An Evaluation of Biomolecular Energetics Learned by AlphaFold

Authors: Ningyi Lyu†[1,2], Siyuan Du†[4,5]*, Jianpeng Ma[1,2,3]*, Dan Herschlag[4,5,6,7]*

Affiliations:

[1]Multiscale Research Institute of Complex Systems, Fudan University, Shanghai 200433, China

[2]Zhangjiang Fudan International Innovation Center, Fudan University, Shanghai 200433, China

[3]Shanghai AI Laboratory, Shanghai 200233, China

[4]Department of Biochemistry, Stanford University, Stanford, California 94305, United States.

[5]Department of Chemistry, Stanford University, Stanford, California 94305, United States.

[6]Department of Chemical Engineering, Stanford University, Stanford, California 94305, United States.

[7]Stanford ChEM-H, Stanford University, Stanford, California 94305, United States.

†These authors contributed equally to this work.

*Corresponding authors. Email: dusiyuan@stanford.edu, jpma@fudan.edu.cn, herschla@stanford.edu

Abstract
Deep learning has revolutionized protein structural prediction, with function prediction on the horizon. Because biomolecular properties emerge from atomic-level interactions and their energetics, models that learn physical principles can deliver accurate structural predictions while also providing the foundation for function prediction. We systematically evaluated the state-of-the-art structural prediction models AlphaFold2 and AlphaFold3, interrogating >3.4 million interactions across 3939 structures. These models capture basic energetic principles but show pervasive biases in the conformational preferences of covalent and non-covalent interactions. The conformational biases manifest as mis-assignments for nearly one-third of side-chain•side-chain interaction partners. Inaccurate energetics are further evidenced by the inability of AlphaFold3 sampling to reproduce experimentally derived ensembles. Our multifaceted, physics-based evaluation identified previously unknown and system-wide limitations in AlphaFold's ability to make physically accurate predictions. This information will allow researchers to judiciously apply AlphaFold and guide next-generation models to learn the energetics needed for biomolecular function prediction.

Introduction

Accurate prediction of biomolecular behaviors—from protein and RNA folding to ligand binding, enzyme catalysis, and allosteric regulation—will transform industry and medicine. Deep learning holds great potential for achieving this goal, as exemplified by the astounding success of AlphaFold and related algorithms in protein structural prediction (*1–4*). This success has led to great excitement and widespread ongoing efforts extending beyond structure to predict biomolecular function [*e.g.* (*5–7*)].

A growing number of studies have used AlphaFold predictions or repurposed the models to predict aspects of protein function, including folding stability (*8*, *9*), ligand binding affinity (*10*, *11*), and alternative conformations (*12–14*); nevertheless, a similar number of tests have suggested a limited ability to extrapolate AlphaFold to these applications [*e.g.* (*15–17*); table S1]. Given the large user base of AlphaFold and rapidly expanding efforts to develop new biomolecular models, there is an urgent need for systematic evaluations that can determine the capabilities and limitations of the models and thus can guide their use and their improvement. Ideally, these evaluations would use metrics that are tightly connected to the properties that reflect the model's predictive power, rather than indirect surrogates or correlates of these properties.

Structures are atoms arranged to form interactions—covalent bonds, backbone and side-chain torsions, hydrogen bonds, van der Waals interactions, *etc.*—with the sum of their energetic contributions representing the minimum on an energy landscape (Fig. 1A). Structural prediction, in its most accurate form, captures this energy minimum. The same physical principles govern functional processes: for example, ligand binding introduces new interactions, and the sum of the energetic differences between the interactions in the apo and bound states determines the ligand's binding affinity (Fig. 1A). To capture molecular interactions, physics-based models specify a set of energetic rules, typically in the form of force fields (*18*) or energy functions (*19*). By contrast, deep learning models need to extract the energetics of molecular interactions from their training data to predict physically accurate structures and functional properties. Yet, we do not know the extent of their knowledge of these physical principles and the prediction errors that arise from limits to this knowledge.

Physics-based evaluations have been critical for developing predictive models even for systems much simpler than the structure and function of biomolecules. Early models of small molecule generation were limited in their value for drug discovery because a large fraction of the generated molecules were found to give invalid valence and highly strained geometries (*20*, *21*). To address this issue, a series of benchmarks placed physics at the forefront of model evaluation (*20*, *22*, *23*) and guided the development of subsequent models that nearly fully eliminated valence errors (*24*, *25*) and greatly reduced high-energy molecular arrangements (*26*, *27*).

Biomolecular models are challenged to learn more complex interactions than small molecules as well as higher-order interaction networks, yet the current evaluation paradigm does not incorporate physics. Indeed, assessments of structure-based models near-universally use root-mean-square deviations (RMSD) and analogous metrics that reduce multi-dimensional conformational differences to a single value and so cannot evaluate the individual interactions and energy terms [*e.g.* (*1–4*, *28*); see also table S2]. Consequently, improvements in RMSDs do not necessarily lead to improvements in the prediction of functional properties. For example, AlphaFold2 predictions gave lower RMSDs than traditional homology models when compared to the experimentally determined structures, but these AlphaFold2 predictions did not give more accurate ligand-bound poses than the homology models (*29*).

To determine the capabilities and limitations of deep learning biomolecular models and guide their application and development, we provide a physics-based framework that systematically evaluates whether models have learned the biomolecular energetics that define structure and function. We applied this approach to evaluate AlphaFold2 and AlphaFold3, interrogating >3.4 million molecular interactions across 3939 structures. We found that the models have extracted energetic properties for multiple types of covalent and non-covalent interactions, despite their RMSD-based training. Nevertheless, our analyses revealed systematic discrepancies in the conformational preferences of interactions derived from both models. These differences are sufficiently consequential to result in prevalent prediction errors in molecular interactions and interaction networks—including the misassignment of nearly one-third of side-chain non-covalent interactions—and an inability to predict conformational ensembles. This evaluation framework is generalizable to any model that delivers atomic-scale structural information, providing a new paradigm to define the capabilities and limitations of current models and guide the development of next-generation models to capture the biomolecular energetics needed for accurate and quantitative function prediction.

