Combining Structure and Sequence for Superior Fitness Prediction

Steffanie Paul[*]
Systems Biology,
Harvard Medical School

Aaron W. Kollasch
Systems Biology,
Harvard Medical School

Pascal Notin
Systems Biology,
Harvard Medical School

Debora S. Marks[*]
Harvard Medical School
Broad Institute

Abstract

Deep generative models of protein sequence and inverse folding models have shown great promise as protein design methods. While sequence-based models have shown strong zero-shot mutation effect prediction performance, inverse folding models have not been extensively characterized in this way. As these models use information from protein structures, it is likely that inverse folding models possess inductive biases that make them better predictors of certain function types. Using the collection of model scores contained in the newly updated ProteinGym, we systematically explore the differential zero-shot predictive power of sequence and inverse folding models. We find that inverse folding models consistently outperform the best-in-class sequence models on assays of protein thermostability, but have lower performance on other properties. Motivated by these findings, we develop StructSeq, an ensemble model combining information from sequence, multiple sequence alignments (MSAs), and structure. StructSeq achieves state-of-the-art Spearman correlation on ProteinGym and is robust to different functional assay types.

1 Introduction

Generative modeling has emerged as a leading approach for mutation effect prediction and protein design. In evolution, protein selection fitness is driven by multiple features including the expressibility and stability of the protein, and the efficacy of its function (e.g. enzymatic activity and binding affinity). Generative sequence models are trained on protein sequences across evolution and capture a rich understanding of protein fitness. Likelihoods from these models have been found to correlate highly with many different aspects of fitness [Sella and Hirsh, 2005, Hopf et al., 2017, Riesselman et al., 2018]. For design, these methods have been used to generate novel sequences and optimize proteins for various functions [Russ et al., 2005, Tian et al., 2018, Shin et al., 2021, Schubert et al., 2018, Blazejewski et al., 2019, Norn et al., 2021, Hie et al., 2023]. While mutation effect prediction and design have often been treated as separate goals, the distribution learned by these generative sequence models draws a link between these tasks. If the probabilities from a model correlate well with a protein property, designs from that model will be enhanced for that property [Tian et al., 2018, Johnson et al., 2023].

While the sequence contains all the information required to fold most proteins [Anfinsen and Scheraga, 1975], and an ideal model could thus derive its predictions from sequence alone, real-world sequence

[*] Correspondence: steffanpaul@g.harvard.edu, debbie@hms.harvard.edu

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

models under-perform due to limited evolutionary data, ill-suited inductive biases, and the inherent performance trade-off between density estimation and fitness estimation [Weinstein et al., 2022]. Therefore, models imbued with task-specific inductive biases may be better suited to predicting specific functions. Furthermore, if protein engineers know which inductive biases correlate with each protein property, they can adapt their design models to promote the properties they are most interested in.

In this work, we delineate the advantages that different generative models have on different functions. We find that, while sequence models outperform inverse folding models on zero-shot prediction of organismal fitness, inverse folding models have superior predictive power on thermostability. Motivated by this, we develop StructSeq, a hybrid model that builds on the concepts developed in TranceptEVE [Notin et al., 2022b]. We combine global protein sequence information from Tranception [Notin et al., 2022a], family-specific sequence information from EVE [Frazer et al., 2021], and structural information from ESM-IF1 [Hsu et al., 2022]. StructSeq has state-of-the-art performance on the ProteinGym DMS substitutions summary benchmark, with top performance for 4/5 of the individual assay function types.

2 Related work

Sequence models Generative sequence models are trained on evolutionary datasets of protein sequences. Large protein-language models are trained on repertoires of millions of sequences and aim to learn a general distribution across all proteins [Nijkamp et al., 2022, Alley et al., 2019, Notin et al., 2022a, Marquet et al., 2022, Rives et al., 2021]. Alignment-based models learn a distribution centered around a particular protein family by training on multiple-sequence alignments (MSAs) [Hopf et al., 2017, Frazer et al., 2021, Laine et al., 2019]. These models benefit from the inductive biases afforded by alignments, such as the importance of conserved residues in columns [Weinstein et al., 2022]. Hybrid sequence models combine general protein information with local family information from alignments [Rao et al., 2021, Notin et al., 2022b]. This combination makes the log-likelihoods more robust, giving these models top performance on mutation effect prediction [Notin et al., 2022b].

