BindingGYM: A Large-Scale Mutational Dataset Toward Deciphering Protein-Protein Interactions

Wei Lu
Aureka Biotechnologies
luwei@aurekabio.com

Jixian Zhang
Aureka Biotechnologies
jixian@aurekabio.com

Ming Gu
Aureka Biotechnologies
guming@aurekabio.com

Shuangjia Zheng
Aureka Biotechnologies
shuangjia@aurekabio.com

Abstract

Protein-protein interactions are crucial for drug discovery and understanding biological mechanisms. Despite significant advances in predicting the structures of protein complexes, led by AlphaFold3, determining the strength of these interactions accurately remains a challenge. Traditional low-throughput experimental methods do not generate sufficient data for comprehensive benchmarking or training deep learning models. Deep mutational scanning (DMS) experiments provide rich, high-throughput data; however, they are often used incompletely, neglecting to consider the binding partners, and on a per-study basis without assessing the generalization capabilities of fine-tuned models across different assays. To address these limitations, we collected over ten million raw DMS data points and refined them to half a million high-quality points from twenty-five assays, focusing on protein-protein interactions. We intentionally excluded non-PPI DMS data pertaining to intrinsic protein properties, such as fluorescence or catalytic activity. Our dataset meticulously pairs binding energies with the *sequences* and *structures of all interacting partners* using a comprehensive pipeline, recognizing that interactions inherently involve at least two proteins. This curated dataset serves as a foundation for benchmarking and training the next generation of deep learning models focused on protein-protein interactions, thereby opening the door to a plethora of high-impact applications including understanding cellular networks and advancing drug target discovery and development.

1 Introduction

Protein-protein interactions (PPI) represent a vital component of the cellular language, mediating communication within and between cells [1, 2]. The strength of these interactions is commonly measured experimentally as the binding free energy, denoted ΔG, or referred to as binding affinity. In antibody drug discovery, a primary optimization goal is to enhance affinity towards desired targets (affinity maturation) while reducing affinity towards non-desired targets. For example, a broad-spectrum neutralizing antibody drug should bind strongly to prevalent variants of COVID-19 virus proteins to prevent immune escape, yet it should not be polyreactive [3, 4].

Despite its importance, progress in predicting binding affinity is limited by the scarcity of publicly available data from low-throughput biophysical experiments. While high-throughput techniques such as Yeast Two-Hybrid and affinity purification-mass spectrometry (AP-MS) provide extensive binary PPI data, these methods only indicate whether proteins bind, lacking details on the strength of the interactions [5, 6]. Additionally, due to the nature of AP-MS, binary PPI data may include pairs

AI for New Drug Modalities at NeurIPS 2024.

that do not directly interact but are connected through a third protein [7]. Another high-throughput method, deep mutational scanning (DMS), maps genotype to phenotype by combining screening techniques such as fluorescence-activated cell sorting (FACS) with next-generation sequencing (NGS) [8, 9]. This method can produce quantitative fitness scores for millions of mutant proteins sequenced by NGS. While DMS has been used to study a wide range of protein properties, including stability and fluorescence, our dataset exclusively includes studies examining protein-protein interactions, where the fitness score correlates with PPI binding affinity. Unlike previous protein-related datasets that only include information about the mutated protein [10, 11, 12, 13], our dataset explicitly incorporates all interacting proteins. The binding affinity between protein A and B differs from that between A and C, and proteins rarely interact with only a single partner. Predicting the binding affinity based on a single protein, such as protein A alone, is often not meaningful; models trained on such incomplete data fail to capture residue-level interactions and struggle to generalize. Including information about all interacting partners enables the training of a more generalizable PPI model across different assays. Furthermore, in some studies, the binding affinities between a mutant protein and multiple distinct interaction partners are individually measured [14]. Previously, this situation could result in multiple, potentially conflicting affinity data points for each mutant. Now, with complete partner information available, these data points become invaluable, enabling models to discern the intricate details of interaction specificity.

Currently, DMS results are predominantly used by sequence-based methods, whereas structure-based methods, traditionally preferred for estimating protein-protein binding energy, rarely leverage such data. To enable structure-based models to also benefit from DMS results and to facilitate the development of structure-based deep learning models, thus ensuring a level playing field between structure and sequence-based approaches, we map the wild-type sequences documented in source papers to their corresponding crystallized complex structures in the Protein Data Bank [15] through a comprehensive pipeline. For sequences that do not precisely match with the sequences in the crystal structures, we employ homology modeling using BioPython and OpenMM [16, 17]. Additionally, to enable our dataset to support baseline models that require both the wild-type and mutant structures, we use FoldX [18, 19] to generate the complex structure for each mutant.

In BindingGYM, we have assembled the largest collection of DMS-based PPI data available, gathering over ten million raw DMS data points and refining them into half a million high-quality data points from twenty research papers. Each entry includes complete data: the binding energy score, sequences of all interacting proteins, and the structure of the entire complex. This completeness allows our dataset to support a broad range of modeling approaches, including both sequence-based and structure-based methods. Additionally, we have introduced two novel and practically important data-splitting strategies: 'Central vs. Extremes Split' and 'Inter-Assay Split'. The first strategy trains models on entries with middle-range binding energies, from the 10th to the 90th percentile, and evaluates them on the extremes, while the second strategy utilizes data from multiple assays to train models that predict outcomes in unseen assays. Similar to how the ImageNet dataset [20] has been foundational in advancing deep learning models for computer vision, and the high-throughput SELEX dataset [21, 22] in training AlphaFold3 for protein-DNA structure prediction [23], BindingGYM is poised to drive significant advancements in the field of protein-protein interactions. All scripts and data are freely accessible at https://anonymous.4open.science/r/BindingGYM-602D/.

2 Related Work

Protein self properties Several datasets are available to evaluate model performance on a range of protein properties, including secondary and tertiary structures, catalytic activity, stability, and expression. Early work, such as TAPE [11], focused primarily on structural properties, designing tasks for secondary structure prediction, contact prediction, and overall structure, as well as two engineering properties: fluorescence and stability. Meanwhile, FLIP [12] introduced multiple splitting schemes for evaluating protein engineering properties but limited its assessments to results from three assays, focusing solely on sequence-based models.

ProteinGYM [13] provides a comprehensive collection of DMS and clinical variants data. It standardizes measurements under a single metric, the fitness score, which is effective for zero-shot evaluation of models across a broad spectrum of protein properties. However, its fine-tuning capabilities are confined to individual assays, lacking generalization to test cross-assay performance for fine-tuned models. Importantly, while ProteinGYM includes binding-related assay results, it only provides data

Table 1: Dataset Comparison: BindingGYM offers the most extensive collection of quantitative protein-protein interaction data currently available. Additionally, each data is meticulously paired with its corresponding protein complex structure, facilitating comparisons between sequence-based and structure-based methods. Detailed definitions of each column are provided in SI. HT: High-throughput Assay, C-Structure: Complex structure, PGYM: ProteinGYM, PW: ProtienWorkshop

Dataset	# of data	HT	C-Structure Available	Quanti tative	ML ready	Multichain support	Design Usecase
BindingGYM	10M	✓	✓	✓	✓	✓	protein-protein interaction
SKEMPI [30]	7K	✗	✓	✓	✗	✓	protein-protein interaction
STRING [35]	300M	✓	✗	✗	✗	✗	protein network
PGYM [13]	2.7M	✓	✗	✓	✓	✗	general protein fitness
FLIP [12]	320K	✓	✗	✓	✓	✗	general protein fitness
TAPE [11]	100K	✗	✗	✓	✓	✗	protein representation learning
PW [38]	2.3M	✓	✗	✓	✓	✗	protein representation learning
FLAb [39]	10K	✗	✗	✓	✓	✗	therapeutic antibody design

for the protein undergoing mutation, neglecting its interacting partners. Additionally, critical data are often missing; for instance, although [14] conducted screenings for the KRAS protein against seven different proteins, ProteinGYM includes results for only one. Similarly, while [24, 25] contain multiple mutations, ProteinGYM only includes data for single mutations.

Low throughput quantitative PPI dataset Due to the importance of protein-protein interactions, many datasets have been carefully curated, collecting binding affinity measurements from hundreds of papers [26, 27, 28, 29]. However, constrained by the low throughput of conventional biophysical methods, the most comprehensive dataset, SKEMPI [30], comprises only 7,085 data points and is heavily biased toward Alanine substitutions. BindingGYM can be considered the next-generation SKEMPI dataset; like SKEMPI, it includes the full complex structure with binding score but provides orders of magnitude more data, enabling the training of advanced deep learning models, as listed in Table 1 It is important to note that the DMS score does not directly equal ΔG but correlates with it. Consequently, the absolute values of DMS scores are not comparable across different assays. To address this, proper grouping of training samples is crucial, and methods such as learning-to-rank techniques [31, 32, 33] should be employed.

High throughput binary PPI dataset Binary PPI provides protein network information at a proteome scale, which is invaluable for identifying key proteins underlying diseases or catalytic pathways. High-throughput methods yield binary (bind/non-bind) data for various model systems, with databases such as STRING and BioGRID containing millions of binary PPI entries [34, 35]. However, the scores in these databases reflect confidence levels in the existence of interactions rather than their strength. Consequently, models trained on this binary data tend to focus on protein-level evolution instead of residue-level physics, which are crucial for generalizing predictions of mutational effects [36]. While datasets like those in [37] exist, BindingGYM distinguishes itself by specifically focusing on mutational effects, rather than broad proteomic network predictions.

3 The BindingGYM dataset

3.1 Data collection

We collected over ten million DMS data points for 41 unique protein-protein complexes from twenty research papers, specifically focusing on protein-protein interactions. The complete list can be found in the supplementary material. For each paper, we manually traced the raw NGS data whenever possible, checked the data distribution to ensure consistency with the conclusions of the original papers, and corrected any errors introduced during data processing. For example, we identified and corrected misaligned entries in the processed data provided by Heredia et al. [40] by consulting the original article and its raw data.

