PNAS RESEARCH ARTICLE | BIOPHYSICS AND COMPUTATIONAL BIOLOGY

Predicting multiple conformations of ligand binding sites in proteins suggests that AlphaFold2 may remember too much

Maria Lazou[a], Omeir Khan[b], Thu Nguyen[c], Dzmitry Padhorny[d,e], Dima Kozakov[d,e,1], Diane Joseph-McCarthy[a,b], and Sandor Vajda[a,b,1]

Affiliations are included on p. 9.

Edited by Barry Honig, Columbia University, New York, NY; received June 24, 2024; accepted October 21, 2024

The goal of this paper is predicting the conformational distributions of ligand binding sites using the AlphaFold2 (AF2) protein structure prediction program with stochastic subsampling of the multiple sequence alignment (MSA). We explored the opening of cryptic ligand binding sites in 16 proteins, where the closed and open conformations define the expected extreme points of the conformational variation. Due to the many structures of these proteins in the Protein Data Bank (PDB), we were able to study whether the distribution of X-ray structures affects the distribution of AF2 models. We have found that AF2 generates both a cluster of open and a cluster of closed models for proteins that have comparable numbers of open and closed structures in the PDB and not too many other conformations. This was observed even with default MSA parameters, thus without further subsampling. In contrast, with the exception of a single protein, AF2 did not yield multiple clusters of conformations for proteins that had imbalanced numbers of open and closed structures in the PDB, or had substantial numbers of other structures. Subsampling improved the results only for a single protein, but very shallow MSA led to incorrect structures. The ability of generating both open and closed conformations for six out of the 16 proteins agrees with the success rates of similar studies reported in the literature. However, we showed that this partial success is due to AF2 "remembering" the conformational distributions in the PDB and that the approach fails to predict rarely seen conformations.

protein structure prediction | binding hot spot | conformational change | machine learning | protein mapping

Significance

After the success of protein structure prediction by AlphaFold2 (AF2), interests turned toward generating realistic conformational ensembles, and running AF2 with stochastic subsampling of the multiple sequence alignment (MSA) received substantial attention. It was shown that the method works only for some fraction of the proteins tested, and the origin of this limitation was not understood. We have shown that predicting multiple conformations requires comparably sized clusters of open and closed structures in the Protein Data Bank (PDB), whereas rarely seen conformations are usually not predicted. Our results emphasize the need for further method development and possibly for a combination of machine learning with physics-based search methods if the goal is generating entire conformational ensembles.

The binding of small molecules to proteins plays important roles in various biological functions, including enzyme catalysis, receptor activation, and drug action, and hence understanding or designing such processes frequently involves the detection and characterization of ligand binding sites (1–4). The release of the AlphaFold2 (AF2) and RoseTTafold programs has opened the possibility that such studies can be extended to previously uncharacterized proteins (5–8). AF2 uses a neural network architecture with attention-based components that take advantage of the evolutionary information extracted from multiple sequence alignments (MSAs), followed by a structural refinement module trained on X-ray crystal structures deposited to the PDB database. The AF2 predictions are primarily determined by the coevolutionary information contained in the MSA but are also influenced by the distribution of protein conformations in the PDB. For example, it was noted that when predicting the structures of the proteins that have both ligand-bound (holo) and ligand-free (apo) structures in the PDB, AF2 predicts the holo form in 70% of cases (9).

The goal of this work is investigating the ability of AF2 to generate conformational ensembles of ligand-binding sites in proteins. As revealed by the variety of X-ray structures available for many proteins (10, 11) and by molecular dynamics simulations (12–18), regions surrounding binding sites may exhibit a high degree of motion, characterized by movements of structural elements on which the binding tends to rely. We selected a benchmark set of proteins with so-called cryptic sites that have both a closed conformation, essentially undetectable in some structures without a bound ligand, and an open conformation, frequently but not necessarily with a bound ligand (10, 19, 20). In particular, the CryptoSite set includes 93 bound–unbound pairs in which each unbound structure had a site considered cryptic due to its low pocket score, and each bound structure had a biologically relevant ligand bound at the site (10). While the original set included only one unbound structure in each pair, it was shown that many of the proteins also have structures in the Protein Data Bank (PDB) with open binding sites without a bound ligand (19). The conformational changes have been studied in some of the proteins by molecular dynamics simulations (11, 15, 21, 22) and also by AF2 (23). Our analysis here

Author contributions: D.K., D.J.-M., and S.V. designed research; M.L., O.K., T.N., D.P., and D.K. performed research; M.L. and S.V. analyzed data; and M.L. and S.V. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: midas@laufercenter.org or vajda@bu.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2412719121/-/DCSupplemental.

Published November 20, 2024.

compares the geometry and druggability features of the experimental X-ray structures available in the PDB and the ensembles of models generated with AF2. Specifically, we explore how closely the diversity of these features in the models conserves the conformational variation seen in X-ray structures of the same protein.

It is well understood that protein function is defined by the existence of conformational ensembles (24–27). Generating multiple conformations of proteins by AF2 has recently received substantial attention (28–33). Heo and Feig employed active and inactive G-protein coupled receptor (GPCR) structures as templates and were able to predict models that accurately captured the main structural changes (28). A more general idea was introduced by Meiler and coworkers, who noticed that multiple conformations of GPCRs can be obtained by reducing the depth of the MSAs (34). Stochastic subsampling of the MSAs led to the generation of conformations that spanned the range between active and inactive structures. An alternative approach was used by Stein and Mchaourab, who manipulated the MSA via in silico mutagenesis (31). This idea was simplified by Kern and coworkers, who employed only naturally occurring mutations but suggested that clustering an MSA based on sequence similarity enables AF2 to sample alternate states (33). However, a follow-up analysis revealed that the clustering method incorrectly predicted some of the structures (35).

The above *methods* of generating multiple protein conformations are based on the hypothesis that the MSA must encode for protein structural heterogeneity, and hence its manipulation by stochastic subsampling or by selecting different clusters of sequences will enable AF2 to sample alternate conformations. Results demonstrate that this approach works in a variety of applications, at least for some of the proteins (36). However, several papers also reported partially negative results. We already mentioned the failures of the sequence clustering approach (35). Meller et al. studied the opening of cryptic pockets in 10 proteins by AF2 with subsampled MSA and found some predicted conformations with less than 1.2 A RMSD from the open (holo) structure in six of the 10 cases (23). However, the center of the clusters of predicted structures satisfied this distance condition only for two of the proteins, and hence, the authors used Markov state modeling to further open the pockets (23). Monteiro de Silva et al. reported generally successful predictions of relative populations of kinase conformations but noted that populations with small occupancy might be missed (37). Xie and Huang attempted capturing alternative conformational states of 16 membrane transporters but reported successful predictions of both inward-facing and outward-facing structures only for seven out of the 16 proteins using stochastic subsampling and only for three by sequence clustering (38).

In this paper, we focus on structures of cryptic binding sites, and hence we have natural extreme points to measure conformational diversity, i.e., the open versus the closed conformations of the pocket. While the opening of cryptic pockets by subsampled AF2 has been studied previously (23), we selected proteins with large numbers of conformations available in the PDB and generate large sets of models. This enables us to explore how well the conformational distributions of the AF2 models reproduce the distributions seen in the X-ray structures. The problem we study is further simplified by the fact that the conformational transition in each protein is primarily caused by the movement of a small segment, which may be a loop, a small secondary structure element, or even a single side chain, and thus, the distance of a particular predicted structure between the open and closed extremes can be easily determined. In view of the studies discussed in the previous paragraph, we try to answer why AF2 with subsampled MSA works for some proteins and not for others. Is it possible that the diversity and populations of the X-ray structures available in the PDB influence the likely outcome in terms of model diversity? In other words, how much does AlphaFold remember, and how well can it generate alternate conformations rarely seen in the PDB?

The second goal of our investigation is exploring the right level of subsampling required for generating multiple conformations. The key parameters for subsampling the MSA in the Colabfold implementation of AF2 are max_seq and max_extra_seq. The first parameter, max_seq, defines the number of sequences randomly selected from the master MSA (the target sequence is always selected). The remaining sequences are then clustered around these selected sequences using a Hamming distance. From each cluster, the cluster center and max_extra_seq additional sequences are used by AF2 for inference. To generate potentially diverse but high-quality models we first use a very conservative approach and run AF2 with the default Colabfold parameters (max_seq = 512 and max_extra_seq = 5120). For each protein target, we perform 100 runs with random initial seeds, each run resulting in five different models. While this protocol is generally used for structure prediction by selecting the highest confidence models (39, 40), it yields substantial conformational diversity for 30% of the proteins in our benchmark set. However, in the remaining 70%, the default parameters produce a single cluster of binding site conformations, and we explored the use of smaller values for max_seq and max_extra_seq. Previous works have shown that a significant reduction in these parameters may improve the diversity in the ensemble prediction (36), but we find that for the problem studied here, the predictions are fairly robust and substantially change only when the MSA becomes very shallow, affecting the overall quality of predictions. Thus, it is not clear whether there exists any general rule for selecting the best max_seq and max_extra_seq parameters. In view of the partial success, we try to identify the main factors that predict whether the subsampling approach will produce conformational diversity.

