Recovering Conformational Heterogeneity from the Protein Data Bank at Scale

An Ensemble Dataset of over 60,000 Structures

Published Aug 10, 2026

Radial

DOI: 10.82153/pff0-ck46

Purpose

Proteins are dynamic ensembles of interconverting conformations. This ensemble, rather than any single structure, governs functions such as catalysis, mutational effects, and molecular recognition. Predicting protein conformational ensembles is a central goal of structural biology, but progress is constrained by a shortage of training data. Current ensemble predictors rely on molecular dynamics simulations, which are limited in number and constrained by force-field accuracy and accessible timescales. Experimental structural data is an untapped source of alternative data for training ensemble predictors of structure. X-ray crystallography and cryo-EM measure many copies of a protein and average over space and time, so the data encodes an ensemble of states. Conventional refinement collapses this signal into a single set of coordinates. Most Protein Data Bank (PDB) depositions, therefore, report a single averaged structure, leaving the underlying heterogeneity unmodeled and hidden in the experimental data. Here, we applied qFit to recover this latent heterogeneous signal at scale. Starting from structures resolved to better than 2 Å with deposited structure factors, we re-refined them and ran qFit multiconformer modeling. This produced over 60,000 completed multiconformer models, the largest dataset of experimentally derived ensemble protein structural models to date, spanning a broad range of sequence and structural diversity. 83.7% of structures have a lower R_{free} with qFit multiconformer models compared to the deposited, re-refined model. These models identified widespread side-chain heterogeneity absent from the deposited models. This resource aims to help address the data bottleneck in ensemble

prediction and reframes the experimental data deposited in the PDB as a source of ensemble information.

Background

Static structural prediction has been transformed by machine-learning predictors [1] [2] [3] [4] [5] [6] [7]. However, proteins are not static. They exist as dynamic ensembles of interconverting conformations [8]. This ensemble, not any single structure, drives biological function, governing catalysis, mutational effects, and molecular recognition [9] [10].

Predicting an ensemble is significantly harder than predicting a single structure, as the target is a distribution of weighted states rather than a single set of coordinates [11]. Beyond the significantly more difficult learning problem, the training data needed to learn that distribution is scarce. Current ensemble structural predictors are trained on and benchmarked against molecular dynamics (MD) data [12] [13] [14] [3]. However, simulations are constrained by force field accuracy, accessible timescales, and the diversity of the training set [15] [16] [17]. Accurately predicting conformational ensembles will require far more training data.

One untapped resource for training data is the experimental data used to derive static structures. X-ray crystallography and cryo-electron microscopy (cryo-EM) measure tens of thousands to billions of copies of the protein. These measurements are then ensemble-averaged over both space and time, so the experimental data carries information about an ensemble of states. While crystal packing and cryogenic temperatures restrict the accessible conformational space, they do not eliminate heterogeneity [18]. Conventional modeling and refinement nonetheless collapse this signal into a single, approximate structure. As a result, most Protein Data Bank (PDB) depositions report a single, averaged set of coordinates [19] [20] [21]. These are the models that helped unlock static structure prediction. The unmodeled heterogeneity in the experimental data is itself a resource for ensembles. Recovering it would substantially increase the available ensemble training data.

However, identifying and modeling the underlying conformational heterogeneity in experimental data is difficult. It is typically done by hand, identifying often sub-Angstrom differences, using visualization tools such as Coot [22]. This is complicated by low density and noise from many sources, including crystal imperfections, radiation damage, and poor initial modeling [23] [24] [18]. Manual modeling would be incredibly challenging to scale across the PDB. To make multiconformer modeling routine and impartial, we previously developed qFit [25] [26] [27]. qFit takes a refined single-conformer structure and a high-resolution X-ray or cryo-EM real-space map as input. It then uses optimization algorithms to identify alternative protein [27] [25] or ligand [26] conformations, supplementing manual modeling.

qFit produces a multiconformer model [28]. A multiconformer model represents conformational heterogeneity within a single set of coordinates. It encodes alternative states locally as altloc-labeled conformations and assigns each a weight in the population proportional to its occupancy. Altlocs are added only where the density requires it, so an ordered region can be modeled by a single conformation, whereas a more heterogeneous area can be modeled by multiple conformations. In contrast, an ensemble model represents heterogeneity as multiple complete copies of the system. A region that adopts several conformations appears as several slightly different versions of itself, one per copy, and the population is read from the spread of the copies rather than from any single one. The number of copies is set by the refinement protocol rather than by the local complexity of the density, so even a well-ordered region is duplicated across every model [29].

Here, we ran qFit on almost 80,000 deposited PDB structures with resolution better than 2 Å and deposited structure factors. In most structures, qFit improves the fit to experimental data and identifies additional conformational heterogeneity latent in the deposited data. This produced the largest dataset of multiconformer protein structures assembled to date, helping address the data bottleneck that constrains ensemble prediction. Beyond ensemble prediction, it also provides a rich resource for systematic studies of how conformational heterogeneity affects allostery, macromolecular interactions, and mutations [30] [31] [32].

Data Generation

The initial dataset includes all structures from PDBRedo with a resolution of 2 Å or better (downloaded December 2024; n=80,876). Structures composed solely of nucleic acids were removed (n=1043). We then re-refined all models using *phenix.refine* with the parameters shown below [33]. This re-refined structure was used as the 'deposited' comparison. We then generated a composite omit map, which was used to run qFit (version 2025.3) with default parameters [28]. Finally, we ran the post-qFit refinement script using *phenix.refine.* The final refinement cycle is outlined below and matches the re-refined model. All code and parameters needed to reproduce this pipeline are available in the qFit repository: https://github.com/ExcitedStates/qfit-3.0. All re-refined models (PDB and CIF), qFit models (PDB and CIF), and FASTA files are included in the Zenodo deposition (https://zenodo.org/records/20801853).

```
phenix.refine \
"${pdb}.pdb" \
"${pdb}.mtz" \
"$ligand_cif" \
refinement.main.number_of_macro_cycles=5 \
refinement.main.nqh_flips=True \
refinement.refine.adp.individual.isotropic=all \
refinement.output.write_maps=False \
refinement.hydrogens.refine=riding \
refinement.main.ordered_solvent=True \
refinement.target_weights.optimize_xyz_weight=true \
refinement.target_weights.optimize_adp_weight=true \
ordered_solvent.mode=every_macro_cycle
```

