Comprehensive Encoding of Conformational and Compositional Protein Structural Ensembles through mmCIF Data Structure

Stephanie A. Wankowicz[1], James S. Fraser[1]

1) Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, San Francisco, CA 94158, USA.

ABSTRACT

In their folded state, biomolecules exchange between multiple conformational states, crucial for their function. However, most structural models derived from experiments and computational predictions only encode a single state. To represent biomolecules more accurately, we must move towards modeling and predicting structural ensembles. Information about structural ensembles exists within experimental data from X-ray crystallography and cryo electron microscopy (cryoEM). While new tools are available to detect conformational and compositional heterogeneity that exist within these ensembles, the legacy PDB data structure does not robustly encapsulate this complexity. We propose modifications to the Macromolecular Crystallographic Information File (mmCIF) to improve the representation and interrelation of conformational and compositional heterogeneity. These modifications will enable improved tools to capture macromolecular ensembles in a way that is human and machine interpretable, potentially catalyzing breakthroughs for ensemble-function predictions, analogous to AlphaFold's achievements with single structure prediction.

Introduction

Most structural models deposited in the Protein Data Bank (PDB)[1] result from experimental X-ray crystallography or single particle CryoEM studies. These methods collect data averaged over tens of thousands to billions of individual copies of the system (containing macromolecules, solvent, ions, small molecules, etc). Each molecule within the system can adopt a different conformation (conformational heterogeneity) and may differ slightly in chemical composition (compositional heterogeneity). However, structural models generally represent the system with a single coordinate set. This simplification overlooks the multiple states present in the experimental data and consequently omits details vital to understanding protein function[2,3]. The single coordinate set convention originated in the specifications of the legacy PDB format and is perpetuated in the current PDBx/mmCIF standard[4]. However, with the universal adoption of PDBx/mmCIF[5], we propose to expand our representations to separate these two aspects of heterogeneity within the sample, enabling more precise and accurate structures to train deep learning approaches that model dynamic biomolecular systems.

Figure 1. Model types to represent conformational heterogeneity. A. Examples of the multiple conformations of protein sidechains captured by CryoEM and X-ray crystallography. **B.** Ensemble representation of sidechains. **C.** Multiconformer representation of sidechains.

In the cases where deposited models go beyond a single set of coordinates, capturing the underlying experimental ensemble is solved in two primary ways: ensemble models that encode many copies of the system, each with a potentially distinct conformation/composition, in a single PDB deposition or muliticonformer models that encode alternate conformational/compositional states only for certain parts of the system (**Figure 1**). Nuclear Magnetic Resonance (NMR) data is typically encoded in ensemble models because it inferentially represents sparse distance and angle restraints[6]. In contrast, CryoEM and X-ray crystallography density maps provide atomistic detail across the entire system, enabling the precise modeling of alternate states directly from a real space signal. As there is no principled way of choosing the most parsimonious ensemble size, encoding cryoEM and X-ray crystallography data in ensemble models can result in an exploding data-to-parameter ratio and overfitting (**Figure 1B**)[7]. Moreover, ensemble models are difficult to modify manually. In contrast, multiconformer models can represent all states within a single model, reducing the data-to-parameter ratio and allowing for more facile human visualization and manipulation[8–10]. A limitation of the current PDBx/mmCIF data structure for ensemble and multiconformer models is that it cannot represent the complex interdependencies of alternative conformational states in the experimental ensemble.

Modeling the experimental ensemble also requires representing the chemical compositional heterogeneity. This heterogeneity can result from covalent modification (e.g. post-translational modification) or the presence of a binding partner stabilized by non-covalent interactions (e.g. a subunit of a macromolecular complex, a small molecule ligand, or even a solvent molecule). Using ensemble-based approaches leads to the same model size selection problems outlined above. Refining the weight of different conformational ensemble members in the modified/bound state also connects this heterogeneity more naturally to the multiconformer format. A major issue for encoding compositional heterogeneity in multiconformer models is that the current formats use the exact same representation for conformational and compositional heterogeneity, creating ambiguity about the various states present in the models and their relationship to the experimental ensemble.

Here, we proposed amendments to the PDBx/mmCIF model format[4] to improve the encoding of the conformational and compositional ensembles in experimental structural biology data. Using the extensible and flexible dictionary-based data structure of the mmCIF/PDBx format, we propose separated entities to capture conformational and compositional heterogeneity that can be layered to show hierarchical relationships (**Figure 1**). These modifications will improve our ability to explain structural ensembles and provide critical training data for new protein ensemble-function predictions.

