Modeling Bias Toward Binding Sites in PDB Structural Models

Stephanie A. Wankowicz[1*]

1) Molecular Physiology and Biophysics, Vanderbilt University, Nashville, TN; Center for Applied AI in Protein Dynamics, Vanderbilt University, Nashville, TN; Center for Structural Biology, Vanderbilt University, Nashville, TN

*Correspondence: stephanie@wankowiczlab.com

Abstract

The protein data bank (PDB) is one of the richest databases in biology. The structural models deposited have provided insights into protein folds, relationships to evolution, energy functions of structures, and most recently, protein structure prediction, connecting sequence to structure. However, the X-ray crystallography (and cryo-EM) models deposited in the PDB are determined by a combination of refinement algorithms and manual modeling. The intervention of human modeling leads to the possibility that within a single structure, there can be differences in how well parts of a structure are modeled and/or fit the underlying experimental data. We identified that small molecule binding sites are more carefully modeled and better match the underlying experimental data than the rest of the protein structural model. This trend persisted regardless of the resolution or Rfree values of the structure. The variation of modeling has implications for how we interpret protein structural models and use structural models as truth in mechanisms, structural bioinformatics, simulations, docking, and structure prediction, especially when comparing binding sites to the rest of the protein.

Introduction

Today, most protein machine learning algorithms rely on the accuracy of existing structures in the protein structural databank (PDB). However, models in the PDB vary in the goodness of fit for experimental data[1]. Global metrics, such as the crystallographic R-factor[2], are commonly used to determine the 'goodness' of a structural model [3]. While the R-value indicates how well a model matches the observed structure factors, it is not always effective in discriminating major errors in a model[4]. Additional metrics, such as geometry, can provide further information on the 'goodness' of a structure, but these still measure values on a global level[5]. While recent efforts have been made to support residue-level information on how well these individual residues fit experimental data[6], how these values change across areas of the protein is unknown. Macromolecular X-ray crystallography models are made through a combination of manual and algorithmic modeling, leading to the possibility that there is bias in how well parts of a single protein structure fit the underlying experimental data that is hidden within these global metrics.

Macromolecular X-ray crystallography models are constructed through an iterative process of manual modeling and refinement to achieve optimal agreement between the atomic model and the experimental data (structure factors). Manual modeling involves the modeler (i.e. human user) examining the atomic model and corresponding real space electron density, adjusting the model to improve the fit to density[7]. The user aims to resolve ambiguities in the map by adjusting the structure by repositioning a portion of the molecule, including fitting ligands or other non-protein atoms, such as water molecules. This updated model is then fed through refinement software, optimizing the model's fit with the experimental data in reciprocal space while maintaining plausible geometry, given user-defined parameters [8,9]. Refinement also updated the phases of the map, providing the user with an improved real-space electron density map to adjust the model manually. However, during the manual modeling step, due to the number of ambiguities in the map, there is a tendency to focus on modeling the 'important' parts of the structure, such as a binding or other functional site, while ignoring other parts of the protein structure. This can introduce human bias in the local quality of deposited structural models, with certain regions of the structure aligning more accurately with the experimental data than others.

Structural models have been invaluable in advancing our understanding of biology. However, recognizing the underlying data and potential biases is crucial, as these factors significantly impact the reliability and uncertainty of the conclusions drawn for bioinformatics or machine-learning algorithms, including differences in prediction algorithms, bias in structural bioinformatics, and incorrect conclusions from single or groups of protein structural models. Since the full structure of a protein is essential for understanding its biological function [10], any biases in the data could lead to inaccurate or misleading interpretations.

Here, we explored whether there are local differences in how well parts of protein structures are modeled, focusing on how binding sites are modeled compared to the rest of the protein. We observed binding site residues fit the experimental data better than residues outside the binding site and this is likely driven by more carefully modeled than the rest of the protein.

Results

Dataset

To create our initial dataset, we collected all X-ray crystallography structures in the PDBRedo, a database of re-refined and re-built structures[11–13](*n=41,374*). PDBRedo pipeline includes an automated re-refinement procedure with REFMAC, rebuilding programs, including loop rebuilding and peptide flips[14,15], and validation. More details can be found in Joosten et al [13].

