Towards automated crystallographic structure refinement with a differentiable pipeline

Minhuan Li
John A. Paulson School of Engineering and Applied Sciences
Harvard University
Cambridge, MA 02138
minhuanli@g.harvard.edu

Doeke R. Hekstra
Department of Molecular & Cellular Biology
John A. Paulson School of Engineering and Applied Sciences
Harvard University
Cambridge, MA 02138
doeke_hekstra@harvard.edu

Abstract

The lack of interfaces between crystallographic data and machine learning methods prevents the application of modern deep learning frameworks to crystal structure determination. Here we present SFcalculator, a differentiable pipeline to generate crystallographic observables (structure factors) from atomistic molecular structures and a bulk solvent model. This calculator fills the gap between the long-established crystallography field and state-of-the-art deep learning algorithms. We discuss the correctness and performance of SFcalculator by comparing with the current most-used tool Phenix. Finally, we demonstrate with an initial try that it enables automated structure refinement in a well-regularized latent space defined by a deep generative model, providing a principled way to impose prior knowledge. We believe this tool paves the way towards fully automated structure refinement and a possible end-to-end model, which is crucial for the next generation of high-throughput diffraction experiments.

1 Introduction

X-ray crystallographic structure refinement is the process of achieving agreement between an atomic model of a structure and the experimental data (structure factor amplitudes)[1]. This often is a complex procedure involving multiple strategies of model parameterization and optimization, which can possibly take days or weeks. Over the past few decades, significant progress on refinement methods has been made, featuring the availability and constant improvement of highly automated model building tools and streamlined software suites like CCP4[2], Phenix[3] and Coot[4]. Although efforts have been made to achieve the automated refinement[5], for most cases it is still inevitable to introduce human insight by manual model building. The development of bright new X-ray sources at synchrotrons and X-ray Free Electron Lasers (XFELs) and ongoing robotization are driving rapid increases in experimental throughput and now enable, for example, the collection of thousands of data sets in crystallographic drug fragment screens. Automated structure determination pipelines will be essential for efficient use of these new capabilities.

On the other hand, structure refinement can also be understood as a sampling problem of macromolecular structures constrained by both crystallographic data and physical priors. In addition

Machine Learning for Structural Biology Workshop, NeurIPS 2022.

to traditional simulation-based conformational sampling methods, 3D structure sampling is under intense study using carefully-designed generative models[6, 7, 8] in the deep learning field, and with notable successes [9, 10, 11, 12]. When it comes to the data-constrained sampling, some researchers have successfully constructed neural network models, like CryoDRGN[13] and Cryo-VAE[14], to reconstruct protein structures from Cryo-EM images. However, to our knowledge, such work is still missing for crystallographic data. One reason is the lack of interfaces connecting crystallography data and modern machine learning methods. We will present our `SFcalculator` in Section 2 to fill the gap. The other reason lies in an intrinsic property of crystallography: the x-ray beam scattering reveals the Fourier transform of the model's electron density, but only the magnitude (or amplitude) of the complex number is recorded. This is the "phase problem"[15]: the phase of the complex number is lost. So more structural information is missing in crystallography compared than Cryo-EM, which will consequently require more prior knowledge. How to impose the prior makes the data-constrained sampling in crystallography hard. Currently most tools adopt geometric constraints[16], coming from geometry statistics of high quality data. An alternative approach is to use the potential energy calculated from a molecular mechanics force field as a more principled restraint[17, 18], but still have some performance issue at the moment. In Section 3, we will propose a framework based on a Boltzmann Generator model[6] as an alternative way to incorporate physical prior knowledge. Together, these two parts constitute an initial try towards machine-learning-assisted automated structure refinement in crystallography.

2 A Differentiable `SFcalculator`

Scattering relates to the Fourier transform of sample's electron density map. Generally, both macromolecular atoms and solvent background will contribute to the process, and the solvent correction plays an significant role in the model building[19]. The total structure factor can be defined as:

$$\mathbf{F}_{\text{model}} = \mathbf{F}_{\text{protein}} + \mathbf{F}_{\text{solvent}} \tag{1}$$

which consists of two parts, $\mathbf{F}_{\text{protein}}$ quantifies the contribution from protein atoms, and $\mathbf{F}_{\text{solvent}}$ serves for the solvent correction. The solvent contribution is often approximated using a so-called bulk-solvent mask correction[20], in which the space outside of the protein atoms is treated as uniformly distributed electron density, represented by a "solvent mask". The relative contributions of protein and solvent are then balanced by a linear scale and an anisotropic factor B_{sol}:

$$\mathbf{F}_{\text{model}} = k_{\text{total}}(\mathbf{F}_{\text{protein}} + k_{\text{solvent}} \exp(-\frac{B_{sol} s^2}{4})\mathbf{F}_{\text{mask}}) \tag{2}$$