The current PDBx/mmCIF format inadequately captures conformational and compositional heterogeneity

The failure of deposited structures in the Protein Data Bank (PDB) to represent the underlying experimental conformational and compositional heterogeneity is partly attributable to the complexity of modeling in the presence of limited signal-to-noise[11]. Noise arises from many sources, including crystal imperfections and radiation damage in X-ray crystallography[12,13], beam-induced motion, and imperfect detector Detector Quantum Efficiency (DQE) in cryo-EM[14]. Additionally, poor modeling resulting from inaccurate phases (for X-ray) and errors in particle alignment and classification (for CryoEM) dominate the imperfect agreement between experiment and model. Further complicating the discovery of heterogeneity is that conformational heterogeneity manifests in many forms. A high amount of harmonic heterogeneity manifests in a fall off of density from a mean atomic position. This type of heterogeneity can be modeled in the PDB format by isotropic, anisotropic, or grouped (e.g. Translation-Libration-Screw[15,16]) B-factors that fit the extent of the disorder[17,18].

Additionally, many macromolecular motions have a highly anharmonic character (e.g. rotamer jumps or sub-domain opening) that manifests in discrete but weaker density around distinct positions with no continuous density connecting the states. This type of heterogeneity is not well fit by B-factors, which leads to underestimating the displacements present in the experimental ensemble[19,20]. To overcome this limitation in the PDB format, atoms can be replicated and labeled with an "alternative location indicator (altloc)," signifying discrete states. Refinement and validation programs treat atoms sharing the same altloc as having the ability to interact with each other and with atoms lacking an altloc, but not with atoms with different altlocs. The lack of a hierarchical relationship between altlocs restricts the complexity of information encoded by the legacy format.

Capturing ensemble information in the legacy PDB format becomes an even more complex problem when considering compositional heterogeneity, which can coexist with conformational heterogeneity. Compositional heterogeneity is often observed with ligands bound at sub-stoichiometric occupancy in X-ray structures[21–23] and with different components in large macromolecular complexes in CryoEM[24]. Compositional heterogeneity is captured using the same “altloc” column as conformational heterogeneity. This ambiguous representation inhibits disentangling compositional and conformational heterogeneity, especially for large data mining efforts.

Computational tools have recently improved at decoding the complex conformational and compositional heterogeneity signal from noise. In CryoEM, human intervention or machine-learning tools can distinguish different large conformations. While many of these tools are primarily used for visualization, some can incorporate discrete states into heterogeneous refinement, moving towards ensemble-based CyroEM models[25–27]. In X-ray crystallography and CryoEM, methods exist that automatically detect subtle conformational shifts, like rotamer jumps, among structural ensemble members through multiconformer approaches[8–10,28,29]. Further, weak signal representing compositional heterogeneity, often seen in X-ray ligand soaking experiments, can now be more easily identified using approaches such as PanDDA[30]. In cryoEM, compositionally heterogeneous models are created by exploring differences in the same or related maps[31].

However, these tools are confined by the data structure that must represent their output in the Protein Data Bank. Failing to account for the diverse conformational and compositional states hinders a thorough understanding of biological functions, the precision of predictive modeling, and the innovation in designing novel proteins and small molecule inhibitors.

Alterations to the existing mmCIF Format can capture conformational and compositional heterogeneity in structures

The PDBx/mmCIF is an extensible data representation built on a flexible dictionary-based system[32,33]. While this data format allows for a more robust representation of many structural models, conformational and compositional heterogeneity is encoded in the same way as the legacy PDB format (altlocs and B-factors). We propose extending the PDBx/mmCIF model to include new conformational and compositional data items linked to atom-level data (**Figure 1**). Additionally, both data items would contain layers. For each atom, the first conformational state would represent the base or first layer of heterogeneity, with subsequent states explaining heterogeneity 'within' the previous state. For example, in the conformational data item, conformational state 1 could contain information on a loop state (in or out), whereas conformational state 2 would contain a peptide flip in the backbone within the out loop state, and conformational state 3 could contain information about a side chain alternative conformation that occurs in a residue in the out loop state and with the peptide flip. This layering would allow for the knowledge of hierarchical conformational heterogeneity. Separating compositional and conformational heterogeneity into individual data items allows us to understand how they are linked, such as multiple conformations populated in the liganded state. Importantly, the inherent flexibility of the mmCIF format is highlighted by the fact that many dictionary items are optional[34], paving the way for a standardized and adaptable format to capture conformational and compositional heterogeneity depending on the experimental data.

