PepDDG: Peptide–Protein Binding $\Delta\Delta G$ Prediction via Information Channel Decomposition

Ruochi Zhang[1,3,*], **Yusi Fan**[3,4,5,*], **Qiong Zhou**[3,*], **Li Jiao**[3,*], **Tian Wang**[3], **Qian Yang**[3,4,5], **Silong Zhai**[3], **Lan Wang**[3], **Fengfeng Zhou**[4,5], **Yajuan Huang**[3], **Liming Guo**[3], **Chang Liu**[2,†], **Xin Gao**[6,7,8,†]

[1]Syneron Opal, George Town, Cayman Island
[2]Beijing Life Science Academy, Beijing, China
[3]Syneron Bio Beijing Co., Ltd.
[4]College of Computer Science and Technology, Jilin University, Changchun, China
[5]Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun, China
[6]Computer Science Program, Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia
[7]Center of Excellence for Smart Health (KCSH), King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia
[8]Center of Excellence on Generative AI, King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia

Abstract

Predicting mutation-induced changes in peptide–protein binding affinity ($\Delta\Delta G$) is central to therapeutic peptide optimization, but peptide-specific predictors remain limited by scarce labels, target overlap in supervised benchmarks, and costly molecular simulations. We introduce PepDDG, a zero-shot, training-free predictor that ranks peptide mutations by decomposing binding perturbations into three complementary information channels: energetic perturbation, geometric environment, and evolutionary compatibility. Each channel is computed from a wild-type complex structure, transformed into rank space, and combined by non-parametric Borda aggregation, without fitting fusion weights or fine-tuning neural predictors. PepDDG is motivated by rank-covariance analysis showing cross-channel fusion can improve rank correlation when channels retain complementary signal, whereas increasingly expensive refinement of a single energetic channel has diminishing returns. On our benchmark of 332 peptide-chain mutations across 33 targets, PepDDG reaches Spearman $\rho = 0.619$; the calibrated PepDDG-Cal variant reaches $\rho = 0.691$. PepDDG supports the recovery of promising candidates in therapeutic peptide screening, achieving 65.8% mean recall and 2.28-fold enrichment for non-destabilizing mutations through top-three selection. PepDDG-Cal achieves $\rho = 0.813$ on short peptides and 0.872 on the mixed four-target diagnostic. These results support complementary evidence fusion on PDB-derived complex structures as a practical alternative to costly same-channel physical refinement for peptide $\Delta\Delta G$ ranking; a separate diagnostic shows that predicted structures can be used when no crystal structure is provided at inference time.

1 Introduction

Short and cyclic peptides are increasingly important therapeutic modalities: they can achieve protein-like affinity and specificity while retaining tractable synthesis, and cyclization strategies improve

*These authors contributed equally to this work.
†Corresponding authors: Chang Liu (liuchang@blsa.com.cn) and Xin Gao (xin.gao@kaust.edu.sa).

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

protease resistance and oral bioavailability [1–3]. In lead optimization, ranking candidate mutations by their effect on binding free energy $\Delta\Delta G$ is a core bottleneck because labeled data are scarce, simulation-based refinement is expensive, and supervised evaluation is prone to leakage. Compared with protein–protein interfaces, peptide–protein complexes are smaller, more dynamic, more sensitive to single-residue substitutions, and often conformationally preorganized [4–6].

Most published $\Delta\Delta G$ predictors are benchmarked on datasets dominated by large protein–protein complexes. Empirical energy methods are cheap but noisy, MD/FEP-style methods are expensive and often unstable, and supervised deep-learning benchmarks trained on SKEMPI v2 [7] are vulnerable to target overlap [8]. This leaves a gap for peptide–protein $\Delta\Delta G$ prediction on new complexes using available complex structures, with predicted-structure inputs as an inference-time extension.

A practical question is: *which additional information improves prediction, and where should we spend compute?* We answer with an *information channel decomposition*: energetic perturbation, geometric environment, and evolutionary compatibility capture complementary aspects of mutational $\Delta\Delta G$. Same-channel refinement suffers diminishing returns (Theorem 2), whereas cross-channel fusion yields an algebraic rank-correlation bound under channel rank-orthogonality assumptions (Theorem 1).

Motivated by this analysis, we introduce **PepDDG**. It computes three channel scores from complex structures, rank-transforms each score to remove scale mismatch, and fuses them with a nonparametric Borda rank-sum. The method uses no learned fusion weights and does not fine-tune any neural predictor; its evolutionary channel queries a frozen, pretrained ProteinMPNN. The output orders a supplied candidate set rather than estimating calibrated values in kcal/mol. We also report **PepDDG-Cal**, a calibrated variant that separates the clean zero-shot algorithm from the empirical benefit of a small residual calibration. Figure 1 summarizes this three-channel decomposition and fusion pipeline.

Our contributions are as follows. **(1) Method.** We introduce PepDDG, a zero-shot rank-fusion algorithm, training-free with respect to binding $\Delta\Delta G$ supervision, for peptide $\Delta\Delta G$ ranking that combines energetic perturbation, geometric environment, and evolutionary compatibility without fine-tuning neural predictors or learning task-specific fusion weights. **(2) Theory.** We provide a rank-covariance analysis showing that cross-channel fusion can improve rank correlation when channels retain complementary signal, whereas repeated refinement within a single channel has diminishing returns. **(3) Empirical validation.** On PepDDG-Bench, zero-shot PepDDG reaches Spearman ρ = 0.619, improving pooled correlation over single-channel variants and the reported empirical and zero-shot protein-model references. Budgeted selection and per-target diagnostics clarify where this advantage supports screening; supervised references are reported with their coverage and training context. PepDDG-Cal further reaches ρ = 0.691. Performance remains strong on short peptides (ρ = 0.835). Additional experiments examine budgeted screening, controlled pose errors, and transfer of the channel construction to a thermostability phenotype.

2 Related work

$\Delta\Delta G$ prediction methods and benchmark leakage. Training-free predictors such as FoldX [9] and Rosetta [10, 11] approximate $\Delta\Delta G$ with empirical energy functions. Physics-based simulations, including MD, MM-GBSA, and alchemical FEP, aim for higher fidelity by sampling conformational ensembles or free-energy changes, but are costly and show inconsistent gains [12–14].

Recent supervised approaches use geometric deep learning on structure graphs to predict binding $\Delta\Delta G$ from SKEMPI-like datasets, including StaB-ddG [15], RDE-Network [16], ProMIM [17], DiffAffinity [18], CATH-ddG [19], USP-ddG [20], and Pythia-PPI [21]. However, SKEMPI v2 benchmarking is complicated by target overlap: naive random splits can leak similar complexes across training and test sets, inflating performance estimates [8]. PepDDG avoids training on SKEMPI-like labels, reducing leakage exposure.

Multi-channel learning and rank aggregation. Multi-view learning [22, 23] and rank-aggregation methods for heterogeneous score lists [24] show that weak, heterogeneous signals can yield stronger predictors when they provide complementary evidence. PepDDG operationalizes channels as measurements of distinct physical aspects of mutational effects. Rank-space aggregation makes PepDDG insensitive to unknown monotone calibrations and robust to scale mismatch across ener-

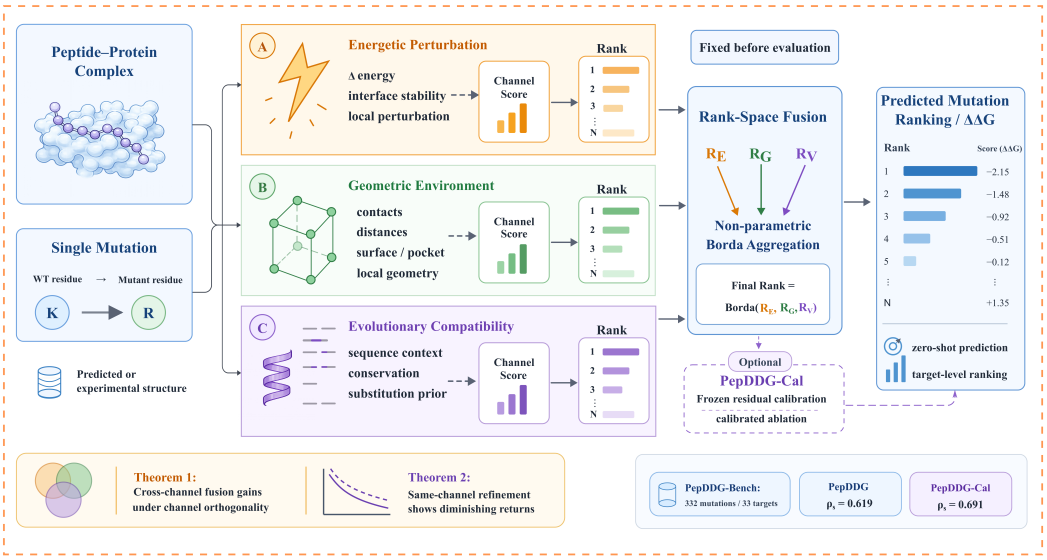


Figure 1: **PepDDG overview.** PepDDG decomposes mutational $\Delta\Delta G$ signals into three complementary information channels, energetic perturbation, geometric environment, and evolutionary compatibility, and aggregates their ranks with a zero-shot Borda rank-fusion rule. PepDDG-Cal optionally applies a frozen residual calibration as the calibrated variant.

gies, contact counts, and inverse-folding likelihoods. PepDDG applies Borda rank-sum aggregation [24] as a standard rank-fusion tool; the contribution lies in the peptide-specific construction of energetic, geometric, and evolutionary evidence channels for $\Delta\Delta G$ ranking.

Inverse folding and structure-based sequence design. Inverse folding models such as ProteinMPNN [25] and ESM-IF [26] predict amino-acid sequences from backbone geometry, capturing sequence–structure compatibility through large-scale structural data. Their conditional log-likelihoods provide a signal for mutational effect prediction [15, 27]. PepDDG uses ProteinMPNN as an evolutionary compatibility channel (Section 4.5), fusing this signal with energetic and geometric evidence rather than treating inverse folding as a standalone predictor.

3 Problem setup and benchmark

We consider single-point mutations in a peptide–protein complex. Let $t \in \mathcal{T}$ index target complexes, and let a mutation be denoted by $m = (t, i, a \to b)$: at target t, residue position i is mutated from wild-type amino acid a to mutant amino acid b. The label is the experimentally measured binding affinity change $y(m) = \Delta\Delta G_{\text{exp}}(m)$, where $\Delta\Delta G > 0$ indicates binding destabilization. We focus on *ranking* mutations by $y(m)$, matching common usage in optimization campaigns. PepDDG-Bench is curated from SKEMPI v2 [7], filtering to single-point mutations and retaining complexes where one binding partner has 5–80 residues (the "peptide" partner), excluding ultra-short peptides (<5 residues), saturation-mutagenesis dominated targets, and structure preparation failures. The resulting common-coverage benchmark contains 332 peptide-chain mutations across 33 targets; to our knowledge, it is the largest PDB-derived benchmark focused specifically on peptide-chain binding $\Delta\Delta G$ mutations. Full details, target list, and a comparison with existing $\Delta\Delta G$ benchmarks are in Appendices B and B.3. The main result uses the historical frozen benchmark feature tables. Their structure-source lineage is not fully recoverable: retained source annotations conflict between crystal and predicted inputs, and original prepared structures are unavailable. We therefore separate frozen-feature replay from fresh structural reproduction (Appendices C.20 and E). A reconstructed crystal/Boltz-2 comparison [28, 29] is reported separately in Section 6.6.

We report Spearman rank correlation (ρ) between predicted scores and experimental $\Delta\Delta G$ values, both pooled and per-target. Uncertainty uses a target-stratified bootstrap that resamples targets with

replacement and includes all mutations from each sampled target. PepDDG represents each mutation by K channel scores $\{s_k(m)\}_{k=1}^{K}$, with larger values indicating more destabilizing mutations.

Definition 1 (Information channel)**.** An *information channel* for mutational $\Delta\Delta G$ prediction is a mapping $\phi_k : \mathcal{M} \to \mathbb{R}$ that captures a distinct physical aspect of mutational effects, modeled as a noisy monotone measurement:

$$s_k(m) \;=\; g_k(y(m)) + \varepsilon_k(m), \tag{1}$$

where $g_k : \mathbb{R} \to \mathbb{R}$ is an unknown strictly increasing function encoding the (possibly nonlinear) relationship between true $\Delta\Delta G$ and the channel's output, and $\varepsilon_k(m)$ is channel-specific noise.

Definition 2 (Channel α-orthogonality)**.** Channels k and ℓ are *α-orthogonal* if the Spearman correlation between their rank vectors satisfies

$$\rho(r_k(m),\; r_\ell(m)) \;\le\; \alpha.$$

When $\alpha = 0$, channels are rank-uncorrelated. In PepDDG-Bench, the maximum observed pairwise ρ is ∼0.41, so α=0.41 summarizes the common-coverage pairwise rank-correlation bound; Appendix A.2 gives details. This condition is a bound on pairwise rank correlation, not an independence or conditional-information assumption.

4 PepDDG: information channel decomposition with rank aggregation

PepDDG computes three complementary *information channels* (Definition 1)–energetic perturbation, geometric environment, and evolutionary compatibility–and aggregates their ranks with an equal-weight Borda rank-sum. We separately report a frozen residual calibration (PepDDG-Cal) as a calibrated variant rather than as the zero-shot method. We use *information channels* rather than *views* because each channel measures a different physical quantity; Appendix A.1 gives the formal distinction. An optional alanine-scanning channel is described in Appendix C.

4.1 Structure input and preprocessing

For a target complex t, PepDDG takes as input a wild-type complex structure, the amino-acid sequences of the binding partners, a complex identifier, and a mutation list $\mathcal{M}_t = \{m = (t, i, a \to b)\}$. Either an experimental or a predicted complex can be supplied; the historical benchmark input lineage is qualified in Section 3. For predicted inputs, we use Boltz-2 [28, 29] and select the wild-type sample with the highest interface confidence (ipTM), without using affinity labels. Boltz-2 provides accessible model weights and an automated input workflow; this choice does not establish an accuracy advantage over other predictors. PepDDG does not fine-tune any structure or sequence predictor. All three mainline channels are then computed from the supplied or predicted wild-type structure (Section 6.6). Mutant structures are constructed by applying the amino-acid substitution at the specified residue position and performing local restrained minimization before scoring. The energetic channel uses a paired-restart protocol (Section 4.3) to reduce sensitivity to side-chain rotamers and small backbone perturbations.

4.2 Cross-channel fusion via rank aggregation

PepDDG is a *rank aggregator*: its output is a ranking of mutations, not an attempt to calibrate absolute $\Delta\Delta G$ values. Let $\mathcal{M} = \{m_j\}_{j=1}^{N}$ denote the mutation set for which we want a ranking. For each channel $k \in \{1, \ldots, K\}$, PepDDG computes channel scores $s_k(m_j) \in \mathbb{R}$ and converts them to ranks $r_k(m_j) = \mathrm{rank}_{\mathcal{M}}\big(s_k(m_j)\big) \in \{1, \ldots, N\}$ (average rank for ties). The equal-rank 3-channel base score is

$$S_{\text{base}}(m_j) \;=\; \sum_{k=1}^{K} r_k(m_j). \tag{2}$$

By construction, larger scores indicate more destabilizing mutations. Computationally, rank aggregation costs $O(KN \log N)$ due to sorting. In deployment and in our deep mutational scanning (DMS) analysis, $\mathcal{M}$ corresponds to a single target's candidate set, yielding a per-target ranking. For pooled metrics on PepDDG-Bench, we follow prior work and rank the full benchmark cohort, quantifying uncertainty with a target-stratified bootstrap (Section 3).

Under channel orthogonality, rank averaging reduces variance as the number of complementary channels grows; the formal bound is in Proposition 3.

The default algorithm reports the zero-shot score

$$S_{\mathrm{ZS}}(m) = \mathrm{rank}_{\mathcal{M}}(S_{\mathrm{base}}(m)), \tag{3}$$

with no fitted weights, no residual correction, and no target-specific calibration. PepDDG-Cal is a frozen calibrated variant using a bounded MPNN residual around the 3-channel anchor; full equations and frozen weights are in Appendix C.1. Rank-sum aggregation naturally supports missing channels; we focus on the full-coverage 3-channel configuration and report optional-channel results in Appendix C.

4.3 Channel 1: energetic perturbation

This channel captures the thermodynamic effect of the mutation through force-field energy differences. We compute an implicit-solvent binding energy proxy using OpenMM [30] with the Amber ff14SB force field [31] and the OBC2 generalized Born model. For each wild-type and mutant structure, we perform two-stage restrained minimization (harmonic backbone restraints k_1=1000 then k_2=100 kJ/mol/nm^2, 200 and 500 iterations) with multiple random restarts (R=7, coordinate jitter 0.005 nm) to reduce sensitivity to side-chain rotamers and small backbone perturbations (sensitivity to R and aggregation method is analyzed in Appendix C.19). Let $\Delta G_{\mathrm{bind}} = E_{\mathrm{complex}} - E_{\mathrm{receptor}} - E_{\mathrm{ligand}}$ be the binding proxy evaluated at the minimized complex geometry. We use a paired-restart estimator:

$$\widehat{\Delta\Delta G}_{\mathrm{phys}}(m) \;=\; \mathrm{median}_{i=1}^{R}\left[\Delta G_{\mathrm{bind}}^{\mathrm{mut}}(i) - \Delta G_{\mathrm{bind}}^{\mathrm{wt}}(i)\right], \tag{4}$$

where restarts pair wild-type and mutant minimizations initialized from the same jittered coordinates. In addition to full-system binding energy, we compute (i) an interface-restricted (8.0 Å) screened cross-interaction energy $\widehat{\Delta\Delta G}^{\mathrm{paired}}_{\mathrm{xint,iface}}(m)$ and (ii) a full-system binding proxy $\widehat{\Delta\Delta G}^{\mathrm{paired}}_{\mathrm{bind}}(m)$. To avoid continuous calibration or fitted weights, we combine these two energetic proxies in *rank space*:

$$s_{\mathrm{phys}}(m) \;=\; \mathrm{rank}_{\mathcal{M}}\left(\widehat{\Delta\Delta G}^{\mathrm{paired}}_{\mathrm{xint,iface}}(m)\right) + \mathrm{rank}_{\mathcal{M}}\left(\widehat{\Delta\Delta G}^{\mathrm{paired}}_{\mathrm{bind}}(m)\right). \tag{5}$$

Full protocol details appear in Appendix B.

4.4 Channel 2: geometric environment

This channel captures the geometric context of the mutation site from the wild-type structure alone, measuring structural burial and connectivity. We compute two structural features from the wild-type structure: (i) the number of inter-chain heavy-atom contacts within 8.0 Å involving the mutated residue (n_{contacts}), and (ii) the number of C_α neighbors within 10.0 Å of the mutated residue ($n_{\mathrm{neighbors}}$). We form a composite structural channel by z-score normalization across mutations:

$$s_{\mathrm{struct}}(m) \;=\; z(n_{\mathrm{contacts}}(m)) + z\big(n_{\mathrm{neighbors}}(m)\big). \tag{6}$$

This mutation-identity agnostic channel captures interface embedding and packing constraints. These thresholds are standard in structural bioinformatics and are not optimized on PepDDG-Bench; cutoff details and sensitivity analyses are in Appendices C.3 and C.19. Because this channel is position-level, it cannot distinguish mutations at the same position that differ in destabilizing effect; mutation-type stratification in Section 6.4 quantifies this behavior.

