PLOS Computational Biology

EDUCATION

Ten rules for a structural bioinformatic analysis

Stephanie A. Wankowicz[1,2]*

[1] Departments of Molecular Physiology and Biophysics, Biochemistry, Computer Science, Vanderbilt University, Nashville, Tennessee, United States of America, [2] Center for Applied AI in Protein Dynamics, Center for Structural Biology, Vanderbilt University, Nashville, Tennessee, United States of America

* stephanie@wankowiczlab.com

Abstract

The Protein Data Bank (PDB) is one of the richest open-source repositories in biology, housing over 242,000 macromolecular structural models alongside much of the experimental data that underpins these models. By systematically collecting, validating, and indexing these models, the PDB has accelerated structural biology discoveries, enabling researchers to compare new entries against a vast archive of solved structures and, more recently, powering protein structure prediction. Leveraging this wealth of data, structural bioinformatics has uncovered patterns, such as conserved protein folds, binding-site features, or subtle conformational shifts among related proteins, that would be impossible to detect from any single structure. Through the democratization of structural data and open-source analytical tools, now amplified by the power of large language models, a broader community of researchers is equipped to drive new scientific discoveries using structural data. However, good structural bioinformatics requires understanding some of the nuances of the underlying experimental data, data encoding conventions, and quality control metrics that can affect a model's precision, fit-to-data, and comparability. This knowledge, combined with developing good controls, statistics, and connections to other databases, is essential for drawing accurate and reliable conclusions from PDB data. Here, we outline 10 recommendations for doing structural bioinformatic analyses crafted to pave the way for others to uncover exciting discoveries.

Author summary

Here, we provide a roadmap for users to leverage the Protein Data Bank's vast collection of protein structural models into reliable and valuable insights. It lays out 10 clear rules that help readers quality control their data, choose fair comparison sets, and judge model quality so results aren't led astray by noise, bias, or overconfidence. The guide also shows how to connect structures to other databases. By highlighting best practices, such as utilizing re-refined models and being aware of common pitfalls, we guide users to leverage this rich data

OPEN ACCESS

Citation: Wankowicz SA (2025) Ten rules for a structural bioinformatic analysis. PLoS Comput Biol 21(10): e1013094. https://doi.org/10.1371/journal.pcbi.1013094

Editor: Patricia M Palagi, SIB Swiss Institute of Bioinformatics, SWITZERLAND

Received: April 30, 2025

Accepted: October 9, 2025

Published: October 31, 2025

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pcbi.1013094

Data availability statement: No data involved.

Funding: The author(s) received no specific funding for this work.

Competing interests: The author has declared that no competing interests exist.

for enhanced biological insights. These guidelines will enable stronger, more reproducible structural analyses that accelerate drug discovery, illuminate disease mechanisms, and make open data broadly useful across the life sciences.

Introduction

In 1971, with only seven structures, the first, and still active, open-access database in biology was created: the Protein Data Bank (PDB) [1]. From these modest beginnings, the PDB has grown to include over 242,000 structures [2], and its systematic archiving of macromolecular models has reshaped the field, transforming our understanding of the relationship between structure and biological function and enabling advances from elucidating enzyme catalysis to the rational design of new therapeutics. Critically, when the Research Collaboratory for Structural Bioinformatics (RCSB) PDB was established, one of their first major undertakings was a large-scale remediation of legacy data, addressing inconsistent formats, incomplete metadata, and nonstandard nomenclature to enable systematic analysis and ensure that the archive could support future large-scale structural bioinformatics. The remediation effort, later extended by the wwPDB partners (PDBe, PDBj, and Biological Magnetic Resonance Bank [BMRB]), standardized chemical components, corrected errors, and transitioned the archive to the current, more robust mmCIF format [2–4]. These efforts laid the foundation for structural bioinformatics by ensuring that PDB data were reliable, interoperable, and machine-readable.

Today, the PDB is maintained as a single, global archive through the Worldwide Protein Data Bank (wwPDB) consortium, which coordinates deposition, validation, and dissemination of macromolecular structures. The consortium comprises regional data centers, RCSB PDB in the United States, PDBe in Europe, PDBj in Japan, each providing unique portals, visualization tools, and database integrations tailored to their respective communities [5–7]. All sites share a unified deposition system, ensuring that structures are consistently validated and mirrored worldwide within 24 hours of release [8]. In addition, the Electron Microscopy Data Bank (EMDB), jointly maintained with the PDB, serves as the central repository for cryo-electron microscopy (cryo-EM) electron potential maps, enabling joint deposition of maps and models. Together, these interconnected entities provide a comprehensive and interoperable ecosystem that continues to accelerate discovery across the structural biology community.

The original vision of the PDB centered on the deposition of individual structures, where each entry told a story about a protein's fold, function, or interaction. This one-structure, one-story mode of structural biology has yielded an enormous wealth of knowledge and fundamentally shaped our understanding of biology. But one of the greatest strengths of the PDB lies in its ability to uncover new biological insights beyond the scope of a single structure. This has spurred the field of structural bioinformatics, which, through examining patterns over tens to thousands of structures, has identified relationships of protein families and folds, the role of evolution driving protein structure and function, and information on macromolecular interactions and catalysis [9–16]. Structural bioinformatics also created the critical foundation for the protein structure prediction breakthrough [17–21]. By analyzing thousands of structures at once, you gain the statistical power to detect subtle variations that, when aggregated, have the potential to reveal robust patterns in allostery, ligand binding, macromolecular assembly, and catalysis [9,12]. These analyses can be incredibly powerful alone or in conjunction with other bioinformatic databases, prospective experiments, or theoretical models.

At the core of every PDB structural model lies the atom table, which records the atomic coordinates along with key attributes such as atom type, residue identity, B-factor (atomic displacement parameter), and occupancy. The adoption of the mmCIF format has provided a far richer and more extensible representation than the legacy PDB format. Unlike the fixed-column limitations of PDB files, mmCIF can accommodate the growth of structural biology, including new ligands with five-character identifiers and very large macromolecular assemblies that exceed the capacity of the original format [22]. Another key advantage of mmCIF is its ability to map these deposited coordinates to canonical protein sequences, enabling seamless integration with UniProt and related databases [23]. mmCIF also underpins emerging resources such as PDB-IHM [24], which supports the deposition of integrative and ensemble models, and it provides the extensibility needed for new schemas, including recent work on hierarchical representations of conformational and chemical heterogeneity [25].

The democratization of coding skills, facilitated by large language models, has enabled more users to delve into structural bioinformatics, build hypotheses, support experimental findings, or make independent discoveries. However, structural data is nuanced and can be challenging to work with; without proper quality control or control analyses, it can lead to inaccurate conclusions. Users should understand limitations, potential pitfalls, and caveats to use the data to its full potential.

The guidelines outlined here stem from the lessons and challenges I and others have encountered while performing structural bioinformatics projects, with many of the lessons being applicable for machine learning applications as well. Although predicted structures are increasingly valuable in bioinformatics research [26–28], this article emphasizes experimentally derived structures. While each rule is not universally applicable, consider each recommendation carefully and evaluate how it may relate or be tweaked to fit their problem.

So, first things first, what question do you want to ask, or hypothesis do you want to test? For example, are you looking at the overall protein fold, or does your question require you to know the rotamer angles and only look at wild-type structures of a specific protein? Knowing the answers to these questions is critical for determining your selection criteria, statistical power, analysis, and controls, as outlined below.

Recommendation #1: Define your biological selection criteria

When starting a structural bioinformatics project, the first step is to define the biological criteria for your study. Consider the structures you need to answer your research question, whether it involves all lysozymes, a specific tyrosine kinase, or all enzymes. Additionally, you may want to further refine your dataset based on ligands. Small molecules such as glycerol or DMSO are often crystallographic additives, while other molecules may be native or synthetic ligands, leading to differences in how you want to classify each structure. Further, assess whether your protein is part of large complexes by examining the identities of other chains. Identical chains typically reflect symmetry-related protomers, whereas distinct macromolecules reveal a multi-protein complex (see more details in Recommendation #4).