Model quality in crystallography is usually measured by R factors:

$$R = \frac{\sum ||F_{\text{obs}}| - |\mathbf{F}_{\text{model}}||}{\sum |\mathbf{F}_{\text{model}}|} \tag{3}$$

where F_{obs} denotes the amplitudes observed from experiments and $\mathbf{F}_{\text{model}}$ denotes the structure factor calculated from the model. R factors usually come with subscripts "work" and "free", indicating training and validation set separately.

2.1 The Protein Contribution F_{protein}

For the part of the structure with an explicit atomic representation, there are two ways to calculate F_{protein}:

- Direct method: As the Fourier transform is a linear transformation, one can directly sum contribution from all atoms[21]:

$$\mathbf{F}_{\text{protein}}(\vec{h}) = \sum_G \sum_j O_j \cdot f_{\vec{h},j} \cdot \text{DWF}(\vec{h}) \cdot \exp\left[2\pi i \vec{h} \cdot \left(\mathbf{R}_G \vec{x}_j + \vec{T}_G\right)\right] \tag{4}$$

  where G is the index of symmetry operations (appearing as rotation matrix $\vec{R}_G$ and translation vector $\vec{T}_G$ given the space group; j is the atom index, O_j is occupancy and $\vec{x}_j$ is the

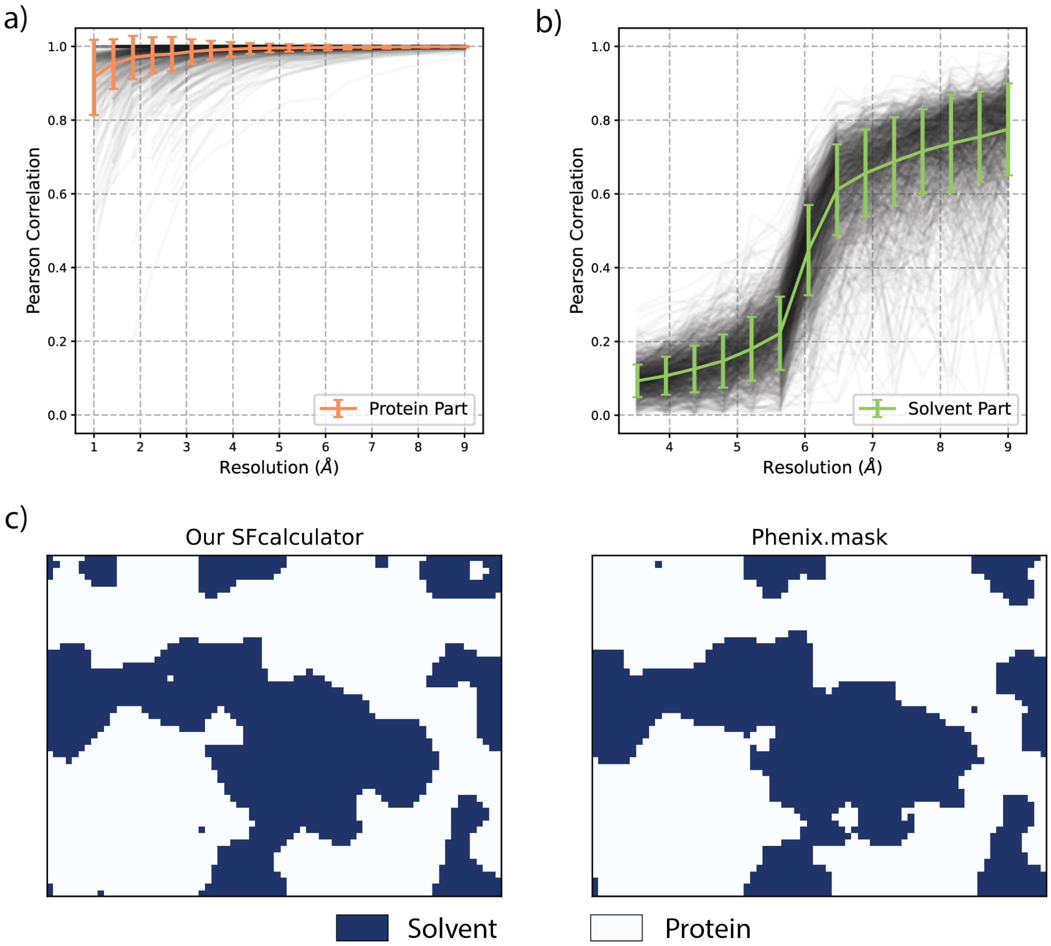

Figure 1: Comparison between SFcalculator and Phenix. a) Statistics of pearson correlation between magnitude of Fprotein calculated by our SFcalculator and Phenix.fmodel. 907 PDB data used here. b) Statistics of pearson correlation between magnitude of Fmask calculated by our SFcalculator and Phenix.fmodel. 907 PDB data used here. c) Visualization of real space solvent masks from two methods. PDBid: 3tsv. The whole unit cell is discretized into grid with size [54, 72, 90], here is the slice of $z = 20$. The two maps look generally similar but differ in high resolution details. The Pearson correlation between the two real space map is 0.83.