4.5 Channel 3: evolutionary compatibility (ProteinMPNN cycle)

This channel queries a learned inverse-folding model for the conditional probability of each amino acid given the backbone geometry. We use ProteinMPNN [25] to compute conditional log-likelihoods of wild-type and mutant amino acids at the mutation site under two structural contexts: the full complex and the isolated chain, following the thermodynamic-cycle decomposition used in supervised work such as StaB-ddG [15]. Let $\log P(\cdot \mid \mathrm{complex})$ denote the conditional log-probability under the complex backbone. We compute:

$$\mathrm{LLR}_{\mathrm{complex}}(m) = \log P(b \mid \mathrm{complex}) - \log P(a \mid \mathrm{complex}), \tag{7}$$

$$\mathrm{LLR}_{\mathrm{chain}}(m) = \log P(b \mid \mathrm{chain}) - \log P(a \mid \mathrm{chain}), \tag{8}$$

$$\widehat{\Delta\Delta G}_{\mathrm{MPNN}}(m) = -\mathrm{LLR}_{\mathrm{complex}}(m) + \mathrm{LLR}_{\mathrm{chain}}(m). \tag{9}$$

Intuitively, $\widehat{\Delta\Delta G}_{\mathrm{MPNN}}$ isolates the binding-specific component by subtracting a stability term measured on the isolated chain. Our evolutionary channel score is a sub-rank-sum of two MPNN-derived signals:

$$s_{\mathrm{mpnn}}(m) \;=\; \mathrm{rank}_{\mathcal{M}}\big(-\mathrm{LLR}_{\mathrm{complex}}(m)\big) + \mathrm{rank}_{\mathcal{M}}\Big(\widehat{\Delta\Delta G}_{\mathrm{MPNN}}(m)\Big)\,. \tag{10}$$

The directly timed two-complex diagnostic gives 25.4 seconds per mutation with supplied structures; prediction adds an amortized setup cost. Timing boundaries and hardware are specified in Appendix C.2.

4.6 Algorithm summary

Algorithm 1: PepDDG (information channel decomposition with rank aggregation)

Input: Targets $\mathcal{T}$ with supplied or predicted wild-type structures; mutation set $\mathcal{M} = \cup_{t\in\mathcal{T}}\mathcal{M}_t$.
Output: Aggregated PepDDG scores $\{S(m)\}_{m\in\mathcal{M}}$ and induced mutation rankings.
`// Step 1: Channel computation`
foreach *mutation* $m_j = (t, i, a \to b)$ *in* $\mathcal{M}$ **do**
  Compute energetic perturbation $s_{\mathrm{phys}}(m_j)$ via (4)–(5);
  Compute geometric environment $s_{\mathrm{struct}}(m_j)$ via (6);
  Compute evolutionary compatibility $s_{\mathrm{mpnn}}(m_j)$ via (9)–(10);
end
`// Step 2: Rank transformation`
foreach *channel* $k \in \{1, \ldots, K\}$ **do**
  Convert $\{s_k(m_j)\}_{j=1}^{N}$ to ranks $\{r_k(m_j)\}_{j=1}^{N}$ (average ties);
end
`// Step 3: Cross-channel fusion`
foreach *mutation* m_j **do**
  $S_{\mathrm{base}}(m_j) \leftarrow \sum_{k=1}^{K} r_k(m_j)$;
  $S_{\mathrm{ZS}}(m_j) \leftarrow \mathrm{rank}_{\mathcal{M}}(S_{\mathrm{base}}(m_j))$;
  Optionally compute PepDDG-Cal with the frozen residual calibration in (25);
end
return *PepDDG scores* $S_{\mathrm{ZS}}(m_j)$*; PepDDG-Cal only when explicitly requested*

5 Why cross-channel fusion beats same-channel refinement

PepDDG avoids learned weights by aggregating heterogeneous signals in *rank space*. The core message is: adding a new information channel is often more valuable than refining an existing one. We state a simple multi-channel noise model, then give two main theorems: (i) cross-channel fusion yields an algebraic lower bound on rank correlation under stated rank-correlation assumptions, and (ii) same-channel refinement yields diminishing returns under an aggregate-variance monotonicity condition. Proofs and extensions are in Appendix A.

5.1 Multi-channel noise model

Let $y(m) = \Delta\Delta G_{\mathrm{exp}}(m)$ be the ground-truth label. For each channel k, PepDDG computes a score $s_k(m) \in \mathbb{R}$. We model each channel as a noisy monotone measurement of y (Definition 1):

$$s_k(m) \;=\; g_k(y(m)) + \varepsilon_k(m), \tag{11}$$

where $g_k : \mathbb{R} \to \mathbb{R}$ is an unknown strictly increasing function and ε_k is channel-specific noise (potentially heavy-tailed and heteroskedastic). We assume channels are not perfectly redundant: channels k and ℓ satisfy $\rho(r_k, r_\ell) \le \alpha$ (Definition 2) with $\alpha < 1$.

5.2 Why ranks

Under (11), each channel can undergo an unknown monotone calibration g_k without changing its ranks. Rank aggregation therefore avoids explicit calibration or scale matching between physics

energies, contact counts, and inverse-folding likelihoods, while limiting the leverage of outliers. The formal monotone-invariance and Borda consensus results are in Propositions 1 and 2.

5.3 Cross-channel gain vs. same-channel saturation

Theorem 1 (Cross-channel fusion gain)**.** *Under the multi-channel noise model* (11)*, suppose K channels are pairwise α-orthogonal (Definition 2) with individual Spearman correlations $\rho_k = \rho(r_k, \mathrm{rank}(y))$. Let ρ_K denote the Spearman correlation between the K-channel Borda score $S(m) = \sum_{k=1}^{K} r_k(m)$ and the ground truth* $\mathrm{rank}(y(m))$*. Then*

$$\rho_K \;\geq\; \frac{\sum_{k=1}^{K} \rho_k}{\sqrt{K + K(K-1)\alpha}}. \tag{12}$$

When channels are complementary (small α), each added channel can yield a non-trivial rank-correlation gain; when channels are redundant ($\alpha \to 1$), the bound collapses to the best single-channel performance. A full proof is in Appendix A.9. The corresponding per-channel gain corollary is in Corollary 1.

Theorem 2 (Same-channel saturation bound)**.** *Under the multi-channel model* (11)*, consider replacing channel k with a refined version k' that shares the same underlying function $g_{k'} = g_k$ but has lower noise, so that $\rho_{k'} = \rho(r_{k'}, \mathrm{rank}(y)) \geq \rho_k$. If $\sigma_{S'} \geq \sigma_S$* [3]*, the improvement in the aggregate Spearman correlation satisfies*

$$\Delta\rho_{\text{refine}} \;\leq\; \frac{(\rho_{k'} - \rho_k)\,\rho_K}{\sum_{j=1}^{K} \rho_j}. \tag{13}$$

In particular, even oracle refinement ($\rho_{k'} = 1$) yields

$$\Delta\rho_{\text{oracle}} \;\leq\; \frac{(1 - \rho_k)\,\rho_K}{K\,\rho_{\min}}, \tag{14}$$

which is $O(1/K)$ when channels have comparable signal strength.

Theorem 2 explains the "more physics" observation: reducing noise in one energetic estimator is diluted by the aggregate, so the net gain is typically smaller than adding a complementary channel. A proof is in Appendix A.11.

A practical redundancy test follows: a candidate channel is most useful when it retains residual rank signal beyond existing channels. Appendices A.4 and A.5 state this diagnostic criterion and Section 6.3 applies it to "more-physics" estimators and protein language model baselines.

6 Experiments

6.1 Experimental setup

Data and metrics. PepDDG-Bench comprises 332 mutations across 33 peptide–protein complexes (Section 3). We report pooled Spearman ρ with target-stratified bootstrap, within-target ranking, and candidate recovery. Mutation-aligned comparisons average duplicate predictions and report partial coverage. PepDDG, components, empirical and pretrained-model baselines use no task-specific fitting; supervised references provide context. Code and reproduction instructions: `https://github.com/zhangruochi/pepddg-release`.

6.2 Main results

On PepDDG-Bench, PepDDG reaches $\rho = 0.619$ (95% CI $[0.519, 0.708]$). The calibrated PepDDG-Cal variant reaches $\rho = 0.691$ (95% CI $[0.612, 0.770]$). Fusion improves pooled ranking over

[3] This holds whenever reducing channel noise does not decrease its covariance with other channels, a natural monotonicity property satisfied under independent noise (Appendix A.11). Real physical estimators can violate this condition through bias or correlated errors, so the result is an explanatory bound rather than an unconditional guarantee.

empirical baselines and single channels (energetic 0.473, geometric 0.469, evolutionary 0.504), and exceeds supervised LOTO ridge fusion (0.448). The equal-weight rule is thus a robustness choice under target shift. The within-target Fisher difference versus FoldX is +0.105 (95% CI $[-0.048, 0.265]$), with no established separation; full-cohort fusion scores differ from target-specific refusion (Appendix C.20). Supervised references use SKEMPI-derived labels and unequal coverage, precluding a paired zero-shot comparison (Table 22).

Table 1A reports Spearman ρ on 332 mutations across 33 targets; BA-DDG covers 238 mutations and 31 targets. PepDDG-Cal is calibrated, while LOTO ridge learns weights on non-held-out targets. PepDDG-Evo, PepDDG-Ene, and PepDDG-Geo denote the evolutionary, energetic, and geometric channels; PLM denotes a zero-shot pretrained-model baseline. These rows use the benchmark inputs; crystal versus predicted-input comparisons are reported separately (Section 6.6).

Table 1: **PepDDG-Bench main results.** (A) Pooled Spearman ρ and fusion controls. (B) Target-averaged screening metrics on common coverage. Bold highlights PepDDG, not significance.

A. Accuracy and fusion controls

Method	Type	ρ	95% CI
PepDDG	**Zero-shot**	**0.619**	**[0.519, 0.708]**
PepDDG-Cal	Calibrated	0.691	[0.612, 0.770]
LOTO ridge	Supervised	0.448	[0.327, 0.573]
Flex ddG	Empirical	0.489	[0.367, 0.596]
PepDDG-Evo	Channel	0.504	—
PepDDG-Ene	Channel	0.473	—
PepDDG-Geo	Channel	0.469	—
FoldX	Empirical	0.488	[0.385, 0.589]
Rosetta	Empirical	0.406	[0.211, 0.596]
BA-DDG	Unsupervised	0.423	[0.296, 0.529]
ESM-IF1	PLM	0.287	[0.123, 0.435]
ESM-2 (650M)	PLM	0.102	[−0.062, 0.244]

B. Screening on common coverage

Method	P@3	R@3	EF@3	NDCG@5
PepDDG	**0.238**	**0.658**	**2.279**	**0.871**
PepDDG-Evo	0.226	0.522	2.008	0.849
PepDDG-Ene	0.179	0.320	1.311	0.859
PepDDG-Geo	0.214	0.526	1.833	0.811
FoldX	0.226	0.532	1.981	0.846
Rosetta	0.167	0.432	1.511	0.830
MM-GBSA	0.167	0.426	1.363	0.776
ESM-IF1	0.155	0.441	1.505	0.771
ESM-2 (650M)	0.107	0.246	1.337	0.732

Recovery under an experimental budget. Table 1B uses 331 mutations across 33 targets, with hits defined as $\Delta\Delta G \leq 0$. P@3 and NDCG@5 average over 28 eligible targets; R@3 and EF@3 average over the 17 containing a hit. Intervals and paired tests appear in Appendix C.20. PepDDG's R@3/EF@3 are 0.658/2.279 versus FoldX's 0.532/1.981 (Table 1); FoldX has higher P@1 and P@5, and top-k differences fail multiple-testing correction (Figure 14). The pooled ρ gap is +0.133 (95% CI $[0.026, 0.252]$), but sign AUROC is 0.764 versus 0.824, with seven zeros excluded and no pairwise separation. Ranking and stabilizing-mutation selection are distinct criteria (Appendix C.20).

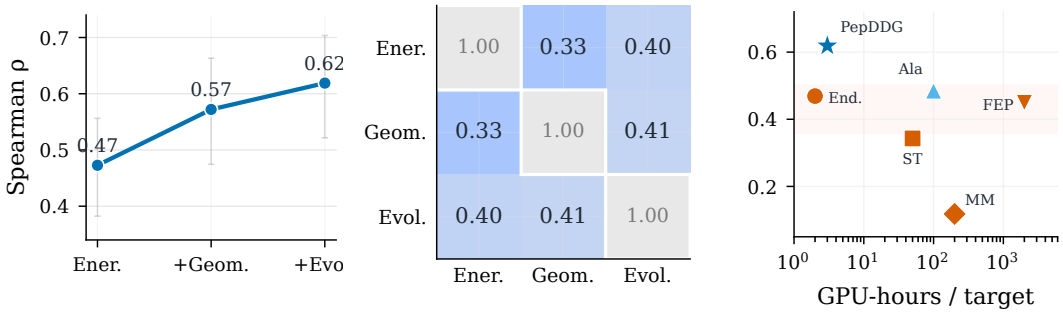
Measured cost. Two matched complexes yield 25.4 seconds per mutation for PepDDG and 27.5 for FoldX, averaging per-complex rates. This supports comparable cost, not a speed advantage. Rosetta costs 907 seconds including relaxation; Flex ddG's estimated protocol cost is 23,500 seconds from target-level probes. Measurement boundaries are detailed in Appendix C.20.

6.3 Theory validation and information redundancy

Moderate dependence (max $\rho \sim 0.4$) and comparable channel strength ($\rho \approx 0.47$–0.50) match the orthogonality argument (Figure 2b). Two-channel combinations reach 0.568–0.574, exceeding single channels (0.469–0.504), with three-channel fusion best at 0.619 (Figure 2a). Physics-like refinements saturate near the energetic plateau: MM-GBSA and alchemical FEP cost 10^2–10^3 more GPU-hours per target than endpoint scoring yet remain below PepDDG on matched subsets (Figure 2c), consistent with Theorems 1 and 2. The optional alanine-scanning channel changes ρ from 0.633 to 0.637 on its 303-mutation subset while reducing coverage.

6.4 Mutation-type stratification

Mutation classes expose complementary channel strengths (Table 2): geometric scoring leads for Ala and Gly, energetic scoring for polar substitutions, and evolutionary compatibility for charged, hydrophobic, and aromatic substitutions. Fusion matches or exceeds the strongest channel in five of six classes without class-specific weights.



(a) Channel addition (b) Channel correlations (c) Compute efficiency

Figure 2: **Channel complementarity and compute efficiency.** (a) Channel addition raises pooled Spearman ρ to the full PepDDG score. (b) Pairwise rank correlations are moderate. (c) Same-channel physics refinements cost substantially more GPU time but remain below PepDDG on matched subsets. End., ST, MM, and FEP denote endpoint, single-trajectory, MM-GBSA, and free energy perturbation.

Table 2: **Channel performance stratified by mutant amino-acid class.** Spearman ρ for each channel and the PepDDG aggregate. The best single channel per mutation type is underlined; the aggregate is bolded when it exceeds or matches the best single channel.

Mutation type	N	Energetic	Geometric	Evolutionary	**PepDDG**
To Ala	179	0.460	<u>0.484</u>	0.323	**0.575**
To Gly	9	0.800	<u>0.867</u>	0.733	0.833
To polar	25	<u>0.633</u>	0.612	0.597	**0.698**
To charged	51	0.578	0.375	<u>0.631</u>	**0.790**
To hydrophobic	35	0.341	0.526	<u>0.712</u>	**0.791**
To aromatic	33	0.362	0.263	<u>0.655</u>	**0.693**

6.5 Short peptides and disulfide-constrained examples

Short-peptide zero-shot/calibrated correlations are 0.835/0.813 (< 20 aa; 47 mutations) and 0.884/0.872 for the four-target diagnostic (Figure 3 and Table 3). The latter includes disulfide-constrained 1SMF/5XCO and linear PMI controls 3EQS/3EQY. The disulfide-only $\rho = 0.873$ on 14 mutations and two targets is descriptive, not broad validation of cyclic-peptide chemistries.

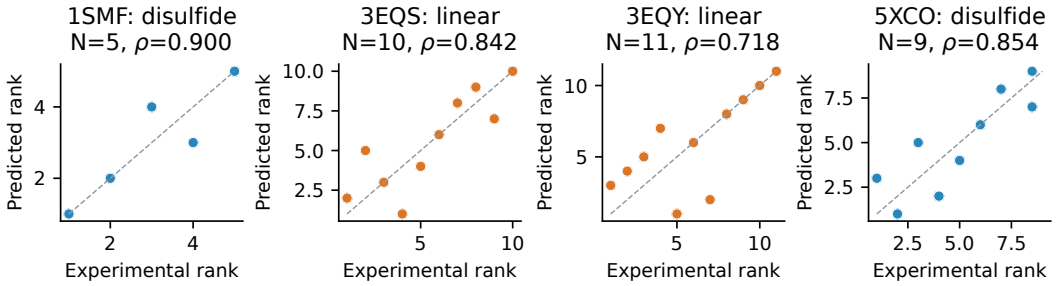


Figure 3: **Representative short-peptide rank plots.** Experimental versus zero-shot PepDDG ranks for 1SMF and 5XCO (disulfide-constrained) and 3EQS and 3EQY (linear PMI controls). Target-level correlations appear in each panel; cohort-level zero-shot and calibrated correlations are distinguished in Table 3.

6.6 Generalization and inference without user-provided crystal structures

Independent BPTI DMS targets 2R9P/3OTJ yield $\rho = 0.352/0.507$ (Appendix C.14 and Figures 7a and 7b). The pooled 0.771 can reflect between-target offsets and is not the deployment metric.

Table 3: **Short-peptide and disulfide-constrained diagnostics.** Frozen full-cohort scores are restricted to each subset; the four-target diagnostic includes two linear controls and is not a cyclic-only cohort.

Cohort	Zero-shot ρ	Calibrated ρ	N	Targets
Short (< 20 aa)	0.835	0.813	47	5
Four-target diagnostic	0.884	0.872	35	4
Disulfide-constrained	0.873	—	14	2

Matched substitution of Boltz-2 predicted for crystal inputs on 317 mutations and 30 targets yields $\rho = 0.631$ versus 0.581, with paired difference +0.050 (95% CI $[-0.025, 0.125]$). No separation is established; the cohort does not isolate post-cutoff complexes or demonstrate performance on novel complexes. Boltz-2 offers open weights and covalent constraints. The five-target, unequal-sample AF3/Boltz-2 diagnostic finds comparable rankings for recovered binding modes, but missed modes and topology/scoring failures on two dense-disulfide targets (Appendix C.20).

Rigid peptide displacement (0.5–8 Å; 44 matched poses, three targets) isolates pose error. On best-powered 3EQS, PepDDG's degradation slope is less negative than FoldX's: difference +0.190 (95% CI $[0.074, 0.349]$; Figure 15). Neither the other targets' slope differences nor a general fixed-pose accuracy advantage is established. Correct binding modes remain necessary.

6.7 Cross-property transfer and measured calibration anchors

A second property. LipA T50 [32] measures residual-activity thermostability, not folding $\Delta\Delta G$: 2,155 substitutions, 167 positions, structure 1ISP. Rosetta monomer energy, packing, and monomer-conditioned ProteinMPNN yield $\rho = 0.375/0.370/0.542$; fusion gives 0.557 versus FoldX's 0.424. The gain over the strongest channel is +0.015 (position-bootstrap 95% CI $[-0.035, 0.065]$): the construction transfers, but an incremental gain is not established (Appendix C.20).