Results

Benchmark Set of Proteins with Cryptic Binding Sites. The 16 proteins in Table 1 have binding sites with both open and closed conformations in the PDB (41), each conformation represented by a reference PDB structure. We focus on the binding sites of the ligands indicated by their three-letter codes, cocrystallized with the protein structures shown by their bound PDB IDs. Table 1 also shows the numbers of unbound and ligand-bound structures and identifies whether the bound structure is open or closed. We restricted consideration to proteins that have both open and closed reference structures without missing residues and have 15 or more structures in the PDB. The difference between the two conformations is primarily due to a moving segment shown in Table 1 (*SI Appendix, Supplementary Methods*). Each sequence, identified with its AF PDB ID, was used for generating 500 models using the Colabfold version of the AF2 program with each of the following (max_seq, max_extra_seq) parameter pairs: (512, 5120), (156, 512), (64, 128), (32, 64), and (8, 16), where (512, 5120) is the default parameter pair. For each protein, we used the FTMove program (*SI Appendix, Supplementary Methods*) to select structures with at least 90% sequence identity to the structure labeled as the FTMove PDB ID. These additional structures form what we call the X-ray or PDB ensemble. The structures in this ensemble were then classified using PyMol as bound if they had a ligand overlapping with the ligand in the bound reference structure (42).

Table 1. Proteins used in the study

Protein	AF PDB ID	FTMove PDB ID	Closed Ref. PDB ID	Open Ref. PDB ID	Bound PDB ID	Bound Struct.	Ligand ID	Moving segment	Closed structures in PDB	Open structures in PDB	Other structures in PDB	Unbound structures in PDB	Bound structures in PDB
Bovine β -lactoglobulin	6GE7.A	1BSQ.A	1BSQ.A	1GX8.A	1GX8.A	Open	RTL	Ile84 - Asn90	42	58	11	60	51
KRAS	4EPW.A	4EPW.A	4EPR.A	4EPV.A	4EPV.A	Open	0QX	Met67, Tyr71	46	154	24	165	59
MAPK	2NPQ.A	2ZB1.A	2ZB1.A	2NPQ.A	2NPQ.A	Open	BOG	Met198	164	105	29	201	97
Pyruvate dehydrogenase kinase	2BU8.A	2BU8.A	2BU8.A	2BU2.A	2BU2.A	Open	TF1	Phe31	21	13	6	34	6
Ribonuclease A	1RHB.A	2W5K.B	1RHB.A	2W5K.B	2W5K.B	Open	NDP	His119	189	49	51	242	47
β-secretase	3IXJ.C	1W50.A	3IXJ.C	1W50.A	3IXJ.C	Closed	586	Gly66 - Glu77	178	105	18	121	180
TEM β-lactamase	1PZ0.A	1PZO.A	1JWP.A	1PZO.A	1PZO.A	Open	CBT	Ala217-Leu225	159	2	20	179	2
cAMP-dependent protein kinase	2GFC.A	2GFC.A	2GFC.A	2JDS.A	2JDS.A	Open	L20	Thr51 - Arg56	19	238	43	51	249
Glutamate receptor 2	1MY0.B	1MY0.B	1MY0.B	1N0T.D	1N0T.D	Open	AT1	Gly136-Ser142	237	60	3	32	268
AMPc beta-lactamase	2BLS.B	2BLS.B	2BLS.B	3GQZ.A	3GQZ.A	Open	GF7	Asn289-Leu293	180	5	57	238	4
Thrombin	1GHY.H	1HAG.E	1GHY.H	1HAG.E	1GHY.H	Closed	121	AKA	26	1	5	10	22
Adipocyte Lipid Droplet Binding Protein (ALDBP)	1ALB.A	1LIC.A	1ALB.A	1LIC.A	1LIC.A	Open	HDS	Phe57	52	8	3	13	50
Myosin II	2AKA.A	2AKA.A	2AKA.A	1YV3.A	1YV3.A	Open	BIT	Leu262, Tyr634	1	9	34	35	9
Ricin	1RTC.A	1RTC.A	1RTC.A	1BR6.A	1BR6.A	Open	PT1	Tyr80	61	24	28	76	37
Androgen receptor	2AX9.A	2AX9.A	2AX9.A	2PIQ.A	2PIQ.A	Open	RB1	Lys720, Met734	33	46	31	102	8
Hsp90	1YES.A	1YES.A	2QFO.B	2WI7.A	2WI7.A	Open	2KL	Asn106-Ile110	79	129	82	30	260

Models Generated by using AF2 with Default MSA Parameters. The most important data in Table 1 are the numbers of structures in the X-ray ensemble with the binding site in open or closed conformation, or in some "other" conformational state not overlapping with either of the two reference structures. As will be shown, these numbers largely determine whether AF2 can reproduce multiple conformations of the binding site, and this result is almost independent of the level of MSA subsampling until very small max_seq and max_extra_seq values are reached. We divide the 16 proteins in Table 1 into three groups based on their numbers of open and closed structures in the PDB. Group 1 is formed by the six proteins (β-lactoglobulin, KRAS, MAPK, pyruvate dehydrogenase kinase, ribonuclease A, and β-secretase) that have comparable numbers of open and closed structures, and hence, we refer to these proteins as having balanced open and closed states. In addition, all proteins in this group have only few "other" conformations that are distant from both open and closed states of the binding sites. Group 2 is formed by the six proteins (TEM β-lactamase, cAMP-dependent protein kinase, glutamate receptor 2, AmpC β-lactamase, thrombin, and adipocyte lipid droplet binding protein) that have imbalanced numbers of open and closed states, since either the closed or the open structures dominate in the PDB. The proteins in this group also have relatively few structures in "other" conformations. Finally, Group 3 consists of the remaining four proteins (myosin II, ricin, androgen receptor, and hsp90) that have both open and closed structures but also have comparable numbers of "other" conformations.

Proteins with Balanced Numbers of Open and Closed States. Fig. 1 shows distributions of binding site conformations and pocket volumes in the X-ray structures and the AF2 models of four Group 1 proteins. For each X-ray structure in the PDB ensemble and for each model in the AF2 ensemble, we determine the RMSD of the moving segment from both the open and the closed reference X-ray structures shown in Table 1. The moving segment is excluded during the alignment to the reference structures and the RMSD is calculated for only the moving residues. For proteins with moving loop segment, only alpha carbons are considered in the RMSD calculations, while for systems with side chains identified as moving segments, all-atom RMSD is calculated. In Figs. 1–3, the structures are represented in a 2D local coordinate system that shows the RMSD of the moving segment from the closed reference structure on the X axis and the RMSD of the moving segment from the open reference structure on the Y axis (Fig. 1, *Top* panels as examples). The figures also show the location of the cluster centers of the AF2 models. The cluster centers are color-coded according to their druggability scores (*SI Appendix*, *Supplementary Methods* for clustering the structures and calculating a druggability score).

We briefly describe the X-ray and AF2 ensembles and pocket volumes for the four Group 1 proteins shown in Fig. 1. In β-lactoglobulin, access to the ligand binding site is modulated by the loop residues Ile84-Asn90 (*SI Appendix*, Fig. S1). The loop opens upon ligand binding or when the pH of the environment is raised from 6 to 8 (43). The PDB ensemble contains more open than closed conformations, and most of the open structures have a small molecule bound at the binding site and a druggable score (Fig. 1 *A*, *Top*). The AF2 ensemble has a similar distribution pattern of conformations, with clusters of open and closed models (Fig. 1 *A*, *Second* row). The pocket volumes for bound and unbound structures in the PDB ensemble reflect the fact that with the open conformation, the pocket volume substantially increases (Fig. 1 *A*, *Third* row). The models in the AF2 ensemble have similar pocket volumes, with the same bimodal distribution, but with several structures exhibiting volumes above the maximum volume threshold in the PDB ensemble, demonstrating AF2's capability of substantially opening the binding site pocket in this protein (Fig. 1 *A*, *Bottom*).

