JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

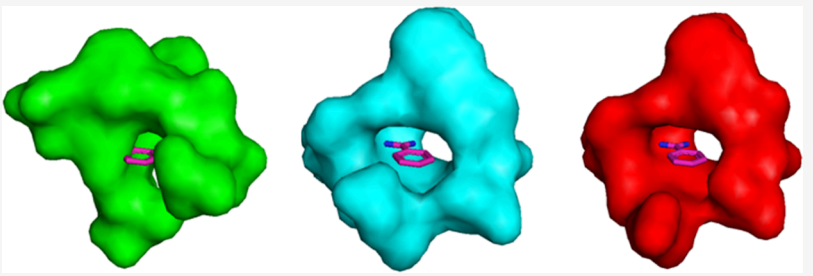
pubs.acs.org/jcim Article

Conservation of Hot Spots and Ligand Binding Sites in Protein Models by AlphaFold2

Ayse A. Bekar-Cesaretli, Omeir Khan, Thu Nguyen, Dima Kozakov, Diane Joseph-Mccarthy, and Sandor Vajda*

Cite This: J. Chem. Inf. Model. 2024, 64, 960−973 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: The neural network-based program AlphaFold2 (AF2) provides high accuracy structure prediction for a large fraction of globular proteins. An important question is whether these models are accurate enough for reliably docking small ligands. Several recent papers and the results of CASP15 reveal that local conformational errors reduce the success rates of direct ligand docking. Here, we focus on the ability of the models to conserve the location of binding hot spots, regions on the protein surface that significantly contribute to the binding free energy of the protein−ligand interaction. Clusters of hot spots predict the location and even the druggability of binding sites, and hence are important for computational drug discovery. The hot spots are determined by protein mapping that is based on the distribution of small fragment-sized probes on the protein surface and is less sensitive to local conformation than docking. Mapping models taken from the AlphaFold Protein Structure Database show that identifying binding sites is more reliable than docking, but the success rates are still 5% to 10% lower than based on mapping X-ray structures. The drop in accuracy is particularly large for models of multidomain proteins. However, both the model binding sites and the mapping results can be substantially improved by generating AF2 models for the ligand binding domains of interest rather than the entire proteins and even more if using forced sampling with multiple initial seeds. The mapping of such models tends to reach the accuracy of results obtained by mapping the X-ray structures.

INTRODUCTION

Machine learning methods in general and the AlphaFold2 (AF2) program in particular represent major advances in protein structure prediction.[1−3] AF2 was shown to yield excellent results in protein structure prediction challenge CASP14 in 2020,[4] and slightly modified versions of the method involving forced sampling also dominated all other prediction tools at CASP15 in 2022.[5,6] The method was also used with success for protein-peptide[6,7] and protein−protein docking,[8−10] and a specific version, AlphaFold-Multimer (AFM), has been developed.[9] AFM yields higher success rates (up to about 65%) for reconstructing complexes in protein−protein benchmark sets, and there are only a few specific applications with limited evolutionary information such as antibody−antigen docking where physics-based methods may be able to compete with AFM. Due to their successes, AF2 and AFM have become the most important tools for computational studies of protein structures and interactions, giving rise to extensive development and use of machine learning in many related areas of biology.

One of the most important potential applications is the use of AF2-generated models for drug discovery.[11,12] The most relevant method in question is the docking of small organic molecules to the models, allowing for high-throughput virtual screening. A number of recent studies have addressed the feasibility and potential accuracy of these applications. Scardino *et* al.[13] compared the performance of AF models in high-throughput docking (HTD) to their corresponding experimental PDB structures using a benchmark set of 16 targets spanning different protein families. They reported that the AF models showed consistently and substantially worse performance than the PDB structures. The shortcoming of this study was that it

Received: November 4, 2023
Revised: December 29, 2023
Accepted: January 3, 2024
Published: January 22, 2024

https://doi.org/10.1021/acs.jcim.3c01761
J. Chem. Inf. Model. 2024, 64, 960−973

Table 1. All-Atom RMSDs of AlphaFold Models to Ligand-Bound and Unbound Structures and Strongest Hot Spots at the Ligand Binding Sites[a,b,c,d,e]

									strongest hot spot		
no.	UniProt ID	Frag. ID	Frag. PDB ID	unbound PDB ID	global RMSD bound	local RMSD bound	global RMSD unbound	local RMSD unbound	AF	bound	unbound
1	P55201	12Q	5T4U_A	4LC2_A	1.277	1.033	1.333	0.996	00(21)	00 (19)	00 (22)
2	Q92831	12Q	5FE1_A	5FE6_A	1.513	0.883	1.698	0.944	01(17)	00 (21)	00 (24)
3	P11142	1LQ	5AQP_E	5AQM_A	3.743	1.219	3.795	0.447	none	04 (10)	none
4	P07900	42C	3HZ1_A	5J80_A	6.697	1.463	6.464	0.888	00(29)	00 (22)	00 (22)
5	P07900	2AE	2YE6_A	5J80_A	6.518	1.855	6.464	1.476	01 (20)	01 (14)	00 (22)
6	P07900	XQ0	2YEC_A	5J80_A	6.409	1.938	6.464	1.569	00 (32)	00 (28)	00 (22)
7	P56817	8AP	2OHM_A	3TPJ_A	1.449	1.064	1.080	0.892	00 (25)	00 (29)	00 (21)
8	P56817	2AQ	2OHL_A	3TPJ_A	1.449	1.579	1.080	0.818	00 (21)	00 (16)	01 (20)
9	P56817	EV0	3HVG_A	3TPJ_A	1.160	1.403	1.080	1.049	00 (20)	00 (24)	00 (21)
10	O60885	3PF	4DON_A	4LYI_A	1.307	0.843	1.469	0.633	03 (11)	00 (26)	00 (27)
11	Q13526	4BX	3KAC_A	2ZQT_A	1.060	0.639	1.747	0.826	00 (16)	00 (20)	00 (17)
12	P08709	7XM	5PAW_B	1JBU_H	1.246	1.065	4.360	4.970	00 (15)	00 (27)	none
13	P08709	AX7	5PAR_C	1JBU_H	1.209	0.878	4.360	4.333	00 (13)	00 (28)	none
14	O95696	8T1	5POE_A	5PQI_B	1.318	0.895	1.324	0.969	00 (23)	03 (10)	00 (24)
15	P25440	A9P	4ALH_A	5IBN_A	1.162	0.824	1.140	0.799	03(12)	00 (29)	00 (27)
16	P25440	TVP	4A9H_A	5IBN_A	1.182	1.135	1.140	0.911	00(23)	00 (24)	00 (27)
17	B9MKT4	ADA	4YZ0_B	3T9G_A	0.817	0.657	0.772	0.766	04 (08)	none	none
18	P00720	ALE	4LDO_A	none	1.352	0.883			01 (13)	04 (09)	none
19	Q7N561	AMG	5ODU_C	5OFZ_B	1.502	0.855	1.465	0.766	03 (14)	01 (15)	03 (13)
20	P28720	AQO	1S39_A	4Q8M_A	0.833	1.011	0.844	0.828	01 (18)	06 (06)	00 (22)
21	P00734	BEN	3P70_H	2UUF_B	2.995	4.658	2.890	4.732		00 (20)	00 (22)
22	P9WIL5	BZ3	3IMC_A	3COV_B	1.552	0.344	1.262	0.371	02 (12)	01 (19)	00 (26)
23	P28482	CAQ	4ZXT_A	4S31_A	2.370	0.559	3.225	0.734	00 (24)	00 (21)	02 (13)
24	P47228	CAQ	1KND_A	1HAN_A	1.222	0.761	1.171	0.854	00 (25)	00 (31)	01 (15)
25	P80188	CAQ	3FW4_C	none	0.930	0.728			01 (23)	00 (22)	
26	Q3JRA0	CYT	3MBM_A	none	0.971	1.211			04 (06)	10 (03)	
27	Q63T71	CYT	3IKE_B	none	1.081	0.559				03 (12)	
28	P15555	DAL	1IKI_A	none	0.559	0.885			02 (15)	00 (22)	
29	P00918	1SA	2HNC_A	3KS3_A	0.950	0.452	0.903	0.650	00 (28)	00 (25)	00 (16)
30	P00918	EVJ	4N0X_B	3KS3_A	0.853	0.608	0.903	0.633	00 (28)	00 (26)	00 (16)
31	P00918	FB2	2WEJ_A	3KS3_A	0.821	0.645	0.903	0.650	00 (27)	00 (23)	00 (16)
32	P00918	M3T	4Q9Y_A	3KS3_A	0.907	0.672	0.903	0.634	00 (31)	02 (15)	00 (16)
33	P00918	RCO	4E49_A	3KS3_A	0.945	1.038	0.903	1.117	01 (21)	02 (16)	none
34	P68400	GAB	5CSV_A	5CVG_A	1.486	1.470	2.363	2.097	00(17)	01 (19)	07 (04)
35	P54818	GAL	4CCE_A	none	0.668	0.637			02 (14)	00 (24)	
36	A0A083Z	GLA	6EQ0_B	none	2.371	1.218			00(22)	00 (21)	none
37	P32890	GLA	1DJR_G	1LTS_D	0.979	0.900	1.039	0.965	03 (14)	01 (18)	02 (15)
38	P42592	GLA	3W7U_B	3D3I_B	1.540	0.413	1.524	0.490	00(23)	00 (18)	01 (17)
39	Q57193	GLA	5ELB_D	5LZJ_B	0.629	0.628	0.993	0.834	02(16)	04 (11)	03 (14)
40	Q9ALJ4	GLA	4FNU_B	4FNQ_A	1.117	0.419	0.944	0.998	02(12)	00 (17)	01 (13)
41	P39900	HAE	1OS2_D	2MLR_A	1.213	0.529	2.433	1.662	00(22)	00 (17)	none
42	P39900	M4S	3LKA_A	2MLR_A	1.022	0.731	2.433	1.190	00(25)	00 (17)	03 (09)
43	Q9H2K2	JPZ	4PNN_B	4PNT_D	7.195	7.520	6.563	7.769		02 (12)	none
44	P24941	LZI	2VTA_A	4EK3_A	2.362	0.785	2.371	0.737	01 (20)	00 (22)	00 (29)
45	P24941	LZ5	2VTL_A	4EK3_A	2.442	0.862	2.371	0.904	01(18)	00 (25)	00 (29)
46	P24941	LZM	2VTM_A	4EK3_A	2.415	0.714	2.371	0.915	00(19)	00 (17)	00 (29)
47	P09874	MEW	4GV7_B	4XHU_A	1.361	1.338	1.312	0.937	02(15)	05 (09)	02 (13)
48	P29477	MR1	2ORQ_A	none	3.108	1.470			none	none	
49	P29477	MSR	2ORQ_A	none	3.108	1.278			01 (11)	01 (16)	
50	Q10588	NCA	1ISM_A	1ISF_B	0.799	0.886	5.841	0.876	00 (21)	00 (25)	00 (19)
51	Q05603	NIO	1L4N_A	none	0.623	0.636			03 (11)	04 (08)	
52	Q08638	NOJ	1OIM_A	5OSS_A	0.957	0.789	0.881	0.772	00 (22)	02 (18)	01 (17)
53	Q4D3W2	ORO	2E6A_B	none	0.784	0.327			02 (12)	01 (16)	
54	P0ABQ4	Q24	3QYO_A	1RA9_A	0.823	0.687	1.174	1.399	00 (25)	00 (19)	00 (32)
55	P19491	SHI	1MS7_A	none	1.325	1.103			00 (22)	04 (11)	
56	P06820	ST3	1IVE_A	4H53_D	0.889	0.873	1.127	1.010	04 (12)	00 (29)	none
57	Q6PL18	TDR	4QSU_A	4QSQ_A	1.620	0.907	1.617	0.736	00 (19)	01 (21)	00 (23)
58	Q6TFC6	TDR	3FS8_B	none	0.893	0.363				none	

