AlphaFold3, a secret sauce for predicting mutational effects on protein-protein interactions

Wei Lu
Shanghai Jiao Tong University
luwei0917@gmail.com

Jixian Zhang
Shanghai Jiao Tong University
jxzly1993@gmail.com

Jiahua Rao
Sun Yat-sen University
raojh6@mail2.sysu.edu.cn

Zhongyue Zhang
Shanghai Jiao Tong University
zhongyuezhang@sjtu.edu.cn

Shuangjia Zheng
Shanghai Jiao Tong University
shuangjia.zheng@sjtu.edu.cn

Abstract

AlphaFold3 has set the new state-of-the-art in predicting protein-protein complex structures. However, the complete picture of biomolecular interactions cannot be fully captured by static structures alone. In the field of protein engineering and antibody discovery, the connection from structure to function is often mediated by binding energy. This work benchmarks AlphaFold3 against SKEMPI, a commonly used binding energy dataset. We demonstrate that AlphaFold3 learns unique information and synergizes with force field, profile-based, and other deep learning methods in predicting the mutational effects on protein-protein interactions. We hypothesize that AlphaFold3 captures a more global effect of mutations by learning a smoother energy landscape, but it lacks the modeling of full atomic details that are better addressed by force field methods, which possess a more rugged energy landscape. Integrating both approaches could be an interesting future direction. All of our benchmark results are openly available at https://github.com/luwei0917/AlphaFold3_PPI.

1 Introduction

The development of AlphaFold and other deep learning methods has revolutionized the study of the protein complex structures [1–7], advancing beyond traditional physics-based docking methods [8–10]. With the advent of AlphaFold3, the success rate of predicting general protein complex structures has reached almost 80%, and that for protein-antibody pairs has exceeded 60%, significantly outperforming its predecessor, AF2-Multimer, previously considered the state-of-the-art [1, 2]. However, as Derek Lowe points out [11], "Structure is not everything." The complete picture of biomolecular interactions cannot be fully captured by static structures alone; it also involves the dynamic association and dissociation of one protein with another protein partner. In an equilibrium state, this interaction property is commonly measured as the binding affinity, K_d, or described as the change in binding free energy, ΔG. In addition, nature is constantly evolving, with mutations regularly introduced that modulate the magnitude of binding affinity, among many other properties, such as the stability of the proteins themselves. A key therapeutic goal is to design molecules that bind strongly enough to either prevent or promote a specific state of its binding partner, such as the inactive or active state of GPCRs. In the field of antibody discovery, the process of mutating a candidate antibody to improve its binding affinity with the target antigen protein is termed antibody maturation. Most antibody drugs in clinical use have picomolar affinity for the target protein [12], while antibodies found through immunization commonly exhibit affinities in the nanomolar or even micromolar range [13]. This means that, during the maturation step, we need to improve the affinity by more than a thousand times, or, if measured in $\Delta\Delta G$, by about 4 kcal/mol.

A variety of methods have been developed to estimate the effects of mutations, ranging from force field-based methods [14, 15], which derive forces from physical interactions such as van der Waals and electrostatic, or from statistical energy, to profile-based methods [16], which query sequence and structure databases. Additionally, there are hybrid methods [17] that combine force field and profile-based information. More recently, deep learning methods [18–21] have emerged, where the underlying energy landscapes are learned through unsupervised pre-training involving perturbations of crystallized protein structures.

In this work, we demonstrate that AlphaFold, although trained as a structure prediction model, learns critical information that is complementary to all types of current methods studying mutational effects on protein-protein interactions. This observation aligns with findings from previous studies of protein-small molecule and protein-peptide interactions [22–25], suggesting that models also trained to predict complex structures demonstrate enhanced capabilities in predicting binding affinity.

2 Related Work

Many works have demonstrated that AlphaFold2 already learns important features useful for other tasks. For example, in the protein design field, RFdiffusion [26] showed that fine-tuning from the pre-trained structure prediction model, RoseTTAFold [4], significantly enhances performance in protein design compared to starting without pre-training. Additionally, Roney [27] demonstrated that AlphaFold2 discerns the underlying physics capable of differentiating decoy structures from native structures, thereby effectively ranking candidate complex structures.