Unlike ProteinGYM [13], which uses UniProt [41] sequences as references, we ensure our reference sequences are those actually used in the experiments by verifying against the original articles,

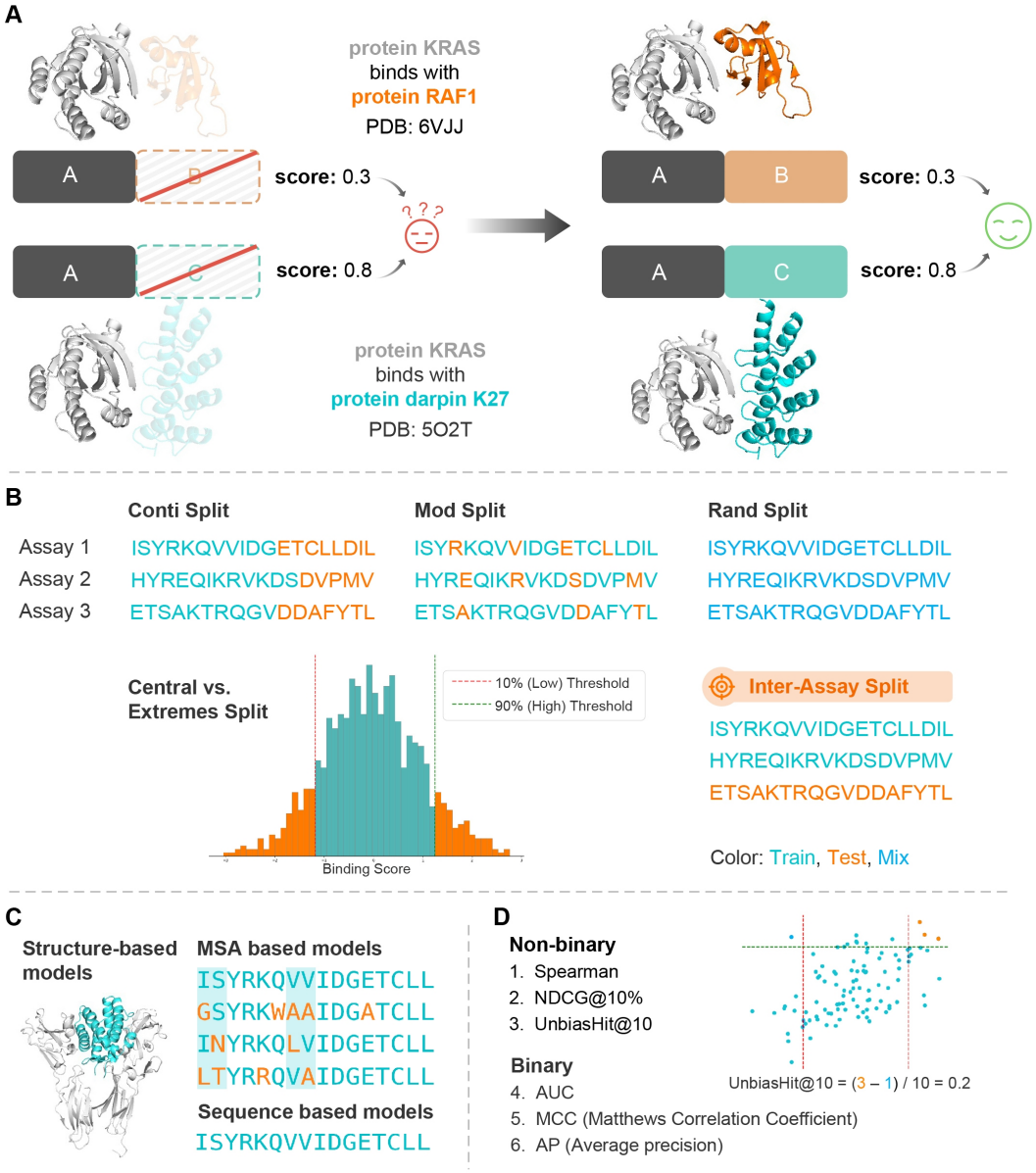

Figure 1: **A**, left: When the interacting partner is not modeled as in most datasets focusing on protein fitness, the same protein can have different binding scores across various assays, leading to confusion. Right: Modeling the full complex, as in BindingGYM, clarifies differences in binding scores, aiding the learning of the underlying physics of protein-protein interactions. **B**, We have implemented five different splits to examine model generalization capabilities, notably 'central vs. extremes split' and 'inter-assay split' to mimic real-world scenarios. **C**, Ten baseline models are included across three categories: structure-based, MSA-based, and sequence-based. **D**, Six evaluation metrics from two groups, binary and non-binary, are used. 'UnbiasHit@10' measures the difference between the proportion of top ten scored mutants in the top 10% and those in the bottom 10%.

appendices, and source codes. This required significant manual effort to identify the actual plasmids and reference sequences used for screening experiments.

In BindingGYM, we find the closest complex structures in the PDB for articles that contain DMS data but lack complex structures, facilitating the application of structure-based methods. When discrepancies arose between PDB structures and actual reference sequences, we employed homology modeling [17, 16] to align the structures with the reference sequences. To support baseline models that require both wild-type and mutant structures, we used FoldX [18, 19] to generate the complex structure for each mutant. The FLIP dataset [12] comprises data from three papers, including one on binding interactions that we have incorporated into our study. Three assays [42, 43, 44] classified as binding-related in ProteinGYM are not actually related to binding and were therefore excluded. Additionally, we noted dataset coverage discrepancies: the original paper [14] reports mutant proteins against seven targets, whereas ProteinGYM includes only one, and covers only single mutations while multiple mutations are documented [24, 25]. We also developed a refined dataset with 508,962 data points, by setting higher NGS count thresholds and applying additional filters, as detailed in the supplementary materials, to facilitate model benchmarking.

3.2 Data splits

As noted in [12, 13], the choice of data splitting scheme is crucial for a fair comparison between models. Random splits often overestimate model generalization because realistic objectives typically involve designing stronger-binding mutants, extrapolating to unexplored fitness landscapes, or applying models to new proteins. Consequently, alternative splitting schemes are necessary to ensure more accurate evaluations. As depicted in Figure 1, we have implemented five distinct data splitting schemes. The first, 'Conti Split', separates continuous blocks in sequence space for both training and testing. The second, 'Mod Split', assigns every n-th residue in sequence space to the test set. The third, 'Rand Split', represents the commonly used random split. These initial three schemes are also used in ProteinGYM.

The fourth scheme, 'Central vs. Extremes Split', sorts the entries by binding score, using the middle 80% for training, while the lowest and highest 10% form the test set. This approach aligns with the common optimization goal of using existing data to design variants with even higher scores. Including the lowest 10% helps reduce bias in evaluation and prevents the trained model from indiscriminately predicting high scores for new mutants.

The fifth, 'Inter-Assay Split', is crucial, where a set of assays is used for training and a different set of assays for testing. This split evaluates the models' ability to generalize to new assays, which holds significant practical significance.

3.3 Baseline models

We include ten baseline models across four categories. Language-based models are notably accessible, requiring only protein sequences as input. These include ProGen2, ESM1v, and ESM2 [45, 46, 47], which exclusively leverage protein sequence data.

Multi-sequence alignment (MSA)-based models such as EVE, Tranception, and TranceptEVE [48, 49, 50] extract evolutionary information from sequence databases. The latter two models also incorporate elements from protein language models, thereby enhancing their prediction capabilities.

Additionally, our dataset includes the structures of protein complexes, supporting the use of structure-based models such as ESM-IF1, ProteinMPNN, PPIformer, and SaProt [51, 52, 53, 54], which leverage these structures to predict protein-protein interactions. With rapid advancements in deep learning for structure prediction [55, 56, 23], the availability of structural data for protein monomers and complexes is expanding, enhancing the applicability of these structure-based models.

For each protein-protein pair, the score for each mutant is defined as the log-ratio of the probability of the mutant to the wild type, $log\frac{p_{mut}}{p_{wt}}$ following [13, 57].

3.4 Metrics

We use six metrics to assess model performance: Spearman, AUC, MCC, NDCG, AP, and a specially designed metric called "UnbiasHit@10". AUC (Area Under the ROC Curve) measures the model's

Table 2: Zero-shot performance on predicting mutational effects on protein-protein interactions.

Category	Model	Spearman	AUC	MCC	NDCG	AP	UnbiasHit@10
Structure-based	ProteinMPNN	**0.40**	**0.69**	**0.15**	**0.72**	**0.22**	0.30
	ESM-if1	0.34	0.66	0.14	0.70	0.20	0.22
	PiFold	0.34	0.66	0.14	0.70	0.20	0.14
	ByProt	0.28	0.62	0.10	0.67	0.17	0.02
	PPIformer	0.19	0.61	0.06	0.59	0.14	0.04
	SaProt	0.27	0.64	0.10	0.67	0.18	0.22
Protein Language-based	ProGen2	0.25	0.61	0.09	0.66	0.16	0.14
	ESM1v	0.26	0.62	0.08	0.66	0.16	0.20
	ESM2	0.29	0.62	0.09	0.67	0.17	0.17
	ESM3	0.27	0.61	0.09	0.66	0.17	0.06
MSA-based	EVE	0.32	0.64	0.12	0.69	0.20	0.28
MSA+Protein Language-based	Tranception	0.32	0.65	0.12	0.69	0.20	**0.31**
	TranceptEVE	0.34	0.66	0.13	0.69	0.20	0.28

ability to discriminate between mutants with higher than binding affinity than the wild type and mutant with lower binding affinity. MCC (Matthews Correlation Coefficient) evaluates the quality of binary classifications, useful in imbalanced datasets. NDCG (Normalized Discounted Cumulative Gain) assesses the ranking quality of the predictions, valuing the order of relevance, we set the threshold at 10%, the same as [13]. AP (Average Precision) calculates the average precision value across different recall levels, highlighting precision-recall trade-offs. "UnbiasHit@10" is useful in practical scenarios where typically only about ten molecules undergo experimental testing with low-throughput, high-accuracy methods. This metric measures the difference between the proportion of the top 10 scored mutants that fall within the top 10% of actual performance and those that fall within the bottom 10%. This metric simulates the situation where we propose ten mutants with the highest predicted binding affinity for experimental validation.

4 Experiments

4.1 Evaluation of zero-shot performances

Due to the high costs associated with setting up experimental assays to quantitatively measure the binding energy between specific protein-protein pairs, zero-shot capability is crucial for discovering potential binders for a protein of interest (POI), studying the mutational effects on a POI, and designing novel binders to a target protein. In this section, we benchmark ten baseline models spanning three categories: structure-based, protein language based, and multiple sequence alignment (MSA) based. Unlike previous studies where predicted monomer structures from AlphaFold were used [13], we input full protein complex structures into our structure-based methods.

