Beyond Random Splits: A Critical Evaluation of Graph Learning Models in Predicting Mutation-Induced Drug Resistance

Zongrui Cheng , Haoxin Wu , Dengming Ming *

Posted Date: 2 April 2026

doi: 10.20944/preprints202604.0069.v1

Keywords: drug resistance prediction; graph neural networks (GNN); inductive generalization; data leakage; ΔΔG prediction

 doi:10.20944/preprints202604.0069.v1

Article

Beyond Random Splits: A Critical Evaluation of Graph Learning Models in Predicting Mutation-Induced Drug Resistance

Zongrui Cheng, Haoxin Wu and Dengming Ming *

College of Biotechnology and Pharmaceutical Engineering, Nanjing Tech University, Nanjing 211816, China

* Correspondence: dming@njtech.edu.cn

Abstract

Background: Deep learning has become an important tool for predicting mutation-induced changes in binding free energy (ΔΔG). However, most current state-of-the-art methods rely heavily on paired wild-type (WT) and mutant (MT) complex structures during both training and inference. This dependence on post-mutation structural information substantially limits their practical utility in real-world scenarios, such as clinical diagnosis and early-stage drug screening, where mutant structures are difficult to obtain experimentally in a timely manner. Methods: To evaluate model performance in more realistic and challenging translational settings, we conducted a systematic benchmark of graph-based deep learning models under a WT-only inductive setting. We constructed a full-protein heterogeneous graph framework that incorporates long-range spatial constraints to implicitly infer mutational effects from static wild-type structures. We compared it against a sequence-based vector baseline model. Results: Through a systematic evaluation on the MdrDB dataset, we revealed a critical generalization gap. Although random splitting yielded relatively high predictive correlation due to homologous data leakage (Pearson $R \approx 0.55$), model performance dropped sharply under a strict UniProt-based cross-protein split designed to simulate prediction on truly unseen targets (Pearson $R \approx 0.15$). Although the absolute performance remained limited, the graph-based model showed a weak but consistent improvement over the sequence baseline, which was close to random guessing (Pearson $R \approx 0.04$). Conclusions: Further analyses suggest that the performance bottleneck may partly arise from intrinsic experimental noise in the dataset (i.e., label inconsistency) and from the absence of conformational entropy (dynamic) information in static WT structures. This study indicates that random splitting can lead to substantial overestimation of model generalizability. It highlights the need to integrate physical priors and dynamic features to overcome the current limitations of drug resistance prediction when explicit mutant structures are unavailable.

Keywords: drug resistance prediction; graph neural networks (GNN); inductive generalization; data leakage; ΔΔG prediction

1. Introduction

The emergence of acquired drug resistance represents a formidable bottleneck in precision oncology and infectious disease management, frequently eroding the clinical efficacy of otherwise potent targeted therapies [1]. At the molecular level, such resistance is often driven by nonsynonymous single-nucleotide mutations in target proteins, which can drastically alter the binding affinity of therapeutic ligands. Quantifying this effect through mutation-induced changes in binding free energy (ΔΔG) provides a key thermodynamic measure for evaluating resistance potential. Physics-based tools such as FoldX [2], as well as more rigorous alchemical free-energy methods, including free energy perturbation (FEP) and thermodynamic integration (TI), can provide highly accurate estimates of binding free-energy changes. However, their substantial computational cost makes them impractical for large-scale mutational scanning [3,4]. Consequently, data-driven

 doi:10.20944/preprints202604.0069.v1

deep learning methods have rapidly emerged as a promising alternative [5,6]. By leveraging large-scale structural and sequence data, these AI-based models aim to efficiently capture complex structure–activity relationships, thereby providing a high-throughput paradigm for predicting drug resistance profiles [7–10].

While current deep learning architectures have significantly advanced ΔΔG prediction, state-of-the-art models—including **CATH-ddG** [11], **DDMut** [12], and **mCSM-lig** [13].—remain inherently dependent on the availability of both wild-type and mutant structures. This reliance on post-mutational conformations creates a significant barrier to their application in rapid clinical screening, where mutant structures are seldom known *a priori.* This assumption often leads to improved predictive performance, suggesting that mutant conformational information contributes significantly to ΔΔG prediction. However, this dependence severely limits the practical applicability of these methods in real-world settings. In clinical diagnosis and early-stage drug screening, only the wild-type complex structure is typically available. Determining a high-resolution three-dimensional structure for every newly emerging clinical mutation through X-ray crystallography or cryo-electron microscopy is not only costly and time-consuming, but often experimentally infeasible. This "scarcity of structural data" renders MT-dependent models difficult to deploy for unseen mutations or large-scale mutational scanning tasks, thereby creating an urgent need for a computational paradigm capable of making effective predictions using only the structural context of the wild-type complex.

Although tools such as AlphaFold can be used to generate mutant structures [14], performing structural prediction for a large number of mutations remains computationally expensive for large-scale mutational scanning. Moreover, structures of point mutants generated by AlphaFold and related methods often tend to regress toward the wild-type conformation, making them insufficiently sensitive to subtle side-chain perturbations and mutation-induced structural changes [15,16]. Therefore, exploring a predictive paradigm that enables millisecond-level inference with zero additional folding cost may have substantial practical value for clinical screening applications. Despite the rapid development of mutation-effect prediction models, a systematic evaluation of their generalization ability in a realistic WT-only prediction setting remains lacking.

To address these practical limitations and assess the true potential of deep learning in translational settings, this study focuses on the wild-type-only prediction task. Unlike conventional approaches that treat ΔΔG prediction as a comparison between two known static structures, our model must predict the energetic impact of a mutation using only the wild-type complex structure and mutation descriptors, without access to the post-mutation geometry. This setting forces the model to implicitly infer complex physicochemical consequences—such as side-chain rearrangement, backbone relaxation, and steric clashes—solely from the local chemical environment and topological constraints present in the wild-type state. Graph neural networks (GNNs) have recently shown strong capability in modeling molecular graphs and protein–ligand interactions [17–20]. Building on this success, we therefore established a rigorous benchmark to test whether current graph learning architectures can learn generalizable physical rules underlying drug resistance, rather than merely memorizing geometric deviations between paired structures.

To investigate this question in depth, we conducted a systematic benchmark on the MdrDB dataset across a range of architectures, from sequence-based vector baselines to full-protein heterogeneous graph models. Our analysis indicates that the strong performance widely reported in the existing literature is often confounded by data leakage inherent to Random split strategies. For example, in DDMut [12], the blind test set is non-redundant at the mutation level, yet the same protein may still appear in both the training and test sets through mutations at different residue positions. Similarly, although DeepDDG [21] applies sequence-similarity-based filtering, it does not explicitly address redundancy at the protein or structural level. Such data partitioning schemes may introduce latent information leakage, thereby leading to an overestimation of model generalizability. Recently, Xie et al. [22] also highlighted the potential impact of dataset partitioning strategies on ΔΔG prediction benchmarks using the MdrDB dataset, demonstrating that sequence-based models relying on protein language model embeddings experience substantial performance degradation under a

 doi:10.20944/preprints202604.0069.v1

strict UniProt-level split. This concern was confirmed in our controlled experiments: the high correlation observed under Random splitting was primarily attributable to homologous sequence overlap between the training and test sets, which in turn led to a systematic overestimation of cross-protein generalization performance. By contrast, when a strict UniProt-based split was imposed to simulate realistic predictions on genuinely unseen targets, the performance of all models declined markedly, highlighting the true difficulty of cross-protein generalization in the absence of explicit structural guidance for mutants. Notably, despite the substantial challenge posed by this stringent setting, models incorporating full-protein graph representations consistently outperformed sequence-only baselines, suggesting that explicit modeling of three-dimensional structural context provides additional predictive signals beyond sequence-derived features. These findings suggest that incorporating explicit structural context through graph-based representations may provide additional predictive signals beyond sequence-derived features, even when mutant conformations are unavailable.

2. Methods

2.1. Dataset

The data on mutation-induced changes in drug binding free energy (ΔΔG) used in this study were obtained from the MdrDB (Mutation-induced Drug Resistance DataBase) [23]. MdrDB integrates multiple public data resources, including the Platinum, AIMMS, TKI, RET, KinaseMD, GDSC, and DepMap datasets. Through a unified preprocessing pipeline, it organizes protein names, UniProt IDs, mutation annotations, drug information, and the corresponding binding free-energy changes (ΔΔG) into a standardized sample format. In addition to the core protein–drug–mutation–ΔΔG information, MdrDB also provides the associated three-dimensional structural data (wild-type and mutant protein–ligand complexes) and structure-related features, thereby offering a structural foundation for the construction and evaluation of machine learning models.

