supplementary Fig. 6 for details). As discussed further below, this offset can be explained by the rocking motion of ubiquitin within the crystal lattice of cubic-PEG-ub.

It has been recently shown that experimental data by MAS NMR and XRD can be successfully reproduced using explicit MD models of protein crystals[27–29]. Towards this goal we have recorded 1-μs-long all-atom MD trajectories representing the two different crystal lattice arrangements of ubiquitin. A block of four crystal unit cells (24 ubiquitin molecules) was simulated for MPD-ub, while one crystal unit cell (48 ubiquitin molecules) was simulated for cubic-PEG-ub. The presence of multiple protein molecules in the simulations effectively improves the statistical properties of the MD models. The results from MD simulations, Fig. 2e–h, nicely reproduce the experimentally observed trends. Consistent with the experimental data, simulated ^{15}N R_1 and S^2 parameters are overall similar in the two crystals, with two notable exceptions found in the β1–β2 loop and residue Q62. On average, the simulated S^2 in cubic-PEG-ub are slightly lower than those in MPD-ub, which is again consistent with the experimental observations.

For the two crystal forms at hand, NMR and MD produce similar R_1 profiles (sensitive primarily to motions on a timescale of tens of picoseconds to ~100 nanoseconds) and S^2 profiles (sensitive to all motions faster than ca. 100 μs). This leads us to suggest that internal dynamics of ubiquitin are similar in the two crystals. Furthermore, site-specific S^2 data in crystals are remarkably similar to those in solution, as confirmed by experimental measurements as well as MD simulations (Fig. 2d,h). These observations are in line with the results from previous studies, which suggested that the crystalline environment has only comparatively minor effect on protein internal dynamics[30–37].

Evidence for overall rocking motion from MAS NMR and MD. Having established that internal motions on ps–ns timescales are generally similar in the two crystals, we then focused on amide-^{15}N $R_{1\rho}$ spin relaxation rate constants. This relaxation parameter is highly sensitive to amplitudes and time constants of reorientational motions occurring on longer timescales—specifically nanosecond to microsecond motions (Supplementary Fig. 5). The experimental $R_{1\rho}$ relaxation rate constants in MPD-ub and cubic-PEG-ub are summarized in Fig. 3a. Interestingly, a clear-cut difference is observed between the two crystal forms. In particular, the 'base' level of $R_{1\rho}$ within secondary structure regions is significantly higher in cubic-PEG-ub ($12\,s^{-1}$) than in MPD-ub ($3.5\,s^{-1}$). To a reasonable approximation this offset is uniform across the sequence, at least for secondary-structure elements. Site-specific differences in $R_{1\rho}$ rates are found mostly in loops, and can be ascribed to nanosecond mobility of these regions[20,26]; differences in loop dynamics have been exposed already by the R_1 and order parameter data discussed above.

The overall offset in the 'base' $R_{1\rho}$ rates of the two crystals points to a global motion that involves the entire molecule. This motion appears to be present in cubic-PEG-ub crystals, but absent or less pronounced in MPD-ub crystals. We attribute this

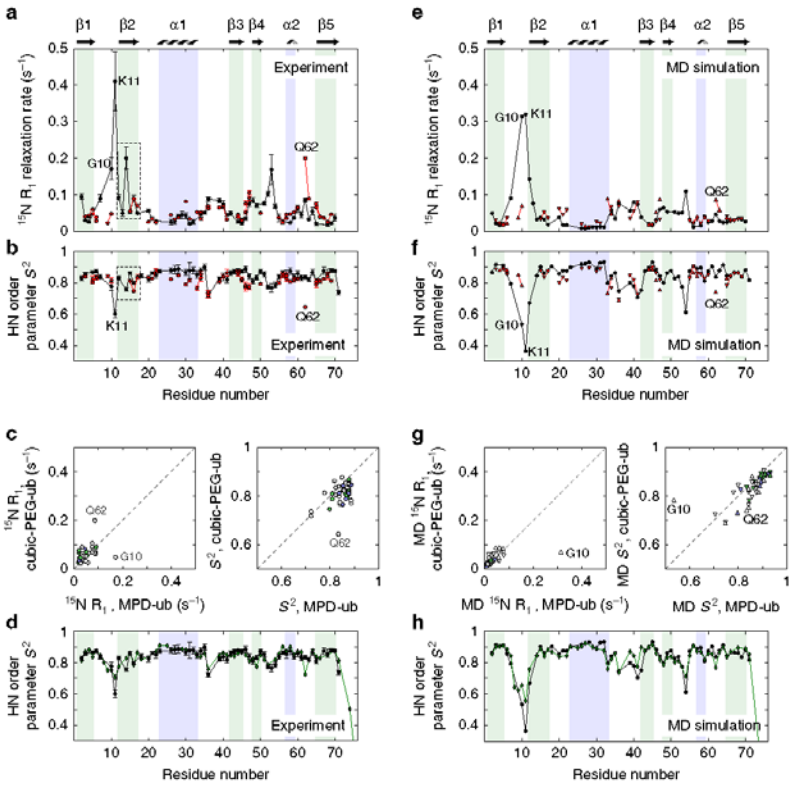

Figure 2 | Site-resolved HN dynamics parameters in two different crystal forms from NMR experiments and MD simulations. Per-residue dynamics data obtained from MPD-ub (black), cubic-PEG-ub (red) and ubiquitin in solution (green) as observed by NMR experiments (**a**–**d**) and MD simulations (**e**–**h**). (**a**) Experimental ^{15}N R_1 rate constants and (**b**) dipolar-coupling derived squared order parameters, S^2. In cases where two data points per residue could be obtained in cubic-PEG-ub, corresponding to the pair of non-equivalent molecules, these are represented by two distinct symbols. Because of the spectral overlaps in spectra of cubic-PEG-ub, it was not possible to unambiguously assign all signals to chain A or B; those data points that have been identified as belonging to the same chain are connected by a solid line. Secondary-structure regions are indicated by the shaded bands and identified above the plot. (**c**) Correlations between the data from two different crystal forms; symbols are coloured according to the secondary-structure classification (α-helix in blue and β-strands in light green). (**d**) Experimental S^2 values measured in MPD-ub crystals (black) juxtaposed on S^2 values from solution-state measurements (green, ref. 57). Supplementary Table 2 lists experimental data for cubic-PEG-ub. Data for MPD-ub have been reported elsewhere[20,26]. Data in **e**–**h** are from MD simulations, plotted using the same template and colouring conventions as in the case of the experimental data (**a**–**d**). The data points from chains A and B in cubic-PEG-ub simulation are plotted with downward- and upward-pointing red triangles, respectively. Supplementary Tables 3–5 list the simulated parameters for MPD-ub, cubic-PEG-ub and ubiquitin in solution, respectively.

effect to relatively slow reorientational fluctuations of the protein molecule embedded in the crystal lattice, that is, to rocking motion. In what follows, we will show that the observed $R_{1\rho}$ offset in cubic-PEG-ub is consistent with a rocking motion having an amplitude of several degrees and a correlation time in the range from hundreds of nanoseconds to tens of microseconds.