Beyond structural selection, sequence-level considerations remove redundancy and drive clustering and alignment analyses. A significant fraction of PDB entries corresponds to homologous proteins or multiple structures of the same protein. Depending on your question, you may want to filter based on sequence or structure. You can cluster by sequence, using MMseq or CDHit [29,30], or based on structure, using TM-score and CATH [31,32], selecting representatives from each cluster for your downstream analysis based on resolution and R-factors or other metrics (see more in Recommendation #2). Tools like the PISCES server automate this by removing sequences above a chosen identity threshold and keeping the highest-quality structure from each group [33].

The RSCB PDB offers precomputed sequence clusters at certain thresholds (100–30%), based on the MMseqs2 algorithm [30], which applies variable identity thresholds based on modeled residues. Note that this does not include unmodeled residues, often in terminal or loop regions, which can impact this analysis. For sequence alignments, the PDBe supports multiple sequence alignment (MSA) using Clustal Omega [34] and allows retrieval of FASTA sequences for custom alignment. By leveraging the SIFTS database [35], you can map PDB entries onto CATH or SCOP structural hierarchies, UniProt sequence records, and select structures by fold, superfamily, or sequence-based functional annotation (see more in Recommendation #9) [23,31,36]. Additionally, you can use structural alignments, such as those performed with FATCAT, TM-align, CE, or Smith-Waterman 3D alignment, to provide insights into sequence and structural relationships [37–41]. These can be powerful in identifying similar shapes of proteins, yet with sequence differences. Many of these selections can be made using one of the three PDB APIs [5,7,42].

Recommendation #2: Determine how you will quality control your data

Beyond determining the biological and sequence selection criteria, it is crucial to consider the experimental data underlying structures to ensure a quality dataset. This begins with identifying the methods of structure determination, such as X-ray crystallography, cryo-EM, nuclear magnetic resonance (NMR), or neutron diffraction. Additional factors include resolution (not applicable for NMR), agreement with structure determination data, and stereochemical accuracy.

Resolution, the most common criterion for structural bioinformatics analysis, sets the theoretical limit on the precision of the structural model and is reported for all structures. High resolution, better than 2.5 Å, is essential for accurate side chain positioning, whereas lower resolution models can still yield valuable insights into overall fold and backbone conformation. In cryo-EM, however, resolution is estimated differently than in crystallography: it is typically calculated using the Fourier Shell Correlation (FSC) between two independently reconstructed half-maps [43]. The FSC curve reflects the degree of agreement between the two maps as a function of spatial frequency, and the resolution is conventionally reported at the point where the correlation falls below a given threshold (commonly 0.143). However, the FSC is not a direct measure of atomic detail in the same way that crystallographic resolution is, but rather a measure of the global similarity between two noisy reconstructions. Complicating matters further, cryo-EM maps often exhibit substantial variation in local resolution across the structure, meaning a single global resolution metric may not faithfully capture the interpretability of all regions of the map [44]. It is also important to note that in both X-ray and cryo-EM, while two structures can have the same resolution, they can be modeled to different levels of accuracy, necessitating the exploration of other metrics.

The PDB publishes global validation metrics, including knowledge-based assessments of atomic models, evaluations of the underlying experimental data, and measures of agreement between model and data, and provides detailed validation reports for each entry, following standards established by the X-ray, NMR, and cryo-EM validation task forces [45–49]. Even at high resolution, nearly all structures have a few local errors, but at lower resolutions, errors become more widespread. As structural models can vary widely in quality, these validation metrics are important to consider maintaining scientific reliability and to minimize the risk of errors propagating into biological interpretation, drug design, or computational modeling.

Geometric metrics, such as Ramachandran outliers, are derived from tools such as MolProbity and PROCHECK [50–52]. In X-ray crystallography, *R*-values quantify the agreement between model and data, with higher values reflecting poorer fits; a value near or above 0.3 is commonly used as a threshold for poor quality [53]. Further, depending on your question, you also may want to examine geometry or fit to real space data of individual residues using tools such as Molprobity, Ringer, or real space correlation coefficients [52,54–57].

In cryo-EM, there is an expanding set of approaches being introduced to evaluate the quality of deposited maps and models [58]. As discussed above, global measures such as FSC between half-maps remain the gold standard for estimating overall map resolution, while model–map FSC curves assess the consistency between the atomic model and experimental maps [47,59]. Increasingly, local validation has become critical. EMRinger evaluates the accuracy of side-chain placement by comparing electron potential peaks with expected rotamer positions, and Q-scores quantify how well the electron potential supports individual atoms and residues [60,61].

You should also be aware of unmodeled regions, often loops or termini. These can be identified by manually comparing the FASTA sequence, representing the input construct, against the sequence in the PDB structural model or with tools like Seqatoms [62]. One may decide to exclude proteins with large missing segments or model in missing loops (see Recommendation #3). The PDB-REDO team has developed an algorithm (Loopwhole) to help fill in many of these missing loops, but it is most effective when high-quality homologous structures are available and when the experimental electron density supports accurate grafting and refinement [63]. Models that include filled loops will be classified as “rebuilt” in the PDB-REDO database. If you include structures with unresolved regions, acknowledge this limitation and adjust your analysis accordingly (see more in Recommendation #8).

Beyond proteins, structures often include small molecules, nucleic acids, carbohydrates, or other molecules of varying quality [64]. Small molecule ligand quality is assessed by agreement with experimental data and geometric accuracy [65], with the latter being evaluated against Cambridge Structural Database reference structures [66]. Metals are also checked by CheckMyMetal, which evaluates metal coordination geometry, bond valency, and potential steric clashes [67]. For nucleic acids, PDB-REDO has introduced validation routines to assess the normality of Watson–Crick base-pair geometry, while DNATCO provides complementary validation of DNA and RNA backbone conformations [68,69].

Ultimately, determining the appropriate experimental selection criteria depends on your research question. For instance, if your research focuses on side chain positioning, higher resolution, lower *R*-values, and precise stereochemical validation are critical. Alternatively, if you are looking for information on the overall protein fold, a broader selection of structures may be acceptable.

Recommendation #3: Re-processing structural model data

Most structural bioinformatic approaches take information directly from coordinate (PDBx/mmCIF) files. By taking information directly from the coordinate files, you are taking on any errors or biases the original modelers had. Where possible, it is recommended to use X-ray structural models from PDB-REDO [70–72], which reanalyzed the majority of structures in the PDB with experimental data (structure factors), providing uniform automated re-refinement, combined with structure validation and difference-density peak analysis. Since the deposition of reflection data was only encouraged beginning in 1998 and became mandatory in 2008, older structures, whose experimental data were less frequently archived in the PDB, are underrepresented in PDB-REDO [73]. While many models without experimental data can be informative, they come with caveats due to different and older data processing pipelines.

It is also possible to re-process all structures yourself [74,75]. If you are new to refinement, there are many tutorials to get you started [76,77]. Re-processing data can ensure that experimental data is processed in the same way, or allow the application of specific modeling modality or tooling within a refinement program, such as multiconformer modeling, ensemble refinement, 3D variation analysis, or quantum refinement [78–80]. To be able to reprocess your data, you need experimental data to be available, such as MTZ files for X-ray crystallography, maps, half maps, or particle stacks for cryo-EM from the EMDB [81,82], or raw NMR data from the BMRB [81]. After re-processing, similar quality control metrics, as described in Recommendation #2, should be used to evaluate structures.

Recommendation #4: The PDB and structural models are weird and biased

The PDB is not a uniform sample of all proteins. Because high-resolution crystallography, which comprises the majority of the PDB, favors small, globular, soluble proteins, membrane and flexible or disordered proteins account for roughly 20%–30% of genes but make up less than 2% of PDB entries [83]. Moreover, publication bias further distorts the distribution of structural models with drug targets, enzymes, and other high-value human proteins accounting for a disproportionate share of PDB entries. This skewing of many structures of the same protein is becoming even more pronounced with an increase in fragment-screening campaigns [84]. As a result, certain protein families dominate the PDB, artificially amplifying their characteristic features in any global analysis. You must consider these redundancies in your analysis, as discussed in Recommendation #2.