In the following sections, we present various examples that contrast the current representation in the mmCIF format (most of which are holdovers from the legacy PDB format) versus our envisioned depiction. We discuss how these changes can be integrated with refinement protocols. This new format would also facilitate a more descriptive representation of conformational heterogeneity. For example, instead of specifying alternative conformer A for a series of residues, different conformations can be encoded as 'Loop Out' or 'Ligand Bound'.

Example 1: simple conformational heterogeneity

The simplest example is an apo protein with alternative conformations of single residues or sections of residues. Currently, these could be captured by alternative conformations or increased B-factors. In our proposed mmCIF format, the first conformational data item would represent the alternative positions of an individual residue (**Figure 1A**) or multiple residues, such as a loop (**Figure 1B**). In this scenario, refinement software would work exactly as it is now by restraining each atom's occupancy to sum to one. We also propose that 'altloc' in the current mmCIF format be moved into the conformational state 1 data item. Further, existing ensemble structures could use this data format by having each model in the ensemble have a different identifier in the conformational state 1 data item.

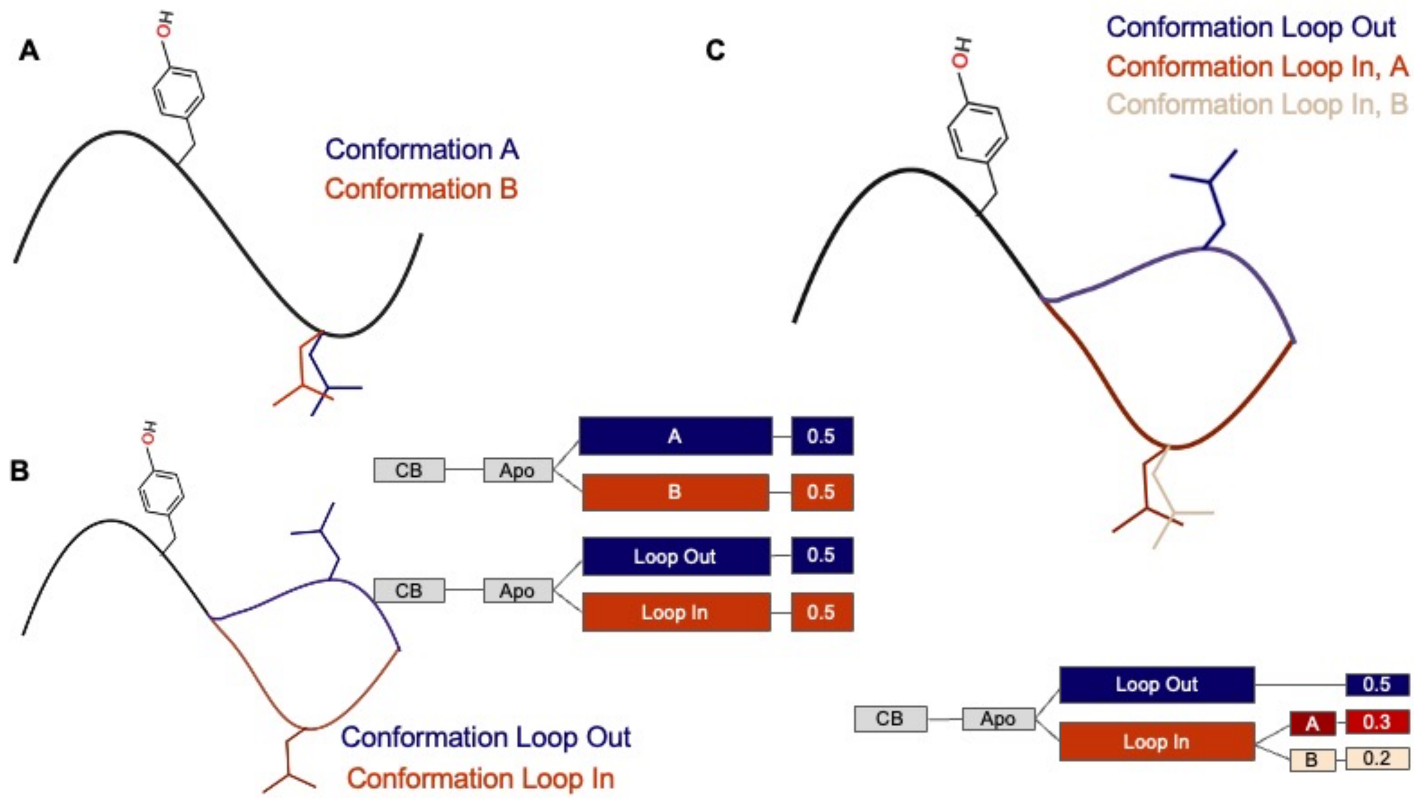

Figure 2: Conformational Heterogeneity. Boxes represent how data items would be connected in the mmCIF format. Shown are: atom, compositional state, conformational state 1, conformational state 2, occupancy. **A.** Example 1: Simple conformational heterogeneity with single residue. **B.** Example 1: Simple conformational heterogeneity with loop. **C.** Example 2: Layer conformational heterogeneity

Example 2: layered conformational heterogeneity

