JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Viewpoint

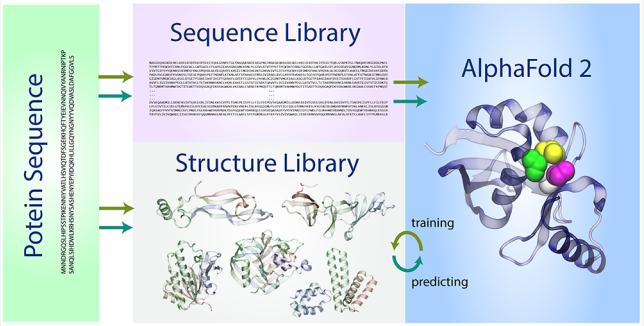
AlphaFold 2: Why It Works and Its Implications for Understanding the Relationships of Protein Sequence, Structure, and Function

Jeffrey Skolnick,* Mu Gao,§ Hongyi Zhou,§ and Suresh Singh

Cite This: J. Chem. Inf. Model. 2021, 61, 4827−4831 Read Online

ACCESS | Metrics & More | Article Recommendations

ABSTRACT: AlphaFold 2 (AF2) was the star of CASP14, the last biannual structure prediction experiment. Using novel deep learning, AF2 predicted the structures of many difficult protein targets at or near experimental resolution. Here, we present our perspective of why AF2 works and show that it is a very sophisticated fold recognition algorithm that exploits the completeness of the library of single domain PDB structures. It has also learned local side chain packing rearrangements that enable it to refine proteins to high resolution. The benefits and limitations of its ability to predict the structures of many more proteins at or close to atomic detail are discussed.



One criticism leveled against deep learning is that it is a black box approach. Input a protein sequence and by "magic" the protein's three-dimensional structure appears. Superficially, that is what the AF2[1] deep learning algorithm for protein structure prediction does. Actually, AF2 figures out the complex interrelationships of the protein's residues that dictate what structure that protein sequence adopts, much like template selection in traditional structure prediction.[2−7] It then iterates to improve the local structural details analogous to traditional refinement.[2−7] This Viewpoint seeks to understand the reasons behind AF2's success and its implications for biology.

The excellent blog of the Oxford Protein Informatics Group[8] describes on a technical level how AF2 works. AF2 starts by employing multiple sequence alignments (MSA) with different regions weighted by importance (Attention). It then uses the Evoformer module to extract information about interrelationships between protein sequences and template structures. The structure module treats the protein as a residue gas moved around by the network to generate the protein's 3D structure followed by local refinement to provide the final prediction.

Assessing the quality of predictions using the TM-score[9] (a structural similarity metric with values [0,1] whose random value is 0.3 and values greater than 0.4 indicate fold similarity), AF2 often produced models with a TM-score greater than 0.9. Going beyond a TM-score of 0.85 indicates that both the global fold and details are correct. In contrast, MSA-based methods generate an average fold and plateau around a TM-score of 0.85. The recent work of the Baker group[10] is an example of this average fold approach.

The key to why AF2[11] works is the fact the library of single domain protein structures is essentially complete.[2,12,13] In 2005,[2] we pointed out that if one had an algorithm that maps a protein's sequence to the correct template structure in the existing PDB[14] library (independent of their evolutionary relationship), the folding problem could be solved. This was the basis of the very early threading/fold recognition ideas.[15−17] Then, it was unclear whether the library of solved structures in the PDB[14] was or could be complete. However, by 2005, the library of single domain protein structures was shown to be essentially complete.[2,12,13] Why was not fold recognition solved in 2005? The problem was the inability to match evolutionarily distant or unrelated proteins to their closest template PDB structure.

Recently, deep learning[18−25] yielded significant improvements in structure prediction. First-generation approaches employed residue covariation as the key feature. The idea of residue covariation (or correlated mutations) that close in space residues in the folded protein covary during evolution[26−28] was introduced decades earlier into structure prediction.[29−31] However, the small sequence libraries used to calculate residue covariation gave uneven performances. With the exponential growth of sequence databases, residue covariation for protein structure prediction experienced a renaissance[32] but still had significant issues.