Other works utilize AF2 outputs directly as inputs for their models. Akdel [28] demonstrated that the AF2-predicted structure could serve as the input for many popular structure-based predictors of protein thermostability, such as FoldX and DynaMut2 [14, 29], achieving results comparable to those obtained with crystal structures. Additionally, Lyu [30] showed that AlphaFold2 structures could be used as inputs for docking programs in small molecule drug discovery. Although Buel [31] showed that AlphaFold2 cannot predict key mutational effects in many cases, McBride [32] demonstrated that with appropriately chosen metrics, such as effective strain in their study, AlphaFold2 can accurately predict the effects of mutations on the intrinsic properties of single proteins, using three experimental datasets: fluorescence, folding, and catalysis.

These works have explored the utility of AlphaFold in many research areas, but none of them have studied the usefulness of AlphaFold in predicting the binding energy and the mutational effects of mutations on protein-protein interactions.

3 Benchmark setups

A commonly used dataset for evaluating methods that predict mutation effects is SKEMPI [33], which manually curates a list of crystallized protein complexes and their mutants with experimentally measured changes in binding affinity, denoted as $\Delta\Delta G$ values, gathered through literature searches. These values are measured using biochemical methods that, while relatively accurate, require considerable effort for each mutant's data. As a result, despite the intensive labor involved, this dataset contains binding data for only 7085 mutations, which is significantly fewer than the total number of crystallized structures, and even less than the number of sequences in the database that enabled the development of AlphaFold. This scarcity of data underscores the value of leveraging information learned from tasks beyond direct protein-protein binding data.

3.1 Dataset definition

In order to benchmark against a wide range of methods, we utilized the common subset of Test Set 1 from SKEMPI as defined in SSIPe [17] and the SKEMPI dataset as employed in DSMBind [19]. The cases where the ranking score predicted by AlphaFold3 is below 0.8 are removed. As a result, our benchmark comprises 475 mutants across 42 unique protein complexes.

3.2 Baselines

We include a comprehensive set of 17 baseline methods for benchmarking. These can be categorized by their types of approaches.

Protein Language-based Models ESM2, ESM1v, and ProGen2 [34–36] are prominent protein language models known for their robust zero-shot performance across multiple tasks, including secondary structure prediction, and the classification of benign and pathogenic mutations. Given that these models typically accept only a single sequence as input, we concatenated the sequences of interacting proteins for our analysis.

Force Field and Profile-based Models Included in our baselines are three popular pure force field-based models: FoldX, FlexddG, and EvoEF [14, 15, 37]; a structure-based profiling baseline, BindProfX [16]; and one hybrid model that combines sequence and structure profiling with force field methods, SSIPe [17]. These models are specifically designed to predict the mutational effects on protein-protein interactions, with SSIPe considered the state-of-the-art model that utilizes the most information available.

Structure-based Deep Learning Models ProteinMPNN [38] is a model that learns to design sequences corresponding to a given input backbone structure, while DSMBind [19] predicts mutational effects by learning to restore a perturbed crystal structure. Both models develop a scoring function that can be used to estimate mutational effects for a given input structure and corresponding sequences.

AlphaFold3, AlphaFold2, and Strain Protein sequences are submitted directly to the AlphaFold3 server in JSON format with the seed set to a fixed arbitrary number, 42, to ensure reproducibility. Five results are downloaded, and the top one is used. For each mutant, the predicted score is calculated by subtracting the ranking score of the wild type from its ranking score. AF2-Multimer v2.3.0 [2] is run locally with default settings, and the ranking score for the top predicted model for each entry is used. As a simple statistical baseline, effective strain, as defined in [27], is computed for each mutant.

