Accurately Predicting Protein Mutational Effects via a Hierarchical Many-Body Attention Network

Dahao Xu[1,*], **Jiahua Rao**[1,*], **Mingming Zhu**[1], **Jixian Zhang**[2], **Wei Lu**[2], **Shuangjia Zheng**[3,†], **Yuedong Yang**[1,†]
[*]Equal Contribution [†]Corresponding Authors
[1]Sun Yat-sen University [2]Aureka Biotechnologies [3]Shanghai Jiao Tong University
shuangjia.zheng@sjtu.edu.cn
yangyd25@mail.sysu.edu.cn

Abstract

Predicting changes in binding free energy ($\Delta\Delta G$) is essential for understanding protein-protein interactions, which are critical in drug design and protein engineering. However, existing methods often rely on pre-trained knowledge and heuristic features, limiting their ability to accurately model complex mutation effects, particularly higher-order and many-body interactions. To address these challenges, we propose **H3-DDG**, a **H**ypergraph-driven **H**ierarchical network to capture **H**igher-order many-body interactions across multiple scales. By introducing a hierarchical communication mechanism, H3-DDG effectively models both local and global mutational effects. Experimental results demonstrate state-of-the-art performance on multiple benchmarks. On the SKEMPI v2 dataset, H3-DDG achieves a Pearson correlation of 0.75, improving multi-point mutations prediction by 12.10%. On the challenging BindingGYM dataset, it outperforms Prompt-DDG and BA-DDG by 62.61% and 34.26%, respectively. Ablation and efficiency analyses demonstrate its robustness and scalability, while a case study on SARS-CoV-2 antibodies highlights its practical value in improving binding affinity for therapeutic design.

1 Introduction

Protein-protein interactions (PPIs) [19, 9, 12, 32] are fundamental to numerous biological processes, driving key cellular functions such as signal transduction [21], immune response [35, 22], and metabolic regulation [40]. A precise understanding of how mutations alter binding free energy ($\Delta\Delta G$) in PPIs is critical for a wide range of applications, including drug design [2, 7, 20, 31], protein engineering [5], and elucidating the molecular basis of disease [33].

Binding free energy quantifies the thermodynamic stability of protein complexes and is inherently governed by the physical interactions between amino acids. These interactions span multiple spatial and structural scales, from pairwise atomic forces—such as hydrogen bonding and van der Waals interactions—to higher-order many-body effects, including hydrogen bond networks and $\pi - \pi$ stacking interactions. Accurately modeling these intricate interactions is essential for predicting the functional and structural consequences of mutations [11, 13]. This becomes particularly challenging in multi-point mutation scenarios, where complex interdependencies between mutation sites often emerge, further complicating the prediction task.

Accurately predicting $\Delta\Delta G$ remains a critical yet challenging task, as existing computational methods face inherent limitations. These methods can be categorized into structure-based approaches and inverse folding-based models. Structure-based approaches leverage well-designed training tasks, such as protein inverse folding [39], side-chain modeling [26, 23, 28], masked modeling [37], and data augmentation [38], to extract protein representations from structural data. However, they often

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

fall short in effectively transferring the learned structural knowledge to $\Delta\Delta G$ prediction. This shortcoming arises from their limited capacity to explicitly model higher-order and many-body interactions, which are crucial for capturing the physicochemical impacts of mutations.

On the other hand, inverse folding-based models, such as ProteinMPNN-DDG [10] and BA-DDG [18], predict $\Delta\Delta G$ by estimating likelihood differences between native and mutant sequences relative to stability. While these models perform well for single-point mutations, they rely on indirect proxies, such as Boltzmann-Alignment, to approximate binding free energy. This reliance limits their effectiveness in multi-point mutation scenarios, which involve intricate interdependencies among mutation sites and require a deeper understanding of protein energetics and many-body interaction dynamics. These limitations underscore the need for an approach capable of capturing higher-order interactions and adapting to the complexity of multi-point mutation scenarios.

In this work, we introduce **H3-DDG**, a **H**ypergraph-driven **H**ierarchical network to capture **H**igher-order many-body interactions across multiple scales for $\Delta\Delta G$ predictions. By leveraging a many-body attention communication mechanism, H3-DDG effectively models higher-order and many-body interactions across multiple scales, enabling precise predictions of binding free energy changes, particularly in challenging multi-point mutation contexts.

Experimental results demonstrate state-of-the-art performance across multiple benchmarks. On the SKEMPI v2 dataset [17], H3-DDG achieves a Pearson correlation of 0.75, improving multi-point mutations prediction by 12.10%. On the challenging BindingGYM dataset [24], it outperforms Prompt-DDG and BA-DDG by 62.61% and 34.26%, respectively. Notably, in ablation studies, the many-body attention mechanism proves to be critical, contributing a 5.4% improvement in prediction accuracy when incorporated. Additionally, the hypergraph-driven hierarchical design enables the model to effectively capture long-range dependencies, which are essential for accurately modeling complex mutation scenarios. A case study on SARS-CoV-2 antibodies further highlights H3-DDG's ability to predict binding affinity changes with high precision, demonstrating its practical utility in real-world applications. These results underscore the capability of H3-DDG to handle intricate mutational landscapes and its potential for broad applicability in protein engineering and drug design. The key contributions of this work are as follows:

- We introduce a hierarchical communication mechanism to model local and global interactions, capturing higher-order effects like hydrogen bond networks and $\pi - \pi$ stacking.
- We propose a many-body attention network to explicitly model higher-order and many-body interactions, enabling robust predictions in complex mutational scenarios.
- H3-DDG achieves state-of-the-art performance on multiple benchmark datasets, outperforming BA-DDG by 12.10% in multi-point mutation prediction on the SKEMPI v2 dataset and by 34.26% on the BindingGYM dataset.

2 Related Work

Efforts to predict the change in binding free energy ($\Delta\Delta G$) have resulted in a variety of computational approaches, broadly categorized into two main types: (1) methods leveraging structure-based training tasks and (2) methods utilizing inverse folding models for $\Delta\Delta G$ prediction. While these approaches have shown promise, they also highlight significant limitations, particularly in capturing higher-order interactions and complex multi-point mutation effects [30, 32].

The first category focuses on designing training tasks that utilize structural information to enhance the modeling of protein energetics. These include techniques such as protein inverse folding [39], side-chain modeling [26, 23, 28], and masked modeling [37]. These approaches attempt to extract meaningful representations of protein structures by optimizing for tasks aligned with physical principles of molecular interactions. However, their reliance on pre-trained knowledge limits their ability to explicitly model higher-order and many-body interactions, which are critical for understanding the thermodynamic effects of mutations. This gap restricts their utility in directly predicting $\Delta\Delta G$, particularly in complex mutation scenarios.

The second category focuses on adapting inverse folding models to predict $\Delta\Delta G$ [25]. Models such as ProteinMPNN-DDG [10] and BA-DDG [18] leverage the ability of inverse folding models to assess the likelihood of native and mutant amino acids within a structural context. These methods predict $\Delta\Delta G$ by estimating the differences in likelihood scores between native and mutant sequences with