https://doi.org/10.1021/acs.jcim.3c01761
J. Chem. Inf. Model. 2024, 64, 960−973

Table 1. continued

									strongest hot spot		
no.	UniProt ID	Frag. ID	Frag. PDB ID	unbound PDB ID	global RMSD bound	local RMSD bound	global RMSD unbound	local RMSD unbound	AF	bound	unbound
59	Q8K4Z3	TDR	3RO7_A	none	0.823	0.482			04 (07)	00 (22)	
60	Q92793	TYL	4A9K_B	5KTU_B	1.049	1.279	1.345	1.127	00 (19)	00 (26)	00 (25)
61	Q9WYE2	ZWZ	2ZWZ_A	1HL8_B	1.093	0.872	1.178	0.732	00 (21)	02 (13)	00 (28)
62	P16083	ZXZ	3NHW_A	none	0.881	0.808			none	none	

[a]PDB ID and chain ID of the X-ray structure with the bound ligand. [b]PDB ID and chain ID of the unbound structure. "None" indicates that no X-ray structure was available in the Protein Data Bank. [c]All-atom RMSD from the sequence and 3D alignment of the bound X-ray structure and the AF model. [d]All-atom RMSD from the sequence and 3D alignment of the unbound X-ray structure and the AF model. [e]Strongest hot spot with at least 50% coverage of the fragment binding site. Hot spots are numbered starting at 00 as established in the FTMap server. The number of probe clusters is given in parentheses. "None" (all columns) indicates that no such hot spot is found. "−" (column "AF") indicates no coverage is detected for any hot spot. "−" (column "Unbound") indicates there is no unbound X-ray structure available.

considered only protein structures cocrystallized with the known ligands, generally providing highly refined binding site conformations for the docked compounds.[13] Similar outcomes were observed by Holcomb *et* al., who redocked ligands in the PDBbind data sets against the experimental cocrystallized receptor structures and against the AF2 structures using AutoDock.[14] The difference in docking success rates was substantial (41% for redocking to X-ray structures versus 17% to AF2 models) and was not predicted by the overall quality of the models. Removing low-confidence regions of the models and making side chains flexible improved the docking results.[14] Holcomb *et* al. also explored docking to apo structures, thus a more realistic situation. In contrast to the results obtained for holo structures, success rates for docking against AF2 predictions were similar and even slightly better than those for docking against apo structures. Based on these results, they concluded that AF2-generated models, to some degree, tend toward holo rather than apo structures and suggested the use of AF2 models alongside apo structures if only the latter were available. An alternative and even more powerful approach, suggested by Zhang et al.,[15] is cominimizing a known ligand with the AF2 model prior to virtual screening in order to move the model toward the ligand-bound conformation.[15] Karelina et al.[16] found that although AF2 models capture binding pocket structures with errors nearly as small as differences between structures of the same protein determined experimentally with different ligands bound, the accuracy of ligand binding poses predicted by docking to AF2 models was much lower than by docking to X-ray structures determined without these ligands bound. These conclusions have been confirmed by the yet unpublished systematic evaluation of the ligand docking experiments at CASP15, revealing that the AF2-generated models, in spite of their overall high accuracy, have local deformations of the binding sites that make accurate placing of ligands by direct docking difficult. In fact, all best-performing groups used template-based docking approaches, thus accounting for information from X-ray structures, in many cases cocrystallized with ligands.

In this paper, we focus on a problem related to docking, namely, the identification of binding hot spots and ligand binding sites. Binding hot spots are regions on the protein surface that significantly contribute to the binding free energy of the protein−ligand interaction,[17−20] and the strength of the hot spots reveals whether there exist potentially druggable sites that are capable of binding ligands with sufficient affinity.[21,22] The knowledge of the binding site is generally required by docking methods that target a selected region frequently called the "docking box".[23,24] The method we use for binding hot spot identification is protein mapping, which is essentially docking a variety of fragment-sized organic molecules.[25,26] The docking is global; thus, the entire protein surface is explored. As previously shown, the binding of fragments is much less sensitive to the geometry of the binding pockets and thus the conformations of the surrounding side chains than docking larger compounds.[27,28] In fact, it is well-known that the hit rates are higher in screening fragment libraries than in direct high-throughput screening of larger ligands.[29] Since fragment binding yields information on the nature of binding sites, including their druggability, and fragments can frequently be extended to larger and higher affinity ligands, our results suggest that the AF2-generated models must be suitable for computational approaches to fragment-based drug discovery. However, we also show that obtaining good results may require generating separate AF2 models for the ligand binding domains, preferably using forced sampling with multiple initial seeds, rather than simply mapping the models downloaded from the AlphaFold Protein Structure Database.[30]

MATERIALS AND METHODS

Benchmark Set. We have previously developed a benchmark set of 62 proteins, listed in Table 1, for testing computational methods for the identification of binding hot spots with an emphasis on fragment-based ligand discovery.[31] Each protein structure in the set has been cocrystallized with a fragment-sized ligand having a molecular weight (MW) under 200 g/mol and with one or more ligands with MW > 250 g/mol in other structures of the protein without substantial change in the binding mode of the smallest ligand as the substructure. In the remainder of this paper, we focus on the binding site of this small ligand, and the latter will be simply referred to as the ligand in order to discriminate it from the fragments used in the mapping process. We note that the targeted binding site in most of the proteins is formed by a single domain. In addition to the benchmark set of ligand-bound structures, we also constructed a benchmark set that included the protein's unliganded structures in the 47 cases when such structures were available. In some cases, such unbound structures may differ from the bound ones in a few residues. Both PDB IDs are shown for each protein in Table 1. More detailed information on the bound and unbound proteins with protein names and references are given in Tables S1 and S2 in the Supporting Information.

Retrieval and Preparation of AlphaFold2 Models. The initial set of AF2 models was downloaded from the AF Protein Structure Database (https://alphafold.ebi.ac.uk/)[30] using the UniProt ID provided for the proteins in the benchmark set mentioned above. The complete sequence from the UniProt ID

https://doi.org/10.1021/acs.jcim.3c01761
J. Chem. Inf. Model. 2024, 64, 960−973

Table 2. Pairwise t-Tests for RMSDs from Alignments of AlphaFold Models with Ligand-Bound and Unbound X-ray Structures

global all-atom: AF to X-ray			local all-atom: AF to X-ray		
(N = 47)	bound X-ray	unbound X-ray	(N = 47)	bound X-ray	unbound X-ray
RMSD mean (Å)	1.8 ± 1.6	2.1 ± 1.7	RMSD mean (Å)	1.1 ± 1.0	1.3 ± 1.0
P value	0.02909		P-value	0.1726	
global backbone: AF to X-ray			local backbone: AF to X-ray		
(N = 47)	bound X-ray	unbound X-ray	(N = 47)	bound X-ray	unbound X-ray
RMSD mean (Å)	1.4 ± 1.8	1.8 ± 1.9	RMSD mean (Å)	0.6 ± 0.8	0.8 ± 1.1
P value	0.04770		P value	0.1086	

was used to generate AF2 models for the few proteins that were not present in the AF database. Sequence alignment was performed for all of the AF2 models and their respective bound crystal structures in PyMOL. The AF models were then truncated to the residues appearing in the ligand-bound crystal structures. The same protocol was implemented to create truncated AF models with respect to unbound crystal structures. Following the truncation of the AF2 models, PyMOL was used to perform sequence-dependent alignments on the following aligned pairs: AF2 model−ligand-bound crystal structure, and AF2 model−unbound crystal structure. Both all-atom and backbone RMSD values were calculated from these alignments. Global RMSD calculations were based on all aligned residue pairs. Local RMSD calculations were restricted to binding site residues, defined as residues on the bound crystal structure that are within a 5 Å radius around the ligand. These same residues are considered to be the binding site residues for the AF2 models.