Table 2 indicates that models leveraging both evolutionary information from MSA and features from protein language models across a broader sequence database outperform those using either source alone. A structure-based method, ProteinMPNN, shows the best performance on our dataset. We anticipate that integrating evolutionary and physical interaction data more effectively could further enhance zero-shot performance.

4.2 Evaluation of intra-assay finetuned performances

In protein optimization, where some experimental data for candidate protein molecules are available, fine-tuning is critical to enhance desirable properties, such as increased binding to target proteins or reduced binding to undesired targets. Due to space constraints, we present the results for two representative intra-assay split schemes here and provide additional results for two other splits in the supplementary materials.

For finetuning all baseline models, we employ learning-to-rank techniques [31, 32], ensuring uniformity across experiments by using the same batch size for each model, with every batch drawn

Table 3: Performance of fine-tuned models on predicting mutational effects in protein-protein interactions, evaluated over five-fold random splits.

Category	Model	Mutational Depth	Spearman	AUC	MCC	NDCG	AP
Structure-based	ProteinMPNN-R	ALL	0.58	0.78	0.25	0.78	0.31
		<3	0.51	0.75	0.20	0.74	0.28
		>=3	0.54	0.80	0.31	0.80	0.38
	ProteinMPNN	ALL	0.75	0.87	0.45	**0.90**	0.51
		<3	0.73	0.87	0.43	0.88	0.49
		>=3	0.63	0.85	0.45	0.88	0.50
Protein Language-based	ESM2-R	ALL	0.36	0.68	0.11	0.69	0.19
		<3	0.29	0.65	0.08	0.65	0.17
		>=3	0.33	0.69	0.14	0.71	0.23
	ESM2	ALL	**0.76**	0.88	0.45	**0.90**	0.53
		<3	0.74	0.88	0.45	0.89	0.52
		>=3	0.66	0.86	0.43	0.87	0.52
OHE	OHE	ALL	**0.76**	**0.89**	**0.49**	**0.90**	**0.56**
		<3	0.74	0.88	0.49	0.89	0.55
		>=3	0.66	0.87	0.45	0.88	0.52

Table 4: Performance of fine-tuned models on predicting mutational effects in protein-protein interactions, evaluated over five-fold contig splits.

Category	Model	Mutational Depth	Spearman	AUC	MCC	NDCG	AP
Structure-based	ProteinMPNN-R	Single	0.22	0.61	0.07	0.61	0.16
	ProteinMPNN	Single	**0.50**	**0.71**	**0.20**	**0.74**	**0.26**
Protein Language-based	ESM2-R	Single	0.18	0.60	0.06	0.63	0.14
	ESM2	Single	0.46	0.67	0.12	0.71	0.18
OHE	OHE	Single	-0.15	0.44	0.00	0.45	0.10

exclusively from the same assay. Specifically, ESM2 and its randomly initialized variant, ESM2-R, are finetuned using LoRA [58] to mitigate overfitting. For further details, refer to the supplementary materials.

Table 3 shows the results for the random split, which is commonly used as a sanity check to verify that the dataset is informative and that the experimental results are learnable, not arbitrary. We randomly divided each assay's data points into five folds. As demonstrated in Table 3, One-hot encoding (OHE) successfully learns from randomly split data. This confirms the reasonableness of the experimental outcomes and aligns with the established understanding that the majority of mutations do not significantly affect the binding score [30]. Once the effects of key mutations are learned, predicting outcomes for mutants with additional non-impacting mutations becomes straightforward.

Structure-based methods and protein language-based methods, such as ESM2, achieve similar performances to OHE, with a Spearman correlation coefficient of 0.76, which approaches the upper limit of what the quality of DMS data allows. Notably, when initialized with random weights, both ProteinMPNN and ESM2 perform worse than One-hot encoding (OHE), likely due to overfitting. Moreover, ESM2, with significantly more parameters, performs even worse.

The second split demonstrated here is the 'Contig Split', where mutated residues are grouped into five contiguous segments in sequence space. We restrict the mutational depth to single, the same as ProteinGYM, to prevent information leak. This arrangement presents a significant challenge as there is no overlap in mutated residues between any two groups, forcing the model to learn transferable features. As shown in Table 4, One-hot encoding (OHE) fails to identify any transferable features. Pre-training proves to be advantageous; both ProteinMPNN and ESM2 significantly outperform their counterparts initialized with random weights.

Table 5: Performance of fine-tuned models on predicting mutational effects in protein-protein interactions, evaluated over five-fold inter-assay splits.

Category	Model	Mutational Depth	Spearman	AUC	MCC	NDCG	AP
Structure-based	ProteinMPNN-R	ALL	0.16	0.57	0.05	0.59	0.14
		<3	0.11	0.56	0.04	0.56	0.15
		>=3	0.19	0.60	0.09	0.63	0.17
	ProteinMPNN	ALL	**0.42**	**0.70**	**0.16**	**0.72**	**0.23**
		<3	0.43	0.70	0.16	0.72	0.22
		>=3	0.30	0.70	0.17	0.69	0.25
Protein Language-based	ESM2-R	ALL	0.09	0.55	0.03	0.57	0.13
		<3	0.09	0.55	0.02	0.56	0.13
		>=3	0.05	0.54	0.03	0.55	0.14
	ESM2	ALL	0.30	0.62	0.10	0.67	0.18
		<3	0.31	0.61	0.08	0.68	0.17
		>=3	0.15	0.60	0.08	0.60	0.18
OHE	OHE	ALL	0.00	0.50	0.00	0.00	0.10
		<3	0.00	0.50	0.00	0.00	0.10
		>=3	0.00	0.50	0.00	0.00	0.10

4.3 Evaluation of inter-assay finetuned performance

A key contribution of our work is the introduction of the inter-assay split, where assays are clustered based on the sequences of the mutated proteins into five distinct groups. Data from one group are used exclusively for testing, while data from the remaining four groups are used for training. This approach aims to evaluate the generalizability of models to unseen protein-protein pairs, thereby improving zero-shot performance in future PPI experiments.

We analyzed three levels of mutational depth: ALL, <3, and >=3, as presented in Table 5. Generally, mutants with more mutations are harder to predict than those with fewer mutations. One-hot encoding (OHE) struggles to transfer knowledge to unseen assays. Random initialization of weights for ProteinMPNN and ESM2 reduces performance, yet these models manage to learn some transferable features. ProteinMPNN achieves the best results, showing a slight improvement from zero-shot performance. Although the improvements are marginal, we anticipate that with more sophisticated model designs, improved quality filtering of existing data, and the continued accumulation of more data, as shown in protein stability prediction[59], we may observe a scaling law effect: as data volume increases, the performance of fine-tuned models also rises.

4.4 Comparison of model performance in zero-shot and finetuned setting

Using a five-fold inter-assay split allows us to generate predictions for all data, enabling a direct comparison between the performance of inter-assay fine-tuned models and their original zero-shot counterparts. As depicted in Fig 2, each dot represents an assay; in most cases, the fine-tuned models outperform the zero-shot models. For this analysis, we have set the performance benchmark of models initialized with random weights to zero in the zero-shot setting. ProteinMPNN demonstrates superior generalization compared to others.

It is noteworthy that the two outlier points for ProteinMPNN, which fall below the diagonal line, are derived from the same study on protein co-evolution [60]. This suggests that these instances may involve more complex changes in interactions. For example, a mutation in protein A that is harmful in the original interaction environment might become benign or neutral when a reciprocal mutation occurs in protein B. Such dynamical change of protein structures underscore the complexity of protein-protein interactions and highlight the need for models that can adequately account for these protein dynamics.

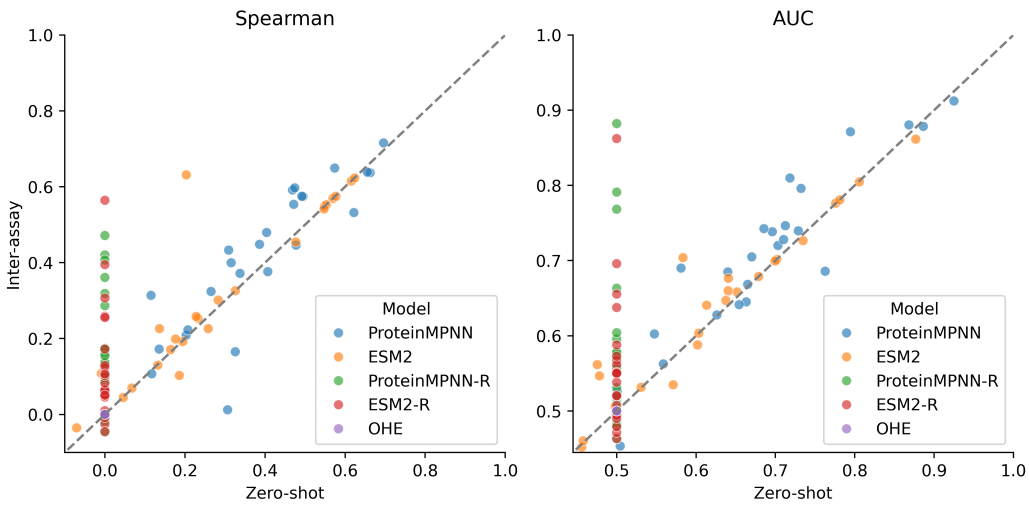

Figure 2: Each dot represents an assay. Finetuned models perform better than zero-shot, especially true for ProteinMPNN, where the dots lie above the diagonal line.

5 Conclusions and Future Work

We have curated BindingGYM, the largest database of quantitative protein-protein interactions to date, highlighting the importance of modeling entire protein complexes. Each assay is meticulously paired with its corresponding complex structure. Five data split schemes were introduced, including two designed to simulate real-world scenarios: the 'Central vs. Extremes Split' for optimizing mutant binding and the 'Inter-Assay Split' for generalizing to new protein pairs. Our evaluation framework includes ten baseline models and six key metrics. Despite the strengths of structure-based models, which outperform sequence- or MSA-based methods and demonstrate superior generalization, there is still room for improvement. Current limitations include noise in DMS-generated data and the relatively limited number of protein-protein pairs studied. Nevertheless, structure-based approaches offer insights into transferable residue-level interactions from numerous mutations.