4 Results

4.1 Comparison of binding energy estimation across all baselines

The results for each baseline are summarized in Table 1. Pearson and Spearman correlation coefficients are two commonly used metrics for assessing continuous variables. The Area under the ROC Curve (AUC) is a statistic used in binary classification, computed by treating all $\Delta\Delta G$ values below zero as positive and those above zero as negative. The results are sorted according to their Pearson correlation coefficients within each category.

Table 1 indicates that protein language models are less effective at predicting the mutational effects on protein-protein interactions. Similarly, AlphaFold2 and Strain show weak correlations with experimentally measured binding affinities. In contrast, the ranking score produced by AlphaFold3 exhibits a significant correlation and is comparable to that of the widely-used FoldX method.

4.2 AlphaFold3 complements other baselines

As demonstrated by SSIPe [17], profile-based and force field-based methods complement each other, producing the most accurate estimators. Similarly, if AlphaFold3 learns a different type of information, it could also complement other models. As shown in Fig 1, a simple ensemble of AlphaFold3's ranking scores boosts performance across all baselines. The ensemble score is computed by adding the equally weighted ranked scores of two models. Notably, the previous state-of-the-art, SSIPe, which is already a combination of models, also experiences a performance boost. AlphaFold3 learns complementary information that is orthogonal to current methods, thereby enhancing the estimation of mutation effects on protein-protein interactions.

Table 1: Comparison of $\Delta\Delta G$ estimation results on our SKEMPI test set using three different metrics.

Category	Method	Pearson	Spearman	AUC
Force Field and Profile-based	SSIPe	0.68	0.62	0.78
	FlexddG	0.62	0.58	0.77
	BindProfX	0.56	0.58	0.74
	EvoEF	0.55	0.51	0.72
	FoldX	0.49	0.54	0.74
Structure-based Deep Learning	DSMBind	0.62	0.53	0.73
	ProteinMPNN	0.51	0.45	0.65
AlphaFold	AF3 ranking_score	0.49	0.51	0.71
	AF3 iptm	0.49	0.50	0.72
	AF3 ptm	0.36	0.33	0.63
	AF3 mean_pae	0.32	0.37	0.64
	AF2 ranking_score	0.21	0.23	0.57
	Effective Strain	0.18	0.31	0.61
	AF2 mean_pae	0.05	0.22	0.54
Protein Language-based	ESM2	0.27	0.35	0.68
	ESM1v	-0.02	0.06	0.52
	ProGen2	-0.09	0.01	0.47

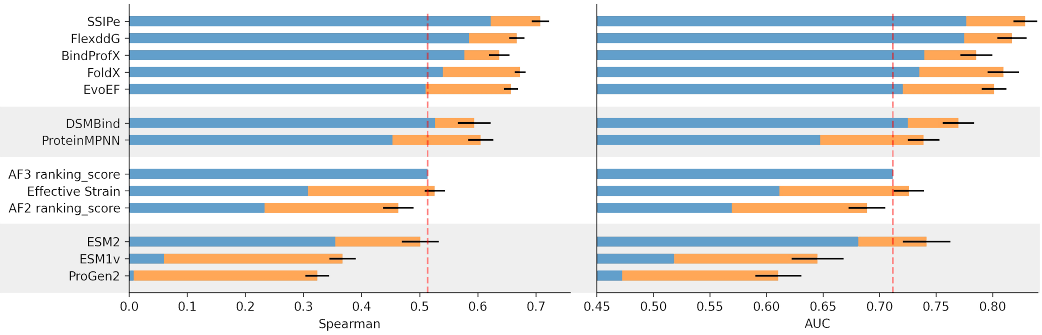

Figure 1: Ensemble with AlphaFold3 boosts performance across all baselines, as evaluated by Spearman correlation, **Left**, and AUC, **Right**, blue is the baseline score, orange is the boost in performance after ensemble with AlphaFold3 score. The dashed line indicates the AlphaFold3 performance.