Received: September 11, 2021
Published: September 29, 2021

https://doi.org/10.1021/acs.jcim.1c01114
J. Chem. Inf. Model. 2021, 61, 4827−4831

Deep learning with residue covariation significantly improved structure prediction.[18,19,23−25,33] Why? The DESTINI convolutional neural network (CNN) deep learning approach to structure prediction[23] demonstrated that deep learning learned the contact map patterns of secondary structure packing and eliminated nonphysical, isolated residue pairs. Thus, it learned long-range (spatially close but far along the sequence), many body interactions among residues that were not captured in early attempts to learn protein-like side chain packing patterns.[34]

Additional insights are provided by the deep learning SAdLSA algorithm[35] which predicts structure alignments between protein sequences without their structures. SAdLSA trained on purely α-helical proteins can align the sequences of distantly related pure β-proteins. On examining self-alignments based on SAdLSA's residue distogram,[36] the closer the template to the target in the library of learned PDB structures is, the better the result. The reason that SAdLSA trained on helical proteins works for β-proteins is that the global topology of β-proteins is in the training set;[2,12,13] the chain contour of helical proteins provides a globally correct fold for β-proteins.[13] As discussed in ref 37, deep learning's ability to learn long-range correlations among residues enables it to map sequences to existing structures by exploiting PDB completeness for single domain proteins.

We turn now to the application of AF2 to predict the structures of the human protein exome.[38] Interestingly, the Pearson correlation coefficient of the average confidence score per protein with contact order is 0.28—a weak correlation of model quality with whether the fold has a sequential or nonsequential arrangement of secondary structural elements. This is different from template-free (TF) approaches.[39] TF is more successful when the fold consists of the sequential arrangement of secondary structural elements. In contrast, AF2 learned information about the relationship of a sequence to its fold regardless of its contact order.

To elucidate how AF2 works, we used TM-align[40] to generate structural alignments of all of its predicted human proteins to a 41,000 member PDB library from 2016. This library was chosen to show that even a half decade ago the existing PDB contains more than enough information to solve the fold recognition problem (at least for single domain proteins). The correlation coefficient of AF2's confidence score with TM-score is 0.67. Using AF2's criterion of a confidence index less than 70 as a nonconfident model, 70−90 a good model, and greater than 90 a highly accurate model, then for low quality predicted structures only 10% of proteins have a closest template in the template library with a TM-score $\geq$ 0.55. In contrast, 53% of proteins with good models and 86% with high accuracy models have a TM-score $\geq$ 0.55. These results are preserved when the current PDB library of almost 90,000 structures clustered at a sequence identity cutoff of 70% is used, with a shift toward higher TM-scores for the various prediction confidence levels. Since AF2 recapitulates the trends of the SAdLSA self-alignment study, albeit with better model accuracy, it appears to have solved the fold recognition problem.

Having selected an approximate fold, AF2 sometimes refines the structure to experimental accuracy, with TM-scores greater than 0.9. Its individual per residue networks learned local, residue specific packing details. Perhaps, the number of local residue packing environments is smaller than might be naively guessed, and thus, AF2's powerful deep learning tools also learned their rules.

Could AF2 predict the structures of multiple domain and multimeric proteins? We suspect this is likely. For some multiple domain CASP14 targets, AF2 made high quality structure predictions. In practice, the structural space of protein−protein interfaces is small (<1000 structurally distinct interfaces),[41,42] with nothing special about protein−protein interfaces.[43] AF2 uses a protein gas so it might identify different patterns of residue connectivity consistent with the same geometric arrangement of residues in space. Thus, for a pair of proteins (or domains), AF2 might find related structures in its training set that are structurally similar to the interfacial region's secondary structure arrangement. Again, this exploits the structural completeness of the PDB and AF2's ability to do sophisticated (possibly sequence order independent) fold recognition.

