Exploring zero-shot structure-based protein fitness prediction

Arnav Sharma[*†] **& Anthony Gitter**[*†‡]
[*] Department of Computer Sciences,
[‡] Department of Biostatistics and Medical Informatics
University of Wisconsin-Madison
[†] Morgridge Institute for Research
arnav@cs.wisc.edu, gitter@biostat.wisc.edu

Abstract

The ability to make zero-shot predictions about the fitness consequences of protein sequence changes with pre-trained machine learning models enables many practical applications. Such models can be applied for downstream tasks like genetic variant interpretation and protein engineering without additional labeled data. The advent of capable protein structure prediction tools has led to the availability of orders of magnitude more precomputed predicted structures, giving rise to powerful structure-based fitness prediction models. Through our experiments, we assess several modeling choices for structure-based models and their effects on downstream fitness prediction. Zero-shot fitness prediction models can struggle to assess the fitness landscape within disordered regions of proteins, those that lack a fixed 3D structure. We confirm the importance of matching protein structures to fitness assays and find that predicted structures for disordered regions can be misleading and affect predictive performance. Lastly, we evaluate an additional structure-based model on the ProteinGym substitution benchmark and show that simple multi-modal ensembles are strong baselines.

Introduction

Designing proteins with desired properties carries immense promise with applications in therapeutic development, agriculture, chemical manufacturing, and biofuels (Liu et al., 2022). Effectively navigating the large design space of proteins requires understanding the relation between a protein's sequence and its function (Romero et al., 2013). The opportunities in this scientific domain and excitement around protein science complemented by the rise in popularity of machine learning methods have driven a wide variety of new methods for quantitative protein function prediction (Yang et al., 2019). Computational methods built using aligned sequences or multiple sequence alignments (MSAs) (Rao et al., 2021) for evolutionarily related proteins can estimate protein fitness scores based on conservation or the likelihood a new sequence belongs to the protein family. In contrast, protein language models (PLMs) (Rives et al., 2021) train on and represent all known natural protein sequences to learn more general statistical patterns of evolution.

Protein structures provide additional information for estimating the fitness consequences of sequence changes. The increase in quality of predicted protein structures from models like AlphaFold 2 (Jumper et al., 2021) has been followed by an increasing number of structure-based models for predicting protein properties (Dauparas et al., 2022; Hsu et al., 2022; Blaabjerg et al., 2023). The initial ProteinGym 0.1 (Notin et al., 2022a), a benchmark for quantitative protein fitness prediction that used deep mutational scanning (DMS) assays for evaluation, did not contain structure-based models. However, ProteinGym 1.0 (Notin et al., 2023) was released with three inverse folding methods that use structures: ESM-IF1 (Hsu et al., 2022), ProteinMPNN (Dauparas et al., 2022), and MIF (Yang et al., 2023). Additional structure-based models were added after the ProteinGym publication. This warrants a closer

examination of structure-based models' behavior and protein fitness prediction performance to determine what modeling choices and biological factors affect their predictions.

Here we explore some design choices around using structure-based models for protein fitness prediction using the ProteinGym benchmark (Notin et al., 2023). We examine how the type of protein structure (predicted or experimental) can affect model predictions. We observe that a number of assays in the ProteinGym benchmark involve proteins containing intrinsically disordered regions (IDRs). These are regions in the protein sequence that lack any rigid 3D structure and can adopt multiple varied conformations (Trivedi & Nagarajaram, 2022). We demonstrate how IDRs can affect model predictions made by various models. Finally, we expand the structure-based models in ProteinGym, comparing them to strong existing MSA- and PLM-based models. We show strengths of uni-modal and multi-modal structure-based models, especially for stability assays. Additionally, we show that simple ensemble models that include structure perform surprisingly well compared to more sophisticated models.

Results

We use ProteinGym (Notin et al., 2023) to explore structure-based predictions of protein fitness. ProteinGym hosts a collection of DMS substitution assays measuring quantitative function of proteins belonging to different taxonomies and five different function types: activity, binding, expression, organismal fitness, and stability (Table 3). Furthermore, the benchmark hosts a variety of models sorted according to their average Spearman correlation with measured function across protein and function types.

Choice of structure affects zero-function protein function prediction

Previous work on PLMs has shown that given a corrupted protein sequence, predicting log-likelihood of mutations at the corrupted residue can be indicative of protein fitness (Rives et al., 2021; Yang et al., 2024). Unlike sequence-based PLMs, inverse folding models take a corrupted sequence as well as the backbone structure of a protein and try to predict the likelihood of the corrupted residue. This explicit conditioning on protein structure has shown to improve recovery of the corrupted residues (Yang et al., 2023) and is therefore indicative of better fitness prediction.

We first benchmark zero-shot predictive performance of the structure-based ESM-IF1 on all DMS substitution assays from ProteinGym for which we could obtain matching experimental structures. One key difference between experimental and predicted structures might be that predicted structures only contain the coordinates of a target chain whereas an experimental structure might contain a protein complex with multiple chains. Therefore, in order to explicitly account for this difference, we separate the results according to whether the experimental structure is a monomer (has a single chain) or a multimer (has multiple chains). We compare the difference of Spearman correlation using predicted versus experimental structures. Previous work has shown that experimental structures improve prediction of protein properties such as Enzyme Commission number (EC) and Gene Ontology term (GO) relative to structures predicted by AlphaFold 2 (Huang et al., 2024). Our results show that, for most assays in ProteinGym with experimental structures, Spearman correlation achieved using AlphaFold 2 predicted structures tends to be higher than that achieved using experimental structures (Table 1). A more detailed breakdown can be found in Figure 1 in the Appendix.

Structure Type	$\rho_{\mathbf{pred}} - \rho_{\mathbf{exp}} \geq 0$	$\rho_{\mathbf{pred}} - \rho_{\mathbf{exp}} < 0$	% **with** $\rho_{\mathbf{pred}} - \rho_{\mathbf{exp}} \geq 0$
Monomers	38	13	74.5%
Multimers	8	2	80%

Table 1: Count of DMS assays in ProteinGym with positive and negative $\rho_{\text{pred}} - \rho_{\text{exp}}$

Intrinsically disordered regions are prevalent in ProteinGym and inform structure-based model performance

52 of the 186 (28%) unique UniProt IDs in ProteinGym are proteins that are annotated as having disordered regions in DisProt in regions of the sequence covered by the DMS assay (Table 4). These proteins are associated with 63 of the 217 (29%) ProteinGym DMS substitution assays. Their disorder content ranges from 1.45% to 100% with a median of 15.67% and a mean of 23.56%. Nine proteins have over 45% disorder content. The Appendix details our process of identifying disordered regions in ProteinGym assays.

We assess the effect disordered regions have on function prediction at the protein and residue-level. We first benchmark models in Table 2 as well as ESM2 650M (Lin et al., 2023) on 63 DMS assays that have disordered regions in the target sequence defined by ProteinGym (Table 4). There is some variation in performance across models, which may be related to the models' different input modalities (Figure 2).

In order to more directly assess how disordered regions affect zero-shot fitness prediction quality, we examine the subset of 43 DMS assays that contain mutations in both disordered and ordered regions corresponding to 36 proteins (Tables 4 and 5). Disordered regions affect not only structure-based models like ESM-IF1 (Figure 3a), but also multi-modal models like ProtSSN (Tan et al., 2024) (Figure 3b), SaProt (Su et al., 2023) (Figure 3g), and TranceptEVE L (Notin et al., 2022b) (Figure 3c) for most function types. Disorder in proteins can also be detrimental to predictions made by PLMs (Figure 3d, 3f, 4). We also observe this phenomenon in ensembles of different models (Figure 3i, 3j, 3k, 3l; ensembles defined in Table 2). Disordered regions tend to be fast-evolving and less conserved (Brown et al., 2002; Light et al., 2013; Fawzy & Marsh, 2025), which may partially explain this observation.

Because many structure-aware protein function prediction methods rely on structures predicted by models like AlphaFold 2, studying the consequences of disordered regions on predicted structures can inform fitness prediction outputs. We take a more in-depth look at two such proteins with disordered regions: P37840 (human α-synuclein) and Q99801 (human NKX3-1). α-synuclein is a protein associated with Parkinson's disease that can take on disordered, helical, or β-sheet-rich aggregate conformations (Newberry et al., 2020b;a). DMS data (Newberry et al., 2020b;a) support that an α-helicial membrane-bound conformation with increasing disorder toward the C-terminal end is relevant for α-synuclein toxicity. Comparing experimental and predicted α-synuclein structures (Figure 5), neither is a perfect match for this membrane-bound conformation, but the helical experimental structure is much more similar. The substantial difference between the two structures likely explains why this DMS assay SYUA_HUMAN_Newberry_2020 has the most negative $\rho_{\text{pred}} - \rho_{\text{exp}}$ in Figure 1.

NKX3-1 has a small positive $\rho_{\text{pred}} - \rho_{\text{exp}}$ for the NKX31_HUMAN_Tsuboyama_2023_2L9R DMS assay (Figure 1). Although the protein contains disordered regions, closer examination shows that the DMS assay only spans 61 positions of the protein that are ordered. The experimental and predicted structures cover this same region of the protein and are quite similar (Figure 6), consistent with the minimal difference in DMS prediction performance.

Multi-modal models are better zero-shot predictors

Besides the choice of input structure, another aspect that affects the performance of structure-based models is how and whether the structural information is combined with other data modalities. Both MSAs (Marks et al., 2011) and PLM embeddings (Rives et al., 2021; Zhang et al., 2024) already provide some information about protein structure. We assess whether multi-modal models that explicitly use protein structure input have different performance characteristics on the protein fitness prediction task than models that implicitly contain structural information via MSA, PLM, or both inputs. We create ensembles of strong existing zero-shot models by choosing the best-performing sequence- and MSA-based models from ProteinGym and combining each with the ESM-IF1 structure-based model to assess how adding the structure input affects predictive performance.

Benchmarking strong models for different input modalities as well as ensembles using these models on the ProteinGym benchmark shows that combining different modalities (through ensembling or using joint architectures) exhibits some of the highest metrics (Tables 2 and 6). Models that incorporate protein structure information tend to perform well on DMS substitution assays measuring stability, consistent with previous results (Paul et al., 2023; Blaabjerg et al., 2024). Furthermore, we see that even simple ensembles of models combining different modalities achieve some of the highest values across both Spearman correlation and Top 10 recall. This suggests that multi-modal models seem to capture complementary predictive signals that models operating on a single modality do not. Furthermore, multi-modal models such as ProtSSN, SSEmb, TranceptEVE L, and SaProt may still underperform simpler ensembled methods like StructSeq (Paul et al., 2023) and others.

