Torsion angles to map and visualize the conformational space of a protein.

Helen Mary Ginn: Diamond Light Source Ltd, Harwell Science and Innovation Campus, Didcot OX11 0DE, United Kingdom; Institute for Nanostructure and Solid State Physics; Hamburg 22761, Germany

1 Abstract

Present understanding of protein structure dynamics trails behind that of static structures. Representation of protein entities (RoPE), based on torsion angles, derives an interpretable conformational space which correlates with data collection temperature, resolution and reaction coordinate. This indicates that torsion angles are a more sensitive and biologically relevant descriptor for protein conformation than atomic coordinates.

2 Introduction

Researchers can now deliver structure solutions for protein targets at remarkable speed, evidenced by the exponentially growing number of Protein Data Bank (PDB) [1] depositions. This database has reached sufficient size to support the development of AlphaFold [2], suggesting that protein folding prediction is a tractable goal [3]. This progress is tempered by the smaller strides made towards understanding allosteric regulation and subtle conformational shifts, which often govern enzyme regulation, activity and signalling [4]. For example, we still do not know whether cryo-cooling samples for data collection, which underpins the most experimental structure determinations, has a deleterious effect on biologically relevant information [5–8], and whether the quest for an interpretable high-resolution map is worth the trade-off in trapping a single conformational state potentially less relevant to biological function. In short, our understanding of dynamics, and hence the essence of biological activity, lags far behind our understanding of static structures.

Beyond experimental data collection, we have computational tools to analyse structures. Individual crystal structures can be analysed using tools such as ECHT [9] or qFit3 [10] to get a sense of dynamic motion in the structure. 2D and 3D classification in cryo-electron microscopy is also able to select enriched conformations from a heterogenous set of particles [11]. Ensemble refinement combines molecular dynamics (MD) and fitting to experimental data [12]. MD has also been adapted to crystal lattices and supercells, as previously reviewed [13]. Tools such as normal mode analysis [14] and elastic network models [15] have a lower computational cost than MD. Diffuse scattering is a potential

experimental source of dynamic information, but application to biomolecules is exceptionally difficult [16]. Despite these tools, we continue to struggle to accurately describe and track subtle conformational states in proteins beyond the easily recognisable large domain or secondary structure movements and rearrangements. These difficult-to-visualise subtle shifts are nevertheless vital for, and intrinsically bound to, activity, efficacy and regulation. In this paper, I introduce a method capable of discerning subtle conformational changes across the entirety of a biomacromolecule. This approach uses torsion angles to describe conformational states, a more biologically relevant quantity than atomic coordinates. This powerful space shows correlations with resolution of the dataset, data collection temperature and space group. The correlations with resolution indicates an explanation by which biological molecules, crystallised under near-equivalent conditions, nevertheless vary widely in resolution.

This multi-dataset analysis is named Representation of Protein Entities (RoPE). The product of this analysis is a representation of protein conformational space (hereby referred to as RoPE space). What were previously considered subtle shifts in the proteins conformational state are starkly separated in RoPE space. Correlation with dataset metadata directly informs questions of data collection temperature, reaction pathways, and what may be contributing to poor resolution in structural biology. Therefore, RoPE is a potentially powerful tool that can also be used to drive experimental decision-making to help experimentalists obtain data suitable to address complex biological questions dominated by protein dynamics. The algorithm is available as an open source web application (https://rope.hginn.co.uk).

3 Results

Calculation of RoPE space. Vagabond refinement [17] of structures already subjected to standardised PDB-redo refinement [18] was used to obtain optimal estimates of torsion angles. All torsion angles, excluding those for hydrogen atoms, contribute to the calculation of the RoPE space, as described in the Methods. Normalisation of torsion angles ensure that uncorrelated highly flexible side chains do not have disproportionate contributions to the distribution (see Methods). Singular value decomposition (SVD) was carried out on the torsion angle difference vectors. Fig. 1 plots are generated by manual rotation of the first three axes to reveal the pertinent information and then projected into 2D. Therefore, the X and Y axes are some combination of the first three principle components, best showing the variation with respect to the metadata. Each data point represents one polypeptide chain.