As will be discussed, a number of models downloaded from the AF Protein Structure Database were constructed for large multidomain proteins and had substantial local errors around the ligand binding site. We considered the proteins with relatively poor mapping results and generated AF2 models for the separate ligand-binding domains using forced sampling by repeated stochastic initialization of the multiple sequence alignment (MSA). For each protein, 100 initial seeds were employed for each of the five AlphaFold parameter sets, resulting in a total of 500 structural models. Neither structural templates nor model refinement were used. Models with the highest confidence according to the predicted local distance difference test (pLDDT) scores were selected for mapping. These "new" models are indicated by the model number followed by _new, and are compared to both the bound and unbound X-ray structures as well as to the "old" AF models downloaded from the AF database and truncated to the region present in the X-ray structures (see Table 4).

Characterization of Binding Hot Spots using FTMap. The binding properties of the proteins in the benchmark sets of X-ray structures and in the AF2-generated models were explored using the FTMap program.[20] FTMap uses a diverse set of 16 small molecular probes with different sizes and polarities to locate binding hot spots on a protein surface. Based on fast Fourier transforms, the algorithm places tens of thousands of copies of each probe throughout the entirety of the protein surface based on favorable energetics regarding probe position and orientation.[20] Clusters of probes are formed within a similar location; then, the clusters are ranked by their average energy, and the lowest energy ones are retained. The second round of clustering is implemented for the low-energy clusters, generating consensus clusters, also called consensus sites. Consensus sites mark the locations of binding hot spots and are ranked based on the number of probe clusters. The strength and importance of a hot spot are reflected in the ranking of a consensus site. Sites that include at least 16 probe clusters have the potential to bind appropriate ligands with low micromolar affinity, whereas high micromolar or millimolar binding requires at least 13 probe clusters in the consensus site.

To locate the ligand binding site on the AF models, the ligand on the bound crystal structure was aligned in space with the truncated AF models. If any atom in an FTMap-derived consensus site is within 2 Å of any atom in the previously aligned ligand, such a consensus site is called a "hit". To quantify the strength of the hits, we calculated the overlaps of the ligand with the hot spot and of the hot spot with the ligand. The overlap percentage of the ligand by a hot spot is calculated using $O_L = (N_L/ N_{LT}) \times 100\%$, where N_{LT} is the total number of atoms in the ligand and N_L is the number of the ligand atoms within 2 Å from any atom of any probe in the hot spot. Conversely, the overlap percentage of a hot spot by the ligand was calculated using $O_{HS} = (N_{HS}/N_{HST}) \times 100\%$, where N_{HST} is the total number of atoms of all probes in the hot spot and N_{HS} is the number of probe atoms that are within 2 Å from any atom of the ligand. In all calculations, only non-hydrogen atoms were considered. We also used a derivative of FTMap called FTMove for the identification of consensus binding sites.[32] FTMove implements FTMap in a high-throughput manner for all proteins in the PDB with at least 90% sequence identity with an input query protein. By characterizing hot spots throughout all highly similar structures, FTMove can cluster similar hot spots across all structures into concatenated larger hot spots referred to as binding sites. This approach to computational solvent mapping overcomes the static nature of hot spots in individual structures by the clustering of common hot spots among many conformers of a single protein. FTMove provides all identified binding sites and FTMap results for all protein conformers as PyMOL sessions and individual structure files.[32]

RESULTS AND DISCUSSION

AF2 Models and Mapping Results. Columns 2 through 5 of Table 1 list the UniProt IDs of the 62 proteins in the benchmark set, the 3-character IDs of the bound small ligands, the PDB IDs of the ligand-bound structures, and the PDB IDs of the unliganded structures. Unliganded structures were found for only 47 of the proteins. As mentioned, the models were downloaded from the AF Protein Structure Database using the UniProt IDs. The models also include confidence scores for each residue. Figure S1 shows the percentages of such scores for each of the 62 proteins in the benchmark set, revealing high confidence (over 90%) for large fractions in most proteins. There are a few exceptions, primarily targets 43 and 21 that will be discussed later in the paper.

https://doi.org/10.1021/acs.jcim.3c01761
J. Chem. Inf. Model. 2024, 64, 960−973

Table 3. Percentages of Proteins That Have Any Hot Spots or the Top Hot Spot with 13+ or 16+ Probe Clusters and at Least 50% or 80% Coverage of the Ligand Binding Site in the AlphaFold Models and X-ray Structures

			any hot spot, %			top hot spot, %		
model	type	N	13+, 50%	13+, 80%	16+, 50%	13+, 50%	13+, 80%	16+, 50%
AlphaFold		62	71.0	58.1	58.1	50.0%	35.5	46.8
X-ray structure	bound	62	77.4	69.3	70.9	56.5	50	56.5
	unbound	47	77.1	62.5	62.5	56.3	43.7	56.3

Table 4. RMSDs for Group 1 Member Entry 20, Group 2 Member Entry 12, Group 3 Member Entry 21, Truncated AlphaFold Model for Entry 21, and Truncated AlphaFold Model for Entry 43

				global alignment RMSD (Å)		local alignment RMSD (Å)	
no.	UniProt ID	PDB ID	kind	all-atom	backbone	all-atom	backbone
20	P28720	1S39_A	bound	0.833	0.364	1.011	0.257
		4Q8M_A	unbound	0.844	0.565	0.828	0.624
12	P08709	5PAW_B	bound	1.246	0.918	1.065	0.624
		1JBU_H	unbound	4.360	3.831	4.970	3.919
21	P00734	3P70_H	bound	2.995	2.454	4.658	3.051
		2UUF_B	unbound	2.890	2.398	4.732	3.111
21_new	P00734	3P70_H	bound	1.180	0.821	0.858	0.255
		2UUF_B	unbound	0.912	0.395	0.767	0.267
			old AF model	3.706	2.377	4.802	2.920
43_new	Q9H2K2	4PNN_B	bound	1.361	0.551	0.971	0.139
		4PNT_D	unbound	1.357	0.671	1.135	0.296
			old AF model	7.332	6.999	7.693	6.070

The models were truncated to have the same residues as in the PDB structures. Columns 6 and 7 of Table 1 show global (all-atom) RMSD values between the truncated AF2 model and the ligand-bound and -unliganded structures, respectively. Columns 8 and 9 show the same RMSD values restricted to the binding site residues. Tables S3 and S4 also include global and local RMSD values for the backbone atoms only. Table 2 shows both global and local average RMSD values between the (truncated) AF2 models and either ligand-bound or unliganded X-ray structures. One interesting question is whether the models are closer to bound or unbound structures. As shown, the global RMSD values are slightly lower for the bound structures, but based on pairwise t tests, the difference is not significant at $p = 0.01$, whereas the local RMSD values do not differ even at $p = 0.05$. Thus, the RMSD values do not provide significant insight into characteristics of AF2 models other than that they are within a good range (RMSDs ~ 1 Å) of resembling the X-ray structures, though it is not clear whether there is a bias toward bound or unbound crystal structure.

The truncated AF2 models were mapped with FTMap and studied for hot spot recovery. Columns 10, 11, and 12 of Table 1 show the strongest hot spot overlapping with the ligand binding site for the AF2 models and the ligand-bound and the unliganded structures, respectively. A consensus site was considered to overlap with a ligand if any atom of any probe in the consensus site was located within 2 Å of any atom of the ligand. Each hot spot is described by its rank starting with 00 as the strongest hot spot with the maximum number of probe clusters placed by FTMap, and in parentheses the number of probe clusters. Detailed mapping results are shown in Table S5. For each entry in the benchmark set, our analysis returned three lines of results, capturing the number and strength of the hot spots (consensus sites) identified by FTMap, the percentage of the ligand covered by probes, and the inverse, i.e., the percentage of the probes covered by the ligand. We further explain the content of this table by discussing a number of examples.

Table 3 is a summary of mapping results for the AF2 models and the fragment bound and unbound X-ray structures. As previously shown, a hot spot with 13 or more probe clusters predicts a site capable of ligand binding, whereas a hot spot with 16 or more clusters is predicted to be druggable.[22] Therefore, in Table 3, we list percentages of proteins that have been found to have hot spots with 13 or more probe clusters and at least 50% or 80% coverage as well as the percentage of proteins in which a hot spot with 16 or more probe clusters covers at least 50% of the ligand binding site. We first show the percentage of proteins that have any hot spot with these properties and then the percentage of proteins in which the strongest hot spot 00 satisfies these conditions. Considering any hot spot with 50% coverage and 13+ probe clusters, FTMap success rates for the models are about 5% lower than for either bound or unbound X-ray structures. Requiring 80% coverage or 16+ probe clusters, which is the condition for druggability,[33] the difference is still about 5% from the unbound structures, but becomes about 10% from the ligand-bound structures, in agreement with the previously reported binding results.[14] Restricting consideration to the top hot spot (identified as hot spot 00) reduces all success rates as expected, but the models perform worse. For both 13+ probe clusters and at least 80% coverage and for 16+ probe clusters and at least 50% coverage, the success rates of mapping the AF2 models are 10% to 15% lower than mapping either the bound or unbound X-ray structures.

Grouping of Proteins with Similar Mapping Results. To understand whether there is a relationship between hot spot recovery and structural quality, i.e., confidence metrics and RMSDs, we grouped the proteins into 3 categories based on the strongest hot spots, defined as the highest-ranking hot spot with at least 50% overlap with the bound ligand. In Group 1 (16 proteins, among them 12 with both bound and unbound

https://doi.org/10.1021/acs.jcim.3c01761
J. Chem. Inf. Model. 2024, 64, 960−973

Table 5. Detailed Mapping Results for Entry 20, AlphaFold (AF) Model with UniProt ID P28720 and Ligand-Bound X-ray Structure with PDB ID 1S39[a]

mapping results for AF model with UniProt ID P08709												
AQO_P28720	1S39_A	hs_lig	00(22)	01(18)	02(15)	03(14)	04(11)	05(08)	06(06)			
AQO_P28720	1S39_A	hs_lig	33%	100%								
AQO_P28720	1S39_A	hs_lig	10%	82%								
mapping results for ligand-bound X-ray structure with PDB ID 5PAW												
AQO_P28720	1S39_A	map	00(23)	01(16)	02(12)	03(12)	04(10)	05(07)	06(06)	07(05)	08(02)	09(02)
AQO_P28720	1S39_A	lig_hs	17%						100%			
AQO_P28720	1S39_A	hs_lig	10%						84%			

[a]Column 1: Ligand PDB ID_UniProt ID. Column 2: "PDB ID_chain" of ligand-bound X-ray structure. Column 3: "map" refers to mapping results with the 10-highest ranked hot spots and the number of probe clusters present in the hot spot (indicated in parentheses) for the consecutive corresponding columns. lig_hs - percentage of ligand covered by the hot spot; hs_lig—percentage of hot spot covered by the ligand. Blank columns indicate no hot spot with overlap detected in that ranking.