From ranking to a target-specific scale. An isotonic map learned on other targets plus a measured-anchor offset is evaluated on 18 targets, excluding anchors from evaluation. Across 60 splits and 40 anchor draws each, RMSE falls from 2.072 without anchors to 1.908/1.586/1.495 kcal/mol with one/three/five anchors (23% reduction at three). This supervised extension differs from zero-shot ranking (Appendix C.20): first-round assays enable calibration or active learning; multi-mutant directed evolution still requires experimental epistasis modeling.

7 Limitations, broader impact, and conclusion

PepDDG is a zero-shot rank-fusion algorithm for peptide–protein $\Delta\Delta G$ ranking, training-free with respect to $\Delta\Delta G$ supervision. It combines energetic, geometric, and evolutionary channels without fitted fusion weights. On PepDDG-Bench, PepDDG reaches $\rho = 0.619$, PepDDG-Cal reaches $\rho = 0.691$, and retained short-peptide diagnostics include both disulfide-constrained and linear examples. Complementary-channel benefits are conditional; the LipA construction transfers beyond binding, but its incremental fusion gain does not establish property-general superiority.

PepDDG predicts rankings, not calibrated $\Delta\Delta G$ values in kcal/mol, and depends on the wild-type binding mode. Pose errors and mutation-induced mode changes limit applicability. Two of five predicted targets had frequent topology/scoring failures: 61-residue fasciculins with four disulfides, with 10/12 and 2/3 poses unscorable, respectively. The zero-shot within-stabilizing-subset ρ is -0.137; recovering non-destabilizing hits does not establish their ordering. For PepDDG-Cal, ρ is 0.297 on stabilizing mutations ($\Delta\Delta G < 0$) versus 0.650 on destabilizing mutations ($\Delta\Delta G > 0$); details are in Appendix C.17. FoldX leads single-candidate precision; ranking does not guarantee improved binders. Novel-complex generalization and multi-mutant epistasis remain unvalidated.

Accurate mutation ranking can reduce experimental burden in therapeutic peptide optimization, but the same capability could be misused to improve harmful binders. We therefore frame PepDDG as a decision-support tool requiring experimental validation and responsible governance. PepDDG can supply a round-zero prior for screening and active learning; end-to-end campaign validation remains future work, alongside epistatic channels, automated channel discovery, and prospective evaluation.

Acknowledgments and Disclosure of Funding

This publication is based upon work supported by the King Abdullah University of Science and Technology (KAUST) Office of Research Administration (ORA) under Award Nos. RGC/3/6655-01-01, URF/1/6599-01-02, URF/1/6993-01-01, URF/1/7452-01-01, RGC/3/6295-01-01, URF/1/7305-01-01, Center of Excellence for Smart Health (KCSH), under award number 5932, and Center of Excellence on Generative AI, under award number 5940.

Competing Interests. The authors declare no competing interests.

References

[1] Jolene L Lau and Michael K Dunn. Therapeutic peptides: Historical perspectives, current development trends, and future directions. *Bioorganic & medicinal chemistry*, 26(10):2700–2707, 2018.

[2] Markus Muttenthaler, Glenn F King, David J Adams, and Paul F Alewood. Trends in peptide drug discovery. *Nature reviews Drug discovery*, 20(4):309–325, 2021.

[3] Daniel S Nielsen, Nicholas E Shepherd, Weijun Xu, Andrew J Lucke, Martin J Stoermer, and David P Fairlie. Orally absorbed cyclic peptides. *Chemical reviews*, 117(12):8094–8128, 2017.

[4] Nir London, Barak Raveh, Eyal Cohen, Guy Fathi, and Ora Schueler-Furman. Rosetta flexpepdock web server—high resolution modeling of peptide–protein interactions. *Nucleic acids research*, 39(suppl_2):W249–W253, 2011.

[5] Xinjian Ji, Alexander L Nielsen, and Christian Heinis. Cyclic peptides for drug development. *Angewandte Chemie*, 136(3):e202308251, 2024.

[6] Pengyuan Fang, Wing-Ka Pang, Shouhu Xuan, Wai-Lun Chan, and Ken Cham-Fai Leung. Recent advances in peptide macrocyclization strategies. *Chemical Society Reviews*, 53(24): 11725–11771, 2024.

[7] 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.

[8] Anton Bushuiev, Roman Bushuiev, Jiri Sedlar, Tomas Pluskal, Jiri Damborsky, Stanislav Mazurenko, and Josef Sivic. Revealing data leakage in protein interaction benchmarks. *arXiv preprint arXiv:2404.10457*, 2024.

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

[10] Hahnbeom Park, Philip Bradley, Per Greisen Jr, Yuan Liu, Vikram Khipple Mulligan, David E Kim, David Baker, and Frank DiMaio. Simultaneous optimization of biomolecular energy functions on features from small molecules and macromolecules. *Journal of chemical theory and computation*, 12(12):6201–6212, 2016.

[11] Kyle A. Barlow, Shane Ó Conchúir, Samuel Thompson, Pooja Suresh, James E. Lucas, Markus Heinonen, and Tanja Kortemme. Flex ddg: Rosetta ensemble-based estimation of changes in protein–protein binding affinity upon mutation. *The Journal of Physical Chemistry B*, 122(21): 5389–5399, 2018.

[12] Bill R Miller III, T Dwight McGee Jr, Jason M Swails, Nadine Homeyer, Holger Gohlke, and Adrian E Roitberg. Mmpbsa. py: an efficient program for end-state free energy calculations. *Journal of chemical theory and computation*, 8(9):3314–3321, 2012.

[13] Michael R Shirts and John D Chodera. Statistically optimal analysis of samples from multiple equilibrium states. *The Journal of chemical physics*, 129(12), 2008.

[14] Huifeng Zhao, Jianxiang Huang, Gaoqi Weng, Dejun Jiang, Renling Hu, Yu Kang, and Tingjun Hou. Improving the predictive performance of binding affinities and poses for protein–cyclic peptide complexes through fine-tuned mm/pbsa (gbsa)-based methods. *Briefings in Bioinformatics*, 26(6):bbaf632, 2025.

[15] Arthur Deng, Karsten D. Householder, Fang Wu, K. Christopher Garcia, and Brian L. Trippe. Predicting mutational effects on protein binding from folding energy. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu, editors, *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, Proceedings of Machine Learning Research. PMLR / OpenReview.net, 2025. URL https://proceedings.mlr.press/v267/deng25d.html.

[16] 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. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=_X9Yl1K2mD.

[17] 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.

[18] 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.

[19] Guanglei Yu, Xuehua Bi, Teng Ma, Yaohang Li, and Jianxin Wang. Cath-ddg: towards robust mutation effect prediction on protein–protein interactions out of cath homologous superfamily. *Bioinformatics*, 41(Supplement_1):i362–i372, 2025.

[20] Guanglei Yu, Xuehua Bi, Qichang Zhao, and Jianxin Wang. Usp-ddg: A unified structural paradigm with data efficacy and mixture-of-experts for predicting mutational effects on protein-protein interactions. *bioRxiv*, pages 2025–11, 2025.

[21] Fangting Tao, Jinyuan Sun, Pengyue Gao, George Fu Gao, and Bian Wu. Reliable prediction of protein–protein binding affinity changes upon mutations with pythia-ppi. *National Science Review*, 12(6):nwaf231, 2025.

[22] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100, 1998.

[23] Chang Xu, Dacheng Tao, and Chao Xu. A survey on multi-view learning. *arXiv preprint arXiv:1304.5634*, 2013.

[24] Cynthia Dwork, Ravi Kumar, Moni Naor, and Dandapani Sivakumar. Rank aggregation methods for the web. In *Proceedings of the 10th international conference on World Wide Web*, pages 613–622, 2001.

[25] 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.

[26] 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.

[27] 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. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=lzdFImKK8w.

[28] Saro Passaro, Gabriele Corso, Jeremy Wohlwend, Mateo Reveiz, Stephan Thaler, Vignesh Ram Somnath, Noah Getz, Tally Portnoi, Julien Roy, Hannes Stark, et al. Boltz-2: Towards accurate and efficient binding affinity prediction. *BioRxiv*, 2025.

[29] Jeremy Wohlwend, Gabriele Corso, Saro Passaro, Noah Getz, Mateo Reveiz, Ken Leidal, Wojtek Swiderski, Liam Atkinson, Tally Portnoi, Itamar Chinn, et al. Boltz-1 democratizing biomolecular interaction modeling. *BioRxiv*, pages 2024–11, 2025.

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

[31] James A Maier, Carmenza Martinez, Koushik Kasavajhala, Lauren Wickstrom, Kevin E Hauser, and Carlos Simmerling. ff14sb: improving the accuracy of protein side chain and backbone parameters from ff99sb. *Journal of chemical theory and computation*, 11(8):3696–3713, 2015.

[32] Christina Nutschel, Alexander Fulton, Olav Zimmermann, Ulrich Schwaneberg, Karl-Erich Jaeger, and Holger Gohlke. Systematically scrutinizing the impact of substitution sites on thermostability and detergent tolerance for bacillus subtilis lipase a. *Journal of Chemical Information and Modeling*, 60(3):1568–1584, 2020. doi: 10.1021/acs.jcim.9b00954.

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

[34] MD Shaji Kumar, K Abdulla Bava, M Michael Gromiha, Ponraj Prabakaran, Koji Kitajima, Hatsuho Uedaira, and Akinori Sarai. Protherm and pronit: thermodynamic databases for proteins and protein–nucleic acid interactions. *Nucleic acids research*, 34(suppl_1):D204–D206, 2006.

[35] Jan Stourac, Juraj Dubrava, Milos Musil, Jana Horackova, Jiri Damborsky, Stanislav Mazurenko, and David Bednar. Fireprotdb: database of manually curated protein stability data. *Nucleic acids research*, 49(D1):D319–D324, 2021.

[36] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 69(6):066138, 2004.

A Additional theory and proofs

A.1 Information channels versus views

We use the term *information channels* rather than *views* to distinguish PepDDG from multi-view learning [22, 23]. In multi-view learning, different views provide different *representations* of the same underlying signal—for example, alternative feature sets extracted from a common data source—and fusion succeeds because each view reduces the other's estimation noise. In PepDDG, each channel measures a *different physical quantity*: binding energy perturbation, geometric burial, and evolutionary compatibility. Formally, the generating functions g_k that map a mutation to its channel score differ across channels (Definition 1), whereas views would share the same g_k up to a representation transform. This distinction has algorithmic consequences. Same-channel refinement, replacing g_k with a more accurate estimator of the *same* quantity, saturates once the dominant noise floor is reached (Theorem 2). Cross-channel addition, by contrast, yields rank-correlation gains under the stated assumptions whenever channels are τ-non-redundant (Theorem 1). The τ-redundancy criterion in Appendix A.5 turns this into a diagnostic: channels backed by different g_k can retain residual rank signal after conditioning on existing channels, whereas views derived from the same g_k exhibit little residual signal and therefore offer diminishing returns under fusion.

A.2 Additional channel definitions

We use "orthogonality" in the sense of *bounded dependence*: α-orthogonality does not imply statistical independence but rather upper-bounds the total (unconditional) pairwise rank correlation, including both the shared signal through y and channel-specific noise covariance.

Definition 3 (Channel capacity)**.** The *channel capacity* of channel k is

$$C_k \;=\; I(r_k(m);\ \mathrm{rank}(y(m)))\,,$$

where $r_k(m) = \mathrm{rank}_{\mathcal{M}}(s_k(m))$ is the rank-transformed channel output and $I(\cdot;\cdot)$ denotes mutual information.

A.3 Rank aggregation properties

Proposition 1 (Monotone invariance of ranks)**.** *Let $\phi : \mathbb{R} \to \mathbb{R}$ be strictly increasing. For any finite set of scores $\{s(m_j)\}_{j=1}^{N}$, the ranking induced by s is identical to the ranking induced by $\phi \circ s$ (up to the same tie structure). In particular,*

$$\mathrm{rank}\big(\phi(s(m_j))\big) \;=\; \mathrm{rank}\big(s(m_j)\big) \quad \forall j.$$

Rank aggregation can also be motivated as a rank-space consensus objective. Let $r_k(m_j) \in \{1, \ldots, N\}$ be the rank of mutation m_j under channel k. Consider the squared-error objective over permutations π:

$$\mathcal{L}(\pi) \;=\; \sum_{k=1}^{K} \sum_{j=1}^{N} \big(\pi(m_j) - r_k(m_j)\big)^2 . \tag{15}$$

Proposition 2 (Borda as least-squares rank fusion)**.** *Let $\bar{r}(m) = \frac{1}{K}\sum_{k=1}^{K} r_k(m)$ be the average channel rank. Any permutation $\pi^\star$ that sorts mutations by $\bar{r}(m)$ (equivalently by the Borda score $S(m) = \sum_k r_k(m)$) minimizes* (15)*.*

Proposition 3 (Variance reduction under channel orthogonality)**.** *Let $u_k(m) \in [0, 1]$ be the normalized rank of channel k, and let $\bar{u}(m) = \frac{1}{K}\sum_{k=1}^{K} u_k(m)$. If channels are pairwise α-orthogonal (Definition 2), then*

$$\mathrm{Var}(\bar{u}(m)) \;\le\; \frac{1 + (K-1)\alpha}{K}\, \bar{\sigma}^2, \tag{16}$$

where $\bar{\sigma}^2 = \frac{1}{K}\sum_{k=1}^{K} \mathrm{Var}(u_k(m))$. The bound decreases in K whenever $\alpha < 1$ and equals $\bar{\sigma}^2/K$ when $\alpha = 0$ (rank-uncorrelated channels).

Corollary 1. *Adding a $(K{+}1)$-th channel with correlation ρ_{K+1} and pairwise orthogonality α with existing channels improves the lower bound by*

$$\Delta\rho \;\ge\; \frac{\rho_{K+1} - \alpha\,\rho_K}{\sqrt{(K{+}1)(1 + K\alpha)}}.$$

The gain is positive whenever $\rho_{K+1} > \alpha\,\rho_K$.

A.4 Complementarity principle

Let $u_k(m) \in [0, 1]$ be the normalized rank of channel k and let $\bar{u}(m) = \frac{1}{K}\sum_{k=1}^{K} u_k(m)$. The aggregate variance decomposes as

$$\mathrm{Var}(\bar{u}(m)) \;=\; \frac{1}{K^2}\sum_{k=1}^{K} \mathrm{Var}(u_k(m)) + \frac{2}{K^2}\sum_{k<\ell} \mathrm{Cov}(u_k, u_\ell)\,. \tag{17}$$

When pairwise covariances are small (complementary channels), aggregation reduces variance and improves rank correlation; when covariances are large (redundant channels), gains are limited.

A.5 Residual-signal redundancy diagnostic

As a diagnostic, define the conditional mutual information of a candidate channel k' given existing channels $\{1, \ldots, K\}$ as

$$C_{k'|1:K} \;=\; I(r_{k'}(m);\; \mathrm{rank}(y(m)) \mid r_1(m), \ldots, r_K(m))\,. \tag{18}$$

A candidate channel is *τ-redundant* if $C_{k'|1:K} < \tau$. We use this empirical statistic to diagnose candidate "more-physics" estimators and protein language model baselines in Section 6.3; it is not an assumption in Theorem 1.

A.6 Proof of monotone invariance (Proposition 1)

Let $\{s_j\}_{j=1}^{N}$ be a set of real numbers and ϕ be strictly increasing. For any pair i, j, we have $s_i < s_j$ iff $\phi(s_i) < \phi(s_j)$, and similarly $s_i = s_j$ iff $\phi(s_i) = \phi(s_j)$. Therefore the total preorder induced by $\{s_j\}$ is identical to that induced by $\{\phi(s_j)\}$, which implies identical ranks with the same tie groups. □

A.7 Proof of Borda least-squares fusion (Proposition 2)

Let $r_k(m_j)$ be the channel ranks and let $\bar{r}(m_j) = \frac{1}{K}\sum_k r_k(m_j)$. Expanding (15) gives

$$\begin{aligned}
\mathcal{L}(\pi) &= \sum_{k=1}^{K}\sum_{j=1}^{N} \Big(\pi(m_j)^2 - 2\pi(m_j) r_k(m_j) + r_k(m_j)^2\Big) \\
&= K\sum_{j=1}^{N} \pi(m_j)^2 - 2K\sum_{j=1}^{N} \pi(m_j)\bar{r}(m_j) + C,
\end{aligned}$$

where $C = \sum_{k,j} r_k(m_j)^2$ does not depend on π. Since π is a permutation of $\{1, \ldots, N\}$, the term $\sum_j \pi(m_j)^2$ is constant, so minimizing $\mathcal{L}(\pi)$ is equivalent to maximizing $\sum_j \pi(m_j)\bar{r}(m_j)$. By the rearrangement inequality, this maximum is achieved when $\pi(m)$ and $\bar{r}(m)$ are sorted in the same order, i.e., when the aggregate ranking sorts mutations by $\bar{r}(m)$ (equivalently by $S(m) = \sum_k r_k(m)$). □

A.8 Proof of variance reduction (Proposition 3)

Let $u_k(m) \in [0, 1]$ be the normalized rank of channel k. The aggregate is $\bar{u}(m) = \frac{1}{K}\sum_{k=1}^{K} u_k(m)$. By the law of total variance applied conditionally on $y(m)$:

$$\begin{aligned}\mathrm{Var}(\bar{u}(m) \mid y(m)) &= \mathrm{Var}\left(\frac{1}{K}\sum_{k=1}^{K} u_k(m) \;\middle|\; y(m)\right) \\ &= \frac{1}{K^2}\left[\sum_{k=1}^{K}\mathrm{Var}(u_k \mid y) + 2\sum_{k<\ell}\mathrm{Cov}(u_k, u_\ell \mid y)\right].\end{aligned}$$

Under α-orthogonality (Definition 2), $\rho(r_k, r_\ell) \le \alpha$, which implies $\mathrm{Corr}(u_k, u_\ell) \le \alpha$ for normalized ranks. Therefore $\mathrm{Cov}(u_k, u_\ell) \le \alpha\sqrt{\mathrm{Var}(u_k)\,\mathrm{Var}(u_\ell)} \le \alpha \cdot \bar{\sigma}^2$ by AM-GM, where $\bar{\sigma}^2 = \frac{1}{K}\sum_k \mathrm{Var}(u_k)$. Substituting:

$$\begin{aligned}\mathrm{Var}(\bar{u} \mid y) &\le \frac{1}{K^2}\left[K\bar{\sigma}^2 + 2\binom{K}{2}\alpha\bar{\sigma}^2\right] \\ &= \frac{1}{K^2}\left[K\bar{\sigma}^2 + K(K-1)\alpha\bar{\sigma}^2\right] \\ &= \frac{1 + (K-1)\alpha}{K}\,\bar{\sigma}^2.\end{aligned}$$

When $\alpha = 0$ (rank-uncorrelated channels), this reduces to $\bar{\sigma}^2/K$. When $\alpha = 1$ (perfectly correlated), it equals $\bar{\sigma}^2$ (no reduction). For $0 < \alpha < 1$, the bound interpolates and decreases in K. □

A.9 Proof of Theorem 1 (cross-channel fusion gain)