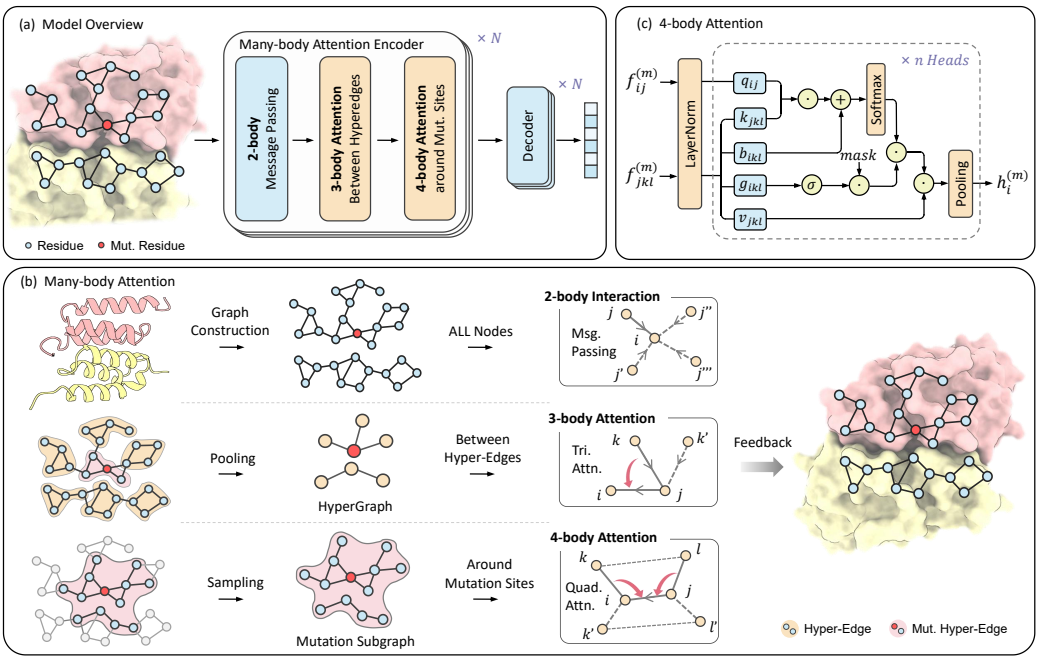

Figure 1: (a) Overall framework of H3-DDG. (b) Many-body attention mechanisms across hierarchical levels and their attention patterns. The first layer constructs a full-residue graph and applies 2-body message passing; the second layer performs spatial pooling to form a mutation-centered hypergraph and applies 3-body attention between hyperedges; the third layer extracts fine-grained subgraphs around mutation sites and applies localized 4-body attention within each mutation subgraph. (c) Detailed computational flow of the 4-body attention module within the mutation subgraph.

their relative stability. However, likelihood-based scoring serves as an indirect proxy for binding free energy changes and often fails to capture explicit physical interactions critical to $\Delta\Delta G$ prediction. Moreover, their performance degrades significantly in multi-point mutation scenarios, where the interdependencies between mutations require a nuanced understanding of higher-order effects.

To address these challenges, we introduces a hierarchical communication mechanism and a many-body attention network to explicitly model local and global interactions across multiple scales. These innovations capture higher-order and many-body effects, such as hydrogen bond networks and $\pi - \pi$ stacking, while explicitly modeling mutation interdependencies in multi-point mutation scenarios.

3 Methodology

In this section, we introduce H3-DDG, a Hypergraph-driven, Hierarchical, and Higher-order interaction network designed to predict binding free energy changes ($\Delta\Delta G$) in protein-protein interactions (PPIs). The overall workflow is illustrated in Figure 1.

3.1 Preliminaries and Notations

We represent a protein graph as $G = (V, E)$, where each node $i \in V$ represents a residue with feature vector $\mathbf{h}_i$, and each edge $(i, j) \in E$ encodes interactions via feature vector $\mathbf{e}_{ij}$.

Building on ProteinMPNN-DDG [10] and BA-DDG [18], which link $\Delta\Delta G$ prediction with inverse folding, we also fine-tune a pretrained inverse folding model (ProteinMPNN [6]) to tackle this task. H3-DDG extends the capabilities of the inverse folding model by introducing hierarchical graph construction and many-body higher-order attention, enabling it to effectively capture complex dependencies across multiple mutation sites and improve protein representation learning.

To predict binding free energy ΔG, we compute amino acid probabilities and use negative log-likelihoods as a surrogate. For a wild-type (wt) complex and its mutant (mut) counterpart, the binding

free energy change is defined as:

$$\Delta\Delta G = \Delta G_{\text{mut}} - \Delta G_{\text{wt}}. \tag{1}$$

3.2 Hierarchical Graph Construction for Multi-Scale Modeling

Residue-Level Graph. As described in Section 3.1, each residue i is represented by a feature vector $\mathbf{h}_i$, and each edge (i, j) encodes spatial or chemical interactions via feature vector $\mathbf{e}_{ij}$.

Following the approch of ProteinMPNN [6], residues are represented using key atoms: N, Cα, C, O, and Cβ, To construct the amino acid-level graph. Edges are formed by identifying the top-k nearest neighbors based on Cα–Cα distances. Each edge (i, j) incorporates geometric features using radial basis functions (RBFs) [4] over 25 atom pair distances (e.g., N–N, C–C). For a given distance d_{ij}, the RBF is defined as:

$$\text{RBF}_{ij}^{(k)} = \exp\left(-\left(\frac{d_{ij} - \mu_k}{\sigma}\right)^2\right), \quad k = 1, \ldots, K, \tag{2}$$

where μ_k are K predefined centers and σ is the bandwidth.

These RBF_{ij} features are concatenated with relative positional encodings PE_{ij}, which include sequence offsets and chain identities. The final edge features $\mathbf{e}_{ij}$ are computed as:

$$\mathbf{e}_{ij} = \text{LayerNorm}\left(W_e \cdot [\text{PE}_{ij} \,\|\, \text{RBF}_{ij}]\right), \tag{3}$$

where W_e is the learnable parameters and $[\cdot \,\|\, \cdot]$ denotes concatenation. This graph effectively captures residue-level spatial and chemical interactions while preserving sequence context.

Mutation-Centered Hypergraph. Directly applying higher-order attention on amino acid-level graphs is computationally expensive, with complexity scaling as $\mathcal{O}(N^3)$ or $\mathcal{O}(N^4)$, as demonstrated in [16, 30]. To address this challenge, we construct a mutation-centered hypergraph that reduces computational complexity while preserving critical local and global interactions.

We construct the hypergraph by clustering residues into hyperedges that capture higher-order spatial dependencies. Clustering is initialized at mutation sites to prioritize biologically salient regions, then expanded by iteratively adding residues with maximal Euclidean distance from existing centroids. Each residue is assigned to its nearest centroid based on Cα coordinates. Hyperedge features $\mathbf{h}^{(e)}$ are computed via mean-pooling over constituent residue embeddings $\mathbf{h}$, yielding a compact representation that preserves structural topology and improves computational efficiency.

This mutation-centered hypergraph construction ensures that the influence of mutation sites is retained during graph compression and facilitates the efficient modeling of higher-order interactions. For further implementation details, refer to Appendix A.1.

Fine-Grained Mutation Subgraph. To capture the local effects of mutations, we extract fine-grained subgraphs centered on each mutation site. These subgraphs include all residues with Cα atoms within an 8Å radius of the mutation centroid in 3D space, preserving key structural and biochemical contexts. Node and edge features in these subgraphs, denoted as $\mathbf{h}^{(m)}$ and $\mathbf{e}^{(m)}$, respectively, enable localized analysis of conformational changes and energetic impacts, providing a focused view of mutation-induced effects on binding affinity.

3.3 Many-Body Attention Network

In this section, we describe the many-body attention mechanisms applied at different hierarchical levels of the graph. These include pairwise (2-body) message passing over the residue-level graph, triplet (3-body) attention between hyperedges in the mutation-centered hypergraph, and quadruplet (4-body) attention within fine-grained mutation subgraphs, as illustrated in Figure 1(b).

2-body Message Passing in Residue-level Graph. We utilize a message passing mechanism to jointly update node features $\mathbf{h}_i$ and edge features $\mathbf{e}_{ij}$ in the residue-level graph (Section 3.2). Node-wise messages $\Delta\mathbf{h}_i$ are computed by projecting concatenated node and edge features through

a multi-layer perceptron (MLP):

$$\Delta \mathbf{h}_i = \sum_{j \in \mathcal{N}(i)} W_{m3} \cdot \phi \left(W_{m2} \cdot \phi \left(W_{m1} \cdot [\mathbf{h}_i \,\|\, \mathbf{e}_{ij}] \right) \right), \tag{4}$$

where $[\cdot \,\|\, \cdot]$ denotes vector concatenation, ϕ is the GELU activation, and W_{m1}, W_{m2}, W_{m3} are learnable projection matrices. The node features $\mathbf{h}_i$ are updated via a residual edge with layer normalization and a position-wise feed-forward network (FFN):

$$\mathbf{h}_i \leftarrow \text{LayerNorm} \left(\mathbf{h}_i + \text{Dropout}(\Delta \mathbf{h}_i) \right), \tag{5}$$

$$\mathbf{h}_i \leftarrow \text{LayerNorm} \left(\mathbf{h}_i + \text{Dropout}(\text{FFN}(\mathbf{h}_i)) \right). \tag{6}$$

Edge features $\mathbf{e}_{ij}$ are updated using the updated node features through a similar MLP-based mechanism:

$$\mathbf{e}_{ij} \leftarrow \text{LayerNorm} \left(\mathbf{e}_{ij} + \text{Dropout} \left(W'_{m3} \cdot \phi \left(W'_{m2} \cdot \phi \left(W'_{m1} \cdot [\mathbf{h}_i \,\|\, \mathbf{e}_{ij}] \right) \right) \right) \right). \tag{7}$$

This mechanism enables the model to integrate residue-level spatial and chemical interactions efficiently across the residue-level graph.