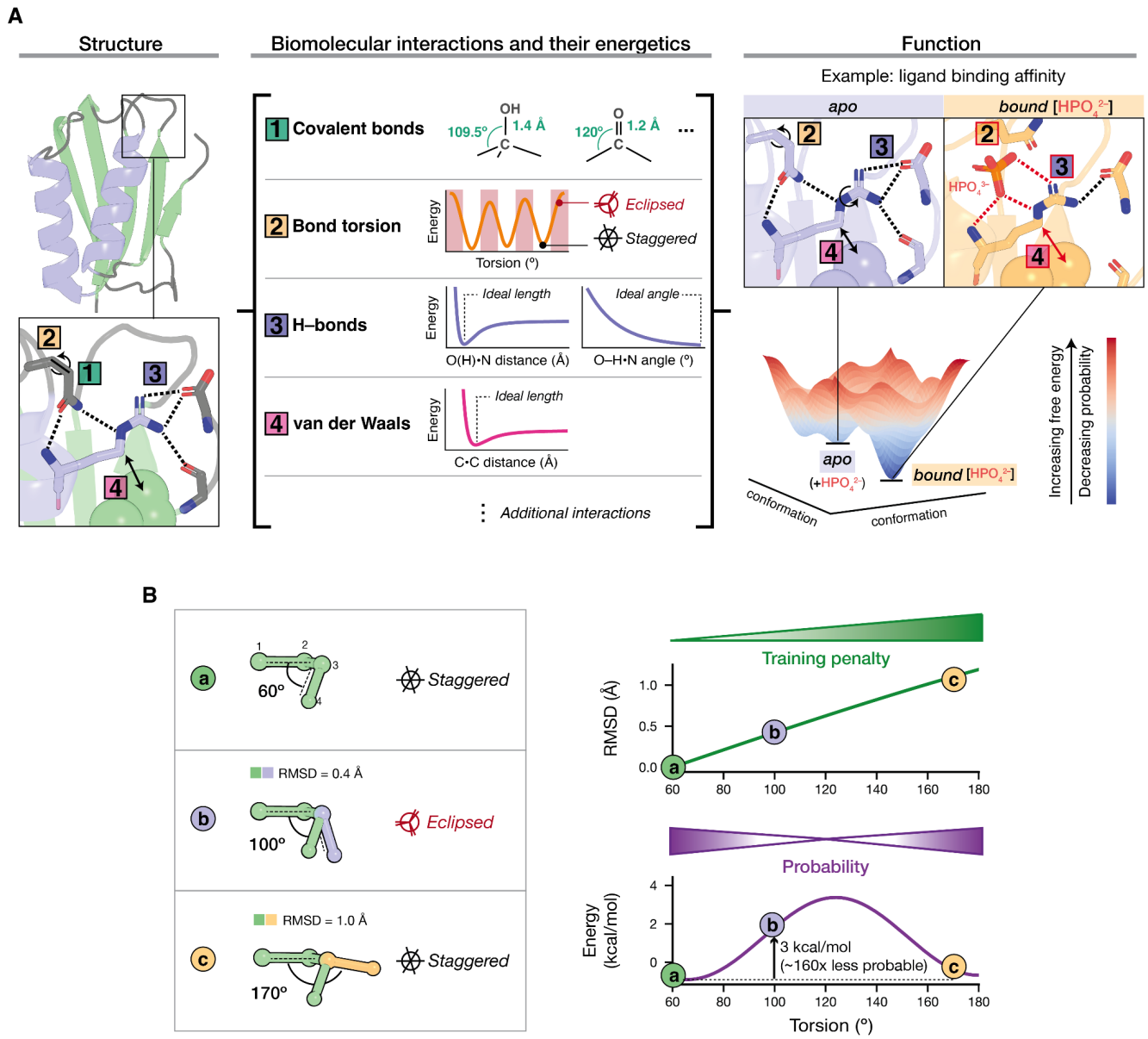

Fig. 1. Molecular interactions and their energetics underlie protein structure and function, but model training and evaluation used RMSD-based metrics rather than interactions and their probabilities. **(A)** Structures are defined by molecular interactions (left; PDB: 2FHM) with each interaction following distinct energetic rules that determine its conformational preferences (middle). When combined, these energetic rules define the multi-dimensional conformational landscape that encodes protein function; a simplified two-dimensional energy landscape for ligand binding is shown as an example (right). Different interactions formed in the apo and the bound state of the protein and determine the free energy difference between the two states (i.e., binding affinity) (acylphosphatase, PDB: 2FHM and 3BR8; red dashed lines and numbered boxes indicate different interactions formed in the bound state *versus* the apo state). The colormap indicates the free energy of the conformational states, from blue (low energy, high probability) to red (high energy, low probability). **(B)** The torsional rotation around a bond and the corresponding changes in RMSD and energy. The left subpanel shows a bond rotating from its original state (a) to states (b) and (c). This rotation is quantified by the torsion angle (defined by atoms 1–2–3–4)

that changes from 60º to 100º and to 170º. The right subpanel compares the changes in RMSD (top) and energy (bottom). The energy values are calculated from a smoothed knowledge-based energy function derived from crystallographic data (T = 298 *K*), as described in *Methods*. The green color map from light to dark green indicates the increased penalty during training; the purple color map from light to dark purple indicates the direction of increased probability (lower energy).

Results and Discussion

An evaluation framework based on molecular interactions and their energetics

To assess the ability of AlphaFold models to reproduce molecular interactions and their energetics, we used evaluation metrics that directly map to energetic properties—covalent bond length, angle and torsion, hydrogen bond length and angle, and van der Waals distance (Fig. 1A). Interaction-based metrics also provide an independent test because they are not directly used in training AlphaFold models (table S2).

Fig. 1B illustrates the distinction of interaction-based metrics from traditional RMSD-based metrics, using bond torsion as an example. A bond rotation of 40º from a staggered to a partially eclipsed state yields a ~3 kcal/mol increase in energy (corresponding to a ~160-fold less probable conformer at 298 *K*), while an additional rotation of 70º leads to a different staggered state with an energy similar to the starting state (Fig. 1B). By contrast, RMSD increases monotonically with the rotation, resulting in an increased penalty for a lower-energy, higher-probability state (Fig. 1B).

We compared 3.4 million interactions predicted from AlphaFold2/3 to those obtained from X-ray crystallographic structures (n = 3949; table S3). These structures were part of the *Top*2018 dataset, with resolutions higher than 2 Å, and they were filtered by model quality at the single residue level (*30*). Identical sequences were used to retrieve AlphaFold2 predictions from its database (*31*) and to predict structures on the AlphaFold3 server (*4*, *32*) (*Methods*). The structures were confidently predicted, with an average predicted-local-distance-difference-test (pLDDT) score of 96 (SEM = 4) across all residues in AlphaFold2 predictions and 94 (SEM = 4) across all atoms in AlphaFold3 predictions (table S3).

AlphaFold extracted basic energetic properties of molecular interactions

AlphaFold models reproduced overall conformational preferences for covalent and non-covalent interactions. We found that the backbone torsion angles of predicted structures follow the classic Ramachandran distributions (Fig. 2A; see fig. S1 for AlphaFold2), consistent with the ability of AlphaFold models to reproduce protein folds (*1*). For side-chains, the predicted bonds showed preferences for staggered over eclipsed conformers, as expected, and the overall distributions agreed with those derived from the PDB (Fig. 2B; see fig. S2 for AlphaFold2). While AlphaFold2 generates structures by rotating covalent bonds of polypeptides, AlphaFold3 employs a diffusion-based architecture that directly generates atomic coordinates so that covalent bond lengths and angles are not predetermined. We found that AlphaFold3 captured basic covalent bond properties, such as the ~0.2 Å shorter C=O compared to C–O bonds (Fig. 2C, figs. S3 to S4 and tables S4 to S5).

AlphaFold models also captured features of non-covalent interactions. The van der Waals interactions from the AlphaFold predictions follow Lennard-Jones-like energetics, where a specific inter-atomic distance is preferred for each interaction type and shorter-than-ideal distances are increasingly disfavored (Fig. 2D; see fig. S5 for AlphaFold2; table S6). Hydrogen bonds were also identified, with specific distance and angle preferences for each donor•acceptor pair, although they diverge from the PDB preferences (Fig. 2E; see fig. S6 for AlphaFold2; table S7 to S8).

Thus, despite using atomic coordinates and multiple sequence alignments as the bases for their training, AlphaFold models have extracted conformational and energetic properties of molecular interactions. Nevertheless, notable discrepancies remain, which we describe and quantify below.

Discrepancies in the conformational distributions derived from AlphaFold models

The peak(s) of each conformational distribution represents the most favored state(s) and the breadth corresponds to the range of conformers and their relative probabilities. Differences in these peaks and distributions would indicate inaccurate conformational preferences derived by the models as well as errors in reproducing at least a subset of the experimentally determined states. To evaluate differences between the AlphaFold2/3 and PDB distributions, we compared the peaks and the dispersions of the distributions of each interaction type, and we found differences for nearly all covalent and non-covalent interactions (Fig. 2F to K).

Favored covalent bond lengths differ by ~0.01–0.03 Å for each bond type between the AlphaFold3 and PDB distributions, with the direction of change dependent on atom type (Fig. 2F and table S4). The AlphaFold3 distributions were consistently narrower than those derived from the PDB distributions, as quantified by their lower median-absolute-deviation values (Fig. 2G and table S4).

Larger differences relative to PDB-derived distributions were found for non-covalent interactions. For van der Waals distances and hydrogen bond lengths, the peaks differ by ~0.1 Å for those between side-chain interacting partners for both AlphaFold2 and AlphaFold3, except for C•C van der Waals interactions, where the peaks agree (Fig. 2H, fig. S7 and table S6 to S7). AlphaFold2/3 hydrogen bonds predominantly adopt more bent geometries compared to those favored in the PDB (Fig. 2J and fig. S7) and in small molecule structures and quantum mechanical calculations (*33–35*). In contrast to covalent bonds, non-covalent distributions in AlphaFold2/3 were broader than those from the PDB (Fig. 2I and K, fig. S7 and table S6 to S8). The increased breadth, in addition to the altered peak positions, was particularly pronounced for hydrogen bonds, indicating that these interactions are more variable in AlphaFold predictions than in the PDB (fig. S6 to S7).

The observed discrepancies are directional for each interaction type, and thus not expected to result from random noise. Discrepancies were also found for nearly all interaction types, suggesting system-level limitations in the AlphaFold models. To investigate the effect of these discrepancies, we next examined how they manifest in the prediction of the individual protein structures and of conformational ensembles.

