USP-ddG: A Unified Structural Paradigm with Data Efficacy and Mixture-of-Experts for Predicting Mutational Effects on Protein-Protein Interactions

Guanglei Yu[1,2,3], Xuehua Bi[3], Qichang Zhao[1,2], and Jianxin Wang[1,2(✉)]

[1] School of Computer Science and Engineering, Central South University, Changsha 410083, China
[2] Hunan Provincial Key Lab on Bioinformatics, Central South University, Changsha 410083, China
jxwang@mail.csu.edu.cn
[3] College of Medical Engineering and Technology, Xinjiang Medical University, Urumqi 830017, China

Abstract. Accurately estimating changes in binding free energy ($\Delta\Delta G$) is essential for understanding protein-protein interactions (PPIs) and guiding rational protein design. Existing deep learning approaches benefit from pre-training on large structural datasets, but they often require heavy computation and overlook data efficacy. In addition, prior models typically assume fixed backbone structures and ignore conformational dynamics, limiting generalization. We present USP-ddG, a unified structural paradigm for $\Delta\Delta G$ prediction. USP-ddG integrates a dual-channel architecture with three complementary components: (i) inverse folding-based log-odds ratio, (ii) empirical energy terms from FoldX, and (iii) a geometric encoder with Gaussian noise to capture relaxed conformations. To enhance representation power, we introduce a framework that integrates feed-forward network (FFN) and Mixture-of-Experts (MoE) to model domain-invariant and -specific features, respectively. We further propose CATH-guided Folding Ordering (CFO), a data efficacy strategy that organizes samples to mitigate catastrophic forgetting and data distribution bias. USP-ddG consistently outperforms existing state-of-the-art (SoTA) methods on the SKEMPI v2.0 benchmark, including the challenging hold-out CATH test set. It achieves superior accuracy on both single- and multi-point mutations and demonstrates strong performance in antibody affinity optimization against H1N1 and HER2, and in assessing SARS-CoV-2 variants to hACE2. Ablation studies confirm the benefit of each component. These results highlight USP-ddG as a robust and data-efficient framework for modeling mutational effects on PPIs.
Availability: USP-ddG is publicly available at https://github.com/ak422/USP-ddG.

Keywords: Unified Structural Paradigm · Data Efficacy · MoE · Mutational Effects.

1 Introduction

As the ultimate executors of cellular structures and functions, proteins communicate through specific physical interactions and assemble into molecular machines, participating in nearly all cellular activities [13,21]. Mechanistic dissection of PPIs has provided a foundation for engineering therapeutics that aim at treating, for example, amyloid-related diseases [12,15] and cancers [22,33].

The binding between protein molecules can be regarded as a reversible and rapid equilibrium process, with the binding strength typically quantified by thermodynamic parameters, such as equilibrium dissociation constant (K_D) or binding free energy (ΔG). Thus, developing methods to predict the changes in binding free energy ($\Delta\Delta G$) caused by amino acid mutations is of critical importance in protein design [27]. This involves screening mutant sequences to identify optimized variants that significantly reduce ΔG and thereby enhance binding strength. However, the potential sequence space of mutations is extremely large, and conventional wet-lab experiments face significant challenges due to limited throughput and high costs. In this context, the development and application of computational methods capable of accurately and efficiently predicting $\Delta\Delta G$ have become an urgent need to bridge this gap [28].

Currently, a variety of computational approaches have been developed for predicting $\Delta\Delta G$. However, traditional methods based on physical energy functions and empirical force fields, as well as experimental techniques, often suffer from inherent limitations such as operational complexity and low throughput. Among the SoTA computational protocols, flex ddG in the Rosetta suite estimates $\Delta\Delta G$ through sampling conformational ensembles and incorporating the "backrub" algorithm to account for protein conformational plasticity [4]. Although physically rigorous, this method suffers from low computational efficiency, with screening

efficiencies up to five orders of magnitude lower than modern deep learning based approaches [5,45]. In contrast, FoldX utilizes its empirical effective energy function (EEEF) to rapidly and quantitatively assess the effects of mutations ($\Delta\Delta G$) [11,14,34]. The parameters of this energy function are optimized through fitting to large-scale experimental datasets using machine learning techniques, rendering it highly reliable for protein engineering applications. Recently, energy terms derived from FoldX have been integrated as effective prior knowledge into deep neural networks, enabling fast and accurate prediction of $\Delta\Delta G$ [45].

With the rapid advancement of deep learning, data-driven approaches have emerged as promising alternatives for $\Delta\Delta G$ prediction [35,42,45]. Despite continuous improvements, predictive performance remains constrained by the scarcity of high-quality experimental annotations. To address this limitation, self-supervised pre-training on large-scale unlabeled structural data has become a prevalent paradigm for enhancing model generalization. Common proxy tasks for pre-training include diverse learning objectives, such as inverse folding modeling [44], atomic coordinate reconstruction [25], side-chain conformation prediction [24,26], and masked language modeling [43]. These pre-training strategies enable the learning of sequence and structure representations imbued with physicochemical constraints, while simultaneously highlighting the critical challenge of effectively transferring such prior knowledge to downstream $\Delta\Delta G$ prediction task [19].

In response, and under the assumption of fixed backbone upon mutations, BA-DDG estimates $\Delta\Delta G$ using the Boltzmann distribution, incorporating the difference in sequence negative log-likelihood scores derived from inverse folding model ProteinMPNN [10], thereby achieving substantially improved predictive accuracy [19]. In contrast, EBM-DDG relaxes the fixed backbone assumption to accommodate structural rearrangements upon mutations. It decomposes $\Delta\Delta G$ into a sequence term, derived from BA-DDG, and a structural term. The structural term is estimated using the diffusion-based score-matching model DSM-Bind [20], which generates mutant conformations and quantifies the associated energy changes, enabling more comprehensive $\Delta\Delta G$ modeling [38]. However, it overlooks the intrinsic conformational dynamics of proteins, and focuses on backbone-only determinants of $\Delta\Delta G$, which may result in the loss of critical information contributed by conformational dynamics and side-chains.

In this work, we propose USP-ddG, a unified structural paradigm for predicting $\Delta\Delta G$ on PPIs. The model integrates both protein sequence and structure as input and decomposes $\Delta\Delta G$ into three components along with an auxiliary task. The overall architecture of our approach is structured as follows:

1) Component 1: leveraging the thermodynamic cycle definition, we use the inverse folding model ProteinMPNN to compute sequence negative log-likelihood scores, enabling USP-ddG to learn the correlations between residue preferences and mutational effects.
2) Component 2: this component focuses on modeling $\Delta\Delta G$ prediction with energy terms generated by force field FoldX, which enables the model to learn physical inductive bias from empirical data, providing to account for the thermodynamic properties of proteins.
3) Component 3: we employ an improved CATH-ddG of our previous work for encoding the protein sequence and 3D structure. To simulate structural dynamics, Gaussian noise is added to atomic coordinates during training. Specifically, for the improved CATH-ddG, we introduce a hybrid architecture that integrates FFN and MoE to capture domain-invariant and domain-specific features, respectively. In addition, we propose CATH-guided Folding Ordering (CFO), a curriculum learning-based strategy to enhance learning efficiency by optimizing the ordering of training data. Unlike random shuffling or order-fixed sampling, CFO aims at mitigating issues such as catastrophic forgetting and data distribution bias.
4) Integration of CATH domain alignment: to integrate protein domain classification information for better representation learning, we formulate the structural constraints derived from CATH domains as an auxiliary multi-label classification task within a self-supervised learning (SSL) framework, enabling the model to learn the structural clusters and functional families within its CATH superfamilies.

We conduct extensive experiments to validate the effectiveness of USP-ddG. Specifically, we adopt the independent hold-out CATH test set introduced in CATH-ddG [45], which is curated from SKEMPI v2.0 [18]. This independent test set contains 813 automatically constructed mutations, which ensures no overlap in CATH superfamily with the training set while also maintaining a maximum TM-score < 0.6 for structural similarity. Our results demonstrate that USP-ddG consistently achieves the best performance. In addition, ablation studies confirm the critical contributions of data efficacy and MoE strategies, as well as the use of inverse folding and improved CATH-ddG in enhancing model performance. Overall, the results strongly support the robustness and superior effectiveness of USP-ddG in predicting $\Delta\Delta G$ in a unified structural paradigm. Independent case studies demonstrate successful enhancement of binding affinity on antibody

variants of CR6261 against H1N1 subtype of human influenza A pandemic, variants of trastuzumab against human epidermal growth factor receptor 2 (HER2), and variants in receptor-binding domain (RBD) of SARS-CoV-2 to human angiotensin-converting enzyme 2 (hACE2) receptor.