Inverse folding models condition their sequence likelihoods on a known structure, allowing them to efficiently model interactions that are local in three-dimensional space but distant in two-dimensional residue space [Ingraham et al., 2019, Hsu et al., 2022, Dauparas et al., 2022, Gao et al., 2022, Yi et al., 2023]. These models take in a protein structure as a graph and leverage graph neural network (GNN) architectures to aggregate information from contacting residues. Some methods construct an input graph using invariant mappings of the residue coordinates [Ingraham et al., 2019, Dauparas et al., 2022], while others leverage SE(3) equivariant GNNs to operate directly on the coordinates themselves [Hsu et al., 2022, Jing et al., 2020]. These models learn a structurally informed representation of the residues, which is autoregressively decoded into a sequence. While they are used increasingly for design tasks [Bennett et al., 2023, Johnson et al., 2023], inverse folding models have not been extensively characterized for predicting protein fitness, and it is not known which aspects of protein fitness are best matched by their inductive biases.

3 Inverse folding models consistently outperform sequence models on stability assays

Inverse folding models have been found to have lower performance than sequence models on an extension to the ProteinGym substitutions dataset [Notin et al., 2023, in submission]. Because inverse folding models have access to structural information, we hypothesized that there might be certain function prediction tasks for which they have an advantage over sequence models.

To investigate this, we looked at the difference in Spearman rank correlation (Spearman Gain) between two inverse folding models (ESM-IF1 [Hsu et al., 2022] and ProteinMPNN [Dauparas et al., 2022]) and the top two models of each sequence model type in the ProteinGym substitutions dataset (Fig 1A). These were GEMME [Laine et al., 2019] and EVE (ens.) [Frazer et al., 2021] (alignment-based); VESPA [Marquet et al., 2022] and ProGen2 (ens.) [Nijkamp et al., 2022] (protein language); and Tranception L [Notin et al., 2022a] and TranceptEVE [Notin et al., 2022b] (hybrid sequence models). For details on the DMSes used and on how the models were run, see the appendix. We find that

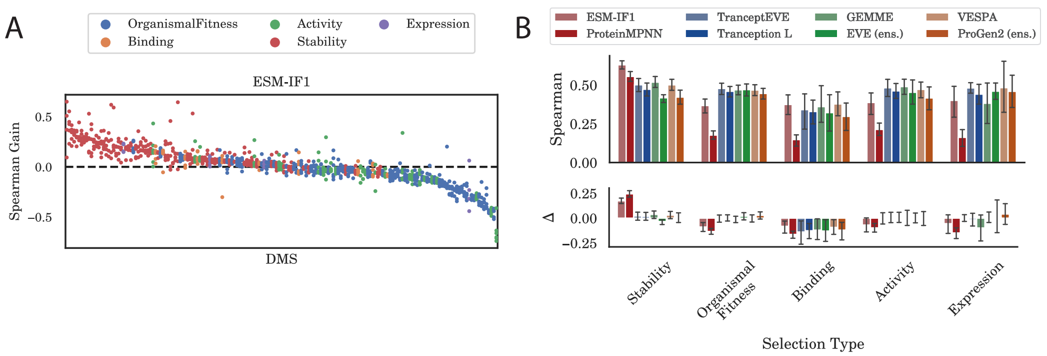

Figure 1: **Spearman differences between Inverse Folding models and top sequence models** (A) The gain in Spearman between ESM-IF1 and the top two sequence models of each sequence model type, calculated for each DMS. We plot the DMSes in order of the median Spearman Gain. The point color corresponds to the type of function assayed in the DMS. (B) Top panel: a barplot of the mean Spearman of each model grouped by function type. Bottom panel: The difference between the Spearman for that function type and the Spearman across all the datasets (Δ). Error bars are SEM for both panels. Models are colored with the primary color corresponding to the model type.

ESM-IF1 and ProteinMPNN consistently outperform the best sequence models on 83/197 and 42/197 of the DMS substitution datasets respectively (Table A2).

We categorized each DMS into 5 function types, corresponding to the kind of function assayed in the DMS [Keeling et al., 2019]. DMSes that could not be clearly put into a category were left out of this analysis (leaving 184 out of 197 assays). Looking at the models' mean Spearman grouped by function type, we find that the sequence models have similar performance across most of the different functions in ProteinGym, while the inverse folding models have much higher Spearman on the protein thermostability assays, and relatively lower performance on other assay types (Fig. 1B). ESM-IF1 outperforms the best sequence models on 55/68 of the stability DMSes in ProteinGym (ProteinMPNN outperforms on 32/68) (Table A2). This suggests that inverse folding models have an advantage over sequence models for predicting mutation effects on stability.