Looking ahead, as innovations in DMS experiments continue and next-generation sequencing becomes more affordable, the volume of available data will increase. We plan to update the BindingGYM database annually to ensure it remains comprehensive, setting the stage for a unified effort to decode the complex language of protein-protein interactions.

References

[1] Begoña Canovas and Angel R Nebreda. Diversity and versatility of p38 kinase signalling in health and disease. *Nature reviews Molecular cell biology*, 22(5):346–366, 2021.

[2] Nicole L La Gruta, Stephanie Gras, Stephen R Daley, Paul G Thomas, and Jamie Rossjohn. Understanding the drivers of mhc restriction of t cell receptors. *Nature Reviews Immunology*, 18(7):467–478, 2018.

[3] Yanjia Chen, Xiaoyu Zhao, Hao Zhou, Huanzhang Zhu, Shibo Jiang, and Pengfei Wang. Broadly neutralizing antibodies to sars-cov-2 and other human coronaviruses. *Nature Reviews Immunology*, 23(3):189–199, 2023.

[4] Edward P Harvey, Jung-Eun Shin, Meredith A Skiba, Genevieve R Nemeth, Joseph D Hurley, Alon Wellner, Ada Y Shaw, Victor G Miranda, Joseph K Min, Chang C Liu, et al. An in silico method to assess antibody fragment polyreactivity. *Nature Communications*, 13(1):7554, 2022.

[5] Anna Brückner, Cécile Polge, Nicolas Lentze, Daniel Auerbach, and Uwe Schlattner. Yeast two-hybrid, a powerful tool for systems biology. *International journal of molecular sciences*, 10(6):2763–2788, 2009.

[6] Andreas Bauer and Bernhard Kuster. Affinity purification-mass spectrometry: Powerful tools for the characterization of protein complexes. *European journal of biochemistry*, 270(4):570–578, 2003.

[7] André C Michaelis, Andreas-David Brunner, Maximilian Zwiebel, Florian Meier, Maximilian T Strauss, Isabell Bludau, and Matthias Mann. The social and structural architecture of the yeast protein interactome. *Nature*, 624(7990):192–200, 2023.

[8] Douglas M Fowler and Stanley Fields. Deep mutational scanning: a new style of protein science. *Nature methods*, 11(8):801–807, 2014.

[9] Kyrin R Hanning, Mason Minot, Annmaree K Warrender, William Kelton, and Sai T Reddy. Deep mutational scanning for therapeutic antibody engineering. *Trends in pharmacological sciences*, 43(2):123–135, 2022.

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

[11] Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Peter Chen, John Canny, Pieter Abbeel, and Yun Song. Evaluating protein transfer learning with tape. *Advances in neural information processing systems*, 32, 2019.

[12] Christian Dallago, Jody Mou, Kadina E Johnston, Bruce J Wittmann, Nicholas Bhattacharya, Samuel Goldman, Ali Madani, and Kevin K Yang. Flip: Benchmark tasks in fitness landscape inference for proteins. *bioRxiv*, pages 2021–11, 2021.

[13] Pascal Notin, Aaron Kollasch, Daniel Ritter, Lood Van Niekerk, Steffanie Paul, Han Spinner, Nathan Rollins, Ada Shaw, Rose Orenbuch, Ruben Weitzman, et al. Proteingym: large-scale benchmarks for protein fitness prediction and design. *Advances in Neural Information Processing Systems*, 36, 2024.

[14] Chenchun Weng, Andre J Faure, Albert Escobedo, and Ben Lehner. The energetic and allosteric landscape for kras inhibition. *Nature*, 626(7999):643–652, 2024.

[15] Stephen K Burley, Helen M Berman, Gerard J Kleywegt, John L Markley, Haruki Nakamura, and Sameer Velankar. Protein data bank (pdb): the single global macromolecular structure archive. *Protein crystallography: methods and protocols*, pages 627–641, 2017.

[16] Peter JA Cock, Tiago Antao, Jeffrey T Chang, Brad A Chapman, Cymon J Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, et al. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422, 2009.

[17] Peter Eastman, Jason Swails, John D Chodera, Robert T McGibbon, Yutong Zhao, Kyle A Beauchamp, Lee-Ping Wang, Andrew C Simmonett, Matthew P Harrigan, Chaya D Stern, et al. Openmm 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS computational biology*, 13(7):e1005659, 2017.

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

[19] Javier Delgado, Leandro G Radusky, Damiano Cianferoni, and Luis Serrano. Foldx 5.0: working with rna, small molecules and a new graphical interface. *Bioinformatics*, 35(20):4168–4169, 2019.

[20] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[21] Yimeng Yin, Ekaterina Morgunova, Arttu Jolma, Eevi Kaasinen, Biswajyoti Sahu, Syed Khund-Sayeed, Pratyush K Das, Teemu Kivioja, Kashyap Dave, Fan Zhong, et al. Impact of cytosine methylation on dna binding specificities of human transcription factors. *Science*, 356(6337):eaaj2239, 2017.

[22] Arttu Jolma, Yimeng Yin, Kazuhiro R Nitta, Kashyap Dave, Alexander Popov, Minna Taipale, Martin Enge, Teemu Kivioja, Ekaterina Morgunova, and Jussi Taipale. Dna-dependent formation of transcription factor pairs alters their binding specificity. *Nature*, 527(7578):384–388, 2015.

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

[24] Justin R Klesmith, Lihe Su, Lan Wu, Ian A Schrack, Fay J Dufort, Alyssa Birt, Christine Ambrose, Benjamin J Hackel, Roy R Lobb, and Paul D Rennert. Retargeting cd19 chimeric antigen receptor t cells via engineered cd19-fusion proteins. *Molecular pharmaceutics*, 16(8):3544–3558, 2019.

[25] Tyler N Starr, Allison J Greaney, Sarah K Hilton, Daniel Ellis, Katharine HD Crawford, Adam S Dingens, Mary Jane Navarro, John E Bowen, M Alejandra Tortorici, Alexandra C Walls, et al. Deep mutational scanning of sars-cov-2 receptor binding domain reveals constraints on folding and ace2 binding. *cell*, 182(5):1295–1310, 2020.

[26] Quanya Liu, Peng Chen, Bing Wang, and Jinyan Li. dbmpikt: a web resource for the kinetic and thermodynamic database of mutant protein interactions. *arXiv preprint arXiv:1708.01857*, 2017.

[27] Sarah Sirin, James R Apgar, Eric M Bennett, and Amy E Keating. Ab-bind: antibody binding mutational database for computational affinity predictions. *Protein Science*, 25(2):393–409, 2016.

[28] Sherlyn Jemimah, K Yugandhar, and M Michael Gromiha. Proximate: a database of mutant protein–protein complex thermodynamics and kinetics. *Bioinformatics*, 33(17):2787–2788, 2017.

[29] Thom Vreven, Iain H Moal, Anna Vangone, Brian G Pierce, Panagiotis L Kastritis, Mieczyslaw Torchala, Raphael Chaleil, Brian Jiménez-García, Paul A Bates, Juan Fernandez-Recio, et al. Updates to the integrated protein–protein interaction benchmarks: docking benchmark version 5 and affinity benchmark version 2. *Journal of molecular biology*, 427(19):3031–3041, 2015.

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

[31] Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the 24th international conference on Machine learning*, pages 129–136, 2007.

[32] Fen Xia, Tie-Yan Liu, Jue Wang, Wensheng Zhang, and Hang Li. Listwise approach to learning to rank: theory and algorithm. In *Proceedings of the 25th international conference on Machine learning*, pages 1192–1199, 2008.

[33] Ziyi Zhou, Liang Zhang, Yuanxi Yu, Mingchen Li, Liang Hong, and Pan Tan. Enhancing the efficiency of protein language models with minimal wet-lab data through few-shot learning. *arXiv preprint arXiv:2402.02004*, 2024.

[34] Rose Oughtred, Chris Stark, Bobby-Joe Breitkreutz, Jennifer Rust, Lorrie Boucher, Christie Chang, Nadine Kolas, Lara O'Donnell, Genie Leung, Rochelle McAdam, et al. The biogrid interaction database: 2019 update. *Nucleic acids research*, 47(D1):D529–D541, 2019.

[35] Damian Szklarczyk, Rebecca Kirsch, Mikaela Koutrouli, Katerina Nastou, Farrokh Mehryary, Radja Hachilif, Annika L Gable, Tao Fang, Nadezhda T Doncheva, Sampo Pyysalo, et al. The string database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic acids research*, 51(D1):D638–D646, 2023.

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

[37] Minghao Xu, Zuobai Zhang, Jiarui Lu, Zhaocheng Zhu, Yangtian Zhang, Ma Chang, Runcheng Liu, and Jian Tang. Peer: a comprehensive and multi-task benchmark for protein sequence understanding. *Advances in Neural Information Processing Systems*, 35:35156–35173, 2022.

[38] Arian Rokkum Jamasb, Alex Morehead, Chaitanya K Joshi, Zuobai Zhang, Kieran Didi, Simon V Mathis, Charles Harris, Jian Tang, Jianlin Cheng, Pietro Liò, et al. Evaluating representation learning on the protein structure universe. In *The twelfth international conference on learning representations*, 2024.

[39] Michael Chungyoun, Jeffrey A Ruffolo, and Jeffrey J Gray. Flab: Benchmarking deep learning methods for antibody fitness prediction. *bioRxiv*, pages 2024–01, 2024.

[40] Jeremiah D Heredia, Jihye Park, Riley J Brubaker, Steven K Szymanski, Kevin S Gill, and Erik Procko. Mapping interaction sites on human chemokine receptors by deep mutational scanning. *The Journal of Immunology*, 200(11):3825–3839, 2018.

[41] UniProt Consortium. Uniprot: a hub for protein information. *Nucleic acids research*, 43(D1):D204–D212, 2015.

[42] Clara J Amorosi, Melissa A Chiasson, Matthew G McDonald, Lai Hong Wong, Katherine A Sitko, Gabriel Boyle, John P Kowalski, Allan E Rettie, Douglas M Fowler, and Maitreya J Dunham. Massively parallel characterization of cyp2c9 variant enzyme activity and abundance. *The American Journal of Human Genetics*, 108(9):1735–1751, 2021.