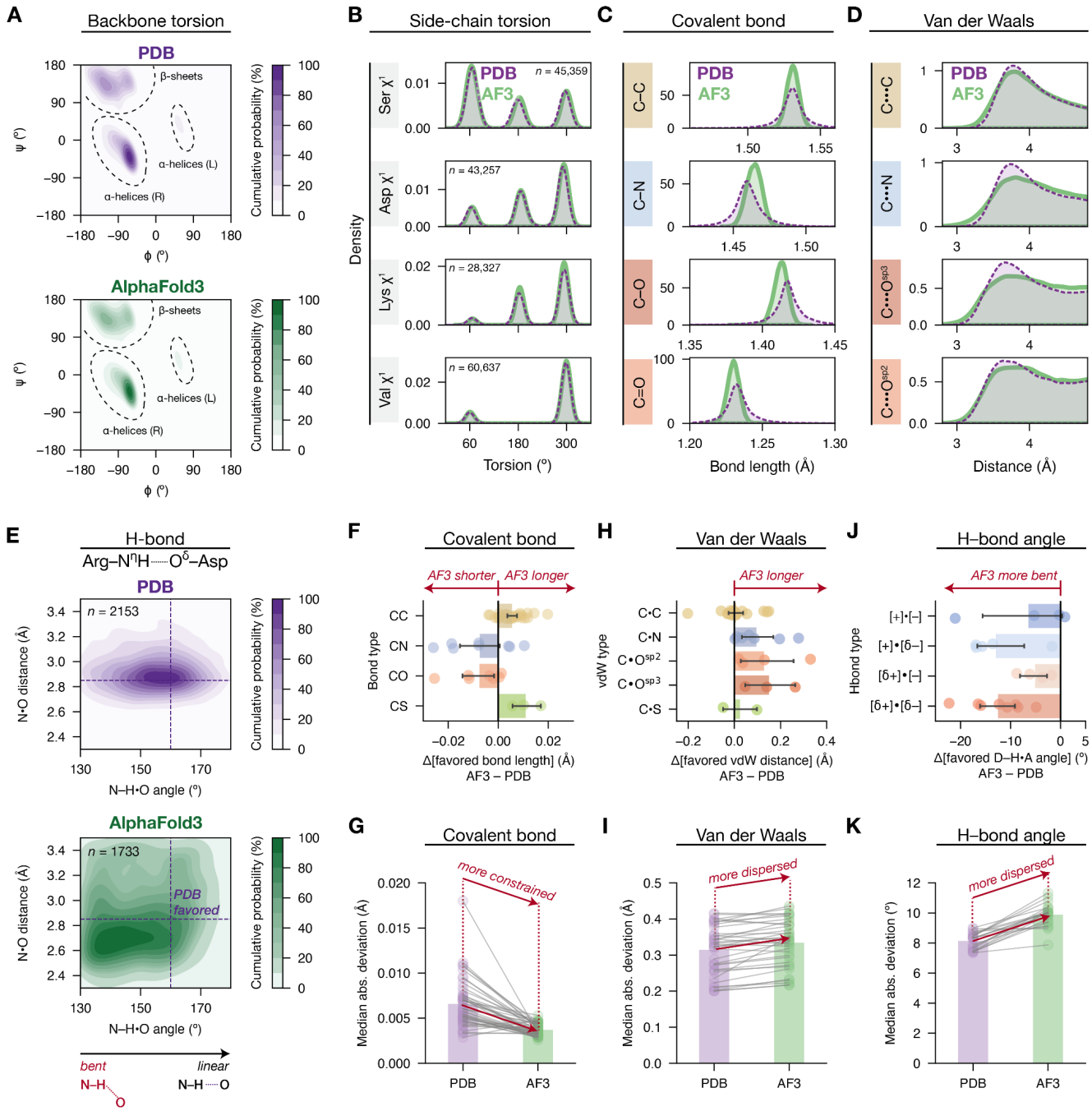

Fig. 2. Biomolecular interaction energetics inferred by AlphaFold3. The analyses for AlphaFold2 gave similar results and are shown in figs. S1 to S7. **(A)** Probability density distributions of backbone ψ and ϕ angles (not including glycine and proline residues; see fig. S1). The color gradients indicate the probability mass under the contour in percentages, with darker shades indicating more probable conformational states. **(B)** Examples of probability density distributions of side-chain torsion angles observed in the PDB (purple) and those from AlphaFold3 predictions (green). **(C)** Examples of probability density distributions of lengths of the specified covalent bonds. The distributions shown were taken from serine residues; distributions for the additional amino acids are shown in fig. S3. **(D)** Probability density distributions of van der Waals distances for different interatomic contacts. Each distribution includes >75,000 van der Waals interactions (table S3). **(E)** Example probability density distributions for hydrogen bond angles (donor–H•acceptor) and distances (donor•acceptor) for arginine•aspartate side-chain hydrogen bonds. The color gradients indicate the probability mass under the contour in percentage, with the darker shades indicating more probable conformational states. This distribution includes Arg•Asp residue pairs

that form a single hydrogen bond between their side-chains (excluding bidentate hydrogen bonds) to ensure an unambiguous comparison. Additional hydrogen bond distributions are shown in fig. S6. **(F)** Differences in the most favored covalent bond lengths in AlphaFold3 predictions *versus* the PDB distributions. Each circle represents the peak of the length distribution for a specific covalent bond type based on the amino acid and atom identities; the bond types and values are given in table S4. **(G)** Differences in the spread of the distributions for covalent bond lengths in AlphaFold3 predictions *versus* the PDB distributions, quantified by the median absolute deviation (MAD). Each circle represents the MAD for a specific covalent bond type; data are provided in table S4. **(H)** Differences in the most favored van der Waals distances in AlphaFold3 predictions *versus* in the PDB distributions, separated by the atoms involved. Each circle represents the peak of the length distribution for a specific van der Waals interaction based on the amino acid and atom identities; the bond types and values are given in table S6. **(I)** Differences in the spread of the distributions for van der Waals distances in AlphaFold3 predictions *versus* in the PDB distributions, quantified by the median absolute deviation (MAD). Each circle represents the MAD for a specific type of van der Waals interaction, and the data are provided in table S6. **(J)** Differences in the most favored hydrogen bond angles in AlphaFold3 predictions *versus* in the PDB distributions, for the following hydrogen bond types: charge•charge ([+]•[–]), charge•dipole ([δ+]•[–], [δ–]•[+]) and dipole•dipole ([δ+]•[δ–]). Each circle represents the peak of the angle distribution for a specific hydrogen bond type, and the data are provided in table S7. Only hydrogen bonds with sp^2 hybridized donor atoms were included so that the hydrogen positions and angles can be confidently calculated (*Methods*). **(K)** Differences in the spread of the distributions for hydrogen bond angles in AlphaFold3 predictions *versus* the PDB distributions, quantified by the median absolute deviation (MAD). Each circle represents the MAD for a specific type of hydrogen bond, and the data are provided in table S7.

Widespread prediction errors of side-chain interactions

Protein structure—and more fundamentally the probability of a conformational state in an ensemble—is determined by the sum of the energetics from all the interactions that are present (Fig. 1A). The discrepancies in the overall conformational preferences identified above could alter the balance of forces so that AlphaFold predicts conformational states with different sets of interactions compared to the corresponding PDB structures. Alternatively, the discrepancies may have less severe consequences, maintaining the same set of interactions but with some geometric distortions.

To distinguish between these possibilities, we first compared each bond torsion angle between AlphaFold2/3 predictions and PDB structures, as they underlie changes in non-covalent interactions and more extensive rearrangements. We then evaluated the energetics consequences of the identified torsion deviations as well as the accompanying changes in hydrogen bonds and van der Waals interactions. This systematic approach allowed us to parse individual interactions, providing a multi-dimensional physical evaluation.

Our comparisons focused on buried residues (relative solvent accessibility <25%, $n \sim 2.2$ million; table S3) as their conformations are determined by the sum of their interactions with surrounding protein groups, and thus can be used to evaluate the energetic interplay of these interactions. Additionally, these residues are the most confidently modeled from the X-ray diffraction data, minimizing errors from X-ray modeling and refinements (*36*, *37*). Below we focus on quantitative results for AlphaFold3 comparisons; AlphaFold2 gave similar results (see figs. S8 to S15), and we compare the two models in a subsequent section (see *Common prediction errors in AlphaFold2 and AlphaFold3*).

Assessing backbone and side-chain bond torsions

Among 1,277,419 backbone torsion angles, 99% of the ψ and ϕ angles match their PDB references within 30º (Fig. 3A; see fig. S8 for AlphaFold2). Correspondingly, 91% of the backbone amide groups are predicted in the correct secondary structure (table S8). The agreement is higher for residues forming alpha helices and beta sheets (97% and 96%) than for residues outside of these secondary structures (85%; table S9). High accuracy of backbone torsion angles is consistent with the excellent performance of AlphaFold models in predicting protein folds in prior evaluations based on backbone or Cα atom RMSDs (*1*).

Bond torsion accuracy is lower for side-chains, and decreases for bonds extending further from the backbone, with 94% to 77% of the predicted χ^1 to χ^4 torsion angles falling within 30º of the corresponding PDB structure (Fig. 3A and fig. S8). The decreasing accuracy may arise from the larger conformational space occupied by longer side-chains, from the increased number of interactions they can form, or from their lower abundance and thus less training opportunity.

Side-chain torsion deviations indicate altered energetic balances in AlphaFold

The favored bond torsions of each buried residue are determined by local molecular forces, and differences in predicted torsions indicate altered energetics of the surrounding environment. Rotameric side-chains can occupy three energetic minima or "wells" (*gauche*–, *trans*, and *gauche*+) (*38*), and rotation to an alternative well can result in steric clashes and accompanying rearrangements of the surrounding atoms. Rotation within rotameric wells, while giving smaller RMSDs than rotation to a different well, can have substantial energetic consequences—e.g., a 40˚ torsional rotation from staggered to partially eclipsed is destabilizing by ~3 kcal/mol, corresponding to ~160-fold difference in probability

at room temperature (Fig. 1B). Thus, reproducing eclipsed conformers provides a sensitive test for AlphaFold's ability to capture accurate energetics. Nevertheless, eclipsed bonds are rare and their effects would be masked by typical comparisons that combine all bonds together (n = 1532 buried eclipsed bonds, compared to n = 501,411 buried staggered bonds, as described below). To overcome this limitation, we analyzed the predicted bond torsions in two stages: (1) across rotameric wells, and (2) within each rotameric well. To quantify AlphaFold's ability to capture the local energetic balance, we compared their predictions to those made by baseline models that contain no information about surrounding forces.

To predict rotameric wells, the constraint-free baseline model randomly samples between the *gauche*–, *trans*, and *gauche*+ conformers, and thus predicts the correct rotameric state one-third of the time (Fig. 3B, baseline #1). AlphaFold3 gave a substantial improvement from this baseline, correctly predicting the rotameric state for 94%, 92%, 78%, and 73% of the χ^1 ,χ^2, χ^3 and χ^4 bonds (Fig. 3B; see fig. S8 for AlphaFold2).