fractional coordinates of atom j; $\vec{h}$ is the miller index, $f_{\vec{h},j}$ is the gaussian approximation for atomic scattering factor for the atom type of atom j, and $\mathrm{DWF}(\vec{h})$ is the debye-waller factor.

- FFT-based method: In this approach, one first calculates the overall electron density from the atom coordinates, followed by a discrete Fourier transform into reciprocal space. In this approach, the continuous integral will be replaced with discrete summation and the atomic density is usually truncated within a sphere.

Most current tools (like Phenix) by default use the FFT-based method, because the direct method scales with N(atoms) × N(HKL) and can be slow for large N(atoms) and N(atoms). However, here we adopt the direct-method in our differentiable `SFcalculator` for two reasons:

1. Equation 4 is intrinsically differentiable, while in the FFT-based method, truncation around atom density requires extra care to realize differentiability.
2. With carefully vectorized codes and GPU acceleration, the direct-method can be very fast–faster than both FFT-based and direct methods running on CPU by Phenix. See computation time comparison in Figure 2(a).

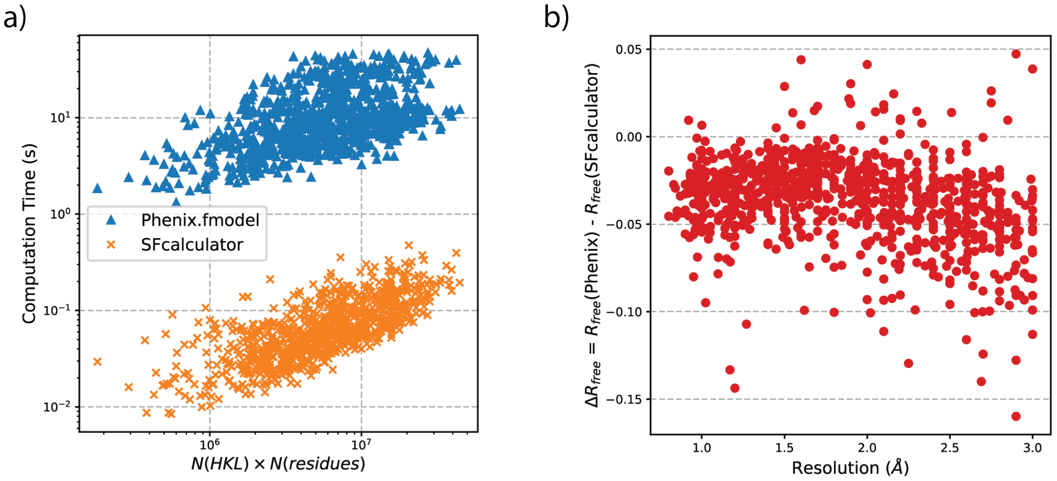

Figure 2: Performance and Correctness Statistics. a) With carefully vectorized codes and GPU acceleration, SFcalculator is much faster than phenix on CPU. b) Due to the approximation of solvent mask, SFcaculator slightly worse R factors. Reproducibility information: Phenix runs on `Intel(R) Xeon(R) Gold 6142 CPU @ 2.60GHz`, no GPU acceleration available; `SFcalculator` runs on a NVIDIA V100 GPU.

To validate our implementation of eq. 4, we computed the correlation between F_{protein} from our `SFcalculator` and the Phenix FFT-based method. As shown in Figure 1(a), the correlation coefficient remains high across the resolution range. This is a convincing evidence that our F_{protein} is correct.