In this study, MdrDB_CoreSet was used as the primary data source. This dataset contains 4,292 protein–ligand mutation records spanning multiple UniProt IDs. We formulated the prediction of mutation-induced changes in binding free energy as a regression task to predict the ΔΔG value for each mutation relative to the wild type (in kcal/mol).

Given that this study focuses on single-point mutations, we further filtered and curated the original dataset as follows:

(1) Only samples involving single amino acid substitution mutations were retained;
(2) Abnormal complex structures lacking valid protein contacts around the ligand were removed;
(3) PDB records with ambiguous structural mapping caused by multichain assemblies or duplicated residue numbering were excluded;
(4) Samples with incomplete or unresolvable structural information were discarded.

After the above filtering procedures, a total of 3,125 valid protein–ligand mutation samples were retained for subsequent model training and evaluation, covering proteins from 166 distinct UniProt IDs.

For the experimental design, we constructed both a Random split and a strict UniProt-based split. The UniProt-based split ensures that no proteins are shared between the training and test sets, thereby simulating a realistic cross-protein generalization scenario and enabling evaluation of the model's ability to generalize to new target prediction tasks.

2.2. Protein Representation

To characterize the contextual semantic information of protein sequences and the local perturbation features induced by mutations, we constructed multiple residue-level representation schemes, including amino acid identity encoding, embedding-based representations derived from a pretrained protein language model [24,25], and mutation-perturbation representations constructed

 doi:10.20944/preprints202604.0069.v1

from embedding differences. Let the wild-type protein sequence be denoted as, and the mutant sequence as S^{MT}.

In this study, we used the pretrained protein language model ESM-C (600M parameters) [26] as the feature extractor, denoted by $f_{\mathrm{ESM}}(\cdot)$. This model is self-supervised and pretrained on large-scale protein sequence data, providing residue-level contextual embeddings. In our framework, the parameters of ESM-C were kept frozen and used only for embedding extraction. The sequence encodings are defined as:

$$\mathrm{E}^{WT} = f_{\mathrm{ESM}}(S^{WT}), \mathrm{E}^{MT} = f_{\mathrm{ESM}}(S^{MT})$$

where $\mathrm{E}^{\mathrm{WT}} = \{\mathrm{e}_{\mathrm{i}}^{\mathrm{WT}}\}_{\mathrm{i}=1}^{\mathrm{n}}$ and $\mathrm{E}^{\mathrm{MT}} = \{\mathrm{e}_{\mathrm{i}}^{\mathrm{MT}}\}_{\mathrm{i}=1}^{\mathrm{n}}$ $\mathrm{n}$is the sequence length, and $\mathrm{e}_{\mathrm{i}} \in \mathbb{R}^{\mathrm{d}}$denotes the embedding vector of the i-th residue.

(1) One-hot representation

A 20-dimensional one-hot vector was used to encode the amino acid identity of each residue:

$$\mathrm{x}_{\mathrm{i}} \in \mathbb{R}^{20}$$

where $\mathrm{i}$denotes the residue index. This representation contains only amino acid category information and serves as a basic sequence feature.

(2) ESM embedding representation

ESM-C was used to extract residue-level embeddings for the wild-type and mutant sequences, denoted as E^{WT} and E^{MT}, respectively, to incorporate evolutionary and contextual semantic information.

(3) ΔESM differential representation

Because $\Delta\Delta\mathrm{G}$ reflects the energetic difference relative to the wild-type state, we further constructed a residue-level differential representation based on ESM-C embeddings:

$$\mathrm{E}^{\Delta} = \mathrm{E}^{\mathrm{WT}} - \mathrm{E}^{\mathrm{MT}}$$

where $\mathrm{E}^{\Delta} = \{\mathrm{e}_{\mathrm{i}}^{\Delta}\}_{\mathrm{i}=1}^{\mathrm{n}}$and $\mathrm{e}_{\mathrm{i}}^{\Delta} \in \mathbb{R}^{\mathrm{d}}$. This representation is used to explicitly characterize mutation-induced perturbations at the representation level, thereby aligning the feature formulation with the relative definition of $\Delta\Delta\mathrm{G}$.

2.3. Mutation-Site Indicator Encoding

In mutation-induced ΔΔG prediction tasks, the perturbation is typically confined to a single amino acid position. However, full-protein sequence- or structure-based representations do not, by default, explicitly distinguish the mutated site from all other residues, which may weaken the model's ability to focus on the functionally critical position. To address this issue, we introduced a mutation-site indicator vector. Let $r_i$denote the residue position associated with the i-th node, and let r_{mut}denote the mutated residue position for the given sample. We define:

$$m_i = \begin{cases} 1, & r_i = r_{\mathrm{mut}} \\ 0, & r_i \neq r_{\mathrm{mut}} \end{cases}$$

where $m_i$is the mutation-site indicator variable for the i-th node. This indicator is appended to the protein node features as an additional input channel to explicitly mark the unique mutated site.

In addition, we constructed a 20-dimensional one-hot difference vector for the amino acid substitution and an 8-dimensional difference vector for physicochemical properties:

$$\Delta\mathrm{o} = \mathrm{o}^{MT} - \mathrm{o}^{WT}, \Delta\mathrm{p} = \mathrm{p}^{MT} - \mathrm{p}^{WT}.$$

Within the model, these difference vectors are first projected into the hidden space and then injected exclusively into the representation of the mutated node under the gating effect of m_i. Furthermore, the original difference vectors are concatenated with the global features at the regression stage to provide sample-level mutational attribute information.

2.4. Ligand Representation

To compare the effects of different representation paradigms on ΔΔG prediction, we adopted two types of ligand representations in this study: one based on vectorized features (non-graph

 doi:10.20944/preprints202604.0069.v1

representation) and the other based on an atom-level graph representation (graph-based model). Both representations were derived by parsing the ligand three-dimensional structure files in SDF format.

2.4.1. Vectorized Ligand Representation

Under the vectorized setting, each ligand was encoded using topology-based Morgan fingerprints (i.e., ECFP) [27], yielding a fixed-length binary vector:

$$\mathrm{f}_{\mathrm{lig}} = ECFP(\mathrm{ligand}; R, B) \in \{0,1\}^{B},$$

where R denotes the fingerprint radius (set to $R = 2$ in this study), and B denotes the fingerprint bit length (set to $B = 1024$ in this study). This vector was used as the ligand-side input feature and was combined with the protein-side representation for downstream regression prediction.

2.4.2. Graph-Based Ligand Representation

In the graph-based setting, each ligand was represented as an atom-level graph $G_{\mathrm{lig}} = (V_{\mathrm{lig}}, E_{\mathrm{bond}})$, where V_{lig} denotes the set of ligand atoms and E_{bond} denotes the set of covalent bond connections. For each atom node, a 40-dimensional atomic feature vector was constructed, and for each bond edge, a 20-dimensional edge feature vector was defined, incorporating information such as bond type and bond length. To ensure consistency among atom indexing, atomic coordinates, and bond connectivity, the ligand molecules were standardized, and explicit hydrogen atoms were removed.

Within the graph-based model, the ligand graph was further connected to the protein residue graph via cross-type edges defined by spatial contact relationships, thereby enabling the explicit representation of protein–ligand interactions. The corresponding graph construction details and the heterogeneous graph neural network architecture are described in full in the following sections.

2.5. Protein Feature Construction Schemes

To systematically assess the effects of different representation strategies on ΔΔG prediction performance, we defined five input feature configurations (Mode 0–4). All modes were evaluated under a unified training framework to quantify the contributions of graph-based modeling and different protein embedding strategies.

Mode 0: Vector Baseline Model

In this mode, no protein–ligand graph structure was constructed. Instead, regression was performed using sample-level vector representations. The protein feature was defined as the global mean-pooled representation of sequence-level embeddings, concatenated with the mutation difference vectors and the ligand molecular fingerprint (ECFP), forming the model input. This mode was used to evaluate baseline performance in the absence of explicit structural information.

Mode 1: Graph + One-hot

In this mode, a protein–ligand heterogeneous graph was constructed, and protein nodes were represented solely by amino acid one-hot encodings. Mutation-related difference vectors were injected into the mutated residue node via a mutation-aware mechanism and subsequently concatenated with global features at the readout stage. This mode was used to evaluate the representational capacity of basic residue identity information within a graph-based framework.