Dataset Description

Of the 79,833 structures that entered the pipeline, 60,549 (85.9%) produced completed models. Among the completed structures, the median resolution was 1.70 Å (range: 0.5–2.0 Å). The median number of residues was 298 (range: 3–2,338) (Figure 1). The distribution of chain count per structure (range

). The re-refined models had a median R_{free} of 0.198 (range 0.06–0.71) and a median R_{work} of 0.17 (range 0.06–0.68). R_{free} rises modestly as resolution worsens (slope 0.062, R^2 = 0.22) (Figure 2). The median R_{free} of about 0.2 corresponds to a residual disagreement of roughly 20% between the experimental data and the model. This residual partially reflects the limitation of reducing ensemble-averaged data to a discrete model rather than experimental error **[34] [35]**. All models are located in the Zenodo deposition (https://zenodo.org/records/20801853).

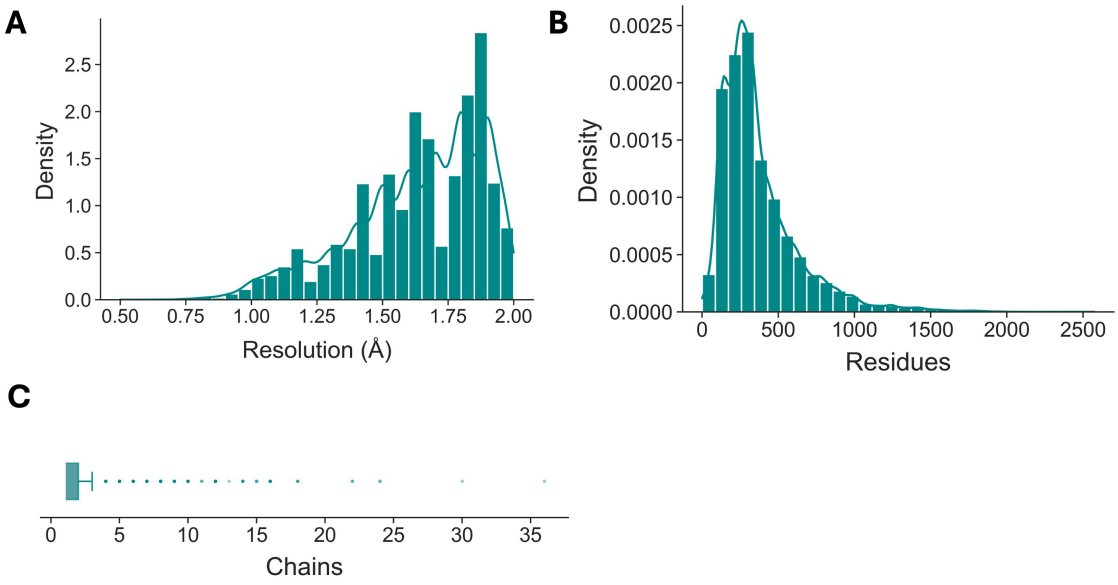

Figure 1.

Overview of the qFit multiconformer model dataset (n = 60,549). A. Distribution of resolutions; median resolution 1.70 Å (range 0.50–2.00 Å). B. Distribution of total residues per structure; median 298 residues (range 3–2,338). C. Distribution of chain count per structure (range).

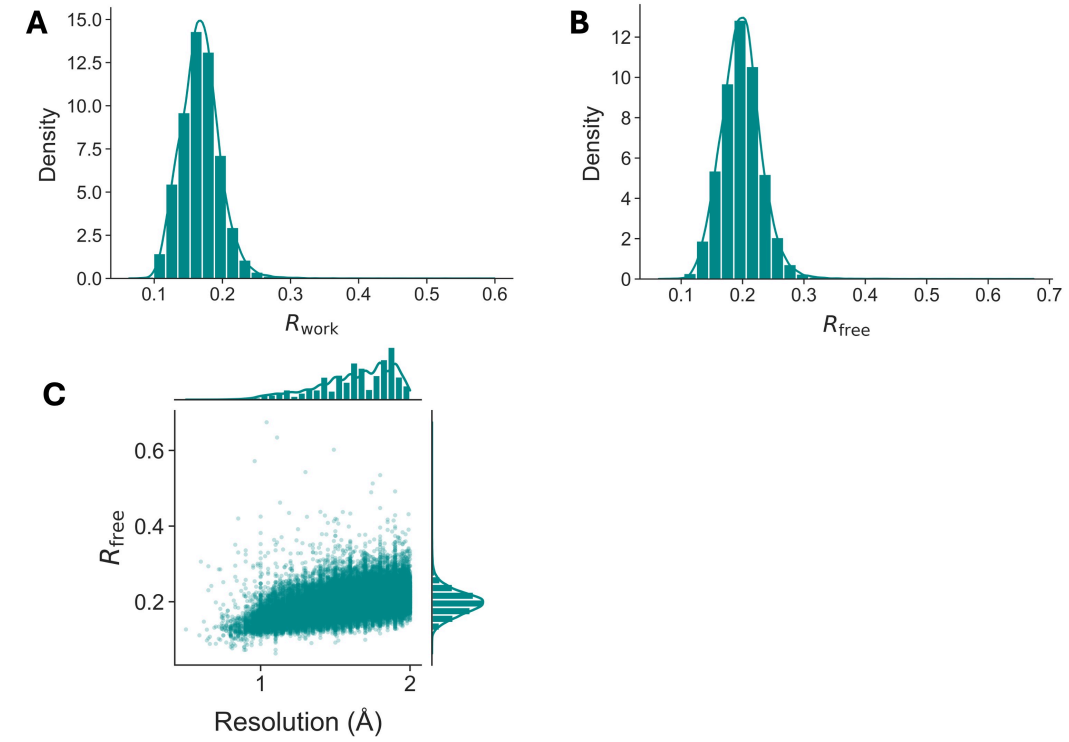

Figure 2. Data Description of qFit multiconformer models (n=60,549). A. Distribution of R_{work}. B. Distribution of R_{free}. C. Relationship between resolution and R_{free} (Slope: 0.062, R^2: 0.22).