2 Method

2.1 Notation and Problem Definition

A protein chain $\mathcal{A}$ is represented by n amino acids $\mathcal{S}_{\mathcal{A}} \in \{1, \ldots, 20\}^n$, and its atomic coordinates $\mathcal{X}_{\mathcal{A}} \in \mathbb{R}^{n \times m \times 3}$. Here, m represents the number of atoms per amino acid. A protein complex $\mathcal{AB}$ is formed by the binding of two chains, $\mathcal{A}$ and $\mathcal{B}$, and exists in a dynamic equilibrium between its bound state $\mathcal{AB}_{\text{bnd}}$ and unbound state $\mathcal{AB}_{\text{unbnd}}$. We refer to the unmutated protein chain as the wild-type (wt) and the mutated chain as the mutant (mut). For simplicity, this section focuses on two-chain complexes, using the notation $\{\mathcal{S}_{\mathcal{AB}}, \mathcal{X}\}$ to represent the sequence and its corresponding structure.

The binding free energy (ΔG) of a protein complex is defined as the difference in Gibbs free energy between the bound state (G_{bnd}) and the unbound state (G_{unbnd}): $\Delta G = G_{\text{bnd}} - G_{\text{unbnd}}$. Based on this definition, the change in binding free energy ($\Delta\Delta G$) upon amino acid mutation is given by:

$$\Delta\Delta G = \Delta G^{\text{mut}} - \Delta G^{\text{wt}}, \tag{1}$$

where ΔG^{mut} and ΔG^{wt} represent the binding free energies of the mutant $\{\mathcal{S}^{\text{mut}}_{\mathcal{AB}}, \mathcal{X}^{\text{mut}}\}$ and wild-type $\{\mathcal{S}^{\text{wt}}_{\mathcal{AB}}, \mathcal{X}^{\text{wt}}\}$ complexes, respectively. Our objective is to predict $\Delta\Delta G$ for a given set of mutations $\{\text{wt} \rightarrow \text{mut}\}$.

2.2 Overview

In Fig. 1 (a), we present an overview of USP-ddG. The workflow consists of the following key steps: (i) Channel 1 (structure-fixed view; Equation 6) decomposes $\Delta\Delta G$ into an inverse folding component, $\log p(\mathcal{S}_{\mathcal{AB}} \mid \mathcal{X}^{\text{FX}})$, derived from the negative log-likelihood scores output by ProteinMPNN, and an energy component, $\log p(\mathcal{X}^{\text{FX}})$, calculated using the energy terms from FoldX. The prior work for estimating $\Delta\Delta G$ using the inverse folding model, as well as the derivation under the structure-fixed assumption, are provided in Appendices A and B, respectively. (ii) Channel 2 (structure-relaxed view; Equation 7) integrates our previous work, CATH-ddG, with an auxiliary SSL task that aligns the latent embeddings to protein domain induction preferences, thereby enhancing both predictive performance and interpretability in downstream $\Delta\Delta G$ prediction. Finally, the loss function is defined as the weighted sum of above supervised learning and SSL tasks by the learnable scalar parameters (Equation 11).

2.3 Modeling with Structure-relaxed Assumptions

Inverse folding-based models for predicting $\Delta\Delta G$ (e.g., BA-DDG and EBM-DDG) typically rely on two assumptions: (1) that protein function is primarily dictated by backbone atoms, and (2) that the protein structure exists in a unique, static conformation. However, protein function is orchestrated by a dynamic conformational landscape that integrates the contributions of both the rigid backbone (BB) and the flexible side-chains (SC). Consequently, prior approaches that rely exclusively on the protein backbone fail to capture the full physicochemical complexity of the system. To this end, we propose a unified $\Delta\Delta G$ prediction framework that incorporates side-chain atoms and dynamic conformations into the structural assumptions. To formalize this, we correspondingly define ΔG as:

$$\Delta G = \Delta G^{\text{FX}} + \Delta G^{\text{RX}}, \tag{2}$$

where ΔG^{FX} and ΔG^{RX} correspond to the binding free energy contributions under the structure-fixed (FX, unique conformation) and structure-relaxed (RX, dynamic conformation) hypotheses, respectively. Accordingly, we define the protein structure under each hypothesis as follows:

$$\mathcal{X}^{\text{FX}} = \{\mathcal{X}^{\text{FX}}_{\text{BB}}, \mathcal{X}^{\text{FX}}_{\text{SC}}\}, \qquad \mathcal{X}^{\text{RX}} = \{\mathcal{X}^{\text{RX}}_{\text{BB}}, \mathcal{X}^{\text{RX}}_{\text{SC}}\}, \tag{3}$$

where $\mathcal{X}^{\text{FX}}$ denotes the protein structure under the structure-fixed hypothesis, defined by a unique backbone $\mathcal{X}^{\text{FX}}_{\text{BB}}$ and side-chains $\mathcal{X}^{\text{FX}}_{\text{SC}}$. In contrast, $\mathcal{X}^{\text{RX}}$ represents the structure under the structure-relaxed hypothesis, characterized by a multi-conformation backbone $\mathcal{X}^{\text{RX}}_{\text{BB}}$ and side-chains $\mathcal{X}^{\text{RX}}_{\text{SC}}$.

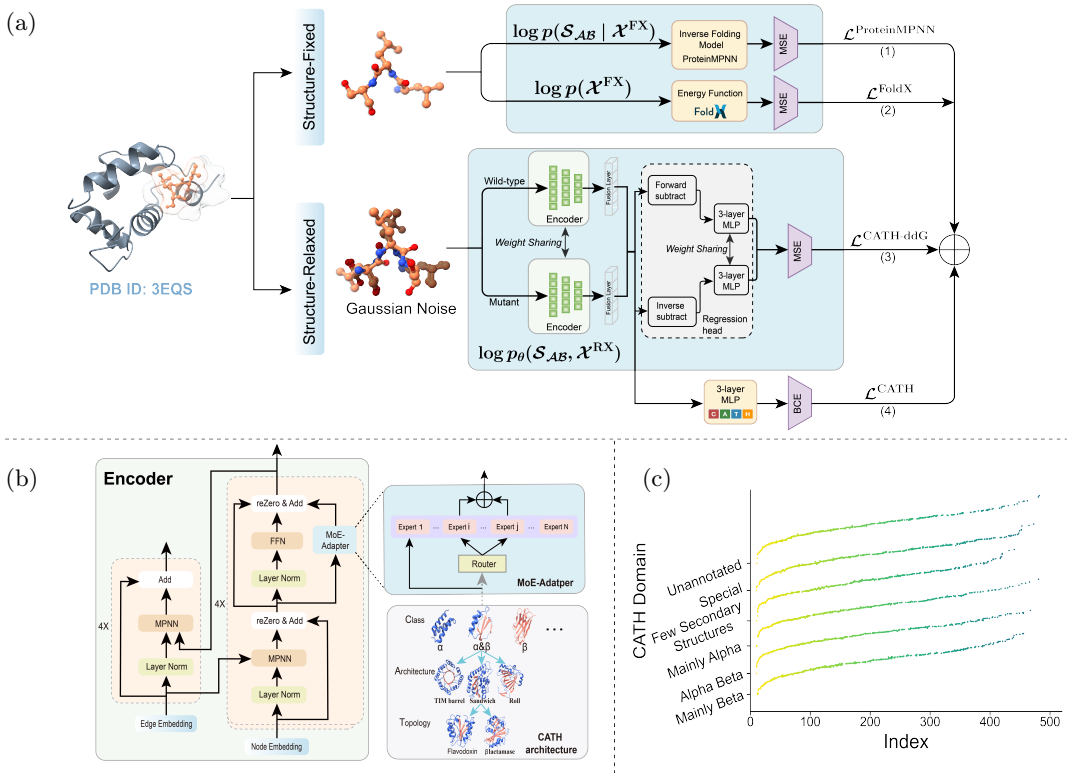

Fig. 1: (a) Overview of the proposed USP-ddG framework. The framework is composed of a dual-channel and four task branches: (1) inverse folding-based supervised learning task, (2) energy function-based supervised learning task, (3) self-supervised learning task for CATH domains, and (4) training with structure noise-based supervised learning task. (b) To utilize the advantage of sparsely activated of MoE to capture the shared and distinctive features across diverse protein domains, we improved the Feed-Forward Network via FFN with MoE to learn domain-invariant and specific features of protein domains. (c) Illustration of CATH-guided Folding Ordering. These results are based on simulated data.