[43] Max V Staller, Alex S Holehouse, Devjanee Swain-Lenz, Rahul K Das, Rohit V Pappu, and Barak A Cohen. A high-throughput mutational scan of an intrinsically disordered acidic transcriptional activation domain. *Cell systems*, 6(4):444–455, 2018.

[44] Kevin S Gill, Kritika Mehta, Jeremiah D Heredia, Vishnu V Krishnamurthy, Kai Zhang, and Erik Procko. Multiple mechanisms of self-association of chemokine receptors cxcr4 and ccr5 demonstrated by deep mutagenesis. *Journal of Biological Chemistry*, 299(10), 2023.

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

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

[47] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

[48] 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, 2021.

[49] 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 *International Conference on Machine Learning*, pages 16990–17017. PMLR, 2022.

[50] 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*, pages 2022–12, 2022.

[51] 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 *International conference on machine learning*, pages 8946–8970. PMLR, 2022.

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

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

[54] Jin Su, Chenchen Han, Yuyang Zhou, Junjie Shan, Xibin Zhou, and Fajie Yuan. Saprot: protein language modeling with structure-aware vocabulary. *bioRxiv*, pages 2023–10, 2023.

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

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

[57] Thomas A Hopf, John B Ingraham, Frank J Poelwijk, Charlotta PI Schärfe, Michael Springer, Chris Sander, and Debora S Marks. Mutation effects predicted from sequence co-variation. *Nature biotechnology*, 35(2):128–135, 2017.

[58] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[59] Henry Dieckhaus, Michael Brocidiacono, Nicholas Z Randolph, and Brian Kuhlman. Transfer learning to leverage larger datasets for improved prediction of protein stability changes. *Proceedings of the National Academy of Sciences*, 121(6):e2314853121, 2024.

[60] Aerin Yang, Kevin M Jude, Ben Lai, Mason Minot, Anna M Kocyla, Caleb R Glassman, Daisuke Nishimiya, Yoon Seok Kim, Sai T Reddy, Aly A Khan, et al. Deploying synthetic coevolution and machine learning to engineer protein-protein interactions. *Science*, 381(6656):eadh1720, 2023.

[61] Martin Steinegger and Johannes Söding. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature biotechnology*, 35(11):1026–1028, 2017.

[62] Angela M Phillips, Katherine R Lawrence, Alief Moulana, Thomas Dupic, Jeffrey Chang, Milo S Johnson, Ivana Cvijovic, Thierry Mora, Aleksandra M Walczak, and Michael M Desai. Binding affinity landscapes constrain the evolution of broadly neutralizing anti-influenza antibodies. *Elife*, 10:e71393, 2021.

[63] Carlos L Araya, Douglas M Fowler, Wentao Chen, Ike Muniez, Jeffery W Kelly, and Stanley Fields. A fundamental protein property, thermodynamic stability, revealed solely from large-scale measurements of protein function. *Proceedings of the National Academy of Sciences*, 109(42):16858–16863, 2012.

[64] Richard N McLaughlin Jr, Frank J Poelwijk, Arjun Raman, Walraj S Gosal, and Rama Ranganathan. The spatial architecture of protein function and adaptation. *Nature*, 491(7422):138–142, 2012.

[65] Lewis Chinery, Alissa M Hummer, Brij Bhushan Mehta, Rahmad Akbar, Puneet Rawat, Andrei Slabodkin, Khang Le Quy, Fridtjof Lund-Johansen, Victor Greiff, Jeliazko R Jeliazkov, et al. Baselining the buzz trastuzumab-her2 affinity, and beyond. *bioRxiv*, page 2024.03.26.586756, 2024.

[66] Mason Minot and Sai T Reddy. Meta learning addresses noisy and under-labeled data in machine learning-guided antibody engineering. *Cell Systems*, 15(1):4–18, 2024.

[67] C Anders Olson, Nicholas C Wu, and Ren Sun. A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain. *Current biology*, 24(22):2643–2651, 2014.

[68] Nicholas C Wu, Lei Dai, C Anders Olson, James O Lloyd-Smith, and Ren Sun. Adaptation in protein fitness landscapes is facilitated by indirect paths. *Elife*, 5:e16965, 2016.

[69] Sanjib Dutta, Stefano Gullá, T Scott Chen, Emiko Fire, Robert A Grant, and Amy E Keating. Determinants of bh3 binding specificity for mcl-1 versus bcl-xl. *Journal of molecular biology*, 398(5):747–762, 2010.

[70] Andrew C McShan, Christine A Devlin, Sarah A Overall, Jihye Park, Jugmohit S Toor, Danai Moschidi, David Flores-Solis, Hannah Choi, Sarvind Tripathi, Erik Procko, et al. Molecular determinants of chaperone interactions on mhc-i for folding and antigen repertoire selection. *Proceedings of the National Academy of Sciences*, 116(51):25602–25613, 2019.

[71] Kui K Chan, Danielle Dorosky, Preeti Sharma, Shawn A Abbasi, John M Dye, David M Kranz, Andrew S Herbert, and Erik Procko. Engineering human ace2 to optimize binding to the spike protein of sars coronavirus 2. *Science*, 369(6508):1261–1265, 2020.

[72] Yunlong Cao, Jing Wang, Fanchong Jian, Tianhe Xiao, Weiliang Song, Ayijiang Yisimayi, Weijin Huang, Qianqian Li, Peng Wang, Ran An, et al. Omicron escapes the majority of existing sars-cov-2 neutralizing antibodies. *Nature*, 602(7898):657–663, 2022.

[73] Thomas A Desautels, Kathryn T Arrildt, Adam T Zemla, Edmond Y Lau, Fangqiang Zhu, Dante Ricci, Stephanie Cronin, Seth J Zost, Elad Binshtein, Suzanne M Scheaffer, et al. Computationally restoring the potency of a clinical antibody against omicron. *Nature*, 629(8013):878–885, 2024.

[74] Bernadeta Dadonaite, Jack Brown, Teagan E McMahon, Ariana G Farrell, Daniel Asarnow, Cameron Stewart, Jenni Logue, Ben Murrell, Helen Y Chu, David Veesler, et al. Full-spike deep mutational scanning helps predict the evolutionary success of sars-cov-2 clades. *bioRxiv*, page 2023.11.13.566961, 2023.

[75] Yufeng Luo, Shuo Liu, Jiguo Xue, Ye Yang, Junxuan Zhao, Ying Sun, Bolun Wang, Shenyi Yin, Juan Li, Yuchao Xia, et al. High-throughput screening of spike variants uncovers the key residues that alter the affinity and antigenicity of sars-cov-2. *Cell Discovery*, 9(1):40, 2023.

[76] Bolun Wang, Junxuan Zhao, Shuo Liu, Jingyuan Feng, Yufeng Luo, Xinyu He, Yanmin Wang, Feixiang Ge, Junyi Wang, Buqing Ye, et al. Ace2 decoy receptor generated by high-throughput saturation mutagenesis efficiently neutralizes sars-cov-2 and its prevalent variants. *Emerging Microbes & Infections*, 11(1):1488–1499, 2022.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] see the experiments section.
   (b) Did you describe the limitations of your work? [Yes] See section Conclusions and Future Work.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See section Supplementary material A.1.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [NA]
   (b) Did you include complete proofs of all theoretical results? [NA]
3. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] https://github.com/luwei0917/BindingGYM
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] in 'the BindingGYM dataset' section and SI
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] In SI.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] In SI.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [NA]
   (b) Did you mention the license of the assets? [NA]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] in SI.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] in SI.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] in SI.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [NA]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [NA]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [NA]

A Appendix

A.1 Broader societal impacts

Our research primarily focuses on the fundamental biophysical aspects of protein-protein interactions (PPI), which, by their nature, are not expected to directly result in negative societal impacts. While the development of improved PPI models could theoretically be applied in various contexts, including the design of proteins with potential biosecurity concerns, our study does not produce any new models specifically tailored for such applications. It is important, however, to acknowledge that any advancement in protein design technology carries potential dual-use concerns. Consequently, we advocate for responsible research and adherence to ethical standards to prevent misuse of scientific discoveries in this field.

A.2 Detailed definitions of each column in Table 1

- **# of data**: Represents the total number of data points in the dataset, each corresponding to a unique entry derived from deep mutational scanning experiments, which capture variations in protein sequences and their respective binding energies.
- **HT (High Throughput)**: Indicates whether the data was generated using high-throughput techniques. A 'Yes' suggests that the dataset includes a large volume of data collected through automated processes, enabling comprehensive analysis at scale. A 'No' indicates traditional, lower-scale data collection methods.
- **C-Structure Available**: Specifies whether the crystal structure of the protein complex is available for the corresponding data entry. 'Yes' indicates that the complex structure data is available. 'No' means the complex structure is not available.
- **Quantitative**: Describes whether the data includes quantitative measurements of binding energies. 'Yes' indicates that the data provides numerical values representing binding affinities. 'No' suggests the data is qualitative or binary.
- **ML Ready**: Indicates if the data has been pre-processed and formatted to be directly used in machine learning models. 'Yes' means that the data is cleaned, normalized, and structured, making it immediately suitable for training predictive models. 'No' means that additional preprocessing might be required.
- **Multichain Support**: Indicates whether the dataset supports modeling of multiple protein chains. 'Yes' signifies that the data includes entries involving complex multichain interactions, which are essential for studying protein interactions. 'No' suggests that the dataset only supports modeling of single protein chains.
- **Design Usecase**: Describes the specific applications for which the dataset is designed, such as studying protein-protein interactions, protein networks, general protein fitness, protein representation learning, and therapeutic antibody design. This column highlights the potential research and development areas that can benefit from the dataset.

A.3 Training Details

A.3.1 Data Partitioning

- **Intra_random**: Data is randomly divided into 5 folds using the `KFold` method from `sklearn.model_selection`, with a set seed of 42 to ensure reproducibility.
- **Intra_contig**: For assays with single-point mutation data count ≥ 100, the data is continuously segmented into 5 sections along the sequence, striving to ensure each segment has a similar amount of data.
- **Intra_mod**: For assays with single-point mutation data count ≥ 100, the data is divided into 5 segments based on the position modulo 5 (`pos % 5`).
- **Intra_two_extreme**: The bottom 10% and top 10% of the data are used as the test set, with the remaining data serving as the training set.
- **Inter_assay**: The target sequences are clustered using MMseqs2 [61] with a stringent sequence identity cutoff of 25%, to ensure meaningful generalization across different assays.