Mode 2: Graph + One-hot + WT-ESM

Based on Mode 1, residue-level ESM embeddings of the wild-type sequence were incorporated to enhance the representation of evolutionary and contextual semantic information.

Mode 3: Graph + One-hot + ΔESM

In this mode, residue-level differential embeddings (ΔESM) were used in place of absolute embedding representations, enabling explicit modeling of mutation-induced perturbation effects.

Mode 4: Graph + One-hot + WT-ESM + MT-ESM

 doi:10.20944/preprints202604.0069.v1

In this mode, both wild-type and mutant residue-level embeddings were retained, allowing the model to learn their relational differences directly within the graph-based architecture.

2.6. Graph Construction

The graph construction pipeline was partially adapted from the open-source GEMS implementation [20]. At the same time, the final representation was redesigned as a heterogeneous graph to explicitly model different node and edge types, following the general paradigm of graph neural network representations for molecular interactions [28–30].

To characterize the structural interaction patterns of protein–ligand complexes in the context of mutation, each complex was represented as a heterogeneous graph.

$$G = (\mathcal{V}, \mathcal{E}),$$

where the node set $\mathcal{V}$ consists of two types of entities: ligand heavy-atom nodes and protein residue nodes.

Unlike approaches that retain only residues within the local binding pocket, this study preserves the full protein structure as graph input. This design was motivated by the observed data distribution (Figure 1), which shows that a substantial proportion of mutation sites in MdrDB are located far from the ligand, exhibiting a pronounced long-tailed distance distribution. For example, samples with mutation–ligand distances greater than 15 Å account for a non-negligible fraction of the dataset. If only a pocket subgraph were constructed, such cases would be difficult to represent adequately because the spatial context surrounding the mutated site would be missing. By retaining the full protein structure, the model has a more complete structural context for capturing the indirect effects of distal mutations, including potential long-range coupling or allosteric effects, thereby improving representational completeness and robustness.

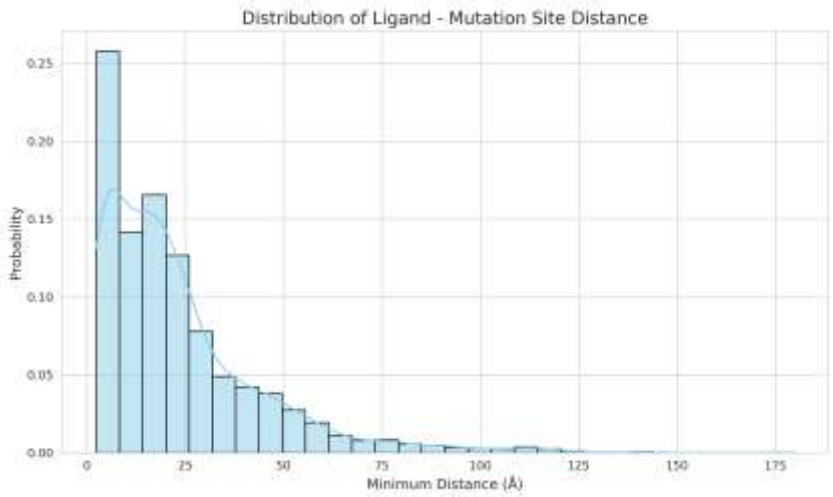

Figure 1. Distribution of distances between the ligand and mutation sites.

Ligands were represented with heavy atoms as nodes, and their three-dimensional coordinates were obtained from the SDF conformations. Ligand node features comprised one-hot and numerical encodings of basic chemical attributes, including atom type, ring membership, hybridization state, formal charge, aromaticity, atomic mass, number of attached hydrogens, atomic degree, and chirality. Proteins were represented at the level of standard amino acid residues. Node coordinates were assigned using the C_α atom position whenever available; if the C_α atom was missing, the geometric center of the residue's atomic coordinates was used instead. The basic protein node feature was a 20-dimensional one-hot encoding of amino acid identity. When language-model-derived information was incorporated, residue-level ESM representations were further appended (in the form of WT, ΔESM, or WT+MT concatenation; see Section 2.5).

 doi:10.20944/preprints202604.0069.v1

The edge set $\mathscr{E}$ consisted of three types of structural relationships, and bidirectional edges were constructed for each relation. In addition, self-loop edges on a dedicated self-relation were introduced to stably preserve node-specific information. All edge features were uniformly encoded as 20-dimensional vectors, including relation-type indicators and geometric distance information, with distance terms normalized by 10 Å to improve numerical scaling.

(1) Within the ligand, atom–atom edges were constructed according to the covalent bonds identified by RDKit [31]. Edge features included bond type (single, double, triple, or aromatic), bond length, conjugation status, ring membership, and stereochemical information.

(2) Within the protein, two types of residue–residue edges were defined. The first type consisted of sequential adjacency edges between neighboring residues on the same chain, preserving primary sequence connectivity. The second type consisted of spatial *k*-nearest-neighbor (*k*-NN) edges constructed from inter-residue distances ($k = 16$ in this study), enabling the representation of the local three-dimensional structural neighborhood. For these intra-protein edges, the edge features primarily encoded the Euclidean distance between nodes.

(3) Contact edges were constructed between the protein and the ligand. Specifically, an interaction edge was introduced whenever the distance between a ligand atom and any protein residue atom was below a predefined threshold (8 Å in this study), indicating a potential spatial interaction. The corresponding edge features included the distances between the ligand atom and key backbone-related atoms of the residue, such as N, C_α, C, and a virtual C_β. All edges were instantiated in bidirectional form, and self-loops were added to enhance the retention of node-specific information.

Global Virtual Master Node. To provide a graph-level contextual representation across node types and strengthen global aggregation during the readout stage, we further introduced a global virtual master node. Its coordinate was placed at the geometric center of the protein and ligand nodes, and edges were constructed between the master node and all ligand atom nodes as well as all protein residue nodes. During readout, the graph-level representation of the master node, $\mathrm{g}_{\mathrm{master}}$, was extracted and incorporated into the final feature fusion.

Through this construction, each complex was represented as a heterogeneous graph that jointly encodes molecular topology and the three-dimensional neighborhood relationships within the protein, thereby providing a unified structural representation for subsequent graph neural network modeling.

2.7. Model Architecture

To characterize the local perturbation induced by a single-point mutation within the protein–ligand interaction network, as well as its propagation through the three-dimensional structural context, we developed a graph neural network model based on a protein–ligand heterogeneous graph, together with a vector-based baseline model as a structure-agnostic control (Figure 2). Both models are built on the central principle of mutation-conditioned modeling, in which mutational attributes are explicitly incorporated during representation learning, enabling the models to directly capture how local residue perturbations influence global binding behavior.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 April 2026 doi:10.20944/preprints202604.0069.v1

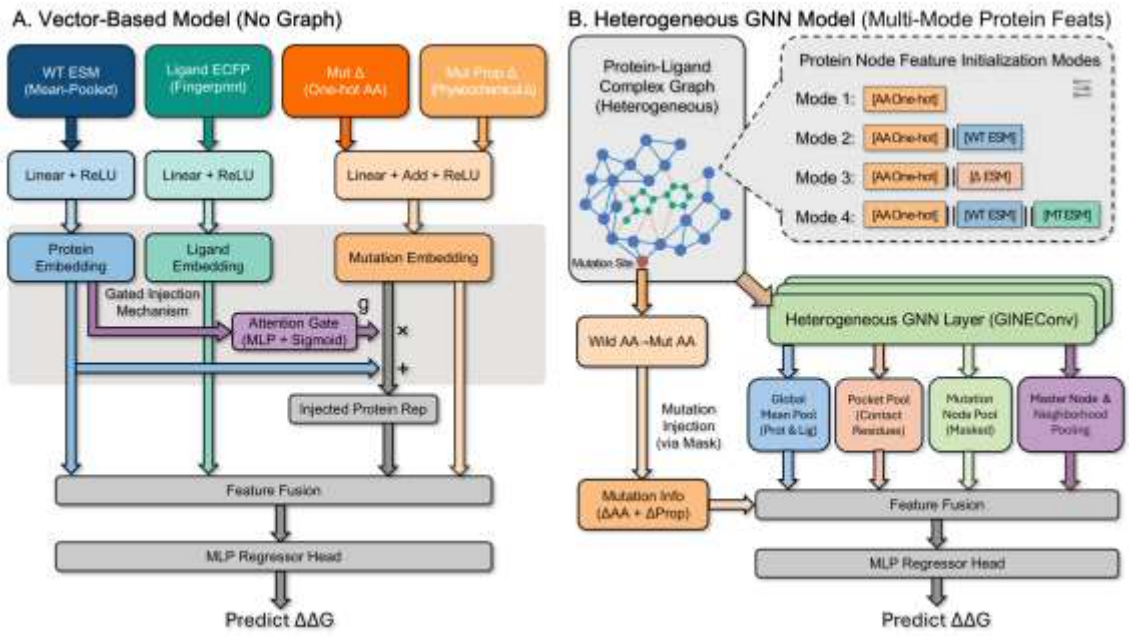

Figure 2. Overview of the model architecture.