To obtain additional insight into rocking motion, we analysed the 1-μs-long MD trajectories of the three crystals (MPD-ub and cubic-PEG-ub, as described previously, as well as rod-PEG-ub). For each trajectory we defined a set of reference coordinates, that is, a block of crystal unit cells constructed from the corresponding crystallographic structures. We further calculated rotation matrices Ξ connecting instantaneous MD coordinates of protein molecules with their respective reference coordinates (Ξ were obtained from least-square fitting of the Cα atoms belonging to the protein secondary structure). A sequence of these small-angle rotation matrices encodes the rocking motion of each individual ubiquitin molecule. Finally, matrices Ξ have been applied to a set of 100 dipolar vectors uniformly distributed on a unit sphere so as to calculate 'isotropic' rocking correlation functions $g_{rock}(\tau)$. The results are shown in Fig. 4 for all individual ubiquitin molecules from MPD-ub, cubic-PEG-ub and rod-PEG-ub simulations. Supplementary Movies 1–3 illustrate rocking motion in MPD-ub, cubic-PEG-ub (chain A) and cubic-PEG-ub (chain B), respectively.

Clearly, the rocking motion found in the MD simulation of cubic-PEG-ub (order parameters 0.982 and 0.957 for chains A and B, respectively) is much more pronounced than for MPD-ub and rod-PEG-ub (average order parameter 0.995 for both systems). This result correlates well with our experimental data

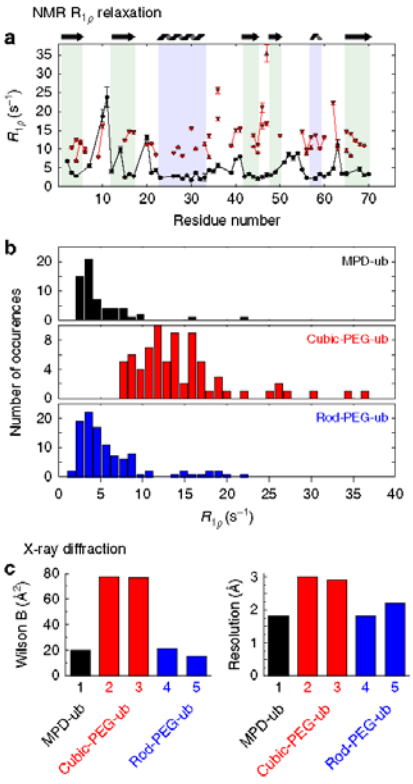

Figure 3 | Evidence for rigid-body motion (rocking) in ubiquitin crystals from NMR and XRD data. (**a**) Residue-wise ^{15}N $R_{1\rho}$ spin relaxation rate constants in MPD-ub (black) and cubic-PEG-ub (red). (**b**) Histograms of per-residue ^{15}N $R_{1\rho}$ relaxation rate constants in the above two crystals, as well as rod-PEG-ub (blue). (**c**) XRD data pointing to different motional behaviour of ubiquitin in the three crystals: Wilson B-factors (left) and structural resolution (right). Shown are the data from the following five PDB structures: 1, 3ONS (ref. 16); 2, 3N30 (ref. 17); 3, 4XOL (this study); 4, 3EHV (ref. 18); 5, 4XOK (this study).

that offer multiple lines of evidence for increased rocking motion in cubic-PEG-ub. The MD simulations also have a potential to shed light on the timescale of rocking dynamics. The simulated correlation functions $g_{rock}(\tau)$ shown in Fig. 4 involve a small-amplitude fast component with the correlation time $\tau_f \sim 1$ ns and the more prominent slow component with τ_s in the range from ~ 0.1 to 1 μs.

It is important to bear in mind, however, that MD simulations offer, at best, a qualitative insight into rocking motions. The effect of crystal packing in protein crystals is governed by a multitude of subtle interactions that involve, in particular, mobile side chains and hydration water. Capturing these interactions in the context of MD modelling remains a challenge even for state-of-the-art force fields. As a consequence, the crystal lattice undergoes slight but progressive distortion during the course of the simulation[38]. Of note, such 'structural drift' has also been observed in MD simulations of globular proteins, even though the determinants of protein structure (for example, amide hydrogen bonds) are generally far better understood than the determinants of crystal packing[39]. This leads to a situation where rocking motion in the MD simulations occurs against the background of gradually deteriorating crystal lattice.

One should also be aware of statistical limitations. Even though each of our 1-μs-long trajectories contains from 24 to 48 ubiquitin molecules, which improves their statistical properties, this would not be sufficient to capture rocking dynamics should it occur on a timescale approaching 100 μs. Note that in this situation it can be difficult to differentiate between 'structural drift' (discussed above) and lack of convergence. The limitations of the MD model can be appreciated from Fig. 4 where one observes a significant spread in the rocking correlation functions belonging to the individual ubiquitin molecules, including a number of outliers (green curves). Under these circumstances it is impossible to meaningfully estimate the anisotropy of rocking motion, although in general rocking is certainly expected to be anisotropic. For further insight into convergence properties of $g_{rock}(\tau)$ see Supplementary Fig. 7.

Finally, one should bear in mind that no attempt has been made to include into MD simulations the crystallization additives, such as 2-methyl-2,4-pentanediol or PEG. These compounds do not appear in the crystallographic coordinates and it is unclear to what degree they are partitioned into the crystals. We also did not include the Zn^{2+} ions, although they are explicitly present in the X-ray structures of cubic-PEG-ub and rod-PEG-ub. There are currently no force field parameters that would be suitable to model Zn^{2+} ions in highly diverse and conformationally

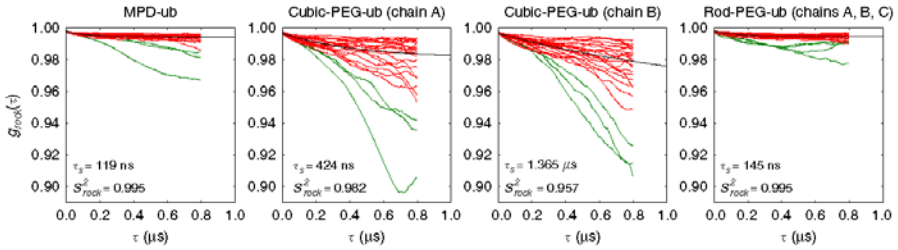

Figure 4 | Rocking correlation functions from three 1-μs-long MD trajectories of ubiquitin crystals. The curves, representing individual ubiquitin molecules in the crystals, were averaged and then fitted using a bi-exponential function with a flat base, $g^{fit}_{rock}(\tau) = c_f \exp(-\tau/\tau_f) + c_s \exp(-\tau/\tau_s) + S^2_{rock}$. The best-fit curve $g^{fit}_{rock}(\tau)$ is shown in the plot (black line), along with the values of the fitted parameters τ_s and S^2_{rock}. In the case of cubic-PEG-ub we have treated two inequivalent molecules, chains A and B, separately, whereas in the case of rod-PEG-ub the data from three inequivalent molecules, chains A, B and C, have been averaged before the fitting. Only red curves have been used in the fitting procedure (green curves have been classified as outliers and set aside).

dynamic adventitious binding sites at protein–protein interfaces. Fundamentally, no single set of force-field parameters would be sufficient in this situation[40–42].