Structure-fixed Assumption (ΔG^{FX}). Under the structure-fixed hypothesis, the binding free energy ΔG^{FX} can be derived from the Boltzmann distribution [2] and is expressed in terms of the probabilities of the bound state and the unbound state, denoted as $p^{\mathrm{FX}}_{\mathrm{bnd}}$ and $p^{\mathrm{FX}}_{\mathrm{unbnd}}$, respectively, as shown below:

$$\Delta G^{\mathrm{FX}} = G^{\mathrm{FX}}_{\mathrm{bnd}} - G^{\mathrm{FX}}_{\mathrm{unbnd}} = -k_B T \cdot \left(\log p^{\mathrm{FX}}_{\mathrm{bnd}} - \log p^{\mathrm{FX}}_{\mathrm{unbnd}}\right), \tag{4}$$

$$= -k_B T \cdot \left(\log p(\mathcal{S}_{\mathcal{AB}}, \mathcal{X}^{\mathrm{FX}}_{\mathrm{bnd}}) - \log p(\mathcal{S}_{\mathcal{AB}}, \mathcal{X}^{\mathrm{FX}}_{\mathrm{unbnd}})\right). \tag{5}$$

Applying Bayes' theorem, similar to the EBM-DDG approach [38], this simplifies to:

$$\Delta G^{\mathrm{FX}} = -k_B T \cdot \left(\log \frac{p(\mathcal{S}_{\mathcal{AB}} \,|\, \mathcal{X}^{\mathrm{FX}}_{\mathrm{bnd}})}{p(\mathcal{S}_{\mathcal{AB}} \,|\, \mathcal{X}^{\mathrm{FX}}_{\mathrm{unbnd}})} + \log \frac{p(\mathcal{X}^{\mathrm{FX}}_{\mathrm{bnd}})}{p(\mathcal{X}^{\mathrm{FX}}_{\mathrm{unbnd}})}\right). \tag{6}$$

The probability of unbound state is approximated as $p(\mathcal{S}_{\mathcal{AB}} \mid \mathcal{X}^{\mathrm{FX}}_{\mathrm{unbnd}}) \approx p(\mathcal{S}_{\mathcal{A}} \mid \mathcal{X}^{\mathrm{FX}}_{\mathcal{A}}) \cdot p(\mathcal{S}_{\mathcal{B}} \mid \mathcal{X}^{\mathrm{FX}}_{\mathcal{B}})$. Furthermore, we leverage the common assumption in protein sequence design that the sequence is largely determined by the backbone structure. Thus, the conditional probabilities $p(\mathcal{S}_{\mathcal{AB}} \,|\, \mathcal{X}^{\mathrm{FX}}_{\mathrm{bnd}})$, $p(\mathcal{S}_{\mathcal{A}} \,|\, \mathcal{X}^{\mathrm{FX}}_{\mathcal{A}})$, and $p(\mathcal{S}_{\mathcal{B}} \,|\, \mathcal{X}^{\mathrm{FX}}_{\mathcal{B}})$ are provided by the pre-trained inverse folding model ProteinMPNN, denoted as $\Delta G_{\mathrm{ProteinMPNN}}$. And the second term in Equation 6, $\log \frac{p(\mathcal{X}^{\mathrm{FX}}_{\mathrm{bnd}})}{p(\mathcal{X}^{\mathrm{FX}}_{\mathrm{unbnd}})}$, which corresponds to the binding free energy of the full complex, is estimated by EBM-DDG under the structure-fixed and backbone-only assumption using the diffusion-based score-matching model DSMBind [20]. In contrast, our method incorporates both backbone and side-chains, and approximates this energy term with the empirical force field FoldX, yielding $\Delta G_{\mathrm{FoldX}}$.

Structure-relaxed Assumption (ΔG^{RX}). Following prior studies and assuming that the unbound state structure is well represented by the bound state conformation, the binding free energy under the relaxed structure hypothesis is formulated by a neural network model:

$$\Delta G^{\mathrm{RX}} = G^{\mathrm{RX}}_{\mathrm{bnd}} - G^{\mathrm{RX}}_{\mathrm{unbnd}} = -k_B T \cdot \log p_\theta(\mathcal{S}_{\mathcal{AB}}, \{\mathcal{X}^{\mathrm{RX}}_{\mathrm{bnd,BB}}, \mathcal{X}^{\mathrm{RX}}_{\mathrm{bnd,SC}}\}), \tag{7}$$

where $p_\theta(\mathcal{S}_{\mathcal{AB}}, \{\mathcal{X}^{\mathrm{RX}}_{\mathrm{bnd,BB}}, \mathcal{X}^{\mathrm{RX}}_{\mathrm{bnd,SC}}\})$ is parameterized by the deep learning neural network CATH-ddG, denoted as $\Delta G_{\text{CATH-ddG}}$, with learnable parameters θ, taking the $\mathcal{S}_{\mathcal{AB}}$ and structure $\{\mathcal{X}^{\mathrm{RX}}_{\mathrm{bnd,BB}}, \mathcal{X}^{\mathrm{RX}}_{\mathrm{bnd,SC}}\}$ with Gaussian noise to computationally simulate its relaxed conformational state as input.

2.4 Unified Structural Paradigm

Under the preceding assumption, we employ a unified structural paradigm $\Delta G = \Delta G^{\mathrm{FX}} + \Delta G^{\mathrm{RX}}$ to achieve the evaluation of $\Delta\Delta G$. This approach, named USP-ddG, integrates the derived components, thereby yielding the overall binding free energy ΔG as follows:

$$\Delta G = \Delta G^{\mathrm{FX}} + \Delta G^{\mathrm{RX}} = \Delta G_{\text{ProteinMPNN}} + \Delta G_{\text{FoldX}} + \Delta G_{\text{CATH-ddG}}, \tag{8}$$

where the first term is predicted by the pre-trained inverse folding model ProteinMPNN, the second term is calculated using FoldX, and the third term is predicted by the improved, trainable neural network model CATH-ddG. Finally, the change in binding free energy $\Delta\Delta G$ is expressed as:

$$\begin{aligned}\Delta\Delta G = &\left(\Delta G^{\mathrm{mut}}_{\text{ProteinMPNN}} + \Delta G^{\mathrm{mut}}_{\text{FoldX}} + \Delta G^{\mathrm{mut}}_{\text{CATH-ddG}}\right) \\ &- \left(\Delta G^{\mathrm{wt}}_{\text{ProteinMPNN}} + \Delta G^{\mathrm{wt}}_{\text{FoldX}} + \Delta G^{\mathrm{wt}}_{\text{CATH-ddG}}\right).\end{aligned} \tag{9}$$

We follow the setup of Yu et al. [45] by incorporating an auxiliary self-supervised Binary Cross-Entropy (BCE) loss to align model training with the inductive biases of protein domains. In our experiments, $-k_B T$ is treated as a learnable parameter. The final objective function is characterized by the following composite Mean Squared Error (MSE) loss and Binary Cross-Entropy (BCE) loss:

$$\mathcal{L} = \mathcal{L}^{\text{ProteinMPNN}} + \mathcal{L}^{\text{FoldX}} + \mathcal{L}^{\text{CATH-ddG}} + \mathcal{L}^{\text{CATH}}, \tag{10}$$

$$\begin{aligned}&= \lambda_1 \mathrm{MSE}(\Delta\Delta G_{\text{ProteinMPNN}}, \Delta\Delta G) + \lambda_2 \mathrm{MSE}(\Delta\Delta G_{\text{FoldX}}, \Delta\Delta G) \\ &\quad + \lambda_3 \mathrm{MSE}(\Delta\Delta G_{\text{CATH-ddG}}, \Delta\Delta G) + \lambda_4 \mathrm{BCE}(\hat{y}_{\text{CATH}}, y_{\text{CATH}}),\end{aligned} \tag{11}$$

where $\{\lambda_i\}_{i=1}^4$ are learnable scalar parameters, as shown in Fig. 1 (a).

2.5 Training with Mixture-of-Experts

We leverage sparse MoE [17,37] to establish a scalable framework aimed at mitigating the issue of structural overlaps across different CATH superfamilies [7,30]. This section first outlines the MoE mechanism and then details its integration into the FFN layer.