In the vector baseline model, each sample is represented by the sequence-embedding-based representation of the wild-type protein, h_p, and the molecular fingerprint representation of the ligand, h_l. For the mutation information, the amino acid identity difference (ΔAA) and the physicochemical property difference ($\Delta Prop$) are independently projected through separate linear layers, then summed and activated to yield a fused mutation embedding h_m. To enhance the modulatory effect of the mutation on the protein representation, a gated injection mechanism is introduced:

$$\mathrm{g} = \sigma(\mathrm{MLP}([\mathrm{h}_p \parallel \mathrm{h}_m])), \tilde{\mathrm{h}}_p = \mathrm{h}_p + \mathrm{g} \odot \mathrm{h}_m$$

Here, σ denotes the sigmoid function, and $\odot$ denotes element-wise multiplication. This mechanism allows the model to adaptively regulate the strength of mutational information injection according to the current state of the protein representation, thereby enabling mutation-conditioned modeling. Finally, the model adopts a residual concatenation strategy, in which the original global protein representation h_p, the ligand representation h_l, the mutation-only representation h_m, and the context-aware protein representation $\tilde{\mathrm{h}}_p$ after gated mutation injection are concatenated and fed into a multilayer perceptron (MLP) to predict the $\Delta\Delta G$ value. This concatenation scheme, which preserves residual access to the original features, ensures direct access to both global contextual information and local perturbation signals.

In the heterogeneous graph model, each protein–ligand complex is represented as a heterogeneous graph $G = (V, E)$, in which the nodes comprise protein residues and ligand atoms. To provide a graph-level contextual representation across node types, an additional global virtual master node is introduced during graph construction and connected to the real nodes; its representation, $\mathrm{g}_{\mathrm{master}}$, is then extracted during the readout stage and incorporated into the final fusion. Node features are first projected through linear layers to obtain the initial representations $\mathrm{h}_i^{(0)}$, after which multiple layers of heterogeneous message-passing operations based on edge features are stacked to iteratively update the node representations, combining HeteroConv and GINEConv [32]. The update rule for a single layer can be abstractly written as:

$$\mathrm{h}_i^{(k+1)} = \phi^{(k)}\left(\mathrm{h}_i^{(k)}, \sum\nolimits_{j \in \mathcal{N}(i)} \psi^{(k)}\left(\mathrm{h}_i^{(k)}, \mathrm{h}_j^{(k)}, \mathrm{e}_{ij}\right)\right)$$

 doi:10.20944/preprints202604.0069.v1

Here, e_{ij} denotes the edge feature, ψ denotes the message function that jointly incorporates node and edge attributes, and ϕ denotes a nonlinear mapping. Through multi-layer propagation, the model progressively integrates local geometric topology and cross-type interaction information.

On the protein node side, the model supports multiple residue input configurations (Modes 1–4), including amino acid identity encoding alone, concatenation with wild-type embedding representations, differential embedding representations, and the simultaneous use of both wild-type and mutant embeddings. This design was introduced to analyze the contribution of different sequence representation strategies within the structural modeling framework. Based on the mutation-site indicator mask $m_i \in \{0,1\}$ defined above, the model constructs a graph-level mutation embedding u, which is then selectively injected into the corresponding residue node representation in conjunction with an independent positional mask bias:

$$\mathrm{h}_i' = \mathrm{h}_i + \mathrm{Linear}_{mask}(m_i) + m_i \cdot \mathrm{u}$$

The mutated site receives an additional mutation-injection term, while the mask channel is incorporated into representation learning as an explicit positional marker.

During the readout stage, the model constructs multi-scale graph-level representations to comprehensively capture structural features ranging from global context to local perturbations. Specifically, the model extracts a globally pooled protein representation $\mathrm{g}_{\mathrm{prot}}$, a pooled ligand representation $\mathrm{g}_{\mathrm{lig}}$, and the representation of the global master node, $\mathrm{g}_{\mathrm{master}}$. At the local level, the single-node representation of the mutated site, $\mathrm{g}_{\mathrm{mut}}$, is extracted, and a binding-pocket representation, $\mathrm{g}_{\mathrm{pocket}}$, is derived based on the contact edges.

More importantly, to explicitly infer mutation-induced steric effects and microenvironmental rearrangements from a static wild-type structure, we further designed a spatial neighborhood pooling mechanism based on Gaussian radial basis functions [33]. Let d_j denote the true three-dimensional Euclidean distance between the j-th residue and the mutated site. The corresponding attention decay weight is defined as

$$w_j \propto \exp\left(-\frac{d_j^2}{2\sigma^2}\right)$$

which is used to perform distance-weighted aggregation over the local physical neighborhood of the mutation, yielding the microenvironment representation $\mathrm{g}_{\mathrm{nei}}$. This mechanism enables the model to more faithfully reflect the prior physical principle that intermolecular interactions decay with spatial distance.

Finally, to prevent critical low-level physicochemical perturbation signals from being excessively smoothed during message propagation through deep graph layers, the model adopts a feature-bypass strategy. Specifically, the multi-scale graph representations described above are directly concatenated with the original lossless mutation descriptors (ΔAA and $\Delta Prop$), which bypass the graph network entirely, to form the final fused representation:

$$\mathrm{z} = [\mathrm{g}_{prot} \parallel \mathrm{g}_{lig} \parallel \mathrm{g}_{master} \parallel \mathrm{g}_{pocket} \parallel \mathrm{g}_{mut} \parallel \mathrm{g}_{nei} \parallel \Delta\mathrm{AA} \parallel \Delta\mathrm{Prop}]$$

The fused representation vector z is subsequently fed into a regression network to generate the final prediction, i.e., $\hat{y} = \mathrm{MLP}(\mathrm{z})$. Model training is optimized using the Huber loss, which improves robustness to outliers and noisy labels commonly present in experimental measurements.

2.8. Training Strategy

Model performance was evaluated using five-fold cross-validation. In each fold, approximately 20% of the samples were held out as the test set. In comparison, the remaining 80% were further divided into training and validation sets for model selection and early stopping. Accordingly, the overall data split was approximately 0.68/0.12/0.20 for the training, validation, and test sets, respectively.

 doi:10.20944/preprints202604.0069.v1

To improve robustness to outlier samples, model training was optimized using the Huber loss [34]. To stabilize the training process, the ΔΔG labels in the training set were standardized via *z*-score normalization, and predictions were rescaled to the original scale during validation and testing before metric calculation. Model parameters were updated using the AdamW optimizer [35], and an early stopping strategy was adopted to prevent overfitting. All experiments were conducted under fixed random seeds to ensure reproducibility.

Model performance was assessed using mean absolute error (MAE), root mean square error (RMSE), the Pearson correlation coefficient, and the Spearman correlation coefficient. The final results are reported as the average test-set performance across the five folds.

All deep learning models in this study were implemented using the PyTorch and PyTorch Geometric frameworks [36,37]. For the graph-based models (Mode 1–4), we adopted a heterogeneous graph neural network architecture comprising two message-passing layers (HeteroConv combined with GINEConv), with a hidden dimension of 128 and a dropout rate of 0.35; inter-layer feature aggregation was performed using sum reduction. The vector-based baseline model (Mode 0) was implemented as a multilayer perceptron (MLP) with a hidden dimension of 512 and a dropout rate of 0.25. All models were optimized using AdamW. To mitigate overfitting, an early stopping mechanism was employed with a patience of 20 epochs. Detailed hyperparameter settings and training configurations are provided in Supplementary Table S1.

3. Results and Discussion

3.1. Overall Predictive Performance under Different Data Splitting Strategies

To systematically evaluate the performance of different feature configuration modes in the ΔΔG prediction task, we conducted five-fold cross-validation experiments under two dataset partitioning schemes: random sample splitting (Random split) and strict UniProt-based grouped splitting (UniProt split). All metrics are reported as the mean ± standard deviation across the five test folds.