To assess the dataset's diversity, we clustered the entries at the sequence and structural levels. Sequence clustering with MMseqs2 **[36]** yielded 16,133 clusters at 30% sequence identity, with many being singletons, and 25,608 clusters at 90% sequence identity. Structural clustering with Foldseek produced 6,590 clusters **[37]**.

Multiconformer models improve fit to experimental data

To assess whether multiconformer modeling improves the fit to the data, we compared the R_{free} of each qFit multiconformer model with that of the corresponding re-refined PDB-Redo model. Across 60,549 models, qFit lowered R_{free} by a mean of −0.009 and a median of −0.01 (Figure 3A/B). Individual outcomes vary widely, ranging from a 0.320 decrease to a 0.310 increase, with a standard deviation of 0.026. However, 54,274 (89.7%) structures have a lower R_{free} with qFit compared to the deposited, re-refined model. While the average

improvement in R_{free} is small in magnitude, it represents a real improvement in fit to the data through identifying and modeling conformational heterogeneity (see below). Further, the improvement in R_{free} is partly obscured by poor solvent fitting around heterogeneous models [38]. Of note, there is a subset of structures (1.3%; n=802) in which qFit yields significantly worse R_{free}, defined as an increase in R_{free} by over 0.05. Examining these structures, we could not find a pattern that explained this increase in R_{free}; this remains the subject of ongoing work.

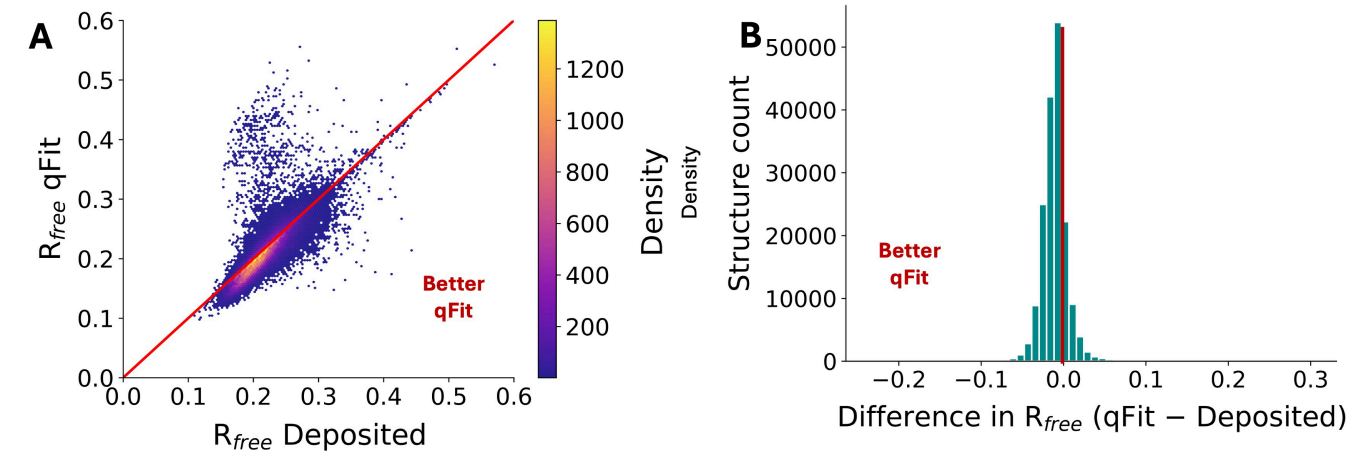

Figure 3. Comparison of R_{free} between deposited and qFit models. A lower R_{free} indicates a better fit to the experimental data. A. Scatterplot of deposited R_{free} versus qFit R_{free}. Points below the diagonal indicate the qFit model fits the data better. B. Histogram of the per-structure R_{free} difference (deposited minus qFit).

Multiconformer models have significantly more conformational heterogeneity compared to static structures

We quantified the additional heterogeneity in qFit models relative to deposited models using root-mean-square fluctuation (RMSF) and the number of alternate conformations (altlocs). RMSF measures the spatial spread of atomic positions across the modeled conformers and captures the magnitude of discrete displacement that qFit introduces through multiconformer modeling. The number of altlocs is the count of discrete alternate conformations assigned per residue

and provides a direct measure of how many distinct conformational states qFit resolves from the density.

The majority of residues had one altloc (90.4%) across deposited and qFit models (Figure 4A). qFit increased the altloc count for 9.4% of residues. Among residues that gained conformations, a single additional altloc was most common at 4.9%, followed by two at 3.6% and three at 0.8%. Alanine, glycine, and proline had the fewest additional altlocs, while leucine and cysteine had the most (Supplementary Figure 1). Only 0.2% of residues also had a reduction in the number of altlocs modeled.

Across the dataset, qFit increased RMSF by a mean of 0.13 Å relative to the deposited model, with a median difference of 0, consistent with the majority of residues not having an alternative conformer modeled (Figure 4B). However, a small subset of residues shows a large increase in RMSF. Unsurprisingly, the longer the amino acid, the larger the qFit RMSF, with lysine, arginine, glutamate, and glutamine showing the largest changes in RMSF (Supplementary Figure 2). As with the removal of altlocs, a minority of residues show reduced RMSF.