Next, a more complicated scenario with the apo protein to demonstrate hierarchically related conformational heterogeneity, such as an alternative conformation within a loop. For example, in the 'loop out' conformation (state 1), there are multiple positions of a single leucine side chain (state 2) (**Figure 2C**). In the current mmCIF format, you could encode the three conformations as A (loop in), B (loop out, position 1), C (loop out, position 2), but this would have no descriptive or hierarchical relationship to each other. In our proposed mmCIF additions, we could encode the loop conformational heterogeneity in the conformational state 1 and the leucine conformational heterogeneity within the conformational state 2. For refinement, restraints would be linked to each level of conformation, such that all conformations in loop in and loop out would have to equal 1, while the nested conformations in loop out would have to sum to an occupancy of 1-loop in. Clashes could be evaluated in PDB deposition validation by all atoms with the same label at the same hierarchy level, extending the current validation scheme[35].

Example 3: simple compositional heterogeneity

Next, we have a protein with a ligand that is partially occupied (i.e. present in 50% of the protein copies) in a space that does not clash with the apo state of the protein. Compositional heterogeneity is almost always observed in high throughput ligand soaking experiments, which now comprise a huge percentage of the PDB depositions[36–38]. However, representing this data has been a topic of great debate[39,40]. In the legacy PDB format, we could encode the compositional heterogeneity by indicating that the small molecule has an occupancy of 0.5 (50%) (**Figure 3A**). However, in the proposed format, we would indicate that this was the 'bound' state in the compositional heterogeneity data type (in addition to the 50% occupancy). Due to the lack of clashes, the compositional state of the protein would be left blank, allowing validation to occur against all compositional states.