Nevertheless, despite all these shortcomings, our MD simulations clearly reproduce the same trend as has been observed experimentally and thus confirm that MPD-ub and rod-PEG-ub form stable crystal arrangements, whereas cubic-PEG-ub is prone to rocking. Furthermore, the MD-derived correlation functions $g_{rock}^{fit}(\tau)$ can be used to calculate the contributions of rocking motion into $R_{1\rho}$ relaxation rate constants. These contributions turn out to be $0.6\,s^{-1}$ for MPD-ub, 9.1 and $63.4\,s^{-1}$ for cubic-PEG-ub (chains A and B, respectively) and $0.7\,s^{-1}$ for rod-PEG-ub. The difference between the first two numbers, $8.5\,s^{-1}$, reproduces quantitatively the difference between the experimentally measured $R_{1\rho}$ rates in MPD-ub (base rate $3.5\,s^{-1}$) and cubic-PEG-ub (base rate $12\,s^{-1}$). Although this result is certainly fortuitous, it demonstrates the potential for quantitative analysis of rocking dynamics using MD models (see Fig. 5 for further details).

In order to obtain better insight into the time scale of the rocking motion, we plot in Fig. 5 the calculated $R_{1\rho}$ relaxation rate constant as a function of the amplitude and time scale of the motion. The black curve shows the solutions (order parameters and correlation times) that are in agreement with the experimentally measured 'base' $R_{1\rho}$ rate in MPD-ub, while the purple curve shows the solutions for cubic-PEG-ub. Furthermore, the black and purple circles illustrate the results obtained from the two respective MD trajectories. If one takes guidance from the MD trajectory of cubic-PEG-ub, and specifically the results for chain A (purple circle in the plot), then one is led to believe that rocking motion is characterized by $S^2 \sim 0.985$, $\tau_s \sim 400$ ns. Indeed, such a scenario would be consistent with all of our existing experimental data (Fig. 5). However, as explained above, the MD simulations offer only qualitative insight into the problem and cannot be viewed in this case as a source of quantitative information. Therefore, we recognize that there is an alternative solution corresponding to the upper branch of the purple curve in Fig. 5: $S^2 \sim 0.985$, $\tau_s \sim 40\,\mu s$. Generally, we can safely conclude that rocking motion in cubic-PEG-ub occurs on the timescale from hundreds of nanoseconds to tens of microseconds. More accurate determination of this important parameter is deferred to future work.

The emerging picture is self-consistent in more ways than one. For instance, MD simulations predict that order parameters in the cubic-PEG-ub crystal should be ~2–3% lower than in MPD-ub due to the intensified rocking motion. This is compatible with our experimental data, which show that cubic-PEG-ub order parameters S^2 are ~4% lower than those in MPD-ub (see above and Supplementary Fig. 6). Furthermore, the MD model predicts the crystallographic B-factors in cubic-PEG-ub to be significantly higher than in MPD-ub, with rocking motion making an important contribution to B-factors in cubic-PEG-ub, but much less in MPD-ub (Supplementary Fig. 8). These predictions are also borne out by the experimental data, as explained below.

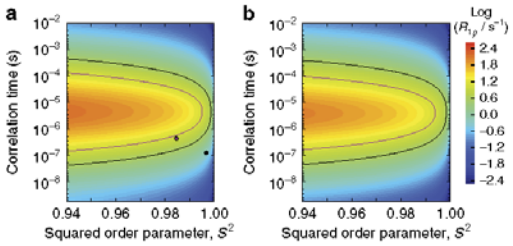

Figure 5 | Estimating the timescale of rocking motion from ^{15}N $R_{1\rho}$ measurements. Plotted is the ^{15}N $R_{1\rho}$ relaxation rate constant as a function of the order parameter S^2 and correlation time τ that describe the motion of the NH vector. (**a**) The calculations were conducted using the Redfield-theory formulas, equations 8 and 18 in ref. 65. (**b**) Alternatively, the calculations were conducted using a numeric model that is also valid outside the Redfield regime; the geometrical details of this two-site jump model are exactly as described in Fig. 2 of ref. 66, and the simulation was implemented in the program GAMMA[67], as described before[68]. The jump angle Φ used in the numerical simulation is related to the order parameter according to $S^2 = (1+3\cos^2\Phi)/4$. Both calculations **a** and **b** assume an MAS frequency of 39.5 kHz and a ^{15}N spin-lock radio-frequency field strength of 15 kHz, the same as in our experimental measurements. The results obtained from the two computational models prove to be similar, thus validating the Redfield-theory based approach for the problem at hand (see Supplemetary Fig. 5 for additional discussion). The black contour line represents the 'base' $R_{1\rho}$ relaxation rate constant as experimentally found in MPD-ub ($3.5\,s^{-1}$), whereas the purple line represents the 'base' rate in cubic-PEG-ub ($12\,s^{-1}$). The black circle represents the relaxation due to rocking motion as obtained from the MD trajectory of MPD-ub, while the purple circle represents the relaxation due to rocking motion in cubic-PEG-ub (chain A). These relaxation rate constants were calculated based on the respective correlation functions $g_{rock}^{fit}(\tau)$, see Fig. 4. In doing so, the small rapidly decaying component of the correlation function, $\tau_f \sim 1$ ns, has been ignored since it makes only negligible contribution to $R_{1\rho}$. Thus, for the purpose of calculating $R_{1\rho}$ we have made the identification $1 - S^2 = c_s$ and $\tau = \tau_s$, where c_s is the amplitude of the slow rocking motion and τ_s is the respective time constant. Note that the experimentally determined relaxation rate constants (black and purple contour lines) reflect both rocking motions and internal protein dynamics, whereas the calculated rates (black and purple circles) are limited to rocking alone.