Under the Random split setting, all models achieved moderate predictive performance, with Pearson correlation coefficients generally ranging from 0.48 to 0.55.

Among them, Mode 0 and Mode 2 showed the best performance, reaching Pearson correlations of 0.5545 ± 0.0581 and 0.5549 ± 0.0271, respectively, with both models yielding RMSE values close to 1.0. By contrast, Mode 1, which used only amino acid one-hot representations, exhibited the weakest performance (Pearson = 0.4873 ± 0.0156, RMSE = 1.0475 ± 0.0392), indicating that residue identity information alone is insufficient to adequately capture the effects of mutations on binding free energy.

The incorporation of pretrained protein language model embeddings (Mode 2–4) consistently improved performance relative to the one-hot-only representation (Mode 1), indicating that sequence contextual information makes a substantial contribution to ΔΔG prediction. However, the performance differences among the different ESM-based representation strategies were relatively limited. Neither the differential embedding representation (Mode 3) nor the WT+MT concatenated representation (Mode 4) showed a clear advantage over the single WT embedding (Mode 2). This may suggest that, under the Random split setting, substantial protein-level similarity exists between the training and test samples, allowing the model to obtain sufficiently informative signals even from relatively simple feature formulations.

Overall, model performance under Random splitting was stable, with relatively small standard deviations across folds, indicating good training consistency (Table 1).

Table 1. Performance metrics under the Random split setting (kcal/mol).

Mode	Pearson	Spearman	MAE	RMSE
Mode0	0.5545±0.0581	**0.5411±0.0399**	**0.7121±0.0158**	**0.9941±0.0303**
Mode1	0.4873±0.0156	0.4554±0.0290	0.7567±0.0359	1.0475±0.0392
Mode2	**0.5549±0.0271**	0.5255±0.0411	0.7223±0.0283	0.9971±0.0525
Mode3	0.5194±0.0385	0.4896±0.0275	0.7471±0.0340	1.0281±0.0525

 doi:10.20944/preprints202604.0069.v1

Mode4	0.5388±0.0438	0.5076±0.0500	0.7356±0.0387	1.0130±0.0574

Under the strict UniProt-based split protocol, the performance of all models declined markedly (Table 2). The Pearson correlation coefficients ranged from 0.04 to 0.15 overall, while the RMSE increased to approximately 1.25–1.31.

Table 2. Performance metrics under the strict UniProt-based split setting (kcal/mol).

Mode	Pearson	Spearman	MAE	RMSE
Mode0	0.0439±0.0774	0.0436±0.0830	0.9308±0.1600	1.3101±0.3731
Mode1	0.0897±0.1472	**0.1418±0.0973**	0.8679±0.0588	1.2558±0.2423
Mode2	0.1001±0.1108	0.0906±0.1160	0.8645±0.0383	**1.2536±0.2421**
Mode3	**0.1526±0.1481**	0.0863±0.1724	0.8838±0.0907	1.2539±0.2749
Mode4	0.1033±0.1163	0.0960±0.1336	**0.8535±0.0405**	1.2458±0.2357

Taking Mode 0 as an example, its Pearson correlation dropped from 0.5545 under the Random split to 0.0439 under the UniProt-based split, indicating that the predictive correlation was nearly lost. A similar trend was observed across the other modes. This finding suggests that, under Random splitting, the models may benefit from protein-level similarity between samples. In contrast, strict UniProt-based partitioning leads to a substantial decline in generalization performance on unseen proteins.

Among the different modes, Mode 3 achieved the highest Pearson correlation (0.1526 ± 0.1481), whereas Mode 4 performed slightly better in terms of MAE and RMSE. However, the overall differences among the modes remained limited, and the standard deviations increased markedly, indicating reduced model stability in the cross-protein prediction setting.

Taken together, we found that under the Random split setting, the vector-based model (Mode 0) and the graph-based models (Modes 2–4) exhibited comparable performance, with no clear advantage of explicit structural modeling. This may be attributable to the high degree of protein-level similarity between training and test samples under random partitioning, such that relatively simple feature representations are sufficient to capture the dominant statistical patterns.

By contrast, under the strict UniProt-based split, the graph-based models consistently outperformed the vector baseline in terms of Pearson correlation. For example, Mode 3 achieved a Pearson correlation of 0.1526, whereas Mode 0 reached only 0.0439. Although the absolute correlations remained low and the standard deviations were larger, this trend suggests that structural modeling may provide complementary information in the cross-protein prediction scenario.

3.2. Scatter Plot and Regression Trend Analysis

To illustrate representative prediction distributions, we visualized the test-set predictions from the best-performing model in a selected fold, with model selection based on validation-set performance in that fold. The ideal diagonal line and the fitted linear regression line were overlaid for reference (Figure 3). The primary conclusions, however, are based on the mean ± standard deviation across all five folds.

Under the Random split setting, the models exhibited an overall favorable fit, with predicted values clustered near the diagonal, indicating reasonable agreement with the experimental measurements. After incorporating graph-structural information and protein embeddings (Mode 2), performance remained comparable to the vector-based baseline, with the highest Pearson correlation reaching 0.578. This suggests that when training and test samples share substantial protein-level background information, the models can already fit the data effectively by exploiting local statistical regularities, and more complex structural modeling confers little additional performance benefit in this setting.

 doi:10.20944/preprints202604.0069.v1

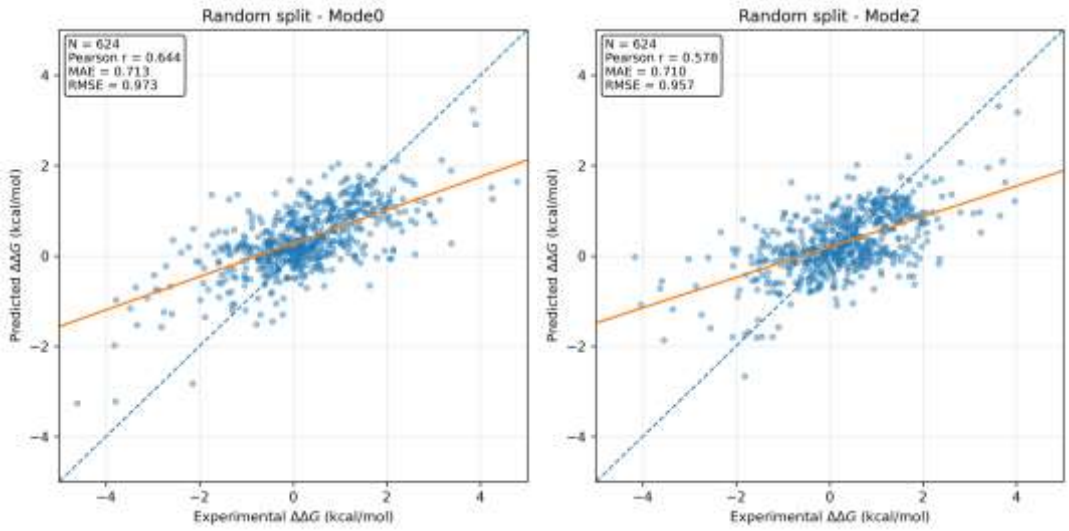

Figure 3. Predicted versus experimental ΔΔG under the Random split setting.

By contrast, under the strict UniProt-based split, model generalization deteriorated substantially. The Pearson correlation of Mode 0 decreased markedly, and the predicted range became noticeably compressed, with a flatter regression slope, indicating that the model struggled to maintain stable predictive performance on previously unseen protein backgrounds. This observation further supports the concern that Random splitting may overestimate model performance. Notably, however, under this more stringent setting, the graph-based model (Mode 3) demonstrated better generalization than the vector-based model. Its prediction distribution was visibly more dispersed than that of Mode 0, although the overall correlation remained weak and the variance was larger. Nevertheless, these results suggest that explicit modeling of protein–ligand structural relationships, together with mutation-conditioned mechanisms, can help mitigate the decline in cross-protein generalization performance.