We defined binding site residues as any residue within 5 angstroms of the largest ligand in the structure, requiring at least ten heavy atoms minus crystallographic additives (**Supplementary Table 1**). These were defined as bound structures (*bound dataset; n=16,203*). We then obtained another dataset deemed 'unbound' structures with no ligand bound. To determine potential binding or functional sites in unbound structures, we used the fpocket algorithm[16], which uses Voronoi tessellation and alpha spheres to identify pockets with structures. Pockets with volumes over 300 were considered large enough to be a possible binding site (*pocket dataset; n=21,381*). We then used the sphere representation of fpocket to determine all residues within the binding pocket, again using 5 Angstroms as the cut-off for pocket residues (**Supplementary Figure 1**).

Density Fitness Calculations

We first determine if there were differences in how well residues in the binding site fit the electron density compared to residues outside the binding site. To explore this, we examined three metrics that measure the fit of real space electron density to the model, real-space correlation coefficient (RSCC), real space R-values (RSR), and the mean electron density metrics for individual atoms (EDIAm) of residues[17].

Overall, we observed differences between binding site residues and the rest of the protein (**Figure 1A-C; Supplementary Table 2**) with binding site residues always fitting to experimental data better. RSCC, which measures the correlation between observed and calculated density (**Methods**). Higher values indicate a better cross correlation between the data and model. We observed that most residues fit the density well. However, there was a difference in how well binding site residues fit the data, with an RSCC average of 0.94 for the entire protein versus 0.96 for the binding site(**Figure 1A**; p=0.0, paired t-test). RSR, which measures the absolute difference between the observed and calculated densities, with lower values indicating improved fit, indicated a larger difference between binding site residues and the entire protein, with an average of 0.058 for binding site residues and 0.076 for all residues (**Figure 1B**; p=0.0, paired t-test). While this metric again indicated that most residues fit well with the data, given it's more sensitive to outliers than RSCC, it is likely identifying that there are more poor-fitting residues outside the binding site. EDIAm showed the same trend as RSR and RSCC (**Figure 1C**; binding site residues: 0.77, all residues: 0.66; p=0.0, paired t-test).

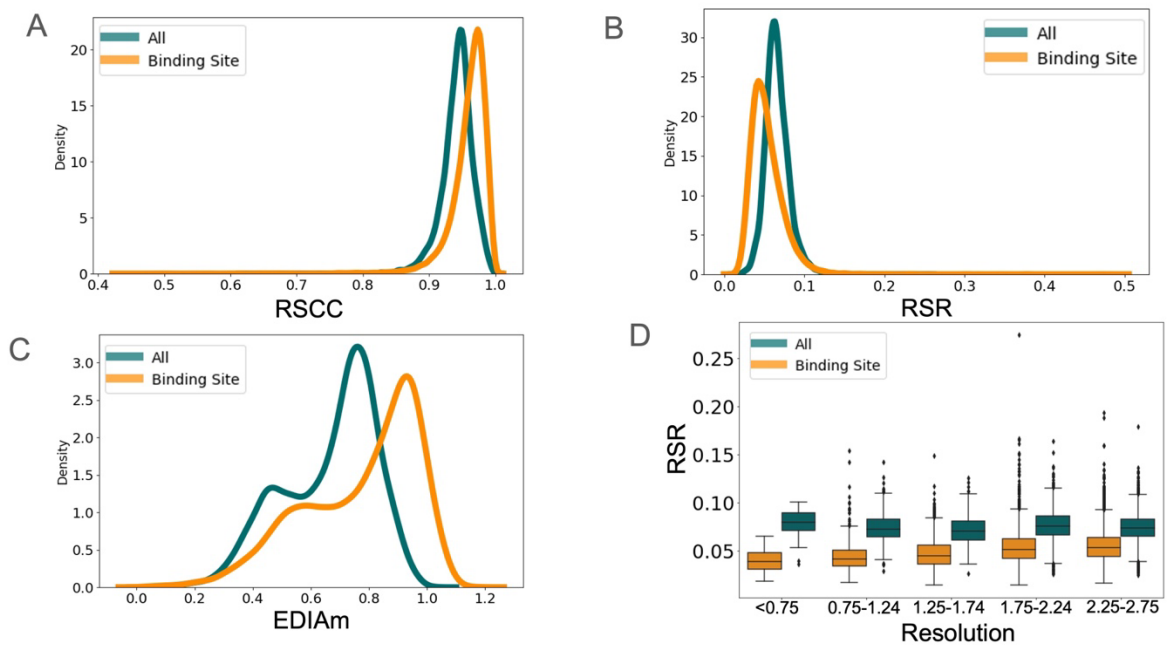


Figure 1. Binding site residues have a much better fit to real space electron density compared to non-binding site residues (**A**) RSCC metrics (binding site median-; non-binding site median; p=0.0), (**B**) RSR metrics (binding site median-; non-binding site median; p=0.0), (**C**) EDIAm metrics (binding site median-; non-binding site median; p=0.0), (**D**) RSR across resolution groups.