Sequence models can learn information about contacting residues that contribute to the overall protein fold from the evolutionary conservation of residue pairs. This makes evolutionary sequence models useful for predicting structure and also identifying mutations that would disrupt protein folds [Lin et al., 2023, Marks et al., 2011]. However, inverse folding models condition directly on the structure. Thus, it stands to reason that they would have an advantage in predicting the effect of structure-disrupting mutations. In contrast, the selective constraints that direct organismal fitness are probably better represented in evolutionary sequence datasets. Thus, sequence models like TranceptEVE would outperform an inverse folding model on these function types.

We find that ProteinMPNN and ESM-IF1 have similar performance on the stability DMS assays, but ESM-IF1 is superior on the non-stability assays (Fig. A1). ProteinMPNN was trained on only structures in the PDB while ESM-IF1 was also trained using predicted structures from AlphaFold2 [Dauparas et al., 2022, Hsu et al., 2022, Jumper et al., 2021]. This suggests that, by being exposed to a broader distribution of sequences across evolution during training, the conditional sequence distribution learned by ESM-IF1 is more predictive of a wider array of protein functions.

4 Combining sequence, alignments, and structure gives state-of-the-art performance

Motivated by this finding, we sought to build a model that would have robust predictive power across a wide array of function types.

We develop StructSeq, a sequence-structure hybrid model that combines best-in-class sequence-based and structure-based models. Consider a protein sequence (X), composed of residues $(x_1, x_2, \cdots, x_L)$. We learn a conditional sequence distribution, conditioned on the protein's structure (Y) and an

	Function type					
	Expression	**Binding**	**Activity**	**Stability**	**Organismal fitness**	
#Assays	6	10	31	68	69	
Model name	**Spearman**					**Mean**
EVE (ens.)	0.460	0.322	0.453	0.418	0.470	0.425
GEMME	0.382	0.361	0.490	0.518	0.469	0.444
ProGen2 (ens.)	0.459	0.298	0.417	0.422	0.446	0.408
VESPA	0.483	0.378	0.472	0.501	0.468	0.460
ProteinMPNN	0.162	0.148	0.213	0.555	0.177	0.251
ESM-IF1	0.401	0.375	0.387	0.630	0.368	0.432
Tranception L	0.441	0.329	0.462	0.473	0.459	0.433
TranceptEVE	0.481	0.341	0.482	0.502	**0.478**	0.457
StructSeq	**0.549**	**0.399**	**0.498**	**0.633**	0.468	**0.509**

Table 1: **Spearman performance by function type** Mean Spearman for each model across the DMS assays grouped under each function type. The number of assays in each group is shown at the top. The top two models of each model type are shown in the order of: Alignment-based models, Protein language models, Inverse folding models, and Sequence hybrid models. The last row contains the performance for our sequence, MSA, and structure ensemble model, StructSeq. We also report the mean of the Spearmans across function types.

MSA containing the protein and homologous sequences (M). Based on Notin et al. [2022b] we linearly combine sequence information from Tranception with retrieval ($\log P_{TR}(x_i|x_{<i})$), with MSA information from EVE ($\log P_E(x_i|M)$ as in Notin et al. [2022b]) and structural information from ESM-IF1 ($\log P_I(x_i|Y)$):

$$\log P(x_i|x_{<i}, Y, M) = \alpha_x \log P_{TR}(x_i|x_{<i}) + \alpha_m \log P_E(x_i|M) + \alpha_y \log P_I(x_i|Y, x_{<i})$$

For our experiments, we set $\alpha_x = \frac{1}{2}(1 - \alpha_P)$ and $\alpha_m = \frac{1}{2}\alpha_P$, where α_P is a constant between 0 and 1, that depends on the MSA depth (as described in Notin et al. [2022b]). We set $\alpha_y = \frac{1}{2}$, so that there is an equal contribution of sequence and structure information to the log-likelihood. Alternatively, to maximize predictive power for a particular property, property-dependent weight sets can be used to upweight the sequence or structure contribution accordingly.

We find that StructSeq has SOTA performance on the DMS assays of each function type in ProteinGym substitutions, except for the organismal fitness assays, for which it has close to top Spearman. Aggregating the mean Spearman for the different selection categories, we find that StructSeq is the best-performing model. Thus, ensembling sequence and structure information produces a model that is robust across multiple protein function types.