In KRAS, a hydrophobic pocket is located between the alpha-2 helix of switch II (residues Gly60-Thr74) and the central beta-sheet. Binding of a small molecule within this pocket causes the alpha-2

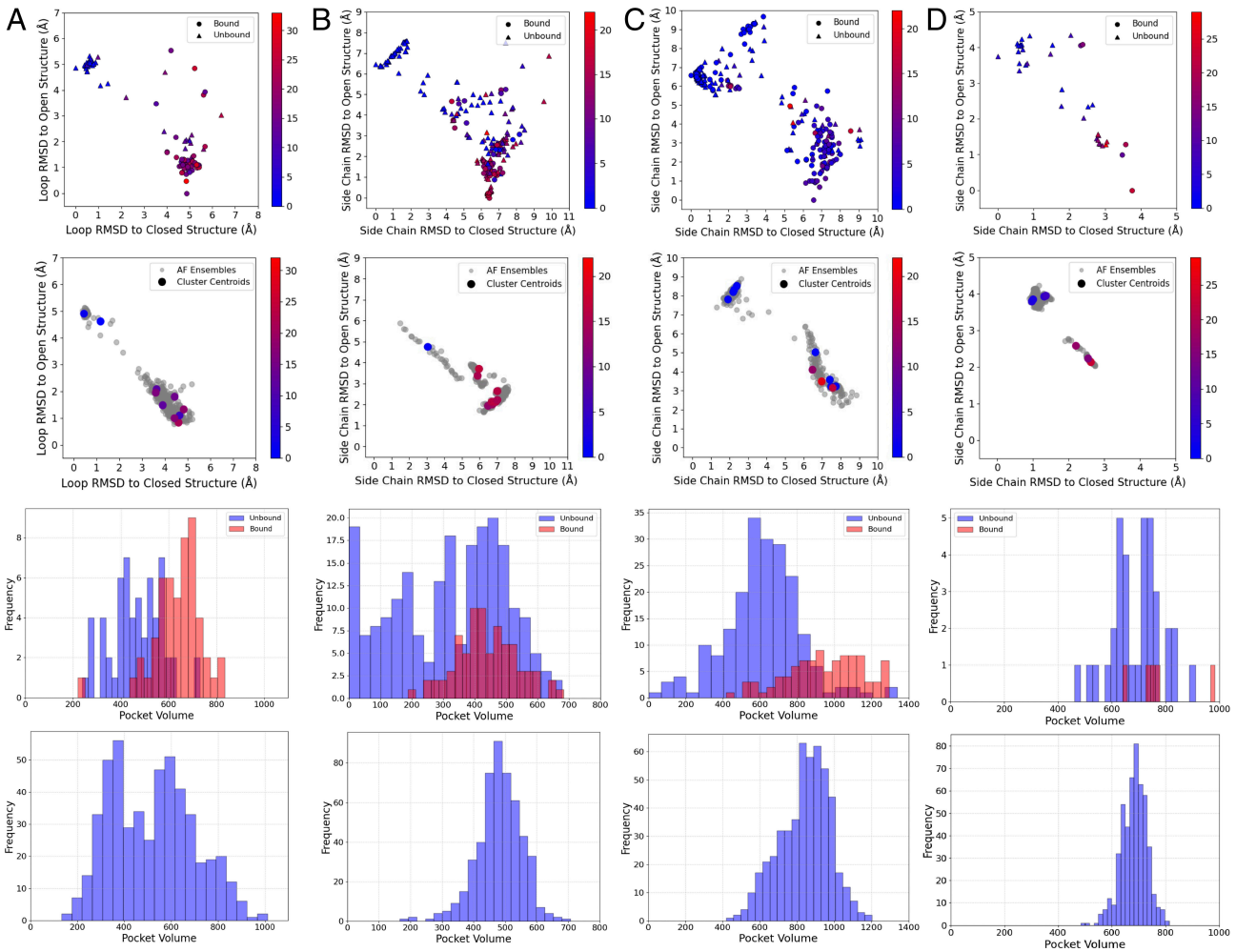

Fig. 1. Distributions of binding site conformations and pocket volumes in X-ray structures and AF2 models of Group 1 proteins with balanced distributions of open and closed states. (*A*) Bovine β lactoglobulin. (*B*) KRAS. (*C*) MAPK. (*D*) Pyruvate Dehydrogenase Kinase. Each column includes the same four subpanels. *Top*: RMSD of the moving fragment to open and closed reference structures in the X-ray structures of the PDB ensemble. Point shapes indicate the presence or absence of a ligand at the bindings site, and color indicates the druggability score. *Second* row: Same as the top panel for the predicted structures in the AF2 ensemble. Larger scatter points indicate cluster centroid structures, with color indicating druggability scores. *Third* row: Binding pocket volumes in the X-ray structures of the PDB ensemble. *Bottom*: Binding pocket volumes in the predicted structures of the AF2 ensemble.

helix to shift away from the beta-sheet. Concurrently, residue Tyr-71 disrupts a hydrogen bond and residue Met-67 rotates outward, thereby creating space within the cavity for an inhibitor to bind. The moving segment is defined by these two residues. In the majority of the bound KRAS structures in the PDB ensemble, the two residues turn away from the beta sheet (*SI Appendix*, Fig. S2), resulting in a larger cluster of open structures (Fig. 1 *B*, *Top*). The models in the AF2 ensemble have a similar pattern (Fig. 1 *B*, *Second* row), with more structures in open or partially open state. The volumes in PDB structures underscore the shallow nature of the pocket since both bound and unbound structures have similar pocket volumes, although some unbound structures are significantly smaller (Fig. 1 *B*, *Third* row). The AF2 set produces similar pocket volumes, but low volumes are absent (Fig. 1 *B*, *Bottom*).

In p38 MAPK, a lipid-binding allosteric site is formed by a local conformational change with an alpha-helix moving further away from the protein core (44). When a ligand binds, the Met198 residue rotates 180°, exposing its side chain, which is buried in the unbound state (*SI Appendix*, Fig. S3). Structures in the PDB have many open and closed conformations (Fig. 1 *C*, *Top*). The AF2 models tend to cluster into two main groups: one closer to the open helix conformation and the other closer to the closed helix (Fig. 1 *C*, *Second* row). The structures exhibit a tendency toward partially open states. Of these structures, three cluster centers displayed high druggability scores. The volume analysis shows that in the bound X-ray structures pocket volume increases and some unbound structures have a very small pocket (Fig. 1 *C*, *Third* row). The AF2 structures display a similar pattern but do not reach the same extremes as the PDB set (Fig. 1 *C*, *Bottom*).

Binding of ligands to pyruvate dehydrogenase kinase (PDHK) causes a hinge motion in helix α2, leading to a shift in the alpha-carbon position of Phe31 (*SI Appendix*, Fig. S4). This opens an induced pocket as the side chain of Phe31 changes from a lid position to an open conformation. Although there are relatively few structures of PDHK in the PDB, the RMSDs of the existing ones show that there is indeed a conformational shift in the position of Phe31 (Fig. 1 *D*, *Top*). The AF2 ensemble has an approximately even split between a cluster of structures with Phe31 residue closing the site and one with more open structures (Fig. 1 *D*, *Second* row). Due to low number of bound structures, it is challenging to determine whether the pocket volume increases with the residue shift. However, a few structures have high pocket volumes (Fig. 1 *D*, *Third* row). The pocket volumes of the AF2 models exhibit a narrower distribution (Fig. 1 *D*, *Bottom*). Placing the remaining two proteins, ribonuclease A and β-secretase, in Group 1 is less certain. For ribonuclease A, *SI Appendix*, Fig. S5 shows that the side chain of