As these real space metrics are likely impacted by resolution and how correct the phases are between the map and model, we separated the real space metrics by resolution and Rfree values. Across resolution ranges, we observed similar trends, with binding site residues fitting better than non-binding site residues. However, these trends were more pronounced in high-resolution structures, indicating that selecting based on resolution would not eliminate this bias (**Figure 1D; Supplementary Figure 2A**; Resolution: <0.75: Binding- 0.039, Non-binding- 0.078; 0.75-1.24: Binding- 0.044, Non-binding- 0.074; 1.25-1.74: Binding- 0.048, Non-binding- 0.074; 1.75-2.24: Binding- 0.054, Non-binding- 0.076). We also observed that this trend hold across Rfree values, indicating that global metrics of data fit does not eliminate this bias (**Supplementary Figure 2B/C**).

We then performed the same analysis in our pocket dataset structures. Pocket residues were better modeled than all residues on average, but there was a smaller difference compared to binding site residues (**Supplementary Figure 3D**). This indicated that the results we observed around ligand binding sites were likely due to modeling differences rather than biological differences.

Number of Alternative Conformations

To try to determine what was driving this difference in fit to density, we explored if it was driven by the number of alternative conformations (altlocs) placed, identifying potential modeling bias. Crystallographers utilize altlocs to describe parts of the structure that could reasonably be modeled in multiple discrete locations. While there are now multiple algorithms to place

alternative conformations automatically[18–20] and help guide users in Coot where alternative conformations should be placed[21], most alternative conformations are placed manually. Therefore, we took this as a metric of modelers paying closer attention to areas of the protein with more alternative conformations.

We observed that over 58% of structures have at least one altloc. In binding site residues, 5.0% of residues had at least one altloc compared to 1.9% of residues elsewhere in the protein (**Figure 2A;** p=5.07e-165, paired t-test). We interpreted this finding as modelers paying more attention to binding site residues than the rest of the protein. Given that altlocs are more likely to be placed at high resolution and in models with overall better fit data, we explored if these trends changed across different resolution ranges and/or Rfree values (**Figure 2B; Supplementary Figure 3A**).

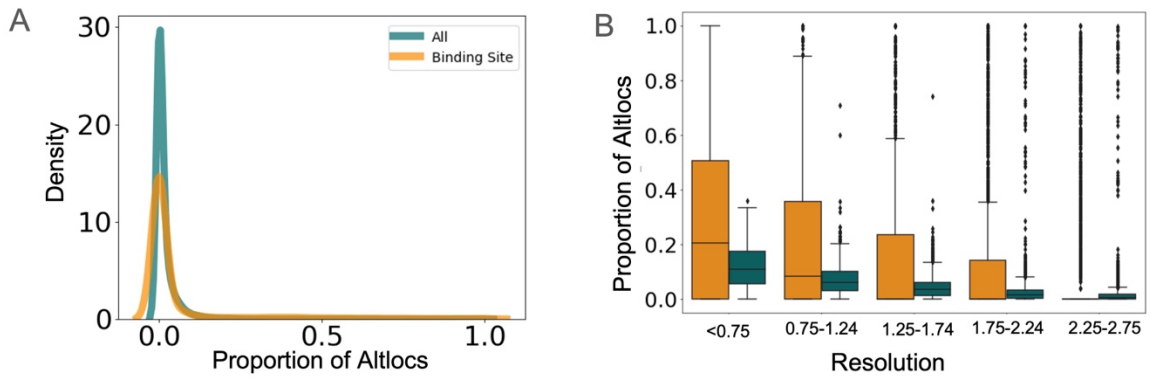


Figure 2. (**A**) Altlocs are more frequently placed in binding site residues compared to all residues. (**B**) Altlocs are more frequently placed in binding sites, regardless of the resolution of the structure.