4.3 AlphaFold3 offers unique information

To investigate whether the predictions made by AlphaFold3 are correlated with those of other methods, we computed the pairwise correlation among all methods, as shown on the left of Fig. 2. AlphaFold3 exhibits very weak correlations with other models, only showing slight correlation with DSMBind. In contrast, other models, such as FlexddG and SSIPe, correlate with many other methods, indicating that AlphaFold3 learns unique features that are orthogonal to those of other methods. As shown on the right side of Fig. 2, protein language models, AlphaFold2, and strain do not provide additional information beyond what AlphaFold3 provides. Conversely, structure-based deep learning, as well as force field and profile-based methods, enhance the predictions made by AlphaFold3.

5 Discussion

Our results indicate that the performance of protein language models in predicting the mutational effects on binding affinity is relatively weak. This finding aligns with a recent study [39], which demonstrates that language models do not scale effectively with model size in prediction tasks that are less dependent on coevolutionary patterns.

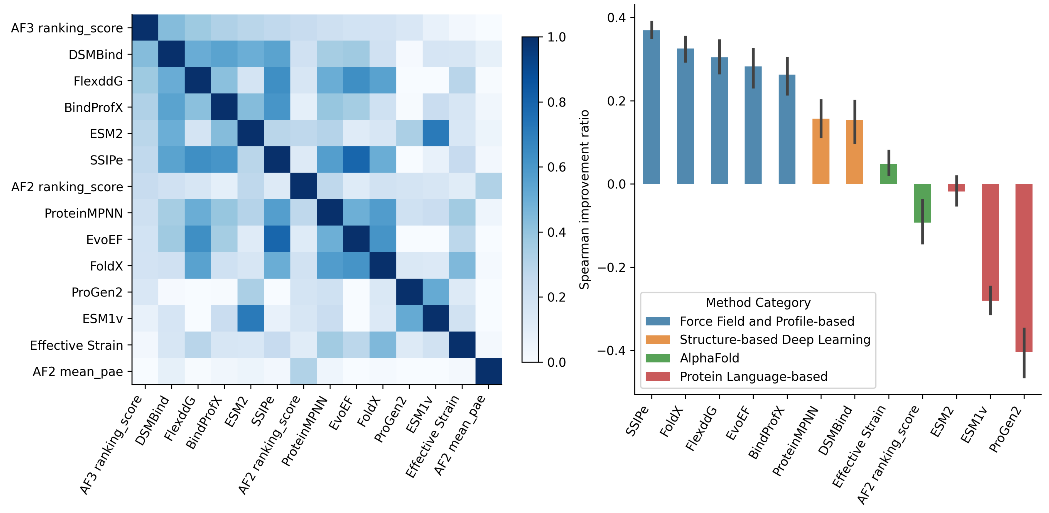

Figure 2: **Left**, correlation between model predictions, sorted by their correlation against AlphaFold3. **Right**, when AlphaFold3 is combined with other models, all methods except protein language models and AlphaFold2 get a significant boost. y axis measures the ratio of improvement relative to the AlphaFold3 baseline, $(r_{m+AF3} - r_{AF3})/r_{AF3}$

An interesting distinction between AlphaFold3 and traditional all-atom force field methods lies in their sensitivity to protein complex conformations. Traditional force fields are highly sensitive to the exact conformation of the protein complex, whereas AlphaFold3 tends to predict similar scores for identical input sequences. This sensitivity to conformation and the difficulty in thoroughly sampling conformations make force field methods prone to inaccurate estimations of the entropy component of the total Gibbs free energy. In contrast, AlphaFold3, as a generative structure prediction model, is capable of learning a smoother energy landscape that more effectively captures the subtle influences of entropy. Fig 3 illustrates how the integration of AlphaFold3 scores can more accurately determine the global relative free energy, ΔG, compared to relying solely on force field methods, which exhibit a more rugged energy landscape.