In addition to redundancy, it is also important to understand what structural unit is represented in a PDB file. The database distinguishes between the asymmetric unit, the crystallographic unit directly observed in the experiment, and the biological assembly, which represents the functional quaternary structure in vivo. The PDB provides separate mmCIF files for biological assemblies, which are either specified by the authors or inferred computationally by tools such as PISA [85]. For most biological analyses, the biological assembly is the appropriate choice, though it should be noted that approximately 20% of these assemblies may be incorrect, with ProtCID and ProtCAD databases being valuable for sorting true assemblies from crystallographic artifacts [86,87].

Beyond the bias of what structural models exist in the PDB, structural models can be odd and biased. First, it is important to remember that PDB models are just models. They do not explain all the underlying experimental data and can vary depending on the processing pipeline (see Recommendation #3). For example, in X-ray crystallography, crystal contacts, nonbiological interactions between symmetry-related molecules within the crystal lattice, can artificially stabilize particular conformations or create interfaces that don't exist in solution, potentially skewing structural bioinformatics analyses of protein dynamics, flexibility, and genuine interaction sites. We also previously showed that binding site residues are often better modeled than residues outside the binding site [88]. Further, regions of unmodeled residues can arise for many reasons, including resolution and subjective modeling, but automated refinement pipelines cannot correct all of them. All of these issues can lead to structures having different biases. In addition, structures often include unmodeled blobs, frequently ligands.

Finally, all structural data contains extensive conformational and compositional heterogeneity modeled with varying accuracy and encoding [25]. These include anisotropic B-factors, alternative atom locations (altlocs), or multiple models. Anisotropic B-factors describe the direction and magnitude of atomic displacement, while alternative atom locations (altlocs) represent multiple conformations modeled for a single atom [78]. Multiple models, often used in ensemble structures, provide different plausible conformations that together capture the underlying structural variability [89]. While there are ways to encode some of these metrics more uniformly, some encodings cannot be interchanged. Additionally, most bioinformatics libraries, including Biopython, strip out much of this encoding, potentially introducing biases into downstream analyses [90]. To guard against these biases, it is essential to document any data exclusions or alterations made to the data, ensuring accurate comparisons downstream.

Recommendation #5: Consider your analysis's sample size, statistics, overfitting, and uncertainty

After dataset selection and quality control comes the fun part, looking at and identifying what drives differences between structures. Descriptive bioinformatic analyses, such as cataloguing residue types and counts within binding pockets, are straightforward, but any comparative study requires careful attention to sample size and statistical power. Smaller groups demand larger effect sizes to achieve significance, and paired comparisons should employ paired statistical tests to account for within-pair correlations. Equally important is judging whether observed differences, such as shifts in binding site residue rotamers or altered pocket volumes, are biologically meaningful [91].

When comparing two unpaired groups, choose parametric or nonparametric tests based on data distribution. Parametric tests assume normality, while nonparametric tests are more flexible when distributions are skewed (e.g., residue B-factor values or pocket volumes). For paired data, for example, wild-type vs. mutant, or bound vs. unbound structures, use paired t-tests or Wilcoxon signed-rank tests. Further, be wary of multiple hypothesis testing. Consider adjusting *p*-values using Bonferroni or false discovery rate corrections. You can also use resampling methods such as jackknife, bootstrap, or cross-validation to help estimate variability and confidence intervals. Applying well-chosen controls helps guard against false positives and ensures that your findings reflect genuine structural phenomena rather than quirks of a particular dataset.

Avoiding overfitting is equally critical, whether you are working in bioinformatics or machine learning. Where possible, never develop and validate hypotheses on the same data without independent testing. Splitting your dataset into train, test, and validation sets, or employing k-fold cross-validation, is even recommended when defining new structural descriptors or clustering algorithms. Further, consider how you partition your test set, whether by sequence similarity, structural features, or other criteria, to avoid overfitting or memorization [92].

Recommendation #6: Determine and apply the correct controls

Choosing the proper controls is one of a bioinformatic study's most challenging and often overlooked aspects. Fortunately, the abundance of publicly available structural data makes incorporating negative and positive controls feasible. Controls must directly address the null hypothesis you wish to reject. Negative control datasets, where no effect is expected, are usually easier to define, while positive control datasets, datasets known to exhibit the effect, can be harder to assemble. For example, if you're testing whether a novel structural motif alters protein function, you might compare your proteins of interest against a set of homologous structures that lack the motif. Differences that persist between the groups are more likely to stem from the motif than background variation. You can also randomize specific features, such as residue type or solvent exposure, to break genuine signals, or selectively choose structures that should not display the phenomenon under study [12]. This strategy ensures that any detected signal isn't merely an artifact of the overall distribution of structural features.

For example, consider a case where you argue that hydrophobic residues in binding sites are inherently less dynamic. Alternative explanations might include differences in solvent exposure, secondary-structure context, or biases introduced by your dataset (for instance, selecting only certain CATH classes or ligand types). A robust negative control would examine hydrophobic residues outside binding pockets matched for solvent accessibility and local secondary structure. While it may be impossible to control every variable perfectly, assessing your metric across complementary subsets is critical for demonstrating that your findings reflect genuine biological effects rather than quirks of data selection.

Recommendation #7: Understand how metrics are compared across your structures

Without careful evaluation, comparison metrics can lead to incorrect conclusions. For instance, larger proteins naturally exhibit higher overall root mean squared distance (RMSD) values, a common metric for comparing the two structures' similarities. Normalizing RMSD by sequence length or reporting RMSD per residue can correct this. Many structure alignment tools, including DALI and TM-align, provide Z-scores indicating the likelihood that an observed similarity would occur by chance [32,93]. Alignment and comparison in torsion space also provide a powerful way to distinguish functionally relevant conformational states. Torsion-angle-based approaches preserve subtle, biologically meaningful differences that are often obscured in atomic coordinate space [94,95].

B-factors, also called temperature factors, atomic displacement parameters, or Debye–Waller factors, estimate each atom's displacement parameter, combining thermal motion of the atom with static disorder from the crystal lattice [96]. Because they arise from the refinement process, B-factors are influenced by data resolution, model bias, occupancy, and lattice packing. As a result, high B-factors do not necessarily guarantee high flexibility in solution. To use them reliably, it's best to normalize B-factors, for example, by Z-scoring within a structure, comparing structures with very similar crystallographic parameters [97,98], and, when possible, corroborating with another metric of flexibility. There are a plethora of other comparison metrics that can be used to compare groups or pairs of PDBs [93,99–101]. Understanding how these metrics are derived and how best to apply them to your analysis is essential to ensuring you use them properly and avoid introducing bias.

Recommendation #8: Appropriately connect and compare structures

When comparing two groups of structures, it is crucial to balance confounding variables to ensure that biological differences, rather than methodological or crystallographic artifacts, drive the observed differences. Differences in resolution, space group, unit-cell parameters, data processing, and data collection parameters can lead to incorrect conclusions. Depending on the question, this can also include differences in local metrics such as MolProbity or validation scores [102]. Even reprocessing identical raw data with identical refinement settings can yield subtly different models due to stochasticity built into those processes to help with the complex refinement optimization process [74,75]. To minimize such artifacts, applying consistent processing pipelines (such as PDB-REDO) and, where possible, matching crystallographic parameters is important.

These controls become even more critical when looking at pairs of structures, such as ligand-bound versus apo or mutant versus wild type. In these analyses, you often look for subtle conformational changes you want to ensure are not driven by nonbiological artifacts. We recommend pairing structures based on biological differences and ensuring that they have similar crystallographic properties. Some general guidelines include using datasets with resolutions within 0.3 Å, identical space groups, and unit cell dimensions that differ by no more than 10%. While these criteria are not always achievable, deviations can introduce artifacts: differences in crystal contacts or solvent volume may affect the conclusions you can draw.

In some cases, it is valuable to collect structures with diverse crystallographic properties from the same or closely related proteins. Such comparisons can provide insight into conformational heterogeneity and, in particular, are useful for studying loop conformations that crystal contacts may influence. By grouping structures into distinct crystal forms, one can analyze loop conformations across different crystallographic contexts and disentangle genuine biological flexibility from artifacts introduced during crystallization [103,104].