Model	**Modality**	**Avg. ρ by Assay Function**					**Avg. ρ**
		Activity	Binding	Expression	Organismal Fitness	Stability	
ESM-IF1*	Structure	0.368	0.389	0.407	0.331	0.624	0.424
GEMME*	MSA	0.482	0.383	0.438	0.453	0.519	0.455
VESPA*	PLM	0.468	0.366	0.404	0.442	0.500	0.436
ProtSSN (ensemble) *	Structure + PLM	0.466	0.366	0.449	0.397	0.568	0.449
SSemb	Structure + MSA	0.436	0.386	0.469	0.422	0.596	0.462
TranceptEVE L*	MSA + PLM	0.487	0.376	0.457	0.460	0.500	0.456
SaProt (650M)*	Structure + PLM	0.458	0.379	0.488	0.373	0.592	0.458
Ensemble 1 StructSeq	Structure + MSA + PLM	0.485	0.432	0.483	0.439	0.639	0.495
Ensemble 2 ESM-IF1, Tranception	Structure + MSA + PLM	0.479	0.418	0.491	0.438	0.621	0.489
Ensemble 3 ESM-IF1, GEMME	Structure + MSA	0.491	0.387	0.440	0.454	0.529	0.460
Ensemble 4 ESM-IF1, VESPA	Structure + PLM	0.485	0.372	0.416	0.449	0.558	0.456

Table 2: Spearman correlation coefficient (ρ) of model predictions versus DMS assay scores according to the function type of each DMS assay. Scores for model names annotated with an asterisk (*) were calculated using their predictions provided in ProteinGym (Notin et al., 2023). Scores highlighted in red, olive, blue represent the 1st, 2nd, and 3rd highest score for each group. See Figure 7 for score distributions.

Methods

We obtain DMS data and AlphaFold 2 predicted protein structures from ProteinGym (Notin et al., 2023), experimental protein structures from the RCSB Protein Data Bank (PDB) (Zardecki et al., 2016), and disordered protein annotations from DisProt (Aspromonte et al., 2024). Predicted monomer structures in the ProteinGym were generated using AlphaFold 2 v2.3.1 with full_dbs and default parameters (Notin, 2025). Additional dataset details are in the Appendix.

There are many protein fitness prediction algorithms included in ProteinGym and far more not integrated into the benchmarking framework. We prioritized representative alignment-based, PLM-based, and structure-based models that performed well on the ProteinGym zero-shot DMS substitution assays as well as additional structure-based models and ensembles. The specific models and ensembles are described in the Appendix.

In order to evaluate the performance of each model on the ProteinGym dataset, we consider two metrics: Spearman correlation (ρ) and Top 10 Recall (Appendix). The average score for each metric follows the convention established by ProteinGym. We first take the average of scores belonging to assays targeting the same protein (determined by UniProt IDs). We

then take the average of the per protein averages across assay functions. Finally, the average column in Tables 2 and 6 represents a mean of per function average scores.

Discussion

Our goal is not to assert that one algorithm is the best overall for zero-shot protein fitness prediction. Rather, we provide insights into the strengths and weaknesses of different algorithms, input protein representations, and their combinations. Although MSA-based models (Marks et al., 2011) and PLMs (Rives et al., 2021; Zhang et al., 2024) implicitly contain information about protein structure, explicit structure-based models do in fact benefit in zero-shot prediction performance on the ProteinGym DMS substitution assays. This finding is further bolstered by observing how combining different modalities can lead to improved predictive performance. In addition, structure-based models using predicted structures as inputs often show better performance on ProteinGym's zero-shot prediction tasks. We note that ESM-IF is explicitly trained on AlphaFold 2 predicted structures (Hsu et al., 2022), which could explain why predicted structures outperform experimental ones in our experiments.

Another reason for the discrepancy between model performance on predicted and experimental structures could be the lack of manual validation of our chosen structures. As detailed in the Appendix, we may have selected experimental structures misaligned with the fitness prediction assays we benchmark in Figure 1. Due to the presence of proteins bound to other molecules, alternative conformations, and protein modifications made to facilitate structure characterization in the PDB, experimental structures can be misaligned with a functional property of interest. However, this problem—as evident in Figure 5—can also be observed in predicted structures. Our finding highlights the need for matching structures with downstream properties when making zero-shot fitness predictions and interpreting structure prediction models like AlphaFold 2 to understand how a protein's structure was predicted. Lastly, Rives et al. (2021); Yang et al. (2024) found that masked language modeling is highly indicative of protein fitness, but this may not be the case for GO and EC prediction tasks used by (Huang et al., 2024).

The best function prediction algorithm can depend on the protein and function (Tables 2 and 6). For some functions such as non-native enzymatic activity, zero-shot predictions are uncorrelated with functional assays (Yang et al., 2025). Furthermore, the ProteinGym assays represent only a fraction of available DMS assays, and our results may change across a larger set of assays. ProtaBank (Wang et al., 2018) and MaveDB (Rubin et al., 2025) catalog additional DMS datasets, and more are published continually.

One limitation of our study is that we did not use the same MSAs for all models. To get the best representation for SSEmb (Blaabjerg et al., 2024) on zero-shot prediction tasks, we generated MSAs using the process chosen by its authors instead of using ProteinGym MSAs. Furthermore, despite the zero-shot setting being a useful paradigm, it is also important to study how these models perform under supervised fine-tuning. Considering only MSAs, structure, and sequence as input modalities leaves out models like METL (Gelman et al., 2025) and RaSP (Blaabjerg et al., 2023) that are trained on alternate modalities like protein stability and biophysical simulations. We could only analyze a small subset of DMS assays with their experimental structures in part due to missing residues in the experimental structures. We could expand this subset by relaxing our criteria for selecting experimental structures to allow a small fraction of missing residues and model the missing residues with Rosetta (Huang et al., 2011; Song et al., 2013; Leman et al., 2020). In addition, we only considered a single experimental structure per protein and predicted structures from a single algorithm. Further exploring the stability DMS assays, which are mostly cDNA display proteolysis assays (Tsuboyama et al., 2023), may explain why the two structure-based models do not see same drop in performance for stability assays that they do in other function types.

Our preliminary examination of intrinsically disordered regions in ProteinGym proteins shows how disorder can affect fitness predictions of not just structure based models, but those based on sequences and MSAs. For example, Figure 4 shows ESM2 650M predictions for the P53_HUMAN_Giacomelli_2018_Null_Etoposide assay separated by mutations in ordered

and disordered regions. In the disordered region, ESM2 predicted pseudo log-likelihood correlates less with DMS scores, and the masked marginal probabilities have a slightly smaller range. Our results on DMS data are consistent with AlphaMissense's lower performance for pathogenic variant prediction within disordered regions (Cheng et al., 2023) and reduced sensitivity in a systematic evaluation of pathogenicity prediction (Fawzy & Marsh, 2025). Out of the structure-based models that we consider, ProtSSN (Tan et al., 2024) and SSEmb (Blaabjerg et al., 2024), do not give special treatment to disordered regions of a protein or low confidence regions of a predicted structure. ESM-IF1 (Hsu et al., 2022) on the other hand, takes the confidence of AlphaFold 2 (Jumper et al., 2021) predicted structures into account by masking low pLDDT coordinates. SaProt (Su et al., 2023) also takes confidence of predicted structures into account by masking structure tokens corresponding to low pLDDT residues and only using sequence tokens. Structure-based models are broadly applied for a variety of tasks beyond fitness prediction (Fout et al., 2017; Gligorijević et al., 2021; Xia et al., 2021; Gao et al., 2023), so our initial exploration of intrinsically disordered regions in ProteinGym presents a general opportunity for future model development. Future structure-based models could consider ways to account for the unique aspects of intrinsically disordered proteins or support conformational ensembles (Lindorff-Larsen & Kragelund, 2021). Lastly, we see that even though multi-modal models capture a more holistic representation of proteins, simple ensembles of different models can often surpass their performance. This motivates further research into architectures and training objectives that better combine information across modalities as well as using simple multi-modal ensembles as baselines when developing those models.

Acknowledgments

We thank Sam Gelman and Bryce Johnson for feedback and modeling suggestions as well as Moses Milchberg for protein structure discussions. This research was supported by National Science Foundation awards CHE 2226451 and OAC 2030508 and computing resources at the University of Wisconsin-Madison Center for High Throughput Computing (Center for High Throughput Computing, 2006).

Code availability

Code to reproduce our results can be found in our GitHub repository (`https://github.com/gitter-lab/benchmarking-structure-based-models`) and is archived on Zenodo at (`https://zenodo.org/doi/10.5281/zenodo.13821572`). MSAs, structures, and additional results are available in the Zenodo repository (`https://doi.org/10.5281/zenodo.13819823`).

Most of our datasets come from the ProteinGym benchmarks and can be downloaded from `https://proteingym.org/download`. As indicated, we used the zero-shot model scores provided by ProteinGym to compute metrics for models that are already part of the benchmark. See the Appendix for details.

References

Maria Cristina Aspromonte, Maria Victoria Nugnes, Federica Quaglia, Adel Bouharoua, DisProt Consortium, Silvio C E Tosatto, and Damiano Piovesan. DisProt in 2024: improving function annotation of intrinsically disordered proteins. *Nucleic Acids Research*, 52(D1):D434–D441, January 2024. ISSN 0305-1048. doi: 10.1093/nar/gkad928. URL https://doi.org/10.1093/nar/gkad928.

Sebastian Bittrich, Joan Segura, Jose M Duarte, Stephen K Burley, and Yana Rose. RCSB Protein Data Bank: exploring protein 3D similarities via comprehensive structural alignments. *Bioinformatics*, 40(6):btae370, June 2024. ISSN 1367-4811. doi: 10.1093/bioinformatics/btae370. URL https://doi.org/10.1093/bioinformatics/btae370.

Lasse M Blaabjerg, Maher M Kassem, Lydia L Good, Nicolas Jonsson, Matteo Cagiada, Kristoffer E Johansson, Wouter Boomsma, Amelie Stein, and Kresten Lindorff-Larsen. Rapid protein stability prediction using deep learning representations. *eLife*, 12:e82593, May 2023. ISSN 2050-084X. doi: 10.7554/eLife.82593. URL https://doi.org/10.7554/eLife.82593. Publisher: eLife Sciences Publications, Ltd.

Lasse M. Blaabjerg, Nicolas Jonsson, Wouter Boomsma, Amelie Stein, and Kresten Lindorff-Larsen. SSEmb: A joint embedding of protein sequence and structure enables robust variant effect predictions. *Nature Communications*, 15(1):9646, November 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-53982-z. URL https://www.nature.com/articles/s41467-024-53982-z. Publisher: Nature Publishing Group.

Antonia Boca and Simon Mathis. Predicting protein variants with equivariant graph neural networks. *arXiv:2306.12231*, July 2023. doi: 10.48550/arXiv.2306.12231. URL http://arxiv.org/abs/2306.12231.

Celeste J. Brown, Sachiko Takayama, Andrew M. Campen, Pam Vise, Thomas W. Marshall, Christopher J. Oldfield, Christopher J. Williams, and A. Keith Dunker. Evolutionary Rate Heterogeneity in Proteins with Long Disordered Regions. *Journal of Molecular Evolution*, 55:104–110, July 2002. ISSN 1432-1432. doi: 10.1007/s00239-001-2309-6. URL https://doi.org/10.1007/s00239-001-2309-6.