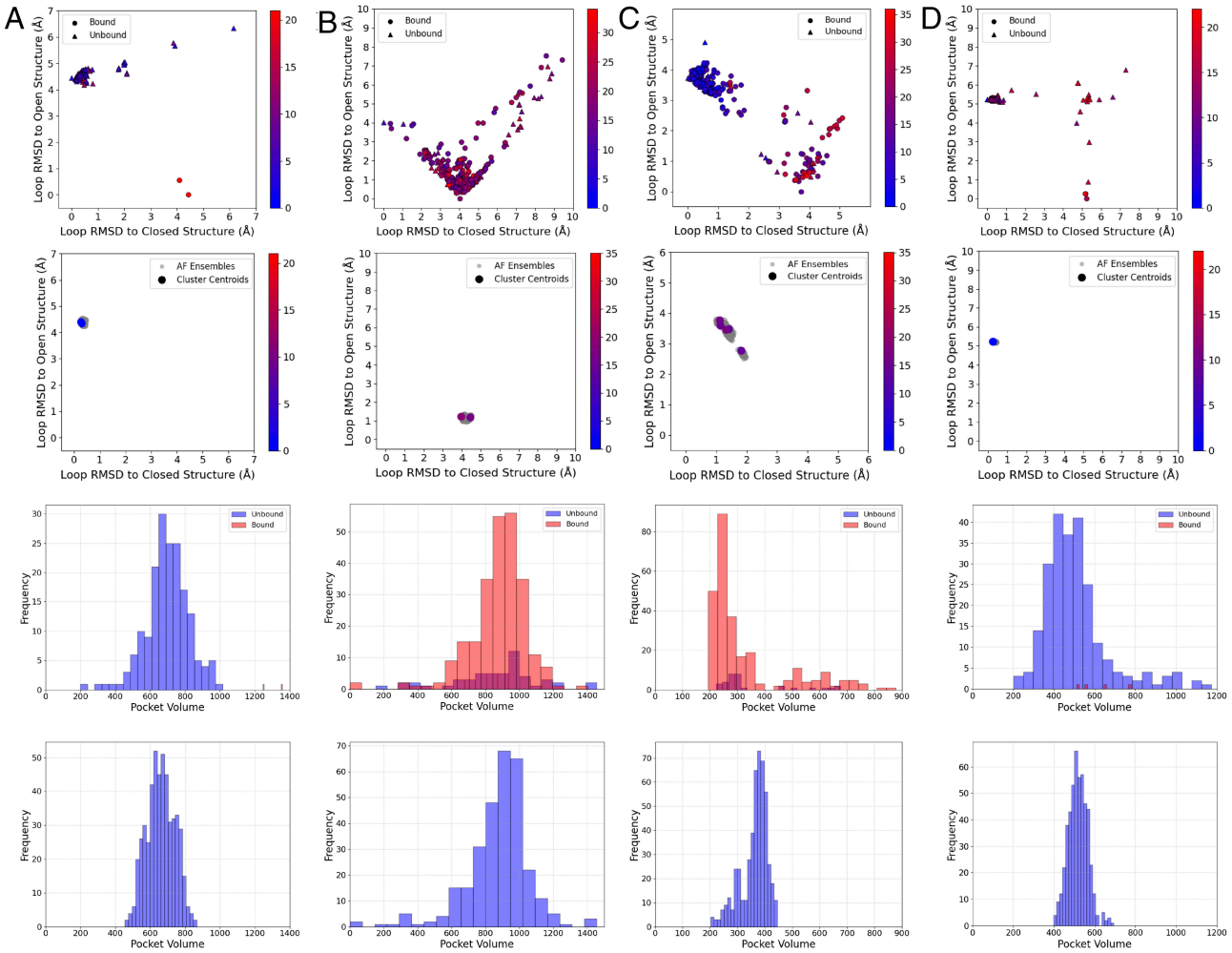

Fig. 2. Distributions of binding site conformations and pocket volumes in X-ray structures and AF2 models of Group 2 proteins with imbalanced distributions of open and closed states. (*A*) TEM β-lactamase. (*B*) cAMP-dependent protein kinase. (*C*) Glutamate receptor 2. (*D*) AMPc β -Lactamase. Each column includes the same four subpanels as in Fig. 1.

His119 moves out of the pocket upon ligand binding. With 189 open and only 49 closed X-ray structures, it is not clear whether these numbers are balanced enough for ribonuclease A to be in Group 1. In fact, AF2 yields only a single cluster of open structures (*SI Appendix*, Fig. S17). However, as will be discussed further in the paper, at reduced MSA depth AF2 yields both open and closed clusters (*SI Appendix*, Fig. S21*D*), thus at that point, ribonuclease A behaves as the other Group 1 proteins. In β-secretase, the pocket opens due to the motion of the β-hairpin loop of residues Gly66–Glu77 that form a mobile flap over the active site (*SI Appendix*, Fig. S6). This results in a fairly continuous distribution of X-ray structures between the open and closed reference states (*SI Appendix*, Fig. S18), in contrast to the other proteins in Group 1 that have fairly distinct clusters around the two states. Although based on our definition (*SI Appendix*, *Supplementary Methods*), β-secretase belongs to Group 1, it has many intermediate structures that are neither open nor closed, and hence could be placed in Group 3. Accordingly, all AF2 generated models form a single diffuse cluster around the closed structure (*SI Appendix*, Fig. S18). The PDB and AF2 structures have similar volumes, with a narrower distribution for the latter.

Proteins with Imbalanced Numbers of Open and Closed States. In contrast to the proteins in Group 1 with fairly similar numbers of open and closed structures in the PDB, the six proteins in Group 2 have dissimilar numbers of structures in open and closed states, in most cases with very limited number of structures available in one of the conformations. As shown in Fig. 2, for these proteins, the AF2 models resemble only one of the conformational states, usually the one with the higher number of X-ray structures in the PDB. We briefly describe the results for four Group 2 proteins shown in Fig. 2. Most PDB structures of TEM β-lactamase are very similar to the closed reference structure (*SI Appendix*, Fig. S7) and there are only two open structures (45), cocrystallized with small ligands that force two helices apart to form an allosteric site (Fig. 2 *A*, *Top*). AF2 is unable to open this cryptic site without the presence of ligands, producing a single cluster of closed models (Fig. 2 *A*, *Second* row). The Thr51–Arg56 loop in cAMP-dependent protein kinase is flexible and has a variety of positions (*SI Appendix*, Fig. S8), but the open conformations dominate both in the PDB (Fig. 2 *B*, *Top*) and in the AF2 ensemble (Fig. 2 *B*, *Second* row). Glutamate receptor 2 has some open structures in the PDB, but much higher number of closed ones (Fig. 2 *C*, *Top*), determined by the conformations of loop Gly136–Ser142 (*SI Appendix*, Fig. S9). AF2 places the loop between the two reference structures, and the models form a single cluster somewhat closer to the closed state (Fig. 2 *C*, *Second* row), with volumes that follow a narrow distribution (Fig. 2 *C*, *Bottom*). For AMPC β lactamase, the closed structures dominate both in the X-ray and AF2 ensembles (Fig. 2 *D*, *Top* and *Second* rows).