Additionally, you must determine how you will compare structures across groups for all comparisons. For most comparisons, you will need to align structures, often based on the alpha carbon; however, other options include aligning the entire structure or taking sequence into account. Global metrics, such as RMSD, allow you to ignore sequence or small length differences, but if you want to compare specific sections of the protein or amino acids, this will take more care and thought. For example, you may want to compare how a specific loop compares among homologs. This will require aligning structures around that loop or to all residues besides the loop, and also ensuring that crystal contacts are not driving these conclusions.

Comparing structures of the same protein, you can compare using chain and residue IDs, but a standard numbering scheme is required. This can be done by manually renumbering chains and residues or employing algorithms such as PDBRenum to map PDB residue numbers onto UniProt numbering, which also allows for integration with other databases (see Recommendation #9) [105]. If PDBs are similar, you can also align them based on a MSA. One thing to note is that while the MSA will enable renumbering, a single residue number may still correspond to different residue types.

Additionally, it is worthwhile to see if existing databases or collections have the comparisons you want. For example, multiple databases pair apo-holo structures together, although depending on your question, you may want to further curate this database down based on crystallographic properties [106].

Recommendation #9: Connect your analysis to other databases or prospective experiments

By connecting PDB structures with other bioinformatics databases, you can enrich your analyses with sequence features, domain architectures, pathway contexts, and chemical insights, uncovering deeper relationships between structure, function, and activity. The PDBe API provides programmatic access to sequence, taxonomy, and functional annotations [5]. Family and domain classifications, including Pfam, SCOP, ECOD, and CATH [14,31,35,107,108], are accessible via SIFTS [35]. SIFTS also offers residue-level mappings between PDB structures and UniProt sequences, enabling the labeling of functional sites onto PDB structures [109]. This facilitates comparative analyses, such as examining conformational changes across a family or correlating structural motifs with functional annotations from Gene Ontology or InterPro [110,111]. PDBs can be connected to pathway and chemical databases such as KEGG and Reactome via UniProt [112,113]. PDBe-KB further consolidates annotations from multiple specialist resources, providing an integrated knowledge base that highlights functional and biological insights mapped onto PDB entries [114]. In addition, the 3D-Beacons network connects structural biology resources across multiple providers, ensuring consistent and federated access to experimental and computational models [115]. While these resources are highly complementary, they are not entirely overlapping, as each database captures different aspects of biological knowledge, and careful integration is often necessary to avoid redundancy or misinterpretation.

Additionally, many PDB structural models have small molecules. PDBe provides excellent ligand pages and tools for analysis within the database [116,117]. Additionally, small molecule information can be linked to existing databases. The PDB's Chemical Component Dictionary assigns ligand IDs that can be cross-referenced with ChEMBL, PubChem, or DrugBank [118–120]. Additionally, external databases such as PDBBind and BindingDB can group chemical or binding information and link it back to PDB information [121,122]. These databases enable easier retrieval of assay data, clinical information, or physicochemical properties of ligands. A growing number of 'curated' databases also look at protein-ligand interactions, post-translational modification, nucleic acid interaction sites, among many others [123–127]. You can then use the pre-calculated metrics or the curated PDB list to calculate the metrics you are interested in.

Additionally, bioinformatics can serve as an excellent partner for hypothesis generation or for supporting prospective experiments. For example, structural bioinformatics can pinpoint the specific residue(s) to mutate to test a desired functional effect, or evaluate whether an experimentally derived hypothesis, such as a loop–domain interaction, holds across homologous structures and influences protein activity.

Recommendation #10: Visualize everything!

One of the best things about structural biology is visualizing what you are discovering. Looking at structures and the metrics you are using via Pymol or Chimera is a powerful quality control tool for your bioinformatic analyses [128,129]. For example, calculating the comparison between two structures and then manually exploring the metric in a visualization software for a given metric. You can ask: Are you aligning the structures or residues correctly? Does the quantification of the metric you are getting make sense? Once you have confirmed that metrics are calculated correctly and you have results you want to show, Pymol, Chimera, or Coot offer various representations for pieces of the molecule, underlying experimental data, and distance measurements [128–130]. PyMOL can also load molecular dynamics trajectories to visualize conformational changes. ChimeraX's plugin infrastructure efficiently handles larger structures.

Discussion

Structural bioinformatics provides a robust framework for identifying patterns in macromolecular structures, integrating with other databases, supporting theoretical approaches, and informing prospective experiments [9,12,131]. For example, overlaying quantitative proteomics and large-scale sequence variation onto structural clusters enables identifying regulatory hotspots and prioritizing functionally relevant variants. Additionally, structural bioinformatics can be incredibly powerful in supporting or refuting hypotheses from prospective experiments. While we did not focus on this, AlphaFold or other structure models can help fill gaps where experimental structures are absent [18–20], including now expanding beyond proteins [18,132]. However, users must remain mindful of the "last-Ångstrom" problem, where these prediction models are often inaccurate in very precise measurement, including molecular interactions, residue networks, and the lack of conformational ensembles stemming from these predicted structures [133,134].

Beyond single-structure analyses, statistical and integrative structural biology approaches can help merge structural models to detect new or more subtle changes in structures or structural ensembles. Further, while most of this article focused on how to detect subtle differences using bioinformatics, these tools can be used to go the other way spatially by integrating cell-scale data to construct multiscale assemblies in their native contexts. We bridge atomistic observations to emergent cellular behaviors, closing the loop between structure, function, and phenotype [135].

Finally, many concepts presented in this paper should also be considered when doing machine learning on protein structures. While protein structure prediction has led to an explosion of machine learning algorithms and approaches applied to structural data, many issues that hinder bioinformatic analyses also arise when splitting datasets in machine learning [92,136,137]. In particular, researchers must carefully avoid information leakage by ensuring that homologous proteins, redundant structures, or closely related crystal forms are not distributed across training and test sets, as this can lead to overly optimistic performance estimates. Incorporating these principles into structural bioinformatics ensures that computational results remain reliable, reproducible, and ultimately informative for guiding experimental design.

Author contributions

Conceptualization: Stephanie A. Wankowicz.

Data curation: Stephanie A. Wankowicz.

Formal analysis: Stephanie A. Wankowicz.

Investigation: Stephanie A. Wankowicz.

Methodology: Stephanie A. Wankowicz.

Project administration: Stephanie A. Wankowicz.

Resources: Stephanie A. Wankowicz.

Software: Stephanie A. Wankowicz.

Validation: Stephanie A. Wankowicz.

Visualization: Stephanie A. Wankowicz.

Writing – original draft: Stephanie A. Wankowicz.

Writing – review & editing: Stephanie A. Wankowicz.

References

1. Bernstein FC, Koetzle TF, Williams GJ, Meyer EF Jr, Brice MD, Rodgers JR, et al. The Protein Data Bank: a computer-based archival file for macromolecular structures. J Mol Biol. 1977;112(3):535–42. https://doi.org/10.1016/s0022-2836(77)80200-3 PMID: 875032
2. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Res. 2000;28(1):235–42. https://doi.org/10.1093/nar/28.1.235 PMID: 10592235
3. Henrick K, Feng Z, Bluhm WF, Dimitropoulos D, Doreleijers JF, Dutta S, et al. Remediation of the protein data bank archive. Nucleic Acids Res. 2008;36(Database issue):D426–33. https://doi.org/10.1093/nar/gkm937 PMID: 18073189
4. Berman H, Henrick K, Nakamura H, Markley JL. The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. Nucleic Acids Res. 2007;35(Database issue):D301–3. https://doi.org/10.1093/nar/gkl971 PMID: 17142228
5. Mir S, Alhroub Y, Anyango S, Armstrong DR, Berrisford JM, Clark AR, et al. PDBe: towards reusable data delivery infrastructure at protein data bank in Europe. Nucleic Acids Res. 2018;46(D1):D486–92. https://doi.org/10.1093/nar/gkx1070 PMID: 29126160
6. Rose Y, Duarte JM, Lowe R, Segura J, Bi C, Bhikadiya C, et al. RCSB Protein Data Bank: architectural advances towards integrated searching and efficient access to macromolecular structure data from the PDB archive. J Mol Biol. 2021;433(11):166704. https://doi.org/10.1016/j.jmb.2020.11.003 PMID: 33186584
7. Kinjo AR, Bekker G-J, Wako H, Endo S, Tsuchiya Y, Sato H, et al. New tools and functions in data-out activities at Protein Data Bank Japan (PDBj). Protein Sci. 2018;27(1):95–102. https://doi.org/10.1002/pro.3273 PMID: 28815765
8. 536–545Young JY, Westbrook JD, Feng Z, Sala R, Peisach E, Oldfield TJ, et al. OneDep: unified wwPDB system for deposition, biocuration, and validation of macromolecular structures in the PDB archive. Structure. 2017;25(3):536–45. https://doi.org/10.1016/j.str.2017.01.004 PMID: 28190782
9. Du S, Kretsch RC, Parres-Gold J, Pieri E, Cruzeiro VWD, Zhu M, et al. Conformational ensembles reveal the origins of serine protease catalysis. Cold Spring Harbor Laboratory; 2024. https://doi.org/10.1101/2024.02.28.582624