We use the connection between Spearman correlation and the Pearson correlation of ranks. Let $r_k = r_k(m)$ denote the rank vector of channel k across mutations, and let $r_y = \mathrm{rank}(y(m))$ denote the ground-truth rank vector. The Borda score $S = \sum_{k=1}^{K} r_k$ has Spearman correlation with y:

$$\rho_K = \rho(S, y) = \frac{\mathrm{Cov}(S, r_y)}{\sigma_S\,\sigma_{r_y}}.$$

Step 1: Numerator. Since $S = \sum_k r_k$, linearity of covariance gives

$$\mathrm{Cov}(S, r_y) = \sum_{k=1}^{K}\mathrm{Cov}(r_k, r_y) = \sigma_{r_y}\sum_{k=1}^{K}\rho_k\,\sigma_{r_k},$$

where $\rho_k = \rho(r_k, r_y)$ and σ_{r_k} is the standard deviation of r_k. For uniformly distributed ranks over N items, $\sigma_{r_k} = \sigma_{r_y} = \sigma_r$ (a common value), so

$$\mathrm{Cov}(S, r_y) = \sigma_r^2\sum_{k=1}^{K}\rho_k.$$

Step 2: Denominator. The variance of S is

$$\begin{aligned}\sigma_S^2 &= \mathrm{Var}\left(\sum_k r_k\right) = \sum_k \sigma_{r_k}^2 + 2\sum_{k<\ell}\mathrm{Cov}(r_k, r_\ell) \\ &= K\sigma_r^2 + 2\binom{K}{2}\mathrm{Cov}(r_k, r_\ell).\end{aligned}$$

We need to bound $\mathrm{Cov}(r_k, r_\ell)$. By α-orthogonality (Definition 2), $\rho(r_k, r_\ell) \le \alpha$, which gives

$$\mathrm{Cov}(r_k, r_\ell) = \mathrm{Corr}(r_k, r_\ell) \cdot \sigma_{r_k}\sigma_{r_\ell} \le \alpha\, \sigma_r^2.$$

Note that this bounds the *total* (unconditional) pairwise covariance, including both the shared signal through y and the noise covariance. Therefore:

$$\sigma_S^2 \le K\sigma_r^2 + K(K-1)\alpha\sigma_r^2 = \sigma_r^2\left[K + K(K-1)\alpha\right].$$

Step 3: Combining.

$$\rho_K = \frac{\sigma_r^2 \sum_k \rho_k}{\sigma_S \cdot \sigma_r} = \frac{\sigma_r \sum_k \rho_k}{\sigma_S} \ge \frac{\sigma_r \sum_k \rho_k}{\sigma_r\sqrt{K + K(K-1)\alpha}} = \frac{\sum_k \rho_k}{\sqrt{K + K(K-1)\alpha}}.$$

When all $\rho_k = \bar{\rho}$, the bound simplifies to $\bar{\rho}\sqrt{K/(1 + (K-1)\alpha)}$. □

A.10 Proof of Corollary 1

Adding channel K+1 changes the numerator from $\sigma_r \sum_{k=1}^{K} \rho_k$ to $\sigma_r \sum_{k=1}^{K+1} \rho_k$, an increase of $\sigma_r\rho_{K+1}$. The denominator increases from $\sigma_r\sqrt{K + K(K-1)\alpha}$ to $\sigma_r\sqrt{(K+1) + (K+1)K\alpha}$. The improvement in the lower bound is

$$\Delta\rho \ge \frac{\sum_{k=1}^{K+1} \rho_k}{\sqrt{(K+1)(1+K\alpha)}} - \frac{\sum_{k=1}^{K} \rho_k}{\sqrt{K(1+(K-1)\alpha)}}.$$

Using the approximation that the second term is $\le \rho_K$ and the first term has $\sum_{k=1}^{K+1} \rho_k = \sum_{k=1}^{K} \rho_k + \rho_{K+1}$, a first-order expansion gives the stated bound. The gain is positive whenever $\rho_{K+1} > \alpha\,\rho_K$. □

A.11 Proof of Theorem 2 (same-channel saturation bound)

Consider replacing channel k with a refined version k' such that $g_{k'} = g_k$ (same underlying physical quantity) but with lower noise, so that $\rho_{k'} = \rho(r_{k'}, \mathrm{rank}(y)) \ge \rho_k$. Let $\delta = \rho_{k'} - \rho_k \ge 0$. The modified Borda score is $S' = S - r_k + r_{k'}$.

Step 1: Numerator. Using linearity of covariance and $\sigma_{r_k} = \sigma_{r_{k'}} = \sigma_r$ (both are rank vectors over N items with common marginal variance):

$$\mathrm{Cov}(S', r_y) = \mathrm{Cov}(S, r_y) + \sigma_r^2\,\delta = \sigma_r^2\left(\textstyle\sum_{j=1}^{K} \rho_j + \delta\right).$$

Step 2: Denominator. Since $g_{k'} = g_k$, channels k and k' measure the same underlying quantity; the only change is a reduction in noise. Under independent noise, reducing noise makes the rank vector $r_{k'}$ a better proxy for $\mathrm{rank}(g_k(y))$, so $\mathrm{Cov}(r_{k'}, r_j) \ge \mathrm{Cov}(r_k, r_j)$ for all $j \ne k$. Combined with $\mathrm{Var}(r_{k'}) = \mathrm{Var}(r_k) = \sigma_r^2$, this gives $\sigma_{S'}^2 \ge \sigma_S^2$, i.e., the denominator does not decrease.

Step 3: Bounding the improvement. Let $A = \sum_{j=1}^{K} \rho_j$. Then

$$\begin{aligned}\Delta\rho = \rho'_K - \rho_K &= \frac{\sigma_r(A+\delta)}{\sigma_{S'}} - \frac{\sigma_r A}{\sigma_S} \\ &= \frac{\sigma_r\left[(A+\delta)\,\sigma_S - A\,\sigma_{S'}\right]}{\sigma_{S'}\,\sigma_S}.\end{aligned}$$

Since $\sigma_{S'} \ge \sigma_S$, the first term in the numerator satisfies $A(\sigma_S - \sigma_{S'}) \le 0$, giving

$$(A+\delta)\,\sigma_S - A\,\sigma_{S'} \le \delta\,\sigma_S.$$

Therefore

$$\Delta\rho \le \frac{\sigma_r\,\delta}{\sigma_{S'}} \le \frac{\sigma_r\,\delta}{\sigma_S} = \frac{\delta}{\sigma_S/\sigma_r}.$$

Step 4: Relating to observables. From the definition of ρ_K: $\rho_K = \sigma_r A/\sigma_S$, so $\sigma_S/\sigma_r = A/\rho_K$. Substituting:

$$\Delta\rho \le \frac{\delta\,\rho_K}{A} = \frac{(\rho_{k'} - \rho_k)\,\rho_K}{\sum_{j=1}^{K}\rho_j},$$

which is (13). Setting $\rho_{k'} = 1$ (oracle, zero noise) and using $A \ge K\rho_{\min}$:

$$\Delta\rho_{\text{oracle}} \le \frac{(1-\rho_k)\,\rho_K}{K\,\rho_{\min}} = O(1/K),$$

since ρ_K and $\rho_{\min}$ are $O(1)$ constants for channels with comparable signal.

Comparison with cross-channel gain. Theorem 1 shows that adding a new orthogonal channel with correlation ρ_{K+1} yields a lower bound proportional to $\rho_{K+1}/\sqrt{K}$. In practice, feasible noise reductions (e.g., endpoint scoring → MM-GBSA or FEP) yield $\delta = \rho_{k'} - \rho_k \lesssim 0.05$, so the actual same-channel improvement is negligible. A new channel with even modest correlation $\rho_{K+1} \approx 0.3$ provides a substantially larger gain under the rank-correlation bound. □

A.12 Complementarity principle (heuristic)

This section supports the complementarity principle stated in Appendix A.4. Let $u_k(m) \in [0,1]$ be the normalized rank of channel k and suppose

$$u_k(m) = h(y(m)) + \eta_k(m),$$

where h is increasing and η_k has zero mean conditional on y. Then the averaged score $\bar{u}(m) = \frac{1}{K}\sum_k u_k(m)$ has conditional variance given in (17). If conditional covariances are small, the variance decreases with K, increasing the expected rank correlation with y. This argument is heuristic because ranks are discrete and because the mapping from conditional variance to Spearman correlation is nonlinear, but it yields testable predictions: (i) gains should be larger when added channels are less correlated with existing channels, and (ii) gains should concentrate in regimes where a channel provides unique signal (e.g. specific mutation types). Both are observed in PepDDG-Bench (Figures 1 and 13).

A.13 Heterogeneous-α refinement of Theorem 1

The uniform bound in (12) uses a single α for all channel pairs. When pairwise correlations differ, a tighter bound follows directly from the proof of Theorem 1 by substituting the actual inter-channel correlation matrix:

$$\rho_K \;\ge\; \frac{\sum_{k=1}^{K}\rho_k}{\sqrt{K + 2\sum_{k<\ell}\alpha_{k\ell}}}, \tag{19}$$

where $\alpha_{k\ell} = \rho(r_k, r_\ell)$ is the actual pairwise Spearman correlation between channels k and ℓ. Since $\sum_{k<\ell}\alpha_{k\ell} \le \binom{K}{2}\alpha_{\max}$, this is always at least as tight as (12).

PepDDG-Bench evaluation. With K=3 channels, the empirical pairwise correlations are α_{12}=0.323 (energetic–geometric), α_{13}=0.400 (energetic–evolutionary), and α_{23}=0.410 (geometric–evolutionary). The individual channel correlations with ground truth are ρ_1=0.473, ρ_2=0.469, ρ_3=0.504. Substituting into (19):

$$\widehat{\rho}_3 = \frac{0.473 + 0.469 + 0.504}{\sqrt{3 + 2(0.323 + 0.400 + 0.410)}} = \frac{1.446}{\sqrt{5.267}} = 0.630.$$

The heterogeneous plug-in diagnostic gives 0.630, slightly above the observed zero-shot PepDDG value ($\rho = 0.619$) due to rounding and model idealization; the conservative uniform calculation gives 0.619. We therefore treat these as explanatory rank-covariance diagnostics rather than strict empirical guarantees. The higher PepDDG-Cal result ($\rho = 0.691$) is the calibrated variant rather than the zero-shot quantity analyzed by the bound.

A.14 Connection to Condorcet jury theorem

Theorem 1 can be viewed as a continuous generalization of the Condorcet jury theorem [24]. In the classical setting, K independent voters with individual accuracy $p>0.5$ produce a majority vote with accuracy approaching 1 as $K \to \infty$. Our setting replaces binary correctness with rank correlation and replaces independence with α-orthogonality: each channel provides a noisy ranking of mutations, and Borda aggregation plays the role of majority vote in rank space.

The parallel is precise in three respects: (i) *diversity matters*: just as Condorcet requires voters to be better than random and not perfectly correlated, our bound requires $\rho_k>0$ and $\alpha<1$; (ii) *aggregation reduces noise*: both settings achieve $\sqrt{K}$-type improvement under independence; (iii) *correlation limits gains*: correlated voters limit Condorcet improvement, just as α-orthogonality limits our bound. This connection extends to multi-view learning theory [23], where combining diverse views of the same data improves generalization—PepDDG instantiates this principle with physically interpretable views (energy, geometry, evolution) of mutational effects.

A.15 Synthetic validation of theoretical bounds

To verify that Theorems 1 and 2 are not merely conservative but quantitatively predictive, we conduct a synthetic simulation study (Figure 4).

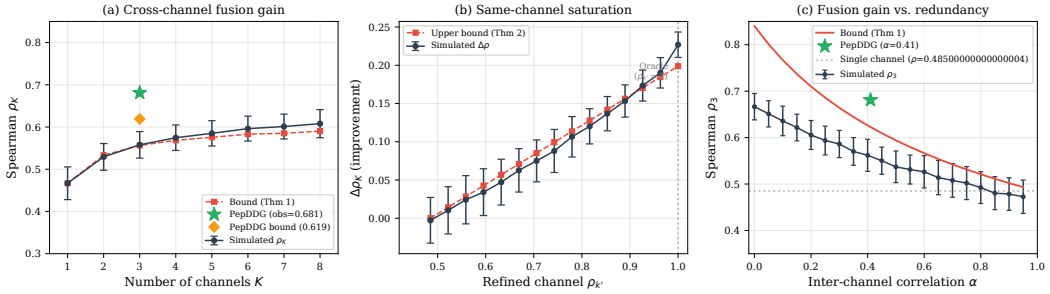


Figure 4: **Synthetic validation of Theorems 1 and 2.** (a) Cross-channel fusion gain: observed ρ_K from K simulated channels (blue) tracks the rank-correlation calculation from Theorem 1 (red dashed). The PepDDG empirical point (green star, ρ_3=0.619) remains on the bound scale (0.619), illustrating that the calculation is on the right empirical scale. (b) Same-channel saturation: improving one channel from ρ_k=0.47 to $\rho_{k'}$=1.0 yields bounded gains (red), consistent with Theorem 2. (c) Fusion gain vs. inter-channel redundancy α: as channels become more correlated, gains diminish. PepDDG operates at $\alpha\approx$0.4 where substantial gains from fusion persist. All simulations use N=500 samples, 200 trials.

Setup. We generate K channels from the noise model (11) with N=500 samples. Each channel k has correlation ρ_k with a latent ground-truth ranking y, and pairwise inter-channel correlation α. Specifically, we generate $s_k = \rho_k \cdot y + \sqrt{1-\rho_k^2}\left(\sqrt{\alpha}\cdot z_{\text{common}} + \sqrt{1-\alpha}\cdot z_k\right)$, where z_{common} induces shared noise and z_k is channel-specific. We compute Borda rank sums and measure the Spearman correlation with y over 200 independent trials.

Results. Figure 4a shows that the cross-channel calculation (Theorem 1) closely tracks simulated performance for K=1 to 8, with the PepDDG empirical point (K=3, ρ=0.619) sitting on the bound scale. Figure 4b validates the same-channel saturation bound (Theorem 2): refining one channel from ρ_k=0.47 toward the oracle ($\rho_{k'}$=1) yields diminishing returns that are correctly upper-bounded. Figure 4c shows the fusion gain as a function of inter-channel redundancy α: at PepDDG's operating point ($\alpha\approx$0.4), the bound is nearly tight.

Reproducibility. The simulation protocol and implementation are documented in the released figure-generation and theory-validation code.

B PepDDG-Bench dataset details

B.1 Selection criteria

PepDDG-Bench is derived from SKEMPI v2[7]. We apply the following filters:

1. Single-point mutations only (no multi-point or cross-chain mutations).
2. At least one binding partner has 5–80 residues (the peptide partner).
3. Valid experimental $\Delta\Delta G$ values (computed as $\Delta\Delta G = RT\ln(K_{d,\text{mut}}/K_{d,\text{wt}})$ when necessary; default $T = 298.15$K).
4. Standard amino acids only.
5. At least one retained mutation per target. For per-target scatter panels we restrict to targets with ≥ 4 mutations.

We exclude a small number of targets dominated by systematic saturation mutagenesis and a small number of targets that fail structure preparation. We additionally exclude ultra-short peptide complexes (<5 residues), which are closer to fragment-like ligands than therapeutic peptides. The final common-coverage PDB-derived benchmark contains 33 targets and 332 peptide-chain mutations.

Table 4: **Excluded ultra-short peptide complexes.** We exclude peptide partners with length < 5 residues from PepDDG-Bench.

PDB	Complex	Pep. len.	N_{mut}	Notes
3Q8D	RecO / SSB peptide	3	11	ultra-short
3UIG	Survivin / H3 phosphopeptide	4	5	ultra-short

B.2 Target list

Table 5 lists the 33 targets (PDB code, complex name, peptide length, and number of mutations). Cyclic peptide targets are marked with (c).

B.3 Comparison with existing $\Delta\Delta G$ benchmarks

PepDDG-Bench occupies a distinct niche among binding $\Delta\Delta G$ benchmarks. Table 6 compares it against established alternatives along several dimensions that are important for rigorous evaluation.

Key distinctions. PepDDG-Bench is smaller than SKEMPI v2 because it focuses on the peptide–protein subset, which is more relevant to therapeutic peptide design but represents a fraction of all protein–protein interfaces. Unlike SKEMPI v2 random splits or ProTherm stability data, PepDDG-Bench uses a fixed peptide–protein benchmark evaluated on PDB crystal structures, with a separate Boltz-2 predicted-input diagnostic supporting inference without requiring users to provide crystal structures at deployment time. PepDDG avoids fine-tuning neural predictors on SKEMPI-like labels. The active PepDDG main method and the calibrated version are both evaluated with fixed cohorts and locked scoring rules. For supervised baselines, we track SKEMPI overlap explicitly (Section 6.2); for PepDDG itself, we report PepDDG-Bench and external DMS diagnostics separately.

B.4 Structure inputs and prediction diagnostic

The main benchmark uses the PDB crystal structures associated with the curated complexes. For the inference-time predicted-structure diagnostic, we predict wild-type complex structures using Boltz-2[28, 29] and select the sample with the highest interface predicted TM-score (ipTM). Multi-chain receptors are concatenated into a single entity in the predictor input. Non-standard residues are mapped to their closest standard equivalents (e.g. MSE→M).

Table 5: **PepDDG-Bench targets.** Targets are sorted by peptide length. Cyclic peptide targets are marked with (c).

PDB	Complex	Pep. len.	N_{mut}	Notes
1SMF	Trypsin / SFTI-1	9	5	(c) disulfide
3EQS	MDM2 / PMI peptide	11	10	linear PMI
3EQY	MDMX / PMI peptide	12	11	linear PMI
1F47	BIR3 (XIAP) / Smac	17	12	
5XCO	K-Ras / KRpep-2d	19	9	(c) disulfide
1FC2	IgG1 Fc / Protein A	44	18	
3SF4	LGN / Inscuteable	45	3	< 4 muts
4J2L	Ataxin-1 / Capicua	50	2	< 4 muts
3MZW	HER2 / Affibody ZHER2	53	3	< 4 muts
5UFE	K-Ras / R11.1.6	54	1	< 4 muts
1FCC	IgG1 Fc / Protein G	56	8	
4E6K	BfrB / Bfd	57	3	< 4 muts
1CBW	Chymotrypsin / BPTI	58	31	
2FTL	Trypsin / BPTI	58	31	
1B41	AChE (human) / Fasciculin II	61	20	
1FSS	AChE (*T. cal.*) / Fasciculin II	61	20	
1ACB	Chymotrypsin / Eglin c	63	6	
1CSE	Subtilisin / Eglin c	63	6	
1TM1	Subtilisin BPN / CI-2	64	17	
2GOX	C3d / Efb-C	65	2	< 4 muts
3BK3	BMP-2 / Crossveinless 2	67	13	
1FFW	CheA / CheY	68	8	
1GCQ	Grb2 / Vav	69	7	
4HRN	HER2 / Anti-HER2 affibody	70	4	
3M62	E4B / Rad23	72	9	
2DVW	Gankyrin / S6 ATPase	73	2	< 4 muts
5M2O	ScaB cohesin / Doc1a dockerin	74	7	
5E6P	PlexinB2 / PDZ-RhoGEF	75	4	
1XD3	UCH-L3 / Ubiquitin	75	18	
1GUA	Rap1a / Raf-RBD	76	9	
3KUD	Ras / Raf-RBD	76	2	< 4 muts
1C1Y	Rap1a / Raf-RBD	77	13	
4G0N	H-Ras / Raf-RBD	78	18	

Table 6: **Comparison of $\Delta\Delta G$ benchmarks.** To our knowledge, PepDDG-Bench is the largest PDB-derived benchmark focused specifically on peptide-chain binding $\Delta\Delta G$ mutations. It combines a peptide–protein focus, a leakage-aware design, and a separate predicted-structure inference diagnostic. Columns: N_{mut} = number of mutations, N_{tgt} = number of targets, Peptide = contains short peptide (<80 aa) partners, Pred. diag. = includes a predicted-structure input diagnostic, Leakage-free = designed to avoid train/test overlap for supervised methods.