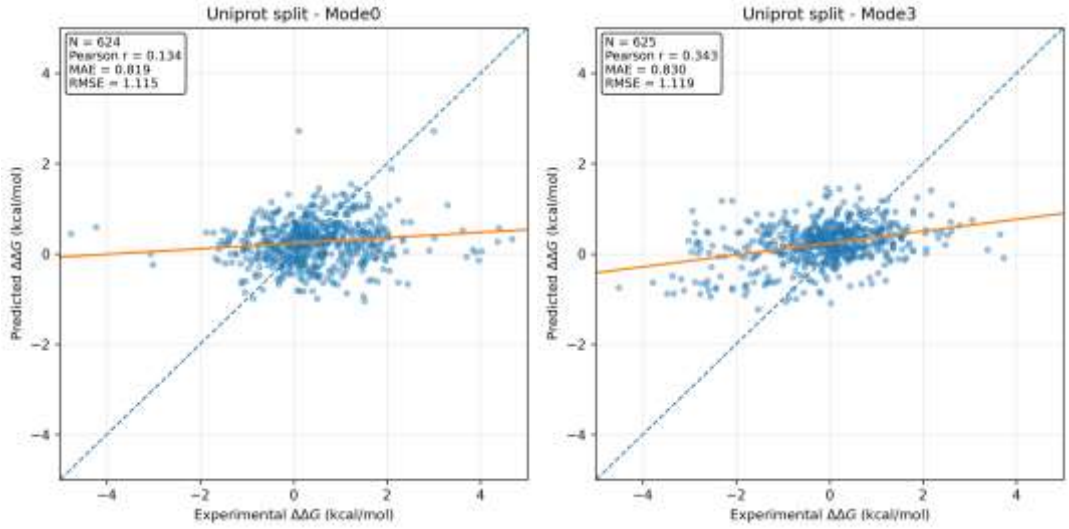

Figure 4. Predicted versus experimental ΔΔG under the UniProt-based split setting.

Taken together, the Random split primarily reflects the models' ability to fit, whereas the UniProt-based split more faithfully evaluates their generalization performance in realistic cross-protein prediction scenarios. Under the strict split protocol, graph-based structural modeling and mutation-differential representations exhibit more evident advantages, supporting their necessity for cross-protein ΔΔG prediction tasks.

 doi:10.20944/preprints202604.0069.v1

3.3. Discussion of Generalization and Data Splitting Strategies

A further comparison of the two splitting strategies reveals that the models achieve relatively high correlations under Random splitting (Pearson > 0.5). In contrast, the correlations decline sharply under the strict UniProt-based split. This pattern suggests that when the training and test samples share protein backgrounds, the models may primarily learn protein-specific statistical patterns rather than transferable structure–energy relationships. Consequently, performance under Random splitting may overestimate the true predictive capacity of the models in realistic cross-protein prediction scenarios, a phenomenon widely reported in recent studies that use machine-learning benchmarks for biological macromolecules [38–40].

To quantify the extent of potential protein-level leakage, we examined whether the proteins appearing in the test set were already present in the training set under different partitioning strategies. Under the Random split protocol, more than 95% of the test samples correspond to proteins present in the training set, indicating extensive protein-level overlap between the two sets. In contrast, the UniProt-based split strictly enforces 0% protein-level overlap, ensuring that all test proteins are completely unseen during training. This fundamental difference in data distribution provides a straightforward explanation for the substantial performance gap observed between the two splitting strategies.

Notably, our observations are consistent with recent findings reported by Xie et al. [22], who also demonstrated that ΔΔG prediction models based on protein language model embeddings exhibit substantial performance degradation under strict protein-level partitioning. Notably, the behavior of our Mode 0 vector baseline—constructed from global protein embeddings and mutation descriptors—closely resembles the modeling paradigm adopted in such sequence-based approaches. The similar performance trends observed in both studies, therefore, provide independent evidence that the apparent predictive success reported under Random split settings is largely driven by protein-level similarity between training and test samples rather than true cross-protein generalization.

Under the UniProt-based split, models relying on sequence-derived and global vector representations showed correlations close to chance. In contrast, graph-based models that explicitly encode protein–ligand structural relationships showed a more stable improvement in predictive capability, with the Pearson correlation increasing from 0.0439 to 0.1526. Although the magnitude of this improvement remains limited, the results indicate that structure-conditioned mechanisms help mitigate the decline in cross-protein generalization performance. Nevertheless, the overall correlation remains low, suggesting that relying solely on the wild-type complex structure and mutation-differential information is insufficient, at the current data scale, to fully capture the mechanisms underlying mutation-induced changes in binding free energy.

3.4. Representation Space Analysis of Wild-Type ESM and ΔESM Embeddings

To further understand why Mode 3 (the ΔESM differential representation) exhibited relatively more robust performance under the strict UniProt-based split, we performed a visualization analysis of the representational space formed by the WT embeddings and the ΔESM embeddings. Specifically, for each sample, the 1152-dimensional ESM-C representation was standardized, subjected to dimensionality reduction, and then projected into a two-dimensional space using UMAP [41]. The resulting embedding distributions were colored according to UniProt ID.

As shown in Figure 5, the WT embeddings exhibit a pronounced protein-identity-driven clustering pattern in the representation space. Samples from the same UniProt cluster tightly together, while samples from different proteins form relatively distinct, separated groups. This indicates that the original WT representations retain substantial protein-identity-related information; that is, the embeddings encode, to a considerable extent, background features associated with which protein a given sample belongs to. Under the Random split setting, such protein-level structure may provide the model with indirect identity cues, thereby improving predictive performance.

 doi:10.20944/preprints202604.0069.v1

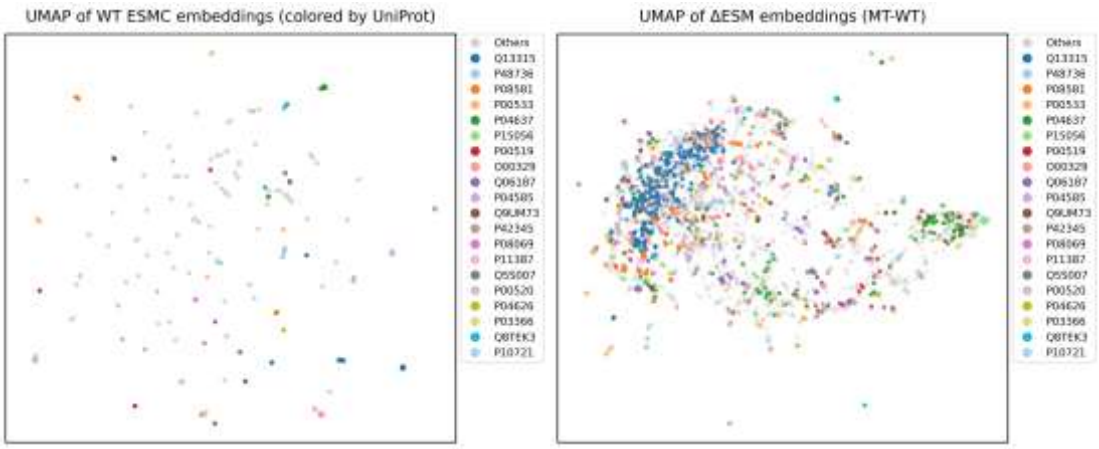

Figure 5. UMAP visualization of WT and ΔESM representations.

In contrast, the ΔESM embeddings display a markedly different geometric organization. Samples from different UniProt IDs are more intermixed in the representation space, and the protein-level clustering pattern is substantially attenuated. This suggests that the differential representation partially suppresses background information associated with protein identity, making the representation more focused on the relative perturbation signal induced by mutation rather than on the overall protein-specific context.

It should be emphasized that this visualization alone does not directly establish a causal explanation for the observed performance differences. However, from the perspective of representation-space geometry, it provides mechanistic support for the relatively greater robustness of ΔESM under the strict cross-protein generalization setting. This observation is consistent with our hypothesis that differential representations partially mitigate the protein-background bias.

3.5. Analysis of Dataset Characteristics and Factors Limiting Model Performance

To investigate the fundamental reasons for the limited model performance under the strict UniProt-based split, we conducted an in-depth analysis not only of the model architecture itself but also of the dataset's intrinsic distributional properties and the physical limitations imposed by the task formulation.

(1) Label noise and data heterogeneity. As an integrated database, MdrDB aggregates measurements collected from multiple literature sources and experimental conditions. Within the MdrDB dataset, we observed that a small subset of identical protein–ligand–mutation samples appear multiple times because the database integrates measurements from different literature sources. These duplicated entries occasionally contain discrepant ΔΔG values. As shown in Table 3 (see also Supplementary Table S2 for the complete data), selected representative mutations in EGFR and BRAF illustrate that repeated records exhibit numerical variation and, in a few cases, even disagree on the direction of activity change (i.e., opposite signs).

Table 3. Examples of duplicated protein–ligand–mutation entries in the MdrDB dataset that exhibit discrepant ΔΔG values across different literature sources.