3-body Attention Mechanism for Hyperedge Interactions. Building on the global structural context learned from 2-body message passing in the full-residue graph, we implement a 3-body attention mechanism to capture higher-order interactions between hyperedges in the mutation-centered hypergraph (Section 3.2). Here, each hyperedge is represented by its feature vector $\mathbf{h}^{(e)}$, capturing the residue cluster-level information.

To model pairwise relationships between hyperedges, we first construct a pairwise interaction tensor $f_{ij}^{(e)} \in \mathbb{R}^{N \times N \times D}$, where N is the number of hyperedges and D is the feature dimension. Here, i and j index hyperedges in the hypergraph. Each element $f_{ij}^{(e)}$ is computed as:

$$f_{ij}^{(e)} = \phi(\mathbf{h}_i^{(e)}) \otimes \phi(\mathbf{h}_j^{(e)}), \tag{8}$$

where ϕ is a non-linear activation function (e.g., GELU) and $\otimes$ denotes the outer product.

An asymmetric attention mechanism is employed by linearly projecting $f_{ij}^{(e)}$ through learnable weight matrices and biases to obtain the query vector $\mathbf{q}_{ij} = W_q f_{ij}^{(e)} + b_q$, key vector $\mathbf{k}_{jk} = W_k f_{jk}^{(e)} + b_k$, and value vector $\mathbf{v}_{jk} = W_v f_{jk}^{(e)} + b_v$. In addition, the mechanism introduces a bias term $\mathbf{b}_{ik}$ and a gating vector $\mathbf{g}_{ik}$ via separate linear projections for adaptive modulation. The triplet attention scores a_{ijk} are then computed using a gated softmax function:

$$a_{ijk} = \text{Softmax} \left(\frac{\mathbf{q}_{ij} \cdot \mathbf{k}_{jk}}{\sqrt{d}} + \mathbf{b}_{ik} \right) \cdot \sigma(\mathbf{g}_{ik}), \tag{9}$$

where Softmax ensures that the attention weights are normalized over hyperedge neighbors, and $\sigma(\mathbf{g}_{ik})$ applies a sigmoid activation to adaptively modulate the attention.

Each hyperedge aggregates attention-weighted messages from its spatial neighbors, and the aggregated features are fused with the corresponding full-residue node feature $\mathbf{h}_{r(i)}$ via residual addition, followed by layer normalization:

$$\mathbf{h}_{r(i)} \leftarrow \text{LayerNorm} \left(\mathbf{h}_{r(i)} + \text{Dropout} \left(\sum_{j \in \mathcal{N}(i)} \sum_{k} a_{ijk} \cdot \mathbf{v}_{jk} \right) \right). \tag{10}$$

Here, $r(i)$ denotes the mapping from hyperedge index i to the corresponding residue index in the residue-level graph.

4-body Attention Mechanism around Mutation Sites. To capture higher-order interactions in the mutation environment, we extend the 3-body attention mechanism to 4-body attention on the Fine-Grained Mutation Subgraph (Section 3.2), where $\mathbf{h}^{(m)}$ and $\mathbf{e}^{(m)}$ denote the node and edge features of the mutation subgraph. This subgraph focuses on residues within an 8Å radius of the mutation site, enabling the model to learn localized interactions among residue quadruples. The

extension increases computational complexity by only a constant factor, preserving scalability while capturing more complex interactions. The full computational process is illustrated in Figure 1(c).

Specifically, we construct two interaction tensors to encode localized multi-residue interactions as inputs to the 4-body attention mechanism:

$$f_{ij}^{(m)} = \phi(\mathbf{h}_i^{(m)}) \odot \phi(\mathbf{h}_j^{(m)}), \quad f_{jkl}^{(m)} = \phi(\mathbf{h}_j^{(m)}) \odot \phi(\mathbf{e}_{kl}^{(m)}). \tag{11}$$

Here, ϕ is a non-linear activation function (e.g., GELU), and $\otimes$ denotes the outer product. The tensors $f_{ij}^{(m)} \in \mathbb{R}^{N\times N\times D}$ and $f_{jkl}^{(m)} \in \mathbb{R}^{N\times E\times D}$ respectively encode node-node and node-edge interactions, where N and E are the number of residues and edges on the mutation subgraph, and D is the feature dimension. Indices i, j, k, l refer to nodes within the mutation subgraph.

Inspired by cross-attention, the tensors $f_{ij}^{(m)}$ and $f_{jkl}^{(m)}$ are projected into query, key, and value vectors:

$$\mathbf{q}_{ij} = W_q \cdot f_{ij}^{(m)}, \quad \mathbf{k}_{jkl} = W_k \cdot f_{jkl}^{(m)}, \quad \mathbf{v}_{jkl} = W_v \cdot f_{jkl}^{(m)} \tag{12}$$

$$\mathbf{b}_{ikl} = W_b \cdot f_{ikl}^{(m)}, \quad \mathbf{g}_{ikl} = W_g \cdot f_{ikl}^{(m)}, \tag{13}$$

where W_q, W_k, W_v, W_b, and W_g are learnable projection matrices.

The quadruplet attention scores a_{ijkl} are computed via a gated softmax function:

$$a_{ijkl} = \text{Softmax}\left(\frac{\mathbf{q}_{ij} \cdot \mathbf{k}_{jkl}}{\sqrt{d}} + \mathbf{b}_{ikl}\right) \cdot \sigma(\mathbf{g}_{ikl}). \tag{14}$$

Each node in the mutation subgraph aggregates attention-weighted messages from its spatial neighbors, and the aggregated features are fused with the original node features via a residual connection, followed by dropout and layer normalization:

$$\mathbf{h}_i^{(m)} \leftarrow \text{LayerNorm}\left(\mathbf{h}_i^{(m)} + \text{Dropout}\Big(\sum_{j\in\mathcal{N}(i)} \sum_{k,l} a_{ijkl} \cdot \mathbf{v}_{jkl}\Big)\right). \tag{15}$$

This 4-body attention mechanism enables the model to encode complex, localized interaction motifs around mutation sites, thereby improving its ability to predict binding free energy changes with structural and biophysical fidelity.

3.4 Prediction Module and Learning Objective

Following the multi-scale representation learning with many-body attention in Section 3.3, we obtain a residue-level representation $\mathbf{h} \in \mathbb{R}^{N\times d}$. Let $s = \{s_1, \ldots, s_N\}$ be the amino acid sequence. Inspired by autoregressive models like ProteinMPNN, we model $P(s_i \mid s_{<i}, \mathbf{h})$ to compute the negative log-likelihoods $\mathcal{E}_{\text{wt}}$ and $\mathcal{E}_{\text{mut}}$ for wild-type and mutant sequences. As in RDE-Network [26] and BA-DDG [18], single-chain contributions are subtracted to isolate binding effects, yielding energy-like scores:

$$\Delta\Delta G_{\text{pred}} = \Delta G_{\text{mut}} - \Delta G_{\text{wt}} = \left(\mathcal{E}_{\text{mut}}^{\text{complex}} - \mathcal{E}_{\text{mut}}^{\text{monomer}}\right) - \left(\mathcal{E}_{\text{wt}}^{\text{complex}} - \mathcal{E}_{\text{wt}}^{\text{monomer}}\right). \tag{16}$$

We minimize the mean squared error between $\Delta\Delta G_{\text{pred}}$ and $\Delta\Delta G_{\text{true}}$ over n training samples:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{n}\sum_{i=1}^{n} \left(\Delta\Delta G_{\text{pred}} - \Delta\Delta G_{\text{true}}\right)^2. \tag{17}$$

4 Experiments

4.1 Experimental Settings

Datasets. We used **SKEMPI v2** [17], a benchmark with 7,085 mutations across 348 protein complexes, to evaluate $\Delta\Delta G$ prediction. Following prior work [26, 37], we split the data into three non-overlapping folds by complex to avoid data leakage. Additionally, we evaluated on **BindingGYM** [24], the largest dataset for protein-protein interactions, with 508,962 curated entries and a high proportion of multi-point mutations. We used the hardest inter-assay split, focusing on the fold with the most multi-point mutations for testing (details in Appendix A.2).