Center for High Throughput Computing. Center for High Throughput Computing, 2006. URL https://chtc.cs.wisc.edu/.

Jun Cheng, Guido Novati, Joshua Pan, Clare Bycroft, Akvilė Žemgulytė, Taylor Applebaum, Alexander Pritzel, Lai Hong Wong, Michal Zielinski, Tobias Sargeant, Rosalia G. Schneider, Andrew W. Senior, John Jumper, Demis Hassabis, Pushmeet Kohli, and Žiga Avsec. Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science*, 381(6664):eadg7492, September 2023. doi: 10.1126/science.adg7492. URL https://www.science.org/doi/10.1126/science.adg7492. Publisher: American Association for the Advancement of Science.

J. Dauparas, I. Anishchenko, N. Bennett, H. Bai, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, A. Courbet, R. J. de Haas, N. Bethel, P. J. Y. Leung, T. F. Huddy, S. Pellock, D. Tischer, F. Chan, B. Koepnick, H. Nguyen, A. Kang, B. Sankaran, A. K. Bera, N. P. King, and D. Baker. Robust deep learning based protein sequence design using ProteinMPNN. *Science (New York, N.Y.)*, 378(6615):49–56, October 2022. ISSN 0036-8075. doi: 10.1126/science.add2187. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9997061/.

Mohamed Fawzy and Joseph A. Marsh. Assessing variant effect predictors and disease mechanisms in intrinsically disordered proteins. *bioRxiv*, pp. 2025.04.01.646619, April 2025. doi: 10.1101/2025.04.01.646619. URL https://www.biorxiv.org/content/10.1101/2025.04.01.646619v1.

Alex Fout, Jonathon Byrd, Basir Shariat, and Asa Ben-Hur. Protein Interface Prediction using Graph Convolutional Networks. In *Advances in Neural Information Processing Systems*, Long Beach, CA, USA, 2017. Curran Associates, Inc. URL https://proceedings.neurips.cc/paper/2017/hash/f507783927f2ec2737ba40afbd17efb5-Abstract.html.

Ziqi Gao, Chenran Jiang, Jiawen Zhang, Xiaosen Jiang, Lanqing Li, Peilin Zhao, Huanming Yang, Yong Huang, and Jia Li. Hierarchical graph learning for protein–protein interaction. *Nature Communications*, 14(1):1093, February 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-36736-1. URL https://www.nature.com/articles/s41467-023-36736-1. Publisher: Nature Publishing Group.

Sam Gelman, Bryce Johnson, Chase Freschlin, Arnav Sharma, Sameer D'Costa, John Peters, Anthony Gitter, and Philip A. Romero. Biophysics-based protein language models for protein engineering. *bioRxiv*, pp. 2024.03.15.585128, January 2025. doi: 10.1101/2024.03.15.585128. URL https://www.biorxiv.org/content/10.1101/2024.03.15.585128v2. Pages: 2024.03.15.585128 Section: New Results.

Vladimir Gligorijević, P. Douglas Renfrew, Tomasz Kosciolek, Julia Koehler Leman, Daniel Berenberg, Tommi Vatanen, Chris Chandler, Bryn C. Taylor, Ian M. Fisk, Hera Vlamakis, Ramnik J. Xavier, Rob Knight, Kyunghyun Cho, and Richard Bonneau. Structure-based protein function prediction using graph convolutional networks. *Nature Communications*, 12(1):3168, May 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-23303-9. URL https://www.nature.com/articles/s41467-021-23303-9. Publisher: Nature Publishing Group.

Chloe Hsu, Robert Verkuil, Jason Liu, Zeming Lin, Brian Hie, Tom Sercu, Adam Lerer, and Alexander Rives. Learning inverse folding from millions of predicted structures. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 8946–8970. PMLR, June 2022. URL https://proceedings.mlr.press/v162/hsu22a.html. ISSN: 2640-3498.

Po-Ssu Huang, Yih-En Andrew Ban, Florian Richter, Ingemar Andre, Robert Vernon, William R. Schief, and David Baker. RosettaRemodel: A Generalized Framework for Flexible Backbone Protein Design. *PLOS ONE*, 6(8):e24109, August 2011. ISSN 1932-6203. doi: 10.1371/journal.pone.0024109. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0024109. Publisher: Public Library of Science.

Yufei Huang, Siyuan Li, Lirong Wu, Jin Su, Haitao Lin, Odin Zhang, Zihan Liu, Zhangyang Gao, Jiangbin Zheng, and Stan Z. Li. Protein 3D Graph Structure Learning for Robust Structure-Based Protein Property Prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(11):12662–12670, March 2024. ISSN 2374-3468. doi: 10.1609/aaai.v38i11.29161. URL https://ojs.aaai.org/index.php/AAAI/article/view/29161. Number: 11.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596 (7873):583–589, August 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03819-2. URL https://www.nature.com/articles/s41586-021-03819-2. Publisher: Nature Publishing Group.

Anastasiya V. Kulikova, Daniel J. Diaz, Tianlong Chen, T. Jeffrey Cole, Andrew D. Ellington, and Claus O. Wilke. Two sequence- and two structure-based ML models have learned different aspects of protein biochemistry. *Scientific Reports*, 13(1):13280, August 2023. ISSN 2045-2322. doi: 10.1038/s41598-023-40247-w. URL https://www.nature.com/articles/s41598-023-40247-w.

Elodie Laine, Yasaman Karami, and Alessandra Carbone. GEMME: A Simple and Fast Global Epistatic Model Predicting Mutational Effects. *Molecular Biology and Evolution*, 36(11):2604–2619, November 2019. ISSN 0737-4038. doi: 10.1093/molbev/msz179. URL https://doi.org/10.1093/molbev/msz179.

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, Benjamin Basanta, Brian J. Bender, Kristin Blacklock, Jaume Bonet, Scott E. Boyken, Phil Bradley, Chris Bystroff, Patrick Conway, Seth Cooper, Bruno E. Correia, Brian Coventry, Rhiju Das, René M. De Jong, Frank DiMaio, Lorna Dsilva, Roland Dunbrack, Alexander S. Ford, Brandon Frenz, Darwin Y. Fu, Caleb Geniesse, Lukasz Goldschmidt, Ragul Gowthaman, Jeffrey J. Gray, Dominik Gront, Sharon Guffy, Scott Horowitz, Po-Ssu Huang, Thomas Huber, Tim M. Jacobs, Jeliazko R. Jeliazkov, David K. Johnson, Kalli Kappel, John Karanicolas, Hamed Khakzad, Karen R. Khar, Sagar D. Khare, Firas Khatib, Alisa Khramushin, Indigo C. King, Robert Kleffner, Brian Koepnick, Tanja Kortemme, Georg Kuenze, Brian Kuhlman, Daisuke Kuroda, Jason W. Labonte, Jason K. Lai, Gideon Lapidoth, Andrew Leaver-Fay, Steffen Lindert, Thomas Linsky, Nir London, Joseph H. Lubin, Sergey Lyskov, Jack Maguire, Lars Malmström, Enrique Marcos, Orly Marcu, Nicholas A. Marze, Jens Meiler, Rocco Moretti, Vikram Khipple Mulligan, Santrupti Nerli, Christoffer Norn, Shane Ó'Conchúir, Noah Ollikainen, Sergey Ovchinnikov, Michael S. Pacella, Xingjie Pan, Hahnbeom Park, Ryan E. Pavlovicz, Manasi Pethe, Brian G. Pierce, Kala Bharath Pilla, Barak Raveh, P. Douglas Renfrew, Shourya S. Roy Burman, Aliza Rubenstein, Marion F. Sauer, Andreas Scheck, William Schief, Ora Schueler-Furman, Yuval Sedan, Alexander M. Sevy, Nikolaos G. Sgourakis, Lei Shi, Justin B. Siegel, Daniel-Adriano Silva, Shannon Smith, Yifan Song, Amelie Stein, Maria Szegedy, Frank D. Teets, Summer B. Thyme, Ray Yu-Ruei Wang, Andrew Watkins, Lior Zimmerman, and Richard Bonneau. Macromolecular modeling and design in Rosetta: recent methods and frameworks. *Nature Methods*, 17(7):665–680, July 2020. ISSN 1548-7105. doi: 10.1038/s41592-020-0848-2. URL https://www.nature.com/articles/s41592-020-0848-2. Number: 7 Publisher: Nature Publishing Group.

Sara Light, Rauan Sagit, Oxana Sachenkova, Diana Ekman, and Arne Elofsson. Protein Expansion Is Primarily due to Indels in Intrinsically Disordered Regions. *Molecular Biology and Evolution*, 30(12):2645–2653, December 2013. ISSN 0737-4038. doi: 10.1093/molbev/mst157. URL https://doi.org/10.1093/molbev/mst157.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, March 2023. doi: 10.1126/science.ade2574. URL https://www.science.org/doi/10.1126/science.ade2574. Publisher: American Association for the Advancement of Science.

Kresten Lindorff-Larsen and Birthe B. Kragelund. On the Potential of Machine Learning to Examine the Relationship Between Sequence, Structure, Dynamics and Function of Intrinsically Disordered Proteins. *Journal of Molecular Biology*, 433(20):167196, October 2021. ISSN 0022-2836. doi: 10.1016/j.jmb.2021.167196. URL https://www.sciencedirect.com/science/article/pii/S0022283621004290.

Rongming Liu, Liya Liang, Maria Priscila Lacerda, Emily F. Freed, and Carrie A. Eckert. Chapter 10 - Advances in protein engineering and its application in synthetic biology. In Vijai Singh (ed.), *New Frontiers and Applications of Synthetic Biology*, pp. 147–158. Academic Press, January 2022. ISBN 978-0-12-824469-2. doi: 10.1016/B978-0-12-824469-2.00013-0. URL https://www.sciencedirect.com/science/article/pii/B9780128244692000130.

Debora S. Marks, Lucy J. Colwell, Robert Sheridan, Thomas A. Hopf, Andrea Pagnani, Riccardo Zecchina, and Chris Sander. Protein 3D Structure Computed from Evolutionary Sequence Variation. *PLOS ONE*, 6(12):e28766, December 2011. ISSN 1932-6203. doi: 10.1371/journal.pone.0028766. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0028766. Publisher: Public Library of Science.

Céline Marquet, Michael Heinzinger, Tobias Olenyi, Christian Dallago, Kyra Erckert, Michael Bernhofer, Dmitrii Nechaev, and Burkhard Rost. Embeddings from protein language models predict conservation and variant effects. *Human Genetics*, 141(10):1629–1647,

October 2022. ISSN 1432-1203. doi: 10.1007/s00439-021-02411-y. URL https://doi.org/10.1007/s00439-021-02411-y.

Milot Mirdita, Konstantin Schütze, Yoshitaka Moriwaki, Lim Heo, Sergey Ovchinnikov, and Martin Steinegger. ColabFold: making protein folding accessible to all. *Nature Methods*, 19 (6):679–682, June 2022. ISSN 1548-7105. doi: 10.1038/s41592-022-01488-1. URL https://www.nature.com/articles/s41592-022-01488-1. Publisher: Nature Publishing Group.