10. Modi V, Dunbrack RL Jr. Defining a new nomenclature for the structures of active and inactive kinases. Proc Natl Acad Sci U S A. 2019;116(14):6818–27. https://doi.org/10.1073/pnas.1814279116 PMID: 30867294

11. Brenner SE, Chothia C, Hubbard TJ. Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. Proc Natl Acad Sci U S A. 1998;95(11):6073–8. https://doi.org/10.1073/pnas.95.11.6073 PMID: 9600919

12. Wankowicz SA, de Oliveira SH, Hogan DW, van den Bedem H, Fraser JS. Ligand binding remodels protein side-chain conformational heterogeneity. Elife. 2022;11:e74114. https://doi.org/10.7554/eLife.74114 PMID: 35312477

13. Finkelstein AV, Ptitsyn OB. Why do globular proteins fit the limited set of folding patterns?. Prog Biophys Mol Biol. 1987;50(3):171–90. https://doi.org/10.1016/0079-6107(87)90013-7 PMID: 3332386

14. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol. 1995;247(4):536–40. https://doi.org/10.1006/jmbi.1995.0159 PMID: 7723011

15. Jones S, Thornton JM. Principles of protein-protein interactions. Proc Natl Acad Sci U S A. 1996;93(1):13–20. https://doi.org/10.1073/pnas.93.1.13 PMID: 8552589

16. Jones DT, Taylor WR, Thornton JM. A new approach to protein fold recognition. Nature. 1992;358(6381):86–9. https://doi.org/10.1038/358086a0 PMID: 1614539

17. Alford RF, Leaver-Fay A, Jeliazkov JR, O'Meara MJ, DiMaio FP, Park H, et al. The rosetta all-atom energy function for macromolecular modeling and design. J Chem Theory Comput. 2017;13(6):3031–48. https://doi.org/10.1021/acs.jctc.7b00125 PMID: 28430426

18. Abramson J, Adler J, Dunger J, Evans R, Green T, Pritzel A, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature. 2024;630(8016):493–500. https://doi.org/10.1038/s41586-024-07487-w PMID: 38718835

19. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596(7873):583–9. https://doi.org/10.1038/s41586-021-03819-2 PMID: 34265844

20. Krishna R, Wang J, Ahern W, Sturmfels P, Venkatesh P, Kalvet I, et al. Generalized biomolecular modeling and design with RoseTTAFold All-Atom. Science. 2024;384(6693):eadl2528. https://doi.org/10.1126/science.adl2528 PMID: 38452047

21. Dunbrack RL Jr, Karplus M. Backbone-dependent rotamer library for proteins. Application to side-chain prediction. J Mol Biol. 1993;230(2):543–74. https://doi.org/10.1006/jmbi.1993.1170 PMID: 8464064

22. Westbrook JD, Fitzgerald PMD. The PDB format, mmCIF, and other data formats. Methods Biochem Anal. 2003;44:161–79.

23. UniProt Consortium. UniProt: a hub for protein information. Nucleic Acids Res. 2015;43(Database issue):D204-12. https://doi.org/10.1093/nar/gku989 PMID: 25348405

24. Vallat B, Webb BM, Westbrook JD, Goddard TD, Hanke CA, Graziadei A, et al. IHMCIF: an extension of the PDBx/mmCIF data standard for integrative structure determination methods. J Mol Biol. 2024;436(17):168546. https://doi.org/10.1016/j.jmb.2024.168546 PMID: 38508301

25. Wankowicz SA, Fraser JS. Comprehensive encoding of conformational and compositional protein structural ensembles through the mmCIF data structure. IUCrJ. 2024;11(Pt 4):494–501. https://doi.org/10.1107/S2052252524005098 PMID: 38958015

26. Nomburg J, Doherty EE, Price N, Bellieny-Rabelo D, Zhu YK, Doudna JA. Birth of protein folds and functions in the virome. Nature. 2024;633(8030):710–7. https://doi.org/10.1038/s41586-024-07809-y PMID: 39187718

27. Monzon V, Haft DH, Bateman A. Folding the unfoldable: using AlphaFold to explore spurious proteins. Bioinform Adv. 2022;2(1):vbab043. https://doi.org/10.1093/bioadv/vbab043 PMID: 36699409

28. Osmanli Z, Falgarone T, Samadova T, Aldrian G, Leclercq J, Shahmuradov I, et al. The difference in structural states between canonical proteins and their isoforms established by proteome-wide bioinformatics analysis. Biomolecules. 2022;12(11):1610. https://doi.org/10.3390/biom12111610 PMID: 36358962

29. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012;28(23):3150–2. https://doi.org/10.1093/bioinformatics/bts565 PMID: 23060610

30. Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol. 2017;35(11):1026–8. https://doi.org/10.1038/nbt.3988 PMID: 29035372

31. 1093–1108Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. CATH—a hierarchic classification of protein domain structures. Structure. 1997;5(8):1093–108. https://doi.org/10.1016/s0969-2126(97)00260-8 PMID: 9309224

32. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res. 2005;33(7):2302–9. https://doi.org/10.1093/nar/gki524 PMID: 15849316

33. Wang G, Dunbrack RL Jr. PISCES: a protein sequence culling server. Bioinformatics. 2003;19(12):1589–91. https://doi.org/10.1093/bioinformatics/btg224 PMID: 12912846

34. Sievers F, Higgins DG. Clustal Omega for making accurate alignments of many protein sequences. Protein Sci. 2018;27(1):135–45. https://doi.org/10.1002/pro.3290 PMID: 28884485

35. Velankar S, Dana JM, Jacobsen J, van Ginkel G, Gane PJ, Luo J, et al. SIFTS: structure integration with function, taxonomy and sequences resource. Nucleic Acids Res. 2013;41(Database issue):D483-9. https://doi.org/10.1093/nar/gks1258 PMID: 23203869

36. Lo Conte L, Ailey B, Hubbard TJ, Brenner SE, Murzin AG, Chothia C. SCOP: a structural classification of proteins database. Nucleic Acids Res. 2000;28(1):257–9. https://doi.org/10.1093/nar/28.1.257 PMID: 10592240

37. Bittrich S, Segura J, Duarte JM, Burley SK, Rose Y. RCSB protein Data Bank: exploring protein 3D similarities via comprehensive structural alignments. Bioinformatics. 2024;40(6):btae370. https://doi.org/10.1093/bioinformatics/btae370 PMID: 38870521

38. Li Z, Jaroszewski L, Iyer M, Sedova M, Godzik A. FATCAT 2.0: towards a better understanding of the structural diversity of proteins. Nucleic Acids Res. 2020;48(W1):W60–4. https://doi.org/10.1093/nar/gkaa443 PMID: 32469061

39. Liu Z, Zhang C, Zhang Q, Zhang Y, Yu D-J. TM-search: an efficient and effective tool for protein structure database search. J Chem Inf Model. 2024;64(3):1043–9. https://doi.org/10.1021/acs.jcim.3c01455 PMID: 38270339

40. Guda C, Lu S, Scheeff ED, Bourne PE, Shindyalov IN. CE-MC: a multiple protein structure alignment server. Nucleic Acids Res. 2004;32(Web Server issue):W100-3. https://doi.org/10.1093/nar/gkh464 PMID: 15215359

41. Topham CM, Rouquier M, Tarrat N, André I. Adaptive Smith-Waterman residue match seeding for protein structural alignment. Proteins. 2013;81(10):1823–39. https://doi.org/10.1002/prot.24327 PMID: 23720362

42. Piehl DW, Vallat B, Truong I, Morsy H, Bhatt R, Blaumann S, et al. rcsb-api: Python toolkit for streamlining access to RCSB Protein Data Bank APIs. J Mol Biol. 2025;437(15):168970. https://doi.org/10.1016/j.jmb.2025.168970 PMID: 39894387

43. Rosenthal PB, Henderson R. Optimal determination of particle orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy. J Mol Biol. 2003;333(4):721–45. https://doi.org/10.1016/j.jmb.2003.07.013 PMID: 14568533

44. Kucukelbir A, Sigworth FJ, Tagare HD. Quantifying the local resolution of cryo-EM density maps. Nat Methods. 2014;11(1):63–5. https://doi.org/10.1038/nmeth.2727 PMID: 24213166

45. Gore S, et al. Validation of structures in the Protein Data Bank. Structure. 2017;25:1916–27.

46. 1563–1570Montelione GT, et al. Recommendations of the wwPDB NMR validation task force. Structure. 2013;21:1563–70.