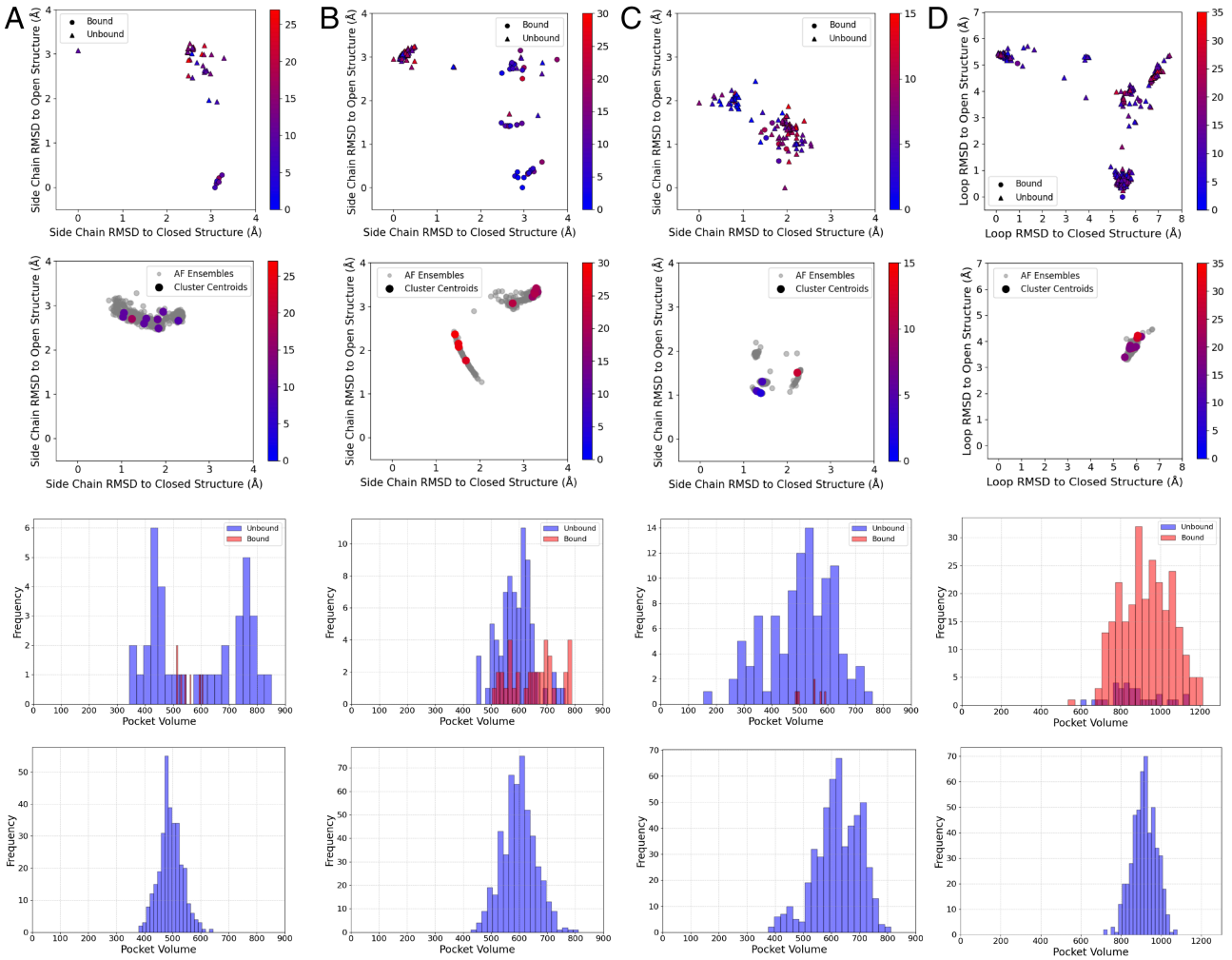

Fig. 3. Distributions of binding site conformations and pocket volumes in X-ray structures and AF2 models of Group 3 proteins with many conformations distant from open and closed states. (*A*) Myosin II. (*B*) Ricin. (*C*) Androgen receptor. (*D*) Hsp90. Each column includes the same four subpanels as in Fig. 1.

The other proteins in Group 2 with results in *SI Appendix* are thrombin and adipocyte lipid droplet binding protein (ALDBP). For thrombin, the impact of the moving Gly216 - Tyr225 loop is shown in *SI Appendix*, Fig. S11. Most X-ray structures are closed, and AF2 yields a single closed cluster (*SI Appendix*, Fig. S19). For ALDBP, the moving segment is the side chain of Phe57 (*SI Appendix*, Fig. S12), and the results disagree with the behavior of the other proteins in Group 2. In fact, while it has substantially more closed X-ray structures than open ones, AF2 generates a larger cluster of open states and a small cluster of closed ones (*SI Appendix*, Fig. S20).

Proteins with Many Structures in Neither Open and nor Closed States. The last four proteins in Table 1 have many structures in the PDB that are not close either to the closed or to the open reference states. AF2 is unable to generate both open and closed models, and in most cases, the majority of models mimic the binding site in the X-ray structures (Fig. 3). Myosin 2 has a single closed structure, nine open structures, and 28 structures that are equally distant from both reference structures (Fig. 3 *A*, *Top*). AF2 creates a single diffuse cluster that is closer to the closed state (Fig. 3 *A*, *Second* row). In the PDB ensemble the pocket volumes are broadly distributed (Fig. 3 *A*, *Third* row), and the distribution becomes more focused in the AF2 models (Fig. 3 *A*, *Bottom*). Ricin is a powerful cytotoxin widely used in the development of therapeutic agents (46). Ligand binding to the active site of Ricin requires the Tyr80 side chain to rotate by approximately 45° (*SI Appendix*, Fig. S14). The PDB structures of ricin form well-defined clusters near both the open and the closed reference structures and a third cluster of "other" structures far from both reference states (Fig. 3 *B*, *Top*). The AF2 models form an intermediate cluster located between closed and open states and a larger cluster close to the location of the "other" structures in the PDB ensemble (Fig. 3 *B*, *Second* row). The side chain conformation of Tyr80 seen in the intermediate cluster of AF2 models is not present in the PDB set. AF2 shows an increase in pocket volumes when the Tyr side chain moves outward (Fig. 3 *B*, *Bottom*). Although the AF2 ensemble lacks a fully open conformation, pocket volumes are similar to those in the PDB set.

The PDB structures of the androgen receptor have a cluster of conformations that are partially closed and a diffuse cluster that is about the same distance from both reference structures (Fig. 3 *C*, *Top*). AF2 reproduces only this diffuse cluster (Fig. 3 *C*, *Second* row). The AF2 models have narrower volume distribution than the PDB structures (Fig. 3 *C*, *Bottom*). In the heat shock protein 90 (hsp90), the 35 amino acid region of the geldanamycin binding domain can exist in open and closed conformations, altering the size of the binding pocket (*SI Appendix*, Fig. S16). In the PDB ensemble, the Asn106–Ile110 segment is observed to have three distinct conformations. The open conformations dominate (Fig. 3 *D*, *Top*), but there are also large clusters of closed and "other" structures. The AF2 models form only a cluster of such

"other" structures (Fig. 3 *D*, *Second* row). AF2 yields a much narrower distribution of pocket volumes than in the PDB ensemble (Fig. 3 *D*, *Third* and *Fourth* rows).

Models Generated by Using AF2 with Subsampled MSA. As discussed above, we were able to generate models with multiple conformations of the ligand binding sites only for four Group 1 proteins that have well-defined clusters of similar sizes near both the open and closed reference states, and not too many conformations anywhere else, plus for adipocyte lipid droplet binding protein (ALDBP). Since ALDBP has substantially fewer open than closed PDB structures, it is classified as a Group 2 protein, but interestingly, AF2 generates clusters of both open and closed models. However, these results are based on using AF2 with the default parameters used for predicting protein structures, whereas the literature describes successful generation of multiple conformations with some level of subsampling. Therefore, the logical next step of our study has been repeating model generation using increasingly subsampled MSAs by reducing the max_seq and max_extra_seq parameters. Based on the relevant reports, for each of the 16 proteins, we generated 500 models each in a series of four AF2 calculations using the following (max_seq, extra_seq) pairs: (156, 512), (64, 128), (32, 64) and (8, 16), in addition to the default values of (512, 5120). To our surprise, as described below, we have found the conformational distributions remarkably stable in spite of the major subsampling.

Fig. 4 shows, from left to right, RMSDs of the moving segment to open and closed reference structures for one example from each of the three protein groups. Results for the remaining 13 proteins are shown in *SI Appendix*. For bovine β lactoglobulin, representing Group 1 with balanced open and closed states, AF2 with the default parameters produces clusters of both open and closed conformations with more models in the open state. Reducing max_seq leaves this result essentially unchanged. The distinction between open and closed clusters becomes less defined, and a number of models diverge from the open and closed clusters at max_seq = 8 (Fig. 4*A*). Fig. 4*B* shows the distributions of AF2 models with reduced MSA for TEM β-lactamase, representing Group 2 with very different numbers of open and closed X-ray structures. As discussed, TEM β-lactamase seems to open only in the presence of ligands, and with the default parameters AF2 produces a single cluster of closed structures (Fig. 2 *A*, *Second* from *Top*) and also with max_seq = 156 (Fig. 4 *B*, *Left*). Reducing the max_seq and max_extra_seq parameters makes this single cluster more diffuse. At max_seq = 8, the cluster includes some partially open structures closer to the ligand-bound state (about 2 Å RMSD instead of the 4.3 Å RMSD obtained with the default parameters), but this is due to some unfolding of the helices rather than their movement seen in the bound structures. From Group 3, we show results for myosin II (Fig. 4*C*). For Myosin II, AF2 with the default MSA yields a diffuse cluster of largely closed structures (Fig. 3 *A*, *Second* row). Reducing MSA results in further widening of the cluster, and with max_seq = 8 yields many conformations equally far from the open and closed states, essentially the misfolding of the protein (Fig. 4 *C*, *Right*). In fact, the average global RMSD reaches 7.5 Å from the bound reference structure (Fig. 5 *C*, *Middle*).