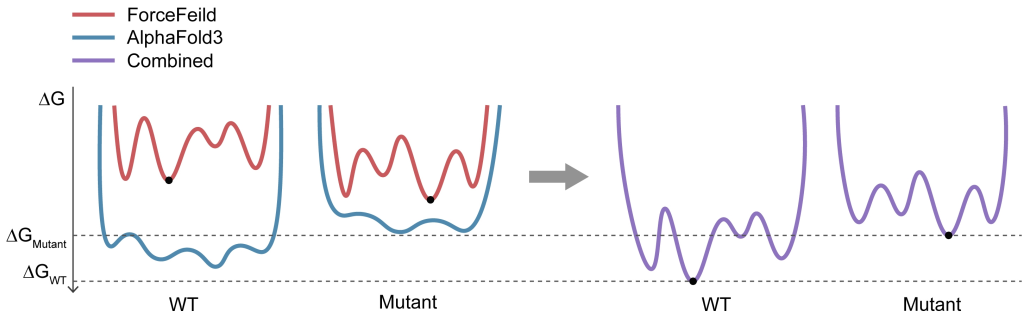

Figure 3: A schematic plot illustrating how the rugged energy landscapes of all-atom force fields (in red) when combined with the globally more accurate energy landscape of AlphaFold3 (in blue), result in a composite energy landscape (in purple) that more closely approximates the ground truth (dashed lines) for estimating ΔG.

6 Conclusion

In this study, we have demonstrated that AlphaFold3 learns unique features beneficial for estimating binding free energy and complements existing models. Looking ahead, a more integrated approach combining folding methods that predict complex structures with inverse-folding models that identify masked residues, along with traditional force field and profile-based models, could significantly revolutionize the field of predicting mutational effects on protein-protein interactions.

Acknowledgments and Disclosure of Funding

We thank our colleagues for submitting the test queries to the AlphaFold3 server. The authors declare that there are no competing interests associated with this work. This work was partly supported by Young Elite Scientists Sponsorship Program by CAST and RayWu Angel Fund.

References

[1] Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, pages 1–3, 2024.

[2] Richard Evans, Michael O'Neill, Alexander Pritzel, Natasha Antropova, Andrew Senior, Tim Green, Augustin Žídek, Russ Bates, Sam Blackwell, Jason Yim, et al. Protein complex prediction with alphafold-multimer. *biorxiv*, pages 2021–10, 2021.

[3] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

[4] Minkyung Baek, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N Kinch, R Dustin Schaeffer, et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557):871–876, 2021.

[5] Kolja Stahl, Andrea Graziadei, Therese Dau, Oliver Brock, and Juri Rappsilber. Protein structure prediction with in-cell photo-crosslinking mass spectrometry and deep learning. *Nature Biotechnology*, 41(12):1810–1819, 2023.

[6] David F Burke, Patrick Bryant, Inigo Barrio-Hernandez, Danish Memon, Gabriele Pozzati, Aditi Shenoy, Wensi Zhu, Alistair S Dunham, Pascal Albanese, Andrew Keller, et al. Towards a structurally resolved human protein interaction network. *Nature Structural & Molecular Biology*, 30(2):216–225, 2023.

[7] Marc F Lensink, Guillaume Brysbaert, Nessim Raouraoua, Paul A Bates, Marco Giulini, Rodrigo V Honorato, Charlotte van Noort, Joao MC Teixeira, Alexandre MJJ Bonvin, Ren Kong, et al. Impact of alphafold on structure prediction of protein complexes: the casp15-capri experiment. *Proteins: Structure, Function, and Bioinformatics*, 91(12):1658–1683, 2023.

[8] Dima Kozakov, David R Hall, Bing Xia, Kathryn A Porter, Dzmitry Padhorny, Christine Yueh, Dmitri Beglov, and Sandor Vajda. The cluspro web server for protein–protein docking. *Nature protocols*, 12(2):255–278, 2017.

[9] Yumeng Yan, Huanyu Tao, Jiahua He, and Sheng-You Huang. The hdock server for integrated protein–protein docking. *Nature protocols*, 15(5):1829–1852, 2020.

[10] Brian G Pierce, Kevin Wiehe, Howook Hwang, Bong-Hyun Kim, Thom Vreven, and Zhiping Weng. Zdock server: interactive docking prediction of protein–protein complexes and symmetric multimers. *Bioinformatics*, 30(12):1771–1773, 2014.