For the second stage of testing, we selected bonds where AlphaFold2/3 predicted the correct rotameric well to test the models' ability to predict accurate states within each well. We compared AlphaFold2/3 to a second baseline model, this time randomly sampling torsion angles within each well, weighted by the probability of occurrence of each angle in the PDB reference dataset (Fig. 3C, upper, baseline #2). We used the Boltzmann relationship to convert probabilities into knowledge-based energies to provide a simple evaluation metric in kcal/mol [*Methods*; (*39*, *40*)]. Because the baseline model most frequently samples the low-energy, staggered states, it is expected to give at most small deviations for staggered bonds in the PDB ($\Delta E_{BL} \approx 0$; defined in Fig. 3C, middle). By contrast, high energy, eclipsed states are rarely sampled by the baseline model and thus unlikely to be predicted correctly by chance—e.g., the model is expected to give a deviation of ΔE_{BL} = 3 kcal/mol for a partially eclipsed state that is 3 kcal/mol higher in energy than the minimum (Fig. 3C, middle).

For staggered states that are within 1 kcal/mol of the minimum, the baseline model and AlphaFold2/3 both gave ΔE values centered around zero (Fig. 3C, lower; see Fig. 5B for AlphaFold2). The AlphaFold models showed a tighter distribution, indicating improved accuracy over the random baseline model. To test eclipsed states, we selected bonds that are 2.5 to 4.0 kcal/mol higher in energy than the minimum. The baseline model gave the expected deviation of ΔE_{BL} = 3.1 kcal/mol (SD = 0.7; Fig. 3C, lower). AlphaFold2/3 also deviated substantially from the PDB, with average prediction deviations of ΔE_{AF3} = 2.4 kcal/mol (SD = 1.1 kcal/mol) and E_{AF2} = 2.1 kcal/mol (SD = 1.2 kcal/mol). These deviations are substantial but less than the baseline model deviation (Fig. 3C; see Fig. 5B for AlphaFold2). The AlphaFold2/3 predictions captured 10% and 6% of the eclipsed bonds within 0.5 kcal/mol of the PDB values, respectively, suggesting that certain structural contexts may be reproduced more accurately.

Overall, AlphaFold2/3 outperformed the baseline model in predicting the correct rotameric wells, where there are relatively large movements and steric factors involved. Nevertheless, the balance of forces responsible for favoring states within a particular rotameric well was not captured in most cases. This limitation presumably arises because bond rotation within a well is highly sensitive to the accuracy of the local energetic balance, despite small changes in atomic positions and thus RMSD metrics.

AlphaFold models *mis-assigned about one-third of side-chain non-covalent interactions*

Rotations of the side-chain bonds are expected to perturb the non-covalent interactions they are involved in. To test the consequence of torsional deviations in non-covalent interactions, we performed

one-to-one comparisons for hydrogen bonds and van der Waals interactions. We determined whether the same interactions were formed in AlphaFold2/3 and PDB structures and quantified the extent of geometric deviations of the interactions that were in common.

Our comparisons included ~440,000 hydrogen bonds where the donor heavy atom is sp^2 hybridized, as we could unambiguously assign the hydrogen position for this subset of hydrogen bonds (table S3). Nearly all (96%) backbone•backbone hydrogen bonds were predicted between the same atom pairs of the same residues as seen in the PDB. The accuracy decreased for hydrogen bonds involving side-chains—with 21% and 32% of the backbone•side-chain and side-chain•side-chain hydrogen bonds predicted to form between incorrect partners by the AlphaFold3 models (Fig. 3D, left; see fig. S9 for AlphaFold2). AlphaFold3 also hallucinated hydrogen bonds such that 4%, 12% and 23% of the backbone•backbone, backbone•side-chain and side-chain•side-chain hydrogen bonds, respectively, were not present in the corresponding PDB structures (Fig. 3D, right; see fig. S9 for AlphaFold2).

Analogous comparisons for side-chain van der Waals interactions were performed by identifying carbon•heteroatom pairs within 0.4 Å of their ideal van der Waals distances (*Methods*). These comparisons revealed similar trends as observed for hydrogen bonds. Across different atom types, from 15% to 31% of the van der Waals interactions in the PDB are missing in AlphaFold3 predictions, and 14% to 22% of the model-predicted van der Waals contacts were hallucinated (Fig. 3E; see fig. S10 for AlphaFold2).

Even when AlphaFold3 predicted the correct interaction pairs, substantial geometric deviations were found. About 39% of the side-chain distances for correctly-predicted hydrogen bond donor•acceptors deviated by more than 0.2 Å from the corresponding PDB distances, and 32% were at least 20º more bent (fig. S11). Similarly, among the ~130,000 correct van der Waals pairs, about 32% of the AlphaFold3 predictions deviated >0.2 Å from the PDB (fig. S12).

Incorrect predictions are distributed throughout the structures and across the dataset

To determine whether there are idiosyncratic features associated with the errors that could help pinpoint specific model limitations, we mapped the incorrect predictions on protein structures. Incorrect interactions appeared to be distributed throughout each structure rather than clustered (Fig. 3F). Quantitative analysis showed that the extent of clustering of incorrectly predicted residues within each protein was indistinguishable from randomly selected residues (fig. S13). The randomly distributed rather than clustered errors suggest multiple individual cases of inaccurate energetic balances, rather than a single or small number of epicenters that propagate errors.

Across the dataset, incorrect interactions were found in nearly all structures across different structural classes, and the number of incorrect interactions in each structure scaled with protein size, suggesting that the prediction errors were not associated with specific proteins (fig. S14 to S15). In addition, the prediction errors could not be identified from AlphaFold confidence scores, as residues involved in incorrect interactions gave average pLDDTs of 95 (AlphaFold2, SD = 7) and 92 (AlphaFold3, SD = 8) (fig. S16). Altogether, these results indicate system-wide, non-idiosyncratic side-chain prediction errors that are not identified as suspect by the model.

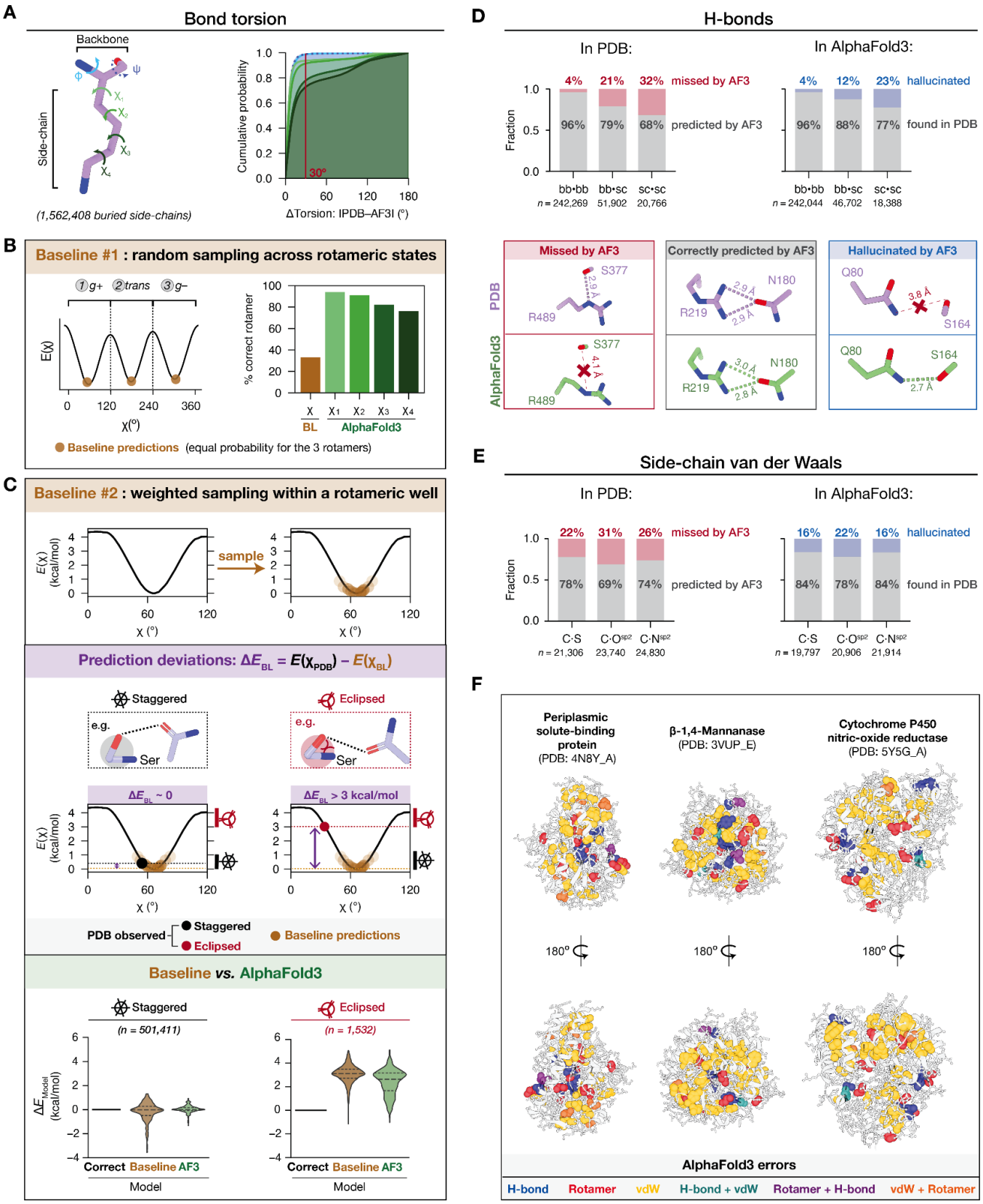

Fig. 3. One-to-one comparisons for interactions in AlphaFold3 predictions *versus* in the corresponding PDB structures. See fig. S8 to S15 for AlphaFold2 results. **(A)** Left: definition of backbone and side-chain torsion angles, using a lysine residue as an example. Right: Cumulative distribution for side-chain torsion angle differences between AlphaFold3 predictions and the corresponding PDBs. **(B)** Prediction accuracy for rotameric states of