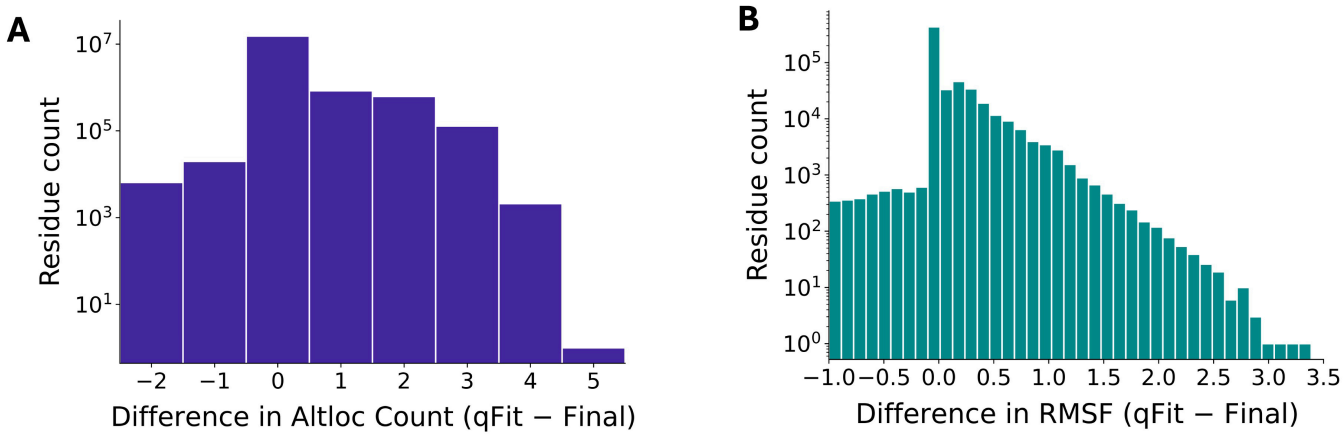

Figure 4. Per-residue changes in altloc count and RMSF between qFit and deposited models. A. Distribution of the per-residue difference in altloc count across all qFit models relative to deposited models. The y-axis shows the number of residues on a log scale. 9.4% of residues gained an altloc in the qFit model relative to the deposited model. B. Distribution of the per-residue difference in RMSF across all qFit models relative to deposited models. The y-axis shows the number of residues on a log scale. qFit increased RMSF by a mean of 0.13 Å relative to the deposited model. The median difference was 0 Å.

Hidden heterogeneity in a high-quality structure

Structures with excellent refinement statistics can still contain conformational heterogeneity that the deposited model does not capture. For example, the crystal structure of death-associated protein kinase 1 (DAPK1) in complex with resveratrol (PDB 7CCU; Figure 5A). This structure has a resolution of 1.65 Å, an R_{free} of 0.189, and an R_{work} of 0.172, about median values for our dataset. By these metrics, the model is of relatively high quality. However, the deposited model has no alternative conformations modeled. After applying the qFit pipeline, we lowered the R_{free} to 0.1812 and the R_{work} to 0.159, indicating a small but meaningful improvement in the fit to the scattering factors. Across the structure, we found that about 60% (164/277) of residues had at least one altloc, and 31% (85/277) had more than 2 altlocs. When removing altlocs that exist within the same rotamer well, we still see that 52% (144/277) and 17.3% (48/277) of residues had one or multiple altlocs. Figure 5B-D shows a visualization of altlocs identified in Phe240 (Figure 5B), Asp220 (Figure 5C), and Glu118 (Figure 5D). All of these had a single conformation modeled in the deposited structure, but clearly support multiple conformations.

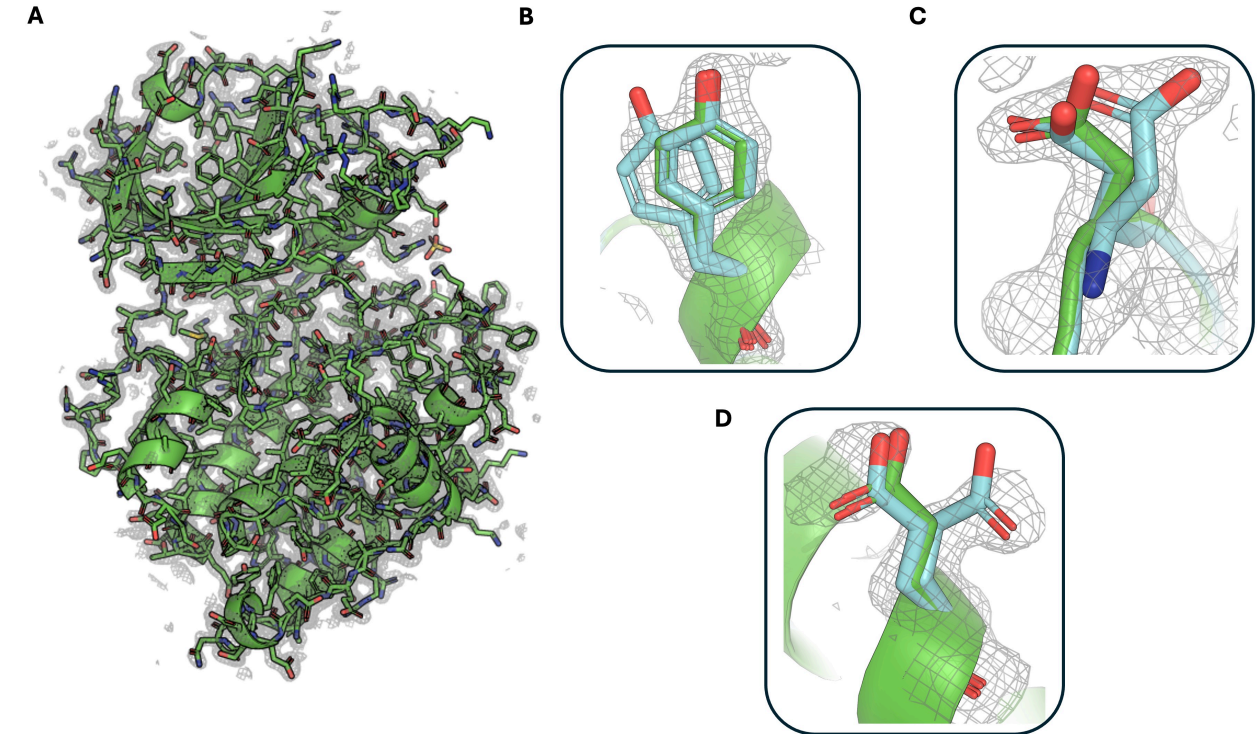

Figure 5. DAPK1 in complex with resveratrol (PDB 7CCU). The deposited model is shown in green and the qFit multiconformer model in blue. Electron density is contoured at 0.5σ (mesh). (A) Deposited model in electron density. (B) Phe240. (C) Asp220. (D) Glu118.