We then performed the same analysis in our pocket dataset structures. We did not observe significant differences in EDIAm between alt locs in pocket residues and all residues (**Supplementary Figure 3B;** pocket residues- 0.72, all residues- 0.66). This indicated that the results we observed around ligand binding sites were likely due to modeling differences rather than biological differences.

Geometry Metrics

We then explored if the differences in fit to density might be driven by improved fitting of individual residues, indicated by geometry metrics. Beyond placing ligands and/or altlocs in protein structures, side-chain rotamers are often manually altered in Coot[7]. While refinement programs restrain structures based on idealized geometric metrics, including fitting into idealized rotamer space, these can be changed manually in Coot. While there are cases, especially in high-resolution structures, that rotamers should fit outside of their 'idealized' position, we expected that there would be fewer outliers in a more carefully modeled area of the protein, such as binding site residues. To explore if the geometry is better in binding site residues, we examined the rotamer status of every residue, examining how many were labeled as 'outliers' or 'allowed' from phenix.rotalyze[22]. Rotamers classified as outliers are those with

less than 0.3% in the reference dataset, whereas rotamers classified in the allowed category indicate a 0.3-2.0% occurrence in the reference data[22].

In binding site residues, 2.4% of residues versus 2.0% of non-binding site residues were classified as outliers (**Figure 3A**; p=0.56), indicating that all residues infrequently had rotamer outliers. We then explored allowed rotamers, which showed that 4.8% of non-binding site residues were categorized here, with 4.2% of binding residues classified as allowed (**Figure 3B;** p=1.38e-08).

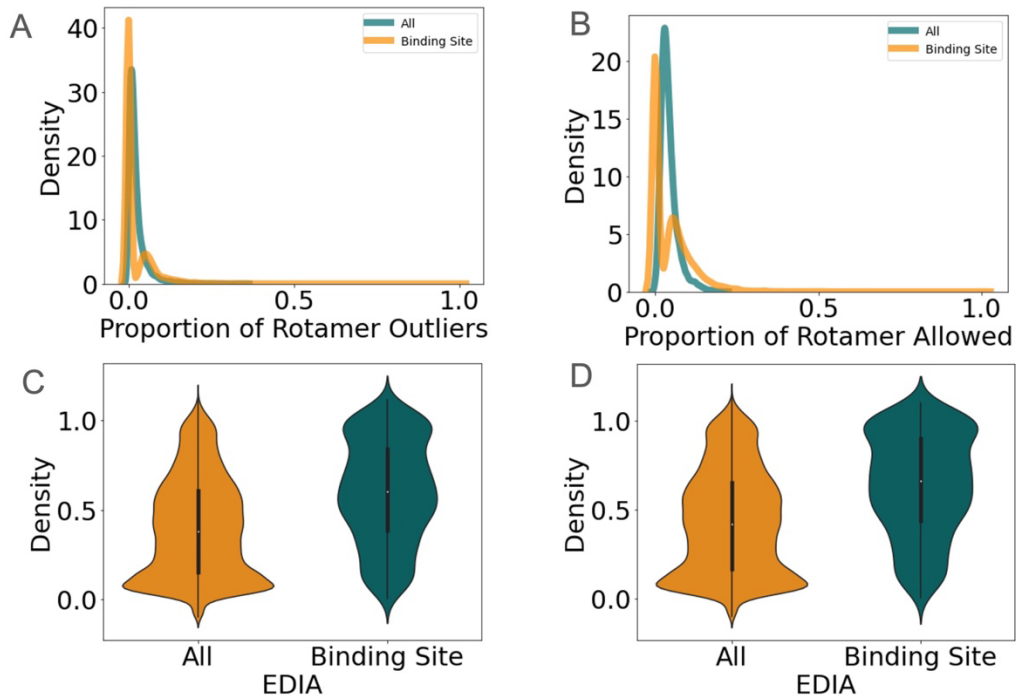


Figure 3. Binding site residues have fewer non-favored rotamer values. (**A**) Proportion of residues with rotamer outliers (binding site- 0.009; all residues- 0.015; p=8.79e-42), (**B**) Proportion of residues with allowed rotamer (binding site- 0.42; all residues- 0.48; p= 8.79e-42) (**C**) Distribution of EDIAm in residues classified as rotamer outliers (Binding site- 0.52; All residues- 0.36; p=1.3e-40) (**D**) Distribution of EDIAm in residues classified as allowed rotamer (binding site-0.64; all residues- 0.43; p=5.89e-95)