Variation with temperature. From 459 crystal structures, 515 molecules of bovine trypsin were extracted (all molecules were used from these crystal structures and some contained non-crystallographic

symmetry). Their RoPE space was determined from all structures collected between 1982 and 2022 which had a corresponding entry in the PDB-redo database [18]. Trypsin structures separated into two categories corresponding to cryo-cooled and room temperature (RT) data collection, along the vertical axis of Fig. 1a. There was greater variation within each temperature group rather than between them, as indicated by the longer horizontal axis. Recognition of cryo- and RT by overlaying structures by eye is very difficult. A linear change in multiple torsion angles often produces non-linear motion in atoms, even causing a single atom's trajectory to double back on itself in a cryptic manner. However, the distinction is clear in RoPE space (Fig. 1a). Notably, there were eight outlier structures where reportedly cryo-cooled molecules were present in the room temperature cluster. Six of these cases were contradicted by the published methods, stating that they were indeed collected at room temperature [19, 20], suggesting that the wrong metadata were deposited in the PDB. Other publications did not specify the collection temperature directly [21, 22]. However, another paper confirmed data collection for two other trypsin outliers was carried out at 100 K [23]. This suggests that these outliers were largely caused by erroneous annotation in the PDB, and that RoPE space is therefore powerful enough to identify these metadata errors from the atomic coordinates alone.

Not all structures have a similar RT/cryo-cooled distinction in RoPE space, and for most there is insufficient sampling to draw a conclusion. Some splits are less extreme; the RT and cryo-cooled regions of lysozyme overlap more than for trypsin (Fig. 1b). In both cases however the torsion angle difference associated with the temperature change is essentially preserved across space groups. RoPE space can therefore be used to determine on a per-protein basis if the biological interpretation of the structures is likely to be perturbed by the RT/cryo-cooled conformational split.

Variation with resolution. Drug or fragment screens provide an excellent supply of multi-datasets from crystals selected specifically for their homogeneity. Despite this, often large variations in resolution between crystals remain, which are largely unexplained. The RoPE space for a two drug or fragment screens was determined. BAZ2B (Fig. 1c) [24] and the DNA crosslink repair protein 1A (Fig. 1d) show a clear dependence of resolution with a molecules position within the RoPE space. In each case, there is a tendency for the high resolution structures to coalesce at the edges of the occupied space. This suggests that molecules that lie in the middle of the RoPE space may actually be sampling a wider distribution of neighbouring conformations in the crystal, blurring the electron density and decreasing the reported resolution. This also indicates the utility of torsion angle space for describing ensembles of structures in the future.

Application to a time-resolved experiment. A time-resolved experiment on the dynamics of photo-excitation of carboxymyoglobin to initiate loss of carbon monoxide consists of timepoints

between -0.1 ps and 150 ps from the point of photoexcitation [25], with a time resolution on the scale of femtoseconds. The RoPE space of these structures showed that, over the first three picoseconds, two torsion angle modes are sufficient to represent a clear trajectory during release of carbon monoxide. Thus, clear and subtle changes can be reliably observed, despite a reported maximum RMSD of 0.11 between the first and last timepoint [25]. Mapping the trajectory of the first SVD component (horizontal axis on Fig. 1e) back onto the dark structure (Fig. 2a) and as a per-residue plot (Fig. 2b) showed similar motions as seen in the original study, highlighting the potential of RoPE for detailed structural analysis of extracted motions. The original analysis suggested that out-of-plane movement of the iron within the haem reaches its final position by 3 picoseconds, concluding the biological motion. The last three timepoints, 10 ps, 50 ps and 150 ps, are therefore beyond the biologically relevant timescales for CO dissociation in myoglobin and in-line with this, they did not strongly correlate with any other timepoints in RoPE space.

4 Conclusions

Torsion angles are significantly more sensitive for describing and evaluating protein conformations than atomic coordinate displacement. Cryptic (non-linear) motions of atoms are described by near-linear motions in torsion angle space. RoPE exposes correlations with temperature and resolution that are often unclear in atomic coordinate space. It is not unreasonable to assume allosteric mechanisms and conformational states in the catalytic cycle of metabolic enzymes will be well-suited to analysis in RoPE space. Future analyses will work towards establishing why proteins are confined to their particular RoPE spaces, each of which have their own idiosyncrasies, and how proteins may find a path in torsion angle space from one conformational state to another.

For trypsin, one of the largest divides is on data collection temperature, and points to the nature of the potential artefact caused by cryo-cooling crystals. However, the preservation of other linear independent motions may prove a tolerable trade-off for improved crystal endurance in face of radiation damage, which is a major concern for most protein crystals. This analysis may be a method to establish a reasonable estimation of how cryo-cooling introduces artefacts for a given protein system.