47. 205–214Henderson R, Sali A, Baker ML, Carragher B, Devkota B, Downing KH, et al. Outcome of the first electron microscopy validation task force meeting. Structure. 2012;20(2):205–14. https://doi.org/10.1016/j.str.2011.12.014 PMID: 22325770

48. Gutmanas A, Adams PD, Bardiaux B, Berman HM, Case DA, Fogh RH, et al. NMR Exchange Format: a unified and open standard for representation of NMR restraint data. Nat Struct Mol Biol. 2015;22(6):433–4. https://doi.org/10.1038/nsmb.3041 PMID: 26036565

49. 1395–1412Read RJ, Adams PD, Arendall WB 3rd, Brunger AT, Emsley P, Joosten RP, et al. A new generation of crystallographic validation tools for the protein data bank. Structure. 2011;19(10):1395–412. https://doi.org/10.1016/j.str.2011.08.006 PMID: 22000512

50. Williams CJ, Headd JJ, Moriarty NW, Prisant MG, Videau LL, Deis LN, et al. MolProbity: more and better reference data for improved all-atom structure validation. Protein Sci. 2018;27(1):293–315. https://doi.org/10.1002/pro.3330 PMID: 29067766

51. Laskowski RA, MacArthur MW, Moss DS, Thornton JM. PROCHECK: a program to check the stereochemical quality of protein structures. J Appl Crystallogr. 1993;26(2):283–91. https://doi.org/10.1107/s0021889892009944

52. Meyder A, Nittinger E, Lange G, Klein R, Rarey M. Estimating electron density support for individual atoms and molecular fragments in X-ray structures. J Chem Inf Model. 2017;57(10):2437–47. https://doi.org/10.1021/acs.jcim.7b00391 PMID: 28981269

53. Brünger AT. Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. Nature. 1992;355(6359):472–5. https://doi.org/10.1038/355472a0 PMID: 18481394

54. Prisant MG, Williams CJ, Chen VB, Richardson JS, Richardson DC. New tools in MolProbity validation: CaBLAM for CryoEM backbone, UnDowser to rethink “waters,” and NGL Viewer to recapture online 3D graphics. Protein Sci. 2020;29(1):315–29. https://doi.org/10.1002/pro.3786 PMID: 31724275

55. Lang PT, Ng H-L, Fraser JS, Corn JE, Echols N, Sales M, et al. Automated electron-density sampling reveals widespread conformational polymorphism in proteins. Protein Sci. 2010;19(7):1420–31. https://doi.org/10.1002/pro.423 PMID: 20499387

56. Tickle IJ. Statistical quality indicators for electron-density maps. Acta Crystallogr D Biol Crystallogr. 2012;68(Pt 4):454–67. https://doi.org/10.1107/S0907444911035918 PMID: 22505266

57. Kleywegt GJ, Harris MR, Zou JY, Taylor TC, Wählby A, Jones TA. The Uppsala electron-density server. Acta Crystallogr D Biol Crystallogr. 2004;60(Pt 12 Pt 1):2240–9. https://doi.org/10.1107/S0907444904013253 PMID: 15572777

58. Lander GC. Single particle cryo-EM map and model validation: It’s not crystal clear. Curr Opin Struct Biol. 2024;89:102918. https://doi.org/10.1016/j.sbi.2024.102918 PMID: 39293191

59. Grigorieff N. Resolution measurement in structures derived from single particles. Acta Crystallogr D Biol Crystallogr. 2000;56(Pt 10):1270–7. https://doi.org/10.1107/s0907444900009549 PMID: 10998623

60. Barad BA, Echols N, Wang RY-R, Cheng Y, DiMaio F, Adams PD, et al. EMRinger: side chain-directed model and map validation for 3D cryo-electron microscopy. Nat Methods. 2015;12(10):943–6. https://doi.org/10.1038/nmeth.3541 PMID: 26280328

61. Pintilie G, Zhang K, Su Z, Li S, Schmid MF, Chiu W. Measurement of atom resolvability in cryo-EM maps with Q-scores. Nat Methods. 2020;17(3):328–34. https://doi.org/10.1038/s41592-020-0731-1 PMID: 32042190

62. Brandt BW, Heringa J, Leunissen JAM. SEQATOMS: a web tool for identifying missing regions in PDB in sequence context. Nucleic Acids Res. 2008;36(Web Server issue):W255-9. https://doi.org/10.1093/nar/gkn237 PMID: 18463137

63. van Beusekom B, Joosten K, Hekkelman ML, Joosten RP, Perrakis A. Homology-based loop modeling yields more complete crystallographic protein structures. IUCrJ. 2018;5(Pt 5):585–94. https://doi.org/10.1107/S2052252518010552 PMID: 30224962

64. Liebeschuetz JW. The good, the bad, and the twisted revisited: an analysis of ligand geometry in highly resolved protein-ligand X-ray structures. J Med Chem. 2021;64(11):7533–43. https://doi.org/10.1021/acs.jmedchem.1c00228 PMID: 34060310

65. 252–262Shao C, Westbrook JD, Lu C, Bhikadiya C, Peisach E, Young JY, et al. Simplified quality assessment for small-molecule ligands in the Protein Data Bank. Structure. 2022;30(2):252-262.e4. https://doi.org/10.1016/j.str.2021.10.003 PMID: 35026162

66. Groom CR, Bruno IJ, Lightfoot MP, Ward SC. The Cambridge structural database. Acta Crystallogr B Struct Sci Cryst Eng Mater. 2016;72(Pt 2):171–9. https://doi.org/10.1107/S2052520616003954 PMID: 27048719

67. Zheng H, Cooper DR, Porebski PJ, Shabalin IG, Handing KB, Minor W. CheckMyMetal: a macromolecular metal-binding validation tool. Acta Crystallogr D Struct Biol. 2017;73(Pt 3):223–33. https://doi.org/10.1107/S2059798317001061 PMID: 28291757

68. de Vries I, Kwakman T, Lu XJ, Hekkelman ML, Deshpande M, Velankar S, et al. New restraints and validation approaches for nucleic acid structures in PDB-REDO. Acta Crystallogr D Struct Biol. 2021;77(Pt 9):1127–41. https://doi.org/10.1107/S2059798321007610 PMID: 34473084

69. Černý J, Božíková P, Schneider B. DNATCO: assignment of DNA conformers at dnatco.org. Nucleic Acids Res. 2016;44(W1):W284-7. https://doi.org/10.1093/nar/gkw381 PMID: 27150812

70. Joosten RP, Vriend G. PDB improvement starts with data deposition. Science. 2007;317(5835):195–6. https://doi.org/10.1126/science.317.5835.195 PMID: 17626865

71. Joosten RP, 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. 2009;65(Pt 2):176–85. https://doi.org/10.1107/S0907444908037591 PMID: 19171973

72. Joosten RP, Salzemann J, Bloch V, Stockinger H, Berglund A-C, Blanchet C, et al. PDB_REDO: automated re-refinement of X-ray structure models in the PDB. J Appl Crystallogr. 2009;42(Pt 3):376–84. https://doi.org/10.1107/S0021889809008784 PMID: 22477769