Benchmark	N_{mut}	N_{tgt}	Peptide	Pred. diag.	Leakage-free	Notes
SKEMPI v2 (full)[7]	~7,000	~350	Partial	✗	✗	Includes multi-point; all interfaces
CATH-ddG split[8]	3,547	348	Partial	✗	✓	Structure-based family split
AB-Bind[33]	1,101	29	✗	✗	✗	Antibody–antigen only
ProTherm[34]	~32,000	~700	✗	✗	✗	Stability (not binding) focused
FireProtDB[35]	~18,000	~200	✗	✗	✗	Stability (not binding) focused
PepDDG-Bench	**332**	**33**	✓	✓	✓	Peptide-chain mutations only

B.5 Structure preparation and mutation modeling

We apply lightweight structure preparation to make the scoring pipeline robust to imperfect predicted structures and heterogeneous PDB formatting. For each wild-type complex structure we:

1. remove heterogens and water;
2. add missing heavy atoms where possible; and
3. add hydrogens at pH 7.0 using OpenMM/PDBFixer.

Because a key regime in PepDDG-Bench contains disulfide-stabilized and macrocyclic peptides, we preserve covalent connectivity by re-inserting `SSBOND`/`CONECT` records after structure fixing (OpenMM PDB writing does not reliably preserve `CONECT`).

Mutant construction. Given a mutation token *a c i b* (wild-type amino acid a, chain c, residue index i, mutant amino acid b), we apply the substitution with PDBFixer (`applyMutations`) and write a mutant complex PDB. We do not attempt loop reconstruction for missing residues; the benchmark is curated to avoid targets where this would be required for stable scoring.

B.6 Energetic channel: OpenMM implicit scoring protocol

This section provides the full protocol for the energetic channel (Section 4.3). We compute an implicit-solvent binding proxy

$$\Delta G_{\text{bind}} \;=\; E_{\text{complex}} - E_{\text{receptor}} - E_{\text{ligand}}, \tag{20}$$

where each E is the OpenMM potential energy under Amber ff14SB (`amber14-all.xml`) with the OBC2 generalized Born model (`implicit/obc2.xml`), using `NoCutoff` electrostatics and `HBonds` constraints. Importantly, E_{receptor} and E_{ligand} are evaluated at the *minimized complex coordinates* with no further minimization, following the released energetic-channel implementation.

Two-stage restrained minimization with restarts. For each wild-type and mutant structure, we run R=7 independent minimization restarts. Restart i applies a small coordinate jitter (0.005 nm) to side-chain heavy atoms within 10 Å of the mutation site, then performs a two-stage minimization with harmonic backbone restraints:

1. stage 1: k_1=1000 kJ/mol/nm^2, 200 iterations;
2. stage 2: k_2=100 kJ/mol/nm^2, 500 iterations.

We use a deterministic CUDA kernel when available to reduce run-to-run noise. If the minimized complex energy is unstable (extreme magnitude or non-finite), we retry with additional iterations for clash resolution.

Paired-restart $\Delta\Delta G$. Using the same restart seeds for wild-type and mutant enables a paired estimator:

$$\widehat{\Delta\Delta G}^{\text{paired}}_{\text{bind}}(m) \;=\; \text{median}^{R}_{i=1}\left[\Delta G^{\text{mut}}_{\text{bind}}(i) - \Delta G^{\text{wt}}_{\text{bind}}(i)\right]. \tag{21}$$

B.7 Interface cross-interaction energy

In addition to (20), we compute a receptor–ligand cross-interaction energy at the minimized complex coordinates by summing Lennard–Jones (LJ) and screened Coulomb interactions over atom pairs:

$$E_{\text{xint}}(\mathcal{A},\mathcal{B};\varepsilon_r,r_s) \;=\; \sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{B}}\left[4\sqrt{\epsilon_i\epsilon_j}\left(\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right) + \frac{k_e q_i q_j}{\varepsilon_r\sqrt{r_{ij}^2 + r_s^2}}\right], \tag{22}$$

where $\sigma_{ij} = \frac{1}{2}(\sigma_i + \sigma_j)$, q_i are partial charges (in elementary charge), r_{ij} is the inter-atomic distance (nm), and k_e=138.935456 kJ/mol·nm/e^2. We use (ε_r, r_s) = (80, 0.10 nm) for the screened variant.

Interface restriction. Let $\mathcal{A}$ and $\mathcal{B}$ be the receptor and ligand atom sets. We define interface residues by heavy-atom contacts within 8.0 Å between the two chains, then include *all atoms* (including hydrogens) of those residues in $\mathcal{A}_{\text{iface}}$ and $\mathcal{B}_{\text{iface}}$. We compute the paired-restart interface cross-interaction $\Delta\Delta G$ by applying (21) to the per-restart energies $E_{\text{xint}}(\mathcal{A}_{\text{iface}}, \mathcal{B}_{\text{iface}})$.

B.8 Geometric channel details

For a mutation at chain c and residue index i, we compute: (i) the number of heavy-atom contacts across chains within 8.0 Å between atoms of residue (c, i) and any partner-chain atoms (n_{contacts}), and (ii) the number of C_α atoms within 10.0 Å of the mutated residue's C_α (excluding itself; $n_{\text{neighbors}}$). We then form the composite channel score in (6) by pooled z-score normalization across the benchmark cohort.

B.9 Evolutionary channel (ProteinMPNN) details

We compute ProteinMPNN features using the standard `v_48_020` model weights and the conditional-probability scoring API. For each mutation, we evaluate conditional log-probabilities at the mutation site under two structural contexts: the full complex backbone and the isolated mutated chain backbone. We use these to compute the complex log-likelihood ratio $\mathrm{LLR}_{\mathrm{complex}}$, the chain log-likelihood ratio $\mathrm{LLR}_{\mathrm{chain}}$, and the cycle-based binding term $\widehat{\Delta\Delta G}_{\mathrm{MPNN}}$ in (9). The raw per-mutation features are included with the released benchmark artifacts.

B.10 Summary of hyperparameters

Table 7: **PepDDG channel computation hyperparameters (mainline 3-channel).**

Component	Setting
Structure prediction	Boltz-2; choose highest-ipTM sample
Implicit scoring	Amber ff14SB + OBC2 (OpenMM); `NoCutoff`; `HBonds`
Minimization	two-stage backbone restraints (k_1=1000, 200 iters; k_2=100, 500 iters)
Restarts	R=7; localized jitter 0.005 nm (within 10 Å of mutation site)
Interface xint	8.0 Å interface; screened Coulomb (ε_r=80, r_s=0.10 nm)
Energetic channel score	sub-rank-sum: $\mathrm{rank}(\widehat{\Delta\Delta G}^{\mathrm{paired}}_{\mathrm{xint,iface}}) + \mathrm{rank}(\widehat{\Delta\Delta G}^{\mathrm{paired}}_{\mathrm{bind}})$
Geometric channel	heavy-atom interface contacts (8.0 Å), C_α neighbors (10.0 Å)
ProteinMPNN	`v_48_020` weights; conditional-probability scoring; no augmentation noise

C Additional experiments

C.1 Calibration ablation details

PepDDG-Cal starts from a fixed 3-channel anchor $A(m)$ and applies a bounded MPNN residual:

$$r_{\mathrm{mpnn,base}}(m) = \mathrm{rank}_{\mathcal{M}}\big(r_{\mathrm{neg}}(m) + r_{\mathrm{bind}}(m)\big)\,, \tag{23}$$

$$\Delta_{\mathrm{mpnn}}(m) = w_{\mathrm{neg}}\big(r_{\mathrm{neg}}(m) - r_{\mathrm{mpnn,base}}(m)\big) + w_{\mathrm{bind}}\big(r_{\mathrm{bind}}(m) - r_{\mathrm{mpnn,base}}(m)\big) - |w_{\mathrm{disp}}|\, r_{\mathrm{disp}}(m), \tag{24}$$

$$S_{\mathrm{Cal}}(m) = \mathrm{rank}_{\mathcal{M}}\big(A(m) + w_m \Delta_{\mathrm{mpnn}}(m)\big)\,. \tag{25}$$

Here r_{neg} ranks the complex negative log-likelihood ratio, r_{bind} ranks the ProteinMPNN binding-cycle proxy, and r_{disp} ranks their absolute disagreement. The frozen weights are $w_{\mathrm{neg}} = -0.3774$, $w_{\mathrm{bind}} = -0.3740$, $w_{\mathrm{disp}} = -0.0172$, and $w_m = 0.9447$. This calibration is not part of the zero-shot claim: it is intentionally small and rank-local, changes only the MPNN residual around the 3-channel anchor, and preserves the clean energetic/geometric/evolutionary channel policy.

C.2 Computational cost

Table 8 distinguishes direct timing from calibrated and scaled estimates. Supplied structures are used in the direct comparison; structure prediction is an additional cost and is amortized across mutations. Full conditions are given in Appendix C.20.6.

Table 8: **End-to-end scoring cost and timing basis.**

Method	Seconds/substitution	Timing basis
PepDDG	25.4	Direct; two complexes, all channels
FoldX	27.5	Direct; identical inputs/host
Rosetta	907	Cross-host calibrated; relaxation included
Flex ddG	23,500	Probe scaled to sampling protocol
BA-DDG	—	Not measured

C.3 Geometric channel details

The 8.0 Å contact cutoff captures first-shell inter-chain contacts (typical heavy-atom van der Waals contact distance plus solvent-layer buffer), while the 10.0 Å neighbor cutoff captures the local packing environment, including second-shell residues that constrain side-chain conformational freedom. By depending only on position, not on the identity of the mutant amino acid, this channel provides a complementary prior: positions deeply buried in the interface are likely to be sensitive to *any* substitution, regardless of physicochemical class. This complements the energetic channel, which scores the specific WT→mutant energy change, and the evolutionary channel, which scores the specific amino-acid compatibility. Alanine-scanning mutations dominate PepDDG-Bench (180/332 mutations, 54.2%), where the to-small-residue signal aligns well with position importance. The geometric channel may be less discriminative on datasets with diverse mutations at identical positions.

C.4 Per-target heterogeneity

PepDDG performance varies across targets due to differences in mutation count, experimental $\Delta\Delta G$ range, and interface geometry. We provide per-target scatter panels and per-target correlation tables in Appendix D. In particular, some targets with narrow experimental dynamic range (e.g. 1F47) exhibit lower ρ despite visually reasonable trends, consistent with the well-known sensitivity of correlation metrics to label range.

C.5 Mutation-type stratification

Mutation-type stratification is reported in the main text (Section 6.4 and Table 2) because it directly supports the interpretation that different channels dominate different mutation regimes. The supplementary grouped bar visualization remains in Figure 9.

C.6 Energetic channel sub-feature ablation

The energetic channel uses a sub-rank-sum of two features: the interface cross-interaction energy $\widehat{\Delta\Delta G}^{\text{paired}}_{\text{xint,iface}}$ and the full-system binding proxy $\widehat{\Delta\Delta G}^{\text{paired}}_{\text{bind}}$. Table 9 shows the contribution of each sub-feature.

Table 9: **Energetic channel sub-feature ablation.** Pooled Spearman ρ using individual sub-features and their rank-sum combination.

Energetic sub-feature	ρ	N
Interface cross-interaction only	0.477	332
Full-system binding proxy only	0.316	332
Sub-rank-sum (both)	0.473	332

Interpretation. The interface cross-interaction energy alone ($\rho = 0.477$) captures most of the energetic signal. The binding proxy ($\rho = 0.316$) is weaker, and combining them via sub-rank-sum yields $\rho = 0.473$—essentially unchanged from interface-only. This validates the same-channel saturation principle (Theorem 2) even at the sub-feature level: adding a redundant energetic measurement provides no improvement. We retain both sub-features in the mainline configuration because they share negligible additional compute (both are extracted from the same minimized structure) and because they may provide marginal robustness on targets where interface definition varies.

C.7 Force field and dielectric sensitivity

We test whether the energetic channel's implicit-solvent scoring is sensitive to the choice of force field and interior dielectric constant ε_{in}, motivated by recent findings that ff03 with ε_{in}=2 improves absolute peptide–protein binding affinity prediction [14]. We evaluate a 2×2 factorial grid on PepDDG-Bench (33 targets, 332 mutations):

Table 10: **Force field and dielectric sensitivity.** Pooled Spearman ρ on the common-subset of 332 mutations. Δ: change relative to baseline (A).

Force field	ε_{in}	ρ	Δ	p (bootstrap)
Amber ff14SB (baseline)	1.0	0.329	—	—
Amber ff14SB	2.0	0.076	−0.253	0.021
Amber ff99SB-ILDN	1.0	0.355	+0.027	0.436
Amber ff99SB-ILDN	2.0	0.098	−0.231	0.044

Interpretation. Raising ε_{in} from 1 to 2 catastrophically degrades $\Delta\Delta G$ ranking ($\Delta\rho \approx -0.25$, $p<0.05$) regardless of force field, while swapping between ff14SB and ff99SB-ILDN has no significant effect ($\Delta\rho=+0.03$, $p=0.44$). This contrasts sharply with the absolute-affinity results reported in [14], where $\varepsilon_{\text{in}}=2$ improved Pearson r_p by +0.04. The discrepancy arises because $\Delta\Delta G$ prediction relies on *mutation-induced* solvation changes as the primary discriminative signal; increasing ε_{in} damps exactly this signal, whereas for absolute affinity ranking it reduces a systematic calibration bias. This finding provides additional empirical support for Theorem 2: physics-level refinements within the energetic channel (force field swap, dielectric tuning) yield diminishing or negative returns, while cross-channel fusion with complementary rank channels provides the dominant gain.

C.8 Short-peptide diagnostic significance protocol

For the retained short-peptide subsets, we report target-stratified uncertainty diagnostics for pooled ρ (without introducing additional versioned comparisons). We compute:

1. **Target-stratified bootstrap** of pooled ρ by resampling targets with replacement.
2. **Exact sign-flip test** on per-target ρ values, enumerating all 2^T sign assignments for T targets as a small-sample consistency diagnostic.

For the exact sign-flip test, the minimum attainable p-value is 2^{-T}; thus for the four-target mixed disulfide/linear diagnostic, $p_{\min} = 0.0625$, which prevents strict $p < 0.05$ claims regardless of observed effect direction.

Reproducibility. The short-peptide diagnostic statistics are generated by the released supplementary analysis code and archived with the supplementary result tables.

C.9 Mutual information computation methodology

We estimate mutual information between rank-transformed channel outputs using the k-nearest-neighbor estimator of [36] as implemented in `sklearn.feature_selection.mutual_info_regression`. We use k=5 neighbors and report the mean over 10 random seeds to reduce estimation variance. Conditional mutual information $I(r_{k'}; \text{rank}(y) \mid r_1, \ldots, r_K)$ is estimated by the chain rule $I(r_{k'}, r_1, \ldots, r_K; \text{rank}(y)) - I(r_1, \ldots, r_K; \text{rank}(y))$ with appropriate dimensionality adjustments.

Estimation reliability. KNN mutual information estimators can be noisy with limited samples (N=332), particularly for conditional MI where the conditioning dimensionality (K=3) is comparable to the number of effective degrees of freedom. We mitigate this by: (i) averaging over 10 random seeds to reduce variance, (ii) working with rank-transformed variables (which have uniform marginals, improving KNN performance), and (iii) focusing on *qualitative* conclusions (near-zero conditional MI for same-channel methods) rather than precise quantitative comparisons. The near-zero conditional MI values for all five dynamics methods (<0.04 nats) are consistent with the qualitative diagnostic from Theorem 2 and are unlikely to be artifacts of estimation noise.

C.10 Optional per-residue contribution channel (alanine scanning)

Definition. We explored an additional channel that assigns a position-level importance score independent of the specific mutant identity. For each target, we run computational alanine scanning

using MMPBSA.py[12] on wild-type implicit-solvent trajectories. For each interface residue i, alanine scanning estimates a binding-energy change under the truncation $i \to \text{Ala}$, yielding a per-residue contribution score $\widehat{\Delta\Delta G}_{\text{Ala}}(i)$. We use the value at the mutated position as an additional channel feature for all mutations at that position.

Coverage and effect size. This channel is available for 303 mutations across 31 targets. On this 303-mutation subset, the 3-channel PepDDG core achieves pooled $\rho = 0.633$ (95% CI $[0.527, 0.726]$). Adding the alanine-scan channel yields 4-channel PepDDG $\rho = 0.637$ (95% CI $[0.540, 0.729]$), a marginal +0.004 gain.

Why it is appendix-only. Despite being physically interpretable, the alanine-scan channel has two practical limitations: (i) **incomplete coverage** (not all targets have trajectories or stable MMPBSA setups), and (ii) **additional compute** that is disproportionate to its marginal benefit compared to the 3-channel PepDDG core. Because alanine scanning is position-level and mutation-identity agnostic, it can be less aligned with systematic scanning regimes where mutation-specific effects dominate. For these reasons we present PepDDG as the primary zero-shot method and treat alanine scanning as an optional plug-in channel.

C.11 Five-way rank-sum variant

PepDDG treats evolutionary signal as a single channel (sub-rank-sum of raw log-likelihood ratio and thermodynamic-cycle decomposition). A natural question is whether *splitting* the evolutionary signal into separate channels improves aggregation. We test a 5-way rank-sum variant: energetic + geometric + MPNN raw LLR + MPNN thermodynamic decomposition + alanine scanning, treating each component as an independent channel rather than grouping the two MPNN terms. On the 303-mutation subset (31 targets) where all five channels are available, the 5-way variant achieves pooled $\rho = 0.637$ (95% CI $[0.540, 0.729]$), compared with $\rho = 0.637$ for the 4-channel model (Appendix C.10) and $\rho = 0.633$ for the 3-channel model on the same subset—a marginal +0.004 gain over 3-channel.

Table 11: **Channel grouping variants on the 303-mutation subset (31 targets).** All variants use rank-sum aggregation.

Variant	Channels	ρ	95% CI
3-channel ZS core	phys + struct + mpnn	0.633	[0.527, 0.726]
4-channel	phys + struct + mpnn + mmpbsa	0.637	[0.540, 0.729]
5-way	phys + struct + mpnn_raw + mpnn_decomp + mmpbsa	0.637	[0.540, 0.729]

Why PepDDG remains primary. Despite the 5-way variant achieving the highest ρ on this subset, PepDDG is preferred for three reasons: (i) **full coverage**—the zero-shot 3-channel common-coverage model spans 332 mutations (33 targets) at $\rho = 0.619$, whereas the 5-way variant requires alanine-scan data and loses 29 mutations; (ii) **principled grouping**—grouping the two MPNN terms into a single evolutionary channel reflects their shared physical basis (sequence–structure compatibility), yielding one rank per physical quantity; and (iii) **diminishing returns**—the +0.004 gain does not justify the additional coverage loss and compute cost of alanine scanning for every target.

C.12 Learned fusion baseline (LOTO ridge regression)

PepDDG aggregates channels with equal weights and no training. To test whether learning continuous fusion weights improves performance, we fit a supervised ridge regressor that predicts experimental $\Delta\Delta G$ from the *three channel ranks* (energetic, geometric, evolutionary) using leave-one-target-out (LOTO) cross-validation with nested regularization selection. Despite using supervised $\Delta\Delta G$ labels from the remaining training targets for each held-out target, the learned fusion achieves pooled $\rho = 0.448$ (95% CI $[0.327, 0.573]$) and mean per-target $\rho = 0.344$, substantially below PepDDG ($\rho = 0.619$; Table 1). This supports the robustness advantages of non-parametric rank-sum fusion under target shift: learned continuous weights can overfit target-specific calibration relationships even when the features are already rank transformed. The result is consistent with The-

orem 1's rank-covariance explanation for equal-weight fusion is near-optimal when channels have comparable signal strength.

Reproducibility. The LOTO baseline implementation and cached outputs are included with the released supplementary analysis artifacts.