[11] Derek Lowe. Alphafold 3 debuts. `https://www.science.org/content/blog-post/alphafold-3-debuts`, 2023. Accessed: 2024-05-15.

[12] Matthew IJ Raybould, Claire Marks, Alan P Lewis, Jiye Shi, Alexander Bujotzek, Bruck Taddese, and Charlotte M Deane. Thera-sabdab: the therapeutic structural antibody database. *Nucleic acids research*, 48(D1):D383–D388, 2020.

[13] James Dunbar, Konrad Krawczyk, Jinwoo Leem, Terry Baker, Angelika Fuchs, Guy Georges, Jiye Shi, and Charlotte M Deane. Sabdab: the structural antibody database. *Nucleic acids research*, 42(D1):D1140–D1146, 2014.

[14] Joost Schymkowitz, Jesper Borg, Francois Stricher, Robby Nys, Frederic Rousseau, and Luis Serrano. The foldx web server: an online force field. *Nucleic acids research*, 33(suppl_2):W382–W388, 2005.

[15] Julia Koehler Leman, Brian D Weitzner, Steven M Lewis, Jared Adolf-Bryfogle, Nawsad Alam, Rebecca F Alford, Melanie Aprahamian, David Baker, Kyle A Barlow, Patrick Barth, et al. Macromolecular modeling and design in rosetta: recent methods and frameworks. *Nature methods*, 17(7):665–680, 2020.

[16] Peng Xiong, Chengxin Zhang, Wei Zheng, and Yang Zhang. Bindprofx: assessing mutation-induced binding affinity change by protein interface profiles with pseudo-counts. *Journal of molecular biology*, 429(3):426–434, 2017.

[17] Xiaoqiang Huang, Wei Zheng, Robin Pearce, and Yang Zhang. Ssipe: accurately estimating protein–protein binding affinity change upon mutations using evolutionary profiles in combination with an optimized physical energy function. *Bioinformatics*, 36(8):2429–2437, 2020.

[18] Anton Bushuiev, Roman Bushuiev, Anatolii Filkin, Petr Kouba, Marketa Gabrielova, Michal Gabriel, Jiri Sedlar, Tomas Pluskal, Jiri Damborsky, Stanislav Mazurenko, et al. Learning to design protein-protein interactions with enhanced generalization. *arXiv preprint arXiv:2310.18515*, 2023.

[19] Wengong Jin, Xun Chen, Amrita Vetticaden, Siranush Sarzikova, Raktima Raychowdhury, Caroline Uhler, and Nir Hacohen. Dsmbind: Se (3) denoising score matching for unsupervised binding energy prediction and nanobody design. *bioRxiv*, pages 2023–12, 2023.

[20] Shitong Luo, Yufeng Su, Zuofan Wu, Chenpeng Su, Jian Peng, and Jianzhu Ma. Rotamer density estimator is an unsupervised learner of the effect of mutations on protein-protein interaction. *bioRxiv*, pages 2023–02, 2023.

[21] Shiwei Liu, Tian Zhu, Milong Ren, Chungong Yu, Dongbo Bu, and Haicang Zhang. Predicting mutational effects on protein-protein binding via a side-chain diffusion probabilistic model. *Advances in Neural Information Processing Systems*, 36, 2024.

[22] Shuya Li, Fangping Wan, Hantao Shu, Tao Jiang, Dan Zhao, and Jianyang Zeng. Monn: a multi-objective neural network for predicting compound-protein interactions and affinities. *Cell Systems*, 10(4):308–322, 2020.

[23] Wei Lu, Qifeng Wu, Jixian Zhang, Jiahua Rao, Chengtao Li, and Shuangjia Zheng. Tankbind: Trigonometry-aware neural networks for drug-protein binding structure prediction. *Advances in neural information processing systems*, 35:7236–7249, 2022.