A.3.2 Training

- **ProteinMPNN & ESM2**: All models are trained using the AdamW optimizer (learning rate = 0.001, weight decay = 0.05, epsilon = 0.00001) combined with the ListMLE loss [32]. Training proceeds for 100 epochs with an early stopping patience of 3 epochs. During training, all mutation positions are masked in the amino acids, and the predicted value is calculated as the difference between the sum of mutant logit probabilities and wild-type logit probabilities ($\sum(mt\_logit\_probs) - \sum(wt\_logit\_probs)$).
- **OHE (One-Hot Encoding)**: Each sequence is transformed into a one-hot encoded feature matrix of dimensions ($\texttt{seq\_len} \times 20$), where `seq_len` is the length of the sequence and 20 represents the number of amino acid types. A Ridge regression model with an alpha value of 0.01 is then trained to directly predict the DMS scores, following methodologies similar to those described in [51].
- **OHE-AA (One-Hot Encoding of Amino Acid Mutations)**: Each sequence is encoded by comparing mutations to the wild-type (WT) sequence, forming a 20x20 matrix for one-hot encoding (e.g., a mutation from A to E at position 2 is encoded with '1' at the respective position in the matrix). This encoded data is then used to train a Ridge regression model with an alpha value of 0.01 to directly predict DMS scores.

A.4 Total Amount of Compute and Type of Resources Used

The computational resources used for training and analysis comprised one NVIDIA A100 GPU, 48 CPUs, and 1024 GB of RAM.

A.5 Ethical Considerations and Data Handling

- **Consent for Data Usage:** [Yes] Consent for using the data in the BindingGYM dataset was obtained through appropriate channels. The data primarily comprises publicly available deep mutational scanning (DMS) results, which are published in scientific literature. For any unpublished or privately sourced data, explicit consent was secured from the original contributors, ensuring compliance with ethical standards and data usage policies.
- **Personal Information and Offensive Content:** [No] The BindingGYM dataset does not include any personally identifiable information.

A.6 Licensing Information

- **Usage License:** [Yes] The BindingGYM dataset is made available under the MIT License. This license permits reuse, distribution, and modification for both academic and commercial purposes, provided that proper credit is given to the original authors and the dataset. The MIT License is chosen for its permissiveness in encouraging open and collaborative scientific research, facilitating the widespread use and adaptation of the dataset in various biotechnological and pharmaceutical applications.

Table 6: Zero-shot std error, computed based on 1000 bootstrap samples from the set of assays.

Category	Model	Spearman	Spearman Std. error	AUC	AUC Std. error
Structure-based	ProteinMPNN	0.40	0.03	0.69	0.02
	ESM-if1	0.34	0.04	0.66	0.02
	PiFold	0.34	0.03	0.66	0.02
	ByProt	0.28	0.04	0.62	0.02
	PPIformer	0.19	0.04	0.61	0.02
	SaProt	0.27	0.04	0.64	0.02
Protein Language-based	ProGen2	0.25	0.04	0.61	0.02
	ESM1v	0.26	0.04	0.62	0.02
	ESM2	0.29	0.04	0.62	0.02
	ESM3	0.27	0.05	0.61	0.03
MSA-based	EVE	0.32	0.06	0.64	0.03
MSA+Protein Language-based	Tranception	0.32	0.04	0.65	0.02
	TranceptEVE	0.34	0.05	0.66	0.02

A.7 Zero-shot performance with standard error bars

To assess the reliability and consistency of our zero-shot predictions, we computed standard error bars using a bootstrapping approach, as in Table 6. This method involved generating 1000 bootstrap samples from each assay to estimate the variability and confidence intervals around the predicted values.

A.8 Finetuned results for continuous split

Table 7 presents the finetuning results for baselines using the continuous split, where mutations within certain contiguous segments of the sequence space are designated for training, with the remaining segments used for testing. The effectiveness of the finetuned models is assessed on these test sets.

ProteinMPNN displayed the best finetuned results, underlining the advantage of pre-training; conversely, ProteinMPNN with randomly initialized weights showed inferior performance. One-hot encoding, including OHE-AA, performed poorly, demonstrating limited ability to generalize to previously unseen regions.

Table 7: Comparison of finetuning performance for continuous split

Category	Model	Mutational Depth	Spearman	AUC	MCC	NDCG	AP
Structure-based	ProteinMPNN-R	Single	0.22	0.61	0.07	0.61	0.16
	ProteinMPNN	Single	0.50	0.71	0.20	0.74	0.26
Protein Language-based	ESM2-R	Single	0.18	0.60	0.06	0.63	0.14
	ESM2	Single	0.46	0.67	0.12	0.71	0.18
OHE	OHE	Single	-0.15	0.44	0.00	0.45	0.10
	OHE-AA	Single	0.12	0.55	0.02	0.59	0.13

A.9 Finetuned results for modulo split

Table 8 demonstrates that one-hot encoding fails to generalize under the modulo split, while both ESM2 and ProteinMPNN achieve similar performance. This similarity likely arises because adjacent amino acids significantly influence the mutational effects on any specific amino acid.

Table 8: Comparison of finetuning performance for modulo split

Category	Model	Mutational Depth	Spearman	AUC	MCC	NDCG	AP
Structure-based	ProteinMPNN-R	Single	0.26	0.63	0.08	0.63	0.16
	ProteinMPNN	Single	0.53	0.73	0.20	0.73	0.25
Protein Language-based	ESM2-R	Single	0.18	0.57	0.04	0.62	0.14
	ESM2	Single	0.52	0.70	0.17	0.74	0.22
OHE	OHE	Single	-0.05	0.46	0.01	0.49	0.10
	OHE-AA	Single	0.19	0.57	0.04	0.65	0.14

A.10 Finetuned results for central vs extremes split

Table 9 illustrates that all baselines provide reasonable predictions in the Central vs Extremes split, where the strongest and weakest binders form the test set. This effectiveness largely stems from the nature of DMS assays, where the test set consists of mutants at the two extremes, often with multiple mutations. Since individual mutations present in test set mutants are likely already encountered in the training set, models are effectively able to identify mutants with improved binding affinities. But as shown in Table 7, 8 and 5 in the main text, generalizing to unseen positions or new assays is still very challenging.

Table 9: Comparison of finetuning performance for central vs extremes split

Category	Model	TopHit@10	BottomHit@10	UnbiasHit@10
Structure-based	ProteinMPNN-R	0.54	0.13	0.41
	ProteinMPNN	0.82	0.02	0.80
Protein Language-based	ESM2-R	0.55	0.10	0.45
	ESM2	0.80	0.02	0.78
OHE	OHE	0.73	0.02	0.71

A.11 Distribution of DMS score by assay

In Figure 3, we present the histograms showing the distribution of DMS scores for each assay included in the benchmark. These distributions are crucial for understanding the variability captured by different assays. The experimental setups vary significantly among assays, leading to notable differences in score distributions. In certain cases, the histogram peaks sharply at a specific value, indicating a high frequency of scores around that point, which may suggest the score for the wild type. Conversely, other assays exhibit bimodal distributions, where two distinct peaks suggest the presence of two different predominant groups. Moreover, the range of scores varies across assays, reflecting differences in the magnitude of mutational impacts or in the sensitivity of the assays. To accommodate these differences effectively and improve the predictive accuracy of our models, we employ ranking-based machine learning techniques, such as learning to rank. This approach allows us to handle the heterogeneity between assays and learn meaningful interactions from the relative rankings of mutations within each specific assay context.

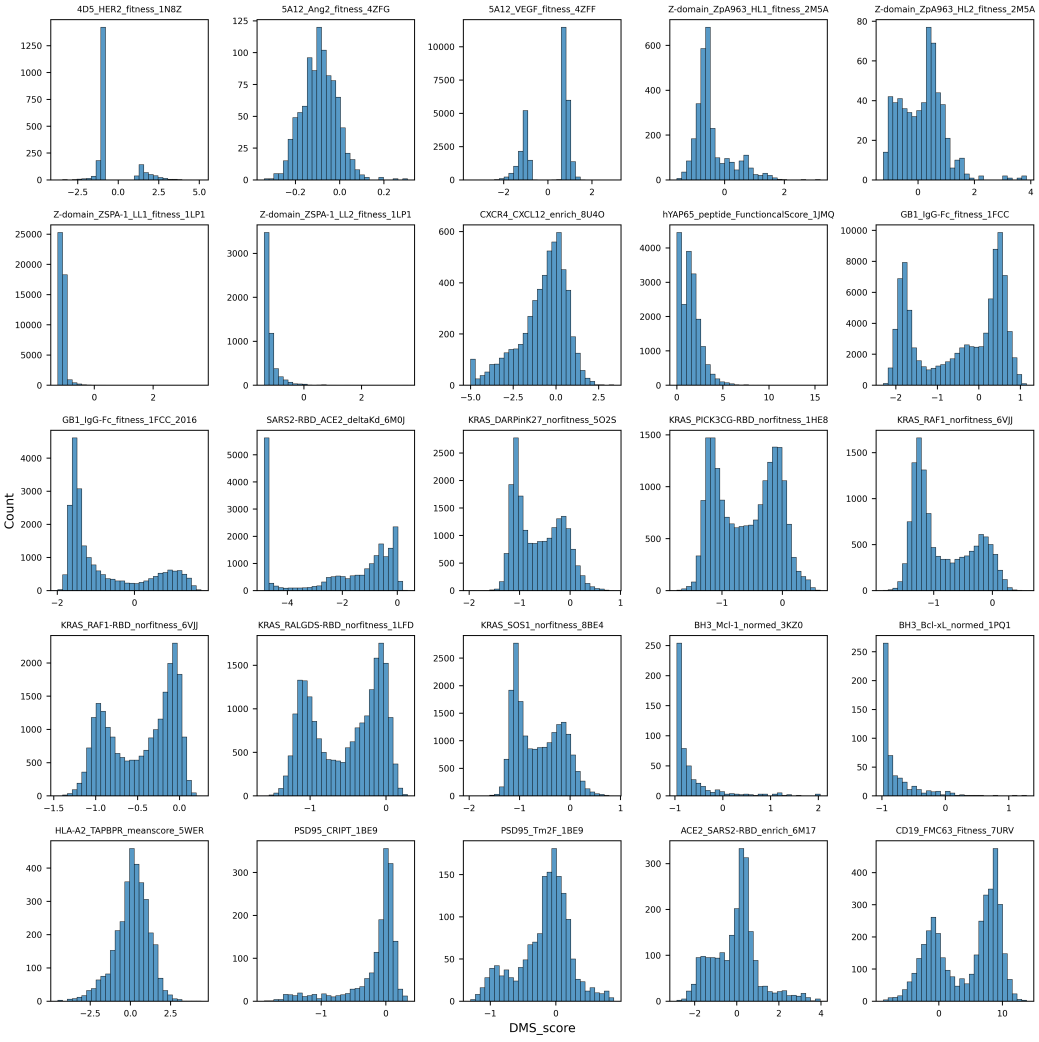

Figure 3: DMS score distribution.