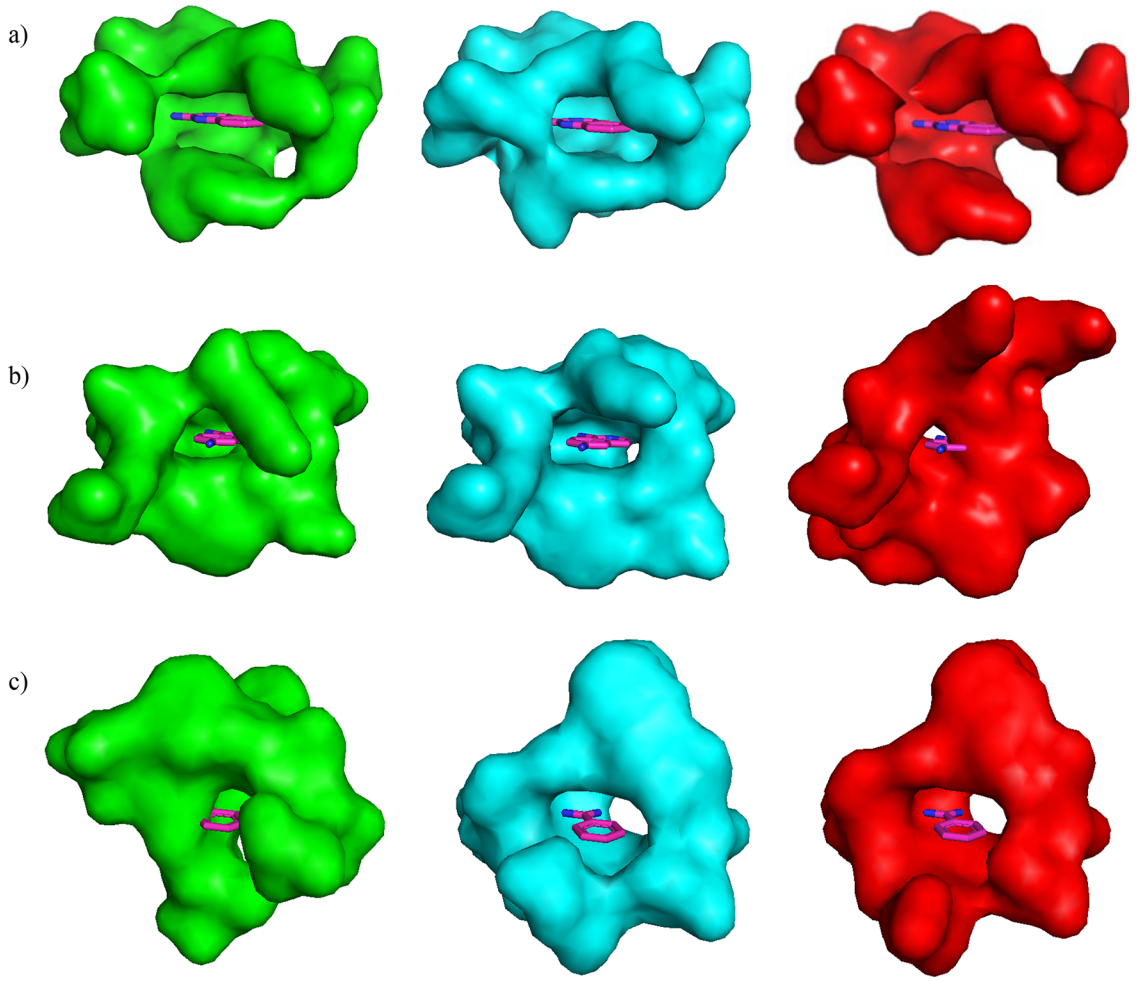


Figure 1. (a) Binding sites in surface representation for Entry 20 of the AlphaFold model P28720 (green), bound structure 1S39 (cyan), and unbound structure 4Q8M (red), with the ligand AQO inside each site, shown in sticks representation. (b) Binding sites in surface representation for Entry 12 of the AlphaFold model P08709 (green), bound structure 5PAW (cyan), and unbound structure 1JBU (red), with the ligand 7XM inside each site, shown in sticks representation. (c) Binding sites in surface representation for Entry 21 of the AlphaFold model P00734 (green), bound structure 3P70 (cyan), and unbound structure 2UUF (red), with the ligand BEN inside each site, shown in sticks representation.

structures), the strongest hot spot of the AF2 model ranks higher than that of the ligand-bound crystal structure. In Group 2 (27 proteins, 22 with unbound structures), the strongest hot spot of the AF2 model ranks about the same as that of the ligand-bound crystal structure. Finally, in Group 3 (19 proteins, 13 with unbound structures), the strongest hot spot of the AF2 model ranks lower than that of the ligand-bound crystal structure (or no hot spot has been detected at all).

Figures S2−S4, respectively, show the distributions of confidence scores for the models in Groups 1, 2, and 3. In all groups, the confidence levels are high for most proteins, with two exceptions in Group 2, and one strong exception in Group 3 (target 43). However, as will be discussed, the differences in

https://doi.org/10.1021/acs.jcim.3c01761
J. Chem. Inf. Model. 2024, 64, 960−973

Table 6. Detailed Mapping Results for Entry 12, AlphaFold Model with UniProt ID P08709 and Ligand-Bound X-ray Structure with PDB ID 5PAW[a]

mapping results for AF model with UniProt ID P08709												
7XM_P08709	5PAW_B	map	00(15)	01(14)	02(12)	03(12)	04(10)	05(09)	06(08)	07(08)	08(03)	09(02)
7XM_P08709	5PAW_B	lig_hs	100%									
7XM_P08709	5PAW_B	hs_lig	87%									
mapping results for ligand-bound X-ray structure with PDB ID 5PAW												
7XM_P08709	5PAW_B	map	00(27)	01(14)	02(10)	03(09)	04(08)	05(06)	06(06)	07(05)	08(03)	09(02)
7XM_P08709	5PAW_B	lig_hs	100%						100%			42%
7XM_P08709	5PAW_B	hs_lig	95%						84%			83%

[a]Column 1: Ligand PDB ID_UniProt ID. Column 2: "PDB ID_chain" of ligand-bound X-ray structure. Column 3: "map" refers to mapping results with the 10-highest ranked hot spots and the number of probe clusters present in the hot spot (indicated in parentheses) for the consecutive corresponding columns. lig_hs - percentage of ligand covered by the hot spot; hs_lig—percentage of hot spot covered by the ligand. Blank columns indicate no hot spot with overlap detected in that ranking.

Table 7. Detailed Mapping Results for Entry 21, AlphaFold Model with UniProt ID P00734[a]

mapping results for the model of UniProt ID P00734 from the AF database												
BEN_P00734	3P70_H	map	00(18)	01(13)	02(11)	03(09)	04(08)	05(07)	06(07)	07(06)	08(06)	09(03)
BEN_P00734	3P70_H	lig_hs										
BEN_P00734	3P70_H	hs_lig										
mapping results for ligand-bound X-ray structure with PDB ID 3P70												
BEN_P00734	3P70_H	map	00(20)	01(19)	02(14)	03(12)	04(09)	05(07)	06(04)	07(03)	08(03)	09(02)
BEN_P00734	3P70_H	lig_hs	100%									
BEN_P00734	3P70_H	hs_lig	83%									
mapping results for the AF model of the ligand binding domain of UniProt ID P00734												
BEN_P00734	3P70_H	map	00(18)	01(16)	02(15)	03(15)	04(07)	05(05)	06(04)	07(04)	08(04)	09(04)
BEN_P00734	3P70_H	lig_hs	100%		33%				78%	67%		
BEN_P00734	3P70_H	hs_lig	89%		18%				65%	82%		
mapping results for ligand-bound X-ray structure with PDB ID 3P70												
BEN_P00734	3P70_H	map	00(20)	01(19)	02(14)	03(12)	04(09)	05(07)	06(04)	07(03)	08(03)	09(02)
BEN_P00734	3P70_H	lig_hs	100%									
BEN_P00734	3P70_H	hs_lig	83%									
mapping results for the AF model of the ligand binding domain of UniProt ID P00734												
BEN_P00734	3P70_H	map	00(18)	01(16)	02(15)	03(15)	04(07)	05(05)	06(04)	07(04)	08(04)	09(04)
BEN_P00734	3P70_H	lig_hs	100%		33%				78%	67%		
BEN_P00734	3P70_H	hs_lig	89%		18%				65%	82%		

[a]Column 1: Ligand PDB ID_UniProt ID. Column 2: "PDB ID chain" of ligand-bound X-ray structure. Column 3: "map" refers to mapping results with the 10-highest ranked hot spots and the number of probe clusters present in the hot spot (indicated in parentheses) for the consecutive corresponding columns. lig_hs - percentage of ligand covered by the hot spot; hs_lig—percentage of hot spot covered by the ligand. Blank columns indicate no hot spot with overlap detected in that ranking.

confidence scores do not affect either the RMSD values or the mapping results. Tables S6 and S7, respectively, show global and local RMSD values between the models and ligand-bound and unbound crystal structures for each of the three groups. In Group 1 both global and local values are small and almost the same for bound and unbound structures. While all values are larger for Group 2, the differences in global RMSDs are larger for the unbound than for the bound structures, and the differences between the two are significant at $p = 0.05$, albeit not at $p = 0.01$. This shows that the similarity of models to bound structures is sufficient for good mapping results. Finally, in Group 3, both global and local RMSDs are relatively high and very similar for bound and unbound structures. We also note that the local RMSD, which is most likely the prime determinant of mapping accuracy, monotonously grows as we go from Group 1 to Group 2 and to Group 3. In Tables S8 and S9, we also compare the RMSDs for the three groups using one-way ANOVA to see whether differences in hot spot conservation throughout Groups 1−3 could be coupled to RMSD data, but the analysis does not show significant differences.