Related datasets of conformational heterogeneity

Several resources catalog the deposited alternative conformations. PDBFlex characterizes flexibility by comparing different PDB structures of the same protein. The Alternate Location Server maintains a list of PDB structures containing alternate locations, and a recent survey has assembled a custom dataset of alternately modeled backbone segments [39]. However, most deposited structures have no alternative conformations, and the difficulty of modeling them often leads to human bias, with some parts of the protein more likely to be modeled as multiconformers while other regions remain single-conformer models [40].

Beyond experimentally derived datasets, most datasets used for training ensemble predictors are MD datasets. The most widely used resources are ATLAS and mdCATH, which provide all-atom trajectories for a broad range of folded domains, spanning 1,390 protein chains in ATLAS and 5,398 CATH domains in

mdCATH [41] [42]. Other resources include MISATO, which simulates approximately 20,000 protein-ligand complexes drawn from PDBbind and focuses on common drug targets [43]. GPCRmd provides MD on GPCRs, with most systems belonging to class A [44]. IDRome covers many disordered regions of the human proteome using coarse-grained simulations [45].

There are a few key differences between the derivatives of these MD datasets and our qFit multiconformer model dataset. MD produces an explicit, time-ordered trajectory. It resolves motions from femtosecond-scale bond vibrations upward but is limited by the simulation length and force-field accuracy. For example, the ATLAS database, which has been used by many ensemble prediction algorithms, provides three 100 ns trajectories per protein [41]. Many functionally important motions, including loop rearrangements and allosteric transitions, occur on microsecond to millisecond timescales and are undersampled at these lengths [46]. In contrast, X-ray crystallographic measurements are inherently time-agnostic, capturing conformational changes regardless of the timescale, provided they occur within the crystal. A single X-ray experimental dataset reports the equilibrium distribution of conformations populated across the crystal, averaged over time and over all unit cells. However, there is no time information, and conformations are biased by the crystalline environment. Further, this data remains latent in the experimental data until modeled.

Impact

X-ray crystallography and cryo-EM data encode rich ensemble information that conventional modeling pipelines largely discard. Beyond the impact of understanding individual proteins' functions, this leaves valuable training data for ensemble methods unused. Here, we applied qFit across high-resolution structures and experimental data in the PDB to recover this heterogeneity. We show that qFit recovers the widespread conformational heterogeneity present in high-resolution X-ray structure factors but absent from the deposited models [28] [26]. The resulting multiconformer models more faithfully represent the true underlying ensemble, as shown by improved R_{free} values. This yields the largest dataset of multiconformer protein models to date. It offers experimentally derived

protein ensemble information at scale, a significant addition to the largely MD-derived datasets used for ensemble training [41] [42] [43] [44]. This foundational resource opens new avenues for studying protein ensembles and advancing computational methods for their prediction [11].

Beyond ensemble training, it is possible that this resource may help the accuracy gap between prediction methods and experiments may result from an incomplete consideration of ensembles [21]. All public predictors are trained on single-conformer models [2] [3] [6] [7]. When a residue populates multiple states, modeling it as a single conformer introduces strain and other local geometric artifacts [28]. Predictors trained on these structures may learn the distortions along with the correct geometry. Beyond incorrect geometry, it is possible that ensemble models will help with “the last angstrom problem”, the gap between the accurate backbones that current predictors produce and the sub-angstrom side-chain and bond geometry needed for ligand docking and mechanistic interpretation [47]. The last angstrom problem may be due to the precision with which the algorithms are built and to underlying uncertainty, but it is also possible that the poor precision reflects the statistical distribution of structural models. Additionally, this data can also help with the analysis and benchmarking of MD data.

More broadly, the dataset supports a shift toward dynamic structural biology. It reframes the PDB as a source of ensemble information. In the vast majority of cases, experimental structural biology data are too rich to be described by a single conformation. Making heterogeneous multiconformer models available at scale lowers the barrier to integrating dynamics into routine structural analysis. This opens up rich possibilities for systematic study across a range of biological phenomena through the lens of conformational ensembles.

Several dataset limitations warrant consideration. First, structures are restricted to those resolved to better than 2 Å with available structure factors, limiting our dataset and hiding heterogeneity in lower-resolution structures. Second, the dataset inherits conformational constraints imposed by crystal packing and cryogenic temperatures in most structures. Third, the conformational heterogeneity identified here arises predominantly from side-chain rather than backbone movements. Beyond these specific constraints, it is important to recognize that structures are models rather than experimental reality [34]. Our

multiconformer models improve fit to the experimental data, as shown by improved R_{free} values, but bridging the remaining gap between deposited models and the full information content of experimental observations remains an important challenge for the field.

Future work will focus on improving these algorithmic approaches and integrating them with guidance methods [48] [49] [50] [51]. This helps establish a self-reinforcing loop (Figure 6). Training data derived from qFit can improve the ability to predict a larger portion of the conformational ensemble. Using these improved structure predictors, we can then use them alongside guidance frameworks, enabling the discovery of richer heterogeneity, which in turn generates higher-quality ensemble training data for subsequent prediction algorithms.