Robert W. Newberry, Taylor Arhar, Jean Costello, George C. Hartoularos, Alison M. Maxwell, Zun Zar Chi Naing, Maureen Pittman, Nishith R. Reddy, Daniel M. C. Schwarz, Douglas R. Wassarman, Taia S. Wu, Daniel Barrero, Christa Caggiano, Adam Catching, Taylor B. Cavazos, Laurel S. Estes, Bryan Faust, Elissa A. Fink, Miriam A. Goldman, Yessica K. Gomez, M. Grace Gordon, Laura M. Gunsalus, Nick Hoppe, Maru Jaime-Garza, Matthew C. Johnson, Matthew G. Jones, Andrew F. Kung, Kyle E. Lopez, Jared Lumpe, Calla Martyn, Elizabeth E. McCarthy, Lakshmi E. Miller-Vedam, Erik J. Navarro, Aji Palar, Jenna Pellegrino, Wren Saylor, Christina A. Stephens, Jack Strickland, Hayarpi Torosyan, Stephanie A. Wankowicz, Daniel R. Wong, Garrett Wong, Sy Redding, Eric D. Chow, William F. DeGrado, and Martin Kampmann. Robust Sequence Determinants of α-Synuclein Toxicity in Yeast Implicate Membrane Binding. *ACS Chemical Biology*, 15(8):2137–2153, August 2020a. ISSN 1554-8929. doi: 10.1021/acschembio.0c00339. URL https://doi.org/10.1021/acschembio.0c00339. Publisher: American Chemical Society.

Robert W. Newberry, Jaime T. Leong, Eric D. Chow, Martin Kampmann, and William F. DeGrado. Deep mutational scanning reveals the structural basis for α-synuclein activity. *Nature Chemical Biology*, 16(6):653–659, June 2020b. ISSN 1552-4469. doi: 10.1038/s41589-020-0480-6. URL https://www.nature.com/articles/s41589-020-0480-6. Publisher: Nature Publishing Group.

Pascal Notin. GitHub issue: AlphaFold2 structure generation details, 2025. URL https://github.com/OATML-Markslab/ProteinGym/issues/65#issuecomment-2640138972.

Pascal Notin, Mafalda Dias, Jonathan Frazer, Javier Marchena Hurtado, Aidan N. Gomez, Debora Marks, and Yarin Gal. Tranception: Protein Fitness Prediction with Autoregressive Transformers and Inference-time Retrieval. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 16990–17017. PMLR, June 2022a. URL https://proceedings.mlr.press/v162/notin22a.html. ISSN: 2640-3498.

Pascal Notin, Lood Van Niekerk, Aaron W. Kollasch, Daniel Ritter, Yarin Gal, and Debora Susan Marks. TranceptEVE: Combining Family-specific and Family-agnostic Models of Protein Sequences for Improved Fitness Prediction. In *NeurIPS 2022 Workshop on Learning Meaningful Representations of Life*, November 2022b. URL https://openreview.net/forum?id=l7Oo9DcLmR1.

Pascal Notin, Aaron Kollasch, Daniel Ritter, Lood van Niekerk, Steffanie Paul, Han Spinner, Nathan Rollins, Ada Shaw, Rose Orenbuch, Ruben Weitzman, Jonathan Frazer, Mafalda Dias, Dinko Franceschi, Yarin Gal, and Debora Marks. ProteinGym: Large-Scale Benchmarks for Protein Fitness Prediction and Design. In *Advances in Neural Information Processing Systems*, volume 36, pp. 64331–64379, December 2023. URL https://papers.nips.cc/paper_files/paper/2023/hash/cac723e5ff29f65e3fcbb0739ae91bee-Abstract-Datasets_and_Benchmarks.html.

Steffanie Paul, Aaron W Kollasch, Pascal Notin, and Debora S Marks. Combining Structure and Sequence for Superior Fitness Prediction. In *NeurIPS 2023 Generative AI and Biology (GenBio) Workshop*, 2023.

Roshan M. Rao, Jason Liu, Robert Verkuil, Joshua Meier, John Canny, Pieter Abbeel, Tom Sercu, and Alexander Rives. MSA Transformer. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8844–8856. PMLR, July 2021. URL https://proceedings.mlr.press/v139/rao21a.html. ISSN: 2640-3498.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jerry Ma, and Rob Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, April 2021. doi: 10.1073/pnas.2016239118. URL https://www.pnas.org/doi/full/10.1073/pnas.2016239118. Publisher: Proceedings of the National Academy of Sciences.

Philip A. Romero, Andreas Krause, and Frances H. Arnold. Navigating the protein fitness landscape with Gaussian processes. *Proceedings of the National Academy of Sciences of the United States of America*, 110(3):E193–E201, January 2013. ISSN 0027-8424. doi: 10.1073/pnas.1215251110. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3549130/.

Alan F. Rubin, Jeremy Stone, Aisha Haley Bianchi, Benjamin J. Capodanno, Estelle Y. Da, Mafalda Dias, Daniel Esposito, Jonathan Frazer, Yunfan Fu, Sally B. Grindstaff, Matthew R. Harrington, Iris Li, Abbye E. McEwen, Joseph K. Min, Nick Moore, Olivia G. Moscatelli, Jesslyn Ong, Polina V. Polunina, Joshua E. Rollins, Nathan J. Rollins, Ashley E. Snyder, Amy Tam, Matthew J. Wakefield, Shenyi Sunny Ye, Lea M. Starita, Vanessa L. Bryant, Debora S. Marks, and Douglas M. Fowler. MaveDB 2024: a curated community database with over seven million variant effects from multiplexed functional assays. *Genome Biology*, 26(1):13, January 2025. ISSN 1474-760X. doi: 10.1186/s13059-025-03476-y. URL https://doi.org/10.1186/s13059-025-03476-y.

Yifan Song, Frank DiMaio, Ray Yu-Ruei Wang, David Kim, Chris Miles, TJ Brunette, James Thompson, and David Baker. High-Resolution Comparative Modeling with RosettaCM. *Structure*, 21(10):1735–1742, October 2013. ISSN 0969-2126. doi: 10.1016/j.str.2013.08.005. URL https://www.sciencedirect.com/science/article/pii/S0969212613002979.

Martin Steinegger and Johannes Söding. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, 35(11):1026–1028, November 2017. ISSN 1546-1696. doi: 10.1038/nbt.3988. URL https://www.nature.com/articles/nbt.3988. Publisher: Nature Publishing Group.

Jin Su, Chenchen Han, Yuyang Zhou, Junjie Shan, Xibin Zhou, and Fajie Yuan. SaProt: Protein Language Modeling with Structure-aware Vocabulary. In *The Twelfth International Conference on Learning Representations*, October 2023. URL https://openreview.net/forum?id=6MRm3G4NiU.

Yang Tan, Bingxin Zhou, Lirong Zheng, Guisheng Fan, and Liang Hong. Semantical and Geometrical Protein Encoding Toward Enhanced Bioactivity and Thermostability. *eLife*, 13, June 2024. doi: 10.7554/eLife.98033.1. URL https://elifesciences.org/reviewed-preprints/98033. Publisher: eLife Sciences Publications Limited.

The UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research*, 51(D1):D523–D531, January 2023. ISSN 0305-1048. doi: 10.1093/nar/gkac1052. URL https://doi.org/10.1093/nar/gkac1052.

Rakesh Trivedi and Hampapathalu Adimurthy Nagarajaram. Intrinsically disordered proteins: An overview. *International Journal of Molecular Sciences*, 23(22), 2022. ISSN 1422-0067. doi: 10.3390/ijms232214050. URL https://www.mdpi.com/1422-0067/23/22/14050.

Kotaro Tsuboyama, Justas Dauparas, Jonathan Chen, Elodie Laine, Yasser Mohseni Behbahani, Jonathan J. Weinstein, Niall M. Mangan, Sergey Ovchinnikov, and Gabriel J. Rocklin. Mega-scale experimental analysis of protein folding stability in biology and design. *Nature*, 620(7973):434–444, August 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06328-6. URL https://www.nature.com/articles/s41586-023-06328-6. Publisher: Nature Publishing Group.

Marcus D. Tuttle, Gemma Comellas, Andrew J. Nieuwkoop, Dustin J. Covell, Deborah A. Berthold, Kathryn D. Kloepper, Joseph M. Courtney, Jae K. Kim, Alexander M. Barclay, Amy Kendall, William Wan, Gerald Stubbs, Charles D. Schwieters, Virginia M. Y. Lee, Julia M. George, and Chad M. Rienstra. Solid-state NMR structure of a pathogenic fibril of full-length human α-synuclein. *Nature Structural & Molecular Biology*, 23(5):409–415,

May 2016. ISSN 1545-9985. doi: 10.1038/nsmb.3194. URL https://www.nature.com/articles/nsmb.3194. Publisher: Nature Publishing Group.

Tobias S. Ulmer, Ad Bax, Nelson B. Cole, and Robert L. Nussbaum. Structure and Dynamics of Micelle-bound Human α-Synuclein. *Journal of Biological Chemistry*, 280 (10):9595–9603, March 2005. ISSN 0021-9258. doi: 10.1074/jbc.M411805200. URL https://www.sciencedirect.com/science/article/pii/S0021925819629790.

Michel van Kempen, Stephanie S. Kim, Charlotte Tumescheit, Milot Mirdita, Jeongjae Lee, Cameron L. M. Gilchrist, Johannes Söding, and Martin Steinegger. Fast and accurate protein structure search with Foldseek. *Nature Biotechnology*, 42(2):243–246, February 2024. ISSN 1546-1696. doi: 10.1038/s41587-023-01773-0. URL https://www.nature.com/articles/s41587-023-01773-0. Publisher: Nature Publishing Group.

Mihaly Varadi, Damian Bertoni, Paulyna Magana, Urmila Paramval, Ivanna Pidruchna, Malarvizhi Radhakrishnan, Maxim Tsenkov, Sreenath Nair, Milot Mirdita, Jingi Yeo, Oleg Kovalevskiy, Kathryn Tunyasuvunakool, Agata Laydon, Augustin Žídek, Hamish Tomlinson, Dhavanthi Hariharan, Josh Abrahamson, Tim Green, John Jumper, Ewan Birney, Martin Steinegger, Demis Hassabis, and Sameer Velankar. AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic Acids Research*, 52(D1):D368–D375, January 2024. ISSN 0305-1048. doi: 10.1093/nar/gkad1011. URL https://doi.org/10.1093/nar/gkad1011.

Connie Y. Wang, Paul M. Chang, Marie L. Ary, Benjamin D. Allen, Roberto A. Chica, Stephen L. Mayo, and Barry D. Olafson. ProtaBank: A repository for protein design and engineering data. *Protein Science*, 27(6):1113–1124, 2018. ISSN 1469-896X. doi: 10.1002/pro.3406. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/pro.3406.