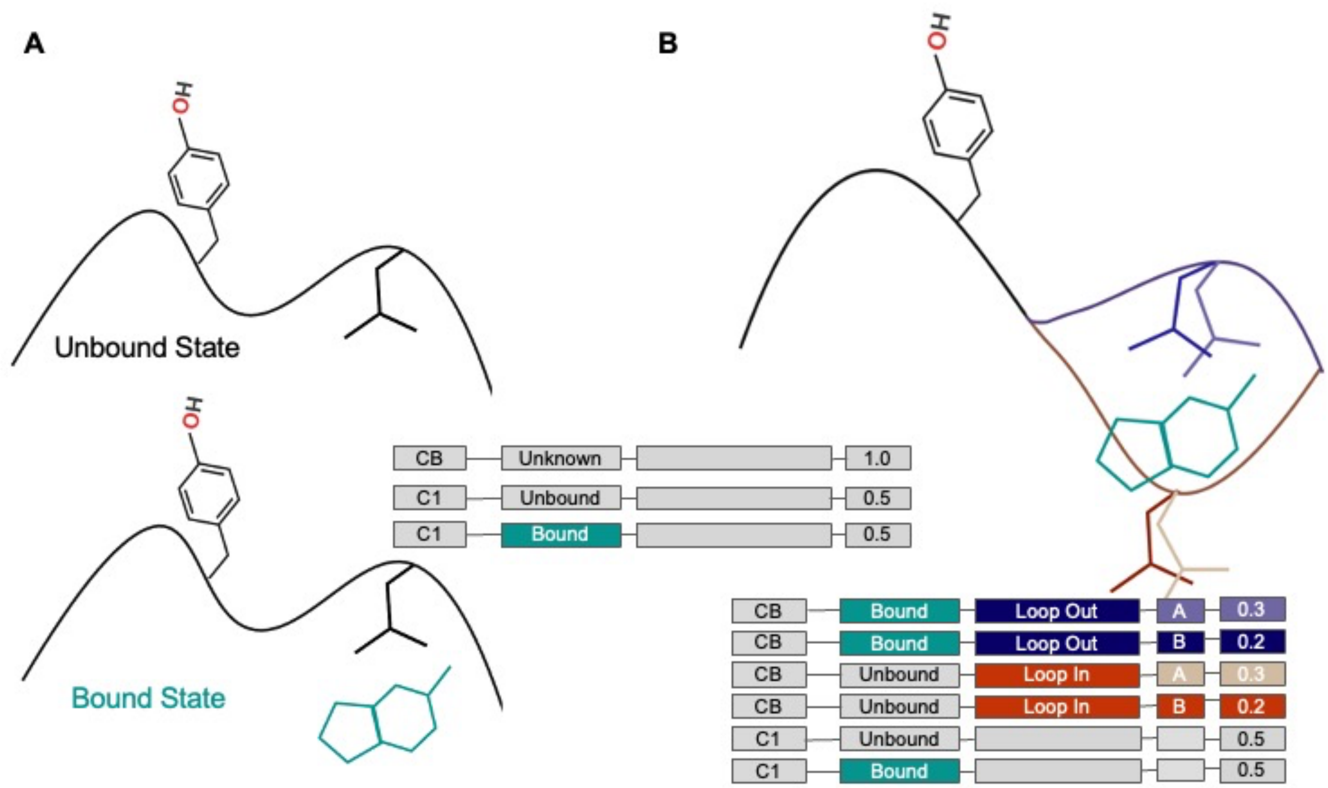

Figure 3: Compositional Heterogeneity. Boxes represent how data items would be connected in the mmCIF format. Shown are: atom, compositional state, conformational state 1, conformational state 2, occupancy. **A.** Example 3: Simple compositional heterogeneity. **B.** Example 4: Compositional and conformational heterogeneity.

Example 4: compositional and conformational heterogeneity

We now consider building on this example of a partially occupied ligand with conformational heterogeneity. The interplay between conformational and compositional heterogeneity is often inferred from the ligand clashing with some protein conformations (**Figure 3B**). In this example, the ligand binding is compatible with the "loop in", but not the "loop out" conformation. In addition, there are individual residues in each loop state. The complexity of this interlinked conformational and compositional heterogeneity would be completely lost in the legacy PDB format. We would have to encode the conformational heterogeneity of the protein with at least four altlocs ids and make copies of the ligand that match the altloc ids of the compatible conformations. There would be no link between how the conformational heterogeneity interacts with the compositional heterogeneity. Furthermore, the hierarchy of conformational heterogeneity is also lost.

In our proposed mmCIF model, we would encode compositional heterogeneity (bound or unbound) in one column and conformational heterogeneity in another. For example, when the compositional column indicated a 'bound state', the corresponding conformational state 1 would indicate the loop out, and then the conformational state 2 could indicate the alternative conformations of the leucine residue. The occupancies of the conformations in the bound state would be restrained to sum to the bound occupancy. For residues that do not interact with the ligand, we would imagine that the compositional column would be blank or unknown. This concept can be extended to subunits in assemblies from CryoEM data or covalent linkages, such as a post-translational modification.

Conclusions

Methods to predict the single structure representations in the PDB have been a breakthrough for structural biology[41,42]. However, the next challenge lies in predicting ensembles. This is important for two reasons; first, ensembles dictate function, second, the accuracy gap between prediction methods and experiments may result from an incomplete consideration of ensembles on both sides[11]. A substantial upgrade in representing our experimental structural data is needed to meet this challenge.

Our proposed amendments to the PDBx/mmCIF model aim to enhance conformational and compositional heterogeneity representation, moving the encoding of the data closer to the underlying experimental ensemble. This improved data structure will accelerate the development of new tools and create representative training datasets for structural ensemble prediction. Alongside this format, infrastructure changes to refinement, visualization, and validation tools are likely needed. This new format should also help interconvert existing multiconformer and ensemble-based models. Such interconversion enables different manual manipulations, such as in Coot[43], or data mining approaches.

We envision refining mmCIF further to more effectively correlate grouped data, including time-resolved techniques, ligand soaking experiments, or EM classification/reconstructions[25,44]. For example, a ‘perturbation’ data item could connect to specific structure factors or real-space maps, enabling restrained refinements of coordinates.