The MoE contains two components: the N experts $\{\mathcal{E}_i\}_{i=1}^N$ and a routing network $\mathcal{R}$. While it maintains efficiency by activating only a subset of experts for each input token, its output is combined dynamically via a weighted strategy by the router $\mathcal{R}$. Inspired by [46], we use the LoRA [16] adapter as the expert in MoE to speed up the adaptation on downstream tasks. Our MoE-Adapters are implemented in all the structured Transformer blocks of the spatial and sequential encoders of CATH-ddG, as shown in Fig. 1 (b). Formally, let $\mathbf{x} \in \mathbb{R}^{L_{\text{patch}} \times d_{\text{model}}}$ be the input tokens. The combined output $\mathbf{h}_{\text{MoE}} \in \mathbb{R}^{L_{\text{patch}} \times d_{\text{model}}}$ is then computed as:

$$\mathbf{h}_{\text{MoE}} = \sum_{i=1}^{N} \omega_i \mathcal{E}_i(\mathbf{x}), \tag{12}$$

where $\mathbf{w} = [\omega_1, \omega_2, \ldots, \omega_N] \in \mathbb{R}^{L_{\text{patch}} \times N}$ represents the gating weights assigned by router $\mathcal{R}$, dictating each expert's contribution to the output $\mathbf{h}_{\text{MoE}}$. The gating weights are computed using a learnable matrix $\mathbf{W} \in \mathbb{R}^{N \times d_{\text{model}}}$ and a learnable matrix for the noise component $\mathbf{W}_{\text{noise}} \in \mathbb{R}^{N \times d_{\text{model}}}$ as follows [37]:

$$\mathbf{w} = \mathrm{Softmax}\left(\mathrm{TopK}\left(H(\mathbf{x})\right)\right), \tag{13}$$

$$H(\mathbf{x}) = \mathbf{W}\mathbf{x} + \mathrm{StandardNormal}() \cdot \mathrm{Softplus}\left(\mathbf{W}_{\text{noise}}\mathbf{x}\right). \tag{14}$$

For the FFN layer, we adopt an architecture inspired by DeepSeek-MoE [8,23], which utilizes a hybrid layer of shared and routed experts. Specifically, the shared expert is dedicated to learning domain-invariant knowledge and reducing redundancy among the routed experts, while the routed experts specialize in domain-specific knowledge of protein structures. Accordingly, the output of FFN layer after the attention head, denoted as $\mathbf{h}'_{\text{FFN}} \in \mathbb{R}^{L_{\text{patch}} \times d_{\text{model}}}$, is subsequently redefined as:

$$\mathbf{h}'_{\text{FFN}} = \mathbf{x} + \mathbf{m} \odot \left(\alpha \cdot \text{Linear}\left(\text{Concat}\left(\mathbf{h}_{\text{FFN}}, \mathbf{h}_{\text{MoE}}\right)\right)\right), \tag{15}$$

where $\mathbf{m} \sim \text{Bernoulli}(p = 0.1)$ is the dropout mask vector, $\odot$ is the Hadamard product operation, α is a learnable scalar for reZero [3], and $\text{Concat}(\cdot)$ represents the concatenation operation, as shown in Fig. 1 (b).

2.6 Training with Data Efficacy

Machine learning models typically achieve reliable performance when the training data accurately capture the real-world distributions. However, even SoTA algorithms often suffer from limited generalization performance when deployed in distribution shifts [39,40]. This raises a critical question: how can we systematically enhance model performance without altering the data scale or model architecture? The answer may lie not in the data itself, but rather in how it is utilized. Currently, data efficiency aims at curating data in terms of both quantity and quality to improve model performance. In contrast, data efficacy is concerned with maximizing performance by optimizing the organization and utilization of training data [9].

Additionally, investigators have demonstrated that while easy negative samples promote shortcut learning, hard negatives are essential for compelling models to learn complex features, thereby ultimately facilitating the discovery of biological mechanistic principles [32,40]. To build models that generalize across diverse real-world scenarios, it is crucial to challenge them with appropriately difficult data. To this end, we propose the CATH-guided Folding Ordering (CFO) strategy, which organizes training data by leveraging evolutionary insights from CATH protein domain annotations to present increasingly difficult samples. More specifically, to construct the training set enriched with hard negatives, we partition the training data into Class-disjoint subsets according to the Class-level categorical definition in the CATH database. The permutation function π_{CFO}, which defines the training order, is formalized as follows:

$$\pi_{\text{CFO}}(\mathcal{D}; L) = \bigcup_{\ell=0}^{L-1} \langle \pi_\ell \,|\, f(\pi_\ell) = \mathcal{CATH}[\ell], \pi_\ell \subseteq \mathcal{D} \rangle \tag{16}$$

where $f(\pi_\ell)$ returns the Class of π_ℓ, $\mathcal{CATH} = \{0, 1, 2, 3, 4, 6\}$ denotes the set of Class-level categories, 0 indicates unannotated entries [45], and $L = |\mathcal{CATH}|$. Thus, based on the Class distributions, an illustration of CFO is provided in Fig. 1 (c), which starts with easier samples and progressively tackles harder ones.

3 Experimental Settings

This section introduces the experimental settings, including the datasets and baseline methods used for comparison. Comprehensive details regarding the evaluation metrics and parameter settings are provided in Appendices C and D, respectively.

3.1 Datasets

To train and evaluate USP-ddG, we utilize the largest available curated dataset, SKEMPI v2.0, which comprises 7,085 mutations. Previous studies have primarily benchmarked models using a PPI-disjoint cross-validation split on this dataset [19,24,26,38]. However, this evaluation strategy is susceptible to significant data leakage and may overestimate model performance. As demonstrated in our prior work [45], an average of 88.70% of mutation entries under this split are classified as easy mutations—defined as those exhibiting a maximum TM-score ≥ 0.6 when aligned to structures in the training set. To enable a more rigorous and functionally relevant assessment, we adopt a partitioning strategy based on CATH homologous superfamilies defined in CATH-ddG [45]. Thus, SKEMPI v2.0 is divided such that the hold-out CATH test set contains only complexes that share no CATH superfamilies with the training set. This procedure yields a challenging test set of 813 mutations, in which none are classified as easy mutations, thereby providing a robust evaluation of the model's generalization towards unseen CATH homologous superfamilies.

3.2 Baselines

For the hold-out CATH test set selected from SKEMPI v2.0, we evaluate our two methods, USP-ddG^{Random} and USP-ddG, against three categories of SoTA baselines, utilizing retrained and retested results whenever available. Here, USP-ddG^{Random} and USP-ddG denote our approach trained on the randomly shuffled training data and the CFO training data $\mathcal{D}$, respectively. Specifically, the first category is traditional energy function approaches, including flex ddG [4] and FoldX [11]. The second category is pre-training based approaches, including RDE-Network [26] and DiffAffinity [24]. The third category is supervised learning approaches, including CATH-ddG [45] and BA-DDG [19].

4 Results

We evaluated our algorithm using both random shuffling and the proposed CFO-based data efficacy strategy, benchmarking it against previous works. The results demonstrate that USP-ddG consistently and substantially outperforms USP-ddG^{Random} across all experiments. This consistent advantage confirms that the CFO-based strategy for organizing training data is significantly more effective than random shuffling. Detailed results are presented in the following sections. In the following tables, the best and second-best results are highlighted in **bold** and <u>underlined</u>, respectively. We also report the performance of USP-ddG for the binding of SARS-CoV-2 RBD to hACE2 in Appendix E.

Table 1: Performance on the hold-out CATH test set under all-, single-, multi-point mutations.