Representative Examples. We discuss one or two structures, listed in Table 4, from each of the three groups to provide more insight. Target 20 with UniProt ID P28720 is a member of Group 1, where the strongest hot spot of the AF2 model overlapping with the fragment ranks higher than that of its corresponding bound crystal structure. As shown in Table 5, the hot spot in question for the AF2 model is hot spot 01 with 18 probe clusters, but in the bound crystal structure, it is the much weaker hot spot 06 with only 6 probe clusters. To investigate why the hot spot overlapping with the ligand is so weak in the crystal structure but not in the AF model, we looked at the confidence metrics for the binding site residues. The AF2 database reports high confidence for 79% of these residues and average confidence for 21% of them, producing the binding site as an overall well-predicted model. When we generated these binding sites for the AF model and crystal structures (Figure 1a) in PyMOL, however, it becomes clear as to why the strongest hot spot has climbed ranks. The AF2 model adopts an intermediate conformation between the bound and unbound crystal structures. Compared to the binding site in the bound crystal

https://doi.org/10.1021/acs.jcim.3c01761
J. Chem. Inf. Model. 2024, 64, 960−973

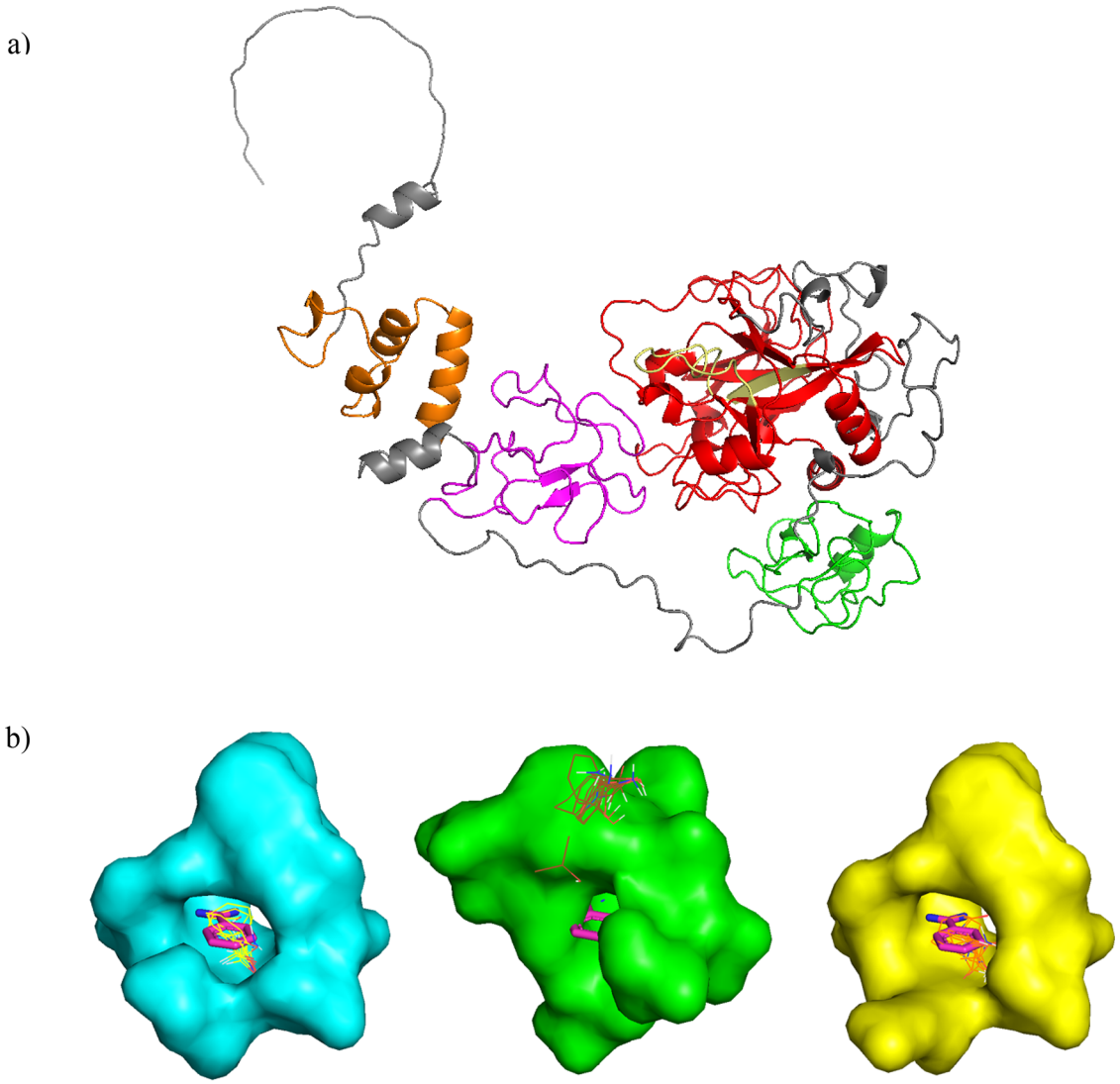


Figure 2. (a) Complete AlphaFold model provided by the AlphaFold database for the protein with UniProt ID P00734. Contains multiple domains denoted by specific colors: orange - Gla domain, magenta - Kringle 1 domain, green - Kringle 2 domain, red - peptidase S1 domain, yellow - high-affinity receptor binding region. (b) Binding sites of the X-ray structure 3P70_H (cyan), of the AF2 model downloaded from the AF database with UniProt ID P00734 (green), and of the AF2 model generated using only the sequence of the ligand-binding domain (yellow). Representative probes in the strongest hot spots predicted by FTMap are shown as wires in yellow, brown, and orange, respectively, for each model.

structure, the placement of residues in the AF2 model creates a wider, more solvent-accessible opening. We assume that this wider mouth allows for more probe molecules to enter the binding cavity in the AF2 model, thus creating a stronger hot spot.

Target 12 with UniProt ID P08709 is a member of Group 2, where the strongest hot spot of the AF model ranks the same as that of the bound crystal structure (Tables 4 and 6). We consulted the pLDDT scores of the binding site residues, which showed that 18% of the residues were predicted with high confidence, 29% with average confidence, and 53% with low confidence. The RMSDs from the alignments of the two structures are shown in Table 4, which reveal better agreement of the AF2 model with the ligand bound crystal structure and a large deviation from the unbound one. Figure 1b shows the binding sites for the AF2 model and the bound and the unbound crystal structures. Visually it is evident that the AF2 model binding site is an intermediate conformation between the two crystal structures. The opening of the AF model is not as closed off as in the unbound crystal structure, so a sufficient number of probes are able to enter the site and identify hot spots with close resemblance to that of the bound crystal structure. As shown in Table 6, the strongest hot spot overlaps well with the bound fragment in both the model and the bound crystal structure, although the one in the latter has more probe clusters. When considering the pLDDT scores, RMSDs, and mapping results altogether, it is surprising that such a combination of data produce a conserved hot spot. More than half of the binding site residues are predicted with low confidence, which is reflected in the slightly high all-atom RMSD values, yet mapping results indicate a conserved hot spot.

We selected Target 21 and another (Target 43) from Group 3 for detailed analysis. For these targets, the local RMSD is higher than the global in the AF2 model. Figure 1c shows the predicted binding site of the AF2 model for Target 21, which does not visually resemble the site in either bound or unbound crystal structures. The difference is not due to a low predicted confidence. Although 7% of binding site residues are predicted