AlphaFold3 compared to a context-free baseline model that randomly samples torsion angles according to the PDB distribution. Left: the baseline model samples three rotameric states that correspond to the local energy minima (*gauche-, trans, gauche+*) with near-equal probabilities; thus, the model predicts the correct rotamer ~1/3 of the time. The three rotameric states are mapped on an energy function $E(\chi)$, obtained by inverting the PDB torsion angle distribution $f(\chi)$ using the Boltzmann relationship (*Methods*). Right: percentage of correct side-chain rotameric states predicted by AlphaFold3 compared to the baseline. **(C)** Comparison of AlphaFold3 to a second context-free baseline model in predicting staggered and eclipsed bonds. Top: baseline model that samples within a rotameric state. The low-energy staggered conformations are more likely sampled than the eclipsed conformations. Middle: expected baseline model accuracy in predicting staggered and eclipsed conformations reported by the energetic deviation of the PDB *versus* AlphaFold-predicted torsional state [$\Delta E_{BL} = E(\chi_{PDB}) - E(\chi_{model})$], obtained by mapping the torsion angles onto the PDB torsion energy function $E(\chi)$. For a staggered PDB bond (e.g., the serine example shaded in grey), the baseline model predicts the correct energy ($\Delta E_{BL} \sim 0$) as it most likely samples a staggered conformer. For an eclipsed conformer constrained by nearby molecular interactions in the PDB (e.g., the serine example shaded in red), the baseline model is expected to give a positive deviation with the deviation scaling proportionally with the energy of the torsion angle [e.g., $\Delta E_{BL} \sim= 2.5$ kcal/mol when $E(\chi_{PDB}) = 2.5$ kcal/mol]. Bottom: AlphaFold3 (light green) compared to baseline (brown) for energy deviations from staggered (bottom left) and eclipsed (bottom right) PDB references. For reference, a theoretical model that predicts all bonds correctly would have $\Delta E_{correct} = 0$, shown as the black lines. **(D)** Comparisons of buried backbone•backbone (bb•bb), backbone•side-chain (bb•sc), and side-chain•side-chain (sc•sc) hydrogen bonds in the PDB structures and in AlphaFold3 predictions. The gray bars indicate the numbers of hydrogen bonds in common in the PDB structure and the corresponding AlphaFold3 prediction ("correctly predicted by AF3"). For the PDB hydrogen bonds, the fraction that were not found in the AlphaFold predictions are shown as red bars ("missed by AF3"); for the AlphaFold3 hydrogen bonds, the fraction that were not found in the PDB structures are shown as blue bars ("hallucinated"). Representative structures for each group are shown below [PDB: 2W3P_B, 1OFL_A and 1KGN_C, from left to right (purple) and their corresponding AlphaFold3 predicted structures (green)]. AlphaFold2 results are shown in fig. S9. **(E)** The same comparisons as in panel D for three types of side-chain•side-chain van der Waals interactions. Additional classes and examples and AlphaFold2 results are shown in fig. S10. **(F)** Examples of prediction errors in buried side-chain interactions mapped on protein structures for the misassignment of rotameric states, hydrogen bond and van der Waals partners, and combinations of these errors, shown in different colors. Buried residues with correctly assigned interactions are shown as white sticks. Additional examples are shown in fig. S17.

AlphaFold3 failed to reproduce the heterogeneity in experimental ensembles

Conformational ensembles are the most direct emergent properties of energy landscapes and provide an independent test for the extent to which the models reproduce protein energetics (*41–44*). Moreover, ensemble generation is a rapidly growing frontier of predictive models , as ensembles provide the states and probabilities that will lead to quantitative prediction of function[(Fig. 1A; (*45*, *46*)].

AlphaFold3's diffusion architecture is well suited for ensemble generation because each reverse-diffusion trajectory samples from an internal probability distribution (*4*, *47*). To evaluate AlphaFold3's ability to generate accurate ensembles, we curated reference experimental datasets collected by non-cryogenic multi-temperature (MT) X-ray crystallography, which reveals a wider range of conformational states than X-ray structures determined at cryogenic temperatures (*48–53*). We collected high-resolution MT datasets (1.0 – 2.2 Å) for nine proteins that were determined above 200 *K* and had been refined into multi-conformer models (Fig. 4A and table S10). We repeatedly sampled from AlphaFold3 for each of these sequences by initializing the predictions with random seeds, and we compared the resulting bond torsion distributions to the corresponding MT datasets.

We found 270 side-chain bonds that occupy two or more conformational states across the total 1361 buried side-chain bonds in the MT multi-conformer models. The majority (85%) of the alternative conformers observed in MT models were not reproduced by AlphaFold3, suggesting an overly restricted probability landscape in the model (Fig. 4B). AlphaFold3 reproduced 15% of the multi-conformer bonds (n = 40), with 93% of these alternative states agreeing with states in the MT model within 30º (Fig. 4C), but we were unable to identify features that might be responsible for these differences.

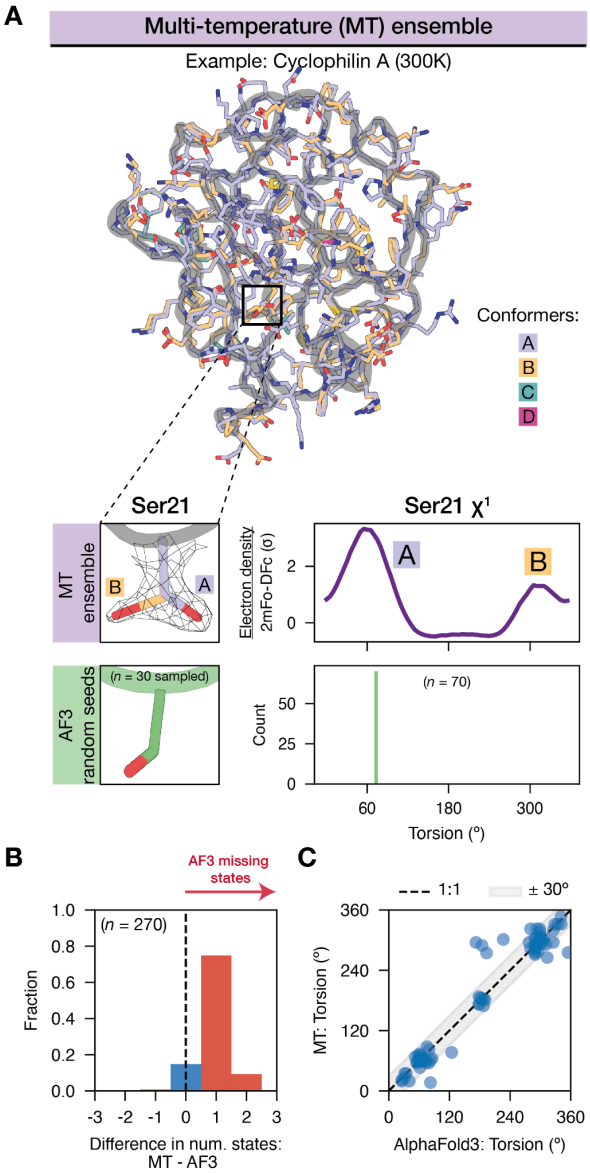

Fig. 4. Comparisons of multi-temperature (MT) crystallographic models and AlphaFold3 ensembles generated by random seeding. **(A)** Top: multi-conformer model of Cyclophilin A refined from an X-ray dataset collected at 300 *K* (PDB: 4YUM; resolution = 1.5 Å). Bottom: an example of the comparison performed for the multi-conformer side-chain of Ser21 from Cyclophilin A. The electron density map around Ser21 (2mFo-DFc; purple line) shows two conformers corresponding to modeled A and B conformers. The torsion distribution for this residue from the AlphaFold3 ensemble is shown below in green. These comparisons were made for all buried residues (with <0.25 relative solvent accessible area) across the nine different proteins in our test set (n = 270). **(B)** Difference in the number of states observed in MT models *versus* AlphaFold3 ensembles for buried residues that occupy two or more states in the MT models (n = 270). The blue bar indicates the fraction of bonds predicted by AlphaFold3 to have the same number of alternative states as in the MT models; the red bars indicate the bonds where AlphaFold3 ensembles missed alternative states. **(C)** One-to-one comparisons for the torsional states predicted by AlphaFold3 and found in the MT model. Only residues where AlphaFold3 predicted the correct number of alternative states were included (i.e., the blue bar in panel B, n = 40 bonds and 81 conformers). The dashed line indicates 1:1 or perfect agreement, and the grey bar indicates a range of ±30º.