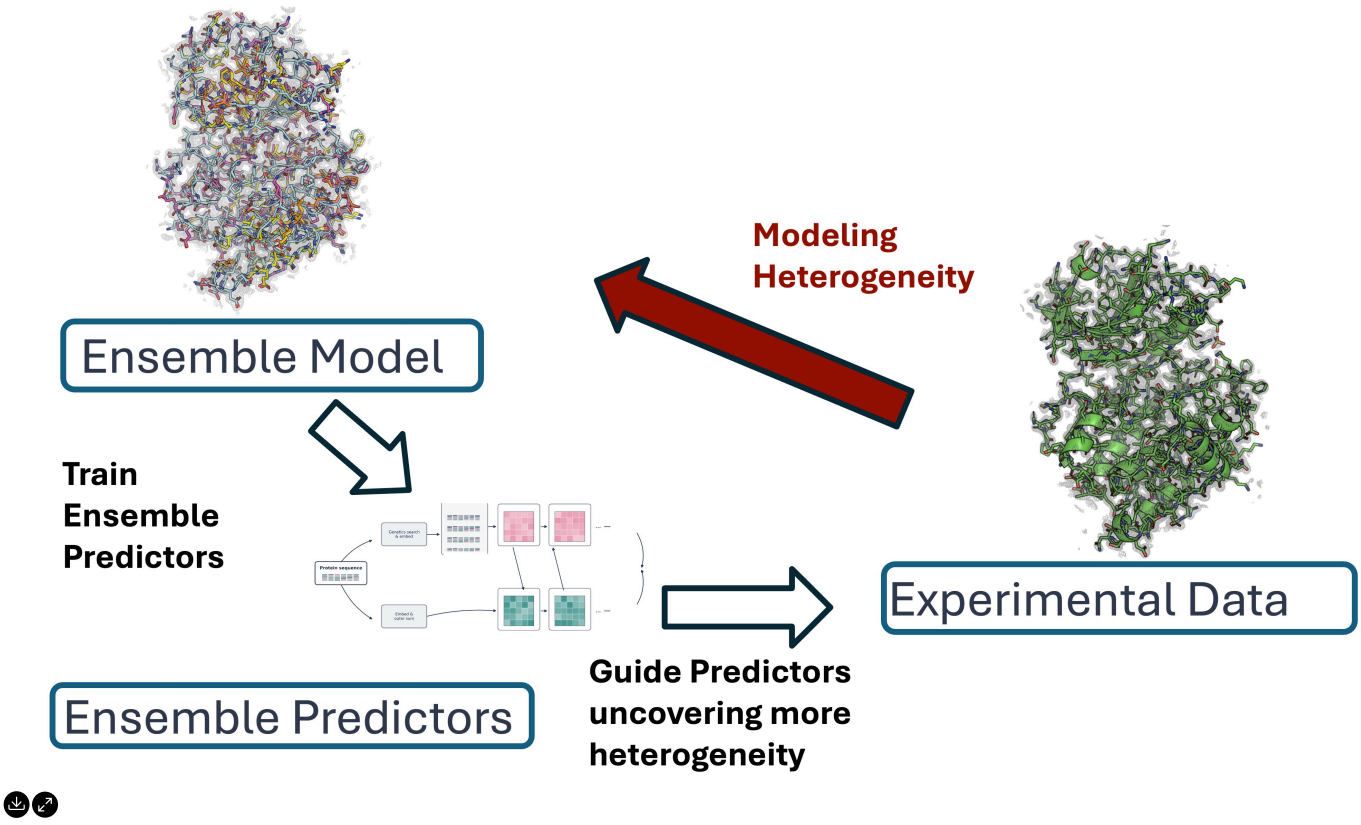

Figure 6. A self-reinforcing loop for conformational ensemble prediction. Training data derived from qFit improves structure predictors' ability to capture conformational ensembles (modeling heterogeneity). These improved predictors, combined with guidance frameworks, enable discovery of richer heterogeneity. This richer heterogeneity, in turn, generates higher-quality training data for the next generation of predictive algorithms.

Realizing this loop also depends on how we encode this information. The experimental data capture conformational heterogeneity spanning length scales

from single atoms to loop movements. There is also considerable compositional heterogeneity within experimental data. While qFit and a growing number of approaches can model this complex heterogeneity landscape, the current PDBx/mmCIF format does not support its description or encoding well [52]. This means that even when heterogeneity is modeled, much of it cannot be communicated through a model. This limits the ensemble information available to describe biological phenomena and the prediction methods used for training.

Finally, closing the loop will also require new infrastructure, including tools to share, query, and compare models, built around a living database that treats experimental data as ground truth and continuously incorporates modeling improvements as they emerge. The circular prediction model we present above depends on capturing this information at each pass (Figure 6). As ensemble-based methods become more accurate and computationally accessible, a unified and continually updated repository would let those improvements propagate easily to downstream users, enabling new biological insights rather than locking them behind the original depositors' modeling choices.

Acknowledgements

This material is based upon work supported by the Defense Advanced Research Projects Agency under this Agreement (SAW received research support). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. This work leveraged the Research Software Engineer Services provided by the Vanderbilt Advanced Computing Center for Research and Education (ACCRE), operated by and for Vanderbilt faculty.

▶

Supplementary Figures

Contributors (A–Z)

- **Stephanie A Wankowicz**: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Writing

References

1. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D. (2021). Highly accurate protein structure prediction with AlphaFold. https://doi.org/10.1038/s41586-021-03819-2
2. Abramson J, Adler J, Dunger J, Evans R, Green T, Pritzel A, Ronneberger O, Willmore L, Ballard AJ, Bambrick J, Bodenstein SW, Evans DA, Hung C-C, O'Neill M, Reiman D, Tunyasuvunakool K, Wu Z, Žemgulytė A, Arvaniti E, Beattie C, Bertolli O, Bridgland A, Cherepanov A, Congreve M, Cowen-Rivers AI, Cowie A, Figurnov M, Fuchs FB, Gladman H, Jain R, Khan YA, Low CMR, Perlin K, Potapenko A, Savy P, Singh S, Stecula A, Thillaisundaram A, Tong C, Yakneen S, Zhong ED, Zielinski M, Žídek A, Bapst V, Kohli P, Jaderberg M, Hassabis D, Jumper JM. (2024). Accurate structure prediction of biomolecular interactions with AlphaFold 3. https://doi.org/10.1038/s41586-024-07487-w
3. Passaro S, Corso G, Wohlwend J, Reveiz M, Thaler S, Somnath VR, Getz N, Portnoi T, Roy J, Stark H, Kwabi-Addo D, Beaini D, Jaakkola T, Barzilay R. (2025). Boltz-2: Towards Accurate and Efficient Binding Affinity Prediction. https://doi.org/10.1101/2025.06.14.659707
4. Team P, Zhang Y, Gong C, Zhang H, Ma W, Liu Z, Chen X, Guan J, Wang L, Yang Y, Xia Y, Xiao W. (2026). Protenix-v1: Toward High-Accuracy Open-Source Biomolecular Structure Prediction. https://doi.org/10.64898/2026.02.05.703733
5. Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, Wang J, Cong Q, Kinch LN, Schaeffer RD, Millán C, Park H, Adams C, Glassman CR, DeGiovanni A, Pereira JH, Rodrigues AV, van Dijk AA, Ebrecht AC, Opperman DJ, Sagmeister T, Buhlheller C, Pavkov-Keller T, Rathinaswamy MK, Dalwadi U, Yip CK, Burke JE, Garcia KC, Grishin NV, Adams PD, Read RJ, Baker D. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. https://doi.org/10.1126/science.abj8754
6. Discovery C, Boitreaud J, Dent J, McPartlon M, Meier J, Reis V, Rogozhnikov A, Wu K. (2024). Chai-1: Decoding the molecular interactions of life. https://doi.org/10.1101/2024.10.10.615955