RoPE analysis is not fundamentally limited to X-ray diffraction studies. The analyses in this paper have benefited from standardisation of refinement of X-ray structures by PDB-redo. RoPE is however generally applicable to atomic models regardless of data source. Researchers who trust their atomic coordinates can proceed with data from any experimental method.

The code is released under the GPL3 license [26]. A compiled web application is available at https://rope.hginn.co.uk covering the entire pipeline.

5 Acknowledgements

Thank you to Dr Gianluca Santoni, Professor David Stuart, Dr Diana Monteiro, Dr Benjamin Krishna and Dr Helen Duyvesteyn for a critical reading through the manuscript, and to Dimitris Triandafillidis for proofreading, helpful discussions, test cases and bug reports.

6 Methods

Definition and population of a protein entity. The UniProt sequence was obtained for trypsin and lysozyme, and the sequences downloaded and referred to as an entity. Sequence similarity was used to search for and download similar protein structures from the PDB. Alternatively, PDB codes associated with a single multi-dataset research project were downloaded for BAZ2B [24], PDB group deposition code G_1002036 and the time-resolved myoglobin CO-dissociation experiment [25] from PDB-redo [18]. The downloaded sequences from the PDB file were aligned against the entity reference sequence, identifying point mutations and correcting the register for insertions and deletions where necessary. Chains identified as having at least 10 residues and 80% identity to a target entity sequence were assigned as a member of the entity group. The mapping of residues to the reference sequence was recorded for each participating chain.

Refinement of torsion angles. Bond lengths and angles were reset to the best estimates from the CCP4 monomer library [27]. The initial torsion angles estimated from the atomic coordinate structure were then optimised using the first step of the Vagabond algorithm as previously described [17] to best match the model atomic coordinates. All torsion angles were then associated with their corresponding residue from the master sequence. Torsion angles which determined the positions of hydrogen atoms, if applicable, were ignored. Torsion angles were wrapped round 360 degrees as required to match the first estimated torsion angle in an analogous set. Torsion angles were considered analogous if they belonged to the same aligned residue and were defined for the same four contiguous atom names, as defined in the PDB and geometry files.

Clustering of torsion angles. Cluster analyses of torsion angles were carried according to the principles of the software package cluster4x [28]. In short, differences were calculated with respect to the average for each identified torsion angle across all participating datasets. Main chain torsion angle variation is much smaller than flexible side chain variation. To prevent highly flexible (but often uncorrelated) flexible side chain motion from dominating the clustering output, torsion angle values were normalised: the standard deviation of values for each analogous torsion angle was determined, and each torsion angle divided through by this standard deviation, thereby converting torsion angle

deviations to Z-scores. A correlation matrix was generated by calculating correlation coefficients between each pair of data sets. Singular value decomposition (SVD) was applied to the resulting matrix and the three top axes of variation were plotted at rotation angles which clearly show the variation in the datasets. The graphical user interface allowed structures to be coloured by metadata categories with numerical values such as collection temperature or resolution. Marker types for structures were changed according to discrete metadata such as space group or point mutation, or series of related datasets were connected by lines.

Mapping motion back onto structure. A principle component of the cluster analysis corresponds to a combination of changes in each torsion angle. A scale factor was applied to bring the average magnitude in torsion angle change to 1°. These torsion angles were then added to the torsion angles of a reference structure (in this case, myoglobin dark structure) and an atomic coordinate structure calculated using the full set of torsion angles. The Kabsch algorithm [29] was used to superimpose the modified structure onto the unmodified reference structure.