73. Jiang J, Abola E, Sussman JL. Deposition of structure factors at the Protein Data Bank. Acta Crystallogr D Biol Crystallogr. 1999;55(Pt 1):4. https://doi.org/10.1107/S0907444998016631 PMID: 10089388

74. Murshudov GN, Skubák P, Lebedev AA, Pannu NS, Steiner RA, Nicholls RA, et al. REFMAC5 for the refinement of macromolecular crystal structures. Acta Crystallogr D Biol Crystallogr. 2011;67(Pt 4):355–67. https://doi.org/10.1107/S0907444911001314 PMID: 21460454

75. Adams PD, Afonine PV, Bunkóczi G, Chen VB, Davis IW, Echols N, et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta Crystallogr D Biol Crystallogr. 2010;66(Pt 2):213–21. https://doi.org/10.1107/S0907444909052925 PMID: 20124702

76. Shabalin IG, Porebski PJ, Minor W. Refining the macromolecular model—achieving the best agreement with the data from X-ray diffraction experiment. Crystallogr Rev. 2018;24(4):236–62. https://doi.org/10.1080/0889311X.2018.1521805 PMID: 30416256

77. Du S, Wankowicz SA, Yabukarski F, Doukov T, Herschlag D, Fraser JS. Refinement of multiconformer ensemble models from multi-temperature X-ray diffraction data. Methods Enzymol. 2023;688:223–54. https://doi.org/10.1016/bs.mie.2023.06.009 PMID: 37748828

78. Wankowicz SA, Ravikumar A, Sharma S, Riley B, Raju A, Hogan DW, et al. Automated multiconformer model building for X-ray crystallography and cryo-EM. Elife. 2024;12:RP90606. https://doi.org/10.7554/eLife.90606 PMID: 38904665

79. Burnley BT, Afonine PV, Adams PD, Gros P. Modelling dynamics in protein crystal structures by ensemble refinement. Elife. 2012;1:e00311. https://doi.org/10.7554/eLife.00311 PMID: 23251785

80. Zubatyuk R, Biczysko M, Ranasinghe K, Moriarty NW, Gokcan H, Kruse H, et al. AQuaRef: machine learning accelerated quantum refinement of protein structures. bioRxiv. 2025;:2024.07.21.604493. https://doi.org/10.1101/2024.07.21.604493 PMID: 39071315

81. Hoch JC, et al. Biological magnetic resonance data bank. Nucleic Acids Res. 2023;51:D368–76.

82. wwPDB Consortium. EMDB-the Electron Microscopy Data Bank. Nucleic Acids Res. 2024;52(D1):D456–65. https://doi.org/10.1093/nar/gkad1019 PMID: 37994703

83. Choy BC, Cater RJ, Mancia F, Pryor EE. A 10-year meta-analysis of membrane protein structural biology: detergents, membrane mimetics, and structure determination techniques. Biochim Biophys Acta Biomembr. 2021;1863:183533.

84. Erlanson D, Burley S, Fraser J, Fearon D, Kreitler D, Nonato MC, et al. Where to house big data on small fragments?. American Chemical Society (ACS); 2025. https://doi.org/10.26434/chemrxiv-2025-hjjnj

85. Krissinel E, Henrick K. Inference of macromolecular assemblies from crystalline state. J Mol Biol. 2007;372(3):774–97. https://doi.org/10.1016/j.jmb.2007.05.022 PMID: 17681537

86. Xu Q, Dunbrack RL Jr. The protein common interface database (ProtCID)—a comprehensive database of interactions of homologous proteins in multiple crystal forms. Nucleic Acids Res. 2011;39(Database issue):D761–70. https://doi.org/10.1093/nar/gkq1059 PMID: 21036862

87. Xu Q, Dunbrack RL. The protein common assembly database (ProtCAD)-a comprehensive structural resource of protein complexes. Nucleic Acids Res. 2023;51(D1):D466–78. https://doi.org/10.1093/nar/gkac937 PMID: 36300618

88. Wankowicz SA. Modeling bias toward binding sites in PDB structural models. Cold Spring Harbor Laboratory; 2024. https://doi.org/10.1101/2024.12.14.628518

89. Woldeyes RA, Sivak DA, Fraser JS. E pluribus unum, no more: from one crystal, many conformations. Curr Opin Struct Biol. 2014;28:56–62. https://doi.org/10.1016/j.sbi.2014.07.005 PMID: 25113271

90. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009;25(11):1422–3. https://doi.org/10.1093/bioinformatics/btp163 PMID: 19304878

91. Gaudreault F, Chartier M, Najmanovich R. Side-chain rotamer changes upon ligand binding: common, crucial, correlate with entropy and rear-range hydrogen bonding. Bioinformatics. 2012;28(18):i423–30. https://doi.org/10.1093/bioinformatics/bts395 PMID: 22962462

92. Buttenschoen M, Morris GM, Deane CM. PoseBusters: AI-based docking methods fail to generate physically valid poses or generalise to novel sequences. Chem Sci. 2023;15(9):3130–9. https://doi.org/10.1039/d3sc04185a PMID: 38425520

93. Holm L, Sander C. Dali: a network tool for protein structure comparison. Trends Biochem Sci. 1995;20(11):478–80. https://doi.org/10.1016/s0968-0004(00)89105-7 PMID: 8578593

94. Ginn HM. Torsion angles to map and visualize the conformational space of a protein. Protein Sci. 2023;32(4):e4608. https://doi.org/10.1002/pro.4608 PMID: 36840926

95. Nicholls RA, Fischer M, McNicholas S, Murshudov GN. Conformation-independent structural comparison of macromolecules with ProSMART. Acta Crystallogr D Biol Crystallogr. 2014;70(Pt 9):2487–99. https://doi.org/10.1107/S1399004714016241 PMID: 25195761

96. Sun Z, Liu Q, Qu G, Feng Y, Reetz MT. Utility of B-factors in protein science: interpreting rigidity, flexibility, and internal motion and engineering thermostability. Chem Rev. 2019;119(3):1626–65. https://doi.org/10.1021/acs.chemrev.8b00290 PMID: 30698416

97. Ringe D, Petsko GA. Study of protein dynamics by X-ray diffraction. Methods Enzymol. 1986;131:389–433. https://doi.org/10.1016/0076-6879(86)31050-4 PMID: 3773767

98. Carugo O. B-factor accuracy in protein crystal structures. Acta Crystallogr D Struct Biol. 2022;78(Pt 1):69–74. https://doi.org/10.1107/S2059798321011736 PMID: 34981763

99. 565–571Tyzack JD, Fernando L, Ribeiro AJM, Borkakoti N, Thornton JM. Ranking enzyme structures in the PDB by bound ligand similarity to biological substrates. Structure. 2018;26(4):565-571.e3. https://doi.org/10.1016/j.str.2018.02.009 PMID: 29551288

100. Yeturu K, Chandra N. PocketMatch: a new algorithm to compare binding sites in protein structures. BMC Bioinformatics. 2008;9:543. https://doi.org/10.1186/1471-2105-9-543 PMID: 19091072

101. Meslamani J, Rognan D, Kellenberger E. sc-PDB: a database for identifying variations and multiplicity of “druggable” binding sites in proteins. Bioinformatics. 2011;27(9):1324–6. https://doi.org/10.1093/bioinformatics/btr120 PMID: 21398668

102. Chen VB, Arendall WB 3rd, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, et al. MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr D Biol Crystallogr. 2010;66(Pt 1):12–21. https://doi.org/10.1107/S0907444909042073 PMID: 20057044

103. Creon A, Scheer TES, Reinke P, Mashhour AR, Günther S, Niebling S, et al. Statistical crystallography reveals an allosteric network in SARS-CoV-2 Mpro. Cold Spring Harbor Laboratory; 2025. https://doi.org/10.1101/2025.01.28.635305

104. Yabukarski F, Doukov T, Pinney MM, Biel JT, Fraser JS, Herschlag D. Ensemble-function relationships to dissect mechanisms of enzyme catalysis. Sci Adv. 2022;8(41):eabn7738. https://doi.org/10.1126/sciadv.abn7738 PMID: 36240280