7. Team BAA, Chen X, Zhang Y, Lu C, Ma W, Guan J, Gong C, Yang J, Zhang H, Zhang K, Wu S, Zhou K, Yang Y, Liu Z, Wang L, Shi B, Shi S, Xiao W. (2025). Protenix - Advancing Structure Prediction Through a Comprehensive AlphaFold3 Reproduction. https://doi.org/10.1101/2025.01.08.631967
8. Hilser VJ, García-Moreno E. B, Oas TG, Kapp G, Whitten ST. (2006). A Statistical Thermodynamic Model of the Protein Ensemble. https://doi.org/10.1021/cr040423+
9. Hilser VJ, Wrabl JO, Millard CE, Schmitz A, Brantley SJ, Pearce M, Rehfus J, Russo MM, Voortman-Sheetz K. (2025). Statistical Thermodynamics of the Protein Ensemble: Mediating Function and Evolution. https://doi.org/10.1146/annurev-biophys-061824-104900
10. Boehr DD, Nussinov R, Wright PE. (2009). The role of dynamic conformational ensembles in biomolecular recognition. https://doi.org/10.1038/nchembio.232
11. Wankowicz SA, Bonomi M. (2026). From possibility to precision in macromolecular ensemble prediction. https://doi.org/10.1038/s41592-026-03084-z
12. Lewis S, Hempel T, Jiménez-Luna J, Gastegger M, Xie Y, Foong AYK, Satorras VG, Abdin O, Veeling BS, Zaporozhets I, Chen Y, Yang S, Foster AE, Schneuing A, Nigam J, Barbero F, Stimper V, Campbell A, Yim J, Lienen M, Shi Y, Zheng S, Schulz H, Munir U, Sordillo R, Tomioka R, Clementi C, Noé F. (2025). Scalable emulation of protein equilibrium ensembles with generative deep learning. https://doi.org/10.1126/science.adv9817
13. Jing B, Berger B, Jaakkola T. (2026). AI-based methods for simulating, sampling, and predicting protein ensembles. https://doi.org/10.1016/j.sbi.2026.103251
14. Jing B, Berger B, Jaakkola T. (2024). AlphaFold Meets Flow Matching for Generating Protein Ensembles. https://doi.org/10.48550/arxiv.2402.04845
15. Hollingsworth SA, Dror RO. (2018). Molecular Dynamics Simulation for All. https://doi.org/10.1016/j.neuron.2018.08.011
16. Robustelli P, Piana S, Shaw DE. (2018). Developing a molecular dynamics force field for both folded and disordered protein states. https://doi.org/10.1073/pnas.1800690115
17. Pedersen KB, Flores Canales JC, Schiøtt B. (2022). Predicting molecular properties of α synuclein using force fields for intrinsically disordered proteins. https://doi.org/10.1002/prot.26409
18. Karplus PA, Diederichs K. (2012). Linking Crystallographic Model and Data Quality. https://doi.org/10.1126/science.1218231

19. Burley SK, Berman HM. (2021). Open-access data: A cornerstone for artificial intelligence approaches to protein structure prediction. https://doi.org/10.1016/j.str.2021.04.010
20. Furnham N, Blundell TL, DePristo MA, Terwilliger TC. (2006). Is one solution good enough?. https://doi.org/10.1038/nsmb0306-184
21. Lane TJ. (2023). Protein structure prediction has reached the single-structure frontier. https://doi.org/10.1038/s41592-022-01760-4
22. Emsley P, Cowtan K. (2004). Coot: model-building tools for molecular graphics. https://doi.org/10.1107/s0907444904019158
23. Weichenberger CX, Afonine PV, Kantardjieff K, Rupp B. (2015). The solvent component of macromolecular crystals. https://doi.org/10.1107/s1399004715006045
24. Kabsch W. (2010). XDS. https://doi.org/10.1107/s0907444909047337
25. Keedy DA, Fraser JS, van den Bedem H. (2015). Exposing Hidden Alternative Backbone Conformations in X-ray Crystallography Using qFit. https://doi.org/10.1371/journal.pcbi.1004507
26. Riley BT, Wankowicz SA, de Oliveira SHP, van Zundert GCP, Hogan DW, Fraser JS, Keedy DA, van den Bedem H. (2020). qFit 3: Protein and ligand multiconformer modeling for X ray crystallographic and single particle cryo EM density maps. https://doi.org/10.1002/pro.4001
27. van den Bedem H, Dhanik A, Latombe J, Deacon AM. (2009). Modeling discrete heterogeneity in X-ray diffraction data by fitting multi-conformers. https://doi.org/10.1107/s0907444909030613
28. Wankowicz SA, Ravikumar A, Sharma S, Riley B, Raju A, Hogan DW, Flowers J, van den Bedem H, Keedy DA, Fraser JS. (2024). Automated multiconformer model building for X-ray crystallography and cryo-EM. https://doi.org/10.7554/elife.90606
29. Woldeyes RA, Sivak DA, Fraser JS. (2014). E pluribus unum, no more: from one crystal, many conformations. https://doi.org/10.1016/j.sbi.2014.07.005
30. Seo L, Farran I, Aslam A, Li X, Jaishankar P, Ashworth A, Fraser JS, Renslo AR, Wankowicz SA. (2025). Crystallographic Ensembles Reveal the Structural Basis of Binding Entropy in SARS-CoV2 Macrodomain. https://doi.org/10.1101/2025.11.25.690589
31. Miller CA, Wankowicz SA. (2026). Binding Entropy Can Be Predicted by Crystallographic Ensembles. https://doi.org/10.7554/elife.111298.1
32. Keedy DA. (2019). Journey to the center of the protein: allostery from multitemperature multiconformer X-ray crystallography.