Ying Xia, Chun-Qiu Xia, Xiaoyong Pan, and Hong-Bin Shen. GraphBind: protein structural context embedded rules learned by hierarchical graph neural networks for recognizing nucleic-acid-binding residues. *Nucleic Acids Research*, 49(9):e51, May 2021. ISSN 0305-1048. doi: 10.1093/nar/gkab044. URL https://doi.org/10.1093/nar/gkab044.

Jason Yang, Ravi G. Lal, James C. Bowden, Raul Astudillo, Mikhail A. Hameedi, Sukhvinder Kaur, Matthew Hill, Yisong Yue, and Frances H. Arnold. Active learning-assisted directed evolution. *Nature Communications*, 16(1):714, January 2025. ISSN 2041-1723. doi: 10.1038/s41467-025-55987-8. URL https://www.nature.com/articles/s41467-025-55987-8. Publisher: Nature Publishing Group.

Kevin K. Yang, Zachary Wu, and Frances H. Arnold. Machine-learning-guided directed evolution for protein engineering. *Nature Methods*, 16(8):687–694, August 2019. ISSN 1548-7105. doi: 10.1038/s41592-019-0496-6. URL https://www.nature.com/articles/s41592-019-0496-6. Publisher: Nature Publishing Group.

Kevin K Yang, Niccolò Zanichelli, and Hugh Yeh. Masked inverse folding with sequence transfer for protein representation learning. *Protein Engineering, Design and Selection*, 36:gzad015, January 2023. ISSN 1741-0126. doi: 10.1093/protein/gzad015. URL https://doi.org/10.1093/protein/gzad015.

Kevin K. Yang, Nicolo Fusi, and Alex X. Lu. Convolutions are competitive with transformers for protein sequence pretraining. *Cell Systems*, 15(3):286–294.e2, March 2024. ISSN 2405-4712, 2405-4720. doi: 10.1016/j.cels.2024.01.008. URL https://www.cell.com/cell-systems/abstract/S2405-4712(24)00029-2. Publisher: Elsevier.

Christine Zardecki, Shuchismita Dutta, David S. Goodsell, Maria Voigt, and Stephen K. Burley. RCSB Protein Data Bank: A Resource for Chemical, Biochemical, and Structural Explorations of Large and Small Biomolecules. *Journal of Chemical Education*, 93(3): 569–575, March 2016. ISSN 0021-9584. doi: 10.1021/acs.jchemed.5b00404. URL https://doi.org/10.1021/acs.jchemed.5b00404. Publisher: American Chemical Society.

Yang Zhang and Jeffrey Skolnick. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research*, 33(7):2302–2309, April 2005. ISSN 0305-1048. doi: 10.1093/nar/gki524. URL https://doi.org/10.1093/nar/gki524.

Zhidian Zhang, Hannah K. Wayment-Steele, Garyk Brixi, Haobo Wang, Dorothee Kern, and Sergey Ovchinnikov. Protein language models learn evolutionary statistics of interacting sequence motifs. *Proceedings of the National Academy of Sciences*, 121(45):e2406285121, November 2024. doi: 10.1073/pnas.2406285121. URL https://www.pnas.org/doi/10.1073/pnas.2406285121. Publisher: Proceedings of the National Academy of Sciences.

Appendix

Datasets

ProteinGym (Notin et al., 2023) is one of the most comprehensive datasets available for evaluating zero-shot protein fitness prediction. It presents a collection of DMS substitution assays measuring various protein functions: activity, binding, expression, organismal fitness, and stability (Notin et al., 2023). Furthermore, the dataset also contains different input data modalities for all the proteins. We used 216 DMS substitution assays from ProteinGym (Table 3). We did not use BRCA2_HUMAN_Erwood_2022_HEK293T. This is because its structure is split across multiple files, and MSA generation for SSEmb proved to be time-consuming.

In order to compute scores for SSemb (Blaabjerg et al., 2024), however, we had to use MMseqs2 (Steinegger & Söding, 2017; Mirdita et al., 2022) to create new MSAs. We used the notebook provided by SSemb to serially generate an alignment for each protein in the ProteinGym DMS substitutions dataset by processing each pdb file from ProteinGym.

To get experimental structures for each assay in ProteinGym we queried the RCSB PDB (Zardecki et al., 2016). Details on the query parameters and selection process can be found below. We ended up with 65 different assays from the ProteinGym benchmark with matching experimental structures. These assays are shown in Figure 1, and their functions are summarized in Table 3. We aligned experimental and predicted structures with the RCSB PDB pairwise structure alignment tool (Bittrich et al., 2024) and the TM-align alignment (Zhang & Skolnick, 2005).

We obtained the disorder content from the 2024_12 DisProt release (Aspromonte et al., 2024). We mapped the ProteinGym UniProt (The UniProt Consortium, 2023) entry names to IDs (that is, accession numbers), replacing the outdated PSAE_SYNP2 with PSAE_PICP2. ANCSZ represents an ancestral sequence, not a UniProt ID, and could not be mapped, so we obtained 186 unique UniProt IDs. We intersected these with all UniProt IDs from the DisProt database.

Protein fitness prediction algorithms

Score for algorithms marked with an asterisk (*) were taken from the ProteinGym benchmark. These scores can be obtained by downloading the zero shot substitution scores provided on `https://proteingym.org`. For all other models, scores were generated manually by running inference on the ProteinGym zero shot substitution benchmark.

- **VESPA* (Marquet et al., 2022):** This model uses protein sequence as an input. It forgoes the need for MSA by using the embeddings from a pretrained PLM along to compute conservation information along with log-probability odds for variants to estimate their fitness effect.
- **GEMME* (Laine et al., 2019):** This model operates on the multiple sequence alignment (MSA) of a protein as input. The key idea is to present sequences homologous to the query as a tree (by sorting on sequence similarities) and extracting conserved residues. This allows the model to predict protein functions based on evolutionary signal for each protein.
- **ESM Inverse Folding* (Hsu et al., 2022):** This model proposes the idea of using a structural encoder in addition to a sequence-to-sequence transformer model in order to consume a protein's folded structure and corrupted sequence to predict the original sequence from it. We chose this model instead of ProteinMPNN (Dauparas et al., 2022) because both the ProteinGym benchmark (Notin et al., 2023) as well as prior work (Paul et al., 2023) have noted that ESM inverse folding provides a stronger baseline representation for structure-based models.
- **ProtSSN* (Tan et al., 2024)** This model uses both protein sequence and structure as input. The model uses a graph neural network to process a protein's structure graph representation. The embeddings from a PLM are used as initial node embeddings and processed over the structure graph.

- **TranceptEVE L* (Notin et al., 2022b):** TranceptEVE combines signal predictive of protein function from the protein's MSA (using EVE and Tranception's MSA retrieval (Notin et al., 2022a)) well as sequence (Tranception's autoregressive transformer). The idea behind TranceptEVE is that sequence and alignment signals are conditionally independent and can be combined as follows:

$$\log \mathbb{P}(x_i|x_{1:i-1}) = (1-\alpha_P)[(1-\beta_P)\mathbb{P}_{\text{Transformer}}(x_i|x_{1:i-1}) + \beta_P \mathbb{P}_{\text{MSA}}(x_i|x_{1:i-1})])$$

$$+\alpha_P \mathbb{P}_{\text{EVE}}(x_i|x_{1:i-1})$$

- **SSemb (Blaabjerg et al., 2024):** SSEmb proposes using a graph neural network to consume protein structure graph and use embeddings from the MSA Transformer model in order to make predictions.
- **SaProt* (Su et al., 2023):** SaProt uses Foldseek (van Kempen et al., 2024) to generate protein residue tokens from its 3D structure. It then uses a PLM to process the tokens and compute probabilities of the masked tokens within a context.
- **ESM2 650M (Lin et al., 2023)**: Is a protein language model trained on protein sequence datasets using masked language modeling. It uses a BERT-style encoder architecture to learn continuous embedding representations of protein sequences.
- **Ensemble 1 (StructSeq)**: This ensemble described in (Paul et al., 2023) assumes that the probability of a masked token belonging to each amino acid class based on its structure and alignment as well as that from a PLM are mutually independent. Therefore, the three log probabilities can be added to compute the total probability. This model combines log probabilities from ESM inverse folding and TranceptEVE. Unlike other ensembling strategies (Kulikova et al., 2023), there is no additional training after combining the base models.
- **Ensemble 2**: We ensemble scores from Tranception, which is a multi-modal model combining PLM and MSA signals and ESM-IF1, which is a structure-based model. We chose this model due to its similarity to Ensemble 1.
- **Ensemble 3**: We ensemble scores from GEMME, which is a MSA-based model, and ESM inverse folding, which is a structure-based model.
- **Ensemble 4**: We ensemble scores from VESPA, which is a PLM, and ESM inverse folding.

Evaluation metrics

Spearman ρ can be defined as:

$$\rho = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N(N^2-1)}$$

where each d_i is a rank difference between i^{th} prediction and i^{th} ground truth data point.

This metric is especially useful when judging how well a model's predictions correlate with a certain quantitative function in a protein. The average value indicates how well the model predicts all protein functions on average.

Top 10 Recall can be defined as:

$$\text{Top-10 Recall} = \frac{\text{No. of data points that lie in the top 10\% of both predicted and measured values}}{\text{No. of data points in the top 10\% of measured values}}$$

The utility of this metric is scenarios where we use a machine learning model to identify mutations in a protein that increase a certain function the most, i.e. identifying the most beneficial mutations. That setting is relevant for protein engineering.

Selection process for disordered proteins

We follow a multi-step process to select proteins containing IDRs from the set of proteins in ProteinGym. First, we select disordered annotations according to protein UniProt IDs present in ProteinGym. Then, we perform a sequence matching operation. We need to do this due to DisProt using protein sequences from UniProt (Aspromonte et al., 2024). This convention is different from ProteinGym, where each DMS assay is mapped to the sequence used to generate the DMS data (Notin et al., 2023). We found that target sequences in ProteinGym can often be different from the UniProt sequences at arbitrary positions. For example, the sequences for α-synuclein in UniProt and ProteinGym are of the same length but differ in residues at arbitrary locations throughout the sequence. We detail the sequence alignment procedure below. Lastly, we select mutations within substitution DMS assays in ProteinGym.

- **1st Stage:** We use the DisProt database to query each UniProt ID in ProteinGym and retrieve the set of proteins that have disordered annotations in DisProt.
- **2nd Stage:**
  - We first align the ProteinGym target sequence for each assay to its UniProt sequence. We do this by handling two cases: sequences differing in arbitrary mutations and target sequences that are subsets of UniProt sequences. If target sequences are of the same length as their corresponding UniProt sequences, we treat them as referring to the same sequence. If the target sequence and UniProt sequence differ in length (the case where target sequences are a subset of the UniProt sequence), we find the location of the target sequence within the UniProt sequence.
  - Using these aligned annotations, we select disordered annotations that lie within the target sequence provided by ProteinGym for each UniProt ID (Notin et al., 2023). At the end of this stage, we are left with DMS assays from ProteinGym that contain at least one disordered region.
- **3rd Stage:** We select DMS assays that have mutations in both ordered and disordered regions.