2.2 Bulk Solvent Correction F_{mask}

A differentiable calculation of the solvent mask part is not trivial. Currently, the most popular approach is the probe-shrink method introduced in CNS[22, 23]. It uses van der Waals (or similar) atomic radii r and two parameters r_{probe} and r_{shrink}:

1. Discretize the real space unit cell into grid. Mark the region within radius $r + r_{\text{probe}}$ of each atom as non-solvent.
2. Shrink the non-solvent are by r_{shrink}. This step will eliminate small solvent islands.

The above approach involves non-differentiable operations like rounding to discretize. Here, we describe a differentiable way to approximate a solvent mask:

1. Do a discrete FFT on the Fprotein we obtained above, resulting in a protein electron density map in real space:
$$\text{protein map} = \text{FFT3d}(\text{Fprotein}) \tag{5}$$
2. Define an electron density cutoff by the solvent volume percentage.
$$\text{cutoff} = \text{percentile}(\text{protein map}, \text{solvent percentage}) \tag{6}$$
3. Apply a sigmoid function to set the mask in high-density regions to approximately 0 and low-density regions to 1:
$$\text{mask map} = \text{sigmoid}((\text{cutoff} - \text{protein map}) * \text{scale}) \tag{7}$$
usually choose scale as 50 to make the mask more binary.
4. Do an inverse discrete FFT on the mask map to output F_{mask}.
$$\text{Fmask} = \text{iFFT3d}(\text{mask map}) \tag{8}$$

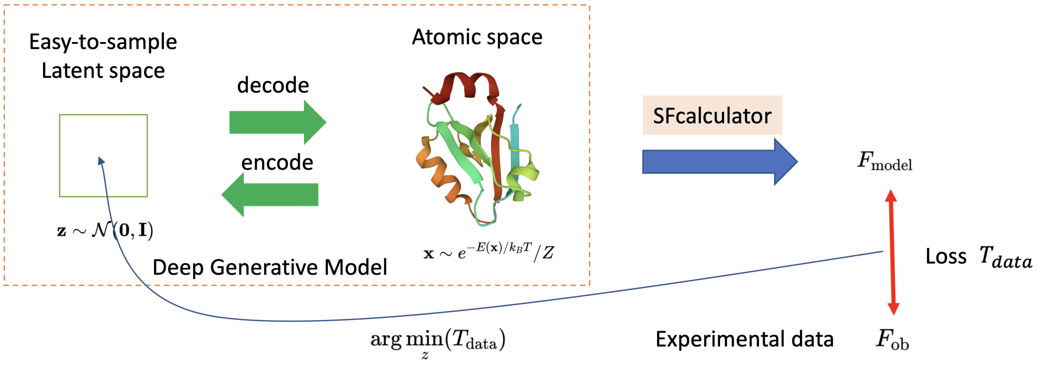

Figure 3: Framework of automated refinement in latent space. We pre-train a deep generative model with physics prior knowledge, encoding a good initial model into the latent space, then do optimization with the experimental likelihood loss.

The pipeline to calculate F_{mask} above is fully differentiable and fast. However, our results differs from the Phenix calculation at high resolution, as shown in Figure 1(b). The correlation of F_{mask} is high at low resolution but quickly decays when it goes to the resolution < 6 Angstrom. This is also clear from visualization of the solvent mask in Figure 1(c). The two maps look generally the same but their high-resolution details are different. However, the contribution of bulk solvent is significant at low resolution (> 8 Angstrom) and becomes weaker at middle and high resolution[20]. As shown in Figure 2(b), this results in slightly worse R factors. However, as shown in the next section, the optimization trajectories demonstrate that this does not preclude optimization.

3 Boltzmann-Generator-Based Refinement

In this section, we present an initial try to connect our `SFcalculator` with deep generative models and do automated structure refinement. As shown in Figure 3, the general framework consists of a pre-trained generative model, combined with `SFcalculator` to connect the atomic model to experimental data. We chose the Boltzmann generator[6] as our first such generative model as its exact likelihood computation enables two resources of physical knowledge during training, one is the likelihood of training data, the other is the potential energy in the real space.

To generate a set of conformations for training of the Boltzmann generator, we ran a single 100ns molecular dynamics (MD) simulation (step size equals 2fs, typically takes around 2-5 hours with OpenMM[24]) started from an AlphaFold[10] prediction generated via ColabFold[25]. We then used these MD conformations along with self-generated samples evaluated by the force field to train the Boltzmann Generator as described by [6]. As shown in Figure4(a), after training, samples drawn from the latent prior of the Boltzmann generator correspond to structures with energy close to the MD samples. The merit of this approach is that sufficiently close to the origin of this latent space points are likely to correspond to physically reasonable structure. In addition, the lossless encoding of the Boltzmann Generator ensures that we can start from a good initial guess.