Table 1: Mean results of 3-fold cross-validation on SKEMPI v2 under single-, multi-, and all-point mutations. **Bold** and underline indicate the best and second-best results.

Method	Mutations	Overall					Per-Structure	
		Pearson↑	Spear.↑	RMSE↓	MAE↓	AUROC↑	Pearson↑	Spear.↑
Rosetta	all	0.3113	0.3468	1.6173	1.1311	0.6562	0.3284	0.2988
	single	0.3250	0.3670	1.1830	0.9870	0.6740	0.3510	0.4180
	multiple	0.1990	0.2300	2.6580	2.0240	0.6210	0.1910	0.0830
FoldX	all	0.3120	0.4071	1.9080	1.3089	0.6582	0.3789	0.3693
	single	0.3150	0.3610	1.6510	1.1460	0.6570	0.3820	0.3600
	multiple	0.2560	0.4180	2.6080	1.9260	0.7040	0.3330	0.3400
RDE-Network	all	0.6447	0.5584	1.5799	1.1123	0.7454	0.4448	0.4010
	single	0.6421	0.5271	1.3333	0.9392	0.7367	0.4687	0.4333
	multiple	0.6288	0.5900	2.0980	1.5747	0.7749	0.4233	0.3926
DiffAffinity	all	0.6609	0.5560	1.5350	1.0930	0.7440	0.4220	0.3970
	single	0.6720	0.5230	1.2880	0.9230	0.7330	0.4290	0.4090
	multiple	0.6500	0.6020	2.0510	1.5400	0.7840	0.4140	0.3870
Prompt-DDG	all	0.6772	0.5910	1.5207	1.0770	0.7568	0.4712	0.4257
	single	0.6596	0.5450	1.3072	0.9191	0.7355	0.4736	0.4392
	multiple	0.6780	0.6433	1.9831	1.4837	0.8187	0.4448	0.3961
ProMIM	all	0.6720	0.5730	1.5160	1.0890	0.7600	0.4640	0.4310
	single	0.6680	0.5340	1.2790	0.9240	0.7380	0.4660	0.4390
	multiple	0.6660	0.6140	1.9630	1.4910	0.8250	0.4580	0.4250
BA-DDG	all	0.7118	0.6346	1.4516	1.0151	0.7726	0.5453	0.5134
	single	0.7321	0.6157	1.1848	0.8409	0.7662	0.5606	0.5192
	multiple	0.6650	0.6293	2.0151	1.4944	0.7875	0.4924	0.4959
H3-DDG	all	**0.7501**	**0.6604**	**1.3665**	**0.9612**	**0.7920**	**0.5686**	**0.5281**
	single	**0.7471**	**0.6374**	**1.1560**	**0.8080**	**0.7803**	**0.5750**	**0.5295**
	multiple	**0.7341**	**0.6913**	**1.8320**	**1.3880**	**0.8309**	**0.5520**	**0.5323**
$\Delta_{\text{BA-DDG}}$	multiple	+10.39%	+9.85%	+9.08%	+7.12%	+5.51%	+12.10%	+7.34%

Baselines. We compared H3-DDG against unsupervised and supervised methods. Unsupervised approaches include energy-based models (e.g., Rosetta [1], FoldX [8]), evolutionary sequence models (e.g., ESM-1v [27], Tranception [29]), and structure-guided pretrained models (e.g., ESM-IF [15], MIF-Δlogits [39]). Supervised methods include end-to-end architectures (e.g., DDGPred [34]) and pretraining-finetuning frameworks (e.g., MIF-Network [39], RDE-Network [26], DiffAffinity [23], Prompt-DDG [37], ProMIM [28], Surface-VQMAE [36], MSM-Mut [14], BA-DDG [18]).

Metrics. We use five metrics: Pearson, Spearman, RMSE, MAE, and AUROC. For SKEMPI, in addition to the global metrics, we also report per-structure metrics. For BindingGYM, metrics are reported per-DMS (deep mutational scanning). Detailed metric definitions and calculations are provided in Appendix A.3.

Implementation details. The hyperparameters and hardware details are provided in Appendix A.4. The code is available at `https://github.com/biomed-AI/H3-DDG`.

4.2 Results on SKEMPI v2 Dataset

As shown in Table 1, H3-DDG outperforms existing baseline approaches across all evaluation metrics in the all-, single-, and multi-point mutation scenarios. Notably, in practical applications where affinity modulation through multi-point amino acid mutations is particularly critical, H3-DDG demonstrates significant advantages in the multi-mutation prediction task: achieving a Spearman correlation coefficient of 0.7341, which is 10.39% higher than the second-best method; at the per-structure level, this metric reaches 0.5520, surpassing the second-best approach by 12.10%. Additional comparisons with more baseline methods are provided in Appendix B.2, Table 6.

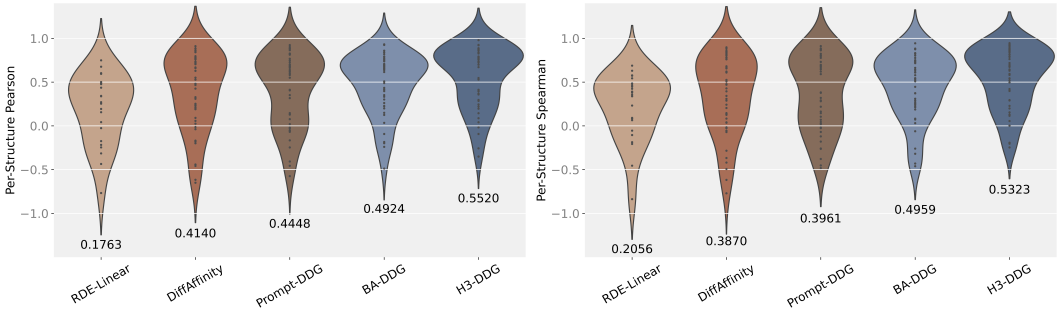

Figure 2: Distribution of per-structure Pearson and Spearman correlation coefficients for multi-point mutations, evaluated across five representative methods.

Table 2: Performance comparison under <3, ≥ 3 and all-point mutations on BindingGYM, where **bold** and underline denotes the best and second-best results under each setting.

Method	Mutations	Per-DMS			
		Pearson↑	Spearman↑	AUROC↑	RMSE↓
ProteinMPNN	ALL	0.0998	0.2050	0.5341	3.4974
	<3	0.1137	0.2439	0.5404	3.2328
	≥ 3	0.0734	0.1614	0.5930	5.5921
BA-Cycle	ALL	0.1320	0.1217	0.5658	1.2419
	<3	0.1385	0.1386	0.5620	1.0925
	≥ 3	0.0830	0.1955	0.6190	2.5822
Prompt-DDG	ALL	0.1880	0.1818	0.5198	1.5216
	<3	0.2160	0.2179	0.5273	1.3499
	≥ 3	0.1841	0.2008	0.6070	3.5747
BA-DDG	ALL	0.2277	0.2142	0.5310	**1.1182**
	<3	0.2553	0.2471	0.5395	**0.9716**
	≥ 3	0.2037	0.2259	0.6307	2.5191
H3-DDG	ALL	**0.3057**	**0.2725**	**0.5703**	1.1294
	<3	**0.3322**	**0.3031**	**0.5745**	1.0758
	≥ 3	**0.2472**	**0.2755**	**0.6734**	**2.4976**
$\Delta_{\text{BA-DDG}}$	≥ 3	+21.35%	+21.96%	+6.77%	+0.85%

These results are enabled by our introduction of a hierarchical communication mechanism combined with a many-body attention network. This proposed framework effectively models both local and global interactions, capturing higher-order effects such as $\pi - \pi$ stacking and explicitly addressing complex synergistic and many-body interactions in mutational scenarios. As further illustrated in Figure 2, the per-structure distributions of Pearson and Spearman correlation coefficients across five representative methods show that H3-DDG achieves significantly higher correlation values, with distributions concentrated in high-correlation regions, highlighting its robustness compared to other approaches.