5 Conclusion and future directions

In this work, we show that inverse folding models are superior zero-shot predictors of protein stability over sequence models. We leverage these findings and present an ensemble method combining sequence, MSAs, and structure which has SOTA summary performance on ProteinGym substitutions and on individual subtypes of DMSes.

As our ensemble approach combines information sources linearly, it is straightforward to up-weight the contribution of the inverse folding model in the fitness predictions and in the designs. This setting would be useful for engineering work in which one is designing de novo stable proteins or optimizing the stability of a protein. As an alternative to ensembling model contributions through their log-likelihoods, it could be of interest to explore learning a joint representation with information flowing from the learned representations of each model sub-unit [Zheng et al., 2023]. These representations could have utility for downstream design and prediction tasks.

We found that all models have low performance on assays of binding relative to other DMS types (Fig 1, Table 1). Predicting binding affinity is of great interest for developing therapeutic antibodies and exploring protein-protein interactions. Future work will develop novel methods with higher gains in predicting binding mutation effects.

Acknowledgments and Disclosure of Funding

Thanks to the Marks Lab mascot, Julius Cheeser!

References

Ethan C Alley, Grigory Khimulya, Surojit Biswas, Mohammed AlQuraishi, and George M Church. Unified rational protein engineering with sequence-based deep representation learning. *Nat. Methods*, 16(12):1315–1322, December 2019.

C.B. Anfinsen and H.A. Scheraga. Experimental and theoretical aspects of protein folding. volume 29 of *Advances in Protein Chemistry*, pages 205–300. Academic Press, 1975. doi: https://doi.org/10.1016/S0065-3233(08)60413-1.

Nathaniel R Bennett, Brian Coventry, Inna Goreshnik, Buwei Huang, Aza Allen, Dionne Vafeados, Ying Po Peng, Justas Dauparas, Minkyung Baek, Lance Stewart, Frank DiMaio, Steven De Munck, Savvas N Savvides, and David Baker. Improving de novo protein binder design with deep learning. *Nat. Commun.*, 14(1):2625, May 2023.

Tomasz Blazejewski, Hsing-I Ho, and Harris H Wang. Synthetic sequence entanglement augments stability and containment of genetic information in cells. *Science*, 365(6453):595–598, August 2019.

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*, 378(6615):49–56, October 2022.

Jonathan Frazer, Pascal Notin, Mafalda Dias, Aidan Gomez, Joseph K Min, Kelly Brock, Yarin Gal, and Debora S Marks. Disease variant prediction with deep generative models of evolutionary data. *Nature*, 599 (7883):91–95, November 2021.

Zhangyang Gao, Cheng Tan, Pablo Chacón, and Stan Z Li. PiFold: Toward effective and efficient protein inverse folding. September 2022.

Brian L Hie, Varun R Shanker, Duo Xu, Theodora U J Bruun, Payton A Weidenbacher, Shaogeng Tang, Wesley Wu, John E Pak, and Peter S Kim. Efficient evolution of human antibodies from general protein language models. *Nat. Biotechnol.*, April 2023.

Thomas A Hopf, John B Ingraham, Frank J Poelwijk, Charlotta P I Schärfe, Michael Springer, Chris Sander, and Debora S Marks. Mutation effects predicted from sequence co-variation. *Nat. Biotechnol.*, 35(2):128–135, February 2017.

Thomas A Hopf, Anna G Green, Benjamin Schubert, Sophia Mersmann, Charlotta P I Schärfe, John B Ingraham, Agnes Toth-Petroczy, Kelly Brock, Adam J Riesselman, Perry Palmedo, Chan Kang, Robert Sheridan, Eli J Draizen, Christian Dallago, Chris Sander, and Debora S Marks. The EVcouplings python framework for coevolutionary sequence analysis. *Bioinformatics*, 35(9):1582–1584, May 2019.

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. April 2022.

John Ingraham, Vikas Garg, Regina Barzilay, and Tommi Jaakkola. Generative models for graph-based protein design. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Bowen Jing, Stephan Eismann, Patricia Suriana, Raphael J L Townshend, and Ron Dror. Learning from protein structure with geometric vector perceptrons. September 2020.

Sean R Johnson, Xiaozhi Fu, Sandra Viknander, Clara Goldin, Sarah Monaco, Aleksej Zelezniak, and Kevin K Yang. Computational scoring and experimental evaluation of enzymes generated by neural networks. April 2023.

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*, July 2021.