Fig. 5 is a summary of features of AF2 models obtained with the series of (max_seq, max_extra_seq) pairs. Results are shown for each of the three groups of proteins separately. The left panels

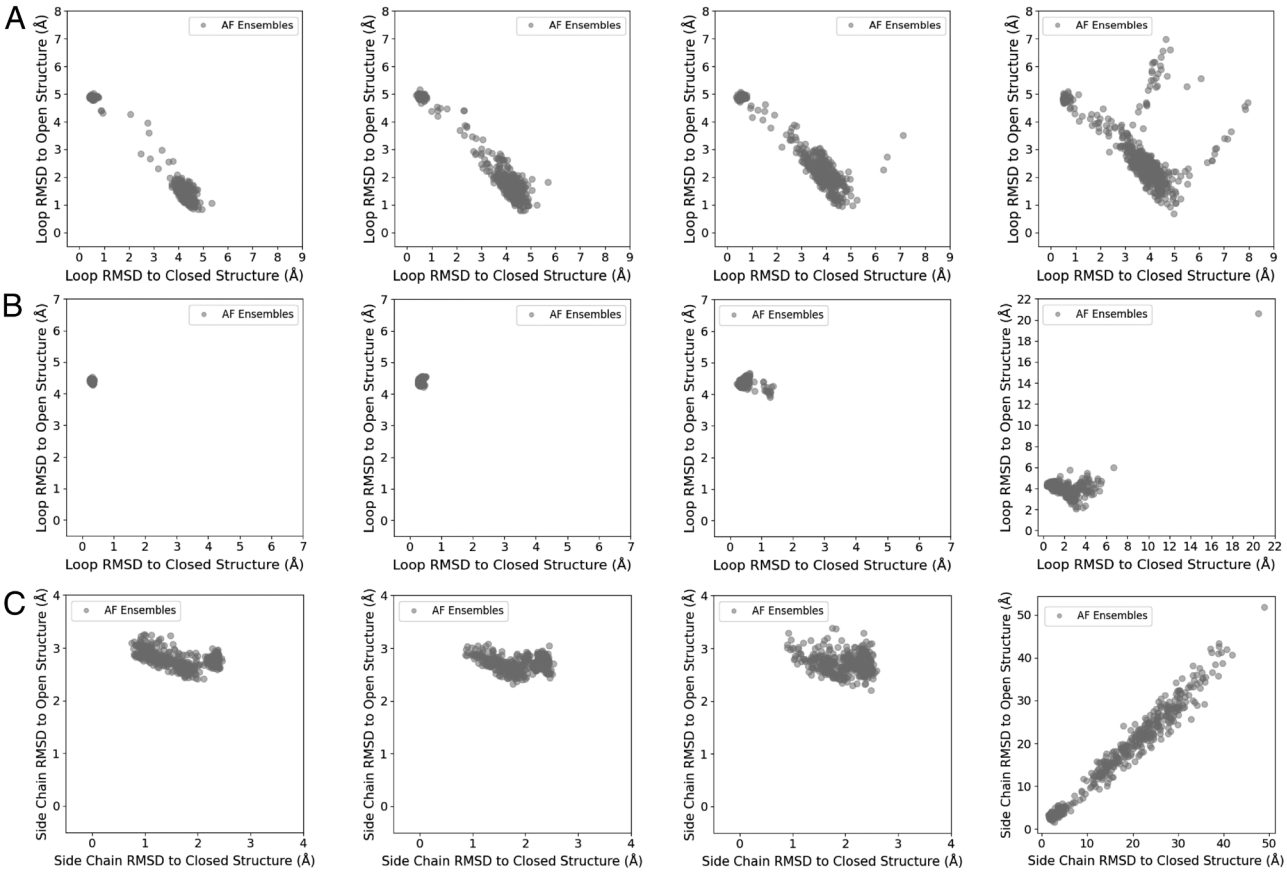

Fig. 4. RMSD of the moving fragment to open and closed reference structures with various levels of MSA reduction. (*A*) Bovine β lactoglobulin. (*B*) TEM β-lactamase. (*C*) Myosin II. For each protein, the four panels show RMSD distributions of AF2 generated models with the (max_seq, max_extra_seq) pairs of (156, 512), (64, 128), (32, 64) and (8, 16) from left to right.

show the normalized diversity distance as a function of the max_seq parameter. The concept of normalized diversity distance (NDD) was introduced to measure the maximum conformational diversity of models. To obtain NDD, we calculate the maximum RMSD of the moving segment between the most closed and most open structures in the AF2 ensemble and normalize this value with the maximum RMSD of the moving segment between the open and closed reference X-ray structures. For most Group 1 proteins, NDD exceeds 0.7, demonstrating that the models reproduce conformational diversity fairly well. The only exception is pyruvate dehydrogenase kinase, for which NDD increases with the decreasing max_seq, but then drops at max_seq = 8. At that point, the open and closed clusters of AF2 models merge (*SI Appendix*, Fig. S21 *C*, *Right* panel). At max_seq = 8, the same happens for KRAS and MAPK (*SI Appendix*, Fig. S21 *A* and *B*, *Right* panels). In fact, at max_seq = 8 most Group 1 protein models cease to form two well-defined clusters, although the global RMSD of the structures does not exceed 2.5 Å (Fig. 5 *A*, *Middle* panel). The only exception is Ribonuclease A. While ribonuclease A is an apparent exception in Group1 as it has a single AF2 cluster at default parameters, at max_seq = 8, AF2 actually yields clusters of open and closed conformations, demonstrating the behavior we had expected for all proteins (*SI Appendix*, Fig. S21 *D*, *Right* panel).

We recall that for proteins in Group 2, AF2 with max_seq = 512 yields single clusters of conformations with the exception of ALDBP, which has two model clusters at all MSA depths (*SI Appendix*, Fig. S22*E*). ALDBP also has a fairly large normalized diversity distance close to 0.8, and this property is conserved at reduced values of max_seq (Fig. 5 *B*, *Left*). For the other proteins in Group 2, the normalized diversity distance, initially small, increases as we reduce max_seq. Additionally, for the other Group 2 proteins AF2 with max_seq = 8 generates conformations that are equally distant from the open and closed reference structures (*SI Appendix*, Fig. S22). In fact, glutamate receptor 2 and AMPC β-lactamase tend to partially unfold in the shallow MSA runs (*SI Appendix*, Fig. S22 *B* and *C*), resulting in large global RMSD from the bound reference structure (Fig. 5 *B*, *Middle*). For the proteins in Group 3, reducing max_seq also yields single clusters (*SI Appendix*, Fig. S23) and increases the NDD values (Fig. 5). However, this is in most cases misleading, since the generated structures, while diverse with large NDD values, may be equally distant from the open and closed states as shown in *SI Appendix*, Fig. S27. No open and closed clusters are produced with reduced MSA for any of the Group 3 proteins (*SI Appendix*, Fig. S23). Reducing the MSAs generally results in higher proportion of incorrectly folded proteins. The average pLDDT is a good predictor of the average global RMSD (*SI Appendix*, Figs. S24–S26). The plDDT values of the moving segments have substantial variation (*SI Appendix*, Table S2), with little impact on the overall RMSD of the models.