While AF2 generates significantly better structure predictions than existing methods on average, for the human exome, it is worthwhile to examine how many additional sequences have confident models compared to the prior art.[44] There are 26,996 consensus human exome sequences between AF2 and a previously predicted set using TASSER-VMT.[44] TASSER-VMT predicts that at least one domain has a TM-score greater than 0.4 (indicating global fold similarity to the native structure) for 89% of sequences. AF2 has a confidence score greater than 70 for 67.4% of sequences, while 86.9% have a confidence score greater than 60. Using the same domain partitioning as TASSER-VMT, AF2 provides an additional 8.7% of human sequences with a confidence score greater than 60. These structures probably have a TM-score to the native $\geq$ 0.40. The union of approaches provides confident models for at least one domain for 97.5% of human sequences. Combining AF2 with a more traditional approach may be a winning strategy to increase exome structure coverage.

How will the availability of higher resolution structures impact biology? First, structure alone does not provide the biochemical function of the protein nor can AF2 predict phenotypical function which is often the result of the complex interplay of proteins, RNA, DNA, and metabolites. However, the structure of a protein can be used to infer its biochemical function. One can also perform virtual ligand screening to predict which small molecules/drugs it might bind and how. Contrary to popular belief that high resolution structures are required, using ligand homology modeling,[45−48] low resolution structures are adequate, with at most a 20% improvement in enrichment factor on going from lower resolution (TM-score > 0.4) to crystal structures. The reason is simple; binding sites are generally the most accurately predicted parts of a protein's structure.

Are the models predicted by AF2 useful for traditional drug discovery where high resolution structures are required?[49] Since AF2 was trained on proteins having particular ions, for example, it might not provide an accurate representation of the binding pocket under other conditions. Using structures for traditional small molecule docking requires evaluation of, for example, the binding pocket shape and protonation states. These issues can be addressed by prepping the protein's structure, solvating the protein, and performing MD simulations perhaps including relevant ligands.[50−52]

Even though AF2 can accurately predict backbone and side chain conformations, it is for a particular conformational state. However, a protein can exist in one conformation in its

https://doi.org/10.1021/acs.jcim.1c01114
J. Chem. Inf. Model. 2021, 61, 4827−4831

"inactive" state and another in its "active" state.[53−56] If an active conformation is predicted by AF2, then using it to target the inactive state will not work. Can AF2 provide information about alternative conformations? AF2's scoring functions are knowledge based. Since more naïve knowledge-based potentials provided insights into aspects of the conformational space that proteins explore,[57−59] perhaps AF2's "trajectory" might generate relevant conformational states. The difference been rapidly converging folds and slower convergence might reflect frustration in choosing the "correct" fold among alternatives and could provide insights into the relevant conformational ensemble.

What are AF2's possible longer-term impacts beyond providing improved structural coverage of exomes?[60] Can AF2 predict the structural effects of missense mutations or splice variants? Can it identify residues in Chameleon sequences which change the global fold and predict the structural changes?[61] Can it predict internal backbone and side chain rearrangements giving rise to cryptic pockets[62] or allosteric transitions, for example, the DFG-in and DFG-out kinase conformations?[63] Clearly, much work needs to be done to establish the strengths and limitations of AF2.

Importantly, while knowledge of a protein's structure alone does not provide its biochemical, much less, phenotypical effects, structure can be a component.[64] The challenge is to develop tools that employ the more accurate protein structures to infer molecular function, ligand binding, and protein−protein interactions and then integrate this information with AI approaches to predict phenotype from genotype. How inclusion of higher quality structures improves the ability to predict drug side effects, efficacy[64] and the role of missense variations in non-Mendelian diseases depends on accurately characterizing the interplay of many molecular players. Clearly, there is dramatic progress, but more remains to be done before the interrelationships of protein sequence, structure, and function are fully understood.