4.3 Results on BindingGYM Dataset

To validate the robustness of H3-DDG, we evaluate it on the larger and more challenging BindingGYM dataset, demonstrating its scalability and generalizability. As shown in Table 2, H3-DDG outperforms all baseline methods across all metrics, achieving the highest Pearson correlation (0.3057), which surpasses the second-best method (BA-DDG) by 34.26%. Particularly in the ≥ 3 mutation scenario, it also achieves the best Spearman correlation (0.2755), exceeding the second-best method by 21.96%. Additionally, H3-DDG achieves the best AUROC across all mutations and demonstrates competitive RMSE. These results highlight H3-DDG's robustness, scalability, and effectiveness for accurately predicting protein-protein binding in complex mutational landscapes, especially in multi-point mutation scenarios.

While H3-DDG's RMSE under single-point mutations is slightly higher than the baseline methods (e.g., BA-DDG), this can be attributed to the BindingGYM dataset spanning different DMS experiments, where the absolute $\Delta\Delta G$ values vary significantly across experiments. However, we focus more on the ranking and correlation within each DMS experiment, making the Pearson and Spearman metrics more critical for evaluation.

4.4 Ablation Study

Ablation Study on Pooling Types. As shown in Table 3, H3-DDG outperforms other graph pooling methods like DiffPool [41] and MinCutPool [3], which rely on predicted assignment matrices and fail to predefine mutation sites as central nodes. By explicitly designating mutation sites as cluster centers, H3-DDG effectively captures long-range dependencies, enabling superior performance in modeling many-body interactions.

Table 3: Ablation study of pooling and many-body attention mechanisms under multi-point mutations on SKEMPI v2.

Pooling Type	Attn. Around Mut. Sites	Overall			
		Pearson↑	Spear.↑	RMSE↓	MAE↓
BA-DDG		0.6650	0.6293	2.0151	1.4944
DiffPool	–	0.6959	0.6558	1.9375	1.4843
MinCutPool	–	0.7004	0.6589	1.9256	1.4693
H3-DDG	–	0.7040	0.6676	1.9161	1.4528
	3-body Attn.	<u>0.7144</u>	<u>0.6732</u>	<u>1.8880</u>	<u>1.4127</u>
	4-body Attn.	**0.7341**	**0.6913**	**1.8320**	**1.3880**

Ablation Study on Attention Mechanisms. Table 9 also demonstrates the importance of many-body attention in $\Delta\Delta G$ prediction. While 3-body attention improves performance by modeling local interactions around mutation sites, extending to 4-body attention further enhances accuracy by capturing more complex dependencies. These findings highlight many-body attention as a key factor in H3-DDG's strong performance and efficient modeling. Our model strikes a balance between efficiency and performance, influenced by two main factors: the number of hyperedges in the hypergraph and the number of edges in the 4-body attention. Notably, our method achieves a Pearson correlation of 0.7341, surpassing BA-DDG's 0.6650, with a relatively small efficiency trade-off. Additional details are provided in Appendix B.4 and B.5.

4.5 SARS-CoV-2 Antibody Optimization

Predicting $\Delta\Delta G$ is essential for identifying affinity-enhancing mutations. We target five SARS-CoV-2 neutralizing mutations [34] within the heavy-chain CDRs (26 residues, 494 single-point variants). Models are fine-tuned on SKEMPI v2.0 to rank mutations by predicted $\Delta\Delta G$, with lower values indicating stronger binding. Table 4 benchmarks performance against top baselines, highlighting mutations ranked in the top 10%. Only our model achieves an average rank below 10%, demonstrating strong generalization and practical utility in antibody design.

Table 4: Rankings of the five favorable mutations on the human antibody against SARS-CoV-2 by various $\Delta\Delta G$ prediction methods.

Method	TH31W	AH53F	NH57L	RH103M	LH104F	Average
MIF-Network	24.49%	**4.05%**	**6.48%**	80.36%	36.23%	30.32%
RDE-Network	**1.62%**	**2.02%**	20.65%	61.54%	**5.47%**	18.26%
DiffAffinity	**7.28%**	**3.64%**	18.82%	81.78%	10.93%	24.49%
Prompt-DDG	**2.02%**	**6.88%**	**3.24%**	34.81%	**6.48%**	10.69%
MSM-Mut	**6.48%**	10.12%	16.19%	19.23%	20.04%	14.41%
BA-DDG	**5.26%**	15.58%	**2.22%**	40.28%	**7.69%**	14.21%
H3-DDG	**3.44%**	**7.48%**	**2.02%**	32.79%	**2.63%**	**9.67%**

5 Conclusion

In this work, we introduce H3-DDG, a novel framework that leverages hierarchical communication mechanisms and many-body attention to tackle the challenges of protein-protein binding prediction. By explicitly modeling higher-order interactions and designating mutation sites as cluster centers,

H3-DDG captures complex dependencies and synergistic effects, particularly excelling in multi-point mutation scenarios. Evaluations on SKEMPI v2 and BindingGYM show that H3-DDG outperforms state-of-the-art methods across most metrics, demonstrating robust performance and scalability. While H3-DDG shows promise for protein design and affinity prediction, its performance on large datasets and diverse mutations needs further study. Integrating it with experimental workflows will be key to validating real-world applicability. Future work will tackle these challenges to expand its impact.

Acknowledgments and Disclosure of Funding

This study has been supported by the Guangdong S&T Program [2024B0101040005], the Guangdong S&T Program [2023B1111030002], the National Natural Science Foundation of China [62041209], the Natural Science Foundation of Shanghai [24ZR1440600], the China Postdoctoral Science Foundation [2025M771540, GZB20250391], the Guangdong Basic and Applied Basic Research Foundation [2025A1515060011], and the Lingang Laboratory [LGL-8888].

References

[1] Rebecca F Alford, Andrew Leaver-Fay, Jeliazko R Jeliazkov, Matthew J O'Meara, Frank P DiMaio, Hahnbeom Park, Maxim V Shapovalov, P Douglas Renfrew, Vikram K Mulligan, Kalli Kappel, et al. The rosetta all-atom energy function for macromolecular modeling and design. *Journal of chemical theory and computation*, 13(6):3031–3048, 2017.

[2] Amro Abd-Al-Fattah Amara. Pharmaceutical and industrial protein engineering: where we are? *Pakistan Journal of Pharmaceutical Sciences*, 26(1), 2013.

[3] Filippo Maria Bianchi, Daniele Grattarola, and Cesare Alippi. Spectral clustering with graph neural networks for graph pooling. In *International conference on machine learning*, pages 874–883. PMLR, 2020.

[4] Martin Dietrich Buhmann. Radial basis functions. *Acta numerica*, 9:1–38, 2000.

[5] Huali Cao, Jingxue Wang, Liping He, Yifei Qi, and John Z Zhang. Deepddg: predicting the stability change of protein point mutations using neural networks. *Journal of chemical information and modeling*, 59(4):1508–1514, 2019.

[6] 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.

[7] Carla CCR De Carvalho. Enzymatic and whole cell catalysis: finding new strategies for old processes. *Biotechnology advances*, 29(1):75–83, 2011.

[8] 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.

[9] Jesse Durham, Jing Zhang, Ian R Humphreys, Jimin Pei, and Qian Cong. Recent advances in predicting and modeling protein–protein interactions. *Trends in biochemical sciences*, 48(6): 527–538, 2023.

[10] Oliver Dutton, Sandro Bottaro, Michele Invernizzi, Istvan Redl, Albert Chung, Falk Hoffmann, Louie Henderson, Stefano Ruschetta, Fabio Airoldi, Benjamin MJ Owens, et al. Improving inverse folding models at protein stability prediction without additional training or data. *bioRxiv*, pages 2024–06, 2024.