Common prediction errors in AlphaFold2 and AlphaFold3

The different architectures of AlphaFold2 and AlphaFold3 could lead to distinct strengths and weaknesses in their respective models, information that would be useful in improving future models. Of the 37,681 and 49,958 mis-assigned side-chain rotamers in AlphaFold2 and AlphaFold3, respectively, most (26,591) were in common. Each model also showed additional unique errors, with AlphaFold3 giving about twice as many unique errors. For nearly all the common errors (96%), the models predicted the same incorrect rotameric states (Fig. 5A, right). In addition, AlphaFold2 and AlphaFold3 also made near-identical (and mainly incorrect) predictions for the eclipsed, high-energy bonds found in the PDB (Fig. 5B).

Mis-assigned noncovalent interactions also tended to be common in AlphaFold2 and AlphaFold3 predictions. PDB hydrogen bonds and van der Waals interactions that were missing in AlphaFold3 predictions were also mainly missing in AlphaFold2 predictions (60%), with AlphaFold3 missing more interactions (Fig. 5C to D). Similarly, a large fraction of hallucinated hydrogen bonds and van der Waals interactions were common for AlphaFold 2 and AlphaFold3 (46% and 62%, respectively; Fig. 5C to D).

Overall, AlphaFold3 produced more interaction errors than AlphaFold2 across each interaction type (Fig. 5E), despite its redesigned architecture and higher reported accuracy (*4*). The larger number of AlphaFold3 errors and the preponderance of common errors with AlphaFold2 could partially arise from the data augmentation procedure that uses AlphaFold2-Multimer predictions to supplement the training set for AlphaFold3 (*4*). Most fundamentally, our physics-based evaluations identified limitations that were not revealed in prior evaluations and can be taken into account as next-generation models are developed (see *Conclusions & Implications*).

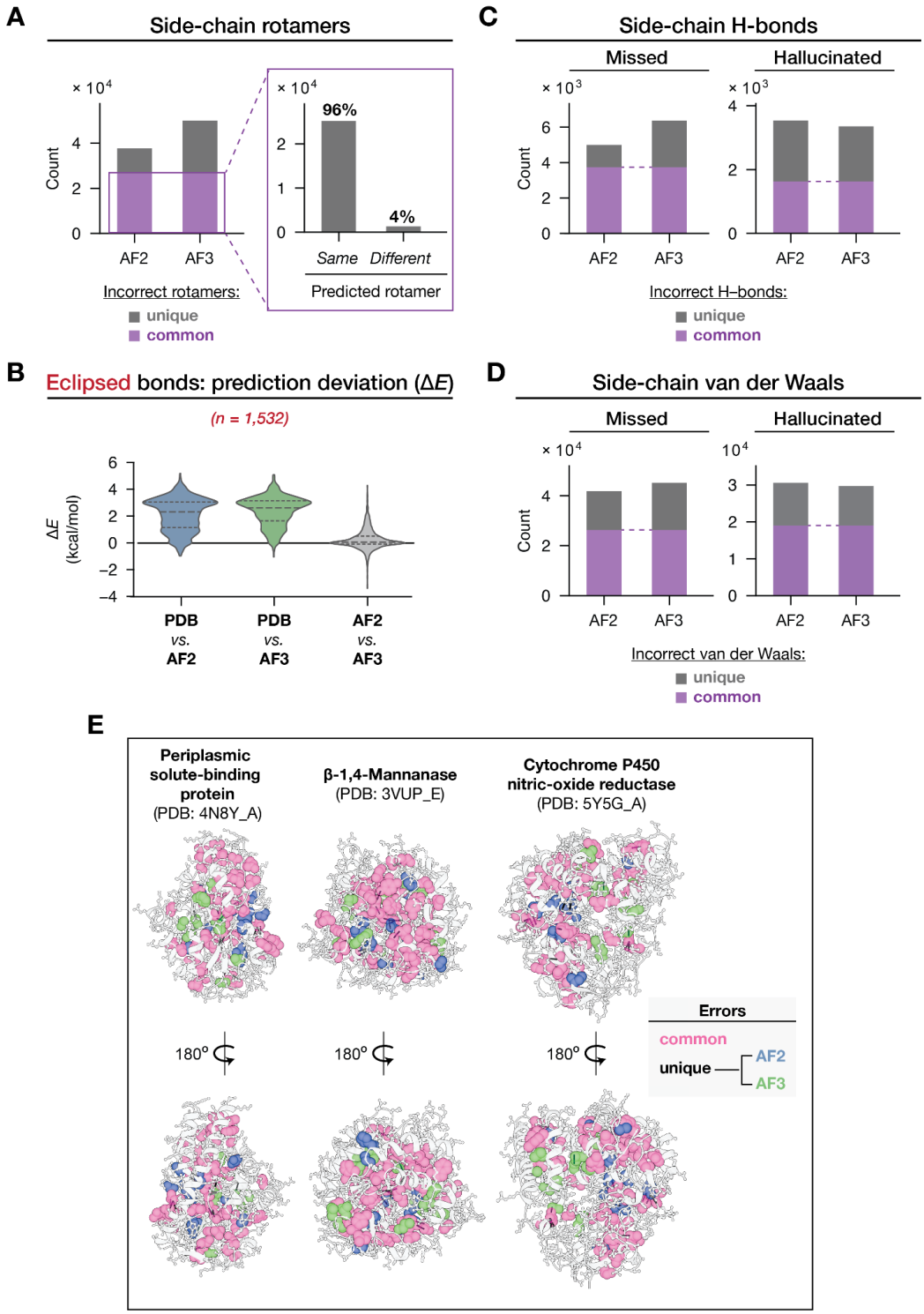

Fig. 5. Comparison of AlphaFold2 and AlphaFold3 prediction errors. **(A)** Rotameric states errors in AlphaFold2 and AlphaFold3. Left: the number of incorrect rotamers that were common (purple) and unique (grey) errors. Right: within the shared-error subset, number of common *versus* different predicted (incorrect) rotameric states. **(B)** Energetic deviation (ΔE) for eclipsed side-chain bonds torsions for PDB conformers to the AlphaFold2 predictions; for PDB conformers *versus* AlphaFold3 predictions; and for AlphaFold2 *versus* AlphaFold3 predictions. **(C)** Side-chain hydrogen bond errors in AlphaFold2 and AlphaFold3 that were common (purple) and unique (grey). Left: hydrogen bonds identified in the PDB structures but not in the corresponding model predictions. Right: hydrogen bonds hallucinated in each model prediction. **(D)** Side-chain van der Waals errors in AlphaFold2 and AlphaFold3

that were common (purple) and unique (grey). Left: van der Waals interactions identified in the PDB structures but not in the corresponding model predictions. Right: van der Waals interactions hallucinated in model predictions. **(E)** Spatial distribution of common (pink) *versus* model-specific side-chain errors (blue: AlphaFold2; green: AlphaFold3), mapped onto three representative protein structures; 180° rotated views are shown for clarity. Additional examples are shown in fig. S18.

Conclusions and Implications

Effective evaluations drive model development by identifying the capabilities and limitations of a model—what the model can reliably do, what it cannot do, and what improvements are needed to expand its capabilities. Current evaluations of deep learning biomolecular models have centered on RMSD-based metrics that are only indirectly related to the physical properties that govern the structure and function of biomolecules; thus, they are limited in their ability to determine what the models can reliably predict and in identifying how to improve the models. We developed a systematic and generalizable evaluation framework that overcomes these limitations, providing a multifaceted assessment of AlphaFold in the natural language of structure and function—molecular interactions and their energetics.

Our evaluation revealed previously unknown capabilities and limitations of the AlphaFold models. These deep learning models extracted basic energetic properties for molecular interactions, including covalent bonds, bond torsions, and van der Waals interactions. Nevertheless, the model-derived energetic rules were flawed, leading to pervasive errors in side-chain interactions of the predicted structures. These incorrect interactions and interaction networks—amounting to nearly one-third of all side-chain hydrogen bonds and van der Waals—may be responsible for the limited applicability of AlphaFold predictions for downstream tasks such as ligand docking [(*29*, *54*, *55*); see also table S2].

AlphaFold3 produced more errors in molecular interactions than AlphaFold2, despite its updated model architecture and better performance in RMSD-based metrics (*4*). This paradoxical behavior may arise from directing model development along a single dimension (i.e., an RMSD/distance-based training objective), leading to overfitting that degrades model performance. The failure to identify these limitations during model development from AlphaFold2 to AlphaFold3 underscores the need for multi-dimensional evaluation metrics beyond RMSDs—in particular, metrics that are tightly coupled to the model's predictive power but different from its training metrics, as for the physics-based metrics provided here.

Incorporating our evaluation into the cycle of model development can identify the effects from each architectural or training modification and reveal which updates truly help the model capture biomolecular physics. To overcome the conformational biases and the accompanying prediction errors we uncovered, future models could select for the highest quality data in their training, incorporate physics-aware training objectives, and/or apply energetic constraints or corrections to the current models. In particular, high-quality, multi-dimensional ensemble data from multi-temperature X-ray, other perturbative X-ray approaches (*4*, *56*), and high-resolution cryo-electron microscopy (*57*) provide richer information about energy landscapes than single PDB structures, and may therefore help models learn energetics. Most generally, a step-wise development approach, combined with rigorous evaluation at each step, will allow next-generation models to achieve the greatest predictive power.