https://doi.org/10.1021/acs.jcim.3c01761
J. Chem. Inf. Model. 2024, 64, 960−973

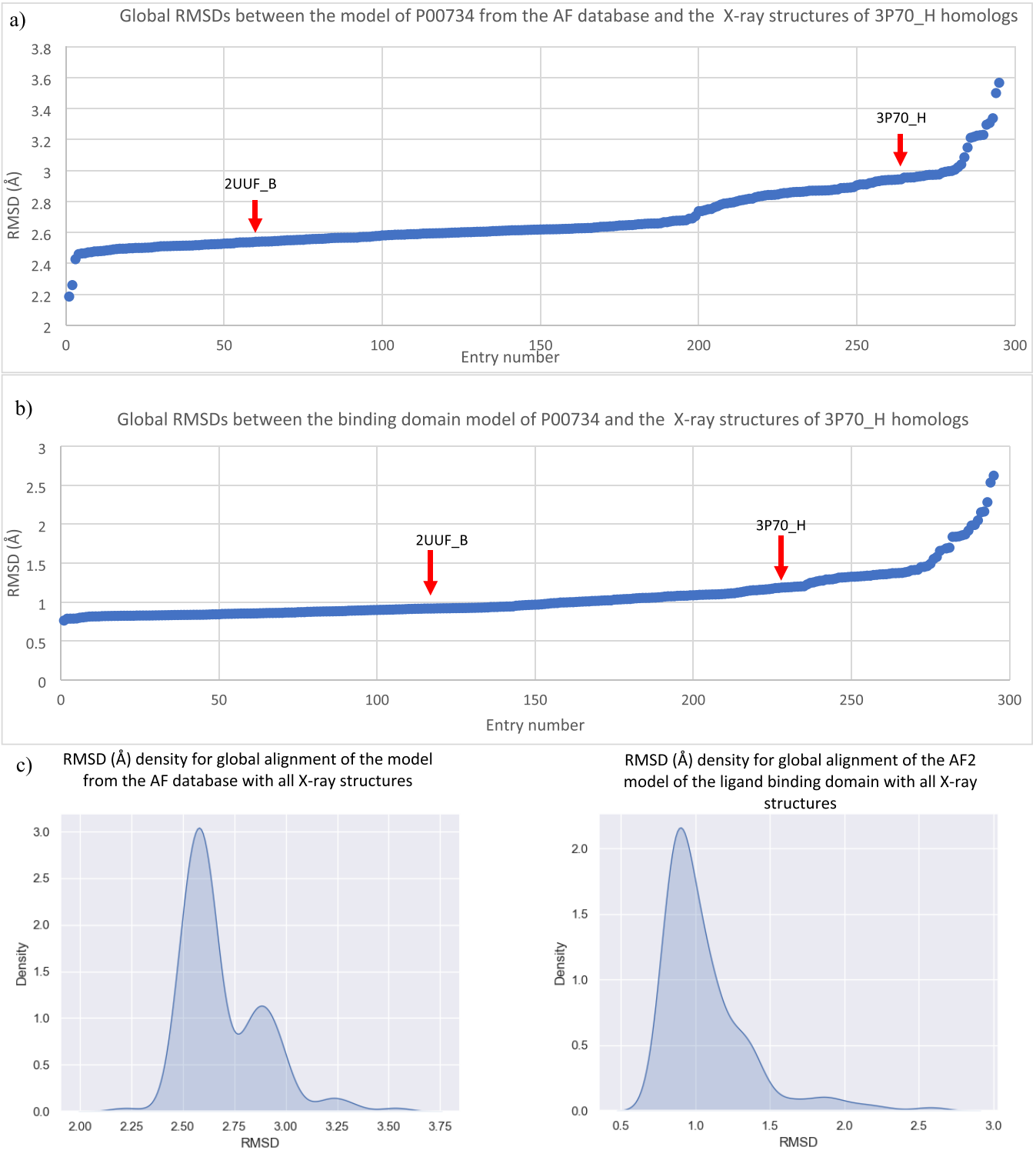


Figure 3. (a) All-atom RMSDs for the global alignment of the AF model P00734 from the AlphaFold database and the X-ray structures of 3P70_H with 90% sequence identity. 3P70_H is the PDB ID of the reference, ligand-bound X-ray structure, and 2UUF_B is the unbound X-ray structure. RMSDs of both structures are depicted in red arrows on the graph. (b) All-atom RMSDs for the global alignment of the AF2 model generated for the ligand binding domain of P00734 and the X-ray structures of 3P70_H with 90% sequence identity. 3P70_H is the PDB ID of the reference, ligand-bound X-ray structure and 2UUF_B is the unbound X-ray structure. RMSDs of both structures are depicted in red arrows on the graph. (c) Density plots for the all-atom RMSDs of the global alignment of the AF models P00734 with X-ray structures of 3P70_H with 90% sequence identity.

with high confidence, 71% are predicted with average confidence, and only 22% are predicted with low confidence, the binding site is distorted. The RMSD values (Table 4) support this argument. The local RMSD is higher than the global, which is not usually the case for the other proteins in the benchmark set. Mapping results in Table 7 indicate that the binding hot spot completely disappears in the AF2 model, even though there exists quite a strong hot spot in the bound crystal structure.

https://doi.org/10.1021/acs.jcim.3c01761
J. Chem. Inf. Model. 2024, 64, 960−973

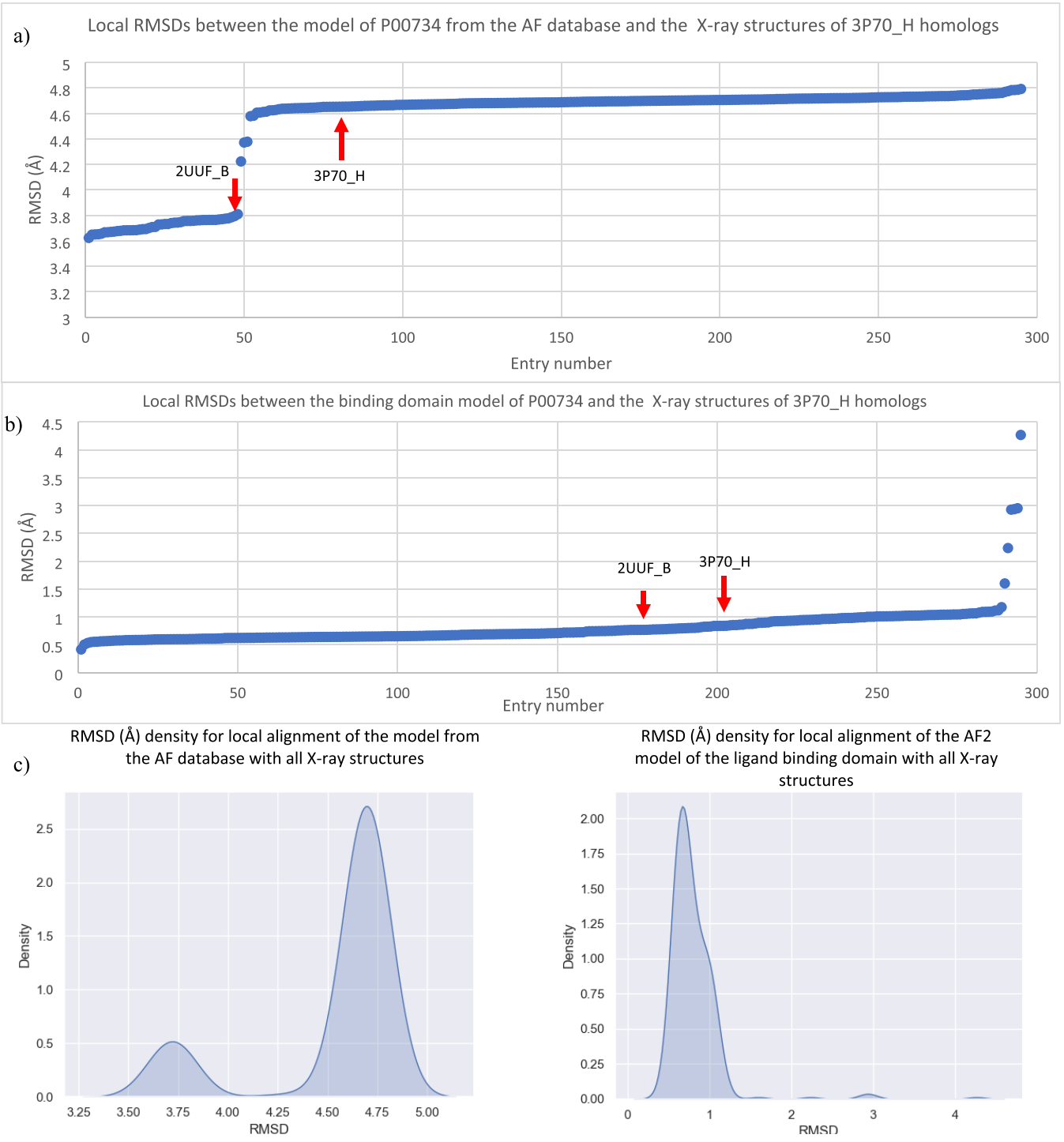


Figure 4. (a) All-atom RMSDs from the local alignment of the AF model P00734 from the AlphaFold database and the X-ray structures of 3P70_H with 90% sequence identity. 3P70_H is the PDB ID of the reference, ligand-bound X-ray structure, and 2UUF_B is the unbound X-ray structure. RMSDs of both structures are depicted in red arrows on the graph. (b) All-atom RMSDs from the local alignment of the AF2 model generated for the ligand binding domain of P00734 and the X-ray structures of 3P70_H with 90% sequence identity. 3P70_H is the PDB ID of the reference, ligand-bound X-ray structure and 2UUF_B is the unbound X-ray structure. RMSDs of both structures are depicted in red arrows on the graph. (c) Density plots for the all-atom RMSDs of the local alignment of the AF models of the P00734 with X-ray structures of 3P70_H with 90% sequence identity.

Mapping Models of Ligand Binding Domains. Targets 21 and 43 that have the most distorted binding site models also happen to be large multidomain proteins. The models in the AF2 database were determined for the entire sequences. As shown in Figure 2a, the model of Target 21 (UniProt ID P00734) downloaded from the database is of the entire human prothrombin (622 residues), whereas the crystal structure (3P70, chain H) is only of the S1 domain of the heavy chain of the human alpha-thrombin (259 residues) that binds the fragment of interest, a benzamidine molecule (BEN). Modeling this domain separately with AF2 yields much better agreement with the bound X-ray structure, reducing both global and local RMSD values (Table 4). Figure 2b shows the binding pockets in crystal structure 3P70_H (cyan), the site in the AF2 model

https://doi.org/10.1021/acs.jcim.3c01761
J. Chem. Inf. Model. 2024, 64, 960−973

Table 8. Percentages of Proteins That Have Any Hot Spots or the Top Hot Spot with 13+ or 16+ Probe Clusters and at Least 50% or 80% Coverage of the Fragment Binding Site in the AlphaFold Models, Including Models Generated with Multiple Random Seeds in Place of Original Group 3 Models and Crystal Structures

			any hot spot, %			top hot spot, %		
model	type	N	13+, 50%	13+, 80%	16+, 50%	13+, 50%	13+, 80%	16+, 50%
AlphaFold		62	79.0	66.1	66.1	58.1	46.8	54.8
X-ray structure	bound	62	77.4	69.3	70.9	56.5	50	56.5
	unbound	47	77.1	62.5	62.5	56.3	43.7	56.3

downloaded from the AF2 database (green), and the site in the "new" AF2 model of the ligand binding S1 domain generated separately by AF2 from the corresponding segment of the sequence (yellow). It is clear that running AF2 on the separate ligand-binding domain provides a much better model of the binding site than the one in the model of the entire multidomain protein deposited in the AF2 database. This observation is supported by the mapping, which places the strongest hot spot 00(18) of the new model overlapping with the ligand binding site (Table 7), and the hot spot is almost the same strength as the hot spot 00(20) in the ligand-bound X-ray structure, both with 100% overlap of the ligand.