The notion that a single, static structure defines a protein is outdated for experimental and structural prediction. Macromolecules adopt an ensemble of conformations, and modeling those structural distributions accurately is now possible. By more correctly encapsulating the underlying experimental data, we can enable both benchmarks for prediction and a new class of “ensemble-function” studies. Moreover, accurately modeling compositional heterogeneity will reveal how ligands interact with the receptors, increasing the potential for an “Alphafold” breakthrough in ligand design. Inevitably, all models are wrong, but we can get closer to useful by taking advantage of the expressive mmCIF format to better model the heterogeneity in the underlying experimental data.

1. Burley, S. K. *et al.* Protein Data Bank (PDB): The Single Global Macromolecular Structure Archive. *Methods Mol. Biol.* **1607**, 627–641 (2017).
2. Furnham, N., Blundell, T. L., DePristo, M. A. & Terwilliger, T. C. Is one solution good enough? *Nat. Struct. Mol. Biol.* **13**, 184–5; discussion 185 (2006).
3. Holton, J. M., Classen, S., Frankel, K. A. & Tainer, J. A. The R-factor gap in macromolecular crystallography: an untapped potential for insights on accurate structures. *FEBS J.* **281**, 4046–4060 (2014).
4. Westbrook, J. D. *et al.* PDBx/mmCIF Ecosystem: Foundational Semantic Tools for Structural Biology. *J. Mol. Biol.* **434**, 167599 (2022).
5. Adams, P. D. *et al.* Announcing mandatory submission of PDBx/mmCIF format files for crystallographic depositions to the Protein Data Bank (PDB). *Acta Crystallogr D Struct Biol* **75**, 451–454 (2019).
6. Rieping, W., Habeck, M. & Nilges, M. Inferential structure determination. *Science* **309**, 303–306 (2005).
7. Burnley, B. T., Afonine, P. V., Adams, P. D. & Gros, P. Modelling dynamics in protein crystal structures by ensemble refinement. *Elife* **1**, e00311 (2012).
8. Wankowicz, S. A. *et al.* Uncovering Protein Ensembles: Automated Multiconformer Model Building for X-ray Crystallography and Cryo-EM. *bioRxiv* (2023) doi:10.1101/2023.06.28.546963.
9. Riley, B. T. *et al.* qFit 3: Protein and ligand multiconformer modeling for X-ray crystallographic and single-particle cryo-EM density maps. *Protein Sci.* **30**, 270–285 (2021).
10. Stachowski, T. R. & Fischer, M. FLEXR: automated multi-conformer model building using electron-density map sampling. *Acta Crystallogr D Struct Biol* **79**, 354–367 (2023).
11. Lane, T. J. Protein structure prediction has reached the single-structure frontier. *Nat. Methods* **20**, 170–173 (2023).

12. Weichenberger, C. X., Afonine, P. V., Kantardjieff, K. & Rupp, B. The solvent component of macromolecular crystals. *Acta Crystallogr. D Biol. Crystallogr.* **71**, 1023–1038 (2015).

13. Karplus, P. A. & Diederichs, K. Linking crystallographic model and data quality. *Science* **336**, 1030–1033 (2012).

14. Glaeser, R. M. How Good Can Single-Particle Cryo-EM Become? What Remains Before It Approaches Its Physical Limits? *Annu. Rev. Biophys.* **48**, 45–61 (2019).

15. Winn, M. D., Isupov, M. N. & Murshudov, G. N. Use of TLS parameters to model anisotropic displacements in macromolecular refinement. *Acta Crystallogr. D Biol. Crystallogr.* **57**, 122–133 (2001).

16. Afonine, P. V., Adams, P. D. & Urzhumtsev, A. From deep TLS validation to ensembles of atomic models built from elemental motions. II. Analysis of TLS refinement results by explicit interpretation. *Acta Crystallogr D Struct Biol* **74**, 621–631 (2018).

17. Konnert, J. H. & Hendrickson, W. A. A restrained-parameter thermal-factor refinement procedure. *Acta Crystallogr. A* **36**, 344–350 (1980).

18. O'Connor, D. A. Thermal Vibrations in Crystallography. *Physician's Bull.* **26**, 498–499 (1975).

19. 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).