Data and Software Availability. This Viewpoint discusses three software packages: Alpha Fold 2 (https://deepmind.com/research/open-source/alphafold), TASSER(https://zhanggroup.org/I-TASSER/download/), and Fr-TM-align (https://sites.gatech.edu/cssb/fr-tm-align/).

AUTHOR INFORMATION

Corresponding Author

Jeffrey Skolnick − *Center for the Study of Systems Biology, School of Biological Sciences, Georgia Institute of Technology, Atlanta, Georgia 30332, United States;* orcid.org/0000-0002-1877-4958; Email: skolnick@gatech.edu

Authors

Mu Gao − *Center for the Study of Systems Biology, School of Biological Sciences, Georgia Institute of Technology, Atlanta, Georgia 30332, United States*

Hongyi Zhou − *Center for the Study of Systems Biology, School of Biological Sciences, Georgia Institute of Technology, Atlanta, Georgia 30332, United States*

Suresh Singh − *Twilight Design, Kendall Park, New Jersey 08824, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.1c01114

Author Contributions

§M. Gao and H. Zhou contributed equally to this work.

Notes

The authors declare no competing financial interest.

REFERENCES

(1) 14th Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction. *Protein Structure Prediction Center*. https://predictioncenter.org/casp14/ (accessed September 2021).

(2) Zhang, Y.; Skolnick, J. The protein structure prediction problem could be solved using the current PDB library. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 1029−1034.

(3) Roy, A.; Kucukural, A.; Zhang, Y. I-TASSER: a unified platform for automated protein structure and function prediction. *Nat. Protoc.* **2010**, *5*, 725−738.

(4) Zhang, Y. I-TASSER server for protein 3D structure prediction. *BMC Bioinf.* **2008**, *9*, 40.

(5) Zhou, H.; Skolnick, J. Template-based protein structure modeling using TASSER(VMT.). *Proteins: Struct., Funct., Genet.* **2012**, *80*, 352−361.

(6) Yang, J.; Zhang, W.; He, B.; Walker, S. E.; Zhang, H.; Govindarajoo, B.; Virtanen, J.; Xue, Z.; Shen, H. B.; Zhang, Y. Template-based protein structure prediction in CASP11 and retrospect of I-TASSER in the last decade. *Proteins: Struct., Funct., Genet.* **2016**, *84*, 233−246.

(7) Anishchenko, I.; Baek, M.; Park, H.; Hiranuma, N.; Kim, D. E.; Dauparas, J.; Mansoor, S.; Humphreys, I. R.; Baker, D. Protein tertiary structure prediction and refinement using deep learning and Rosetta in CASP14. *Proteins: Struct., Funct., Genet.* **2021**, na.

(8) AF2 is here: what's behind the structure prediction miracle. *Oxford Protein Informatics Group*. https://www.blopig.com/blog/2021/07/alphafold-2-is-here-whats-behind-the-structure-prediction-miracle/ (accessed September 2021).

(9) Zhang, Y.; Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **2005**, *33*, 2302−2309.

(10) Baek, M.; DiMaio, F.; Anishchenko, I.; Dauparas, J.; Ovchinnikov, S.; Lee, G. R.; Wang, J.; Cong, Q.; Kinch, L. N.; Schaeffer, R. D.; Millan, C.; Park, H.; Adams, C.; Glassman, C. R.; DeGiovanni, A.; Pereira, J. H.; Rodrigues, A. V.; van Dijk, A. A.; Ebrecht, A. C.; Opperman, D. J.; Sagmeister, T.; Buhlheller, C.; Pavkov-Keller, T.; Rathinaswamy, M. K.; Dalwadi, U.; Yip, C. K.; Burke, J. E.; Garcia, K. C.; Grishin, N. V.; Adams, P. D.; Read, R. J.; Baker, D. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **2021**, *373*, 871−876.

(11) 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.

(12) Zhang, Y.; Hubner, I. A.; Arakaki, A. K.; Shakhnovich, E.; Skolnick, J. On the origin and highly likely completeness of single-domain protein structures. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103*, 2605−2610.