Category	Method	Mutations	Overall					Per-PPI	
			PearsonR↑	SpearmanR↑	RMSE↓	MAE↓	AUROC↑	PearsonR↑	SpearmanR↑
Energy Function	FoldX	all	0.4603	0.5246	2.3083	1.6837	0.7541	0.4932	0.4303
		single	0.4507	0.4919	1.7453	1.3017	0.7125	0.5001	0.4458
		multiple	0.4029	0.3698	3.1460	2.4289	0.8353	0.5831	0.5479
	flex ddG	all	0.6081	0.6096	2.0643	1.4946	0.7637	0.5083	0.4542
		single	0.5697	0.5330	1.6068	1.2002	0.7073	0.5681	0.4633
		multiple	0.6198	0.5616	2.6975	2.0545	0.8278	0.6161	0.6399
Pre-training based	RDE-Network	all	0.4588	0.4703	2.3103	1.7029	0.7452	0.3500	0.2880
		single	0.3830	0.3756	1.8061	1.3547	0.6899	0.3465	0.2534
		multiple	0.4498	0.3789	3.0700	2.4025	0.8196	0.5032	0.4729
	DiffAffinity	all	0.3144	0.3626	2.4684	1.7393	0.6251	0.3233	0.2494
		single	0.2639	0.2927	1.8859	1.3838	0.594	0.3358	0.2500
		multiple	0.2458	0.2602	3.3318	2.4426	0.6729	0.2903	0.2882
Supervised	CATH-ddG	all	0.6150	0.6269	2.0504	1.5047	0.7805	0.5260	<u>0.4940</u>
		single	<u>0.6250</u>	<u>0.5937</u>	<u>1.5263</u>	**1.1377**	0.7464	0.5690	0.5090
		multiple	0.5746	0.5275	2.8131	2.2518	0.8325	0.5965	0.5507
	BA-DDG	all	0.4143	0.4964	2.3646	1.6962	0.7321	0.4641	0.4021
		single	0.4455	0.4817	1.7499	1.2746	0.7112	0.5024	0.4229
		multiple	0.3355	0.3088	3.2339	2.4416	0.7393	0.6603	0.6180
	USP-ddG^{Random}	all	<u>0.6550</u>	<u>0.6352</u>	<u>1.9649</u>	<u>1.4560</u>	<u>0.7937</u>	**0.5418**	**0.5048**
		single	0.6019	0.5784	1.5614	1.1681	<u>0.7612</u>	<u>0.5836</u>	<u>0.5195</u>
		multiple	<u>0.6641</u>	<u>0.5767</u>	<u>2.5699</u>	<u>2.029</u>	<u>0.8382</u>	<u>0.6712</u>	<u>0.6534</u>
	USP-ddG	all	**0.6767**	**0.6537**	**1.9145**	**1.4223**	**0.8015**	<u>0.5416</u>	0.4927
		single	**0.6288**	**0.5959**	**1.5202**	<u>1.1402</u>	**0.7721**	**0.6023**	**0.5361**
		multiple	**0.6888**	**0.6111**	**2.4919**	**1.9898**	**0.8483**	**0.6949**	**0.6681**

4.1 Performance Comparison on SKEMPI v2.0 dataset

We evaluate our USP-ddG deep learning framework on the hold-out CATH test set curated from SKEMPI v2.0, and present the results in Table 1 for three categories of methods across seven evaluation metrics. To ensure robust assessment, five independent experiments with different random seeds are conducted. The key findings are summarized below: (1) USP-ddG consistently outperforms all baselines on both single- and

multi-point mutation subsets, achieving superior performance in 5 out of 7 metrics when evaluated across all mutations. (2) On this non-superfamily leakage test set, traditional force field simulators (e.g., the SoTA flex ddG) are surpassed by USP-ddG not only in computational efficiency but also in predictive accuracy. (3) Without utilizing any additional pre-training data, USP-ddG exceeds all pre-training based methods across all metrics, suggesting that our unified structural framework is intrinsically more effective for $\Delta\Delta G$ prediction than approaches leveraging external pre-training knowledge. (4) USP-ddG achieves clear improvements over BA-DDG, the strongest baseline, which has substantially outperformed other methods under PPI-disjoint cross-validation. (5) Finally, USP-ddG demonstrates the largest gains on multi-point mutations, a challenging due to the complexity of epistatic effects, thereby highlighting its strong practical applicability.

4.2 Evaluation on the antibody CR2621 against H1N1

CR6261 is a broadly neutralizing antibody (bnAb) against influenza A. Phillips et al. [31] systematically measured the K_D of combinatorially complete mutational libraries for 14 heavy-chain mutations, reconstructing all possible evolutionary intermediates to the germline sequence. Leveraging these experimental data, AbBiBench curated a benchmark of 1,887 mutants targeting A/New Caledonia/20/1999 (H1N1) for evaluating model performance in designing high-affinity antibodies (PDB ID: 3GBN) [47]. Crucially, none of these mutants appears in any public training databases, ensuring the absence of mutation-level data leakage and enabling reliable assessment of generalization capability. Accurately predicting $\Delta\Delta G$ for the variants of antibody CR6261 presents a challenge due to the prevalence of multi-point mutations and epistatic effects. Epistasis introduces non-additive interactions between mutations, which not only constrains the evolutionary paths accessible under selective pressure but also complicates the difficulty of accurately predicting $\Delta\Delta G$. The results in Table 2 show that USP-ddG achieves the best performance measured by PearsonR and SpearmanR, with PearsonR and SpearmanR values exceeding the SoTA baseline BA-DDG by 12.86% and 10.78%, respectively, demonstrating its superior generalization ability in capturing epistatic effects.

Table 2: Evaluation of the binding affinity of CR6261 against H1N1.

Category	Method	PearsonR↑	SpearmanR↑
Energy Function	FoldX	0.577	0.593
	flex ddG	0.515	0.516
Pre-training based	RDE-Network	0.370	0.422
	DiffAffinity	0.351	0.323
Supervised	CATH-ddG	0.499	0.502
	BA-DDG	0.583	0.566
	USP-ddG^{Random}	0.608	0.584
	USP-ddG	**0.658**	**0.627**

4.3 Evaluation on the antibody trastuzumab against HER2

The HER2 binders test set, sourced from [6,36] (PDB ID: 1N8Z), comprises 419 variants featuring de novo designed complementarity determining region (CDR) loops. These variants were computationally designed to enhance the quality and controllability of the trastuzumab antibody using generative AI and have been experimentally validated by surface plasmon resonance (SPR), thus establishing the reliability for benchmarking AI-based antibody design methods.The antibody trastuzumab variants exhibit a high average edit distance of 7.6, presenting a significant challenge for $\Delta\Delta G$ predictors which are primarily trained on the low edit distance SKEMPI v2.0 dataset [6]. As shown in Table 3, USP-ddG achieves the best performance on this independent test set, with PearsonR and SpearmanR values that are 3.79% and 3.44% higher, respectively, than the SoTA baseline CATH-ddG. In contrast, traditional energy function methods such as flex ddG and FoldX are substantially outperformed by machine learning-based approaches on this HER2 binder test set.

4.4 Ablation study

To elucidate the contribution of each component in USP-ddG, we performed an ablation study on the hold-out CATH test set. The model was systematically evaluated by progressively deactivating individual

Table 3: Evaluation on the binding affinity of HER2 binders.

Category	Method	PearsonR↑	SpearmanR↑
Energy Function	FoldX	0.246	0.329
	flex ddG	0.301	0.339
Pre-training based	RDE-Network	0.465	0.481
	DiffAffinity	0.590	0.524
	GearBind+Ensemble[a]	0.579	0.608
Supervised	CATH-ddG	0.607	0.639
	BA-DDG	0.326	0.362
	USP-ddG^{Random}	0.619	0.656
	USP-ddG	**0.630**	**0.661**

[a] Results are from GearBind [6].

Table 4: Impact of individual module exclusion on performance drop in the hold-out CATH test set.

Model	Mutations	Overall					Per-PPI	
		PearsonR↑	SpearmanR↑	RMSE↓	MAE↓	AUROC↑	PearsonR↑	SpearmanR↑
w/o MoE	all	0.6392	0.6251	1.9996	1.4707	0.7855	0.4467	0.4119
	single	0.5804	0.5591	1.5921	1.1965	0.7569	0.5088	0.4376
	multiple	0.6521	0.5908	2.606	2.0224	0.8241	0.6528	0.6473
w/o Inverse Folding	all	0.6265	0.6065	2.0267	1.5058	0.7781	0.4963	0.4597
	single	0.5578	0.5292	1.6227	1.2169	0.7310	0.5328	0.4805
	multiple	0.6456	0.5551	2.6249	2.0805	0.8534	0.6061	0.5762
w/o CATH-ddG	all	0.5963	0.6178	2.0873	1.5287	0.7837	0.4709	0.3955
	single	0.5523	0.5530	1.6300	1.2231	0.7488	0.5120	0.4328
	multiple	0.5888	0.5497	2.7783	2.1615	0.8311	0.6800	0.6552
w/o Sampler	all	0.6550	0.6352	1.9649	1.456	0.7937	**0.5418**	**0.5048**
	single	0.6019	0.5784	1.5614	1.1681	0.7612	0.5836	0.5195
	multiple	0.6641	0.5767	2.5699	2.029	0.8382	0.6712	0.6534
USP-ddG	all	**0.6767**	**0.6537**	**1.9145**	**1.4223**	**0.8015**	0.5416	0.4927
	single	**0.6288**	**0.5959**	**1.5202**	**1.1402**	**0.7721**	**0.6023**	**0.5361**
	multiple	**0.6888**	**0.6111**	**2.4919**	**1.9898**	**0.8483**	**0.6949**	**0.6681**

modules, followed by retraining and assessment. As illustrated in Table 4, we observed consistent declines in correlation metrics, underscoring the importance of each module in achieving optimal performance. Our results indicate that the inverse folding $\log p(\mathcal{S}_{\mathcal{AB}} \,|\, \mathcal{X}^{\text{FX}})$ and CATH-ddG $\log p_\theta(\mathcal{S}_{\mathcal{AB}}, \mathcal{X}^{\text{RX}})$ modules contribute nearly equally to model performance. Specifically, CATH-ddG tends to offer a slightly greater contribution in scenarios involving multiple-point mutations, whereas inverse folding contributes more in single-point mutation cases. In contrast, the impact of data efficacy module (w/o Sampler) is less pronounced than that of the MoE module. Nevertheless, we observed that incorporating data efficacy improves generalization, which is evidenced in the case studies discussed above, thereby confirming its positive effect on model robustness.