We investigated whether the nearly 300 other structures homologous to 3P70 also deviate from the AF2 models, either downloaded from the AF2 database or generated only for the ligand-binding domain.[32] As shown in Figure 3a , 3P70 is somewhat of an outlier, but many other structures also have RMSD values from the AF2 model downloaded from the AF database close to or over 3 Å. The figure also shows that the model is closer but still at about 2.5 Å from the unbound structure 2UUF. In contrast, Figure 3b shows that the model generated by AF2 for the separate ligand binding domain (the heavy chain of the human Å-thrombin S1 domain) has less than 1.2 Å global RMSD from the bound structure and less than 1.0 Å RMSD from the unbound one. As shown in Figure 3c, the RMSD values between the model from the database and the X-ray structures have a bimodal distribution, with even the smaller RMSD peak being around 2.6 Å. In contrast, the distribution of RMSDs from the new model of the separate domain has a single peak at around 1 Å. The improvement of the AF2 model is even more significant around the binding site as shown by the local RMSD values. Indeed, for the model from the database, most of the local RMSD values are as high as 4.6 Å (Figure 4a), whereas for the new model, most values are below 1 Å (Figure 4b). As shown in Figure 4c, the distribution of RMSD values changes shape and shifts to the left by about 3 Å.

To emphasize that modeling the ligand binding domain separately increases the accuracy of the binding site, we performed a similar analysis for Target 43 (human tankyrase 2, PDB ID 4PNN, UniProt ID P00734), another multidomain protein in Group 3. Results are presented in the Supporting Information. Similar to Target 21, Target 43 also has a higher local than global RMSD, indicating a poor model of the binding site in the AF2 model downloaded from the AF database (Table 4). The left panel in Figure S5 shows that the model has a very distorted pocket at the ligand binding site. As shown in Table S10, no hot spot of the model overlaps with the small ligand quinazolin-4(1H)-one (JPZ). In the X-ray structure 4PPN the strongest hot spot overlapping with the ligand is 02(12), which is not very strong, but hot spot 06(8) is also at an overlapping location, and the site has a well-defined pocket (Figure S5). The RMSDs for the AF2 model downloaded from the AF database and truncated to the domain of interest are around 7 Å for both global and local alignments (Table 4). In fact, the binding site residue 1138 is completely misplaced in this model. Overall, the binding site of the model is highly altered from the crystal structure, and the binding hot spot is not conserved. In the model from the AF2 database, 20% of residues are predicted with average confidence, 20% with low confidence, and 60% with very low confidence, already suggesting poor quality. Figure S6a shows the complete AF2 model provided by the AlphaFold database for the protein with UniProt ID Q9H2K2, poly [ADP-ribose] polymerase tankyrase-2. In contrast, the PDB structure 4PNN is only the catalytic domain of the protein, cocrystallized with the small ligand quinazolin-4(1*H*)-one. As for Target 21, we used AF2 to model the domain separately with the sequence from 4PNN. As shown in Figure S6b, the binding site in this new model is substantially improved with good similarity to the one in the bound crystal structure. We also explored the RMSDs of homologues of tankyrase-2 in the PDB from both the original model in the AF database (Figure S7a) and the model of the separate ligand binding domain (Figure S7b), as well as the distributions of these RMSD values (Figure S7c).

Generating and mapping improved AF2 models. As shown, using AF2 models generated for the ligand binding domains of two multidomain proteins rather than the models downloaded from the AF database substantially improved the accuracy of the mapping results. Since it was not clear whether generating models would also improve results for other targets, we decided to run AF2 for the truncated sequences of all targets in Group 3, i.e., for cases when the mapping results were less accurate for the model than for the ligand-bound crystal structure. For each protein, 500 models were generated using 100 random seed for each of the five AF parameter sets, and models with the highest confidence were selected according to the predicted local distance difference test (pLDDT) scores. Tables S11 and S12, respectively, show global and local RMSD values between the models generated for the truncated sequences in Group 3 and the ligand-bound and unbound crystal structures, while not changing the models for Groups 1 and 2. Comparison to the RMSD values based on models from the AF2 database for Group 3 (see Figures S4 and S5) reveals that generating separate models for the relevant regions invariably reduces both global and local RMSDs by about 0.1 to 0.3 Å. While these changes are small, Table S13 shows that the recalculation has a major impact on the mapping results. Indeed, the number of probe clusters located at the ligand binding sites increases for 13 of the 19 proteins in Group 3, in some cases substantially, remains the same for 2 proteins, and decreases for 4 (Figures 3 and 4).

Table 8 shows a summary of mapping results with the new models included in the analysis. Comparing these results to those for the models downloaded from the AF2 database listed in Table 3 reveals that modeling of the shorter sequences with

https://doi.org/10.1021/acs.jcim.3c01761
J. Chem. Inf. Model. 2024, 64, 960−973

improved sampling yields substantially better mapping results. The success rates of finding the ligand binding sites are close to those obtained by mapping the ligand-bound X-ray structures and in most cases are similar or even slightly better than those from mapping the unbound structures

CONCLUSIONS

We compared the ligand binding properties of AF2 models to those of X-ray crystal structures. The focus was the conservation of binding hot spots, which are regions of the protein surface with a large contribution to the free energy of ligand binding and locate the potential binding sites. The hot spots can be detected as clusters of small molecular probes globally docked to the protein. The binding of the fragment-sized probes does not require the steric complementarity seen in ligand−receptor interactions, and hence, mapping is less sensitive to local conformational changes than docking. In agreement with this expectation, our study shows better results than those reported for docking to AF2 models. Nevertheless, the success rates were still lower than for mapping unliganded or ligand-bound X-ray structures, respectively, by about 5% and 10%. In particular, a large drop in quality was seen for large multidomain models directly downloaded from the AF2 database. Further analysis revealed that the binding cavity in some of these models was substantially distorted. However, both the accuracy of the binding sites and the quality of mapping results were substantially improved by building AF2 models only for the ligand-binding domains. In addition, using the multiseed approach in the AF2 calculations improved the results for most proteins with relatively poor conservation of the binding sites, bringing the success rates of mapping the models very close to those of the X-ray structures. Since we studied only a few multidomain proteins, we do not know whether the binding sites are generally distorted in the models of such proteins in the AF2 database. However, it is clear that modeling only the ligand-binding domains, particularly when using forced sampling, improves the accuracy of the ligand-binding sites, and therefore, the approach is preferable to simply downloading the models from the AF2 database.

ASSOCIATED CONTENT

Data Availability Statement

All PDB and UniProt accession codes for the structures and AlphaFold models studied in this work are provided within the published article and its Supporting Information. AlphaFold models were downloaded from the AF Protein Structure Database at https://alphafold.ebi.ac.uk/. The AlphaFold open-source code can be accessed from https://github.com/deepmind/alphafold. Models created for this study can be found in the following repository: 10.5281/zenodo.10064299. Specifications for how and why each model was created can be found within the published article. The list in the attached Excel file Data S1.xlsx provides the names and explanations of the AlphaFold models in the repository. The FTMap server is available to use free of charge for academic and governmental purposes at https://ftmap.bu.edu.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.3c01761.

Table S1: names and references of proteins studied with corresponding UniProt IDs and PDB IDs of bound X-ray structures; Table S2: names and references of proteins studied with corresponding UniProt IDs and PDB IDs of unbound X-ray structures; Table S3: RMSDs (Å) for the global all-atom (AA) and backbone (BB) alignments of the AlphaFold model to the bound X-ray structure and the AlphaFold model to the unbound X-ray structure; Table S4: RMSDs (Å) for the local all-atom (AA) and backbone (BB) alignments of the AlphaFold model to the bound X-ray structure and the AlphaFold model to the unbound X-ray crystal structure; Figure S1: histogram of confidence metrics for binding site residues in AlphaFold models; Figure S2: histogram of confidence metrics for binding site residues in Group 1 AlphaFold models; Figure S3: histogram of confidence metrics for binding site residues in Group 2 AlphaFold models; Figure S4: histogram of confidence metrics for binding site residues in Group 3 AlphaFold models; Table S5: detailed mapping results for all AlphaFold models in Acpharis benchmark set; Table S6: pairwise *t* tests for RMSDs from global alignment of AlphaFold models with ligand-bound and unbound X-ray structures − Group studies; Table S7: pairwise *t* tests for RMSDs from local alignment of AlphaFold models with ligand-bound and unbound X-ray structures − Group studies; Table S8: one-way ANOVA for RMSDs from global alignments of AlphaFold models with ligand-bound and unbound X-ray structures − Group studies; Table S9: one-way ANOVA for RMSDs from local alignments of AlphaFold models with ligand-bound and unbound X-ray structures − Group studies; Table S10: detailed mapping results for entry 43, AlphaFold (AF) model with UniProt ID Q9H2K2 and ligand-bound X-ray structure with PDB ID 4PNN; Figure S5: binding sites in surface representation for entry 43; Figure S6: complete AlphaFold model for entry 43 with different domains colored; binding sites of entry 43 crystal structure and corresponding AlphaFold models; Figure S7: all-atom RMSDs for the global alignment of Entry 43 AlphaFold model and X-ray structures with 90% sequence identity to structure with PDB ID 4PNN, chain B; all-atom RMSDs for the global alignment of the ligand-binding domain of entry 43 AlphaFold model and X-ray structures with 90% sequence identity to structure with PDB ID 4PNN, chain B; density plots for the all-atom RMSDs of the global alignments of entry 43 AlphaFold model with X-ray structures of 4PNN, chain B with 90% sequence identity; Figure S8: all-atom RMSDs for the local alignment of entry 43 AlphaFold model and X-ray structures with 90% sequence identity to structure with PDB ID 4PNN, chain B; all-atom RMSDs for the local alignment of the ligand-binding domain of entry 43 AlphaFold model and X-ray structures with 90% sequence identity to structure with PDB ID 4PNN, chain B; density plots for the all-atom RMSDs of the local alignments of Entry 43 AlphaFold model with X-ray structures of 4PNN, chain B with 90% sequence identity;Table S11: pairwise *t* tests for RMSDs from global alignment of recalculated AlphaFold models with ligand-bound and unbound X-ray structures − Group studies; Table S12: pairwise *t* tests for RMSDs from local alignment of recalculated AlphaFold models with ligand-bound and unbound X-ray structures − Group studie; Table S13: ligand-bound structures, ligand IDs for the (NEW) AlphaFold models, and strongest hot spots at the ligand binding sites with all-atom RMSDs (Å) for the (NEW) AlphaFold models versus (OLD)

https://doi.org/10.1021/acs.jcim.3c01761
J. Chem. Inf. Model. 2024, 64, 960−973

AlphaFold models and bound and unbound X-ray structures (PDF)

Names and explanations of the AlphaFold models deposited (XLSX)

AUTHOR INFORMATION

Corresponding Author

Sandor Vajda − *Department of Chemistry and Department of Biomedical Engineering, Boston University, Boston, Massachusetts 02215, United States;* orcid.org/0000-0003-1540-8220; Email: vajda@bu.edu

Authors

Ayse A. Bekar-Cesaretli − *Department of Chemistry, Boston University, Boston, Massachusetts 02215, United States*

Omeir Khan − *Department of Chemistry, Boston University, Boston, Massachusetts 02215, United States;* orcid.org/0000-0002-5822-1549

Thu Nguyen − *Department of Computer Science, Stony Brook University, Stony Brook, New York 11794, United States*

Dima Kozakov − *Department of Applied Mathematics and Statistics and Laufer Center for Physical and Quantitative Biology, Stony Brook University, Stony Brook, New York 11794, United States*

Diane Joseph-Mccarthy − *Department of Biomedical Engineering, Boston University, Boston, Massachusetts 02215, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.3c01761

Notes

The authors declare no competing financial interest.