(13) Skolnick, J.; Zhou, H.; Brylinski, M. Further evidence for the likely completeness of the library of solved single domain protein structures. *J. Phys. Chem. B* **2012**, *116*, 6654−6664.

(14) Segura, J.; Rose, Y.; Westbrook, J.; Burley, S. K.; Duarte, J. M. RCSB Protein Data Bank 1D Tools and Services. *Bioinformatics* **2021**, *36*, 5526.

(15) Finkelstein, A. V.; Reva, B. A. A search for the most stable folds of protein chains. *Nature* **1991**, *351*, 497−9.

(16) Chothia, C.; Finkelstein, A. V. The classification and origins of protein folding patterns. *Annu. Rev. Biochem.* **1990**, *59*, 1007−1035.

(17) Bowie, J. U.; Luthy, R.; Eisenberg, D. A method to identify protein sequences that fold into a known three-dimensional structure. *Science* **1991**, *253*, 164−170.
(18) Xu, J. Distance-based protein folding powered by deep learning. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116*, 16856−16865.
(19) Liu, J.; Wu, T.; Guo, Z.; Hou, J.; Cheng, J. Improving protein tertiary structure prediction by deep learning and distance prediction in CASP14. *Proteins: Struct., Funct., Genet.* **2021**, na.
(20) Xu, J.; McPartlon, M.; Li, J. Improved protein structure prediction by deep learning irrespective of co-evolution information. *Nat. Mach Intell* **2021**, *3*, 601−609.
(21) Wu, T.; Guo, Z.; Hou, J.; Cheng, J. DeepDist: real-value inter-residue distance prediction with deep residual convolutional network. *BMC Bioinf.* **2021**, *22*, 30.
(22) Senior, A. W.; Evans, R.; Jumper, J.; Kirkpatrick, J.; Sifre, L.; Green, T.; Qin, C.; Zidek, A.; Nelson, A. W. R.; Bridgland, A.; Penedones, H.; Petersen, S.; Simonyan, K.; Crossan, S.; Kohli, P.; Jones, D. T.; Silver, D.; Kavukcuoglu, K.; Hassabis, D. Improved protein structure prediction using potentials from deep learning. *Nature* **2020**, *577*, 706−710.
(23) Gao, M.; Zhou, H.; Skolnick, J. DESTINI: A deep-learning approach to contact-driven protein structure prediction. *Sci. Rep.* **2019**, *9*, 3514.
(24) Yang, J.; Anishchenko, I.; Park, H.; Peng, Z.; Ovchinnikov, S.; Baker, D. Improved protein structure prediction using predicted interresidue orientations. *Proc. Natl. Acad. Sci. U. S. A.* **2020**, *117*, 1496−1503.
(25) Li, Y.; Hu, J.; Zhang, C.; Yu, D.-J.; Zhang, Y. ResPRE: high-accuracy protein contact prediction by coupling precision matrix with deep residual neural networks. *Bioinformatics* **2019**, *35*, 4647−4655.
(26) Sander, C.; Schneider, R. Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins: Struct., Funct., Genet.* **1991**, *9*, 56−68.
(27) Gobel, U.; Sander, C.; Schneider, R.; Valencia, A. Correlated mutations and residue contacts in proteins. *Proteins: Struct., Funct., Genet.* **1994**, *18*, 309−317.
(28) Mumenthaler, C.; Braun, W. Predicting the helix packing of globular proteins by self-correcting distance geometry. *Protein Sci.* **1995**, *4*, 863−871.
(29) Olmea, O.; Valencia, A. Improving contact predictions by the combination of correlated mutations and other sources of sequence information. *Folding Des.* **1997**, *2*, S25−S32.
(30) Ortiz, A. R.; Kolinski, A.; Skolnick, J. Nativelike topology assembly of small proteins using predicted restraints in Monte Carlo folding simulations. *Proc. Natl. Acad. Sci. U. S. A.* **1998**, *95*, 1020−1025.
(31) Skolnick, J.; Kolinski, A.; Ortiz, A. R. MONSSTER: a method for folding globular proteins with a small number of distance restraints. *J. Mol. Biol.* **1997**, *265*, 217−241.
(32) Marks, D. S.; Hopf, T. A.; Sander, C. Protein structure prediction from sequence variation. *Nat. Biotechnol.* **2012**, *30*, 1072−1080.
(33) Chen, C.; Wu, T.; Guo, Z.; Cheng, J. Combination of deep neural network with attention mechanism enhances the explainability of protein contact prediction. *Proteins: Struct., Funct., Genet.* **2021**, *89*, 697−707.
(34) Milik, M.; Kolinski, A.; Skolnick, J. Neural network system for the evaluation of side-chain packing in protein structures. *Protein Eng., Des. Sel.* **1995**, *8*, 225−236.
(35) Gao, M.; Skolnick, J. A novel sequence alignment algorithm based on deep learning of the protein folding code. *Bioinformatics* **2021**, *37*, 490−496.