4.5 Domain Specialization

Following [29], domain specialization characterizes the functional preference of expert $\mathcal{E}_i$ for CATH domain $\mathcal{C}_j$, representing the percentage of tokens from $\mathcal{C}_j$ assigned to $\mathcal{E}_i$, with 100% indicating exclusive dedication and 0% identifying prunable experts for that domain. Fig. 2 reveals that many experts are significantly activated deviating from the uniform routing baseline on specific CATH domains. A notable example is the EphA2-SAM/SHIP2-SAM complex, where C-terminal sterile alpha motifs (SAM) form a heterodimer that negatively regulates receptor endocytosis and degradation [41], resulting in the 2nd and 3rd experts in each layer specialize over 50%. This reflects that the domain specialization in USP-ddG results in low functional

overlap, with dedicated experts exhibiting distinct domain preferences, thereby effectively alleviating structural overlaps among protein domains as intended. In addition, as the E9 DNase/Im9 and E9 DNase/Im2 complexes share similar CATH domains mediated by structurally analogous PPIs, they exhibit closely aligned domain specializations. Likewise, EphA3/ephrinA5 and EphA4/ephrinB2 PPIs also exhibit similar domain specializations. Consequently, as shown in Fig. 2 below, the MoE framework demonstrates the ability to precisely delineate these domain-specific features, this highlights its advantage in identifying determinants of PPIs and showcasing its potential to advance the development of protein domain representation learning.

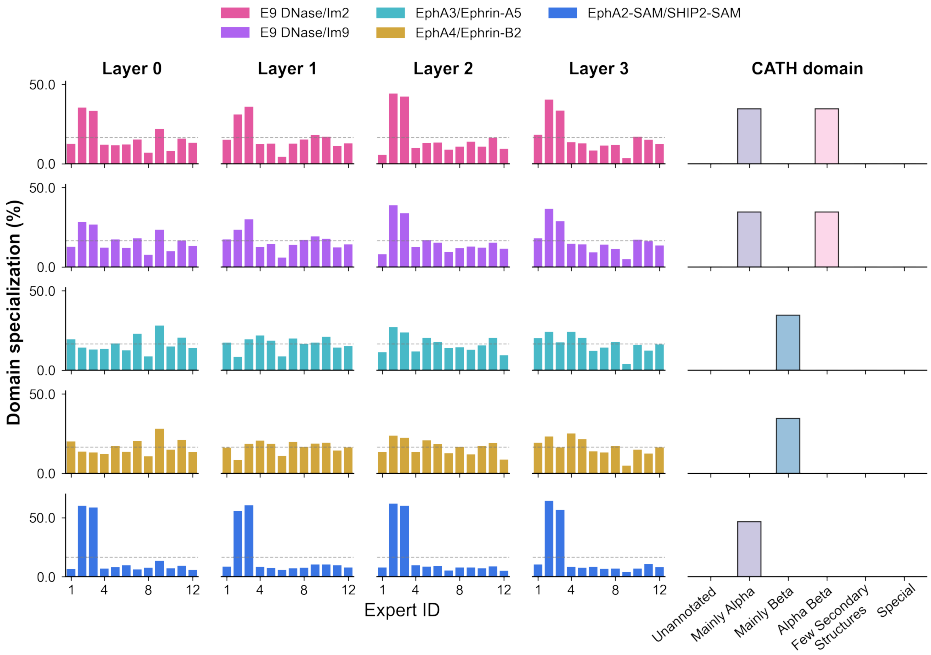

Fig. 2: Domain specialization of USP-ddG. We visualize the routing distribution of tokens from different protein domains across each layer to the 12 experts for the five PPIs in the hold-CATH test set, considering the top-2 active experts per token. The horizontal gray dashed lines represent the uniform routing baseline of 16.67% (2/12) per expert, which results from activating 2 out of 12 experts each token in every layer, providing reference for assessing the learned domain specialization.

5 Discussion

Predicting $\Delta\Delta G$ for PPIs with the unified structural framework is not only a critical computational task, but also facilitates to a comprehensive understanding of proteins as the ultimate executors of cellular structures and functions. However, current methods lack a systematic consensus on mutation-induced conformational changes and often overlook the necessity of establishing a unified structural paradigm. In this work, we present USP-ddG, a pioneering dual-channel architecture that significantly advances the prediction of the changes in binding free energy. By integrating structure-fixed and -relaxed assumptions via a hybrid layer of FFN and MoE, coupled with data efficacy training strategy, USP-ddG outperforms existing approaches and offers a robust tool for $\Delta\Delta G$ prediction.

Extensive experimental results demonstrate that USP-ddG achieves outstanding performance across various metrics. Our ablation studies further confirm the importance of each component to the overall performance. While USP-ddG represents a significant advancement, certain limitations remain. Two principal bottlenecks currently limit the accuracy and scalability of $\Delta\Delta G$ prediction. First, the scarcity of high-resolution complex structures necessitates the use of structure prediction tools such as AlphaFold3 [1]. Second, the reliance on tools like FoldX for generating mutant variants incurs computationally burden when applied at scale, emphasizing the need for end-to-end approaches capable of inferring mutational effects directly from wild-type structures. Addressing these challenges would enhance the efficiency and scalability of the proposed *in silico* antibody affinity maturation pipeline, thereby advancing therapeutic antibody engineering and contributing to disease therapy and biomedical research.