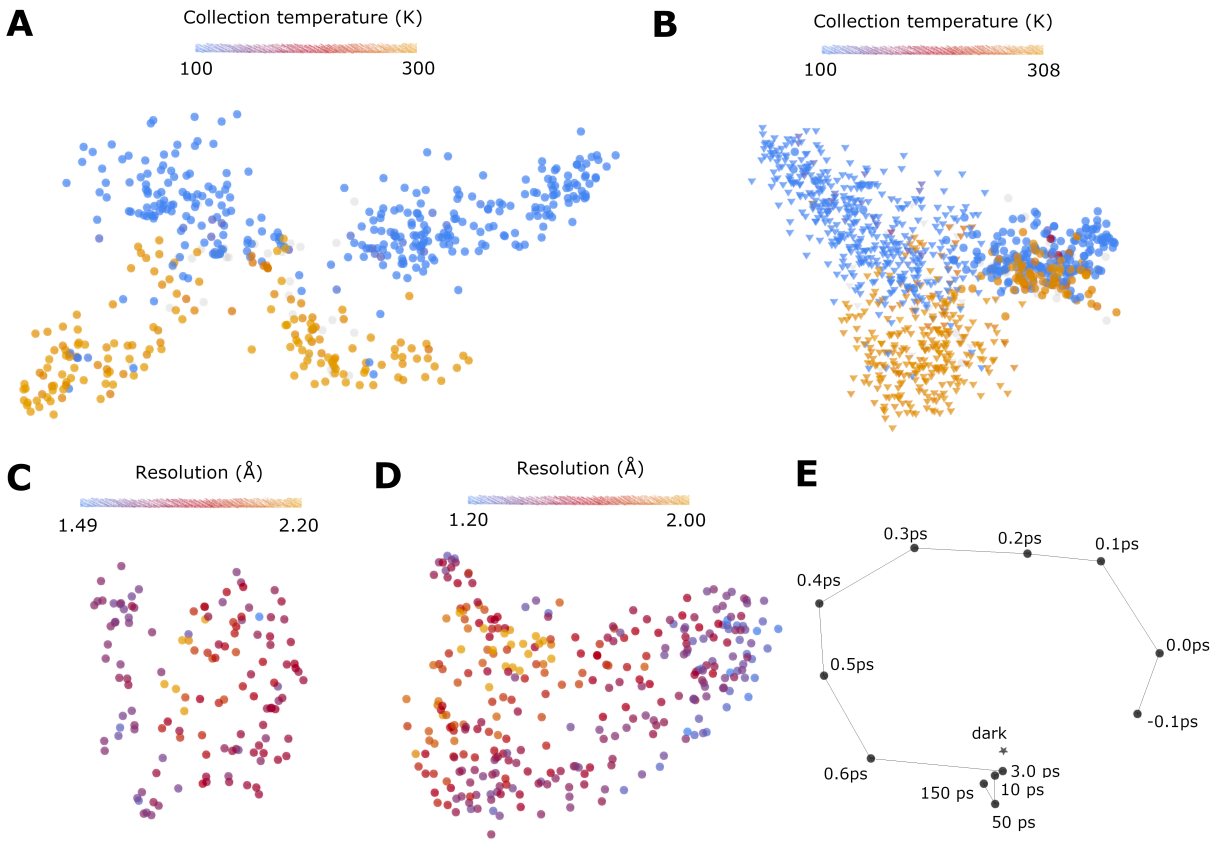

Figure 1: RoPE spaces for (A) trypsin (Uniprot ID P00760, 459 structures, 511 molecules) showing split on collection temperature. (B) lysozyme (Uniprot ID P00698, 939 structures, 1038 molecules) with the majority space group P43212 marked as triangles. (C) BAZ2B, drug fragment screen for PanDDA analysis, 132 structures/molecules, coloured by resolution. (D) DNA cross-link repair protein 1A, deposition ID G_1002036, 312 structures/molecules, coloured by resolution. (E) myoglobin time series dataset. Dark structure marked as a star, -0.1 ps to 150 ps timepoints marked as a line series.