105. Faezov B, Dunbrack RL Jr. PDBrenum: a webserver and program providing Protein Data Bank files renumbered according to their UniProt sequences. PLoS One. 2021;16(7):e0253411. https://doi.org/10.1371/journal.pone.0253411 PMID: 34228733

106. Feidakis CP, Krivak R, Hoksza D, Novotny M. AHoJ-DB: a PDB-wide assignment of apo & holo relationships based on individual protein-ligand interactions. J Mol Biol. 2024;436(17):168545. https://doi.org/10.1016/j.jmb.2024.168545 PMID: 38508305

107. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, et al. Pfam: the protein families database in 2021. Nucleic Acids Res. 2021;49(D1):D412–9. https://doi.org/10.1093/nar/gkaa913 PMID: 33125078

108. Cheng H, Schaeffer RD, Liao Y, Kinch LN, Pei J, Shi S, et al. ECOD: an evolutionary classification of protein domains. PLoS Comput Biol. 2014;10(12):e1003926. https://doi.org/10.1371/journal.pcbi.1003926 PMID: 25474468

109. Choudhary P, Anyango S, Berrisford J, Tolchard J, Varadi M, Velankar S. Unified access to up-to-date residue-level annotations from UniProtKB and other biological databases for PDB data. Sci Data. 2023;10(1):204. https://doi.org/10.1038/s41597-023-02101-6 PMID: 37045837

110. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000;25(1):25–9. https://doi.org/10.1038/75556 PMID: 10802651

111. Paysan-Lafosse T, et al. InterPro in 2022. Nucleic Acids Res. 2023;51:D418–27.

112. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 2017;45(D1):D353–61. https://doi.org/10.1093/nar/gkw1092 PMID: 27899662

113. Gillespie M, et al. The reactome pathway knowledgebase 2022. Nucleic Acids Res. 2022;50:D687–92.

114. PDBe-KB consortium. PDBe-KB: a community-driven resource for structural and functional annotations. Nucleic Acids Res. 2020;48:D344–53.

115. Varadi M, Nair S, Sillitoe I, Tauriello G, Anyango S, Bienert S, et al. 3D-Beacons: decreasing the gap between protein sequences and structures through a federated network of protein structure data resources. Gigascience. 2022;11:giac118. https://doi.org/10.1093/gigascience/giac118 PMID: 36448847

116. Choudhary P, Kunnakkattu IR, Nair S, Lawal DK, Pidruchna I, Afonso MQL, et al. PDBe tools for an in-depth analysis of small molecules in the Protein Data Bank. Protein Sci. 2025;34(4):e70084. https://doi.org/10.1002/pro.70084 PMID: 40100137

117. Kunnakkattu IR, Choudhary P, Pravda L, Nadzirin N, Smart OS, Yuan Q, et al. PDBe CCDUtils: an RDKit-based toolkit for handling and analysing small molecules in the Protein Data Bank. J Cheminform. 2023;15(1):117. https://doi.org/10.1186/s13321-023-00786-w PMID: 38042830

118. Gaulton A, et al. The ChEMBL database in 2017. Nucleic Acids Res. 2017;45:D945–54.

119. Knox C, Wilson M, Klinger CM, Franklin M, Oler E, Wilson A, et al. DrugBank 6.0: the DrugBank knowledgebase for 2024. Nucleic Acids Res. 2024;52(D1):D1265–75. https://doi.org/10.1093/nar/gkad976 PMID: 37953279

120. Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, et al. PubChem 2019 update: improved access to chemical data. Nucleic Acids Res. 2019;47(D1):D1102–9. https://doi.org/10.1093/nar/gky1033 PMID: 30371825

121. Liu Z, Li Y, Han L, Li J, Liu J, Zhao Z, et al. PDB-wide collection of binding data: current status of the PDBbind database. Bioinformatics. 2015;31(3):405–12. https://doi.org/10.1093/bioinformatics/btu626 PMID: 25301850

122. Liu T, Hwang L, Burley SK, Nitsche CI, Southan C, Walters WP, et al. BindingDB in 2024: a FAIR knowledgebase of protein-small molecule binding data. Nucleic Acids Res. 2025;53(D1):D1633–44. https://doi.org/10.1093/nar/gkae1075 PMID: 39574417

123. Zhang C, Zhang X, Freddolino L, Zhang Y. BioLiP2: an updated structure database for biologically relevant ligand-protein interactions. Nucleic Acids Res. 2024;52(D1):D404–12. https://doi.org/10.1093/nar/gkad630 PMID: 37522378

124. Joosten RP, te Beek TAH, Krieger E, Hekkelman ML, Hooft RWW, Schneider R, et al. A series of PDB related databases for everyday needs. Nucleic Acids Res. 2011;39(Database issue):D411-9. https://doi.org/10.1093/nar/gkq1105 PMID: 21071423

125. Li F, Fan C, Marquez-Lago TT, Leier A, Revote J, Jia C, et al. PRISMOID: a comprehensive 3D structure database for post-translational modifications and mutations with functional impact. Brief Bioinform. 2020;21(3):1069–79. https://doi.org/10.1093/bib/bbz050 PMID: 31161204

126. Mitra R, Cohen AS, Tang WY, Hosseini H, Hong Y, Berman HM, et al. RNAproDB: a webserver and interactive database for analyzing protein-RNA interactions. J Mol Biol. 2025;:169012. https://doi.org/10.1016/j.jmb.2025.169012 PMID: 40126909

127. Jendele L, Krivak R, Skoda P, Novotny M, Hoksza D. PrankWeb: a web server for ligand binding site prediction and visualization. Nucleic Acids Res. 2019;47(W1):W345–9. https://doi.org/10.1093/nar/gkz424 PMID: 31114880

128. Lill MA, Danielson ML. Computer-aided drug design platform using PyMOL. J Comput Aided Mol Des. 2011;25(1):13–9. https://doi.org/10.1007/s10822-010-9395-8 PMID: 21053052

129. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, et al. UCSF Chimera—a visualization system for exploratory research and analysis. J Comput Chem. 2004;25(13):1605–12. https://doi.org/10.1002/jcc.20084 PMID: 15264254

130. Casañal A, Lohkamp B, Emsley P. Current developments in Coot for macromolecular model building of Electron Cryo-microscopy and Crystallographic Data. Protein Sci. 2020;29(4):1069–78. https://doi.org/10.1002/pro.3791 PMID: 31730249

131. Lovell SC, Word JM, Richardson JS, Richardson DC. The penultimate rotamer library. Proteins. 2000;40(3):389–408. https://doi.org/10.1002/1097-0134(20000815)40:3<389::aid-prot50>3.3.co;2-u PMID: 10861930

132. Hekkelman ML, de Vries I, Joosten RP, Perrakis A. AlphaFill: enriching AlphaFold models with ligands and cofactors. Nat Methods. 2023;20(2):205–13. https://doi.org/10.1038/s41592-022-01685-y PMID: 36424442

133. Lyu N, Du S, Ma J, Herschlag D. An evaluation of biomolecular energetics learned by AlphaFold. Cold Spring Harbor Laboratory; 2025. https://doi.org/10.1101/2025.06.30.662466

134. Wankowicz SA, Bonomi M. From possibility to precision in macromolecular ensemble prediction. 2025. https://doi.org/10.48550/ARXIV.2505.01919

135. Young LN, Villa E. Bringing structure to cell biology with cryo-electron tomography. Annu Rev Biophys. 2023;52:573–95. https://doi.org/10.1146/annurev-biophys-111622-091327 PMID: 37159298

136. Chakravarty D, Schafer JW, Chen EA, Thole JF, Ronish LA, Lee M, et al. AlphaFold predictions of fold-switched conformations are driven by structure memorization. Nat Commun. 2024;15(1):7296. https://doi.org/10.1038/s41467-024-51801-z PMID: 39181864

137. Sellner MS, Lill MA, Smieško M. Quality matters: deep learning-based analysis of protein-ligand interactions with focus on avoiding bias. Cold Spring Harbor Laboratory; 2023. https://doi.org/10.1101/2023.11.13.566916