UniProt ID	Mutation	Drug	ΔΔG (Source 1)	ΔΔG (Source 2)	Abs. Diff
P00533 (EGFR)	L858R	Gefitinib	-4.04	-1.6	2.44
P00533 (EGFR)	G719S	Gefitinib	-1.53	0.48	2.01

 doi:10.20944/preprints202604.0069.v1

P15056 (BRAF)	L597V	PLX-4720	-1.65	0.35	2.00
P15056 (BRAF)	A728V	PLX-4720	-2.51	-0.64	1.87

We therefore further examined the consistency of ΔΔG annotations among these repeated experimental records. Under identical conditions (UniProt, mutation, drug), a small number of samples showed substantial discrepancies in their annotated ΔΔG values. Such conflicts are more likely attributable to genuine systematic variation—arising, for example, from differences in assay methodology, temperature, pH, or other experimental conditions [42–44]—and thus represent an unavoidable form of observational noise in real-world data. In our dataset, these samples account for approximately 0.5% of all entries. For two reasons, we chose to retain these noisy samples rather than remove them manually. First, their exclusion could introduce selection bias and artificially inflate model performance. Second, in real deployment settings, a useful model must be robust to this type of noise. Accordingly, all samples were retained in the main experiments, and all discrepant records are additionally reported in the supplementary analysis. This observation also suggests that intrinsic experimental variability in ΔΔG measurements may impose a practical upper bound on achievable prediction accuracy.

(2) Sparse coverage of protein space. The dataset contains only approximately 3,000–4,000 samples, and under the UniProt-based split, the training and test sets often correspond to entirely different protein families or fold types. Given the vastness of protein chemical space, the current data density is far from sufficient to support the learning of general physical rules that transfer across families. The strong performance observed under Random splitting, therefore, obscures the model's vulnerability to out-of-distribution (OOD) samples.

(3) Physical limitations of static structures. This represents a more fundamental theoretical bottleneck. Changes in binding free energy are jointly determined by enthalpic (ΔH) and entropic (ΔS) contributions [45]. Under the WT-only setting, the graph-based model primarily captures the enthalpic component arising from steric and electrostatic interactions encoded in the static structure, but has limited ability to infer mutation-induced changes in conformational dynamics (i.e., entropic effects). The absence of dynamic information therefore imposes an intrinsic physical constraint, particularly when predicting resistance mutations driven predominantly by allosteric effects or changes in molecular flexibility [46].

Taken together, the intrinsic noise and sparse coverage of the dataset, along with the loss of dynamic information under the WT-only setting, define the current performance boundary for existing methods in strict cross-protein prediction tasks.

4. Conclusions

In this study, we systematically evaluated the effects of data partitioning strategies and protein representation schemes on mutation-induced ΔΔG prediction in the more practically relevant setting where only the wild-type complex structure is available (WT-only). Controlled experiments on the MdrDB CoreSet showed that the models achieved moderate correlation under Random splitting. In contrast, performance declined sharply under a strict UniProt-based cross-protein split, indicating that random partitioning may overestimate cross-target generalization due to protein-level overlap between the training and test sets.

Under the strict UniProt-based split, graph-based models that explicitly encode protein–ligand geometric relationships retained weak but nonzero predictive correlation relative to the vector baseline. In contrast, purely sequence-based or global-vector baselines performed close to random. This result suggests that, in the absence of post-mutation structural information, geometric constraints and structural context can provide complementary signals for cross-protein prediction. However, reliance on only the static WT structure and mutation descriptors remains insufficient for reliable out-of-distribution (OOD) ΔΔG prediction. Although the UniProt-based split effectively

 doi:10.20944/preprints202604.0069.v1

prevents direct memorization of identical targets, latent homologous sequence similarity across target families may persist. Future benchmarking should therefore adopt more stringent partitioning strategies based on sequence-similarity clustering to probe a more realistic lower bound on generalization.

Taken together with our analyses of data consistency and the task's physical properties, we conclude that the current bottlenecks arise primarily from three factors: intrinsic experimental noise and heterogeneity in the dataset, sparse coverage of protein space, and the absence of conformational and entropic information in the WT-only static-structure setting. Overall, this study emphasizes the importance of enforcing strict protein- and family-level partitioning and reporting cross-protein evaluation results to assess model generalizability more objectively. At the same time, it suggests that future progress in WT-only drug resistance prediction will likely require stronger physical priors, dynamics-related features, or higher-quality cross-protein data coverage to break through the current performance ceiling.

5. Data Availability

The primary dataset used in this study is obtained from the Mutation-induced Drug Resistance Database (MdrDB), available at https://quantum.tencent.com/mdrdb/.

All processed datasets, model implementations, and evaluation scripts are publicly available at https://github.com/LakiAo/WT-ddG-CriticalEval to ensure reproducibility.

The pretrained protein language model ESM-C (600M parameters) used for feature extraction is available on Hugging Face at https://huggingface.co/EvolutionaryScale/esmc-600m-2024-12.

Supplementary Materials: The following supporting information can be downloaded at the website of this paper posted on Preprints.org.

References

1. Vasan N, Baselga J, Hyman DM. A view on drug resistance in cancer. Nature. 2019 Nov 1;575(7782):299–309. doi:10.1038/s41586-019-1730-1
2. Guerois R, Nielsen JE, Serrano L. Predicting Changes in the Stability of Proteins and Protein Complexes: A Study of More Than 1000 Mutations. Journal of Molecular Biology. 2002 Jul 5;320(2):369–87. doi:10.1016/S0022-2836(02)00442-4
3. Cournia Z, Allen B, Sherman W. Relative Binding Free Energy Calculations in Drug Discovery: Recent Advances and Practical Considerations. J Chem Inf Model. 2017 Dec 26;57(12):2911–37. doi:10.1021/acs.jcim.7b00564
4. Wang L, Wu Y, Deng Y, Kim B, Pierce L, Krilov G, et al. Accurate and Reliable Prediction of Relative Ligand Binding Potency in Prospective Drug Discovery by Way of a Modern Free-Energy Calculation Protocol and Force Field. J Am Chem Soc. 2015 Feb 25;137(7):2695–703. doi:10.1021/ja512751q
5. Atz K, Grisoni F, Schneider G. Geometric deep learning on molecular representations. Nat Mach Intell. 2021 Dec;3(12):1023–32. doi:10.1038/s42256-021-00418-8
6. Gaudelet T, Day B, Jamasb AR, Soman J, Regep C, Liu G, et al. Utilizing graph machine learning within drug discovery and development. Brief Bioinform. 2021 Nov 1;22(6):bbab159. doi:10.1093/bib/bbab159
7. Ferreira FJN, Carneiro AS. AI-Driven Drug Discovery: A Comprehensive Review. ACS Omega. 10(23):23889–903. doi:10.1021/acsomega.5c00549 PubMed PMID: 40547666; PubMed Central PMCID: PMC12177741.
8. Wang H. Prediction of protein–ligand binding affinity via deep learning models. Brief Bioinform. 2024 Mar 1;25(2):bbae081. doi:10.1093/bib/bbae081
9. Zhao L, Zhu Y, Wang J, Wen N, Wang C, Cheng L. A brief review of protein–ligand interaction prediction. Computational and Structural Biotechnology Journal. 2022 Jan 1;20:2831–8. doi:10.1016/j.csbj.2022.06.004
10. Verburgt J, Jain A, Kihara D. Recent Deep-Learning Applications to Structure-Based Drug Design. Methods Mol Biol. 2024;2714:215–34. doi:10.1007/978-1-0716-3441-7_13 PubMed PMID: 37676602; PubMed Central PMCID: PMC10578466.