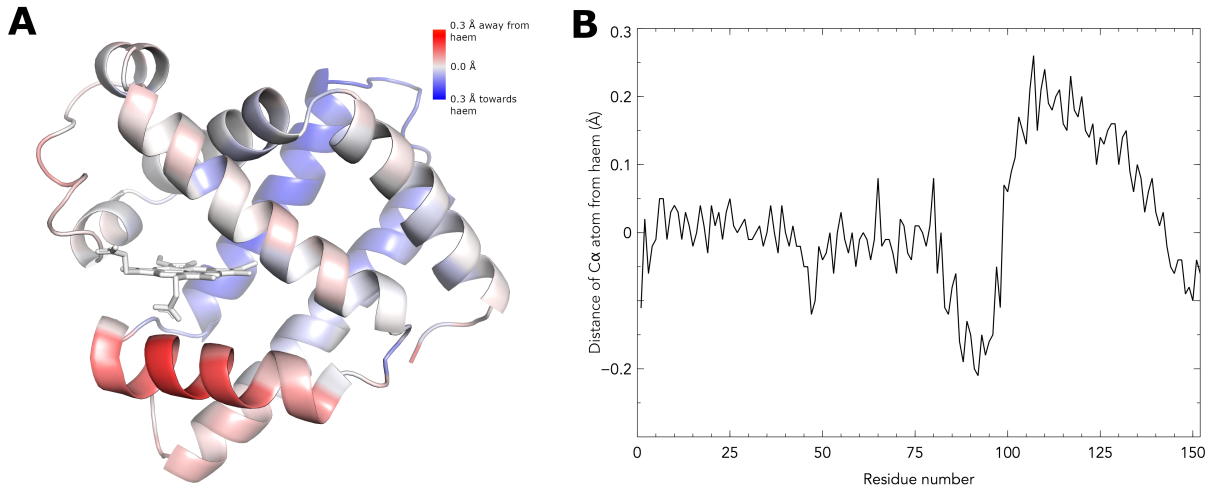

Figure 2: (A) for the time-resolved myoglobin dataset, the first principle component from SVD (horizontal axis, Fig. 1e), corresponding to a set of torsion angle differences, was mapped onto the dark structure of myoglobin. This structure is coloured according to the change in distance of each Cα atom to the haem iron. (B) change in distance to haem as a function of C-alpha residue number.

References

[1] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The protein data bank," *Nucleic acids research*, vol. 28, no. 1, pp. 235–242, 2000.

[2] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, *et al.*, "Highly accurate protein structure prediction with alphafold," *Nature*, vol. 596, no. 7873, pp. 583–589, 2021.

[3] P. B. Moore, W. A. Hendrickson, R. Henderson, and A. T. Brunger, "The protein-folding problem: Not yet solved," *Science*, vol. 375, no. 6580, pp. 507–507, 2022.

[4] H. N. Motlagh, J. O. Wrabl, J. Li, and V. J. Hilser, "The ensemble nature of allostery," *Nature*, vol. 508, no. 7496, pp. 331–339, 2014.

[5] J. S. Fraser, H. van den Bedem, A. J. Samelson, P. T. Lang, J. M. Holton, N. Echols, and T. Alber, "Accessing protein conformational ensembles using room-temperature x-ray crystallography," *Proceedings of the National Academy of Sciences*, vol. 108, no. 39, pp. 16247–16252, 2011.

[6] S. Russi, A. González, L. R. Kenner, D. A. Keedy, J. S. Fraser, and H. v. d. Bedem, "Conformational variation of proteins at room temperature is not dominated by radiation damage," *Journal of synchrotron radiation*, vol. 24, no. 1, pp. 73–82, 2017.

[7] D. W. Kneller, G. Phillips, H. M. ONeill, R. Jedrzejczak, L. Stols, P. Langan, A. Joachimiak, L. Coates, and A. Kovalevsky, "Structural plasticity of sars-cov-2 3cl mpro active site cavity revealed by room temperature x-ray crystallography," *Nature communications*, vol. 11, no. 1, pp. 1–6, 2020.

[8] S. M. Keable, A. Kölsch, P. S. Simon, M. Dasgupta, R. Chatterjee, S. K. Subramanian, R. Hussein, M. Ibrahim, I.-S. Kim, I. Bogacz, *et al.*, "Room temperature xfel crystallography reveals asymmetry in the vicinity of the two phylloquinones in photosystem i," *Scientific reports*, vol. 11, no. 1, pp. 1–14, 2021.

[9] N. M. Pearce and P. Gros, "A method for intuitively extracting macromolecular dynamics from structural disorder," *Nature communications*, vol. 12, no. 1, pp. 1–11, 2021.

[10] B. T. Riley, S. A. Wankowicz, S. H. de Oliveira, G. C. van Zundert, D. W. Hogan, J. S. Fraser, D. A. Keedy, and H. van den Bedem, "qfit 3: Protein and ligand multiconformer modeling for

x-ray crystallographic and single-particle cryo-em density maps," *Protein Science*, vol. 30, no. 1, pp. 270–285, 2021.

[11] S. H. Scheres, "Processing of structurally heterogeneous cryo-em data in relion," *Methods in enzymology*, vol. 579, pp. 125–157, 2016.

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

[13] D. S. Cerutti and D. A. Case, "Molecular dynamics simulations of macromolecular crystals," *Wiley Interdisciplinary Reviews: Computational Molecular Science*, vol. 9, no. 4, p. e1402, 2019.

[14] D. A. Case, "Normal mode analysis of protein dynamics," *Current Opinion in Structural Biology*, vol. 4, no. 2, pp. 285–290, 1994.

[15] A. R. Atilgan, S. Durell, R. L. Jernigan, M. C. Demirel, O. Keskin, and I. Bahar, "Anisotropy of fluctuation dynamics of proteins with an elastic network model," *Biophysical journal*, vol. 80, no. 1, pp. 505–515, 2001.