The multifaceted, physics-based evaluation framework developed here can be applied to any model that operates on atomistic structural information and can be readily expanded to include interactions involving non-amino acid groups (nucleic acids, solvent, and ligands). This generalizable evaluation approach can aid in the evolution of deep learning models from memorization and interpolation of structural data to the extraction of fundamental physics universal to all biomolecular behaviors. Ultimately, learning these universal physical rules will allow models to extrapolate beyond their training to predict functions of more complex systems where biological insights are needed yet the scale of the training data required for interpolation may be unattainable.

Acknowledgments: We thank S. Costello, J. Yoon, P. Almhjell and S. Wankowicz for helpful discussions and feedback on the manuscript. We thank the following individuals for their help generating AlphaFold3 predictions using the online AlphaFold server before the code was released: Yaming Yan, Ben Ma, Chen Zhou, Garrett Wang, Shawn Costello, Patrick Almhjell, Lauren Hagler, Abigail Chiang, and Allison Rabbani.

Funding:

National Key Research and Development Program of China (No. 2024YFA1307502) (JM)

The Science and Technology Innovation Plan of Shanghai Science and Technology Commission (No. 23JS1400200) (JM)

The Research Fund for International Senior Scientists (No. W2431060) (JM)

National Science Foundation grant MCB2322069 (DH)

Stanford Center for Molecular Analysis and Design Fellowship (SD)

Author contributions:

Conceptualization: SD, DH, NL, JM

Investigation: NL, SD

Visualization: NL, SD

Funding acquisition: DH, JM

Writing – original draft: SD, NL, DH

Writing – review & editing: SD, DH, NL, JM

Competing interests: Authors declare that they have no competing interests.

Data and materials availability: All data, code, and materials used in the analysis are available at Zenodo: https://doi.org/10.5281/zenodo.15770723

References

1. J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

2. Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, A. Dos Santos Costa, M. Fazel-Zarandi, T. Sercu, S. Candido, A. Rives, Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).

3. R. Krishna, J. Wang, W. Ahern, P. Sturmfels, P. Venkatesh, I. Kalvet, G. R. Lee, F. S. Morey-Burrows, I. Anishchenko, I. R. Humphreys, R. McHugh, D. Vafeados, X. Li, G. A. Sutherland, A. Hitchcock, C. Neil Hunter, A. Kang, E. Brackenbrough, A. K. Bera, M. Baek, F. DiMaio, D. Baker, Generalized biomolecular modeling and design with RoseTTAFold All-Atom. *Science*, doi: 10.1126/science.adl2528 (2024).

4. J. Abramson, J. Adler, J. Dunger, R. Evans, T. Green, A. Pritzel, O. Ronneberger, L. Willmore, A. J. Ballard, J. Bambrick, S. W. Bodenstein, D. A. Evans, C.-C. Hung, M. O'Neill, D. Reiman, K. Tunyasuvunakool, Z. Wu, A. Žemgulytė, E. Arvaniti, C. Beattie, O. Bertolli, A. Bridgland, A. Cherepanov, M. Congreve, A. I. Cowen-Rivers, A. Cowie, M. Figurnov, F. B. Fuchs, H. Gladman, R. Jain, Y. A. Khan, C. M. R. Low, K. Perlin, A. Potapenko, P. Savy, S. Singh, A. Stecula, A. Thillaisundaram, C. Tong, S. Yakneen, E. D. Zhong, M. Zielinski, A. Žídek, V. Bapst, P. Kohli, M. Jaderberg, D. Hassabis, J. M. Jumper, Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).

5. J. L. Watson, D. Juergens, N. R. Bennett, B. L. Trippe, J. Yim, H. E. Eisenach, W. Ahern, A. J. Borst, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, N. Hanikel, S. J. Pellock, A. Courbet, W. Sheffler, J. Wang, P. Venkatesh, I. Sappington, S. V. Torres, A. Lauko, V. De Bortoli, E. Mathieu, S. Ovchinnikov, R. Barzilay, T. S. Jaakkola, F. DiMaio, M. Baek, D. Baker, De novo design of protein structure and function with RFdiffusion. *Nature* **620**, 1089–1100 (2023).

6. K. Tunyasuvunakool, J. Adler, Z. Wu, T. Green, M. Zielinski, A. Žídek, A. Bridgland, A. Cowie, C. Meyer, A. Laydon, S. Velankar, G. J. Kleywegt, A. Bateman, R. Evans, A. Pritzel, M. Figurnov, O. Ronneberger, R. Bates, S. A. A. Kohl, A. Potapenko, A. J. Ballard, B. Romera-Paredes, S. Nikolov, R. Jain, E. Clancy, D. Reiman, S. Petersen, A. W. Senior, K. Kavukcuoglu, E. Birney, P. Kohli, J. Jumper, D. Hassabis, Highly accurate protein structure prediction for the human proteome. *Nature* **596**, 590–596 (2021).

7. Y. Song, Q. Yuan, S. Chen, Y. Zeng, H. Zhao, Y. Yang, Accurately predicting enzyme functions through geometric graph learning on ESMFold-predicted structures. *Nature Communications* **15**, 1–11 (2024).

8. M. Akdel, D. E. V. Pires, E. P. Pardo, J. Jänes, A. O. Zalevsky, B. Mészáros, P. Bryant, L. L. Good, R. A. Laskowski, G. Pozzati, A. Shenoy, W. Zhu, P. Kundrotas, V. R. Serra, C. H. M. Rodrigues, A. S. Dunham, D. Burke, N. Borkakoti, S. Velankar, A. Frost, J. Basquin, K. Lindorff-

Larsen, A. Bateman, A. V. Kajava, A. Valencia, S. Ovchinnikov, J. Durairaj, D. B. Ascher, J. M. Thornton, N. E. Davey, A. Stein, A. Elofsson, T. I. Croll, P. Beltrao, A structural biology community assessment of AlphaFold2 applications. *Nat Struct Mol Biol* **29**, 1056–1067 (2022).

9. J. M. McBride, T. Tlusty, AI-Predicted Protein Deformation Encodes Energy Landscape Perturbation. *Phys Rev Lett* **133**, 098401 (2024).

10. L. Chang, A. Perez, Ranking Peptide Binders by Affinity with AlphaFold. *Angew Chem Int Ed Engl* **62**, e202213362 (2023).

11. A. Motmaen, J. Dauparas, M. Baek, M. H. Abedi, D. Baker, P. Bradley, Peptide-binding specificity prediction using fine-tuned protein structure prediction networks. *Proc Natl Acad Sci U S A* **120**, e2216697120 (2023).

12. G. Monteiro da Silva, J. Y. Cui, D. C. Dalgarno, G. P. Lisi, B. M. Rubenstein, High-throughput prediction of protein conformational distributions with subsampled AlphaFold2. *Nat Commun* **15**, 2464 (2024).

13. Modeling conformational states of proteins with AlphaFold. *Current Opinion in Structural Biology* **81**, 102645 (2023).

14. H. K. Wayment-Steele, A. Ojoawo, R. Otten, J. M. Apitz, W. Pitsawong, M. Hömberger, S. Ovchinnikov, L. Colwell, D. Kern, Predicting multiple conformations via sequence clustering and AlphaFold2. *Nature* **625**, 832–839 (2024).

15. V. Agarwal, A. C. McShan, The power and pitfalls of AlphaFold2 for structure prediction beyond rigid globular proteins. *Nat Chem Biol* **20**, 950–959 (2024).

16. D. Chakravarty, J. W. Schafer, E. A. Chen, J. F. Thole, L. A. Ronish, M. Lee, L. L. Porter, AlphaFold predictions of fold-switched conformations are driven by structure memorization. *Nat Commun* **15**, 7296 (2024).

17. G. R. Buel, K. J. Walters, Can AlphaFold2 predict the impact of missense mutations on structure? *Nat Struct Mol Biol* **29**, 1–2 (2022).

18. A. R. Leach, *Molecular Modelling: Principles and Applications* (Pearson Education, 2001).

19. Effective energy functions for protein structure prediction. *Current Opinion in Structural Biology* **10**, 139–145 (2000).

20. K. Preuer, P. Renz, T. Unterthiner, S. Hochreiter, G. Klambauer, Fréchet ChemNet Distance: A Metric for Generative Models for Molecules in Drug Discovery. *J Chem Inf Model* **58**, 1736–1741 (2018).

21. R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams, A. Aspuru-Guzik, Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent Sci* **4**, 268–276 (2018).

22. N. Brown, M. Fiscato, M. H. S. Segler, A. C. Vaucher, GuacaMol: Benchmarking Models for de Novo Molecular Design. *J Chem Inf Model* **59**, 1096–1108 (2019).

23. M. Buttenschoen, G. M. Morris, C. M. Deane, PoseBusters: AI-based docking methods fail to generate physically valid poses or generalise to novel sequences. *Chem Sci* **15**, 3130–3139 (2024).

24. M. Xu, A. S. Powers, R. O. Dror, S. Ermon, J. Leskovec, "Geometric Latent Diffusion Models for 3D Molecule Generation" in *International Conference on Machine Learning* (PMLR, 2023), pp. 38592–38610.

25. L. Wu, C. Gong, X. Liu, M. Ye, Q. Liu, Diffusion-based Molecule Generation with Informative Prior Bridges. *Advances in Neural Information Processing Systems* **35**, 36533–36545 (2022).