Overall rocking impacts resolution in XRD experiments. Both the NMR and MD data indicate that ubiquitin molecules arranged in a crystal lattice experience varying degree of rocking motion at room temperature. But is this rocking motion impacting the XRD data collected at 100 K? Figure 3c shows that this is indeed the case. The Wilson B-factor in cubic-PEG-ub is almost fourfold higher than in MPD-ub and the resolution is significantly lower, which we propose to arise from differences in the respective rocking dynamics. This correlation between NMR ^{15}N $R_{1\rho}$ relaxation data and XRD resolution is further substantiated by the third crystal form, rod-PEG-ub, which displays lower ^{15}N $R_{1\rho}$ rates, suggesting that rocking motions are of low amplitude (blue bars in Fig. 3b). Correspondingly, these rod-PEG-ub crystals display a lower Wilson B, and they diffract to high resolution (blue bars in Fig. 3c).

Similar conclusions can also be reached if a TLS model is used to account for rigid-body motion of proteins in the crystals[9]. In XRD refinement, TLS modelling is one of the ways by which collective and local motions can be separated. As expected, cubic-PEG-ub shows the highest librational as well as translational amplitude among the three crystal structures (Supplementary Fig. 9), in good qualitative agreement with our NMR and MD data. At this stage, it should be reminded that the TLS model is based on certain simplifying assumptions. If a protein molecule experiences a series of small rotations with different pivot points (a likely scenario in the protein crystal lattice), the TLS model may interpret this dynamics as translation. In this sense, the information content of the TLS parameters is not very different from that of the Wilson B-factor insofar as it is difficult to disentangle libration and translation.

It is interesting to examine why the same molecule, with overall identical structure and internal dynamics, exhibits more rocking motion in one of the examined crystals than in others. A direct influence on rocking of the precipitating agent used for crystallization can be excluded on the basis that both cubic-PEG-ub and rod-PEG-ub crystals crystallize in essentially the same condition (sometimes even in the same crystallization drop). The amplitude of the rocking motion is likely to be influenced by the crystal packing density—increased contact surface area is generally expected to offer more resistance to rocking. In our case, the packing density is indeed lowest for the crystal with the most pronounced rocking motion, with solvent content V_s of 58% for cubic-PEG-ub, 49% for MPD-ub and 40% for rod-PEG-ub, respectively. These values follow the expected trend—lower packing density allows for more overall motion. However, given the small size of this data set, the correspondence of rocking motion and packing density may as well be fortuitous. We thus performed a wider analysis seeking to determine whether there is a correlation between packing density and rocking dynamics (as manifested in XRD resolution and B-factors). A comprehensive search of the Protein Data Bank indeed shows that high solvent content correlates with low resolution and high Wilson B, with correlation coefficients of 0.39 and 0.36, respectively (Supplementary Fig. 10a). As expected, these dependencies are subject to strong scatter, reflecting the intricate and complex nature of the crystallization process and the large diversity of the shapes and properties of the analysed structures[43,44]. We have also repeated this analysis for the subset of crystallographic structures in the Protein Data Bank that have been solved at room temperature. The results prove to be very similar (cf. Supplementary Fig. 10a,b). Although not a direct proof, this finding suggests that the spread of orientations observed at cryo-temperatures (typically 100 K) reflects qualitatively the amplitudes of rocking motions at room temperature. In other words, the disorder associated with rocking motion also persists under cryo-cooling conditions.

Discussion

We have shown here that three independent and complementary techniques, NMR, MD and XRD, all provide evidence for an overall rocking motion in protein crystals. The rocking motion is (i) observed by NMR, through the increased $R_{1\rho}$ rates, as well as a slight decrease of order parameters; (ii) reproduced by MD in all-atom crystal lattice simulations; and (iii) confirmed by XRD through the decreased resolution and increased atomic displacement factors. We have been able to provide for the first time a measure of the timescale at which this motion takes place at room temperature, which turned out to be hundreds of nanoseconds to tens of microseconds. Our data suggest that rigid-body motion is an important determinant for the resolution achieved in X-ray crystallography and may explain at least partly why visually perfect crystals do not always produce high-resolution XRD data[45].

Methods

Sample preparation. Uniformly [^{2}H,^{13}C,^{15}N]-labelled ubiquitin was obtained by bacterial overexpression in *Escherichia coli* and purified using ion-exchange and size-exclusion chromatography. The protein was dialysed against water, lyophilized and then resuspended in 20 mM ammonium acetate at pH 4.3 with protein concentration of 20 mg ml^{-1}. All crystals were obtained using a sitting-drop crystallization plate with 47–50 µl protein drops and 500 µl reservoir buffer. In all protein drops except MPD-ub, the protein solution was mixed with reservoir buffer at a ratio of 1:1. All NMR samples have been prepared with H_2O:D_2O ratio of 1:1 (taking into account the exchangeable protons on precipitation agents).

For generating MPD-ub crystals, described before[19], the ubiquitin solution was mixed with reservoir buffer at a ratio of 3.7:1. The reservoir buffer was a mixture of 20 mM citric acid, pH 4.2 and 2-methyl-2,4-pentanediol (MPD) at a ratio of 40:60. Needle-shaped crystals were obtained at 4 °C after about 1–2 weeks.

Cubic-PEG-ub crystals (PDB ID code 4XOL) were obtained with a reservoir buffer of 100 mM 2-(N-morpholino)ethanesulfonic acid (MES), pH 6.3, 20% PEG 3350 and 100 mM zinc acetate. Cubic-shape crystals were obtained within 1 week at 23 °C.

Rod-PEG-ub crystals (PDB ID code 4XOK) were obtained with a reservoir buffer of 50 mM 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES), pH 7.0, 25% PEG 1500 and 25 mM zinc acetate. Long-rod-shape crystals were obtained after 2 weeks at 23 °C.

In addition to these three crystal forms, we also obtained a fourth crystal, from unlabelled ubiquitin. This crystal, rod-PEG-ub-II, (PDB ID code 4XOF) was obtained with a reservoir buffer of 50 mM MES, pH 6.3, 25% PEG 2000 and 1 mM zinc acetate, after 1 month at 23 °C. The amount of crystals obtained was insufficient for NMR analyses, but we were able to determine its structure by XRD.

For the preparation of NMR samples, protein crystals with their crystallization solution were pipetted into an in-house made centrifugation device (funnel) that was adapted to a 1.6-mm solid-state NMR rotor. The device, similar to a recently reported filling tool[46], was spun in a Beckman SW41 rotor at 10,000 r.p.m. (about 15,000*g*) for 10 min to pellet the protein crystals into the NMR rotor. Typical samples contained ~4–5 mg of material (total mass, including the solvent).