[16] S. P. Meisburger, D. A. Case, and N. Ando, "Diffuse x-ray scattering from correlated motions in a protein crystal," *Nature communications*, vol. 11, no. 1, pp. 1–13, 2020.

[17] H. M. Ginn, "Vagabond: bond-based parametrization reduces overfitting for refinement of proteins," *Acta Crystallographica Section D: Structural Biology*, vol. 77, no. 4, pp. 424–437, 2021.

[18] R. P. Joosten, F. Long, G. N. Murshudov, and A. Perrakis, "The pdb_redo server for macromolecular structure model optimization," *IUCrJ*, vol. 1, no. 4, pp. 213–220, 2014.

[19] E. Toyota, K. K. Ng, H. Sekizaki, K. Itoh, K. Tanizawa, and M. N. James, "X-ray crystallographic analyses of complexes between bovine β-trypsin and schiff base copper (ii) or iron (iii) chelates," *Journal of Molecular Biology*, vol. 305, no. 3, pp. 471–479, 2001.

[20] J. Cui, F. Marankan, W. Fu, D. Crich, A. Mesecar, and M. E. Johnson, "An oxyanion-hole selective serine protease inhibitor in complex with trypsin," *Bioorganic & medicinal chemistry*, vol. 10, no. 1, pp. 41–46, 2002.

[21] J. R. Pruitt, D. J. Pinto, R. A. Galemmo, R. S. Alexander, K. A. Rossi, B. L. Wells, S. Drummond, L. L. Bostrom, D. Burdick, R. Bruckner, *et al.*, "Discovery of 1-(2-aminomethylphenyl)-3-trifluoromethyl-n-[3-fluoro-2 -(aminosulfonyl)[1, 1 -biphenyl)]-4-yl]-1 h-pyrazole-5-carboxyamide (dpc602), a potent, selective, and orally bioavailable factor xa inhibitor," *Journal of medicinal chemistry*, vol. 46, no. 25, pp. 5298–5315, 2003.

[22] J. Fokkens and G. Klebe, "A simple protocol to estimate differences in protein binding affinity for enantiomers without prior resolution of racemates," *Angewandte Chemie International Edition*, vol. 45, no. 6, pp. 985–989, 2006.

[23] H. Nar, M. Bauer, A. Schmid, J.-M. Stassen, W. Wienen, H. W. Priepke, I. K. Kauffmann, U. J. Ries, and N. H. Hauel, "Structural basis for inhibition promiscuity of dual specific thrombin and factor xa blood coagulation inhibitors," *Structure*, vol. 9, no. 1, pp. 29–37, 2001.

[24] N. M. Pearce, T. Krojer, A. R. Bradley, P. Collins, R. P. Nowak, R. Talon, B. D. Marsden, S. Kelm, J. Shi, C. M. Deane, *et al.*, "A multi-crystal method for extracting obscured crystallographic states from conventionally uninterpretable electron density," *Nature communications*, vol. 8, no. 1, pp. 1–8, 2017.

[25] T. R. Barends, L. Foucar, A. Ardevol, K. Nass, A. Aquila, S. Botha, R. B. Doak, K. Falahati, E. Hartmann, M. Hilpert, *et al.*, "Direct observation of ultrafast collective motions in co myoglobin upon ligand dissociation," *Science*, vol. 350, no. 6259, pp. 445–450, 2015.

[26] H. M. Ginn, "helenginn/rope: Baseline rope analysis. doi 10.5281/zenodo.6958155," Aug. 2022.

[27] M. D. Winn, C. C. Ballard, K. D. Cowtan, E. J. Dodson, P. Emsley, P. R. Evans, R. M. Keegan, E. B. Krissinel, A. G. Leslie, A. McCoy, *et al.*, "Overview of the ccp4 suite and current developments," *Acta Crystallographica Section D: Biological Crystallography*, vol. 67, no. 4, pp. 235–242, 2011.

[28] H. M. Ginn, "Pre-clustering data sets using cluster4x improves the signal-to-noise ratio of high-throughput crystallography drug-screening analysis," *Acta Crystallographica Section D: Structural Biology*, vol. 76, no. 11, pp. 1134–1144, 2020.

[29] W. Kabsch, "A solution for the best rotation to relate two sets of vectors," *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, vol. 32, no. 5, pp. 922–923, 1976.