[11] Bowen Gao, Yinjun Jia, Yuanle Mo, Yuyan Ni, Weiying Ma, Zhiming Ma, and Yanyan Lan. Profsa: Self-supervised pocket pretraining via protein fragment-surroundings alignment. *arXiv preprint arXiv:2310.07229*, 2023.

[12] Ziqi Gao, Chenran Jiang, Jiawen Zhang, Xiaosen Jiang, Lanqing Li, Peilin Zhao, Huanming Yang, Yong Huang, and Jia Li. Hierarchical graph learning for protein–protein interaction. *Nature Communications*, 14(1):1093, 2023.

[13] Jiaqi Guan, Jiahan Li, Xiangxin Zhou, Xingang Peng, Sheng Wang, Yunan Luo, Jian Peng, and Jianzhu Ma. Group ligands docking to protein pockets. *arXiv preprint arXiv:2501.15055*, 2025.

[14] Ruihan Guo, Rui Wang, Ruidong Wu, Zhizhou Ren, Jiahan Li, Shitong Luo, Zuofan Wu, Qiang Liu, Jian Peng, and Jianzhu Ma. Enhancing protein mutation effect prediction through a retrieval-augmented framework. *Advances in Neural Information Processing Systems*, 37: 49130–49153, 2024.

[15] 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.

[16] Md Shamim Hussain, Mohammed J Zaki, and Dharmashankar Subramanian. Triplet interaction improves graph transformers: Accurate molecular graph learning with triplet graph transformers. *arXiv preprint arXiv:2402.04538*, 2024.

[17] 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.

[18] Xiaoran Jiao, Weian Mao, Wengong Jin, Peiyuan Yang, Hao Chen, and Chunhua Shen. Boltzmann-aligned inverse folding model as a predictor of mutational effects on protein-protein interactions. *arXiv preprint arXiv:2410.09543*, 2024.

[19] Susan Jones and Janet M Thornton. Principles of protein-protein interactions. *Proceedings of the National Academy of Sciences*, 93(1):13–20, 1996.

[20] Edward King, Erick Aitchison, Han Li, and Ray Luo. Recent developments in free energy calculations for drug discovery. *Frontiers in molecular biosciences*, 8:712085, 2021.

[21] Xue Li, Peifu Han, Gan Wang, Wenqi Chen, Shuang Wang, and Tao Song. Sdnn-ppi: self-attention with deep neural network effect on protein-protein interaction prediction. *BMC genomics*, 23(1):474, 2022.

[22] Yang Li, Min Li, Caijie Qu, Yongxi Li, Zhanli Tang, Zhike Zhou, Zengzhao Yu, Xu Wang, Linlin Xin, and Tongxin Shi. The polygenic map of keloid fibroblasts reveals fibrosis-associated gene alterations in inflammation and immune responses. *Frontiers in Immunology*, 12:810290, 2022.

[23] Shiwei Liu, Tian Zhu, Milong Ren, Chungong Yu, Dongbo Bu, and Haicang Zhang. Predicting mutational effects on protein-protein binding via a side-chain diffusion probabilistic model. *Advances in Neural Information Processing Systems*, 36:48994–49005, 2023.

[24] Wei Lu, Jixian Zhang, Ming Gu, and Shuangjia Zheng. Bindinggym: A large-scale mutational dataset toward deciphering protein-protein interactions. *bioRxiv*, pages 2024–12, 2024.

[25] Wei Lu, Jixian Zhang, Jiahua Rao, Zhongyue Zhang, and Shuangjia Zheng. Alphafold3, a secret sauce for predicting mutational effects on protein-protein interactions. *bioRxiv*, pages 2024–05, 2024.

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

[27] 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.

[28] Yuanle Mo, Xin Hong, Bowen Gao, Yinjun Jia, and Yanyan Lan. Multi-level interaction modeling for protein mutational effect prediction. *arXiv preprint arXiv:2405.17802*, 2024.

[29] 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.

[30] Jiahua Rao, Dahao Xu, Wentao Wei, Yicong Chen, Mingjun Yang, and Yuedong Yang. Quadruple attention in many-body systems for accurate molecular property predictions. In *Forty-second International Conference on Machine Learning*.

[31] Jiahua Rao, Jiancong Xie, Qianmu Yuan, Deqin Liu, Zhen Wang, Yutong Lu, Shuangjia Zheng, and Yuedong Yang. A variational expectation-maximization framework for balanced multi-scale learning of protein and drug interactions. *Nature Communications*, 15(1):4476, 2024.

[32] Jiahua Rao, Deqin Liu, Xiaolong Zhou, Qianmu Yuan, Wentao Wei, Wei Lu, Jixian Zhang, Yu Rong, Yuedong Yang, and Shuangjia Zheng. Accurate protein-protein interactions modeling through physics-informed geometric invariant learning. *bioRxiv*, pages 2025–07, 2025.

[33] Jeffrey A Ruffolo, Lee-Shin Chu, Sai Pooja Mahajan, and Jeffrey J Gray. Fast, accurate antibody structure prediction from deep learning on massive set of natural antibodies. *Nature communications*, 14(1):2389, 2023.

[34] Sisi Shan, Shitong Luo, Ziqing Yang, Junxian Hong, Yufeng Su, Fan Ding, Lili Fu, Chenyu Li, Peng Chen, Jianzhu Ma, et al. Deep learning guided optimization of human antibody against sars-cov-2 variants with broad neutralization. *Proceedings of the National Academy of Sciences*, 119(11):e2122954119, 2022.

[35] Changfa Shu, Jianfeng Li, Jin Rui, Dacheng Fan, Qiankun Niu, Ruiyang Bai, Danielle Cicka, Sean Doyle, Alafate Wahafu, Xi Zheng, et al. Uncovering the rewired iap-jak regulatory axis as an immune-dependent vulnerability of lkb1-mutant lung cancer. *Nature Communications*, 16 (1):2324, 2025.

[36] Fang Wu and Stan Z Li. Surface-vqmae: Vector-quantized masked auto-encoders on molecular surfaces. In *Forty-first International Conference on Machine Learning*, 2024.

[37] Lirong Wu, Yijun Tian, Haitao Lin, Yufei Huang, Siyuan Li, Nitesh V Chawla, and Stan Z Li. Learning to predict mutation effects of protein-protein interactions by microenvironment-aware hierarchical prompt learning. *arXiv preprint arXiv:2405.10348*, 2024.

[38] Lirong Wu, Yunfan Liu, Haitao Lin, Yufei Huang, Guojiang Zhao, Zhifeng Gao, and Stan Z Li. A simple yet effective ddg predictor is an unsupervised antibody optimizer and explainer. *arXiv preprint arXiv:2502.06913*, 2025.

[39] Fang Yang, Kunjie Fan, Dandan Song, and Huakang Lin. Graph-based prediction of protein-protein interactions with attributed signed graph embedding. *BMC bioinformatics*, 21:1–16, 2020.

[40] Deqi Yin, Ning Jiang, Chang Cheng, Xiaoyu Sang, Ying Feng, Ran Chen, and Qijun Chen. Protein lactylation and metabolic regulation of the zoonotic parasite toxoplasma gondii. *Genomics, proteomics & bioinformatics*, 21(6):1163–1181, 2023.

[41] Zhitao Ying, Jiaxuan You, Christopher Morris, Xiang Ren, Will Hamilton, and Jure Leskovec. Hierarchical graph representation learning with differentiable pooling. *Advances in neural information processing systems*, 31, 2018.

A Additional Details

A.1 Graph Pooling Method of H3-DDG

To enable scalable learning on biomolecular graphs, H3-DDG employs a deterministic graph pooling strategy that compresses node representations while preserving spatial and functional context. Given residue-wise features $\mathbf{h}$ and their Cα coordinates $\mathbf{x} \in \mathbb{R}^{N\times 3}$, we form a reduced set of hyper-nodes via Farthest Point Sampling (FPS) with mutation-aware initialization. Specifically, mutation sites are first selected as initial cluster centroids c_k, and additional centroids are iteratively sampled to maximize coverage under Euclidean distance.