NMR spectroscopy. All dynamics experiments were performed on an Agilent VNMRS spectrometer operating at a ^{1}H Larmor frequency of 600 MHz, equipped with a 1.6 mm HXY MAS probe tuned to ^{1}H, ^{13}C and ^{15}N frequencies. HN dipolar couplings as well as ^{15}N R_1 and ^{15}N $R_{1\rho}$ relaxation rate constants were measured using proton-detected two-dimensional HN correlation experiments, identical to those used before, employing MAS frequencies between 37.0 (dipolar-coupling measurement) and 39.5 kHz ($R_{1\rho}$ measurement, using a ^{15}N spin-lock with radio-frequency field strength of 15 kHz)[20]. The REDOR scheme[47] was used to measure HN dipolar couplings; this experiment was shown to be particularly robust with respect to systematic errors[48]. Dipolar couplings were fitted based on peak volumes in a series of two-dimensional HN spectra with variable recoupling time. The employed χ^2 fitting procedure explicitly takes into consideration the radio-frequency field inhomogeneity across the sample as described[20] and utilizes full-scale numerical simulations of the REDOR recoupling element conducted on a grid which samples different coupling strengths. Error margins were obtained from Monte Carlo analyses, based on three times the spectral noise level. Relaxation rate constants were obtained through numerical fits using a single-exponential function and their associated error margins were also obtained from Monte Carlo analysis.

Resonance assignment of MPD-ub has been reported before[19,26]. Assignment of cubic-PEG-ub has been achieved using a series of three-dimensional correlation spectra based on ^{13}C detection (NCACX with 50 ms DARR CC transfer, NCOCX with 50 ms DARR CC transfer and CANCO, NCACB with DREAM transfer) and spectra with ^{1}H detection (hCONH, hCANH, hcoCAcoNH)[49]. For a number of residues two sets of spectral correlations were identified, resulting from the two non-equivalent molecules in the unit cell (chains A and B). It was possible to obtain partial connectivities for certain groups of peaks representing chain A or, alternatively, chain B. It was not possible to unambiguously identify the two sets of resonances, because of the extensive chemical shift overlap between the two sub-spectra. The obtained partial connectivities are shown by red lines in Figs 2 and 3.

MD simulations and analysis. The initial coordinates for the MPD-ub simulation were obtained from the crystallographic structure 3ONS (ref. 16). Four flexible C-terminal residues of ubiquitin were rebuilt as described previously[28]. To determine the protonation status of ionizable residues, we performed the PROPKA[50] calculations for ubiquitin in the relevant crystal-lattice environment. The effective pH was assumed to be 4.2, same as in the crystallization buffer of 3ONS. The original dimensions of the unit crystal cell were all multiplied by a factor 1.016 to account for thermal expansion of the protein crystal on transition from 100 (temperature at which 3ONS was solved) to 301 K[51]. The unit crystal cell was hydrated using SPC/E water[52]; in doing so, the crystallographic water molecules have been retained in their original positions. The system was neutralized by adding Cl^- ions. The periodic boundary box was defined as a block of four crystal unit cells, containing 24 ubiquitin molecules and 8,772 water molecules, for the total of 56,244 atoms. The simulations were conducted under Amber ff99SB*-ILDN force field using Amber 11 program[53–55]. The trajectory was recorded at 301 K, using isothermal-isobaric (NPT) ensemble. The volume of the simulation box remains stable throughout the simulation within 0.5% of its target value (on average, there is a slight uniform expansion as described by linear factor 1.0009). The production rate with NVIDIA GeForce GTX580 cards was 9 ns per card per day. The net length of the trajectory was 1 µs.

The same approach was employed to record the cubic-PEG-ub trajectory. In this case the initial coordinates were derived from the crystallographic structure 3N30 (ref. 17). The periodic boundary box was modelled after a single crystal unit cell, containing 48 ubiquitin molecules (equally divided between chains A and B) and 23,419 water molecules. The net length of the trajectory was 1 µs. The volume of the simulation box remains stable throughout the simulation within 0.7% of its target value (on average, there is a slight uniform contraction as described by linear factor 0.9986). Note that the statistical sampling for both chain A and chain B is the same as for the single ubiquitin chain in the MPD-ub trajectory. Finally, the rod-PEG-ub trajectory was designed based on the crystallographic coordinates 3EHV

Table 1 | X-ray data collection and refinement statistics.

	Rod-PEG-ub	Rod-PEG-ub II	Cubic-PEG-ub
Data collection			
Space group	P 2_1 2_1 2_1	P 2_1 2_1 2_1	P 4_3 3 2
Cell dimensions			
a, *b*, *c* (Å)	43.72, 50.36, 93.46	27.94, 43.30, 50.19	104.95, 104.95, 104.95
α, β, γ (°)	90, 90, 90	90, 90, 90	90, 90, 90
Resolution (Å)	46.73–2.2 (2.279–2.2)	32.78–1.15 (1.191–1.15)	34.98–2.91 (3.013–2.91)
R_{merge}	0.08323 (0.1753)	0.0609 (0.8113)	0.06642 (0.7768)
$I/\sigma I$	16.04 (7.59)	14.10 (1.93)	16.46 (2.11)
Completeness (%)	92.91 (62.00)	99.68 (98.12)	98.83 (99.34)
Redundancy	5.6 (4.9)	7.0 (6.7)	5.1 (5.1)
Refinement			
Resolution (Å)	46.73–2.2 (2.279–2.2)	32.78–1.15 (1.191–1.15)	34.98–2.91 (3.013–2.91)
No. of reflections	56,289 (3144)	155,489 (14390)	23,513 (2321)
R_{work}	0.3015 (0.3538)	0.1369 (0.2230)	0.2372 (0.3805)
R_{free}	0.3249 (0.3776)	0.1713 (0.2605)	0.2689 (0.4189)
No. of non-H atoms	1,791	789	1,191
Protein	1,703	663	1,176
Ligand/ion	6		5
Water	82	125	10
B-factors			
Protein	26.30	14.60	87.70
Ligand/ion	23.90	NA	87.60
Water	19.70	28.00	37.30
R.m.s deviations			
Bond lengths (Å)	0.007	0.010	0.005
Bond angles (°)	1.36	1.27	0.93

NA, not applicable; R.m.s., root mean squared.

(ref. 18). The periodic boundary box was defined as a block of two crystal unit cells, containing 24 ubiquitin molecules (equally divided between chains A, B and C, which comprise the asymmetric unit) and 6,198 water molecules, for the total of 48,234 atoms.

The solution trajectory was based on the coordinate file 1UBQ[55]; this crystal structure has an excellent record in terms of interpreting the solution NMR data. The sample conditions were assumed to be pH 4.7, 300 K, matching those in the experimental study[57]. The truncated octahedral periodic boundary box contained a single ubiquitin molecule and 3,572 water molecules. The net length of the solution trajectory was 2 μs.