A.12 Pairwise target sequence similarity

Some assays involve the same types of proteins, such as KRAS and its binding partners, leading to shared similarities among entries. While a model trained on one such assay may perform well on others, our primary interest lies in the model's ability to generalize to new targets. Therefore, we cluster target sequences and assess the inter-assay fine-tuned results exclusively across these clusters. Figure 4 illustrates the pairwise sequence similarity among all assays, highlighting potential overlaps and distinctions.

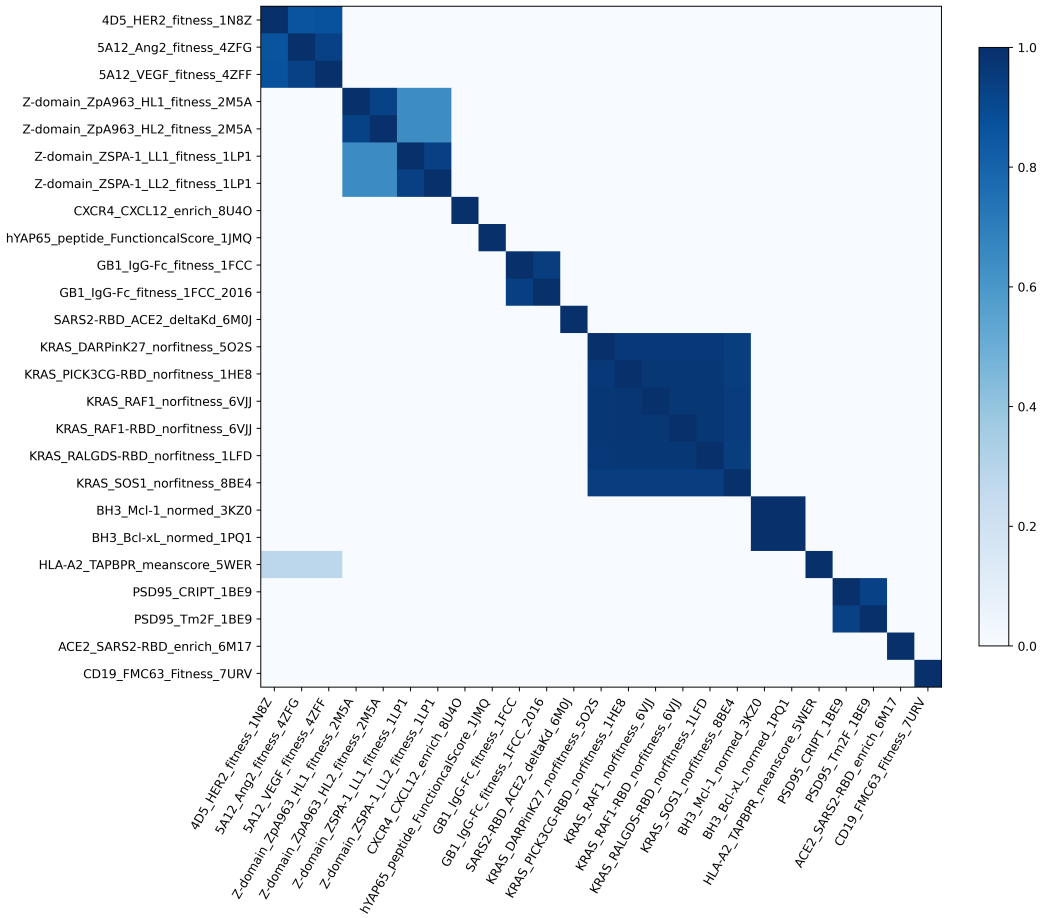

Figure 4: Target sequence similarity across all assays.

A.13 Source and Target Amino Acids

Figure 5 displays the frequency of each specific amino acid being mutated to every other specific amino acid.

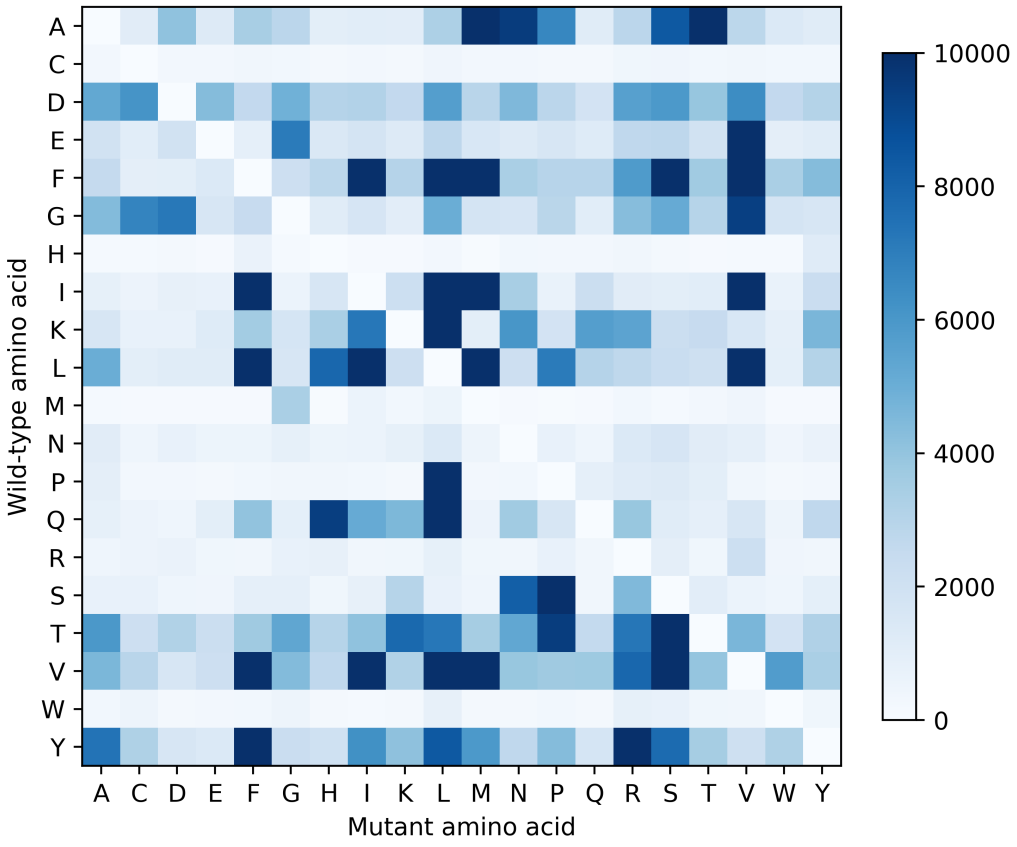

Figure 5: Distribution of amino acid mutations.

A.14 List of data sources

Various studies on protein functions and interactions have been compiled, gathering diverse protein-protein interaction data, as in Table 10. These studies cover a wide range of protein types, including immunoglobulins, chemokine receptors, cytokine receptors, growth factors, regulatory proteins, and virus-related proteins. Additionally, the data includes interactions of the same protein with different other proteins, helping models understand different interaction regions of the same protein and enhancing the model's generalization ability.

Table 10: List of data sources

Protein1	Protein2	Title	DOI
KRAS	PICK3CG-RBD	The energetic and allosteric landscape for KRAS inhibition[14]	10.1038/s41586-023-06954-0
KRAS	RAF1	The energetic and allosteric landscape for KRAS inhibition[14]	10.1038/s41586-023-06954-0
KRAS	RAF1-RBD	The energetic and allosteric landscape for KRAS inhibition[14]	10.1038/s41586-023-06954-0
KRAS	RALGDS-RBD	The energetic and allosteric landscape for KRAS inhibition[14]	10.1038/s41586-023-06954-0
KRAS	SOS1	The energetic and allosteric landscape for KRAS inhibition[14]	10.1038/s41586-023-06954-0
KRAS	DARPinK27	The energetic and allosteric landscape for KRAS inhibition[14]	10.1038/s41586-023-06954-0
CD19	FMC63	Retargeting CD19 Chimeric Antigen Receptor T Cells via Engineered CD19-Fusion Proteins[24]	10.1021/acs.molpharmaceut.9b00418
CXCR4	CXCL12	Mapping Interaction Sites on Human Chemokine Receptors by Deep Mutational Scanning[40]	10.4049/jimmunol.1800343
Z-domain	ZpA963	Deploying synthetic coevolution and machine learning to engineer protein-protein interactions[60]	10.1126/science.adh1720
Z-domain	ZSPA-1	Deploying synthetic coevolution and machine learning to engineer protein-protein interactions[60]	10.1126/science.adh1720
CR6261	FluAH1	Binding affinity landscapes constrain the evolution of broadly neutralizing anti-influenza antibodies[62]	10.7554/eLife.71393
CR9114	FluAH3	Binding affinity landscapes constrain the evolution of broadly neutralizing anti-influenza antibodies[62]	10.7554/eLife.71393
hYAP65	peptide	A fundamental protein property, thermodynamic stability, revealed solely from large-scale measurements of protein function[63]	10.1073/pnas.1209751109
PSD95	CRIPT	The spatial architecture of protein function and adaptation[64]	10.1038/nature11500
Trastuzumab	HER2	Baselining the Buzz Trastuzumab-HER2 Affinity, and Beyond[65]	10.1101/2024.03.26.586756
4D5	HER2	Meta learning addresses noisy and underlabeled data in machine learning-guided antibody engineering[66]	10.1016/j.cels.2023.12.003