Each residue is then assigned to the nearest cluster centroid based on the Euclidean distance between its Cα coordinates and the centroids. Feature aggregation is performed within each cluster by computing the mean of the features of all residues assigned to that cluster:

$$\mathbf{h}_k = \frac{1}{|S_k|} \sum_{i \in S_k} \mathbf{h}_i, \tag{18}$$

where S_k is the set of residues assigned to cluster k. This produces a set of K cluster-level representations $\{\mathbf{h}_k\}_{k=1}^{K}$, with K dynamically determined as $K = \lfloor L/R \rfloor$, where L is the number of valid residues and R is a predefined reduction ratio.

Anchored to mutation sites and guided by spatial diversity, our pooling method preserves locality and biological relevance. It is non-parametric and efficient, suitable for large or variable-length proteins.

A.2 Details of the Inter-Assay Split in BindingGYM

To evaluate generalization to unseen protein-protein interactions, we adopt the inter-assay split strategy from the BindingGYM dataset, following the approach of [24]. In this setting, assays are first clustered into five groups based on the sequences of their mutated proteins. Data from one cluster is held out for testing, while the remaining four are used for training. This split evaluates the models' ability to generalize to new assays, which holds significant practical significance.

We evaluate performance under three levels of mutational depth: **ALL** (all mutants), **<3** (mutants with fewer than 3 mutations), and **≥3** (mutants with 3 or more mutations).

A.3 Details of metric definitions and calculations

For the **SKEMPI v2** dataset, we use seven quantitative metrics, including five standard global criteria: Pearson and Spearman correlation coefficients, root mean square error (RMSE), mean absolute error (MAE), and area under the ROC curve (AUROC). To calculate AUROC, mutations are classified based on the sign of AUROC, mutations are classified based on the sign of $\Delta\Delta G$. In practice, correlations within individual protein complexes are often more relevant. To this end, following RDE-Network [26], we group mutations by protein structure, exclude groups with fewer than 10 mutations, and compute correlations separately for each structure. For the **BindingGYM** dataset, we use per-DMS (deep mutational scanning) metrics, including Pearson, Spearman, AUROC, and RMSE.

A.4 Training Details

A.4.1 Hyper-parameters

We used the Adam optimizer with a learning rate of 4e-4 and a batch size of 1, 2, depending on GPU memory and graph size. The model was trained for 20,000 iterations with 4 attention heads and a hidden dimension of 128. The number of hyperedges was selected from $L/10$, $L/6$, $L/4$, and the number of edges in the 4-body attention module from $1N$, $2N$, $3N$, where L and N denote the numbers of residues and nodes, respectively. The pre-trained ProteinMPNN module used its default 3-layer configuration.

A.4.2 Hardware

Experiments were run on dual Xeon Gold 6248R CPUs and an RTX 4090 GPU under Ubuntu 22.04.

B Additional Results

B.1 H3-DDG Performance on Predicted Complex Structures

To evaluate robustness, we used AlphaFold 3 (AF3) to predict the complex structures for SKEMPI v2 and applied H3-DDG for prediction. While AF3 achieves high-quality structure prediction, its accuracy may still be lower than experimental structures. Importantly, H3-DDG showed only a slight decrease in performance on AF3-predicted structures, with a drop of 5.4% in Pearson correlation, demonstrating strong robustness and practical applicability.

B.2 Extended Baseline Comparisons on SKEMPI v2

Table 6 reports extended baseline results on SKEMPI v2, including additional methods to complement the main text comparisons. These include energy function-based, sequence-based, and unsupervised approaches, providing a broader context for evaluating performance across diverse methodological categories.

B.3 Ablation Results under Different Mutation Depths

Table 7 presents the full ablation results of different pooling strategies and many-body attention mechanisms across mutation types, including single-point, multi-point, and all-point mutations on SKEMPI v2. The results demonstrate that our proposed H3-DDG consistently outperforms baseline pooling methods (e.g., DiffPool [41] and MinCutPool [3]) under all mutation settings. Furthermore, incorporating 4-body attention around mutation sites yields the best performance across all metrics, validating the effectiveness of higher-order attention in capturing complex interaction patterns.

B.4 Ablation study on Edge Count in 4-body Attention

We investigate the impact of varying the number of edges in the 4-body attention mechanism around mutation sites on $\Delta\Delta G$ prediction. The number of edges in the 4-body attention is directly proportional to the computational complexity. As shown in Table 8, where $\mathcal{N}$ denotes the number of nodes involved in the 4-body attention computation, selecting $2 \cdot \mathcal{N}$ edges yields a computational complexity of $2 \cdot |\mathcal{N}|^3$. The results indicate that increasing the number of edges leads to only a constant-factor increase in computational cost, yet yields sustained improvements in predictive performance. For example, the Pearson correlation, which was already high at 0.7352, further increases to 0.7501. These findings highlight that adding more edges in the many-body attention mechanism near mutation sites significantly enhances the model's capacity to capture complex many-body interactions, while maintaining manageable computational overhead, thereby driving further gains in prediction accuracy.

B.5 Efficiency Analysis

Our model balances efficiency and performance. As shown in Table 9, increasing hyperedges from $L/10$ to $L/4$ steadily improves Pearson correlation, reaching a peak of 0.7501 with $L/4$ hyperedges and $3N$ 4-body edges. This setting maintains acceptable efficiency (4.34 it/s), only slightly slower than BA-DDG (6.25 it/s, Pearson: 0.7118). Similarly, increasing 4-body edges from $1.5N$ to $3N$ yields incremental gains, indicating denser subgraph modeling enhances accuracy at manageable cost. Additionally, as shown in Table 10, larger cutoff radii improve performance by capturing richer structural context, but gains diminish beyond 8 Å while cost rises. Thus, 8 Å is chosen as the optimal cutoff, offering the best accuracy-efficiency trade-off.

Table 5: H3-DDG Performance on Experimental and AF3-Predicted Structures.

Structure	**Pearson↑**	**RMSE↓**	**AUROC↑**
AlphaFold3	0.7117	1.4518	0.7755
Experimental	0.7501	1.3665	0.7920

Table 6: Mean results of 3-fold cross-validation on SKEMPI v2 under single-, multi-, and all-point mutations. **Bold** and underline indicate the best and second-best results.

Method	Mutations	Overall					Per-Structure	
		Pearson↑	Spear.↑	RMSE↓	MAE↓	AUROC↑	Pearson↑	Spear.↑
Rosetta	all	0.3113	0.3468	1.6173	1.1311	0.6562	0.3284	0.2988
	single	0.3250	0.3670	1.1830	0.9870	0.6740	0.3510	0.4180
	multiple	0.1990	0.2300	2.6580	2.0240	0.6210	0.1910	0.0830
FoldX	all	0.3120	0.4071	1.9080	1.3089	0.6582	0.3789	0.3693
	single	0.3150	0.3610	1.6510	1.1460	0.6570	0.3820	0.3600
	multiple	0.2560	0.4180	2.6080	1.9260	0.7040	0.3330	0.3400
DDGPred	all	0.6580	0.4687	1.4998	1.0821	0.6992	0.3750	0.3407
	single	0.6515	0.4390	1.3285	0.9618	0.6858	0.3711	0.3427
	multiple	0.5938	0.5150	2.1813	1.6699	0.7590	0.3912	0.3896
End-to-End	all	0.6373	0.4882	1.6198	1.1761	0.7172	0.3873	0.3587
	single	0.6605	0.4594	1.3148	0.9569	0.7019	0.3818	0.3426
	multiple	0.5858	0.4942	2.1971	1.7087	0.7532	0.4178	0.4034
RDE-Network	all	0.6447	0.5584	1.5799	1.1123	0.7454	0.4448	0.4010
	single	0.6421	0.5271	1.3333	0.9392	0.7367	0.4687	0.4333
	multiple	0.6288	0.5900	2.0980	1.5747	0.7749	0.4233	0.3926
DiffAffinity	all	0.6609	0.5560	1.5350	1.0930	0.7440	0.4220	0.3970
	single	0.6720	0.5230	1.2880	0.9230	0.7330	0.4290	0.4090
	multiple	0.6500	0.6020	2.0510	1.5400	0.7840	0.4140	0.3870
Prompt-DDG	all	0.6772	0.5910	1.5207	1.0770	0.7568	0.4712	0.4257
	single	0.6596	0.5450	1.3072	0.9191	0.7355	0.4736	0.4392
	multiple	0.6780	0.6433	1.9831	1.4837	0.8187	0.4448	0.3961
ProMIM	all	0.6720	0.5730	1.5160	1.0890	0.7600	0.4640	0.4310
	single	0.6680	0.5340	1.2790	0.9240	0.7380	0.4660	0.4390
	multiple	0.6660	0.6140	1.9630	1.4910	0.8250	0.4580	0.4250
BA-DDG	all	0.7118	0.6346	1.4516	1.0151	0.7726	0.5453	0.5134
	single	0.7321	0.6157	1.1848	0.8409	0.7662	0.5606	0.5192
	multiple	0.6650	0.6293	2.0151	1.4944	0.7875	0.4924	0.4959
H3-DDG	all	**0.7501**	**0.6604**	**1.3665**	**0.9612**	**0.7920**	**0.5686**	**0.5281**
	single	**0.7471**	**0.6374**	**1.1560**	**0.8080**	**0.7803**	**0.5750**	**0.5295**
	multiple	**0.7341**	**0.6913**	**1.8320**	**1.3880**	**0.8309**	**0.5520**	**0.5323**
$\Delta_{\text{BA-DDG}}$	all	+5.38%	+4.07%	+5.86%	+5.31%	+2.51%	+4.27%	+2.86%
	single	+2.05%	+3.52%	+2.43%	+3.91%	+1.84%	+2.57%	+1.98%
	multiple	+10.39%	+9.85%	+9.09%	+7.12%	+5.51%	+12.10%	+7.34%