References

1. Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A.J., Bambrick, J., Bodenstein, S.W., Evans, D.A., Hung, C.C., O'Neill, M., Reiman, D., Tunyasuvunakool, K., Wu, Z., Žemgulytė, A., Arvaniti, E., Beattie, C., Bertolli, O., Bridgland, A., Cherepanov, A., Congreve, M., Cowen-Rivers, A.I., Cowie, A., Figurnov, M., Fuchs, F.B., Gladman, H., Jain, R., Khan, Y.A., Low, C.M.R., Perlin, K., Potapenko, A., Savy, P., Singh, S., Stecula, A., Thillaisundaram, A., Tong, C., Yakneen, S., Zhong, E.D., Zielinski, M., Žídek, A., Bapst, V., Kohli, P., Jaderberg, M., Hassabis, D., Jumper, J.M.: Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature **630**(8016), 493–500 (2024). https://doi.org/10.1038/s41586-024-07487-w
2. Atkins, P.W., Paula, J.D., Keeler, J.: Atkins' Physical Chemistry. Oxford University Press, New York, NY, United states of Amerima (2023)
3. Bachlechner, T., Majumder, B.P., Mao, H., Cottrell, G., McAuley, J.: ReZero is all you need: Fast convergence at large depth. In: Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence. pp. 1352–1361. UAI 2021, Baltimore, MD, USA (2021). https://doi.org/10.48550/arXiv.2003.04887
4. Barlow, K.A., Conchúir, S.Ó., Thompson, S., Suresh, P., Lucas, J.E., Heinonen, M., Kortemme, T.: Flex ddG: Rosetta ensemble-based estimation of changes in protein–protein binding affinity upon mutation. The journal of physical chemistry B **122**(21), 5389–5399 (2018). https://doi.org/10.1021/acs.jpcb.7b11367
5. Bushuiev, A., Bushuiev, R., Kouba, P., Filkin, A., Gabrielova, M., Gabriel, M., Sedlar, J., Pluskal, T., Damborsky, J., Mazurenko, S., Sivic, J.: Learning to design protein-protein interactions with enhanced generalization. In: Proceedings of The Twelfth International Conference on Learning Representations. ICLR 2024, Vienna, Austria (2024). https://doi.org/10.48550/arXiv.2310.18515
6. Cai, H., Zhang, Z., Wang, M., Zhong, B., Li, Q., Zhong, Y., Wu, Y., Ying, T., Tang, J.: Pretrainable geometric graph neural network for antibody affinity maturation. Nature Communications **15**(1), 7785 (2024). https://doi.org/10.1038/s41467-024-51563-8
7. Cuff, A.L., Sillitoe, I., Lewis, T., Redfern, O.C., Garratt, R., Thornton, J., Orengo, C.A.: The CATH classification revisited–architectures reviewed and new ways to characterize structural divergence in superfamilies. Nucleic Acids Research **37**(Database), D310–D314 (2009). https://doi.org/10.1093/nar/gkn877
8. Dai, D., Deng, C., Zhao, C., Xu, R., Gao, H., Chen, D., Li, J., Zeng, W., Yu, X., Wu, Y., Xie, Z., Li, Y., Huang, P., Luo, F., Ruan, C., Sui, Z., Liang, W.: DeepSeekMoE: Towards ultimate expert specialization in Mixture-of-Experts language models. In: Proceedings of The 62nd Annual Meeting of the Association for Computational Linguistics. vol. 1, pp. 1280–1297. ACL 2024, Bangkok, Thailand (2024). https://doi.org/10.18653/v1/2024.acl-long.70
9. Dai, Y., Huang, Y., Zhang, X., Wu, W., Li, C., Lu, W., Cao, S., Dong, L., Li, S.: Data efficacy for anguage model training. arXiv preprint (arXiv:2506.21545) (2025). https://doi.org/10.48550/arXiv.2506.21545
10. Dauparas, J., Anishchenko, I., Bennett, N., Bai, H., Ragotte, R.J., Milles, L.F., Wicky, B.I.M., Courbet, A., de Haas, R.J., Bethel, N., Leung, P.J.Y., Huddy, T.F., Pellock, S., Tischer, D., Chan, F., Koepnick, B., Nguyen, H., Kang, A., Sankaran, B., Bera, A.K., King, N.P., Baker, D.: Robust deep learning based protein sequence design using ProteinMPNN. Science **378**(6615), 49–56 (2022). https://doi.org/10.1126/science.add2187
11. Delgado, J., Radusky, L.G., Cianferoni, D., Serrano, L.: FoldX 5.0: Working with RNA, small molecules and a new graphical interface. Bioinformatics **35**(20), 4168–4169 (2019). https://doi.org/10.1093/bioinformatics/btz184
12. Eisenberg, D., Jucker, M.: The amyloid state of proteins in human diseases. Cell **148**(6), 1188–1203 (2012). https://doi.org/10.1016/j.cell.2012.02.022
13. Greenblatt, J.F., Alberts, B.M., Krogan, N.J.: Discovery and significance of protein-protein interactions in health and disease. Cell **187**(23), 6501–6517 (2024). https://doi.org/10.1016/j.cell.2024.10.038
14. Guerois, R., Nielsen, J.E., Serrano, L.: Predicting changes in the stability of proteins and protein complexes: A study of more than 1000 mutations. Journal of Molecular Biology **320**(2), 369–387 (2002). https://doi.org/10.1016/S0022-2836(02)00442-4
15. Härd, T., Lendel, C.: Inhibition of amyloid formation. Journal of Molecular Biology **421**(4-5), 441–465 (2012). https://doi.org/10.1016/j.jmb.2011.12.062
16. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models (2021). https://doi.org/10.48550/arXiv.2106.09685
17. Jacobs, R.A., Jordan, M.I., Nowlan, S.J., Hinton, G.E.: Adaptive Mixtures of Local Experts. Neural Computation **3**(1), 79–87 (1991). https://doi.org/10.1162/neco.1991.3.1.79
18. Jankauskaite, J., Jiménez-García, B., Dapkunas, J., Fernández-Recio, J., Moal, I.H.: SKEMPI 2.0: An updated benchmark of changes in protein–protein binding energy, kinetics and thermodynamics upon mutation. Bioinformatics **35**(3), 462–469 (2019). https://doi.org/10.1093/bioinformatics/bty635
19. Jiao, X., Mao, W., Jin, W., Yang, P., Chen, H., Shen, C.: Boltzmann-aligned inverse folding model as a predictor of mutational effects on protein-protein interactions. In: Proceedings of the Thirteenth International Conference

on Learning Representations. pp. 56810–56827. ICLR 2025, Singapore (2025). https://doi.org/10.48550/arXiv.2410.09543
20. Jin, W., Chen, X., Vetticaden, A., Sarzikova, S., Raychowdhury, R., Uhler, C., Hacohen, N.: DSMBind: SE(3) denoising score matching for unsupervised binding energy prediction and nanobody design. In: Proceedings of the NeurIPS 2023 Generative AI and Biology (GenBio) Workshop. NeurIPS 2023, New Orleans, LA, USA (2023). https://doi.org/10.1101/2023.12.10.570461
21. Kastritis, P.L., Bonvin, A.M.J.J.: On the binding affinity of macromolecular interactions: Daring to ask why proteins interact. Journal of the Royal Society Interface **10**(79), 20120835 (2013). https://doi.org/10.1098/rsif.2012.0835
22. Lapenna, S., Giordano, A.: Cell cycle kinases as therapeutic targets for cancer. Nature Reviews Drug Discovery **8**(7), 547–566 (2009). https://doi.org/10.1038/nrd2907
23. Liu, A., Feng, B., Wang, B., Wang, B., Liu, B., Zhao, C., Dengr, C., Ruan, C., Dai, D., Guo, D., Yang, D., Chen, D., Ji, D., Li, E., Lin, F., Luo, F., Hao, G., Chen, G., Li, G., Zhang, H., Xu, H., Yang, H., Zhang, H., Ding, H., Xin, H., Gao, H., Li, H., Qu, H., Cai, J.L., Liang, J., Guo, J., Ni, J., Li, J., Chen, J., Yuan, J., Qiu, J., Song, J., Dong, K., Gao, K., Guan, K., Wang, L., Zhang, L., Xu, L., Xia, L., Zhao, L., Zhang, L., Li, M., Wang, M., Zhang, M., Zhang, M., Tang, M., Li, M., Tian, N., Huang, P., Wang, P., Zhang, P., Zhu, Q., Chen, Q., Du, Q., Chen, R.J., Jin, R.L., Ge, R., Pan, R., Xu, R., Chen, R., Li, S.S., Lu, S., Zhou, S., Chen, S., Wu, S., Ye, S., Ma, S., Wang, S., Zhou, S., Yu, S., Zhou, S., Zheng, S., Wang, T., Pei, T., Yuan, T., Sun, T., Xiao, W.L., Zeng, W., An, W., Liu, W., Liang, W., Gao, W., Zhang, W., Li, X.Q., Jin, X., Wang, X., Bi, X., Liu, X., Wang, X., Shen, X., Chen, X., Chen, X., Nie, X., Sun, X., Wang, X., Liu, X., Xie, X., Yu, X., Song, X., Zhou, X., Yang, X., Lu, X., Su, X., Wu, Y., Li, Y.K., Wei, Y.X., Zhu, Y.X., Xu, Y., Huang, Y., Li, Y., Zhao, Y., Sun, Y., Li, Y., Wang, Y., Zheng, Y., Zhang, Y., Xiong, Y., Zhao, Y., He, Y., Tang, Y., Piao, Y., Dong, Y., Tan, Y., Liu, Y., Wang, Y., Guo, Y., Zhu, Y., Wang, Y., Zou, Y., Zha, Y., Ma, Y., Yan, Y., You, Y., Liu, Y., Ren, Z.Z., Ren, Z., Sha, Z., Fu, Z., Huang, Z., Zhang, Z., Xie, Z., Hao, Z., Shao, Z., Wen, Z., Xu, Z., Zhang, Z., Li, Z., Wang, Z., Gu, Z., Li, Z., Xie, Z.: DeepSeek-V2: A strong, economical, and efficient mixture-of-experts language model. arXiv preprint (arXiv:2405.04434) (2024). https://doi.org/10.48550/arXiv.2405.04434
24. Liu, S., Zhu, T., Ren, M., Chungong, Y., Bu, D., Zhang, H.: Predicting mutational effects on protein-protein binding via a side-chain diffusion probabilistic model. In: Proceedings of Advances in Neural Information Processing Systems 36. vol. 36, pp. 48994–49005. NeurIPS 2023, New Orleans, LA, USA (2023). https://doi.org/10.48550/arXiv.2310.19849
25. Liu, X., Luo, Y., Li, P., Song, S., Peng, J.: Deep geometric representations for modeling effects of mutations on protein-protein binding affinity. PLoS computational biology **17**(8), e1009284 (2021). https://doi.org/10.1371/journal.pcbi.1009284
26. Luo, S., Su, Y., Wu, Z., Su, C., Peng, J., Ma, J.: Rotamer density estimator is an unsupervised learner of the effect of mutations on protein-protein interaction. In: Proceedings of The Eleventh International Conference on Learning Representations. ICLR 2023, Kigali, Rwanda (2023). https://doi.org/10.1101/2023.02.28.530137
27. Mandell, D.J., Kortemme, T.: Computer-aided design of functional protein interactions. Nature Chemical Biology **5**(11), 797–807 (2009). https://doi.org/10.1038/nchembio.251
28. Marchand, A., Van Hall-Beauvais, A.K., Correia, B.E.: Computational design of novel protein-protein interactions - An overview on methodological approaches and applications. Current Opinion in Structural Biology **74**, 102370 (2022). https://doi.org/10.1016/j.sbi.2022.102370
29. Muennighoff, N., Soldaini, L., Groeneveld, D., Lo, K., Morrison, J., Min, S., Shi, W., Walsh, P., Tafjord, O., Lambert, N., Gu, Y., Arora, S., Bhagia, A., Schwenk, D., Wadden, D., Wettig, A., Hui, B., Dettmers, T., Kiela, D., Farhadi, A., Smith, N.A., Koh, P.W., Singh, A., Hajishirzi, H.: OLMoE: Open mixture-of-experts language models. In: Proceedings of the Thirteenth International Conference on Learning Representations. pp. 97116–97176. ICLR 2025, Singapore (2025). https://doi.org/10.48550/arXiv.2409.02060
30. Nallapareddy, V., Bordin, N., Sillitoe, I., Heinzinger, M., Littmann, M., Waman, V.P., Sen, N., Rost, B., Orengo, C.: CATHe: Detection of remote homologues for CATH superfamilies using embeddings from protein language models. Bioinformatics **39**(1), btad029 (2023). https://doi.org/10.1093/bioinformatics/btad029
31. Phillips, A.M., Lawrence, K.R., Moulana, A., Dupic, T., Chang, J., Johnson, M.S., Cvijovic, I., Mora, T., Walczak, A.M., Desai, M.M.: Binding affinity landscapes constrain the evolution of broadly neutralizing anti-influenza antibodies. eLife **10**, e71393 (2021). https://doi.org/10.7554/eLife.71393
32. Robinson, J., Sun, L., Yu, K., Batmanghelich, K., Jegelka, S., Sra, S.: Can contrastive learning avoid shortcut solutions? In: Proceedings of Advances in Neural Information Processing Systems 34. vol. 34, pp. 4974–4986. NeurIPS 2021, New Orleans, LA, USA (2021). https://doi.org/10.48550/arXiv.2106.11230
33. Schreiber, R.D., Old, L.J., Smyth, M.J.: Cancer immunoediting: Integrating immunity's roles in cancer suppression and promotion. Science **331**(6024), 1565–1570 (2011). https://doi.org/10.1126/science.1203486
34. Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., Serrano, L.: The FoldX web server: An online force field. Nucleic Acids Research **33**(Web Server), W382–W388 (2005). https://doi.org/10.1093/nar/gki387