To calculate ^{15}N–^{1}H dipolar order parameters from the MPD-ub trajectory, we first superimposed all ubiquitin molecules in the periodic boundary box by applying the appropriate crystal symmetry transformations. Then ^{15}N–^{1}H^{N} vectors were extracted from the transformed coordinates; the vectors pertaining to each individual residue were arranged to the form of a long array (corresponding to the effective 24 μs time span). Finally, the Brüschweiler–Wright formula has been applied to these arrays to calculate S^2 (ref. 58). To calculate the ^{15}N relaxation rate constants, the ^{15}N–^{1}H dipolar correlation functions have been computed on a non-linear grid[59]. They were subsequently averaged over 24 equivalent ubiquitin molecules, as found in the crystal trajectory. The resulting curves were fitted to a combination of six exponentials and a constant. The upper bound was imposed on the fitted correlation times: they were not allowed to be longer than the length of the trajectory, that is, 1 μs. The time-modulated portion of the correlation function (that is, the six weighted exponentials) was then used to evaluate the spectral density functions and subsequently calculate the per-residue ^{15}N R_1 rates[60]. The same strategies were used for the other trajectories.

XRD data collection and processing. Before being flash frozen in the cryogenic N_2 stream on the beamline, crystals were cryoprotected with a brief soaking in a solution composed of the mother liquor complemented with 20% glycerol. Data were collected at 100 K on the ESRF ID29 (cubic-PEG-ub and rod-PEG-ub) and ID23-2 (rod-PEG-ub II) beamlines. Diffraction frames were processed with XDS[61] and intensities were further processed with XSCALE and XDSCONV. All structures were solved using the molecular replacement technique with PHASER[62].

Molecular replacement and model refinement. The initial search models were ubiquitin models obtained under identical crystallization conditions, that is, 3N30 (ref. 17) and 3EHV (ref. 18) for cubic-PEG-ub and rod-PEG-ub, respectively. As expected, two and three molecules of ubiquitin were found in the molecular replacement solutions for cubic-PEG-ub and rod-PEG-ub. Rod-PEG-ub-II crystals grew in the same space group as rod-PEG-ub (P 2_1 2_1 2_1), but with different unit cell parameters and diffracted up to 1.15 Å (Table 1). Only one ubiquitin molecule is present in the asymmetric unit of this crystal form. The refinement was conducted with PHENIX[63]. Following an initial rigid body minimization, the refinement procedure was identical for cubic-PEG-ub and rod-PEG-ub models and consisted of refinement of atomic displacement and individual isotropic B-factors. Water molecules were added to the rod-PEG-ub model using the automated water-picking option in PHENIX and were checked manually for possible close contacts with the protein. For the model of rod-PEG-ub-II, similar refinement strategy was used with the exception of anisotropic refinement of B-factors for all protein atoms, as well as water molecules. Five and six Zn^{2+} ions were modelled in cubic-PEG-ub and rod-PEG-ub coordinates, respectively, based on the presence of large positive peaks in the mFo-DFc map and taking into consideration Zn^{2+} chemical coordination. Model building was carried out with COOT[64]. For rod-PEG-ub, unexpectedly high R_{free} and R_{work} values were obtained (0.325 and 0.302, respectively). Various refinement strategies were attempted without success (for example, multiple models, TLS refinement, use of a reference model). To validate the correctness of our molecular replacement solution, we carried out a *de novo* model building, using the autobuild function of PHENIX. The initial map was computed using our experimental data and the refined ubiquitin model obtained under identical crystallization conditions (3EHV). The automated procedure was able to reconstruct 99% of the backbone and 84% of the side chains confirming the correctness of the molecular replacement solution. Cubic-PEG-ub, rod-PEG-ub and rod-PEG-ub-II have been deposited to the Protein Data Bank under the codes 4XOL, 4XOK and 4XOF, respectively.

MPD-ub crystals grew as sea urchins composed of thousands of extremely thin rods (~100–200 × 5 × 5 μm), impossible to isolate and loop individually. We therefore performed a powder diffraction experiment, to confirm that our crystals have the same space group as the previously reported PDB entry 3ONS (which was obtained under identical conditions and comprehensively characterized by NMR). Details of the powder diffraction experiment are reported in the Supporting Information (Supplementary Fig. 11).

Stereo view images of the electron density maps are provided as Supplementary Fig. 12.