20. Kuriyan, J., Petsko, G. A., Levy, R. M. & Karplus, M. Effect of anisotropy and anharmonicity on protein crystallographic refinement. An evaluation by molecular dynamics. *J. Mol. Biol.* **190**, 227–254 (1986).

21. Danley, D. E. Crystallization to obtain protein-ligand complexes for structure-aided drug design. *Acta Crystallogr. D Biol. Crystallogr.* **62**, 569–575 (2006).

22. Turnbull, A. P. & Emsley, P. Studying protein-ligand interactions using X-ray crystallography. *Methods Mol. Biol.* **1008**, 457–477 (2013).

23. Müller, I. Guidelines for the successful generation of protein-ligand complex crystals. *Acta*

Crystallogr D Struct Biol **73**, 79–92 (2017).

24. Webster, S. M., May, M. B., Powell, B. M. & Davis, J. H. Imaging structurally dynamic ribosomes with cryogenic electron microscopy. *ArXiv* (2023).
25. Zhong, E. D., Bepler, T., Berger, B. & Davis, J. H. CryoDRGN: reconstruction of heterogeneous cryo-EM structures using neural networks. *Nat. Methods* **18**, 176–185 (2021).
26. Punjani, A. & Fleet, D. J. 3D variability analysis: Resolving continuous flexibility and discrete heterogeneity from single particle cryo-EM. *J. Struct. Biol.* **213**, 107702 (2021).
27. Serna, M. Hands on Methods for High Resolution Cryo-Electron Microscopy Structures of Heterogeneous Macromolecular Complexes. *Front Mol Biosci* **6**, 33 (2019).
28. Keedy, D. A., Fraser, J. S. & van den Bedem, H. Exposing Hidden Alternative Backbone Conformations in X-ray Crystallography Using qFit. *PLoS Comput. Biol.* **11**, e1004507 (2015).
29. Ginn, H. M. Vagabond: bond-based parametrization reduces overfitting for refinement of proteins. *Acta Crystallogr D Struct Biol* **77**, 424–437 (2021).
30. Pearce, N. M. *et al.* A multi-crystal method for extracting obscured crystallographic states from conventionally uninterpretable electron density. *Nat. Commun.* **8**, 15123 (2017).
31. Punjani, A. & Fleet, D. J. 3DFlex: determining structure and motion of flexible proteins from cryo-EM. *Nat. Methods* **20**, 860–870 (2023).
32. Bourne, P. E. *et al.* Macromolecular Crystallographic Information File. *Methods Enzymol.* **277**, 571–590 (1997).
33. Westbrook, J.D., 1997. Solvation effects on electronically excited states and a dictionary description language for macromolecular structure applications.
34. Westbrook, J.D., Berman, H.M. and Hall, S.R., 2006. Specification of a relational dictionary definition language (DDL2). International Tables for Crystallography, pp.61-70.
35. Read, R. J. *et al.* A new generation of crystallographic validation tools for the protein data bank. *Structure* **19**, 1395–1412 (2011).

36. Gahbauer, S. *et al.* Iterative computational design and crystallographic screening identifies potent inhibitors targeting the Nsp3 macrodomain of SARS-CoV-2. *Proc. Natl. Acad. Sci. U. S. A.* **120**, e2212931120 (2023).

37. Barthel, T., Wollenhaupt, J., Lima, G. M. A., Wahl, M. C. & Weiss, M. S. Large-Scale Crystallographic Fragment Screening Expedites Compound Optimization and Identifies Putative Protein-Protein Interaction Sites. *J. Med. Chem.* **65**, 14630–14641 (2022).

38. Skaist Mehlman, T. *et al.* Room-temperature crystallography reveals altered binding of small-molecule fragments to PTP1B. *Elife* **12**, (2023).

39. Weiss, M. S. *et al.* Of problems and opportunities-How to treat and how to not treat crystallographic fragment screening data. *Protein Sci.* **31**, e4391 (2022).

40. Jaskolski, M., Wlodawer, A., Dauter, Z., Minor, W. & Rupp, B. Group depositions to the Protein Data Bank need adequate presentation and different archiving protocol. *Protein Sci.* **31**, 784–786 (2022).

41. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

42. Baek, M. *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).

43. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 486–501 (2010).

44. Zhong, E. D., Lerer, A., Davis, J. H. & Berger, B. CryoDRGN2: Ab initio neural reconstruction of 3D protein structures from real cryo-EM images. in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (IEEE, 2021). doi:10.1109/iccv48922.2021.00403.