Stage 1 yields 58 of the 186 (31%) unique UniProt IDs in ProteinGym that are proteins that are annotated as having disordered regions in DisProt (Table 4). These proteins are associated with 69 of the 217 (32%) ProteinGym DMS substitution assays.

After Stage 2, we obtain 52 of the 186 (28%) unique UniProt IDs in ProteinGym associated with 63 of the 217 (29%) of the DMS substitution assays. These assays contain a valid disordered region in Table 4. We use the subset of assays obtained after Stage 2 in our analysis in Figure 2 for benchmarking how different models score proteins which contain disordered regions.

Lastly, in order to benchmark how well models predict functional effects of mutations in ordered versus disordered regions, we select assays that contain mutations in both ordered and disordered regions during Stage 3. After this stage, we are left with 36 unique UniProt IDs corresponding to 43 different assays. The set of assays obtained after Stage 3 are marked with a † in Table 4 and are used in Figure 3.

Selection process for experimental structures

We queried the RCSB PDB (Bittrich et al., 2024) to get structures that exactly match the entire target sequence in ProteinGym. We used the logic to query PDB by performing a search using sequence similarities from the code provided by Boca & Mathis (2023).

Our filtering process follows two steps:

- As our first filtering criteria, we discard proteins that do not have any matching experimental structures. We used an E-value cutoff of 1.0 and identity cutoff of 90%. We found 82 such experimental structures.

- Our second filtering criteria is to drop any proteins that match to experimental structures that do not contain coordinates for all residues in the ProteinGym target sequence. This is due to the fact that the ESM-IF1 model would not be able to process such structures.

A limitation of our approach is that we did not manually confirm that the selected structures match the functional assays in the ProteinGym benchmark. This can be detrimental to making fitness predictions. For example, if an experimental structure represents a conformation that does not expose an active site, it will be less informative in predicting protein function.

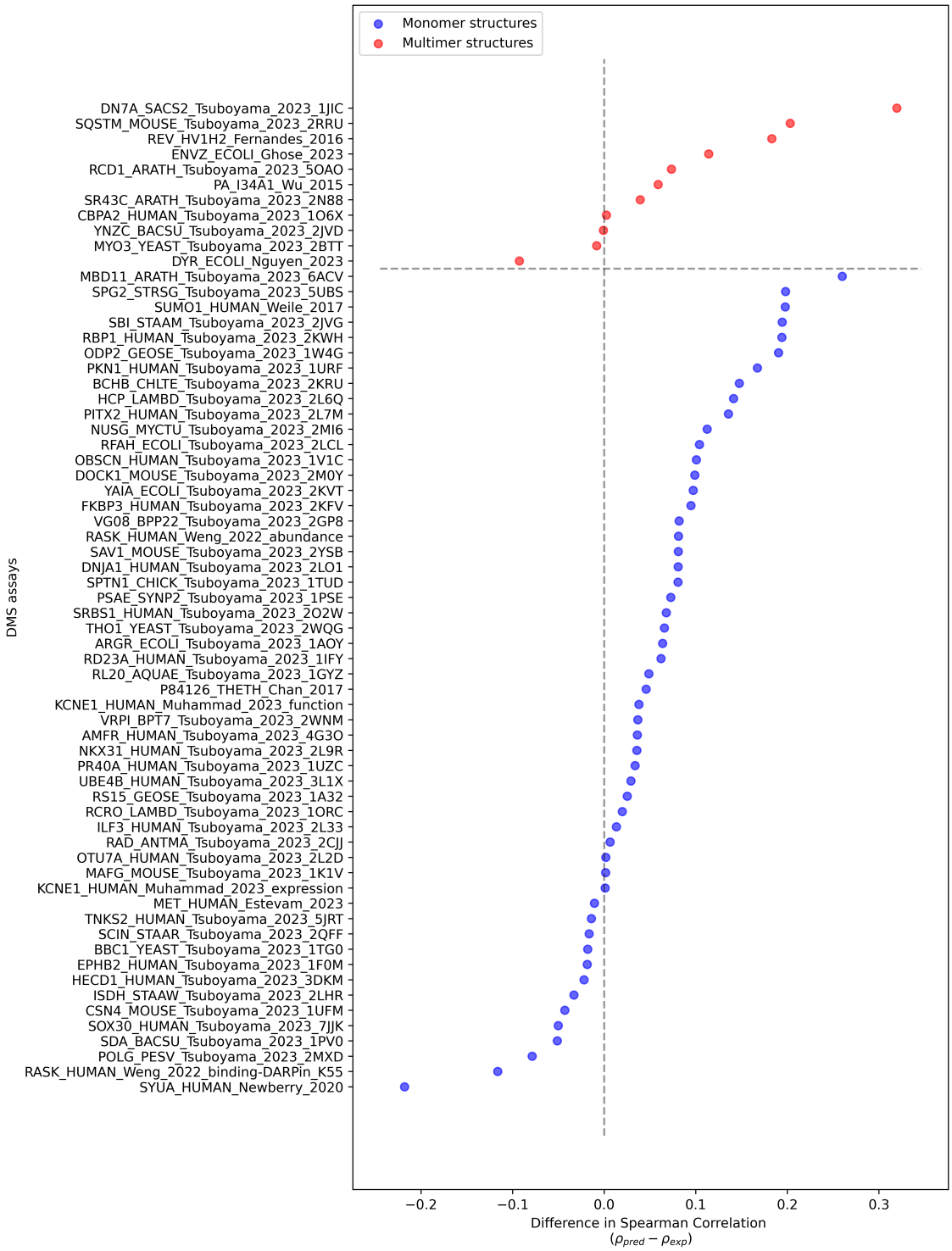


Figure 1: Difference in Spearman correlation between predictions made using predicted and experimental structures. Each datapoint represents a DMS assay for which an experimental structure exists.

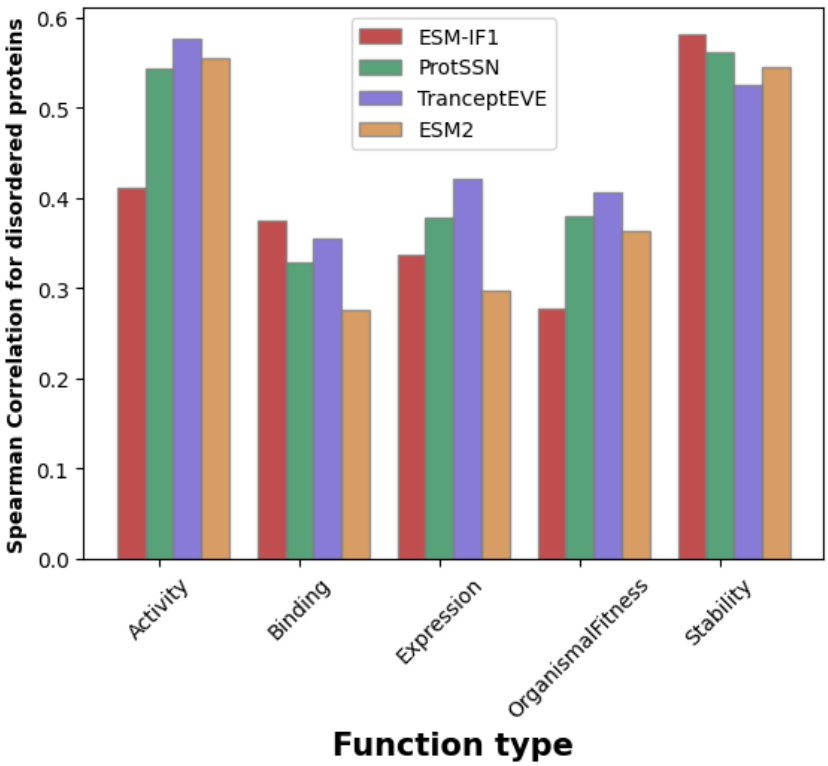
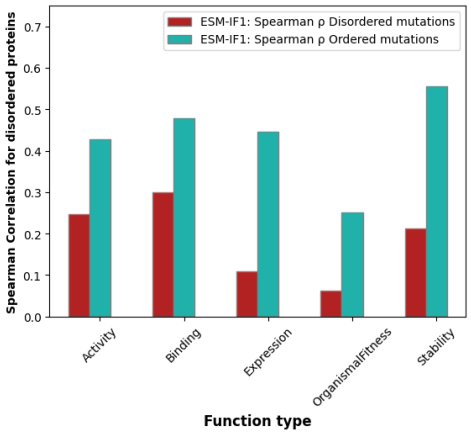
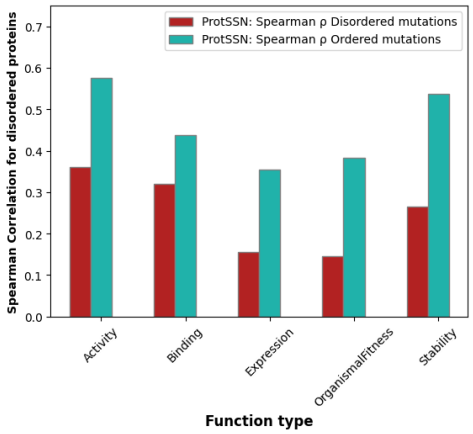


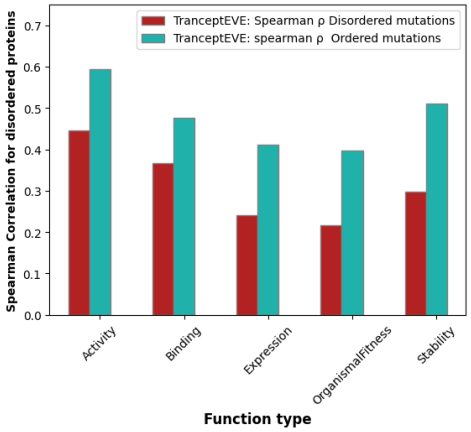
Figure 2: Protein-level Spearman ρ predicting function of proteins comprising some form of disordered regions in the ProteinGym assay target sequence according to the DisProt database.