 doi:10.20944/preprints202604.0069.v1

11. 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. 2025 Jul 1;41(Supplement_1):i362–72. doi:10.1093/bioinformatics/btaf228
12. Zhou Y, Pan Q, Pires DEV, Rodrigues CHM, Ascher DB. DDMut: predicting effects of mutations on protein stability using deep learning. Nucleic Acids Res. 2023 Jul 5;51(W1):W122–8. doi:10.1093/nar/gkad472
13. Pires DEV, Blundell TL, Ascher DB. mCSM-lig: quantifying the effects of mutations on protein-small molecule affinity in genetic disease and emergence of drug resistance. Sci Rep. 2016 Jul 7;6(1):29575. doi:10.1038/srep29575
14. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021 Aug;596(7873):583–9. doi:10.1038/s41586-021-03819-2
15. Pak MA, Markhieva KA, Novikova MS, Petrov DS, Vorobyev IS, Maksimova ES, et al. Using AlphaFold to predict the impact of single mutations on protein stability and function. PLoS One. 2023 Mar 16;18(3):e0282689. doi:10.1371/journal.pone.0282689 PubMed PMID: 36928239; PubMed Central PMCID: PMC10019719.
16. Buel GR, Walters KJ. Can AlphaFold2 predict the impact of missense mutations on structure? Nat Struct Mol Biol. 2022 Jan;29(1):1–2. doi:10.1038/s41594-021-00714-2 PubMed PMID: 35046575; PubMed Central PMCID: PMC11218004.
17. Nguyen T, Le H, Quinn TP, Nguyen T, Le TD, Venkatesh S. GraphDTA: predicting drug–target binding affinity with graph neural networks. Bioinformatics. 2021 May 23;37(8):1140–7. doi:10.1093/bioinformatics/btaa921
18. Li S, Zhou J, Xu T, Huang L, Wang F, Xiong H, et al. Structure-aware Interactive Graph Neural Networks for the Prediction of Protein-Ligand Binding Affinity. In: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining [Internet]. New York, NY, USA: Association for Computing Machinery; 2021 [cited 2026 Mar 4]. p. 975–85. (KDD '21). Available from: https://dl.acm.org/doi/10.1145/3447548.3467311 doi:10.1145/3447548.3467311
19. Satorras VG, Hoogeboom E, Welling M. E(n) Equivariant Graph Neural Networks [Internet]. arXiv; 2022 [cited 2026 Mar 5]. Available from: http://arxiv.org/abs/2102.09844 doi:10.48550/arXiv.2102.09844
20. Graber D, Stockinger P, Meyer F, Mishra S, Horn C, Buller R. Resolving data bias improves generalization in binding affinity prediction. Nat Mach Intell. 2025 Oct;7(10):1713–25. doi:10.1038/s42256-025-01124-5
21. Cao H, Wang J, He L, Qi Y, Zhang JZ. DeepDDG: Predicting the Stability Change of Protein Point Mutations Using Neural Networks. J Chem Inf Model. 2019 Apr 22;59(4):1508–14. doi:10.1021/acs.jcim.8b00697
22. Xie L, Bao G, Zhang D, Xu L, Xu X, Chang S. Evaluating data partitioning strategies for accurate prediction of protein-ligand binding free energy changes in mutated proteins. Computational and Structural Biotechnology Journal. 2025 Jan 1;27:4418–30. doi:10.1016/j.csbj.2025.10.020 PubMed PMID: 41169760.
23. Yang Z, Ye Z, Qiu J, Feng R, Li D, Hsieh C, et al. A mutation-induced drug resistance database (MdrDB). Commun Chem. 2023 Jun 14;6(1):123. doi:10.1038/s42004-023-00920-7
24. Rives A, Meier J, Sercu T, Goyal S, Lin Z, Liu J, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proceedings of the National Academy of Sciences. 2021 Apr 13;118(15):e2016239118. doi:10.1073/pnas.2016239118
25. Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science. 2023 Mar 17;379(6637):1123–30. doi:10.1126/science.ade2574
26. Hayes T, Rao R, Akin H, Sofroniew NJ, Oktay D, Lin Z, et al. Simulating 500 million years of evolution with a language model. Science. 2025 Feb 21;387(6736):850–8. doi:10.1126/science.ads0018
27. Rogers D, Hahn M. Extended-Connectivity Fingerprints. J Chem Inf Model. 2010 May 24;50(5):742–54. doi:10.1021/ci100050t
28. Wang K, Zhou R, Tang J, Li M. GraphscoreDTA: optimized graph neural network for protein–ligand binding affinity prediction. Bioinformatics. 2023 Jun 1;39(6):btad340. doi:10.1093/bioinformatics/btad340
29. Yang D, Kuang L, Hu A. Edge-enhanced interaction graph network for protein-ligand binding affinity prediction. PLOS ONE. 2025 Apr 8;20(4):e0320465. doi:10.1371/journal.pone.0320465

 doi:10.20944/preprints202604.0069.v1

30. Samudrala MV, Dandibhotla S, Kaneriya A, Dakshanamurthy S. PLAIG: Protein–Ligand Binding Affinity Prediction Using a Novel Interaction-Based Graph Neural Network Framework. ACS Bio Med Chem Au. 2025 Jun 18;5(3):447–63. doi:10.1021/acsbiomedchemau.5c00053
31. Landrum G. RDKit: Open-source cheminformatics [Internet]. 2016. Available from: https://www.rdkit.org
32. Xu* K, Hu* W, Leskovec J, Jegelka S. How Powerful are Graph Neural Networks? In. 2018 [cited 2026 Mar 4]. Available from: https://openreview.net/forum?id=ryGs6iA5Km
33. Schütt K, Kindermans PJ, Sauceda Felix HE, Chmiela S, Tkatchenko A, Müller KR. SchNet: A continuous-filter convolutional neural network for modeling quantum interactions. In: Advances in Neural Information Processing Systems [Internet]. Curran Associates, Inc.; 2017 [cited 2026 Mar 5]. Available from: https://proceedings.neurips.cc/paper/2017/hash/303ed4c69846ab36c2904d3ba8573050-Abstract.html
34. Huber PJ. Robust Estimation of a Location Parameter. The Annals of Mathematical Statistics. 1964 Mar;35(1):73–101. doi:10.1214/aoms/1177703732
35. Loshchilov I, Hutter F. Decoupled Weight Decay Regularization. In. 2018 [cited 2026 Mar 5]. Available from: https://openreview.net/forum?id=Bkg6RiCqY7
36. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In: Advances in Neural Information Processing Systems [Internet]. Curran Associates, Inc.; 2019 [cited 2026 Mar 4]. Available from: https://papers.neurips.cc/paper_files/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html
37. Fey M, Lenssen JE. Fast Graph Representation Learning with PyTorch Geometric [Internet]. arXiv; 2019 [cited 2026 Mar 5]. Available from: http://arxiv.org/abs/1903.02428 doi:10.48550/arXiv.1903.02428
38. Bernett J, Blumenthal DB, List M. Cracking the black box of deep sequence-based protein–protein interaction prediction. Brief Bioinform. 2024 Mar 5;25(2):bbae076. doi:10.1093/bib/bbae076 PubMed PMID: 38446741; PubMed Central PMCID: PMC10939362.
39. Wallach I, Heifets A. Most Ligand-Based Classification Benchmarks Reward Memorization Rather than Generalization. J Chem Inf Model. 2018 May 29;58(5):916–32. doi:10.1021/acs.jcim.7b00403
40. Bushuiev A, Bushuiev R, Sedlar J, Pluskal T, Damborsky J, Mazurenko S, et al. Revealing data leakage in protein interaction benchmarks [Internet]. arXiv; 2024 [cited 2026 Mar 5]. Available from: http://arxiv.org/abs/2404.10457 doi:10.48550/arXiv.2404.10457
41. McInnes L, Healy J, Saul N, Großberger L. UMAP: Uniform Manifold Approximation and Projection. Journal of Open Source Software. 2018 Sep 2;3(29):861. doi:10.21105/joss.00861
42. Kramer C, Kalliokoski T, Gedeck P, Vulpetti A. The Experimental Uncertainty of Heterogeneous Public Ki Data. J Med Chem. 2012 Jun 14;55(11):5165–73. doi:10.1021/jm300131x
43. Gilson MK, Zhou HX. Calculation of Protein-Ligand Binding Affinities*. Annual Review of Biophysics. Annual Reviews; 2007. p. 21–42. doi:https://doi.org/10.1146/annurev.biophys.36.040306.132550
44. Cournia Z, Chipot C, Roux B, York DM, Sherman W. Free Energy Methods in Drug Discovery—Introduction. In: Free Energy Methods in Drug Discovery: Current State and Future Directions [Internet]. American Chemical Society; 2021 [cited 2026 Mar 4]. p. 1–38. (ACS Symposium Series; no. 1397). Available from: https://doi.org/10.1021/bk-2021-1397.ch001 doi:10.1021/bk-2021-1397.ch001
45. Mobley DL, Dill KA. Binding of Small-Molecule Ligands to Proteins: "What You See" Is Not Always "What You Get." Structure. 2009 Apr 15;17(4):489–98. doi:10.1016/j.str.2009.02.010
46. Boehr DD, Nussinov R, Wright PE. The role of dynamic conformational ensembles in biomolecular recognition. Nat Chem Biol. 2009 Nov;5(11):789–96. doi:10.1038/nchembio.232