26. A. Morehead, J. Cheng, Geometry-complete diffusion for 3D molecule generation and optimization. *Commun Chem* **7**, 150 (2024).

27. A. Schneuing, C. Harris, Y. Du, K. Didi, A. Jamasb, I. Igashov, W. Du, C. Gomes, T. L. Blundell, P. Lio, M. Welling, M. Bronstein, B. Correia, Structure-based drug design with equivariant diffusion models. *Nat Comput Sci* **4**, 899–909 (2024).

28. M. Baek, F. DiMaio, I. Anishchenko, J. Dauparas, S. Ovchinnikov, G. R. Lee, J. Wang, Q. Cong, L. N. Kinch, R. Dustin Schaeffer, C. Millán, H. Park, C. Adams, C. R. Glassman, A. DeGiovanni, J. H. Pereira, A. V. Rodrigues, A. A. van Dijk, A. C. Ebrecht, D. J. Opperman, T. Sagmeister, C. Buhlheller, T. Pavkov-Keller, M. K. Rathinaswamy, U. Dalwadi, C. K. Yip, J. E. Burke, K. Christopher Garcia, N. V. Grishin, P. D. Adams, R. J. Read, D. Baker, Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, doi: 10.1126/science.abj8754 (2021).

29. M. Karelina, J. J. Noh, R. O. Dror, How accurately can one predict drug binding modes using AlphaFold models? *Elife* **12** (2023).

30. C. J. Williams, D. C. Richardson, J. S. Richardson, The importance of residue-level filtering and the Top2018 best-parts dataset of high-quality protein residues. *Protein Sci* **31**, 290–300 (2022).

31. M. Varadi, D. Bertoni, P. Magana, U. Paramval, I. Pidruchna, M. Radhakrishnan, M. Tsenkov, S. Nair, M. Mirdita, J. Yeo, O. Kovalevskiy, K. Tunyasuvunakool, A. Laydon, A. Žídek, H. Tomlinson, D. Hariharan, J. Abrahamson, T. Green, J. Jumper, E. Birney, M. Steinegger, D. Hassabis, S. Velankar, AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic Acids Res* **52**, D368–D375 (2023).

32. J. Abramson, J. Adler, J. Dunger, R. Evans, T. Green, A. Pritzel, O. Ronneberger, L. Willmore, A. J. Ballard, J. Bambrick, S. W. Bodenstein, D. A. Evans, C.-C. Hung, M. O'Neill, D. Reiman, K. Tunyasuvunakool, Z. Wu, A. Žemgulytė, E. Arvaniti, C. Beattie, O. Bertolli, A. Bridgland, A. Cherepanov, M. Congreve, A. I. Cowen-Rivers, A. Cowie, M. Figurnov, F. B. Fuchs, H. Gladman, R. Jain, Y. A. Khan, C. M. R. Low, K. Perlin, A. Potapenko, P. Savy, S. Singh, A. Stecula, A. Thillaisundaram, C. Tong, S. Yakneen, E. D. Zhong, M. Zielinski, A. Žídek, V. Bapst, P. Kohli, M. Jaderberg, D. Hassabis, J. M. Jumper, Addendum: Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **636**, E4 (2024).

33. A. V. Morozov, T. Kortemme, K. Tsemekhman, D. Baker, Close agreement between the orientation dependence of hydrogen bonds observed in protein structures and quantum mechanical calculations. *Proc Natl Acad Sci U S A* **101**, 6946–6951 (2004).

34. P. A. Wood, F. H. Allen, E. Pidcock, Hydrogen-bond directionality at the donor H atom—analysis of interaction energies and database statistics. *CrystEngComm* **11**, 1563–1571 (2009).

35. D. Santos-Martins, S. Forli, Charting Hydrogen Bond Anisotropy. *Journal of Chemical Theory and Computation*, doi: 10.1021/acs.jctc.9b01248 (2020).

36. S. Parthasarathy, M. R. Murthy, Analysis of temperature factor distribution in high-resolution protein structures. *Protein Sci* **6**, 2561–2567 (1997).

37. P. Radivojac, Z. Obradovic, D. K. Smith, G. Zhu, S. Vucetic, C. J. Brown, J. David Lawson, A. Keith Dunker, Protein flexibility and intrinsic disorder. *Protein Science* **13**, 71–80 (2004).

38. Backbone-dependent Rotamer Library for Proteins Application to Side-chain Prediction. *Journal of Molecular Biology* **230**, 543–574 (1993).

39. T. Kortemme, A. V. Morozov, D. Baker, An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes. *J Mol Biol* **326**, 1239–1259 (2003).

40. M.-Y. Shen, A. Sali, Statistical potential for assessment and prediction of protein structures. *Protein Science* **15**, 2507–2524 (2006).

41. S. Li, M. Li, Y. Wang, X. He, N. Zheng, J. Zhang, P.-A. Heng, Improving AlphaFlow for Efficient Protein Ensembles Generation (2024). http://arxiv.org/abs/2407.12053.

42. B. Jing, B. Berger, T. Jaakkola, AlphaFold Meets Flow Matching for Generating Protein Ensembles (2024). http://arxiv.org/abs/2402.04845.

43. S. Lewis, T. Hempel, J. Jiménez Luna, M. Gastegger, Y. Xie, A. Y. K. Foong, V. García Satorras, O. Abdin, B. S. Veeling, I. Zaporozhets, Y. Chen, S. Yang, A. Schneuing, J. Nigam, F. Barbero, V. Stimper, A. Campbell, J. Yim, M. Lienen, Y. Shi, S. Zheng, H. Schulz, U. Munir, C. Clementi, F. Noé, Scalable emulation of protein equilibrium ensembles with generative deep learning, *bioRxiv* (2024)p. 2024.12.05.626885.

44. M. Cagiada, F. E. Thomasen, S. Ovchinnikov, C. M. Deane, K. Lindorff-Larsen, AF2χ: Predicting protein side-chain rotamer distributions with AlphaFold2, *bioRxiv* (2025)p. 2025.04.16.649219.

45. R. Nussinov, P. G. Wolynes, A second molecular biology revolution? The energy landscapes of biomolecular function. *Phys Chem Chem Phys* **16**, 6321–6322 (2014).

46. S. Du, R. C. Kretsch, J. Parres-Gold, E. Pieri, V. W. D. Cruzeiro, M. Zhu, M. M. Pinney, F. Yabukarski, J. P. Schwans, T. J. Martínez, D. Herschlag, Conformational ensembles reveal the origins of serine protease catalysis. *Science* **387**, eado5068 (2025).

47. J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, S. Ganguli, "Deep Unsupervised Learning using Nonequilibrium Thermodynamics" in *International Conference on Machine Learning* (PMLR, 2015), pp. 2256–2265.

48. S. Du, S. A. Wankowicz, F. Yabukarski, T. Doukov, D. Herschlag, J. S. Fraser, Refinement of multiconformer ensemble models from multi-temperature X-ray diffraction data. *Methods Enzymol.* **688**, 223–254 (2023).

49. D. A. Keedy, Z. B. Hill, J. T. Biel, E. Kang, T. J. Rettenmaier, J. Brandão-Neto, N. M. Pearce, F. von Delft, J. A. Wells, J. S. Fraser, An expanded allosteric network in PTP1B by multitemperature crystallography, fragment screening, and covalent tethering. *Elife* **7** (2018).

50. F. Yabukarski, J. T. Biel, M. M. Pinney, T. Doukov, A. S. Powers, J. S. Fraser, D. Herschlag, Assessment of enzyme active site positioning and tests of catalytic mechanisms through X-ray-derived conformational ensembles. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 33204–33215 (2020).

51. J. T. Biel, M. C. Thompson, C. N. Cunningham, J. E. Corn, J. S. Fraser, Flexibility and Design: Conformational Heterogeneity along the Evolutionary Trajectory of a Redesigned Ubiquitin. *Structure (London, England : 1993)* **25** (2017).

52. S. Russi, A. González, L. R. Kenner, D. A. Keedy, J. S. Fraser, H. van den Bedem, Conformational variation of proteins at room temperature is not dominated by radiation damage. *J. Synchrotron Radiat.* **24**, 73–82 (2017).

53. D. A. Keedy, L. R. Kenner, M. Warkentin, R. A. Woldeyes, J. B. Hopkins, M. C. Thompson, A. S. Brewster, A. H. Van Benschoten, E. L. Baxter, M. Uervirojnangkoorn, S. E. McPhillips, J. Song, R. Alonso-Mori, J. M. Holton, W. I. Weis, A. T. Brunger, S. M. Soltis, H. Lemke, A. Gonzalez, N. K. Sauter, A. E. Cohen, H. van den Bedem, R. E. Thorne, J. S. Fraser, Mapping the conformational landscape of a dynamic enzyme by multitemperature and XFEL crystallography. *Elife* **4** (2015).

54. How good are AlphaFold models for docking-based virtual screening? *iScience* **26**, 105920 (2023).

55. M. Holcomb, Y.-T. Chang, D. S. Goodsell, S. Forli, Evaluation of AlphaFold2 structures as docking targets. *Protein Science* **32**, e4530 (2023).

56. D. R. Hekstra, K. I. White, M. A. Socolich, R. W. Henning, V. Šrajer, R. Ranganathan, Electric-field-stimulated protein mechanics. *Nature* **540**, 400–405 (2016).