References

1. DePristo, M., de Bakker, P. & Blundell, T. Heterogeneity and inaccuracy in protein structures solved by X-ray crystallography. *Structure* **12,** 831–838 (2004).
2. Fraser, J. S. *et al.* Accessing protein conformational ensembles using room-temperature X-ray crystallography. *Proc. Natl Acad. Sci. USA* **108,** 16247–16252 (2011).
3. de Bakker, P., Furnham, N., Blundell, T. & DePristo, M. Conformer generation under restraints. *Curr. Opin. Struct. Biol.* **16,** 160–165 (2006).
4. Soheilifard, R., Makarov, D. E. & Rodin, G. J. Critical evaluation of simple network models of protein dynamics and their comparison with crystallographic B-factors. *Phys. Rev.* **5,** 026008 (2008).
5. Li, D.-W. & Brüschweiler, R. All-atom contact model for understanding protein dynamics from crystallographic B-factors. *Biophys. J.* **96,** 3074–3081 (2009).
6. Song, G. & Jernigan, R. L. vGNM: a better model for understanding the dynamics of proteins in crystals. *J. Mol. Biol.* **369,** 880–893 (2007).
7. Stec, B., Zhou, R. & Teeter, M. M. Full-matrix refinement of the protein crambin at 0.83 Å and 130 K. *Acta Crystallogr. D* **51,** 663–681 (1995).
8. Parak, F. Physical aspects of protein dynamics. *Rep. Prog. Phys.* **66,** 103–129 (2003).
9. Schomaker, V. & Trueblood, K. N. On the rigid-body motion of molecules in crystals. *Acta Crystallogr. B Struct. Crystallogr. Cryst. Chem.* **24,** 63–76 (1968).
10. Wall, M. E. *et al.* Conformational dynamics of a crystalline protein from microsecond-scale molecular dynamics simulations and diffuse X-ray scattering. *Proc. Natl Acad. Sci. USA* **111,** 17887–17892 (2014).
11. Lewandowski, J. R., Sein, J., Blackledge, M. & Emsley, L. Anisotropic collective motion contributes to nuclear spin relaxation in crystalline proteins. *J. Am. Chem. Soc.* **132,** 1246–1248 (2010).
12. Banigan, J. R., Gayen, A. & Traaseth, N. J. Correlating lipid bilayer fluidity with sensitivity and resolution of polytopic membrane protein spectra by solid-state NMR spectroscopy. *Biochim. Biophys. Acta* **1848,** 334–341 (2014).
13. Park, S. H., Das, B. B., De Angelis, A. A., Scrima, M. & Opella, S. J. Mechanically, magnetically, and 'rotationally aligned' membrane proteins in phospholipid bilayers give equivalent angular constraints for NMR structure determination. *J. Phys. Chem. B* **114,** 13995–14003 (2010).
14. Aisenbrey, C. & Bechinger, B. Investigations of polypeptide rotational diffusion in aligned membranes by ^{2}H and ^{15}N solid-state NMR spectroscopy. *J. Am. Chem. Soc.* **126,** 16676–16683 (2004).
15. Luo, W., Cady, S. D. & Hong, M. Immobilization of the influenza A M2 transmembrane peptide in virus envelope – mimetic lipid membranes: a solid-state NMR investigation. *Biochemistry* **48,** 6361–6368 (2009).
16. Huang, K.-Y., Amodeo, G. A., Tong, L. & McDermott, A. The structure of human ubiquitin in 2-methyl-2,4-pentanediol: a new conformational switch. *Protein Sci.* **20,** 630–639 (2011).
17. Arnesano, F. *et al.* Crystallographic analysis of metal-ion binding to human ubiquitin. *Chemistry* **17,** 1569–1578 (2011).
18. Falini, G., Fermani, S., Tosi, G., Arnesano, F. & Natile, G. Structural probing of Zn(II), Cd(II) and Hg(II) binding to human ubiquitin. *Chem. Commun. (Camb.)* **45,** 5960–5962 (2008).
19. Igumenova, T. *et al.* Assignments of carbon NMR resonances for microcrystalline ubiquitin. *J. Am. Chem. Soc.* **126,** 6720–6727 (2004).
20. Haller, J. D. & Schanda, P. Amplitudes and time scales of picosecond-to-microsecond motion in proteins studied by solid-state NMR: a critical evaluation of experimental approaches and application to crystalline ubiquitin. *J. Biomol. NMR* **57,** 263–280 (2013).
21. Schneider, R. *et al.* Probing molecular motion by double-quantum (^{13}C,^{13}C) solid-state NMR spectroscopy: application to ubiquitin. *J. Am. Chem. Soc.* **132,** 223–233 (2010).
22. Paulson, E. *et al.* Sensitive high-resolution inverse detection NMR spectroscopy of proteins in the solid state. *J. Am. Chem. Soc.* **125,** 15831–15836 (2003).
23. Seidel, K., Etzkorn, M., Heise, H., Becker, S. & Baldus, M. High-resolution solid-state NMR studies on uniformly [^{13}C,^{15}N]-labeled ubiquitin. *Chembiochem* **6,** 1638–1647 (2005).
24. Schmidt, H. L. P. *et al.* Crystal polymorphism of protein GB1 examined by solid-state NMR spectroscopy and X-ray diffraction. *J. Phys. Chem. B* **111,** 14362–14369 (2007).
25. Paßhuber, H. K. *et al.* Structural heterogeneity in microcrystalline ubiquitin studied by solid-state NMR. *Protein Sci.* **24,** 592–598 (2015).
26. Schanda, P., Meier, B. H. & Ernst, M. Quantitative analysis of protein backbone dynamics in microcrystalline ubiquitin by solid-state NMR spectroscopy. *J. Am. Chem. Soc.* **132,** 15957–15967 (2010).
27. Mollica, L. *et al.* Atomic-resolution structural dynamics in crystalline proteins from NMR and molecular simulation. *J. Phys. Chem. Lett.* **3,** 3657–3662 (2012).
28. Xue, Y. & Skrynnikov, N. R. Ensemble MD simulations restrained via crystallographic data: accurate structure leads to accurate dynamics. *Protein Sci.* **23,** 488–507 (2014).
29. Kuzmanic, A., Pannu, N. S. & Zagrovic, B. X-ray refinement significantly underestimates the level of microscopic heterogeneity in biomolecular crystals. *Nat. Commun.* **5,** 3220 (2014).
30. Sahu, S., Bhuyan, A., Majumdar, A. & Udgaonkar, J. Backbone dynamics of barstar: a ^{15}N NMR relaxation study. *Proteins* **41,** 460–474 (2000).
31. Kordel, J., Skelton, N., Akke, M., Palmer, A. G. & Chazin, W. Backbone dynamics of calcium-loaded calbinding-D_{9k} studied by two-dimensional proton-detected ^{15}N NMR spectroscopy. *Biochemistry* **31,** 4856–4866 (1992).
32. Powers, R., Clore, G., Garrett, D. & Gronenborn, A. Relationships between the precision of high-resolution protein NMR structures, solution order parameters and crystallographic B factors. *J. Magn. Reson. B* **101,** 325–327 (1993).
33. Halle, B. Flexibility and packing in proteins. *Proc. Natl Acad. Sci. USA* **99,** 1274–1279 (2002).
34. Yang, L. *et al.* Insights into equilibrium dynamics of proteins from comparison of NMR and X-ray data with computational predictions. *Structure* **15,** 741–749 (2007).
35. Eastman, P., Pellegrini, M. & Doniach, S. Protein flexibility in solution and in crystals. *J. Chem. Phys.* **110,** 10141–10152 (1999).
36. Stocker, U., Spiegel, K. & van Gunsteren, W. On the similarity of properties in solution or in the crystalline state: a molecular dynamics study of hen lysozyme. *J. Biomol. NMR* **18,** 1–12 (2000).
37. Rueda, M. *et al.* A consensus view of protein dynamics. *Proc. Natl Acad. Sci. USA* **104,** 796–801 (2007).
38. Janowski, P. A., Liu, C., Deckman, J. & Case, D. A. Molecular dynamics simulation of triclinic lysozyme in a crystal lattice. *Protein Sci.* (In the press) doi:10.1002/pro.2713 (2015).
39. Raval, A., Piana, S., Eastwood, M. P., Dror, R. O. & Shaw, D. E. Refinement of protein structure homology models via long, all-atom molecular dynamics simulations. *Proteins* **80,** 2071–2079 (2012).
40. Hoops, S. C., Anderson, K. W. & Merz, K. M. Force field design for metalloproteins. *J. Am. Chem. Soc.* **113,** 8262–8270 (1991).
41. Peters, M. B. *et al.* Structural survey of zinc containing proteins and the development of the zinc AMBER force field (ZAFF). *J. Chem. Theory Comput.* **6,** 2935–2947 (2010).
42. Li, P., Roberts, B. P., Chakravorty, D. K. & Merz, K. M. Rational design of particle mesh Ewald compatible Lennard-Jones parameters for +2 metal cations in explicit solvent. *J. Chem. Theory Comput.* **9,** 2733–2748 (2013).
43. Matthews, B. W. X-ray crystallographic studies of proteins. *Annu. Rev. Phys. Chem.* **27,** 493–493 (1976).
44. Kantardjieff, K. A. & Rupp, B. Matthews coefficient probabilities: improved estimates for unit cell contents of proteins, DNA, and protein-nucleic acid complex crystals. *Protein Sci.* **12,** 1865–1871 (2003).
45. Newman, J. A review of techniques for maximizing diffraction from a protein crystal in stilla. *Acta Crystallogr. D* **62,** 27–31 (2006).
46. Böckmann, A. *et al.* Characterization of different water pools in solid-state NMR protein samples. *J. Biomol. NMR* **45,** 319–327 (2009).
47. Gullion, T. & Schaefer, J. Detection of weak heteronuclear dipolar coupling by rotational-echo double-resonance nuclear-magnetic-resonance. *Adv. Magn. Reson.* **13,** 57–83 (1988).
48. Schanda, P., Meier, B. H. & Ernst, M. Accurate measurement of one-bond H-X heteronuclear dipolar couplings in MAS solid-state NMR. *J. Magn. Reson.* **210,** 246–259 (2011).
49. Barbet-Massin, E. *et al.* Rapid proton-detected NMR assignment for proteins with fast magic angle spinning. *J. Am. Chem. Soc.* **136,** 12489–12497 (2014).
50. Bas, D. C., Rogers, D. M. & Jensen, J. H. Very fast prediction and rationalization of pKa values for protein-ligand complexes. *Proteins* **73,** 765–783 (2008).
51. Juers, D. H. & Matthews, B. W. Reversible lattice repacking illustrates the temperature dependence of macromolecular interactions. *J. Mol. Biol.* **311,** 851–862 (2001).
52. Cerutti, D. S., Le Trong, I., Stenkamp, R. E. & Lybrand, T. P. Simulations of a protein crystal: explicit treatment of crystallization conditions links theory and experiment in the streptavidin-biotin complex. *Biochemistry* **47,** 12065–12077 (2008).
53. Hornak, V. *et al.* Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins* **65,** 712–725 (2006).
54. Lindorff-Larsen, K. *et al.* Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins* **78,** 1950–1958 (2010).
55. Best, R. B. & Hummer, G. Optimized molecular dynamics force fields applied to the helix – coil transition of polypeptides. *J. Phys. Chem. B* **113,** 9004–9015 (2009).
56. Vijay-Kumar, S., Bugg, C. E. & Cook, W. J. Structure of ubiquitin refined at 1.8 Å resolution. *J. Mol. Biol.* **194,** 531–544 (1987).
57. Lienin, S., Bremi, T., Brutscher, B., Brüschweiler, R. & Ernst, R. Anisotropic intramolecular backbone dynamics of ubiquitin characterized by NMR relaxation and MD computer simulation. *J. Am. Chem. Soc.* **120,** 9870–9879 (1998).
58. Brüschweiler, R. & Wright, P. E. NMR order parameters of biomolecules: a new analytical representation and application to the gaussian axial fluctuation model. *J. Am. Chem. Soc.* **116,** 8426–8427 (1994).