However, with both metrics, the binding site residues show a bimodal distribution. As some rotamer outliers represent true unusual positioning of residues, especially when interacting with a substrate or small molecules, we suspected that this was driving the bimodal distribution only observed in binding site residues. We hypothesized that binding site residues with rotamer outliers would have better evidence from the electron density to support their outlier status. Therefore, we explored whether outlier rotamers in the binding site had better evidence from the experimental data, comparing the EDIAm values of outliers found in binding sites versus all residues sites. We observed that binding site residues classified as outliers (**Figure 3C**) and those modeled as allowed (**Figure 3D**) had an improved fit to the experimental data compared

to all residues. This indicated that non-ideal rotamers in binding sites are more likely driven by true outliers, with density support, compared to non-ideal rotamers outside of the binding site.

Discussion

Over multiple metrics, we observed that binding site residues tend to fit experimental data more accurately. Our data indicates that this is likely driven due to more carefully manually modeled in binding sire residues as observed by increased number of alternative conformations, the reduction in geometry outliers, and the fit of geometry outliers to the underlying experimental data in binding site residues. Overall, this highlights a strong bias in the accuracy of structural models towards binding sites, with implications on drawing conclusions from a single structural model to machine learning algorithms. We call on the structural biology community to try to reduce this bias modeling moving forward by paying attention to areas outside of the binding site and for the need for more automated or guided modeling algorithms[19–21,23]. Although manual modeling will likely remain a key component of experimental structural biology, the shift towards higher-throughput techniques and the increasing complexity of large cryo-EM complexes necessitate more automated experimental modeling methods, as manually handling larger models is becoming progressively more challenging.

While we focused our analysis on PDBRedo to eliminate errors due to varied refinement, this trend is likely worse within non-PDBRedo structures. PDBRedo has provided individual residue metrics, and these metrics should be used to highlight local errors in models. However, echoing what has been called for before[4], within the PDB there is a pressing need for more useful measures to highlight local errors in structural models both upon deposition and for people who are downloading and using the data. These metrics are likely even more important for other frequently mismodeled molecules, such as ligands and waters[39].

Further, we and others have demonstrated that there is a lot of importance in the structures and dynamics outside of the binding site! These areas are critical for allostery, binding, and cooperativity to name a few. Belittling these areas in our structural models can perpetuate biological conclusions about the importance of binding sites over the rest of the protein[10,36–38].

These biased structural models often serve as ‘truth’ for simulations, docking, bioinformatic studies, and, most recently, machine learning algorithms[24–30]. However, our findings emphasize we must be careful in how we use and explain biological results from PDB models. *It is critical to understand that PDB models are just that,* ***models***. They do not fully explain the underlying experimental data, have nuances in how they are representing the data, including alternative conformations, anisotropic B-factors, or translation-liberation-screw parameters, and as we show, can vary across the model in how well the underlying data is explained due to user error or biases[31–33]. While there are efforts to harmonize and improve the encoding of this data to capture more of these nuances[34], many structural biology machine learning algorithms directly use these biased models to draw biological conclusions, potentially perpetuating errors. The modeling biases that we present here, adds to additional biases in the PDB based on the ability to capture states of a macromolecule using X-ray crystallography or cryoEM[35]. Understanding

the errors, biases, and uncertainties in data is as important as the computational algorithms when drawing conclusions.