By connecting the pre-trained generative model to the `SFcalculator`, we can navigate in the latent space by minimizing the crystallographic loss:

$$T_{data} = \sum_{h} \left(\frac{F_{\text{ob}}^{h} - |\mathbf{F}_{\text{model}}^{h}|}{\sigma_{F}^{h}} \right)^{2} \tag{9}$$

where h is the miller index in the working set, σ is the observed standard deviation. To illustrate the progress of structure refinement, we show an example of optimization trajectory in Figure4 (b) with model system PDB 3tsv. Initial model is the still AlphaFold2 prediction. As a benchmark, we used the PDB deposited model (PDB ID 3tsv, previously refined in Phenix) stripped small molecules and explicit waters to make sure the force field could work. As the objective loss went down, the R factors in the test set as a measurement of model quality was also going better, resulting in a model comparable to the benchmark.

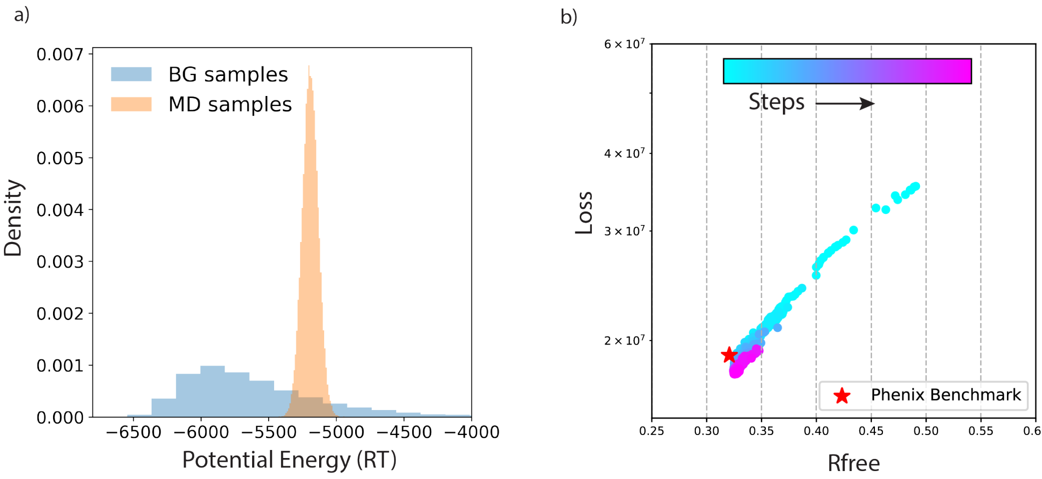

Figure 4: Results. a) Distribution of samples' potential energy after training. BG samples are generated from the latent space gaussian prior. b) Optimization trajectory of the automated refinement in latent space. Model system: PDB 3tsv. Benchmark is PDB deposited model striped out of small molecules and explicit waters. We can see the automated refinement is getting a model close to the benchmark quality.

4 Conclusion

In this work, we presented a differentiable `SFcalculator` to allow direct optimization of atomic models against crystallographic data using machine learning methods. And we show its usefulness by connecting it to a pre-trained Boltzmann Generator and perform an automated refinement. We believe this tool could fill the gap between crystallography field and the fast evolving deep learning field, especially generative molecular models, and pave the path towards better data-constrained sampling models.

5 Code Availability

`SFcalculator` is available in Tensorflow2, Pytorch and Jax on Github: `https://github.com/Hekstra-Lab/SFcalculator`

6 Acknowledgements

The authors thank Kevin Dalton for helpful conversations and Harrison Wang for initial tests of the solvent mask. This work was supported by an NIH Director's New Innovator Award (DP2-GM141000) to D.R.H. Funding to support this research was provided for M.H.L. by the Eric and Wendy Schmidt Center.

References

[1] Ivan G Shabalin, Przemyslaw J Porebski, and Wladek Minor. Refining the macromolecular model–achieving the best agreement with the data from x-ray diffraction experiment. *Crystallography reviews*, 24(4):236–262, 2018.