Continued on next page

Continued from previous page

Protein1	Protein2	Title	DOI
5A12	Ang2	Meta learning addresses noisy and under-labeled data in machine learning-guided antibody engineering[66]	10.1016/j.cels.2023.12.003
5A12	VEGF	Meta learning addresses noisy and under-labeled data in machine learning-guided antibody engineering[66]	10.1016/j.cels.2023.12.003
GB1	IgG-Fc	A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain[67]	10.1016/j.cub.2014.09.072
GB1	IgG-Fc	Adaptation in protein fitness landscapes is facilitated by indirect paths[68]	10.7554/eLife.16965
BH3	Mcl-1	Determinants of BH3 binding specificity for Mcl-1 versus Bcl-xL[69]	10.1016/j.jmb.2010.03.058
BH3	Bcl-xL	Determinants of BH3 binding specificity for Mcl-1 versus Bcl-xL[69]	10.1016/j.jmb.2010.03.058
HLA-A2	TAPBPR	Molecular determinants of chaperone interactions on MHC-I for folding and antigen repertoire selection[70]	10.1073/pnas.1915562116
SARS2-RBD	ACE2	Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding[25]	10.1016/j.cell.2020.08.012
ACE2	SARS2-RBD	Engineering human ACE2 to optimize binding to the spike protein of SARS coronavirus 2[71]	10.1126/science.abc0870
SARS2-RBD	COVOX-150	Omicron escapes the majority of existing SARS-CoV-2 neutralizing antibodies[72]	10.1038/s41586-021-04385-3
COV2-2130	SARS2	Computationally restoring the potency of a clinical antibody against Omicron[73]	10.1038/s41586-024-07385-1
SARS2	ACE2	Full-spike deep mutational scanning helps predict the evolutionary success of SARS-CoV-2 clades[74]	10.1101/2023.11.13.566961
SARS2	ACE2	High-throughput screening of spike variants uncovers the key residues that alter the affinity and antigenicity of SARS-CoV-2[75]	10.1038/s41421-023-00534-2
ACE2	SARS2	ACE2 decoy receptor generated by high-throughput saturation mutagenesis efficiently neutralizes SARS-CoV-2 and its prevalent variants[76]	10.1080/22221751.2022.2079426

Table 11: Zero-shot Performance of models with error bar.

Category	Model	Spearman	AUC	MCC	NDCG	AP	UnbiasHit@10
Structure-based	ProteinMPNN	0.40 ± 0.03	0.69 ± 0.02	0.15 ± 0.03	0.72 ± 0.02	0.22 ± 0.03	0.30 ± 0.05
	ESM-if1	0.34 ± 0.04	0.66 ± 0.02	0.14 ± 0.03	0.70 ± 0.03	0.20 ± 0.02	0.22 ± 0.06
	PiFold	0.34 ± 0.04	0.66 ± 0.02	0.14 ± 0.03	0.70 ± 0.02	0.20 ± 0.02	0.14 ± 0.06
	PPIformer	0.19 ± 0.04	0.61 ± 0.02	0.06 ± 0.01	0.59 ± 0.02	0.14 ± 0.01	0.04 ± 0.05
	SaProt	0.27 ± 0.04	0.64 ± 0.02	0.10 ± 0.02	0.67 ± 0.03	0.18 ± 0.02	0.22 ± 0.05
Protein Language-based	ProGen2	0.25 ± 0.04	0.61 ± 0.02	0.09 ± 0.02	0.66 ± 0.03	0.16 ± 0.02	0.14 ± 0.06
	ESM1v	0.26 ± 0.04	0.62 ± 0.02	0.08 ± 0.02	0.66 ± 0.03	0.16 ± 0.02	0.20 ± 0.06
	ESM2	0.29 ± 0.04	0.62 ± 0.02	0.09 ± 0.03	0.67 ± 0.03	0.17 ± 0.02	0.17 ± 0.06
	ESM3	0.27 ± 0.05	0.61 ± 0.03	0.09 ± 0.03	0.66 ± 0.03	0.17 ± 0.02	0.06 ± 0.06
MSA-based	EVE	0.32 ± 0.06	0.64 ± 0.03	0.12 ± 0.03	0.69 ± 0.03	0.20 ± 0.02	0.28 ± 0.06
MSA+Protein Language-based	Tranception	0.32 ± 0.04	0.65 ± 0.02	0.12 ± 0.03	0.69 ± 0.03	0.20 ± 0.02	0.31 ± 0.07
	TranceptEVE	0.34 ± 0.05	0.66 ± 0.02	0.13 ± 0.03	0.69 ± 0.03	0.20 ± 0.02	0.28 ± 0.05

Table 12: Performance of models with error bar, evaluated over five-fold random splits.

Category	Model	Mutational Depth	Spearman	AUC	MCC	NDCG	AP
Structure-based	ProteinMPNN-R	ALL	0.58 ± 0.04	0.78 ± 0.02	0.25 ± 0.03	0.78 ± 0.02	0.31 ± 0.03
		<3	0.51 ± 0.04	0.75 ± 0.02	0.20 ± 0.03	0.74 ± 0.02	0.28 ± 0.02
		>=3	0.54 ± 0.07	0.80 ± 0.04	0.31 ± 0.06	0.80 ± 0.04	0.38 ± 0.06
	ProteinMPNN	ALL	0.75 ± 0.04	0.87 ± 0.02	0.45 ± 0.03	0.90 ± 0.01	0.51 ± 0.03
		<3	0.73 ± 0.04	0.87 ± 0.02	0.43 ± 0.03	0.88 ± 0.02	0.49 ± 0.03
		>=3	0.63 ± 0.06	0.85 ± 0.04	0.45 ± 0.06	0.88 ± 0.03	0.50 ± 0.07
Protein Language-based	ESM2-R	ALL	0.36 ± 0.03	0.68 ± 0.02	0.11 ± 0.02	0.69 ± 0.02	0.19 ± 0.02
		<3	0.29 ± 0.03	0.65 ± 0.02	0.08 ± 0.02	0.65 ± 0.02	0.17 ± 0.01
		>=3	0.33 ± 0.05	0.69 ± 0.03	0.14 ± 0.04	0.71 ± 0.04	0.23 ± 0.03
	ESM2	ALL	0.76 ± 0.04	0.88 ± 0.02	0.45 ± 0.03	0.90 ± 0.02	0.53 ± 0.04
		<3	0.74 ± 0.04	0.88 ± 0.02	0.45 ± 0.03	0.89 ± 0.02	0.52 ± 0.03
		>=3	0.66 ± 0.06	0.86 ± 0.04	0.43 ± 0.06	0.87 ± 0.03	0.52 ± 0.07
OHE	OHE	ALL	0.76 ± 0.03	0.89 ± 0.02	0.49 ± 0.04	0.90 ± 0.02	0.56 ± 0.04
		<3	0.74 ± 0.04	0.88 ± 0.02	0.49 ± 0.04	0.89 ± 0.02	0.55 ± 0.04
		>=3	0.66 ± 0.05	0.87 ± 0.03	0.45 ± 0.06	0.88 ± 0.03	0.52 ± 0.07

Table 13: Performance of models with error bar, evaluated over five-fold contig splits.

Category	Model	Mutational Depth	Spearman	AUC	MCC	NDCG	AP
Structure-based	ProteinMPNN-R	Single	0.22 ± 0.03	0.61 ± 0.02	0.07 ± 0.02	0.61 ± 0.02	0.16 ± 0.02
	ProteinMPNN	Single	0.50 ± 0.04	0.71 ± 0.03	0.20 ± 0.03	0.74 ± 0.02	0.26 ± 0.03
Protein Language-based	ESM2-R	Single	0.18 ± 0.02	0.60 ± 0.01	0.06 ± 0.02	0.63 ± 0.02	0.14 ± 0.01
	ESM2	Single	0.46 ± 0.04	0.67 ± 0.02	0.12 ± 0.03	0.71 ± 0.02	0.18 ± 0.02
OHE	OHE	Single	-0.15 ± 0.04	0.44 ± 0.03	0.00 ± 0.01	0.45 ± 0.03	0.10 ± 0.01

Table 14: Performance of models with error bar, evaluated over five-fold inter-assay splits.

Category	Model	Mutational Depth	Spearman	AUC	MCC	NDCG	AP
Structure-based	ProteinMPNN-R	ALL	0.16 ± 0.03	0.57 ± 0.02	0.05 ± 0.02	0.59 ± 0.02	0.14 ± 0.02
		<3	0.11 ± 0.02	0.56 ± 0.02	0.04 ± 0.02	0.56 ± 0.02	0.15 ± 0.02
		>=3	0.19 ± 0.05	0.60 ± 0.04	0.09 ± 0.04	0.63 ± 0.04	0.17 ± 0.03
	ProteinMPNN	ALL	0.42 ± 0.04	0.70 ± 0.02	0.16 ± 0.03	0.72 ± 0.02	0.23 ± 0.02
		<3	0.43 ± 0.04	0.70 ± 0.02	0.16 ± 0.02	0.72 ± 0.02	0.22 ± 0.02
		>=3	0.30 ± 0.06	0.70 ± 0.04	0.17 ± 0.05	0.69 ± 0.04	0.25 ± 0.05
Protein Language-based	ESM2-R	ALL	0.09 ± 0.03	0.55 ± 0.02	0.03 ± 0.02	0.57 ± 0.02	0.13 ± 0.01
		<3	0.09 ± 0.03	0.55 ± 0.02	0.02 ± 0.02	0.56 ± 0.02	0.13 ± 0.01
		>=3	0.05 ± 0.06	0.54 ± 0.04	0.03 ± 0.03	0.55 ± 0.04	0.14 ± 0.03
	ESM2	ALL	0.30 ± 0.04	0.62 ± 0.02	0.10 ± 0.02	0.67 ± 0.03	0.18 ± 0.02
		<3	0.31 ± 0.05	0.61 ± 0.03	0.08 ± 0.03	0.68 ± 0.03	0.17 ± 0.02
		>=3	0.15 ± 0.06	0.60 ± 0.04	0.08 ± 0.04	0.60 ± 0.04	0.18 ± 0.04