References

1. Joosten, R. P., Joosten, K., Murshudov, G. N. & Perrakis, A. PDB_REDO: constructive validation, more than just looking for errors. *Acta Crystallogr D Biol Crystallogr* **68**, 484–496 (2012).
2. Brünger, A. T. Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature* **355**, 472–475 (1992).
3. Urzhumtseva, L., Afonine, P. V., Adams, P. D. & Urzhumtsev, A. Crystallographic model quality at a glance. *Acta Crystallogr D Biol Crystallogr* **65**, 297–300 (2009).
4. Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr A* **47 (Pt 2)**, 110–119 (1991).
5. Read, R. J. *et al.* A new generation of crystallographic validation tools for the protein data bank. *Structure* **19**, 1395–1412 (2011).
6. Shao, C., Bittrich, S., Wang, S. & Burley, S. K. Assessing PDB macromolecular crystal structure confidence at the individual amino acid residue level. *Structure* **30**, 1385–1394.e3 (2022).
7. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr D Biol Crystallogr* **66**, 486–501 (2010).
8. Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr D Biol Crystallogr* **66**, 213–221 (2010).
9. Murshudov, G. N. *et al.* REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallogr D Biol Crystallogr* **67**, 355–367 (2011).
10. Wankowicz, S. A., de Oliveira, S. H., Hogan, D. W., van den Bedem, H. & Fraser, J. S.

Ligand binding remodels protein side-chain conformational heterogeneity. *Elife* **11**, (2022).

11. Joosten, R. P. & Vriend, G. PDB improvement starts with data deposition. *Science* **317**, 195–196 (2007).
12. Joosten, R. P., Womack, T., Vriend, G. & Bricogne, G. Re-refinement from deposited X-ray data can deliver improved models for most PDB entries. *Acta Crystallogr D Biol Crystallogr* **65**, 176–185 (2009).
13. Joosten, R. P. *et al.* PDB_REDO: automated re-refinement of X-ray structure models in the PDB. *J Appl Crystallogr* **42**, 376–384 (2009).
14. Joosten, R. P., Joosten, K., Cohen, S. X., Vriend, G. & Perrakis, A. Automatic rebuilding and optimization of crystallographic structures in the Protein Data Bank. *Bioinformatics* **27**, 3392–3398 (2011).
15. van Beusekom, B., Joosten, K., Hekkelman, M. L., Joosten, R. P. & Perrakis, A. Homology-based loop modeling yields more complete crystallographic protein structures. *IUCrJ* **5**, 585–594 (2018).
16. Le Guilloux, V., Schmidtke, P. & Tuffery, P. Fpocket: an open source platform for ligand pocket detection. *BMC Bioinformatics* **10**, 168 (2009).
17. Tickle, I. J. Statistical quality indicators for electron-density maps. *Acta Crystallogr D Biol Crystallogr* **68**, 454–467 (2012).
18. 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).
19. Wankowicz, S. A. *et al.* Automated multiconformer model building for X-ray crystallography and cryo-EM. *Elife* **12**, (2024).
20. Flowers, J. *et al.* Expanding Automated Multiconformer Ligand Modeling to Macrocycles and Fragments. *bioRxiv* (2024) doi:10.1101/2024.09.20.613996.
21. 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).

22. Hintze, B. J., Lewis, S. M., Richardson, J. S. & Richardson, D. C. Molprobity's ultimate rotamer-library distributions for model validation. *Proteins* **84**, 1177–1189 (2016).
23. Jamali, K. *et al.* Automated model building and protein identification in cryo-EM maps. *Nature* **628**, 450–457 (2024).
24. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
25. Krishna, R. *et al.* Generalized biomolecular modeling and design with RoseTTAFold All-Atom. *Science* **384**, eadl2528 (2024).
26. Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).
27. Sadybekov, A. V. & Katritch, V. Computational approaches streamlining drug discovery. *Nature* **616**, 673–685 (2023).
28. Cai, F., Wei, Y., Kirchhofer, D., Chang, A. & Zhang, Y. Rapid prediction of key residues for foldability by machine learning model enables the design of highly functional libraries with hyperstable constrained peptide scaffolds. *PLoS Comput Biol* **20**, e1012609 (2024).
29. Anand, N. *et al.* Protein sequence design with a learned potential. *Nat Commun* **13**, 746 (2022).
30. Dauparas, J. *et al.* Robust deep learning-based protein sequence design using ProteinMPNN. *Science* **378**, 49–56 (2022).
31. Woldeyes, R. A., Sivak, D. A. & Fraser, J. S. E pluribus unum, no more: from one crystal, many conformations. *Curr Opin Struct Biol* **28**, 56–62 (2014).
32. 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).
33. Fraser, J. S. & Murcko, M. A. Structure is beauty, but not always truth. *Cell* **187**, 517–520 (2024).

34. Wankowicz, S. A. & Fraser, J. S. Comprehensive encoding of conformational and compositional protein structural ensembles through the mmCIF data structure. *IUCrJ* **11**, 494–501 (2024).