(36) Gao, M.; Skolnick, J. A General Framework to Learn Tertiary Structure for Protein Sequence Characterization. *Frontiers in Bioinformatics* **2021**, *1*, na DOI: 10.3389/fbinf.2021.689960.
(37) Skolnick, J.; Gao, M. The role of local versus nonlocal physicochemical restraints in determining protein native structure. *Curr. Opin. Struct. Biol.* **2021**, *68*, 1−8.
(38) Tunyasuvunakool, K.; Adler, J.; Wu, Z.; Green, T.; Zielinski, M.; Zidek, A.; Bridgland, A.; Cowie, A.; Meyer, C.; Laydon, A.; Velankar, S.; Kleywegt, G. J.; Bateman, A.; Evans, R.; Pritzel, A.; Figurnov, M.; Ronneberger, O.; Bates, R.; Kohl, S. A. A.; Potapenko, A.; Ballard, A. J.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Clancy, E.; Reiman, D.; Petersen, S.; Senior, A. W.; Kavukcuoglu, K.; Birney, E.; Kohli, P.; Jumper, J.; Hassabis, D. Highly accurate protein structure prediction for the human proteome. *Nature* **2021**, *596*, 590−596.
(39) Zhou, H.; Skolnick, J. Ab initio protein structure prediction using chunk-TASSER. *Biophys. J.* **2007**, *93*, 1510−8.
(40) Pandit, S. B.; Skolnick, J. Fr-TM-align: a new protein structural alignment method based on fragment alignments and the TM-score. *BMC Bioinf.* **2008**, *9*, 531.
(41) Gao, M.; Skolnick, J. Structural space of protein-protein interfaces is degenerate, close to complete, and highly connected. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107*, 22517−22.
(42) Gao, M.; Skolnick, J. iAlign: a method for the structural comparison of protein-protein interfaces. *Bioinformatics* **2010**, *26*, 2259−65.
(43) Tonddast-Navaei, S.; Skolnick, J. Are protein-protein interfaces special regions on a protein's surface? *J. Chem. Phys.* **2015**, *143*, 243149.
(44) Zhou, H.; Gao, M.; Skolnick, J. Comprehensive prediction of drug-protein interactions and side effects for the human proteome. *Sci. Rep.* **2015**, *5*, 11090.
(45) Brylinski, M.; Skolnick, J. A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 129−34.
(46) Zhang, C.; Freddolino, P. L.; Zhang, Y. COFACTOR: improved protein function prediction by combining structure, sequence and protein-protein interaction information. *Nucleic Acids Res.* **2017**, *45*, W291−W299.
(47) Zhou, H.; Cao, H.; Skolnick, J. FINDSITE(comb2.0): A New Approach for Virtual Ligand Screening of Proteins and Virtual Target Screening of Biomolecules. *J. Chem. Inf. Model.* **2018**, *58*, 2343−2354.
(48) Zhou, H.; Cao, H.; Skolnick, J. FRAGSITE: A Fragment-Based Approach for Virtual Ligand Screening. *J. Chem. Inf. Model.* **2021**, *61*, 2074−2089.
(49) Goodsell, D. S.; Sanner, M. F.; Olson, A. J.; Forli, S. The AutoDock suite at 30. *Protein Sci.* **2021**, *30*, 31.
(50) Torrens-Fontanals, M.; Stepniewski, T. M.; Aranda-Garcia, D.; Morales-Pastor, A.; Medel-Lacruz, B.; Selent, J. How Do Molecular Dynamics Data Complement Static Structural Data of GPCRs. *Int. J. Mol. Sci.* **2020**, *21*, 5933.
(51) Clark, J. J.; Orban, Z. J.; Carlson, H. A. Predicting binding sites from unbound versus bound protein structures. *Sci. Rep.* **2020**, *10*, 15856.
(52) Lakkaraju, S. K.; Yu, W.; Raman, E. P.; Hershfeld, A. V.; Fang, L.; Deshpande, D. A.; MacKerell, A. D., Jr. Mapping functional group free energy patterns at protein occluded sites: nuclear receptors and G-protein coupled receptors. *J. Chem. Inf. Model.* **2015**, *55*, 700−8.
(53) Ramanathan, A.; Savol, A.; Burger, V.; Chennubhotla, C. S.; Agarwal, P. K. Protein conformational populations and functionally relevant substates. *Acc. Chem. Res.* **2014**, *47*, 149−56.
(54) Modi, V.; Dunbrack, R. L., Jr. Defining a new nomenclature for the structures of active and inactive kinases. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116*, 6818−6827.
(55) Weis, W. I.; Kobilka, B. K. The Molecular Basis of G Protein-Coupled Receptor Activation. *Annu. Rev. Biochem.* **2018**, *87*, 897−919.
(56) Xie, T.; Saleh, T.; Rossi, P.; Kalodimos, C. G. Conformational states dynamically populated by a kinase determine its function. *Science* **2020**, *370*, eabc2754.
(57) Godzik, A. Knowledge-based potentials for protein folding: what can we learn from known protein structures? *Structure* **1996**, *4*, 363−6.
(58) Mohanty, D.; Dominy, B. N.; Kolinski, A.; Brooks, C. L., 3rd; Skolnick, J. Correlation between knowledge-based and detailed