[24] Wei Lu, Jixian Zhang, Weifeng Huang, Ziqiao Zhang, Xiangyu Jia, Zhenyu Wang, Leilei Shi, Chengtao Li, Peter G Wolynes, and Shuangjia Zheng. Dynamicbind: predicting ligand-specific protein-ligand complex structure with a deep equivariant generative model. *Nature Communications*, 15(1):1071, 2024.

[25] Amir Motmaen, Justas Dauparas, Minkyung Baek, Mohamad H Abedi, David Baker, and Philip Bradley. Peptide-binding specificity prediction using fine-tuned protein structure prediction networks. *Proceedings of the National Academy of Sciences*, 120(9):e2216697120, 2023.

[26] Joseph L Watson, David Juergens, Nathaniel R Bennett, Brian L Trippe, Jason Yim, Helen E Eisenach, Woody Ahern, Andrew J Borst, Robert J Ragotte, Lukas F Milles, et al. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100, 2023.

[27] James P Roney and Sergey Ovchinnikov. State-of-the-art estimation of protein model accuracy using alphafold. *Physical Review Letters*, 129(23):238101, 2022.

[28] Mehmet Akdel, Douglas EV Pires, Eduard Porta Pardo, Jürgen Jänes, Arthur O Zalevsky, Bálint Mészáros, Patrick Bryant, Lydia L Good, Roman A Laskowski, Gabriele Pozzati, et al. A structural biology community assessment of alphafold2 applications. *Nature Structural & Molecular Biology*, 29(11):1056–1067, 2022.

[29] Carlos HM Rodrigues, Douglas EV Pires, and David B Ascher. Dynamut2: Assessing changes in stability and flexibility upon single and multiple point missense mutations. *Protein Science*, 30(1):60–69, 2021.

[30] Jiankun Lyu, Nicholas Kapolka, Ryan Gumpper, Assaf Alon, Liang Wang, Manish K Jain, Ximena Barros-Álvarez, Kensuke Sakamoto, Yoojoong Kim, Jeffrey DiBerto, et al. Alphafold2 structures guide prospective ligand discovery. *Science*, page eadn6354, 2024.

[31] Gwen R Buel and Kylie J Walters. Can alphafold2 predict the impact of missense mutations on structure? *Nature structural & molecular biology*, 29(1):1–2, 2022.

[32] John M McBride, Konstantin Polev, Amirbek Abdirasulov, Vladimir Reinharz, Bartosz A Grzybowski, and Tsvi Tlusty. Alphafold2 can predict single-mutation effects. *Physical Review Letters*, 131(21):218401, 2023.

[33] Justina Jankauskaitė, Brian Jiménez-García, Justas Dapkūnas, Juan Fernández-Recio, and Iain H Moal. Skempi 2.0: an updated benchmark of changes in protein–protein binding energy, kinetics and thermodynamics upon mutation. *Bioinformatics*, 35(3):462–469, 2019.

[34] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Sal Candido, et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *BioRxiv*, 2022:500902, 2022.

[35] Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alex Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. *Advances in neural information processing systems*, 34:29287–29303, 2021.

[36] Erik Nijkamp, Jeffrey A Ruffolo, Eli N Weinstein, Nikhil Naik, and Ali Madani. Progen2: exploring the boundaries of protein language models. *Cell systems*, 14(11):968–978, 2023.

[37] Xiaoqiang Huang, Robin Pearce, and Yang Zhang. Evoef2: accurate and fast energy function for computational protein design. *Bioinformatics*, 36(4):1135–1142, 2020.

[38] Justas Dauparas, Ivan Anishchenko, Nathaniel Bennett, Hua Bai, Robert J Ragotte, Lukas F Milles, Basile IM Wicky, Alexis Courbet, Rob J de Haas, Neville Bethel, et al. Robust deep learning–based protein sequence design using proteinmpnn. *Science*, 378(6615):49–56, 2022.

[39] Francesca-Zhoufan Li, Ava P Amini, Yisong Yue, Kevin K Yang, and Alex X Lu. Feature reuse and scaling: Understanding transfer learning with protein language models. *bioRxiv*, pages 2024–02, 2024.