35. Peng, K., Obradovic, Z. & Vucetic, S. Exploring bias in the Protein Data Bank using contrast classifiers. *Pac Symp Biocomput* 435–446 (2004).

36. Wankowicz, S. & Fraser, J. Making sense of chaos: uncovering the mechanisms of conformational entropy. *ChemRxiv* (2024) doi:10.26434/chemrxiv-2023-9b5k7-v3.

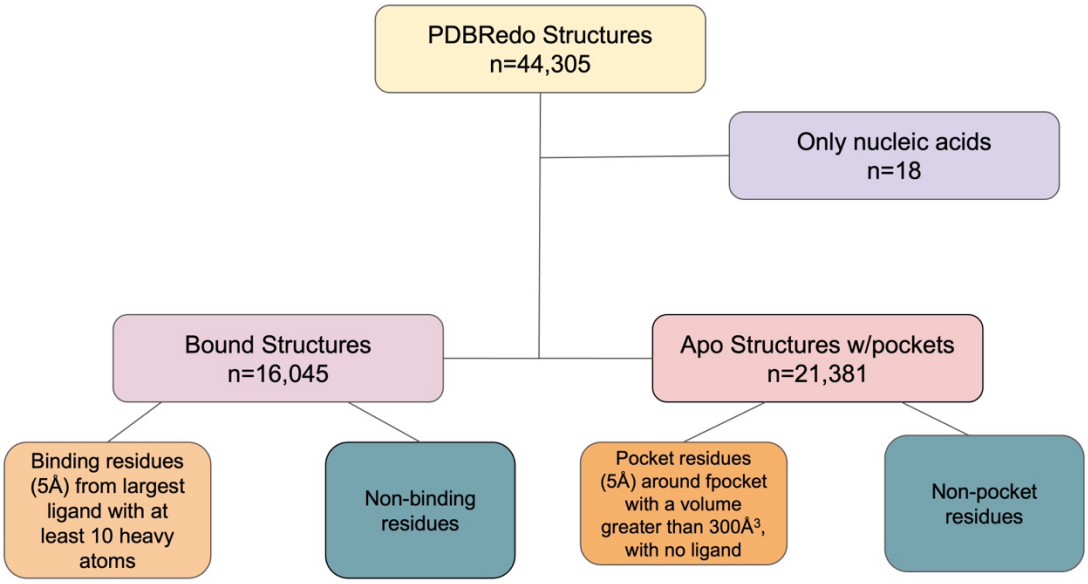
37. Perdikari, A. *et al.* Structures of human PTP1B variants reveal allosteric sites to target for weight loss therapy. *bioRxiv* (2024) doi:10.1101/2024.08.05.603709.

38. Markin, C. J. *et al.* Decoupling of catalysis and transition state analog binding from mutations throughout a phosphatase revealed by high-throughput enzymology. *Proc Natl Acad Sci U S A* **120**, e2219074120 (2023).