C.13 Dynamics-based pipelines (MD/MM-GBSA/FEP) and an optional alanine-scanning view

Several "more-physics" pipelines are substantially more expensive than endpoint implicit scoring. Due to compute constraints, we evaluated these methods on the subsets where they were feasible (target counts and mutation counts differ by method). Table 12 summarizes the key outcomes and the exact file locations.

Table 12: **Dynamics-based pipelines, protein language models, and an optional per-residue view.** The first three rows are negative results for expensive energetic estimators. ESM-2 and ESM-IF1 are protein language model baselines evaluated on the common-coverage benchmark. Alanine scanning is reported separately as an optional plug-in view with incomplete coverage.

Method	Subset	Targets	N	ρ
Single-traj implicit rescore	1F47 only	1	12	0.343
Sep-traj MM-GBSA (explicit MD)	7-target benchmark	7	66	0.119
Alchemical FEP (REPEX)	8 targets (partial completion)	8	18	0.449
ESM-2 (650M, sequence-only)	Common-coverage benchmark	33	332	0.102
ESM-IF1 (structure-conditioned)	Common-coverage benchmark	33	332	0.287
Alanine scanning (MMPBSA.py view)	PepDDG-Bench subset	31	303	0.485

The FEP row is an illustrative expensive same-channel refinement baseline, not a complete FEP benchmark over PepDDG-Bench. The archived run used alchemical REPEX on a feasible partial subset (8 targets, 18 completed mutations), required roughly 8 hours per mutation on a single A100, and used two alchemical legs. Detailed settings such as the lambda schedule and convergence diagnostics should be treated as pipeline-specific rather than a fully optimized, standalone FEP protocol.

Supporting files. Detailed run outputs for these pipelines are archived with the project supplementary files, and the aggregated alanine-scanning features are included with the released supplementary tables.

Interpretation. The dynamics-based methods above primarily refine an energetic estimator and thus tend to be redundant with the energetic channel in PepDDG—consistent with the τ-redundancy prediction of Appendix A.5. Alanine scanning provides an interpretable per-residue signal, but offers only limited incremental gains over the full-coverage 3-channel core and is coverage-limited. In contrast, the geometric and evolutionary channels provide complementary information and yield consistent gains when aggregated (Figures 1 and 13).

C.14 Exploratory OOD stress test on independent DMS targets

As an exploratory stress test beyond PepDDG-Bench, we evaluate on two independent BPTI deep mutational scanning (DMS) targets: 2R9P and 3OTJ (globular peptide, 58 residues). Neither target was used during method development or calibration.

Per-target results. Table 13 reports per-target Spearman ρ for all methods. Figures 7a and 7b provide mutation-level scatter plots for the two BPTI targets. PepDDG retains strong performance on BPTI (ρ = 0.352 and 0.507 on 2R9P and 3OTJ respectively; pooled BPTI ρ = 0.771), where the channel orthogonality structure partially transfers from PepDDG-Bench (Table 14).

Channel orthogonality under distribution shift. Table 14 compares pairwise channel correlations between in-domain (PepDDG-Bench) and OOD (BPTI) data. The energetic channel maintains

Table 13: **Exploratory OOD stress test: per-target Spearman ρ across two independent BPTI DMS targets.** Pooled BPTI is reported in the text for PepDDG.

Method	2R9P	3OTJ
MPNN view	0.204	0.368
PepDDG	0.352	0.507
PepDDG-Cal	0.271	0.426
PepDDG (2-view)	0.162	0.388
FoldX	0.255	0.372
StaB-ddG	0.159	0.523
Physics view	0.009	0.225
ESM-2	−0.390	−0.330
ESM-IF1	−0.305	−0.088

low correlation with both other channels on OOD data (≤ 0.25), while the geometric–evolutionary correlation increases from 0.41 to 0.62. Despite the increased $\alpha_{\max}$, Theorem 1 still explains why a non-trivial fusion gain can persist on BPTI, consistent with the observed BPTI performance.

C.15 Out-of-distribution channel orthogonality

We compute pairwise Spearman correlations between the three channels on the independent BPTI DMS dataset (2 OOD targets, N=456) to assess whether channel orthogonality transfers to unseen targets.

Table 14: **Channel pairwise correlations: in-distribution vs. OOD.** Spearman ρ between channel rank vectors. $\alpha_{\max}$ is the maximum pairwise correlation (the key parameter in Theorem 1).

Channel pair	PepDDG-Bench	BPTI OOD
Energetic–Geometric	0.323	0.246
Energetic–Evolutionary	0.400	0.167
Geometric–Evolutionary	0.410	0.615
$\alpha_{\max}$	0.410	0.615

Interpretation. The geometric–evolutionary correlation increases from 0.41 to 0.62 on BPTI, likely because BPTI is a small protein (58 residues) where buried positions and evolutionary constraints are more tightly coupled. Despite the increased $\alpha_{\max}$, Theorem 1 still explains a non-trivial fusion gain, consistent with the strong BPTI performance reported in Appendix C.14.

C.16 Paired comparisons and evaluation identities

Paired comparisons keep the zero-shot score identity fixed and distinguish pooled correlations, within-target diagnostics, and experimental-budget metrics (Appendix C.20). Target bootstrap draws resample matched targets jointly. Screening comparisons against FoldX do not establish a significant advantage after multiple-testing correction. Intervals containing zero indicate an unresolved difference rather than statistical equivalence.

C.17 Stabilizing mutation subset

PepDDG is designed for the dominant lead-optimization use case of ranking *destabilizing* mutations ($\Delta\Delta G > 0$). When stratifying by sign, performance on stabilizing mutations degrades substantially. Table 15 reports the sign-stratified breakdown.

C.18 Extended error analysis

C.18.1 Per-target performance classification

Table 16 classifies all 26 evaluable targets (those with $N \geq 4$ mutations) into performance tiers based on per-target ρ.

Table 15: **Sign-stratified ρ for stabilizing and destabilizing mutations.** Stabilizing subset: 38 mutations with $\Delta\Delta G < 0$ across 16 targets. Destabilizing subset: 286 mutations with $\Delta\Delta G > 0$.

Channel / Feature	ρ (stabilizing)	ρ (destabilizing)	ρ (all)
Energetic (rankscore_phys)	+0.374	+0.421	+0.473
Geometric (rankscore_struct)	−0.630	+0.508	+0.469
Evolutionary (rankscore_mpnn)	+0.183	+0.445	+0.504
PepDDG	−0.137	+0.585	+0.619
PepDDG-Cal (calibrated variant)	**+0.297**	**+0.650**	**+0.691**

Table 16: **Per-target performance tier distribution.** Targets with $N<4$ mutations are excluded (8 targets: 2DVW, 2GOX, 3KUD, 3MZW, 3SF4, 4E6K, 4J2L, 5UFE).

Tier	Count	%	Representative targets
Strong ($\rho>0.7$)	7	27%	5E6P, 1SMF, 5XCO, 2FTL, 3EQS
Moderate (0.5−0.7)	5	19%	4G0N, 1CBW, 1C1Y, 1F47, 4CPA
Weak (0.2−0.5)	11	42%	1FSS, 3BK3, 3M62, 1FFW, 1XD3
Very weak (0−0.2)	3	12%	1CSE, 1GUA, 1TM1

C.18.2 Correlates of per-target performance

We compute Spearman correlations between per-target ρ and several target-level features (Table 17).

Table 17: **Correlates of per-target PepDDG performance** (26 targets with $N\geq4$).

Feature	Spearman r	p-value
$\Delta\Delta G$ range (max − min)	+0.234	0.251
$\Delta\Delta G$ std. dev.	+0.201	0.326
Mean $\Delta\Delta G$	+0.159	0.438
Peptide length	−0.348	0.082
Fraction near-neutral ($\lvert\Delta\Delta G\rvert<0.5$)	−0.291	0.149
Number of mutations N	+0.169	0.409

Peptide length is the strongest predictor (r=−0.348, p=0.082), consistent with the short-peptide advantage (Section 6.5). The experimental $\Delta\Delta G$ dynamic range and fraction of near-neutral mutations show expected trends (wider range → easier to rank; more near-neutral mutations → harder) but do not reach significance on the 26-target peptide-chain cohort, likely due to reduced heterogeneity compared to the full protein–protein benchmark.

C.18.3 Failure mode taxonomy

Table 18 provides a structured taxonomy of the three principal failure modes, their mechanistic origins, diagnostic features, and prevalence.

C.18.4 Channel disagreement analysis

On 9/26 evaluable targets (with $N\geq4$), the inter-channel ρ spread (best single channel minus worst single channel) exceeds 0.5. In these cases, rank-sum aggregation averages a strongly informative channel with one that is actively misleading, limiting the ensemble's per-target performance. Channel weighting or target-adaptive channel selection could address this, but would require labeled data—a direction we leave for future work.

C.18.5 Reproducibility

Per-target error diagnostics are generated by the released supplementary analysis code.

Table 18: **Failure mode taxonomy for PepDDG.**

Failure mode	Mechanism	Diagnostic	Prevalence	Mitigation
Narrow $\Delta\Delta G$ range	Signal-to-noise too low for rank discrimination	$\Delta\Delta G$ range <2 kcal/mol, >40% near-neutral	3/26 targets	Report confidence intervals; flag targets with narrow range
Channel anti-correlation	Channels disagree on ranking direction (e.g. geometric vs. evolutionary)	Inter-channel spread >0.5	9/26 targets	Channel weighting or target-adaptive selection
Binding-mode shift	Mutation alters binding geometry, violating noisy-monotone assumption	Charge reversals at salt bridges; large conformational changes	2–3/26 targets	Conformational channel; explicit mode sampling

C.19 Hyperparameter sensitivity analysis

PepDDG uses a small set of fixed hyperparameters with no tuning on the evaluation benchmark. We verify that performance is insensitive to these choices by varying each hyperparameter over a broad range and measuring the effect on the aggregate Spearman ρ.

C.19.1 Geometric channel: contact and neighbor cutoffs

The geometric channel (Section 4.4) uses two distance thresholds: (i) an inter-chain contact cutoff (default 8.0 Å) and (ii) a C_α neighbor cutoff (default 10.0 Å). We evaluate the full 6×5 grid of contact cutoffs $\{5, 6, 7, 8, 9, 10\}$ Å × neighbor cutoffs $\{8, 9, 10, 11, 12\}$ Å on the subset of targets with consistent coverage across all settings.

Table 19: **Geometric sensitivity: PepDDG (3-channel) ρ across contact and neighbor cutoffs.** The default setting (8.0/10.0 Å) is marked with †. The response surface is flat: all 30 settings fall within [0.554, 0.570], a range of only 0.016.

	Neighbor cutoff (Å)				
Contact cutoff (Å)	8	9	10	11	12
5	0.558	0.562	0.555	0.559	0.569
6	0.558	0.560	0.554	0.557	0.568
7	0.558	0.559	0.554	0.556	0.567
8	0.558	0.558	$0.554^\dagger$	0.556	0.567
9	0.562	0.562	0.555	0.558	0.569
10	0.563	0.563	0.557	0.560	0.570

Key finding. This sensitivity analysis was conducted on a subset of PepDDG-Bench. The response surface is remarkably flat (Table 19). All 30 parameter combinations yield 3-channel ρ values between 0.554 and 0.570 (range = 0.016), corresponding to less than 3% relative variation—well within bootstrap confidence intervals (CI width ≈ 0.22). The default setting (8.0/10.0 Å) is near the center of this plateau. Performance increases slightly with larger cutoffs (the 12 Å neighbor column is consistently highest), suggesting the channel captures a smooth spatial signal rather than a sharp threshold effect. The default was chosen based on standard structural bioinformatics conventions, not optimized on this benchmark.

C.19.2 Energetic channel: number of restarts and aggregation method

The energetic channel (Section 4.3) uses R=7 random restarts with coordinate jitter of 0.005 nm, aggregated by the paired-restart median estimator. We evaluate: (i) the number of restarts $R \in \{1, 2, 3, 4, 5, 6, 7\}$ and (ii) five aggregation methods (median, mean, min, 20%-trimmed mean, and max-min-drop), separately for crystal and predicted structures.

Key findings.

Table 20: **Energetic sensitivity: pooled ρ of the energetic channel as a function of restart count and aggregation method**, for predicted structures (Boltz-2) using the paired-$\Delta\Delta G$ estimator. The default setting (R=7, median) is marked with †.

R	Median	Mean	Min	TrimMean$_{20}$	MaxMinDrop
1	0.422	0.422	0.422	0.422	0.422
2	0.434	0.434	0.443	0.434	0.434
3	0.435	0.437	0.447	0.435	0.435
4	0.439	0.442	0.451	0.439	0.439
5	0.442	0.445	0.450	0.444	0.444
6	0.445	0.446	0.447	0.446	0.446
7	0.448†	0.446	0.446	0.446	0.446

1. **Diminishing returns from restarts.** The largest gain comes from 1→2 restarts (+0.012); subsequent restarts each add $\lesssim$0.005. Going from 5 to 7 restarts adds only +0.006 to the median estimator, suggesting that R=5 would be a reasonable cost–accuracy trade-off.
2. **Aggregation method is largely irrelevant.** At R=7, all five aggregation methods yield ρ within $[0.446, 0.448]$—a spread of only 0.002. The median is marginally best but the difference is negligible.
3. **Predicted > crystal.** On crystal structures, the energetic channel achieves $\rho\approx$ 0.367 at R=7 (vs. 0.448 on predicted), consistent with the main-text finding of higher point estimates on predicted inputs (Section 6.6).

C.19.3 Reproducibility

Sensitivity analyses are implemented in the released supplementary analysis code.

C.20 Additional experimental protocols

Engineering use and extension scope. In an experimental screening workflow, the initial ranks select a small single-substitution library before target labels exist. In directed evolution, they can prioritize position–residue choices for the first library, while later rounds must learn interactions among measured combinations. For active learning, the ranks supply a round-zero prior rather than an uncertainty or acquisition function; end-to-end campaign benefits remain future work. Energetic and geometric channels can in principle be reconstructed for other ligand classes, but the ProteinMPNN channel requires a protein backbone and does not directly transfer to non-protein ligands. The LipA experiment tests a second protein property rather than validating every interaction class.

C.20.1 Screening metrics and decision thresholds

The common-method screening cohort contains 331 substitutions across 33 targets. Within each target, candidates are selected in ascending predicted score order; ties use stable ordering. A non-destabilizing hit satisfies $\Delta\Delta G \leq 0$, which includes neutral substitutions and differs from a strictly improving hit. The separate tolerated tier uses $\Delta\Delta G < 0.5$ kcal/mol. Precision at budget k is the selected hit fraction; recall is the selected hit count divided by the target's total hit count; enrichment is precision divided by that target's hit prevalence. Targets require $n \geq \max(k, 3)$ for these budget metrics. Targets without hits remain in precision but are excluded from recall and enrichment because their denominators vanish. Thus, the non-destabilizing top-three analysis has 28 targets for precision and 17 for recall and enrichment; top-five precision has 23 targets. Values are means over eligible targets rather than mutation-weighted pooled hit fractions.

For NDCG, the relevance of substitution i is $g_i = \max(0, \max_j \Delta\Delta G_j - \Delta\Delta G_i)$ within its target. The discount at rank r is $1/\log_2(r+1)$; DCG is divided by the ideal DCG, with an effective budget $\min(k, n)$. Targets with zero ideal DCG are omitted. The analyzed targets have at least three substitutions. Uncertainty uses 1,000 target-bootstrap draws, and paired comparisons use 10,000 draws with seed 20260302. Reported adjusted comparisons use the Holm procedure; nominal point-estimate advantages are not evidence of a significant gain. In particular, the tested top-k comparisons do not establish superiority over FoldX after adjustment.

Sign evaluation excludes exact-zero experimental labels and uses $\Delta\Delta G < 0$ as the positive class. AUROC is computed with the negative predicted score. The archived accuracy, sensitivity, balanced accuracy, and MCC diagnostics choose the predicted-positive count to equal the observed positive count. This prevalence-matched threshold uses experimental labels and is not an available deployment rule for an unmeasured library. AUROC is therefore the threshold-independent sign diagnostic; rank scores alone do not specify a physically calibrated zero-energy cutoff.

C.20.2 Descriptive within-target aggregation

Correlations are calculated within each target from the archived full-cohort rank-fusion scores. This analysis removes between-target contributions to the correlation calculation, but does not re-run channel rank normalization on each target's candidate set. It should therefore be distinguished from a fresh deployment in which fusion ranks are constructed within a new library.

For a comparator, let $r_{m,t}$ be the within-target Spearman correlation for method m, and let $w_t = n_t - 3$. We report the descriptive aggregate

$$\rho_{z,m} = \tanh\left(\frac{\sum_t w_t \operatorname{atanh}(r_{m,t})}{\sum_t w_t}\right), \qquad \Delta\rho_z = \rho_{z,\text{PepDDG}} - \rho_{z,\text{reference}}.$$

Finite correlations are clipped to $[-0.999999, 0.999999]$ before the transform. Targets with three substitutions are excluded from this aggregate because their weight is zero; undefined correlations are excluded. The $n_t - 3$ weighting is a descriptive Fisher approximation for Spearman correlations, rather than an exact sampling-variance model. Each comparison uses its own common eligible targets. The two methods are transformed back separately: $\tanh(\bar{z}_A - \bar{z}_B)$ is not a difference of correlations. Intervals below resample matched targets jointly in 10,000 bootstrap draws (seed 20260302); these descriptive intervals have no multiplicity adjustment and do not replace the primary paired screening comparisons.

Comparator	Targets	PepDDG ρ_z	Reference ρ_z	$\Delta\rho_z$ [95% CI]
FoldX	25	0.606	0.502	+0.105 [-0.048, +0.265]
Rosetta	25	0.606	0.451	+0.156 [+0.025, +0.300]
MM-GBSA	22	0.615	0.466	+0.149 [+0.028, +0.267]
ESM-IF1	25	0.606	0.216	+0.391 [+0.243, +0.539]
ESM-2	25	0.606	0.027	+0.579 [+0.312, +0.814]

Table 21: Descriptive within-target aggregation on archived rank-fusion scores. Coverage varies with the reference. The FoldX difference interval includes zero; this is insufficient evidence of a difference and does not establish equivalence.

The unweighted mean and median within-target PepDDG–FoldX correlation differences are 0.076 and 0.077, respectively; PepDDG has the higher correlation on 15 of the 25 targets eligible for the weighted comparison. These quantities address a different estimand from a pooled correlation.

C.20.3 Supervised methods as contextual references

The historical supervised-reference archive has method-dependent coverage, shown explicitly in Table 22. These figures are not paired comparisons with the primary 332-substitution benchmark. Methods trained with SKEMPI-derived labels also have a different supervision and potential overlap setting from PepDDG's zero-shot use. Accordingly, the table supplies context without establishing a clean out-of-target or zero-shot superiority claim.

C.20.4 Transfer to a thermostability phenotype

The *Bacillus subtilis* lipase A scan of Nütschel et al. [32] provides 2,155 single substitutions over 167 positions with all transferred channels available. The label is T_{50}, the temperature at which residual enzymatic activity falls to 50% after a 20-minute heat challenge. This phenotype is not folding $\Delta\Delta G$. The structure is 1ISP, with DMS position equal to PDB residue number plus 31. The energetic channel is a Rosetta ref2015 folding-energy proxy obtained by mutation and repacking an 8 Å shell; the geometric channel uses monomer burial and mutation volume change; the evolutionary

Method	Substitutions	Targets	Pooled ρ	Within-target ρ
ProMIM	340	34	0.515	0.551
RDE-Network	340	34	0.468	0.515
PepDDG (reference archive)	340	34	0.612	0.514
StaB-ddG	371	36	0.638	0.448
DiffAffinity	448	40	0.515	0.444

Table 22: Historical reference correlations on each method's own coverage. The within-target column is the archived per-target summary, not the Fisher aggregate in Table 21.

channel uses monomer-conditioned ProteinMPNN. These are property-specific replacements for the binding channels, with orientation chosen so that higher scores predict higher T_{50}.