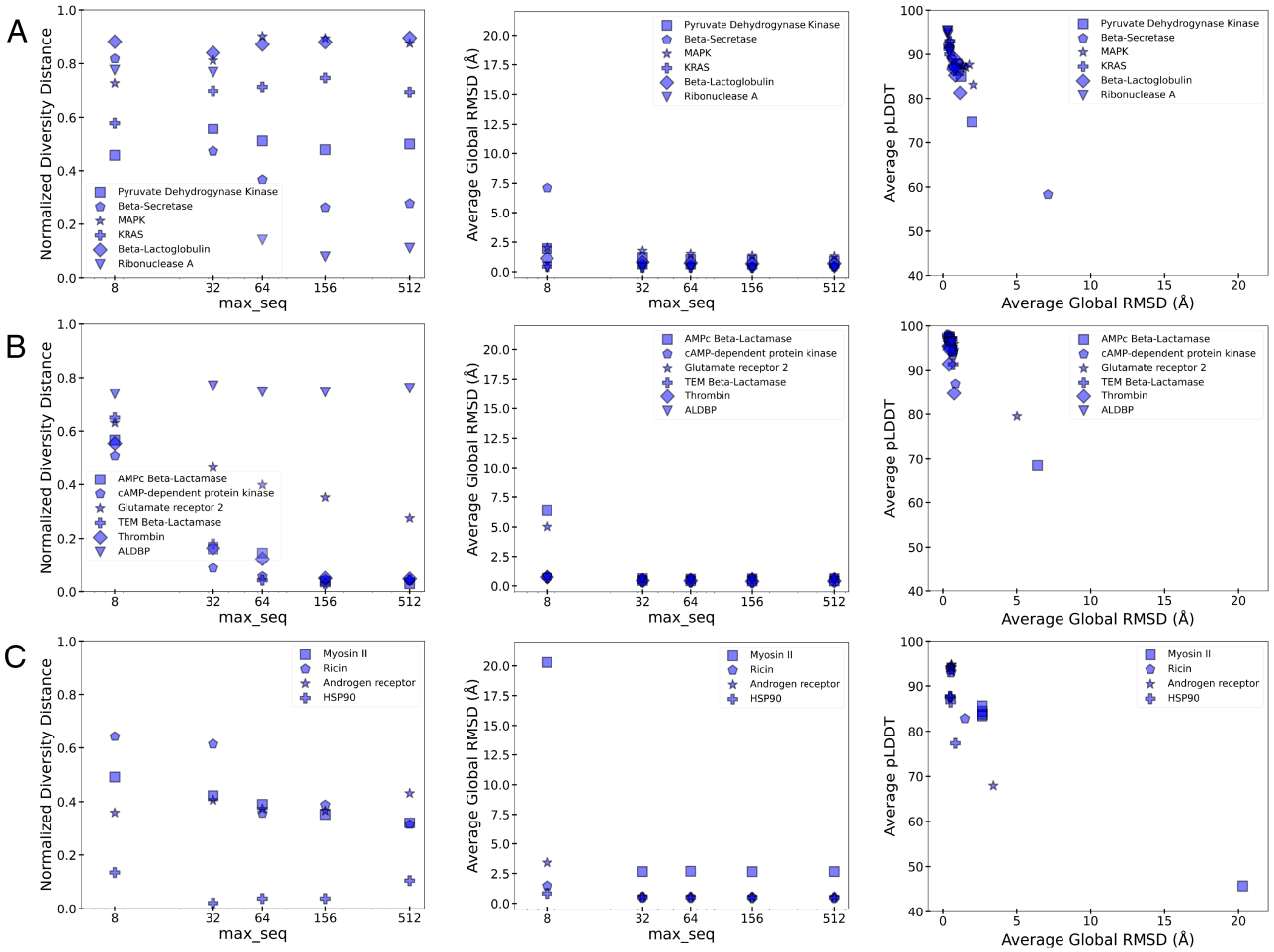

Fig. 5. Properties of AF2 models obtained using reduced MSAs for the three groups of proteins considered in this study. (*A*) Group 1 proteins with balanced numbers of open and closed states. (*B*) Group 2 proteins with imbalanced numbers of open and closed states. (*C*) Group 3 proteins with many structures in neither open and nor closed states. *Left* panels show the normalized diversity distances as a functions of the max_seq parameter. *Middle* panels show the average global RMSD values to the ligand-bound reference state, also as functions of max_seq. *Right* panels show average plDDT values as functions of the global RMSD.

Discussion

It has been reported in recent papers that running AF2 with subsampled MSA can produce multiple conformations, but only for some of the proteins studied (23, 35, 37, 38). Thus, it is an open question what properties of a protein predict success. We set out to explore the opening of cryptic ligand binding sites in 16 proteins, where the closed and open conformations define the expected extreme points of the conformational variation. Due to the many structures in the PDB, we were able to study whether the distribution of X-ray structures between closed and open states affect the distribution of AF2 models. We have found AF2 generates multiple clusters of models for proteins that have comparable numbers of closed and open structures in the PDB and not too many other conformations. Surprisingly, this property was observed both without and with subsampling of the MSA. The multiple conformations in the models cease to exist at very shallow MSA as the proteins start to misfold. The exceptions to this rule are ribonuclease A, which has a single AF2 generated cluster at default MSA (*SI Appendix*, Fig. S17), but clusters of open and closed structures at (max_seq, max_extra_seq) = (8, 16) (*SI Appendix*, Fig. S21*D*); thus, it behaves the way we had expected all proteins to in this study. As we mentioned, β-secretase has a continuum of conformations in the X-ray ensemble between open and closed states rather than two distinct clusters as the other Group 1 proteins, which may be the reason why its models form a single cluster at all parameters (*SI Appendix*, Fig. S21*E*)

In contrast to the Group 1 proteins shown in Fig. 1, AF2 generally does not yield multiple clusters of conformations for the proteins that have imbalanced numbers of open and closed structures in the PDB (Fig. 2) or have substantial numbers of structures distant from these states (Fig. 3). Thus, we conclude that AF2 seems to have a strong "memory" and fails to generate rarely seen conformations (37). Among the 16 proteins studied, we have found only one, adipocyte lipid droplet binding protein (ALDBP) that does not fully comply with this observation. In fact, ALDBP has many closed X-ray structures and only a few open ones, and yet, AF2 generates a larger cluster of open and a smaller cluster of closed models (*SI Appendix*, Fig. S20).

Following the substantial body of published work, we assumed that the right level of subsampling promotes generating multiple conformations. However, we obtained multiple conformations only for the proteins that also have multiple conformations in the PDB, and that this property was independent of the level of subsampling, except for using a very shallow MSA that may misfold the protein. In contrast, subsampling did not help for proteins that did not have models in multiple conformations using the default MSA. The only exception among the 16 proteins was the already-mentioned ribonuclease A. Since AF2 with subsampled MSA provided multiple conformations for six of the 16 proteins (i.e., five proteins in Group 1 and ALDBP), our success rate seems to agree with the rates reported in some earlier publications (23, 30, 34, 35, 37, 38). The difference to these previous papers is our ability to show that the success or failure for a particular protein is heavily dependent on the conformational distributions of the X-ray structures in the PDB, suggesting that AF2 has a nonnegligible level of memory, limiting the ability of the program to model rarely occurring conformations. A number of recent publications support this conclusion. Skolnick et al. have shown that AF2 predictions are determined by the TM-score, a structural similarity metric, of the closest structure in the training library to that of the target protein's structure (47). The Grigoryan group analyzed antibody modeling success with AF2 and found that it works mostly for cases where TERtiary Motifs (TERMs) from natural other proteins can be used to represent the CDR3 loops (48). Thus, in view of the limitation of learning due to incomplete training sets, it is possible that for generating conformational ensembles of proteins, it may be necessary to combine the machine learning approach with physics-based methods that are independent of the distributions of existing protein structures. It is important to note that our paper examines the relationships between the statistics of protein structures in the PDB used as the training set and the ability of AF2 of generating multiple conformations. The number of conformations of a protein in the PDB may depend on the interests in cocrystallizing the protein with many ligands, but also on existing physical constraints. For example, only a few ligands are known to open the cryptic allosteric site of TEM β-lactamase, and opening of the pocket could not be accomplished by molecular dynamics simulations without the presence of such ligands (11, 23). Investigating the relationships among physical reality, training set statistics, and machine learning memory is an important general problem for further studies,

Methods

Model Generation. The selected protein sequences were used as input for the ColabFold version of AF2 (35, 49). We have run AF2 using 100 random seeds, generating five models per seed and thus resulting in a total of 500 structural models for each (max_seq, extra_seq) parameter pairs.

Calculation of Pocket Volumes. We used the dpocket option of the Fpocket program to determine the volume of ligand binding pockets formed around the identified ligand (50).

Data, Materials, and Software Availability. Code for our RMSD and clustering analyses, FTMove calculations, details about bound proteins, MSAs utilized for Colabfold generations, and the Global RMSD/pLDDT data for each structure; Fpocket algorithm; FTMove server data have been deposited in github; FTmove (https://github.com/marialzs/AF_multiconformation; https://github.com/Discngine/fpocket; https://ftmove.bu.edu) (51). All study data are included in the article and/or *SI Appendix*.

ACKNOWLEDGMENTS. This work was supported by Grants R35GM118078 and RM1135136 from the National Institute of General Medical Sciences.

Author affiliations: [a]Department of Biomedical Engineering, Boston University, Boston, MA 02215; [b]Department of Chemistry, Boston University, Boston, MA 02215; [c]Department of Computer Science, Stony Brook University, Stony Brook, NY 11794; [d]Department of Applied Mathematics and Statistics, Stony Brook University, Stony Brook, NY 11794; and [e]Laufer Center for Physical and Quantitative Biology, Stony Brook University, Stony Brook, NY 11794

1. M. Nayal, B. Honig, On the nature of cavities on protein surfaces: Application to the identification of drug-binding sites. *Proteins* **63**, 892–906 (2006).
2. H. Hwang, F. Dey, D. Petrey, B. Honig, Structure-based prediction of ligand-protein interactions on a genome-wide scale. *Proc. Natl. Acad. Sci. U.S.A.* **114**, 13685–13690 (2017).
3. M. Gao, J. Skolnick, A comprehensive survey of small-molecule binding pockets in proteins. *PLoS Comput. Biol.* **9**, e1003302 (2013).
4. M. N. Wass, L. A. Kelley, M. J. Sternberg, 3DLigandSite: Predicting ligand-binding sites using similar structures. *Nucleic Acids Res.* **38**, W469–473 (2010).
5. J. Jumper *et al.*, Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
6. J. M. Thornton, R. A. Laskowski, N. Borkakoti, AlphaFold heralds a data-driven revolution in biology and medicine. *Nat. Med.* **27**, 1666–1669 (2021).
7. R. Tejero, Y. J. Huang, T. A. Ramelot, G. T. Montelione, AlphaFold models of small proteins rival the accuracy of solution NMR structures. *Front. Mol. Biosci.* **9**, 877000 (2022).
8. M. Baek *et al.*, Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).
9. T. Saldano *et al.*, Impact of protein conformational diversity on AlphaFold predictions. *Bioinformatics* **38**, 2742–2748 (2022).
10. P. Cimermancic *et al.*, CryptoSite: Expanding the druggable proteome by characterization and prediction of cryptic binding sites. *J. Mol. Biol.* **428**, 709–719 (2016).