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

[3] Paul D Adams, Pavel V Afonine, Gábor Bunkóczi, Vincent B Chen, Ian W Davis, Nathaniel Echols, Jeffrey J Headd, L-W Hung, Gary J Kapral, Ralf W Grosse-Kunstleve, et al. Phenix: a

comprehensive python-based system for macromolecular structure solution. *Acta Crystallographica Section D: Biological Crystallography*, 66(2):213–221, 2010.

[4] Paul Emsley, Bernhard Lohkamp, William G Scott, and Kevin Cowtan. Features and development of coot. *Acta Crystallographica Section D: Biological Crystallography*, 66(4):486–501, 2010.

[5] Pavel V Afonine, Ralf W Grosse-Kunstleve, Nathaniel Echols, Jeffrey J Headd, Nigel W Moriarty, Marat Mustyakimov, Thomas C Terwilliger, Alexandre Urzhumtsev, Peter H Zwart, and Paul D Adams. Towards automated crystallographic structure refinement with phenix. refine. *Acta Crystallographica Section D: Biological Crystallography*, 68(4):352–367, 2012.

[6] Frank Noé, Simon Olsson, Jonas Köhler, and Hao Wu. Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science*, 365(6457):eaaw1147, 2019.

[7] Fabian Fuchs, Daniel Worrall, Volker Fischer, and Max Welling. Se (3)-transformers: 3d roto-translation equivariant attention networks. *Advances in Neural Information Processing Systems*, 33:1970–1981, 2020.

[8] Bowen Jing, Gabriele Corso, Jeffrey Chang, Regina Barzilay, and Tommi Jaakkola. Torsional diffusion for molecular conformer generation. *arXiv preprint arXiv:2206.01729*, 2022.

[9] Yilun Du, Joshua Meier, Jerry Ma, Rob Fergus, and Alexander Rives. Energy-based models for atomic-resolution protein conformations. *arXiv preprint arXiv:2004.13167*, 2020.

[10] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

[11] Minkyung Baek, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N Kinch, R Dustin Schaeffer, et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557):871–876, 2021.

[12] Namrata Anand and Tudor Achim. Protein structure and sequence generation with equivariant denoising diffusion probabilistic models. *arXiv preprint arXiv:2205.15019*, 2022.

[13] Ellen D Zhong, Tristan Bepler, Bonnie Berger, and Joseph H Davis. Cryodrgn: reconstruction of heterogeneous cryo-em structures using neural networks. *Nature methods*, 18(2):176–185, 2021.

[14] Dan Rosenbaum, Marta Garnelo, Michal Zielinski, Charlie Beattie, Ellen Clancy, Andrea Huber, Pushmeet Kohli, Andrew W Senior, John Jumper, Carl Doersch, et al. Inferring a continuous distribution of atom coordinates from cryo-em images using vaes. *arXiv preprint arXiv:2106.14108*, 2021.

[15] Robert W Harrison. Phase problem in crystallography. *JOSA a*, 10(5):1046–1055, 1993.

[16] Jeffrey J Headd, Nathaniel Echols, Pavel V Afonine, Ralf W Grosse-Kunstleve, Vincent B Chen, Nigel W Moriarty, David C Richardson, Jane S Richardson, and Paul D Adams. Use of knowledge-based restraints in phenix. refine to improve macromolecular refinement at low resolution. *Acta Crystallographica Section D: Biological Crystallography*, 68(4):381–390, 2012.

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

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

[19] PV Afonine, RW Grosse-Kunstleve, PD Adams, and A Urzhumtsev. Bulk-solvent and overall scaling revisited: faster calculations, improved results. *Acta Crystallographica Section D: Biological Crystallography*, 69(4):625–634, 2013.

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

[21] Bernhard Rupp. *Biomolecular crystallography: principles, practice, and application to structural biology*. Garland Science, 2009.

[22] Axel T Brünger, Paul D Adams, G Marius Clore, Warren L DeLano, Piet Gros, Ralf W Grosse-Kunstleve, J-S Jiang, John Kuszewski, Michael Nilges, Navraj S Pannu, et al. Crystallography & nmr system: A new software suite for macromolecular structure determination. *Acta Crystallographica Section D: Biological Crystallography*, 54(5):905–921, 1998.

[23] Marcin Wojdyr. Gemmi: A library for structural biology. *Journal of Open Source Software*, 7(73):4200, 2022.

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

[25] Milot Mirdita, Konstantin Schütze, Yoshitaka Moriwaki, Lim Heo, Sergey Ovchinnikov, and Martin Steinegger. Colabfold: making protein folding accessible to all. *Nature Methods*, pages 1–4, 2022.