B.6 Scatter Plot of Prediction Results

In Figure 3, we present scatter plots comparing experimental and predicted $\Delta\Delta G$ values for the three methods under multi-point mutation scenarios.

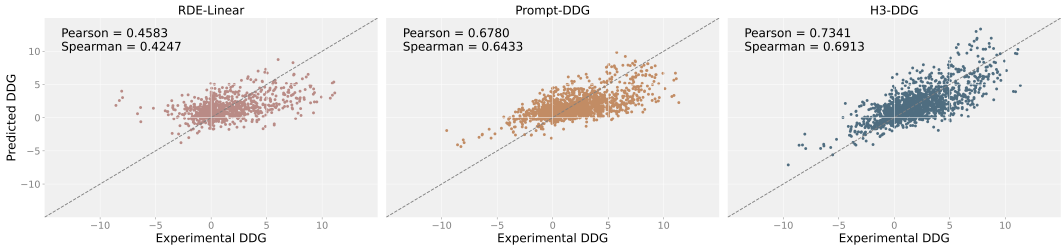

Figure 3: Comparison of predicted and experimental $\Delta\Delta$G across methods.

Table 7: Ablation study of pooling types and many-body attention mechanisms across mutation types on SKEMPI v2.

Pooling Type	Attn. Between Hyperedges	Attn. Around Mutation Sites	Mutations	Overall			
				Pearson↑	Spear.↑	RMSE↓	MAE↓
DiffPool	3-body Attn.	–	all	0.7261	0.6407	1.4209	1.0079
			single	0.7344	0.6256	1.1804	0.8336
			multiple	0.6959	0.6558	1.9375	1.4843
MinCutPool	3-body Attn.	–	all	0.7275	0.6456	1.4178	1.0027
			single	0.7324	0.6333	1.1841	0.8342
			multiple	0.7004	0.6589	1.9256	1.4693
H3-DDG	3-body Attn.	–	all	0.7317	0.6485	1.4085	0.9923
			single	0.7383	0.6273	1.1728	0.8242
			multiple	0.7040	0.6676	1.9161	1.4528
H3-DDG	3-body Attn.	3-body Attn.	all	0.7352	0.6509	1.4007	0.9805
			single	0.7355	0.6326	1.1782	0.8243
			multiple	0.7144	0.6732	1.8880	1.4127
H3-DDG	3-body Attn.	4-body Attn.	all	**0.7501**	**0.6604**	**1.3665**	**0.9612**
			single	**0.7471**	**0.6374**	**1.1560**	**0.8080**
			multiple	**0.7341**	**0.6913**	**1.8320**	**1.3880**

Table 8: Ablation on the number of edges near mutations within 4-body attention on SKEMPI v2.

Complexity	Mutations	Overall					Per-Structure	
		Pearson↑	Spear.↑	RMSE↓	MAE↓	AUROC↑	Pearson↑	Spear.↑
$\mathcal{O}(\lvert\mathcal{N}\rvert^3)$	all	0.7352	0.6509	1.4007	0.9805	0.7895	0.5663	0.5227
	single	0.7355	0.6326	1.1782	0.8243	**0.7851**	0.5758	0.5246
	multiple	0.7144	0.6732	1.8880	1.4127	0.8098	0.5372	0.5156
$\mathcal{O}(2 \cdot \lvert\mathcal{N}\rvert^3)$	all	0.7461	0.6570	1.3760	0.9719	**0.7941**	0.5660	0.5231
	single	0.7455	0.6365	1.1591	0.8128	0.7850	**0.5790**	**0.5341**
	multiple	0.7279	0.6815	1.8500	1.4079	0.8262	0.5449	0.5310
$\mathcal{O}(3 \cdot \lvert\mathcal{N}\rvert^3)$	all	**0.7501**	**0.6604**	**1.3665**	**0.9612**	0.7920	**0.5686**	**0.5281**
	single	**0.7471**	**0.6374**	**1.1560**	**0.8080**	0.7803	0.5750	0.5295
	multiple	**0.7341**	**0.6913**	**1.8320**	**1.3880**	**0.8309**	**0.5520**	**0.5323**

Table 9: Impact of hypergraph and subgraph configurations on efficiency and performance. L is the residue count, and $L/\,k$ denotes the number of hyperedges (floor division). $3N$ means the 4-body attention has three times as many edges as nodes.

Method	Number of Hyperedges	Number of Edges in 4-body Attn.	Pearson↑	Training Speed (iterations/sec)	Training Time (mins/epoch)
BA-DDG	-	-	0.7118	6.25	5.94
H3-DDG	$L/10$	$3N$	0.7418	5.11	7.27
	$L/6$		0.7482	4.68	7.93
	$L/4$		0.7501	4.34	8.55
	$L/4$	$1.5N$	0.7420	4.40	8.44
		$2N$	0.7461	4.38	8.48
		$3N$	0.7501	4.34	8.55

Table 10: Impact of Cutoff Radius on Model Performance and Efficiency.

Cutoff of Mut. Subgraph(Å)	Pearson↑	MAE↓	Training Speed (iterations/sec)
5	0.7378	0.9834	4.52
8	0.7501	0.9612	4.32
12	0.7511	0.9602	4.03

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We confirm that the main claim in the abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have discussed the limitations of the work in the conclusion section. Specifically, we note that H3-DDG requires further evaluation on large-scale datasets and highly diverse mutation scenarios. Additionally, integration with experimental workflows is needed to validate its real-world applicability. These challenges are acknowledged as directions for future work.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: Our paper does not include complex theoretical results, and therefore, there are no assumptions or proofs to provide.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The training details and datasets are provided in the paper and supplementary material, and we also plan to open-source our model.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We have provided the source code in the supplementary material and will open-source our data and code.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We have specified all the training and test details in Section 4.1 and Appendix A.4

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We have shown the complete distribution of our results in Fig. 2 and Fig. 3, highlighting our robustness compared to other approaches.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We have provided sufficient information on the computer resources in Appendix A.4.2 and B.5.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our research conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We have discussed it in the introduction section. These results underscore the capability of H3-DDG to handle intricate mutational landscapes and its potential for broad applicability in protein engineering and drug design.

    Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This paper does not have such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All the assets are properly cited in the paper.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We introduce a new model to capture higher-order many-body interactions across multiple scales for $\Delta\Delta G$ predictions, and we have provided the details of the model in this paper.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: This paper dose not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.