Energetic, geometric, and evolutionary Spearman correlations are 0.375, 0.370, and 0.542; three-channel fusion reaches 0.557. FoldX scores all 2,155 substitutions without failure and reaches 0.424. Fusion's gain over the strongest single channel is 0.015, with a position-bootstrap 95% interval $[-0.035, 0.065]$, which includes zero. Geometric–evolutionary fusion alone reaches 0.557 (0.556868 versus 0.556949 for all three channels). Thus, the experiment demonstrates a transferable construction, while providing no detected three-channel gain over the best single channel and little point estimate gain from the third channel on this assay. It neither establishes a property-general advantage nor identifies a biological mechanism.

C.20.5 Using a few measured anchors for quantitative prediction

This extension first learns an isotonic rank-to-energy map from labeled substitutions in other targets under leave-one-target-out evaluation. It then uses measured substitutions from the held-out target to estimate an offset as the median anchor residual. Both stages use experimental labels; this is a supervised quantitative extension of the ranker.

Eighteen of the 33 targets support reserving five anchor candidates while retaining at least three evaluation substitutions, yielding 185 evaluation substitutions per split across the usable targets. For each target, 60 random evaluation/anchor-pool splits are generated; 40 random anchor draws are used per split. The five-candidate anchor pool is excluded from evaluation, and the evaluation set is fixed across anchor budgets within each split. With seed 20260302, offset-only RMSEs for zero, one, two, three, four, and five anchors are 2.072, 1.908, 1.658, 1.586, 1.515, and 1.495 kcal/mol. Three anchors reduce RMSE by approximately 23% relative to the zero-anchor isotonic prediction; the training-set-mean reference has RMSE 2.777 kcal/mol. Estimating both offset and scale is unstable at small budgets (RMSE 9.902 with two anchors and 3.814 with three), supporting the simpler offset-only extension for this diagnostic. These results do not establish prospective performance on a newly discovered complex.

C.20.6 Runtime scope and measurement conditions

Direct PepDDG–FoldX timing uses identical complex inputs on a shared RTX 4090 host under four-way contention. For 3EQS (10 substitutions), complete PepDDG scoring takes 193.8 s (19.38 s/substitution) and FoldX takes 159.1 s (15.91 s/substitution). For 5XCO (nine substitutions), the respective costs are 283.3 s (31.47 s/substitution) and 352.3 s (39.14 s/substitution). PepDDG totals cover all three channels. The median across the two per-complex rates is 25.42 s/substitution for PepDDG and 27.52 for FoldX; this is a two-complex diagnostic, not an independently timed median over the full benchmark. Contention affects both measurements and these costs do not establish a speed advantage over FoldX.

The Rosetta estimate is 907 s/substitution with required relaxation included and uses cross-host timing calibration. The Flex ddG estimate is approximately 23,500 s/substitution, obtained by scaling probes to the full sampling protocol; it is not direct wall time for a completed full-protocol benchmark. BA-DDG has no corresponding end-to-end timing. Ratios involving these estimates should retain these different measurement boundaries; model loading, structural prediction, and shared-host conditions must be specified when transferring costs to another deployment.

C.20.7 Structure sources, matched comparisons, and topology failures

The original prepared structures are unavailable, and the retained feature tables contain conflicting crystal and predicted-input source annotations. The primary result therefore evaluates frozen scores rather than a verified all-crystal reconstruction. The matched experiment below is separate and does not recover the missing full-cohort structural inputs.

The revised matched comparison contains 317 substitutions across 30 targets. Experimental labels from the benchmark cohort are used for both arms: 34 rows disagreed with labels in the alternate archive, with a maximum discrepancy of 1.545 kcal/mol. The same benchmark label is used for a matched substitution in both arms. Crystal and Boltz-2 Spearman correlations are 0.581 and 0.631 for fusion, 0.442 and 0.479 for the energetic channel, 0.458 and 0.475 for the geometric channel, and 0.467 and 0.518 for the evolutionary channel. The paired fusion difference is approximately 0.050, with 95% interval $[-0.025, 0.125]$. No degradation is detected on this cohort; the interval also permits a negative difference. Possible structure-model training overlap and the absence of a verified post-training-cutoff cohort prevent a claim of prospective generalization to novel complexes.

The AF3–Boltz-2 diagnostic covers five targets, with 15 AF3 and 25 Boltz-2 poses per target. It is unequal-sampling evidence, not a predictor-wide benchmark. Boltz-2 was selected for open weights and support for covalent peptide specification. For 1B41, 1FSS, and 5XCO, median AF3 interface-mode errors are 18.44, 18.27, and 11.06 Å, versus 1.72, 2.41, and 1.77 Å for Boltz-2. These target-specific differences do not demonstrate superiority across general predictor workloads.

Topology failure has separate coverage from pose generation. Among poses selected for the scoring diagnostic, 10/12 for 1B41 and 2/3 for 1FSS are unscorable. Both are acetylcholinesterase–fasciculin complexes; fasciculin has four disulfides within 61 residues. The predicted sulfur arrangement causes the OpenMM preparation to reject the topology. The 1B41 failures occur for both predictors; the three 1FSS scoring attempts in this archive are AF3 poses. Missing scores are failures rather than low correlations and must not be silently dropped to imply support for complex disulfide networks.

Finally, controlled rigid peptide displacements span 44 poses over 1XD3, 3EQS, and 5XCO, with within-target RMSD ranging from 0.5 to 8 Å. The best-powered target, 3EQS (24 poses), has a PepDDG-minus-FoldX degradation slope difference of 0.190, with bootstrap interval $[0.074, 0.349]$. The 5XCO interval includes zero, and 1XD3 has only seven poses. Across all 44 fixed poses PepDDG has the higher correlation on 16; the archived sign test gives $p = 0.096$. A slower deterioration on 3EQS therefore does not establish a general accuracy advantage at arbitrary pose error. Uncontrolled cross-target pose collections additionally confound pose quality with target identity and are not substitutes for this controlled comparison.

D Additional figures and tables

D.1 Supplementary scatter plots

Figures 5 and 6 provide the pooled and per-target scatter diagnostics underlying the supplementary mutation-level comparisons.

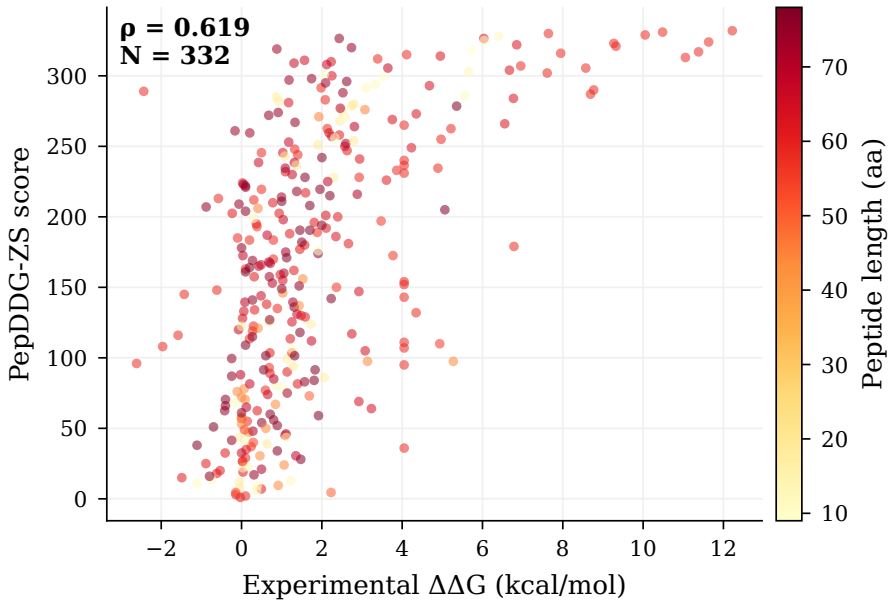


Figure 5: **Pooled mutation-level scatter (supplementary).**

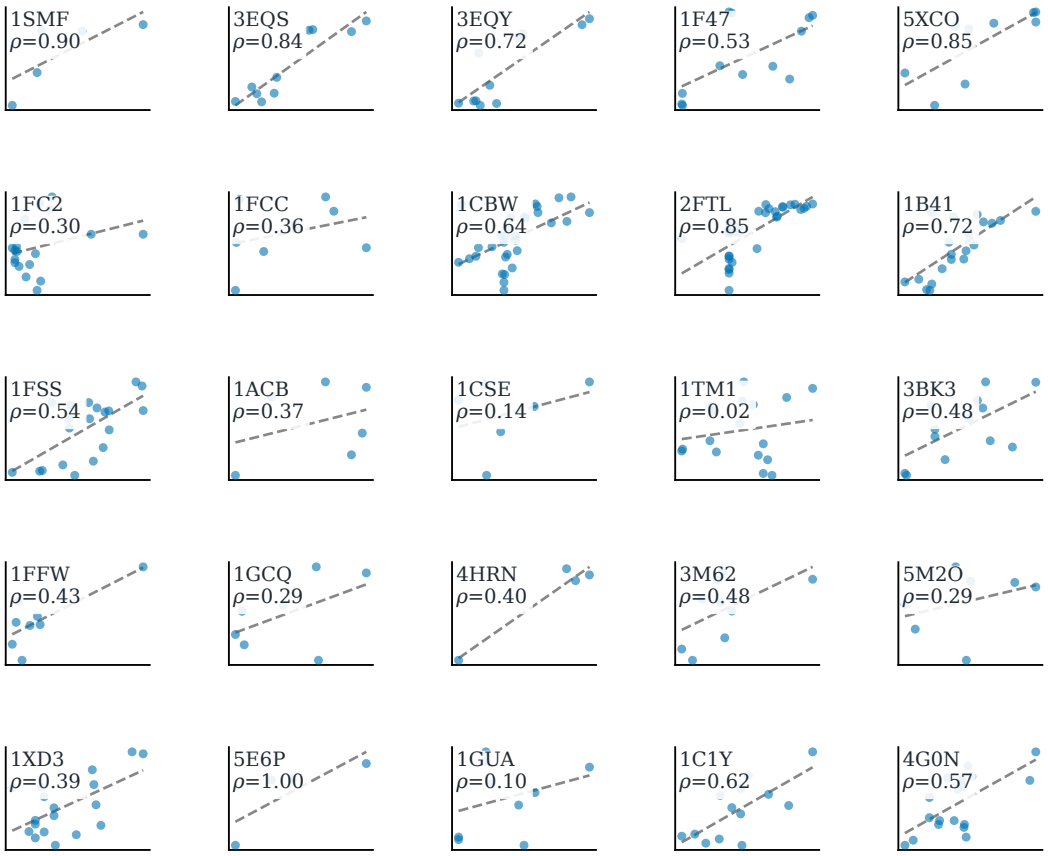
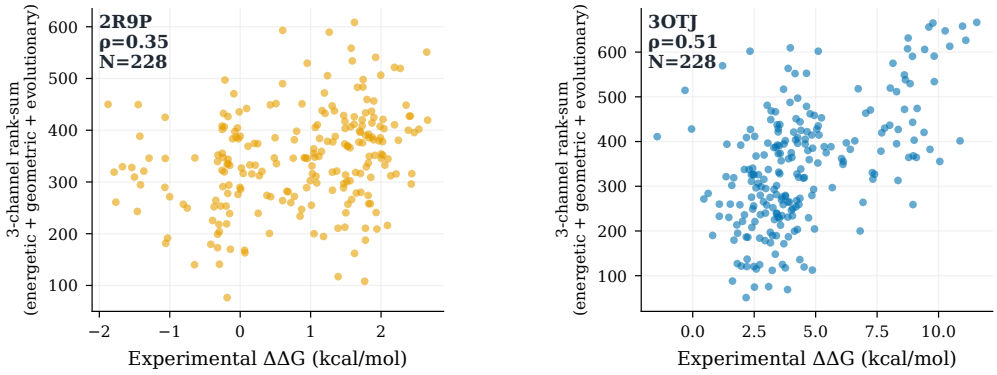


Figure 6: **Per-target scatter panels (supplementary).**

D.2 Independent validation plots

Figure 7 summarizes the independent BPTI validation targets used for the out-of-distribution stress test.



(a) 2R9P. (b) 3OTJ.

Figure 7: **BPTI DMS per-target scatter (supplementary).** 3-channel rank-sum (energetic + geometric + evolutionary) on the two OOD targets (2R9P, 3OTJ), matching the OOD stress test in Appendix C.14. We exclude 1CBW due to overlap with PepDDG-Bench.

D.3 Crystal vs. predicted structures (per-target)

The reconstructed comparison contains 317 mutations across 30 targets, using the same experimental label in both arms. The per-target table reports the 23 targets with at least four mutations (301 mutations); seven smaller targets (16 mutations) contribute to the 317/30 pooled comparison but are omitted from this per-target diagnostic.

Table 23: Matched per-target structure diagnostic: 301 mutations, 23 targets with $N \geq 4$.

Target	N	Crystal ρ	Predicted ρ	Target	N	Crystal ρ	Predicted ρ
1ACB	6	0.143	0.371	1SMF	5	0.900	0.900
1B41	20	0.645	0.725	1TM1	17	0.121	0.001
1C1Y	13	0.369	0.624	1XD3	18	0.306	0.391
1CBW	31	0.140	0.621	2FTL	31	0.778	0.846
1CSE	6	0.086	0.143	3BK3	13	0.495	0.503
1F47	12	0.566	0.559	3EQS	10	0.782	0.855
1FC2	18	0.626	0.297	3EQY	11	0.700	0.718
1FCC	8	0.738	0.357	4G0N	18	0.428	0.666
1FFW	8	0.524	0.429	4HRN	4	0.200	0.400
1FSS	20	0.460	0.504	5M2O	7	-0.071	0.286
1GCQ	7	0.679	0.286	5XCO	9	0.803	0.828
1GUA	9	-0.042	0.101				

D.4 Per-target channel performance heatmap

Figure 8 breaks down channel and aggregate performance by target.

Per-target observations. The heatmap reveals several patterns: (i) PepDDG is not uniformly best on every target—it is the top-performing method on 11/23 eligible targets (counting ties; 10/23 unique wins, with $N \geq 5$), while individual channels win on the remaining targets. (ii) For targets where a single channel is dominant (e.g. 3EQS: Evolutionary $\rho = 0.92$), rank aggregation still preserves most of the signal because the weaker channels add moderate noise but do not cancel the strong one. (iii) The short-peptide block (blue box) shows consistently high correlations across all channels, supporting the length-dependent capacity analysis in Figure 11 (bottom-left panel).

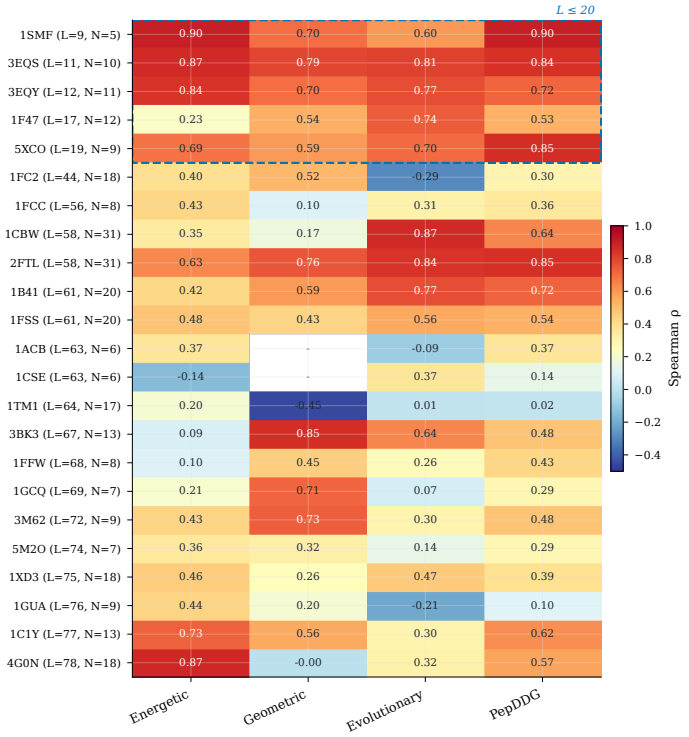


Figure 8: **Per-target Spearman ρ for individual channels and PepDDG.** Targets are sorted by peptide length (ascending). Each cell shows ρ for targets with $N \geq 5$ mutations. The dashed blue box highlights short-peptide targets ($L \leq 20$), where all channels and PepDDG achieve their highest correlations. Target labels include peptide length (L) and mutation count (N).

D.5 Channel performance by mutation type

Figure 9 shows how individual channel performance varies across mutant amino-acid classes.

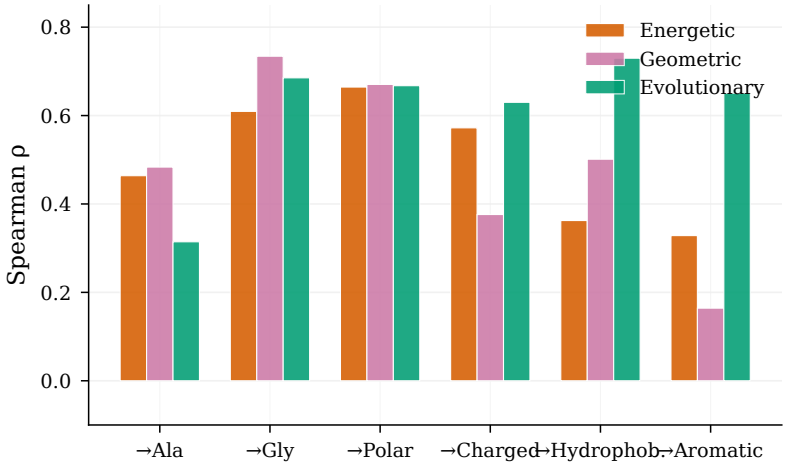


Figure 9: **Single-channel performance stratified by mutant amino-acid class.** Grouped bar chart showing Spearman ρ for energetic, geometric, and evolutionary channels within each mutation type. Different channels dominate different regimes: geometric for to-Ala/to-Gly, energetic for to-polar, evolutionary for to-hydrophobic and to-aromatic. See Table 2 for exact values including the PepDDG aggregate.

D.6 Error analysis: rank parity plot

Figure 10 localizes the largest rank-disagreement cases in the common coverage split.

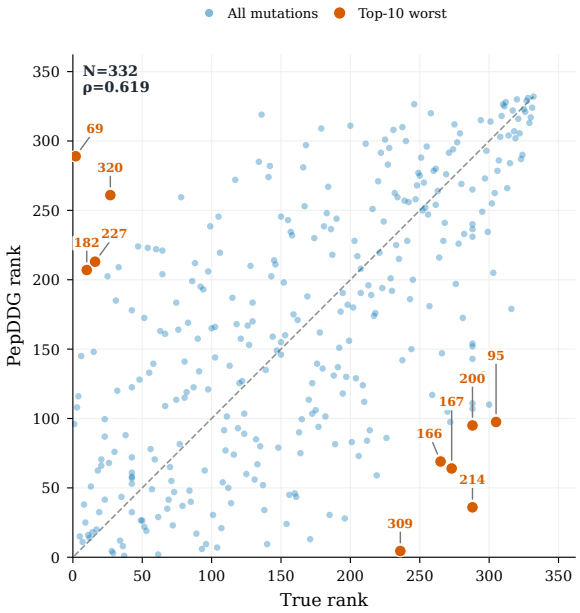


Figure 10: **Rank parity plot with worst-predicted mutations.** True rank vs PepDDG rank for all 332 mutations in the common-coverage split. Points near the diagonal are well-predicted; the top-10 worst-predicted mutations (red) are annotated. Systematic off-diagonal deviations indicate mutation types or targets where channel complementarity is insufficient.