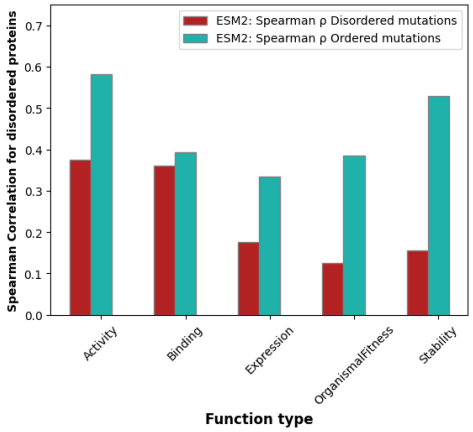
(a) ESM-IF1 spearman ρ



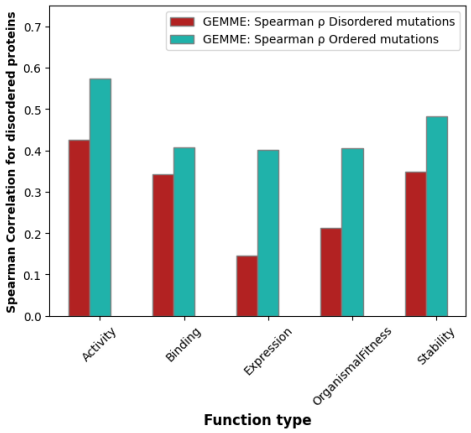
(b) ProtSSN spearman ρ



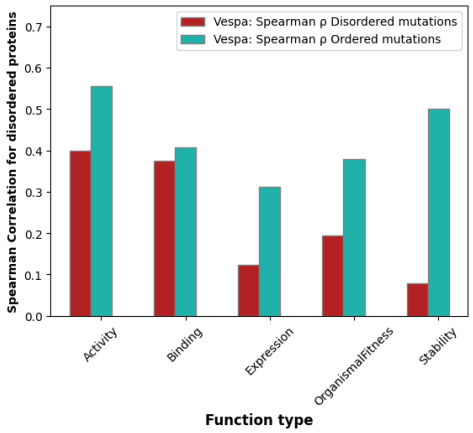
(c) TranceptEVE L spearman ρ



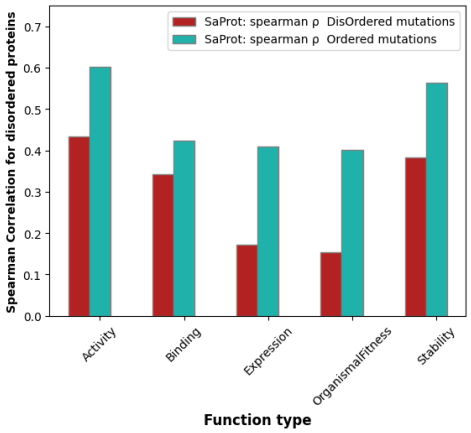
(d) ESM2 650M spearman ρ



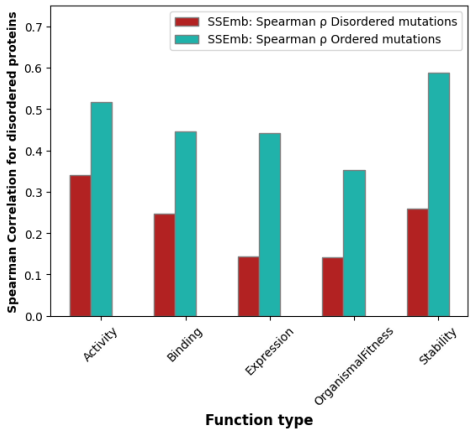
(e) GEMME spearman ρ



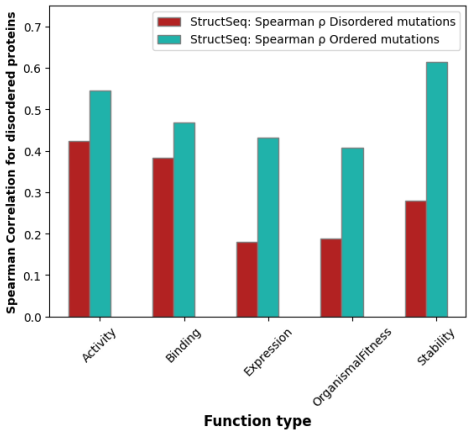
(f) Vespa spearman ρ



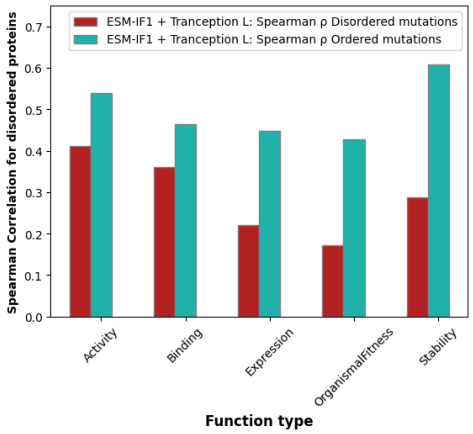
(g) SaProt 650M spearman ρ



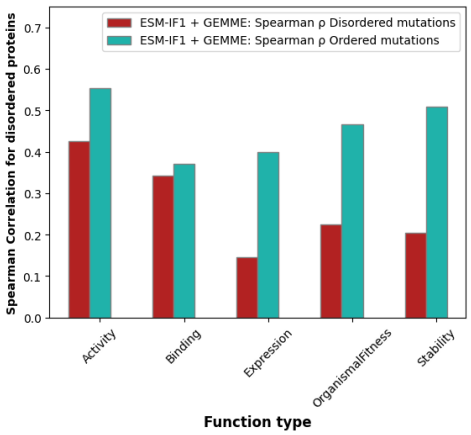
(h) SSEmb spearman ρ



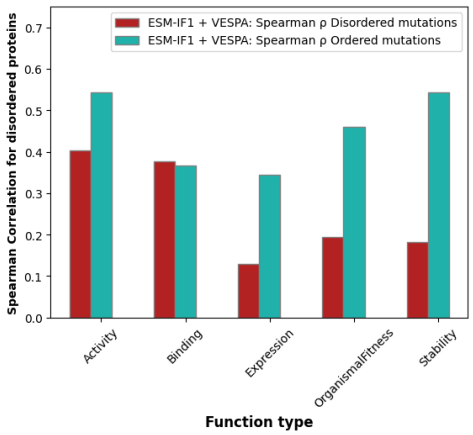
(i) StructSeq Ensemble spearman ρ



(j) ESM-IF1 + Tranception Ensemble spearman ρ

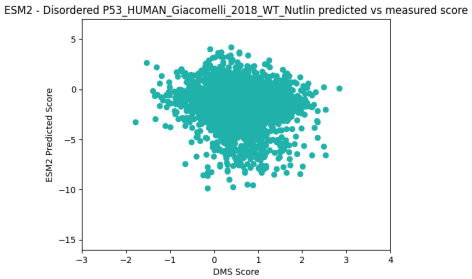


(k) ESM-IF1 + GEMME Ensemble spearman ρ

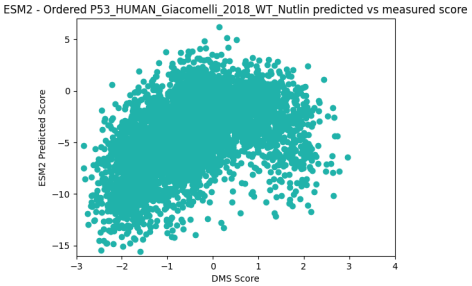


(l) ESM-IF1 + Vespa Ensemble spearman ρ

Figure 3: Spearman correlation averaged across UniProtID and then function type for disordered and ordered regions. These scores were computed across proteins that contain mutations in both disordered and ordered regions.



(a) ESM2 predictions for mutations in the disordered regions



(b) ESM2 predictions for mutations in the ordered regions

Figure 4: Predictions made by ESM2 650M on the P53_HUMAN_Giacomelli_2018_WT_Nutlin dataset separated by disordered and ordered regions.

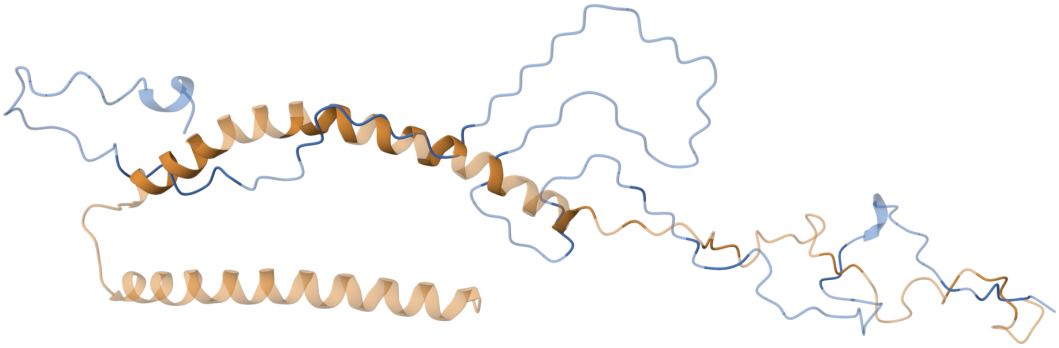


Figure 5: Protein structure alignment between the experimental (orange, PDB 1XQ8 (Ulmer et al., 2005)) and predicted (blue) structures for P37840 (human α-synuclein). The ProteinGym AlphaFold 2 predicted structure resembles an α-synuclein fibril, PDB 2N0A (Tuttle et al., 2016), which is a different conformation than the predicted α-synuclein structure in the AlphaFold Protein Structure Database (AF-P37840-F1-v4) (Varadi et al., 2024).

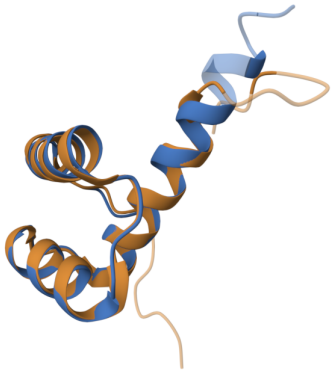


Figure 6: Protein structure alignment between the experimental (orange, PDB 2L9R) and predicted (blue) structures for Q99801 (human NKX3-1). The experimental structure has sequence length of 69 versus 61 for the predicted structure.

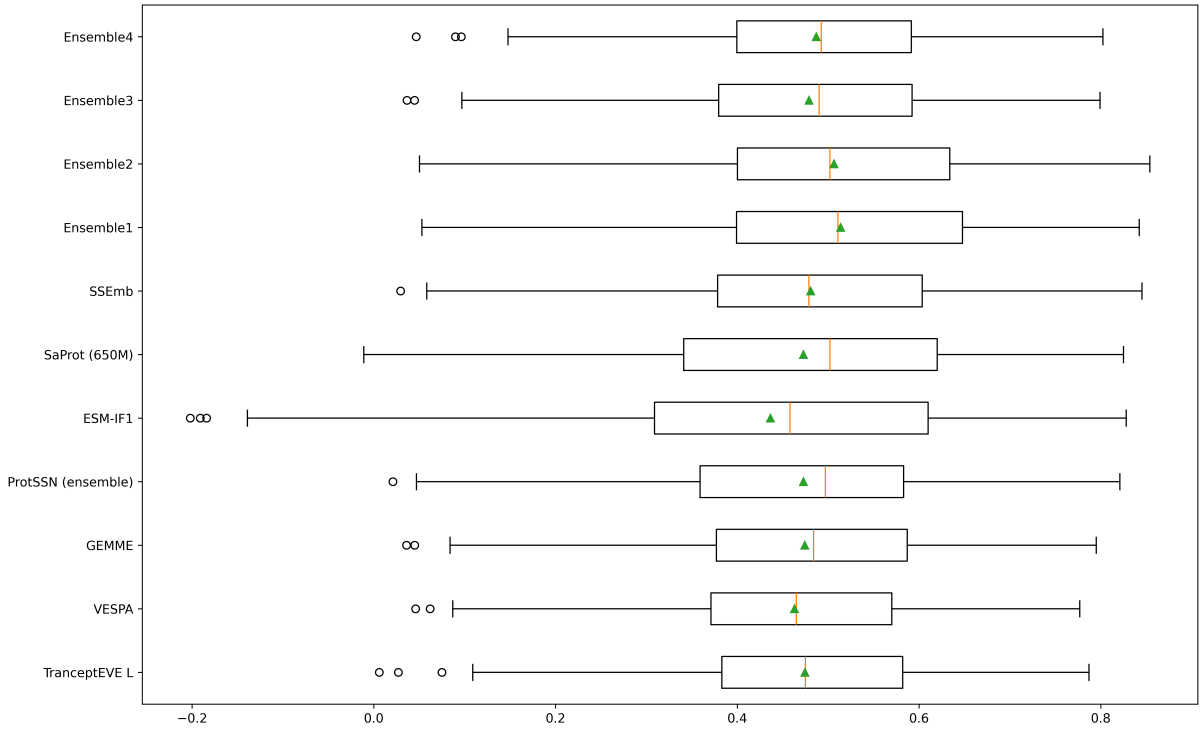


Figure 7: Distribution of Spearman correlation of various models being considered across 216 DMS assays from ProteinGym. Triangles denote the mean correlation. Ensemble1 is the StructSeq method proposed by (Paul et al., 2023).