35. Shan, S., Luo, S., Yang, Z., Hong, J., Su, Y., Ding, F., Fu, L., Li, C., Chen, P., Ma, J., Shi, X., Zhang, Q., Berger, B., Zhang, L., Peng, J.: Deep learning guided optimization of human antibody against SARS-CoV-2 variants with broad neutralization. Proceedings of the National Academy of Sciences of the United States of America **119**(11), e2122954119 (2022). https://doi.org/10.1073/pnas.2122954119
36. Shanehsazzadeh, A., Bachas, S., Kasun, G., Sutton, J.M., Steiger, A.K., Shuai, R., Kohnert, C., Morehead, A., Brown, A., Chung, C., Luton, B.K., Diaz, N., McPartlon, M., Knight, B., Radach, M., Bateman, K., Spencer, D.A., Cejovic, J., Kopec-Belliveau, G., Haile, R., Yassine, E., McCloskey, C., Natividad, M., Chapman, D., Stojanovic, L., Rakocevic, G., Hannum, G., Yapici, E., Moran, K., Caguiat, R., Abdulhaqq, S., Guo, Z., Klug, L.R., Gander, M., Meier, J.: Unlocking de novo antibody design with generative artificial intelligence. bioRxiv (2023). https://doi.org/10.1101/2023.01.08.523187
37. Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., Dean, J.: Outrageously large neural networks: The sparsely-gated mixture-of-experts layer (2017). https://doi.org/10.48550/arXiv.1701.06538
38. Soga, P., Lei, Z., He, Y., Bilodeau, C., Li, J.: Energy-based models for predicting mutational effects on proteins. In: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining. vol. 12, pp. 2597–2608. KDD 2025, Toronto, ON, Canada (2025). https://doi.org/10.1145/3711896.3736931
39. Ta, W., Stokes, J.M.: The importance of negative training data for robust antibody binding prediction. Nature Machine Intelligence **7**(8), 1192–1194 (2025). https://doi.org/10.1038/s42256-025-01080-0
40. Ursu, E., Minnegalieva, A., Rawat, P., Chernigovskaya, M., Tacutu, R., Sandve, G.K., Robert, P.A., Greiff, V.: Training data composition determines machine learning generalization and biological rule discovery. Nature Machine Intelligence **7**(8), 1206–1219 (2025). https://doi.org/10.1101/2024.06.17.599333
41. Vincenzi, M., Mercurio, F.A., Palumbo, R., La Manna, S., Pirone, L., Marasco, D., Pedone, E.M., Leone, M.: Inhibition of the EphA2-Sam/Ship2-Sam association through peptide ligands: Studying the combined effect of charge and aromatic character. Journal of Medicinal Chemistry **67**(18), 16649–16663 (2024). https://doi.org/10.1021/acs.jmedchem.4c01459
42. Wang, G., Liu, X., Wang, K., Gao, Y., Li, G., Baptista-Hon, D.T., Yang, X.H., Xue, K., Tai, W.H., Jiang, Z., Cheng, L., Fok, M., Lau, J.Y.N., Yang, S., Lu, L., Zhang, P., Zhang, K.: Deep-learning-enabled protein–protein interaction analysis for prediction of SARS-CoV-2 infectivity and variant evolution. Nature Medicine **29**(8), 2007–2018 (2023). https://doi.org/10.1038/s41591-023-02483-5
43. Wu, L., Tian, Y., Lin, H., Huang, Y., Li, S., Chawla, N.V., Li, S.Z.: Learning to predict mutation effects of protein-protein interactions by microenvironment-aware hierarchical prompt learning. In: The Forty-First International Conference on Machine Learning. ICML 2024, Vienna, Austria (2024). https://doi.org/10.48550/arXiv.2405.10348
44. Yang, F., Fan, K., Song, D., Lin, H.: Graph-based prediction of protein-protein interactions with attributed signed graph embedding. BMC Bioinformatics **21**(1), 323 (2020). https://doi.org/10.1186/s12859-020-03646-8
45. Yu, G., Bi, X., Ma, T., Li, Y., Wang, J.: CATH-ddG: Towards robust mutation effect prediction on protein–protein interactions out of CATH homologous superfamily. Bioinformatics **41**(Supplement_1), i362–i372 (2025). https://doi.org/10.1093/bioinformatics/btaf228
46. Yu, J., Zhuge, Y., Zhang, L., Hu, P., Wang, D., Lu, H., He, Y.: Boosting continual learning of vision-language models via mixture-of-experts adapters. In: Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23219–23230. IEEE, Seattle, WA, USA (2024). https://doi.org/10.1109/CVPR52733.2024.02191
47. Zhao, X., Tang, Y.C., Singh, A., Cantu, V.J., An, K., Lee, J., Stogsdill, A.E., Ramesh, A.K., An, Z., Jiang, X., Kim, Y.: Benchmark for antibody binding affinity maturation and design (2025). https://doi.org/10.48550/arXiv.2506.04235