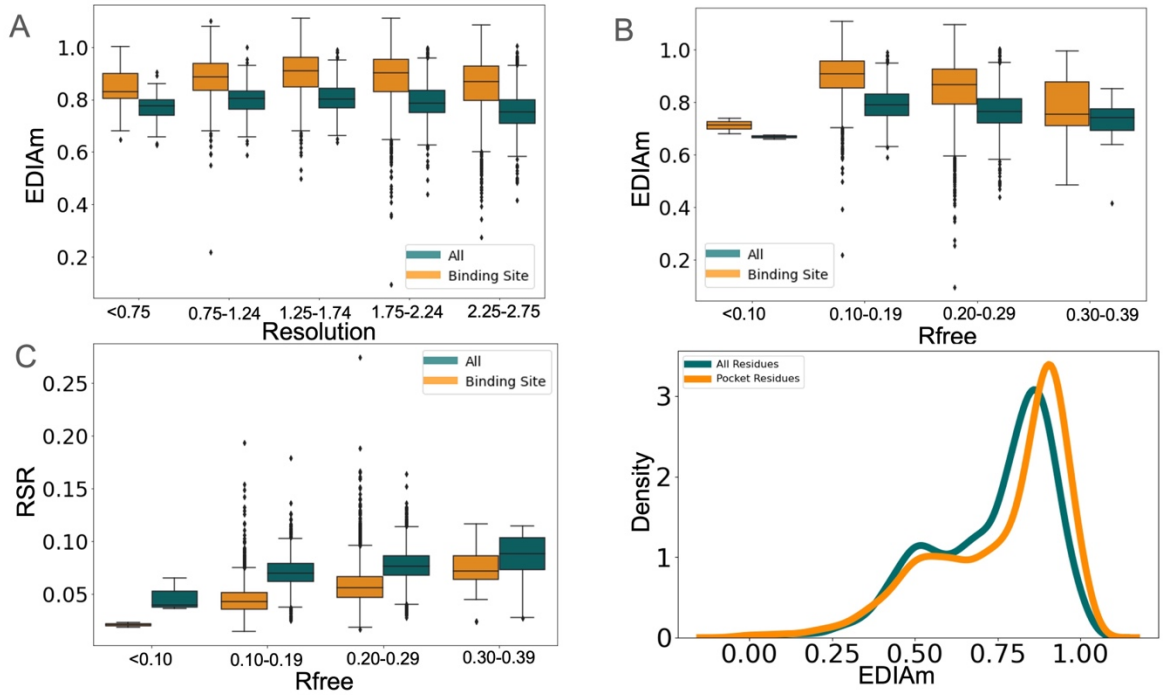
39. Nittinger, E., Schneider, N., Lange, G. & Rarey, M. Evidence of water molecules--a statistical evaluation of water molecules based on electron density. *J Chem Inf Model* **55**, 771–783 (2015).

40. Murshudov, G. N., Vagin, A. A. & Dodson, E. J. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr D Biol Crystallogr* **53**, 240–255 (1997).

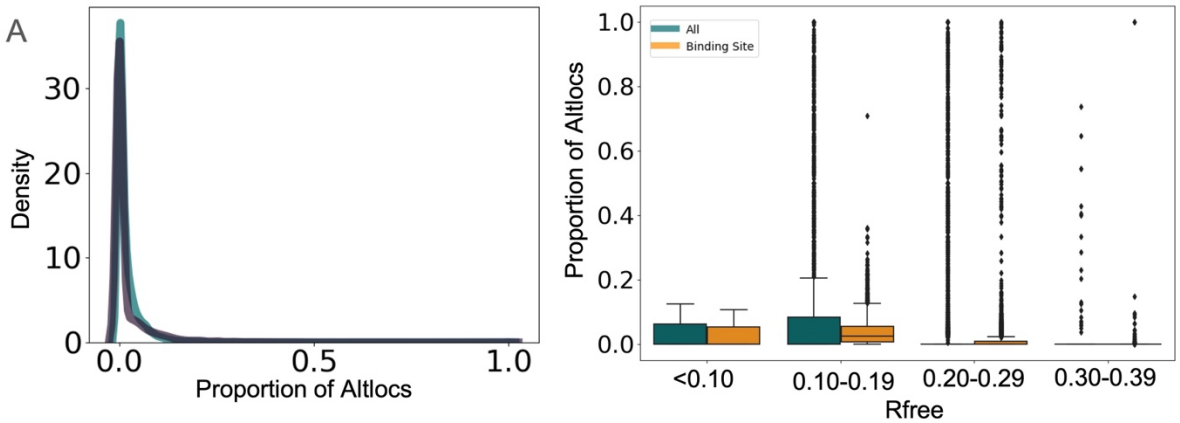
Supplementary Figures



Supplementary Figure 1. Workflow for determining PDBs and residues in comparison groups



Supplementary Figure 2. (**A**) EDIAm differences by resolution. While all groups were statistically significant, the largest difference were in the higher resolution groups. (**B**) EDIAm differences by Rfree value. We observed similar trend across all Rfree ranges, with Rfree between 0.10 and 0.19 having the most extreme differences (binding- 0.89, all-0.79). (**C**) RSR differences by Rfree values. (D) EDIAm values in pocket residues versus all residues (pocket- 0.72, all-0.66).



Supplementary Figure 3. (**A**) The number of altlocs places in pockets without ligands occurs more often than non-pocket residues, but less frequently than in binding site residues (Pocket residues median-2.0%; all residues median-1.9%; p=3.38e-07). (**B**) Increased number of altlocs modeled in binding sites occurs regardless of Rfree values.