atomic potentials: application to the unfolding of the GCN4 leucine zipper. *Proteins: Struct., Funct., Genet.* **1999**, *35*, 447−52.
(59) Mullinax, J. W.; Noid, W. G. Recovering physical potentials from a model protein databank. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107*, 19867−19872.
(60) Alpha Fold Protein Structure Database. *DeepMind*. https://deepmind.com/research/open-source/alphafold-protein-structure-database (accessed September 2021).
(61) Li, W.; Kinch, L. N.; Karplus, P. A.; Grishin, N. V. ChSeq: A database of chameleon sequences. *Protein Sci.* **2015**, *24*, 1075−86.
(62) Vajda, S.; Beglov, D.; Wakefield, A. E.; Egbert, M.; Whitty, A. Cryptic binding sites on proteins: definition, detection, and druggability. *Curr. Opin. Chem. Biol.* **2018**, *44*, 1−8.
(63) Chen, J.; Wang, W.; Sun, H.; Pang, L.; Bao, H. Binding mechanism of inhibitors to p38alpha MAP kinase deciphered by using multiple replica Gaussian accelerated molecular dynamics and calculations of binding free energies. *Comput. Biol. Med.* **2021**, *134*, 104485.
(64) Zhou, H.; Cao, H.; Matyunina, L.; Shelby, M.; Cassels, L.; McDonald, J.; Skolnick, J. MEDICASCY: A Machine Learning Approach for Predicting Small Molecule Drug Side Effects, Indications, Efficacy and Mode of Action. *Mol. Pharmaceutics* **2020**, *17*, 1558−1574.

https://doi.org/10.1021/acs.jcim.1c01114
J. Chem. Inf. Model. 2021, 61, 4827−4831