Assay function type	Activity	Binding	Expression	Organismal	Stability	**Total**
Assays with predicted structures	43	13	18	76	66	216
Assays with experimental structures	3	1	2	6	53	65

Table 3: Count of DMS assays by function type.

ProteinGym DMS ID	**UniProt**	**Disorder content (%)**	**Disordered regions**
A4_HUMAN_Seuma_2022†	P05067	5.19	671-710
ACE2_HUMAN_Chan_2020	Q9BYF1	4.60	768-804
ADRB2_HUMAN_Jones_2020†	P07550	17.43	341-412
B2L11_HUMAN_Dutta_2010_binding-Mcl-1	O43521	55.56	0-40,96-164
BRCA1_HUMAN_Findlay_2018†	P38398	83.20	99-1648
BRCA2_HUMAN_Erwood_2022_HEK293T	P51587	6.93	47-283
CALM1_HUMAN_Weile_2017†	P0DP23	6.71	75-84
CASP3_HUMAN_Roychowdhury_2020†	P42574	16.28	28-37,172-182,136-156
CASP7_HUMAN_Roychowdhury_2020†	P55210	5.34	173-187
CATR_CHLRE_Tsuboyama_2023_2AMI†	P05434	4.17	0-2
CBS_HUMAN_Sun_2020†	P35520	7.26	0-39
CD19_HUMAN_Klesmith_2019_FMC_singles†	P15391	2.70	137-151
CUE1_YEAST_Tsuboyama_2023_2MYX†	P38428	9.62	0-4
DLG4_HUMAN_Faure_2021	P78352	9.81	0-70
DYR_ECOLI_Nguyen_2023†	P0ABQ4	27.04	8-23,62-71,115-131
DYR_ECOLI_Thompson_2019†	P0ABQ4	27.04	8-23,62-71,115-131
ERBB2_HUMAN_Elazar_2016	P04626	21.35	987-1254
GAL4_YEAST_Kitzman_2015	P04386	4.31	106-143
GCN4_YEAST_Staller_2018†	P03069	47.69	0-133
HMDH_HUMAN_Jiang_2019†	P04035	3.15	860-887
KCNE1_HUMAN_Muhammad_2023_expression†	P15382	50.39	23-44,71-90,106-128
KCNE1_HUMAN_Muhammad_2023_function†	P15382	50.39	23-44,71-90,106-128
MET_HUMAN_Estevam_2023†	P08581	6.97	166-185
MK01_HUMAN_Brenan_2016†	P28482	8.33	0-14,174-188
MTHR_HUMAN_Weile_2021†	P42898	1.68	160-170
NKX31_HUMAN_Tsuboyama_2023_2L9R	Q99801	0	Disordered region not in target seq.
NPC1_HUMAN_Erwood_2022_HEK293T†	O15118	12.83	287-332,373-382,604-619, 772-812,959-982,1251-1277
NPC1_HUMAN_Erwood_2022_RPE1	O15118	12.83	287-332,373-382,604-619, 772-812,959-982,1251-1277
NUSA_ECOLI_Tsuboyama_2023_1WCL	P0AFF6	100.0	0-68
P53_HUMAN_Giacomelli_2018_Null_Etoposide†	P04637	37.66	0-92,290-311,360-392
P53_HUMAN_Giacomelli_2018_Null_Nutlin†	P04637	37.66	0-92,290-311,360-392
P53_HUMAN_Giacomelli_2018_WT_Nutlin†	P04637	37.66	0-92,290-311,360-392
P53_HUMAN_Kotler_2018†	P04637	37.66	0-92,290-311,360-392
PABP_YEAST_Melamed_2013	P04147	14.56	418-501
PAI1_HUMAN_Huttinger_2021†	P05121	3.98	354-369
PA_I34A1_Wu_2015†	P03433	3.63	371-396
PIN1_HUMAN_Tsuboyama_2023_1I6C†	Q13526	15.38	33-38
POLG_DEN26_Suphatrakul_2023	P29990	0	Disordered region not in target seq.

POLG_HCVJF_Qi_2014	Q99IB8	6.73	2223-2316,2330-2439
PPARG_HUMAN_Majithia_2016†	P37231	59.21	8-210,237-297,442-476
PRKN_HUMAN_Clausen_2023†	O60260	16.13	72-98,108-141,377-390
PTEN_HUMAN_Matreyek_2021†	P60484	18.61	285-308,352-402
PTEN_HUMAN_Mighell_2018†	P60484	18.61	285-308,352-402
R1AB_SARS2_Flynn_2022	P0DTD1	0	Disordered region not in target seq.
RAF1_HUMAN_Zinkus-Boltz_2019	P04049	3.09	235-254
RASH_HUMAN_Bandaru_2017	P01112	10.58	169-188
RASK_HUMAN_Weng_2022_abundance†	P01116	17.55	59-69,166-187
RASK_HUMAN_Weng_2022_binding-DARPin_K55†	P01116	17.55	59-69,166-187
RCD1_ARATH_Tsuboyama_2023_5OAO†	Q8RY59	1.75	0-0
RCRO_LAMBD_Tsuboyama_2023_1ORC†	P03040	15.87	53-62
RD23A_HUMAN_Tsuboyama_2023_1IFY†	P54725	9.09	40-43
RFAH_ECOLI_Tsuboyama_2023_2LCL†	P0AFW0	12.73	0-6
RL20_AQUAE_Tsuboyama_2023_1GYZ†	O67086	49.15	0-0, 31-58
SHOC2_HUMAN_Kwon_2022†	Q9UQ13	14.60	0-84
SPIKE_SARS2_Starr_2020_binding	P0DTC2	39.59	0-25,66-79,141-163,172-184, 245-261,318-540,620-639, 672-686,827-852,1146-1272
SPIKE_SARS2_Starr_2020_expression	P0DTC2	39.59	0-25,66-79,141-163,172-184, 245-261,318-540,620-639, 672-686,827-852,1146-1272
SRBS1_HUMAN_Tsuboyama_2023_2O2W	Q9BX66	0	Disordered region not in target seq.
SRC_HUMAN_Ahler_2019	P12931	15.67	0-83
SRC_HUMAN_Chakraborty_2023_binding-DAS_25uM	P12931	15.67	0-83
SRC_HUMAN_Nguyen_2022	P12931	15.67	0-83
SUMO1_HUMAN_Weile_2017†	P63165	17.82	0-17
SYUA_HUMAN_Newberry_2020	P37840	100.0	0-139
TADBP_HUMAN_Bolognesi_2019	Q13148	36.71	262-413
TAT_HV1BR_Fernandes_2016	P04610	100.0	0-85
UBE4B_HUMAN_Tsuboyama_2023_3L1X†	O95155	1.45	0-0
UBR5_HUMAN_Tsuboyama_2023_1I2T	O95071	0	Disordered region not in target seq.
VILI_CHICK_Tsuboyama_2023_1YU5†	P02640	3.08	0-1
YAP1_HUMAN_Araya_2012†	P46937	24.21	49-170
YNZC_BACSU_Tsuboyama_2023_2JVD	O31818	0	Disordered region not in target seq.

Table 4: DisProt disorder content for ProteinGym proteins and the associated ProteinGym substitution DMS assays. The disorder content and 0-indexed disordered regions pertain only to the subset of the protein sequence covered by the DMS assay (target sequence). Disordered regions for proteins with disordered annotation outside of the target sequence are marked as such. Additionally, DMS assays used in Figure 3 containing mutations in both ordered and disordered regions marked with (†).

Assay function type	Activity	Binding	Expression	Organismal	Stability	**Total**
DMS ID Count	9	4	4	15	11	**43**

Table 5: Count of DMS IDs corresponding to disordered proteins identified by 36 unique UniProt IDs that have mutations in both ordered and disordered regions.

Model	Modality	Top 10 Recall (R) by Assay Function					Avg. Top 10 R
		Activity	Binding	Expression	Organismal Fitness	Stability	
ESM-IF1*	Structure	0.180	0.207	0.205	0.173	0.344	0.222
GEMME*	MSA	0.194	0.200	0.198	0.215	0.233	0.208
VESPA*	PLM	0.175	0.192	0.184	0.208	0.242	0.200
ProtSSN* (ensemble)	Structure + PLM	0.187	0.197	0.220	0.191	0.337	0.227
SSemb	Structure + MSA	0.198	0.222	0.204	0.215	0.334	0.235
TranceptEVE L*	MSA + PLM	0.197	0.221	0.224	0.229	0.278	0.230
SaProt (650M)*	Structure + PLM	0.187	0.232	0.236	0.175	0.331	0.232
Ensemble 1 StructSeq	Structure + MSA + PLM	0.208	0.238	0.232	0.221	0.335	0.247
Ensemble 2 ESM-IF1, Tranception	Structure + MSA + PLM	0.206	0.224	0.221	0.221	0.328	0.240
Ensembled 3 ESM-IF1, GEMME	Structure + MSA	0.197	0.199	0.198	0.223	0.244	0.212
Ensemble 4 ESM-IF1, VESPA	Structure + PLM	0.184	0.194	0.186	0.213	0.278	0.211

Table 6: Top 10 Recall of predictions made by different models sorted according to function type of each DMS assay. Scores for model names annotated with an asterisk (*) were calculated using their predictions provided in the ProteinGym benchmark (Notin et al., 2023). Scores highlighted in red, olive, blue represent the 1st, 2nd, and 3rd highest score for each group. Modalities represented in the table are protein language models (PLMs), structure-based models (Structure), Multiple Sequence Alignment (MSA).