Diane Marie Keeling, Patricia Garza, Charisse Michelle Nartey, and Anne-Ruxandra Carvunis. The meanings of 'function' in biology and the problematic case of de novo gene emergence. *Elife*, 8, November 2019.

Elodie Laine, Yasaman Karami, and Alessandra Carbone. GEMME: A simple and fast global epistatic model predicting mutational effects. *Mol. Biol. Evol.*, 36(11):2604–2619, November 2019.

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.

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.

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. *Hum. Genet.*, 141(10):1629–1647, October 2022.

Erik Nijkamp, Jeffrey Ruffolo, Eli N Weinstein, Nikhil Naik, and Ali Madani. ProGen2: Exploring the boundaries of protein language models. June 2022.

Christoffer Norn, Basile I M Wicky, David Juergens, Sirui Liu, David Kim, Doug Tischer, Brian Koepnick, Ivan Anishchenko, Foldit Players, David Baker, and Sergey Ovchinnikov. Protein sequence design by conformational landscape optimization. *Proc. Natl. Acad. Sci. U. S. A.*, 118(11), March 2021.

Pascal Notin, Mafalda Dias, Jonathan Frazer, Javier Marchena-Hurtado, Aidan Gomez, Debora S Marks, and Yarin Gal. Tranception: protein fitness prediction with autoregressive transformers and inference-time retrieval. May 2022a.

Pascal Notin, Lood Van Niekerk, Aaron W Kollasch, Daniel Ritter, Yarin Gal, and Debora S Marks. TranceptEVE: Combining family-specific and family-agnostic models of protein sequences for improved fitness prediction. *bioRxiv*, page 2022.12.07.519495, December 2022b.

Pascal Notin, Aaron W. Kollasch, Daniel Ritter, Lood van Niekerk, Steffanie Paul, Hansen Spinner, Nathan Rollins, Ada Shaw, Ruben Weitzman, Jonathan Frazer, Mafalda Dias, Dinko Franceschi, Rose Orenbuch, Yarin Gal, and Debora S. Marks. Proteingym: Large-scale benchmarks for protein fitness prediction and design. In *Advances in Neural Information Processing Systems*, 2023. URL `https://openreview.net/forum?id=URoZHqAohf`.

Roshan Rao, Jason Liu, Robert Verkuil, Joshua Meier, John F Canny, Pieter Abbeel, Tom Sercu, and Alexander Rives. MSA transformer. August 2021.

Adam J Riesselman, John B Ingraham, and Debora S Marks. Deep generative models of genetic variation capture the effects of mutations. *Nat. Methods*, 15(10):816–822, October 2018.

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. *Proc. Natl. Acad. Sci. U. S. A.*, 118(15), April 2021.

William P Russ, Drew M Lowery, Prashant Mishra, Michael B Yaffe, and Rama Ranganathan. Natural-like function in artificial WW domains. *Nature*, 437(7058):579–583, September 2005.

Benjamin Schubert, Charlotta Schärfe, Pierre Dönnes, Thomas Hopf, Debora Marks, and Oliver Kohlbacher. Population-specific design of de-immunized protein biotherapeutics. *PLOS Computational Biology*, 14(3): 1–19, 03 2018. doi: 10.1371/journal.pcbi.1005983. URL `https://doi.org/10.1371/journal.pcbi.1005983`.

Guy Sella and Aaron E Hirsh. The application of statistical physics to evolutionary biology. *Proc. Natl. Acad. Sci. U. S. A.*, 102(27):9541–9546, July 2005.

Jung-Eun Shin, Adam J Riesselman, Aaron W Kollasch, Conor McMahon, Elana Simon, Chris Sander, Aashish Manglik, Andrew C Kruse, and Debora S Marks. Protein design and variant prediction using autoregressive generative models. *Nat. Commun.*, 12(1):2403, April 2021.

Pengfei Tian, John M Louis, James L Baber, Annie Aniana, and Robert B Best. Co-Evolutionary fitness landscapes for sequence design. *Angew. Chem. Int. Ed Engl.*, 57(20):5674–5678, May 2018.

Eli Weinstein, Alan Amin, Jonathan Frazer, and Debora Marks. Non-identifiability and the blessings of misspecification in models of molecular fitness. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 5484–5497. Curran Associates, Inc., 2022.

Kai Yi, Bingxin Zhou, Yiqing Shen, Pietro Liò, and Yu Guang Wang. Graph denoising diffusion for inverse protein folding. June 2023.

Zaixiang Zheng, Yifan Deng, Dongyu Xue, Yi Zhou, Fei Ye, and Quanquan Gu. Structure-informed language models are protein designers. February 2023.