11. Z. Sun, A. E. Wakefield, I. Kolossvary, D. Beglov, S. Vajda, Structure-based analysis of cryptic-site opening. *Structure* **28**, 223–235 e222 (2020).
12. J. D. Durrant, J. A. McCammon, Molecular dynamics simulations and drug discovery. *BMC Biol.* **9**, 71 (2011).
13. F. R. Salsbury Jr., Molecular dynamics simulations of protein dynamics and their relevance to drug discovery. *Curr. Opin. Pharmacol.* **10**, 738–744 (2010).
14. A. Ivetac, J. A. McCammon, A molecular dynamics ensemble-based approach for the mapping of druggable binding sites. *Methods Mol. Biol.* **819**, 3–12 (2012).
15. V. Oleinikovas, G. Saladino, B. P. Cossins, F. L. Gervasio, Understanding cryptic pocket formation in protein targets by enhanced sampling simulations. *J. Am. Chem. Soc.* **138**, 14257–14263 (2016).
16. G. R. Bowman, E. R. Bolin, K. M. Hart, B. C. Maguire, S. Marqusee, Discovery of multiple hidden allosteric sites by combining Markov state models and experiments. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 2734–2739 (2015).
17. C. D. Wassman *et al.*, Computational identification of a transiently open L1/S3 pocket for reactivation of mutant p53. *Nat. Commun.* **4**, 1407 (2013).
18. P. Ghanakota, H. A. Carlson, Moving beyond active-site detection: MixMD applied to allosteric systems. *J. Phys. Chem. B* **120**, 8685–8695 (2016).
19. D. Beglov *et al.*, Exploring the structural origins of cryptic sites on proteins. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E3416–E3425 (2018).
20. S. Vajda, D. Beglov, A. E. Wakefield, M. Egbert, A. Whitty, Cryptic binding sites on proteins: Definition, detection, and druggability. *Curr. Opin. Chem. Biol.* **44**, 1–8 (2018).
21. G. R. Bowman, P. L. Geissler, Equilibrium fluctuations of a single folded protein reveal a multitude of potential cryptic allosteric sites. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 11681–11686 (2012).
22. S. R. Kimura, H. P. Hu, A. M. Ruvinsky, W. Sherman, A. D. Favia, Deciphering cryptic binding sites on proteins by mixed-solvent molecular dynamics. *J. Chem. Inf. Model.* **57**, 1388–1401 (2017).
23. A. Meller, S. Bhakat, S. Solieva, G. R. Bowman, Accelerating cryptic pocket discovery using AlphaFold. *J. Chem. Theory Comput.* **19**, 4355–4363 (2023). 10.1021/acs.jctc.2c01189.
24. D. D. Boehr, R. Nussinov, P. E. Wright, The role of dynamic conformational ensembles in biomolecular recognition. *Nat. Chem. Biol.* **5**, 789–796 (2009).
25. I. H. Moal, R. Agius, P. A. Bates, Protein-protein binding affinity prediction on a diverse set of structures. *Bioinformatics* **27**, 3002–3009 (2011).
26. J. O. Wrabl *et al.*, The role of protein conformational fluctuations in allostery, function, and evolution. *Biophys. Chem.* **159**, 129–141 (2011).
27. T. Bohnuud, D. Kozakov, S. Vajda, Evidence of conformational selection driving the formation of ligand binding sites in protein-protein interfaces. *PLoS Comput. Biol.* **10**, e1003872 (2014).
28. L. Heo, M. Feig, Multi-state modeling of G-protein coupled receptors at experimental accuracy. *Proteins* **90**, 1873–1885 (2022).
29. G. Janson, G. Valdes-Garcia, L. Heo, M. Feig, Direct generation of protein conformational ensembles via machine learning. *Nat. Commun.* **14**, 774 (2023).
30. D. Chakravarty, L. L. Porter, AlphaFold2 fails to predict protein fold switching. *Protein. Sci.* **31**, e4353 (2022).
31. R. A. Stein, H. S. McHaourab, SPEACH_AF: Sampling protein ensembles and conformational heterogeneity with Alphafold2. *PLoS Comput. Biol.* **18**, e1010483 (2022).
32. L. E. Zheng, S. Barethiya, E. Nordquist, J. Chen, Machine learning generation of dynamic protein conformational ensembles. *Molecules* **28**, 4047 (2023).
33. H. K. Wayment-Steele *et al.*, Predicting multiple conformations via sequence clustering and AlphaFold2. *Nature* **625**, 832–839 (2024).
34. D. Del Alamo, D. Sala, H. S. McHaourab, J. Meiler, Sampling alternative conformational states of transporters and receptors with AlphaFold2. *eLife* **11**, e75751 (2022).
35. J. W. Schafer, D. Chakravarty, E. A. Chen, L. L. Porter, Sequence clustering confounds AlphaFold2. bioRxiv [Preprint] (2024). https://doi.org/10.1101/2024.01.05.574434 (Accessed 1 August 2024).
36. D. Sala, F. Engelberger, H. S. McHaourab, J. Meiler, Modeling conformational states of proteins with AlphaFold. *Curr. Opin. Struct. Biol.* **81**, 102645 (2023).
37. 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).
38. T. Xie, J. Huang, Can protein structure prediction methods capture alternative conformations of membrane transporters? *J. Chem. Inf. Model.* **64**, 3524–3536 (2024).
39. I. Johansson-Akhe, B. Wallner, Improving peptide-protein docking with AlphaFold-Multimer using forced sampling. *Front. Bioinform.* **2**, 959160 (2022).
40. B. Wallner, AFsample: Improving multimer prediction with AlphaFold using massive sampling. *Bioinformatics* **39**, btad573 (2023).
41. H. M. Berman *et al.*, The protein data bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
42. Anonymous, The PyMOL Molecular Graphics System. Version 1.2. (Schrödinger, LLC, 2000).
43. G. Kontopidis, C. Holt, L. Sawyer, The ligand-binding site of bovine beta-lactoglobulin: Evidence for a function? *J. Mol. Biol.* **318**, 1043–1055 (2002).
44. R. Diskin, D. Engelberg, O. Livnah, A novel lipid binding site formed by the MAP kinase insert in p38 alpha. *J. Mol. Biol.* **375**, 70–79 (2008).
45. J. R. Horn, B. K. Shoichet, Allosteric inhibition through core disruption. *J. Mol. Biol.* **336**, 1283–1291 (2004).
46. K. Jasheway, J. Pruet, E. V. Anslyn, J. D. Robertus, Structure-based design of ricin inhibitors. *Toxins (Basel)* **3**, 1233–1248 (2011).
47. J. Skolnick, M. Gao, H. Zhou, S. Singh, AlphaFold 2: Why it works and its implications for understanding the relationships of protein sequence, structure, and function. *J. Chem. Inf. Model.* **61**, 4827–4831 (2021).
48. K. M. McCoy, M. E. Ackerman, G. Grigoryan, A comparison of antibody-antigen complex sequence-to-structure prediction methods and their systematic biases. *Protein. Sci.* **33**, e5127 (2024).
49. M. Mirdita *et al.*, ColabFold: Making protein folding accessible to all. *Nat. Methods* **19**, 679–682 (2022).
50. P. Schmidtke, V. Le Guilloux, J. Maupetit, P. Tuffery, fpocket: Online tools for protein ensemble pocket detection and tracking. *Nucleic Acids Res.* **38**, W582–W589 (2010).
51. M. Lazou, Predicting multiple conformations of ligand binding sites in proteins suggests that AlphaFold2 may remember too much. Additional Data GitHub Repository. https://github.com/marialzs/AF_multiconformation. Deposited 1 June 2024.