https://doi.org/10.1107/s2059798318017941

33. Afonine PV, Grosse-Kunstleve RW, Echols N, Headd JJ, Moriarty NW, Mustyakimov M, Terwilliger TC, Urzhumtsev A, Zwart PH, Adams PD. (2012). Towards automated crystallographic structure refinement with phenix.refine. https://doi.org/10.1107/s0907444912001308
34. Fraser JS, Murcko MA. (2024). Structure is beauty, but not always truth. https://doi.org/10.1016/j.cell.2024.01.003
35. Holton JM, Classen S, Frankel KA, Tainer JA. (2014). The R factor gap in macromolecular crystallography: an untapped potential for insights on accurate structures. https://doi.org/10.1111/febs.12922
36. Steinegger M, Söding J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. https://doi.org/10.1038/nbt.3988
37. van Kempen M, Kim SS, Tumescheit C, Mirdita M, Lee J, Gilchrist CLM, Söding J, Steinegger M. (2023). Fast and accurate protein structure search with Foldseek. https://doi.org/10.1038/s41587-023-01773-0
38. Holton J, Wankowicz SA. (2026). Supplying a user-defined bulk solvent map for refinement. https://doi.org/10.82153/2me0-hd96
39. Hrabe T, Li Z, Sedova M, Rotkiewicz P, Jaroszewski L, Godzik A. (2015). PDBFlex: exploring flexibility in protein structures. https://doi.org/10.1093/nar/gkv1316
40. Wankowicz SA. (2024). Modeling Bias Toward Binding Sites in PDB Structural Models. https://doi.org/10.1101/2024.12.14.628518
41. Vander Meersche Y, Cretin G, Gheeraert A, Gelly J, Galochkina T. (2023). ATLAS: protein flexibility description from atomistic molecular dynamics simulations. https://doi.org/10.1093/nar/gkad1084
42. Mirarchi A, Giorgino T, De Fabritiis G. (2024). mdCATH: A Large-Scale MD Dataset for Data-Driven Computational Biophysics. https://doi.org/10.1038/s41597-024-04140-z
43. Siebenmorgen T, Menezes F, Benassou S, Merdivan E, Didi K, Mourão ASD, Kitel R, Liò P, Kesselheim S, Piraud M, Theis FJ, Sattler M, Popowicz GM. (2024). MISATO: machine learning dataset of protein–ligand complexes for structure-based drug discovery. https://doi.org/10.1038/s43588-024-00627-2
44. Rodríguez-Espigares I, Torrens-Fontanals M, Tiemann JKS, Aranda-García D, Ramírez-Anguita JM, Stepniewski TM, Worp N, Varela-Rial A, Morales-Pastor A, Medel-Lacruz B, Pándy-Szekeres G, Mayol E, Giorgino T, Carlsson J, Deupi X, Filipek S, Filizola M, Gómez-Tamayo JC, Gonzalez A, Gutiérrez-

de-Terán H, Jiménez-Rosés M, Jespers W, Kapla J, Khelashvili G, Kolb P, Latek D, Marti-Solano M, Matricon P, Matsoukas M, Miszta P, Olivella M, Perez-Benito L, Provasi D, Ríos S, R. Torrecillas I, Sallander J, Sztyler A, Vasile S, Weinstein H, Zachariae U, Hildebrand PW, De Fabritiis G, Sanz F, Gloriam DE, Cordomi A, Guixà-González R, Selent J. (2020). GPCRmd uncovers the dynamics of the 3D-GPCRome. https://doi.org/10.1038/s41592-020-0884-y

45. Cao F, von Bülow S, Tesei G, Lindorff Larsen K. (2024). A coarse grained model for disordered and multi domain proteins. https://doi.org/10.1002/pro.5172
46. Henzler-Wildman K, Kern D. (2007). Dynamic personalities of proteins. https://doi.org/10.1038/nature06522
47. Lyu N, Du S, Shao Q, Yang Z, Ma J, Herschlag D. (2025). Physics-Grounded Evaluation to Guide Accurate Biomolecular Prediction. https://doi.org/10.1101/2025.06.30.662466
48. Chrispens K, Collins M, Fraser JS, Mai D, van den Bedem H, Wankowicz SA. (2026). _sampleworks:_ A Modular Platform for Experimentally Guided Biomolecular Ensemble Generation. https://doi.org/10.82153/jkxj-tw08
49. Maddipatla A, Rzayev A, Pegoraro M, Pacesa M, Schanda P, Marx A, Vedula S, Bronstein AM. (2026). Inference-time optimization for experiment-grounded protein ensemble generation. https://doi.org/10.48550/arxiv.2602.24007
50. Raghu R, Levy A, Wetzstein G, Zhong ED. (2025). Multiscale guidance of protein structure prediction with heterogeneous cryo-EM data. https://doi.org/10.48550/arxiv.2506.04490
51. Fadini A, Li M, McCoy AJ, Banjara S, Okumura H, Napier E, Fontana P, Khan AR, Jovine L, Terwilliger TC, Read RJ, Hekstra DR, AlQuraishi M. (2026). AlphaFold as a prior: experimental structure determination conditioned on a pretrained neural network. https://doi.org/10.1038/s41592-026-03047-4
52. Wankowicz SA, Fraser JS. (2024). Comprehensive encoding of conformational and compositional protein structural ensembles through the mmCIF data structure. https://doi.org/10.1107/s2052252524005098