6 Appendix

6.1 Summary of data and models evaluated in ProteinGym v1.0

Here we briefly describe the data and model scores used from ProteinGym v1.0. ProteinGym v1.0 is an extensive dataset for benchmarking mutation effect prediction models. It comprises over 250 standardized Deep Mutational Scanning (DMS) assays which include over 3.5 million measurements of mutant sequences. While ProteinGym contains benchmark datasets for Clinical Variants and Insertion-Deletion mutants, we focus on the 197 substitution DMS assays included. These datasets comprise measurements of various functions of mutant sequences. All mutation effects are calculated with respect to a single wild-type sequence, specific for that dataset.

Function type annotation We grouped the substitution DMSes into one of five function types. While each assay may have been run differently, these annotations serve as a coarse-grained grouping to understand what sort of protein feature is captured by a model's learned distribution. A description of the rationale behind each annotation is described in Table A1. Any assay that could not satisfactorily be categorized was discarded from this analysis, leaving 184 out of 197 assays.

Implementation details for Zero-shot models All models used in this analysis were run using their default settings according to their publicly available implementations. For our evaluation of alignment-based methods, all alignments were retrieved by querying the target protein sequence from each DMS assay using the Jackhmmer protocol as implemented in the EVcouplings pipeline [Hopf et al., 2019]. For the inverse folding models, input structures were obtained with AlphaFold2 (with templates and amber relaxation) [Jumper et al., 2021]. The target wild-type sequence associated with each DMS was folded and the top predicted fold by AF2 plddt was used as the input to the inverse folding models.

Function type	# Assays	Description
Activity	31	Assays that directly or indirectly measure a protein's catalytic (or otherwise biochemical) activity
Binding	10	Assays that measure the affinity or the degree to which a protein binds its target
Expression	6	Assays that measure how much the protein is expressed in a cell
Organismal fitness	69	Assays that measure how much changes in the protein affect an organisms growth rate
Stability	68	Assays that measure how thermostable a protein is

Table A1: **DMS assay function types** The number of assays in each of the 5 function type categories and a general description used to categorize the assays.

Function type	ESM-IF1	ProteinMPNN
Activity	9	5
Binding	2	0
Expression	4	2
Organismal fitness	12	2
Stability	55	32
Unannotated	1	1
Total	83	42

Table A2: **Top inverse folding model statistic** The number of assays for which each inverse folding model had higher Spearman than all 6 best-in-class sequence models analyzed here, grouped by function type. The performances of ESM-IF1 and ProteinMPNN were not compared to each other here.

Model name	Activity	Binding	Expression	OrganismalFitness	Stability
EVE (ens.)	2	1	0	1	0
GEMME	4	1	1	7	3
ProGen2 (ens.)	3	0	0	9	0
VESPA	5	4	1	18	1
ProteinMPNN	0	0	0	0	2
ESM-IF1	9	2	4	12	53
Tranception L	3	1	0	7	4
TranceptEVE	5	1	0	15	5

Table A3: **Top model statistics** Each row shows the number of assays for which that model was the top performer by Spearman, grouped by assay function type. Here all 8 models were compared against each other.

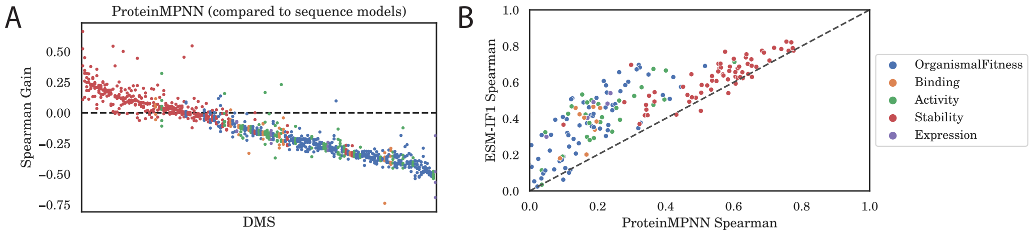

Figure A1: **Spearman comparisons for ProteinMPNN** (A) The gain in Spearman between ProteinMPNN and the top two sequence models of each sequence model type, calculated for each DMS. We plot the DMSes in order of the median Spearman Gain. (A) The ProteinMPNN Spearman compared to the ESM-IF1 Spearman on the DMS assays. The dotted line corresponds to the identity line. The point color for both panels corresponds to the type of function assayed in the DMS.