Error case characterization. The worst-predicted mutations tend to cluster in two failure modes: (i) **long-peptide targets with flexible interfaces**, where all channels have lower signal-to-noise ratios, and (ii) **mutations that alter binding mode** (e.g. charge reversals at critical salt bridges), which violate the noisy-monotone assumption (Definition 1) because the mutant adopts a qualitatively different binding geometry. These failure modes suggest that a future "conformational" channel—capturing binding-mode changes—could provide complementary information not present in the current three channels.

D.7 Additional theory and redundancy diagnostics

Figures 11 and 12 collect the supplementary theory-validation and redundancy diagnostics.

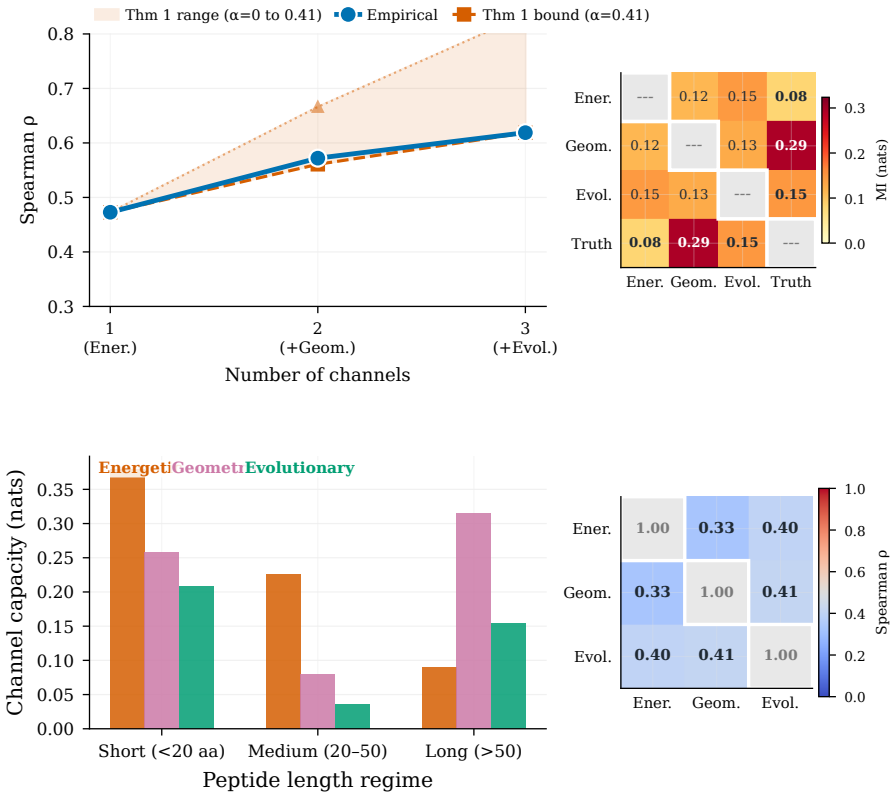


Figure 11: **Supplementary theory-validation diagnostics.** Panels include mutual-information and peptide-length capacity diagnostics.

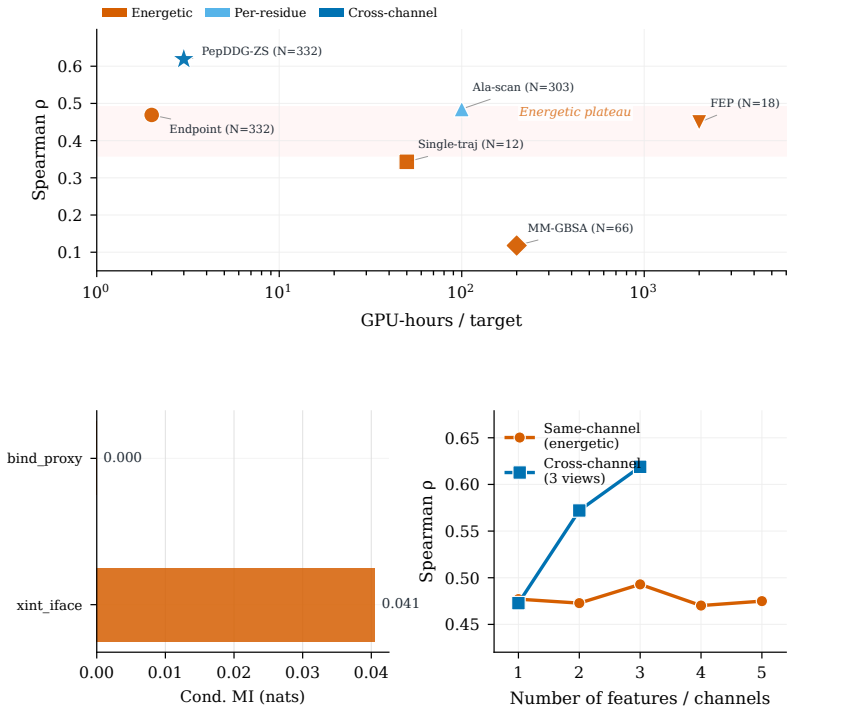
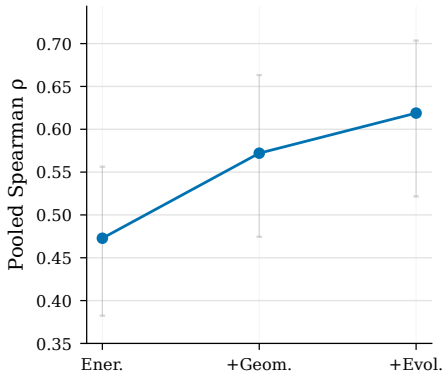


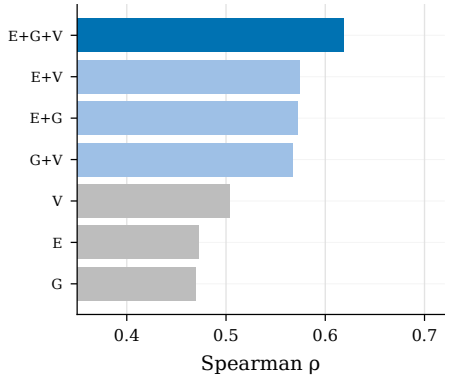
Figure 12: **Supplementary redundancy diagnostics.** Same-channel refinement panels include residual-rank-signal and energetic-saturation diagnostics.

D.8 All channel subset ablation

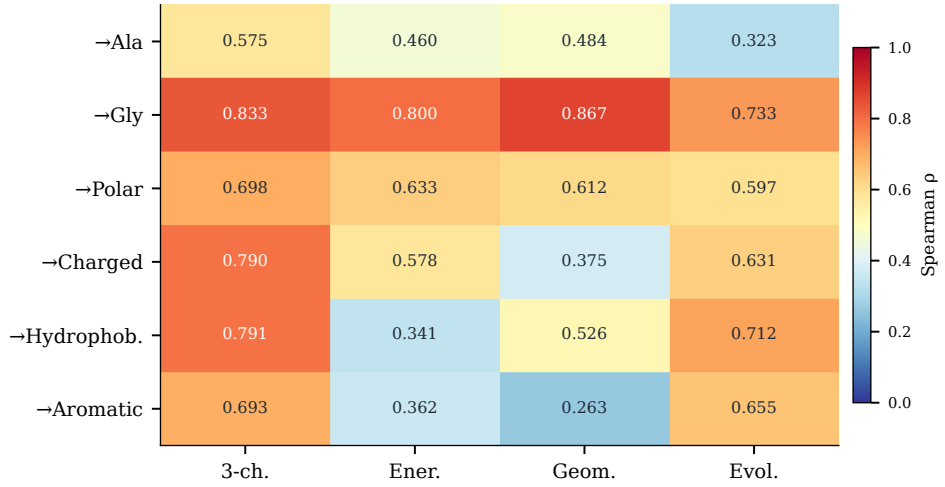
Figure 13 and Table 24 summarize all channel-subset ablations.



(a) Cumulative channel addition from energetic-only to PepDDG (3-channel).



(b) Top non-empty channel subsets by pooled ρ.



(c) Channel performance stratified by mutation type (different channels dominate in different regimes).

Figure 13: **Ablations support multi-channel complementarity.**

Table 24: **All non-empty channel subsets ranked by pooled ρ.** Common-coverage split (N=332, T=33).

Channel combination	K	ρ	95% CI low	95% CI high	N
Energetic + Geometric + Evolutionary (PepDDG)	3	**0.619**	0.519	0.708	332
Energetic + Evolutionary	2	0.574	—	—	332
Energetic + Geometric	2	0.572	—	—	332
Geometric + Evolutionary	2	0.568	—	—	332
Evolutionary only	1	0.504	—	—	332
Energetic only	1	0.473	—	—	332
Geometric only	1	0.469	—	—	332

Monotonic gains from channel addition. Table 24 shows that (i) all three single channels have comparable correlation (ρ = 0.469–0.504), supporting the equal-weight assumption in our Borda aggregation; (ii) every two-channel combination outperforms every single channel; (iii) the 3-channel model outperforms all two-channel subsets; and (iv) all three pairwise combinations yield similar

performance (ρ = 0.568–0.574), indicating that no single pairing is disproportionately synergistic. This monotonic improvement pattern is consistent with the rank-covariance diagnostic in Proposition 3. PepDDG-Cal is reported separately in Table 1 because it is a calibrated ablation rather than a channel-subset ablation.

D.9 Channel summary table

Table 25 defines the channel-to-feature mapping used throughout the main paper and appendix.

Table 25: **PepDDG channels and corresponding features.** The main paper uses the first three channels (full coverage). The optional per-residue channel is explored in the appendix.

Channel	Scientific question	Method	Feature(s)
Energetic perturbation	Binding energy change?	Implicit-solvent scoring	`ddg_paired_xint_iface`, `ddg_paired_bind_proxy`
Geometric environment	Position importance?	Contacts + packing	`struct_composite`
Evolutionary compatibility	AA compatibility?	ProteinMPNN cycle	`neg_llr`, `ddg_bind`
Per-residue contribution (optional)	Residue importance?	Alanine scanning	`mmpbsa_ddg_ala`

D.10 Mutual information matrix

Table 26 reports the pairwise mutual information estimates between rank-transformed channels and the ground-truth ranking.

Table 26: **Pairwise mutual information (nats) between rank-transformed channels and ground truth.** Estimated via a KNN estimator (k=5; 10 random seeds).

	Energetic	Geometric	Evolutionary	Ground truth
Energetic	—	0.116	0.148	0.084
Geometric	0.116	—	0.129	0.294
Evolutionary	0.148	0.129	—	0.150
Ground truth	0.084	0.294	0.150	—

D.11 Experimental budgets and controlled structural perturbations

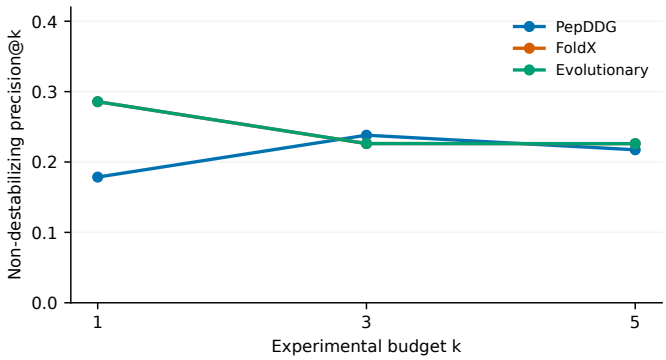


Figure 14: Non-destabilizing precision under experimental budgets of one, three, and five substitutions. Eligible targets differ with budget (Appendix C.20.1); the curves are retrospective averages. Point-estimate differences from FoldX do not establish a significant gain after multiple-testing correction.

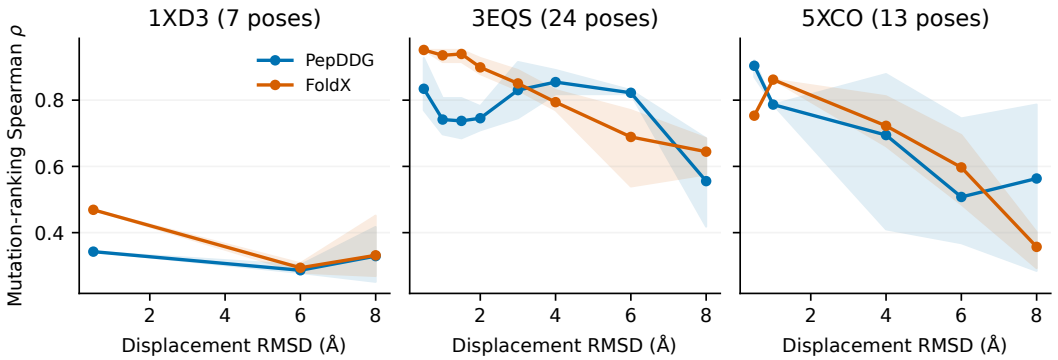


Figure 15: Mutation-ranking correlations under controlled rigid peptide displacements: all 44 poses across three targets. Lines denote means at each displacement, and shading shows the observed range, not confidence intervals. A slower degradation slope is detected for 3EQS; the other targets do not establish the same advantage (Appendix C.20.7).

E Reproducibility and Artifact Availability

Environment. All Python analyses are run with Python 3.12. Figure generation, statistical tests, and benchmark-table construction use the same locked dependency set as the evaluation pipeline. LaTeX compilation assumes a modern TeX Live installation.

Table 27: **Key software versions in our evaluation environment.**

Package	Version	Used for
Python	3.12.7	analysis + figure generation
OpenMM	8.2	implicit minimization + scoring
PDBFixer	1.12.0	structure fixing + mutation application
PyTorch	2.6.0	ProteinMPNN inference
NumPy / SciPy / pandas	1.26.4 / 1.13.1 / 2.3.3	metrics + bootstraps

Evaluation protocol. All headline metrics are computed on fixed mutation cohorts before model comparison. The primary in-domain result uses the common-coverage PepDDG-Bench split with complete channel and comparator availability (332 peptide-chain mutations across 33 targets). The OOD analysis reported in the paper uses two independent BPTI DMS targets (456 mutations total). Confidence intervals are computed by target-stratified bootstrap so that targets, rather than individual mutations, are the resampling unit.

Baseline processing. For supervised baselines, we align each method's prediction dump to the fixed cohort at the mutation level (by `(target, mut)` when available, or by a chain-independent mutation key for methods that report `mutstr`). If a baseline provides multiple predictions for the same mutation, we average `ddG_pred` across duplicates before computing pooled metrics.

Artifact release and replay boundaries. The public research release is available at `https://github.com/zhangruochi/pepddg-release`, with the recorded software release v0.1.2. It includes a direct wild-type-complex/mutation-list Python API, a CLI, environment installation instructions, and a four-target structural reproduction example. New code and documentation are provided under the repository's noncommercial license; commercial use requires authorization. Historical MIT-covered code and third-party data retain their stated rights.

The following commands run from the installed release repository:

```
python -m pepddg.frozen_benchmark --repo-root . --output /tmp/pepddg-frozen
bash examples/skempi_cyclic/run.sh /tmp/pepddg-structural
```

The first command replays the 332/33 frozen feature benchmark; it verifies scoring arithmetic, not historical structural execution. The second performs fresh structural inference for 35 mutations on four prepared official crystal complexes: two disulfide systems and two linear PMI controls. Seven paired WT/mutant restarts are used. Both examples distinguish inference inputs from experimental labels used only in evaluation. Full 332/33 fresh structural reproduction has not been completed, and original prepared structural inputs are unavailable. The small-cohort consistency exercise does not prove full-benchmark reproduction or experimental affinity.

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: The abstract and introduction state the zero-shot rank-fusion method, the peptide–protein evaluation setting, the theoretical analysis, and the main empirical results. The claims are scoped to peptide–protein $\Delta\Delta G$ ranking, with assumptions and limitations discussed in the limitations and broader impact section.

   Guidelines:

   - The answer [N/A] 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 [N/A] 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: The paper includes a dedicated limitations and broader impact section discussing reliance on predicted structures, weaker performance on stabilizing mutations, distribution-shift behavior, single-point mutation scope, pretrained-component biases, and assumptions behind the channel model.

   Guidelines:

   - The answer [N/A] 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: [Yes]

   Justification: The theoretical results state their assumptions in the multi-channel noise model and theorem statements. Complete proofs and extensions are provided in the theory appendix.

   Guidelines:

   - The answer [N/A] 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 paper describes dataset construction, common-coverage evaluation, channel computations, rank aggregation, metrics, target-stratified bootstrap, and comparator alignment. Additional protocol and environment details are provided in the appendices.

   Guidelines:

   - The answer [N/A] 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: The paper provides open access to the data and code through the public research repository at `https://github.com/zhangruochi/pepddg-release`, with instructions for frozen-feature replay and a four-target structural example. Full structural regeneration of the 332-mutation benchmark is unavailable; this boundary is documented in Appendix E.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.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://neurips.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) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies the data split, metrics, bootstrap procedure, channel definitions, OpenMM minimization protocol, ProteinMPNN scoring, Boltz-2 structure prediction use, and rank aggregation procedure in the method and experimental sections and appendices.

   Guidelines:

   - The answer [N/A] 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: Main results report target-stratified 95% confidence intervals, and paired application-oriented comparisons distinguish adjusted significance from descriptive intervals; their cohorts and timing boundaries are reported in Appendix C.20.

Guidelines:

- The answer [N/A] 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: The paper reports per-target runtime and hardware for the main components, including Boltz-2 structure prediction, OpenMM energy scoring, geometry scoring, ProteinMPNN scoring, and rank aggregation.

   Guidelines:

   - The answer [N/A] 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: The work uses public protein-structure and mutation data, reports limitations and dual-use considerations, and does not involve private data or human subjects. We have reviewed the NeurIPS Code of Ethics and believe the work conforms to it.

   Guidelines:

   - The answer [N/A] 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: The limitations and broader impact section discusses potential benefits for therapeutic peptide discovery and possible misuse for designing harmful binders, together with a recommendation for responsible use and governance.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] 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., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release a high-risk generative model, scraped sensitive dataset, or autonomous design system; the method ranks specified point mutations for peptide–protein complexes.

    Guidelines:

    - The answer [N/A] 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: The paper cites the original sources of all existing datasets, software packages, and pretrained components used in the study, including SKEMPI v2, Boltz-2, ProteinMPNN, OpenMM, FoldX, and Rosetta. The reproducibility appendix records key software versions used in evaluation; release documentation will state asset versions and applicable license or terms-of-use details for redistribution.

    Guidelines:

    - The answer [N/A] 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: The paper introduces PepDDG-Bench and documents its construction, filtering criteria, splits, evaluation protocol, target list, and limitations in the main text and appendices.

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification: The work does not involve crowdsourcing experiments or research with human subjects.

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification: The work does not involve human subjects and therefore does not require IRB approval or equivalent review.

    Guidelines:

    - The answer [N/A] 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 rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not used as an important, original, or non-standard component of the core scientific method; any ordinary writing or formatting assistance does not affect the methodology or scientific claims.

    Guidelines:

    - The answer [N/A] 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 in the NeurIPS handbook for what should or should not be described.