59. Xue, Y., Pavlova, M. S., Ryabov, Y. E., Reif, B. & Skrynnikov, N. R. Methyl rotation barriers in proteins from ^{2}H relaxation data: implications for protein structure. *J. Am. Chem. Soc.* **129,** 6827–6838 (2007).
60. Bremi, T., Brüschweiler, R. & Ernst, R. A protocol for the interpretation of side-chain dynamics based on NMR relaxation: application to phenylalanines in antamanide. *J. Am. Chem. Soc.* **119,** 4272–4284 (1997).
61. Kabsch, W. XDS. *Acta Crystallogr. D* **66,** 125–132 (2010).
62. McCoy, A. J. *et al.* Phaser crystallographic software. *J. Appl. Crystallogr.* **40,** 658–674 (2007).
63. Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66,** 213–221 (2010).
64. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66,** 486–501 (2010).
65. Kurbanov, R., Zinkevich, T. & Krushelnitsky, A. The nuclear magnetic resonance relaxation data analysis in solids: general $R_1/R_{1\rho}$ equations and the model-free approach. *J. Chem. Phys.* **135,** 184104 (2011).
66. Skrynnikov, N. Asymmetric doublets in MAS NMR: coherent and incoherent mechanisms. *Magn. Reson. Chem.* **45,** S161–S173 (2007).
67. Smith, S., Levante, T., Meier, B. & Ernst, R. Computer simulations in magnetic resonance. An object-oriented programming approach. *J. Magn. Reson.* **106,** 75–105 (1994).
68. Ma, P. *et al.* Probing transient conformational states of proteins by solid-state $R_{1\rho}$ relaxation-dispersion NMR spectroscopy. *Angew. Chem. Int. Ed.* **53,** 4312–4317 (2014).

Acknowledgements

This work was financially supported by the European Research Council (ERC-Stg-2012-311318-ProtDyn2Function), the French Research Agency ANR (ANR 10-PDOC-011-01), as well as Commissariat à l'énergie atomique et aux énergies alternatives (CEA), Centre National de la Recherche Scientifique (CNRS) and Université Grenoble Alpes. This work used the platforms of the Grenoble Instruct Center (ISBG; UMS 3518 CNRS-CEA-UJF-EMBL) with support from FRISBI (ANR-10-INSB-05-02) and GRAL (ANR-10-LABX-49-01) within the Grenoble Partnership for Structural Biology (PSB). Funding from the National Science Foundation (Grant MCB 1158347) to N. R. S. is acknowledged. We are grateful to the ESRF for beam-time under long-term projects MX722, MX1464 and MX1583 (IBS BAG). N.C. is supported by a fellowship from the Fondation France Alzheimer. We thank Florian Schmitzberger (Research Institute of Molecular Pathology, Vienna, Austria), Matthias Huber, Beat H. Meier and Jason Greenwald (ETH Zürich) for insightful discussions.

Author contributions

P.M. crystallized protein, performed and analysed NMR experiments; Y.X. designed, performed and analysed MD simulations; N.C. performed XRD experiments, determined the crystal structures and performed statistical analyses of the Protein Data Bank; J.D.H. analysed NMR experiments and performed the NMR resonance assignment of cubic-PEG-ub; T.Y. performed and analysed MD simulations; I.A. produced and purified protein samples; O.M. analysed MD data and produced the rocking animations; D.W. designed research; J.-P.C. designed, performed and analysed XRD experiments and wrote parts of the paper; N.R.S. designed the research, directed and analysed MD simulations and co-wrote the paper; P.S. designed the research, recorded NMR experiments, coordinated the project and co-wrote the paper.

Additional information

Accession codes: Coordinates and structure factors for cubic-PEG-ub, rod-PEG-ub and rod-PEG-ub-II have been deposited in the RCSB Protein Data Bank under accession codes 4XOL, 4XOK and 4XOF, respectively.

Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing financial interests: The authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

How to cite this article: Ma, P. *et al.* Observing the overall rocking motion of a protein in a crystal. *Nat. Commun.* 6:8361 doi: 10.1038/ncomms9361 (2015).