Crystal structures studied were accessed and downloaded from https://rcsb.org.

ACKNOWLEDGMENTS

This work was supported by grants R35GM118078, R01GM140098 and R01GM102864 from the National Institute of General Medical Sciences; and grants 2200052 and 2054251 from the National Science Foundation.

REFERENCES

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

(2) Skolnick, J.; Gao, M.; Zhou, H.; Singh, S. Alphafold 2: Why It Works and Its Implications for Understanding the Relationships of Protein Sequence, Structure, and Function. *J. Chem. Inf Model* **2021**, *61*, 4827−4831.

(3) Thornton, J. M.; Laskowski, R. A.; Borkakoti, N. Alphafold Heralds a Data-Driven Revolution in Biology and Medicine. *Nat. Med.* **2021**, *27*, 1666−1669.

(4) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Zidek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Applying and Improving Alphafold at Casp14. *Proteins* **2021**, *89*, 1711−1721.

(5) Elofsson, A. Progress at Protein Structure Prediction, as Seen in Casp15. *Curr. Opin Struct Biol.* **2023**, *80*, No. 102594.

(6) Johansson-Akhe, I.; Wallner, B. Improving Peptide-Protein Docking with Alphafold-Multimer Using Forced Sampling. *Front Bioinform* **2022**, *2*, No. 959160.

(7) Tsaban, T.; Varga, J. K.; Avraham, O.; Ben-Aharon, Z.; Khramushin, A.; Schueler-Furman, O. Harnessing Protein Folding Neural Networks for Peptide-Protein Docking. *Nat. Commun.* **2022**, *13*, 176.

(8) Ghani, U.; Desta, I.; Jindal, A.; Khan, O.; Jones, G.; Hashemi, N.; Kotelnikov, S.; Padhorny, D.; Vajda, S.; Kozakov, D. Improved Docking of Protein Models by a Combination of Alphafold2 and Cluspro. *BioRxiv* **2021**, 2021-09 DOI: .

(9) Evans, R.; O'Neill, M.; Pritzel, A.; Antropova, N.; Senior, A.; Green, T.; Žídek, A.; Bates, R.; Blackwell, S.; Yim, J. Protein Complex Prediction with Alphafold-Multimer. *BioRxiv* **2021**, 2021.10 DOI: .

(10) Yin, R.; Feng, B. Y.; Varshney, A.; Pierce, B. G. Benchmarking Alphafold for Protein Complex Modeling Reveals Accuracy Determinants. *Protein Sci.* **2022**, *31*, No. e4379.

(11) Mullard, A. What Does Alphafold Mean for Drug Discovery? *Nat. Rev. Drug Discov* **2021**, *20*, 725−727.

(12) Nussinov, R.; Zhang, M.; Liu, Y.; Jang, H. Alphafold, Allosteric, and Orthosteric Drug Discovery: Ways Forward. *Drug Discov Today* **2023**, *28*, No. 103551.

(13) Scardino, V.; Di Filippo, J. I.; Cavasotto, C. N. How Good Are Alphafold Models for Docking-Based Virtual Screening? *iScience* **2023**, *26*, No. 105920.

(14) Holcomb, M.; Chang, Y. T.; Goodsell, D. S.; Forli, S. Evaluation of Alphafold2 Structures as Docking Targets. *Protein Sci.* **2023**, *32*, No. e4530.

(15) Zhang, Y.; Vass, M.; Shi, D.; Abualrous, E.; Chambers, J. M.; Chopra, N.; Higgs, C.; Kasavajhala, K.; Li, H.; Nandekar, P.; Sato, H.; Miller, E. B.; Repasky, M. P.; Jerome, S. V. Benchmarking Refined and Unrefined Alphafold2 Structures for Hit Discovery. *J. Chem. Inf Model* **2023**, *63*, 1656−1667.

(16) Karelina, M.; Noh, J. J.; Dror, R. O. How Accurately Can One Predict Drug Binding Modes Using Alphafold Models? *bioRxiv* **2023**, 2023.05 DOI: .

(17) DeLano, W. L. Unraveling Hot Spots in Binding Interfaces: Progress and Challenges. *Curr. Opin Struct Biol.* **2002**, *12*, 14−20.

(18) Ciulli, A.; Williams, G.; Smith, A. G.; Blundell, T. L.; Abell, C. Probing Hot Spots at Protein-Ligand Binding Sites: A Fragment-Based Approach Using Biophysical Methods. *J. Med. Chem.* **2006**, *49*, 4992−5000.

(19) Lal Gupta, P.; Carlson, H. A. Cosolvent Simulations with Fragment-Bound Proteins Identify Hot Spots to Direct Lead Growth. *J. Chem. Theory Comput* **2022**, *18*, 3829−3844.

(20) Kozakov, D.; Grove, L. E.; Hall, D. R.; Bohnuud, T.; Mottarella, S. E.; Luo, L.; Xia, B.; Beglov, D.; Vajda, S. The FTMap Family of Web Servers for Determining and Characterizing Ligand-Binding Hot Spots of Proteins. *Nat. Protoc* **2015**, *10*, 733−755.

(21) Hajduk, P. J.; Huth, J. R.; Fesik, S. W. Druggability Indices for Protein Targets Derived from NMR-based Screening Data. *J. Med. Chem.* **2005**, *48*, 2518−2525.

(22) Kozakov, D.; Hall, D. R.; Napoleon, R. L.; Yueh, C.; Whitty, A.; Vajda, S. New Frontiers in Druggability. *J. Med. Chem.* **2015**, *58*, 9063−9088.

(23) Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. Automated Docking Using a Lamarckian Genetic Algorithm and an Empirical Binding Free Energy Function. *J. Comput. Chem.* **1998**, *19*, 1639−1662.

(24) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J. Med. Chem.* **2004**, *47*, 1739−1749.

(25) Dennis, S.; Kortvelyesi, T.; Vajda, S. Computational Mapping Identifies the Binding Sites of Organic Solvents on Proteins. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 4290−4295.
(26) Landon, M. R.; Lancia, D. R., Jr; Yu, J.; Thiel, S. C.; Vajda, S. Identification of Hot Spots within Druggable Binding Regions by Computational Solvent Mapping of Proteins. *J. Med. Chem.* **2007**, *50*, 1231−1240.
(27) Hall, D. R.; Kozakov, D.; Vajda, S. Analysis of Protein Binding Sites by Computational Solvent Mapping. *Methods Mol. Biol.* **2012**, *819*, 13−27.
(28) Hall, D. R.; Ngan, C. H.; Zerbe, B. S.; Kozakov, D.; Vajda, S. Hot Spot Analysis for Driving the Development of Hits into Leads in Fragment-Based Drug Discovery. *J. Chem. Inf Model* **2012**, *52*, 199−209.
(29) Erlanson, D. A. Introduction to Fragment-Based Drug Discovery. *Top Curr. Chem.* **2011**, *317*, 1−32.
(30) Varadi, M.; Anyango, S.; Deshpande, M.; Nair, S.; Natassia, C.; Yordanova, G.; Yuan, D.; Stroe, O.; Wood, G.; Laydon, A.; Zidek, A.; Green, T.; Tunyasuvunakool, K.; Petersen, S.; Jumper, J.; Clancy, E.; Green, R.; Vora, A.; Lutfi, M.; Figurnov, M.; Cowie, A.; Hobbs, N.; Kohli, P.; Kleywegt, G.; Birney, E.; Hassabis, D.; Velankar, S. Alphafold Protein Structure Database: Massively Expanding the Structural Coverage of Protein-Sequence Space with High-Accuracy Models. *Nucleic Acids Res.* **2022**, *50*, D439−D444.
(31) Wakefield, A. E.; Yueh, C.; Beglov, D.; Castilho, M. S.; Kozakov, D.; Keseru, G. M.; Whitty, A.; Vajda, S. Benchmark Sets for Binding Hot Spot Identification in Fragment-Based Ligand Discovery. *J. Chem. Inf Model* **2020**, *60*, 6612−6623.
(32) Egbert, M.; Jones, G.; Collins, M. R.; Kozakov, D.; Vajda, S. Ftmove: A Web Server for Detection and Analysis of Cryptic and Allosteric Binding Sites by Mapping Multiple Protein Structures. *J. Mol. Biol.* **2022**, *434*, No. 167587.
(33) Abbotts, R.; Madhusudan, S. Human AP Endonuclease 1 (APE1): From Mechanistic Insights to Druggable Target in Cancer. *Cancer Treat Rev.* **2010**, *36*, 425−435.

https://doi.org/10.1021/acs.jcim.3c01761
J. Chem. Inf. Model. 2024, 64, 960−973