AFFINITYSYM: SEQUENCE-BASED INTERPRETABLE PROTEIN–PROTEIN BINDING AFFINITY PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Protein–protein binding affinity, often expressed as a change in Gibbs free energy (ΔG), quantifies the stability of a complex. Most predictors require a three-dimensional structure and return a scalar without explaining the estimate. AffinitySym is a multimodal, sequence-first deep neural model with frozen ESM-3 and MINT encoders. Bidirectional cross-attention combines partner sequences with additional equivariant graph neural network (EGNN) and invariant point attention (IPA) features when a structure is supplied; learned-query attention conditions an autoregressive Transformer decoder that generates 16 executable symbolic formulas. Dual-path training lets the same checkpoint predict from partner sequences alone or from those sequences paired with a supplied structure, making AffinitySym the only compared method to support both inference modes. Formula-conditioned residue attributions identify amino acid residues that contribute to each prediction. We train on the protein–protein subset of PDBbind v2020 and evaluate on two test sets adopted from prior studies: Test Set 1 (82 complexes) and Test Set 2 (166 wild-type and mutant complexes from SKEMPI v2.0). On Test Set 1, the best-observed sequence-only formula achieves an MAE of 1.2982 kcal/mol and a PCC of 0.7034, compared with an MAE of 1.36 kcal/mol and a PCC of 0.629 for the strongest listed baseline. The same checkpoint also performs competitively on Test Set 2, while a separate checkpoint achieves an MAE of 0.9960 kcal/mol and a PCC of 0.8493. Detailed comparisons on both test sets and multi-seed results are presented in the main paper. Frozen encoders reduce training and gradient-memory costs, while local inference lets researchers keep sensitive protein data on their own systems. Residue-level attributions can help prioritize binding-critical residues for drug discovery.

Code Availability (Anonymous): `https://anonymous.4open.science/r/AffinitySymProject-D964`

1 INTRODUCTION

1.1 BACKGROUND

Proteins carry out many cellular functions by binding to other proteins. These protein–protein interactions (PPIs) assemble molecular machines, transmit signals, regulate transcription, and control immune responses (Jones & Thornton, 1996; Akbarzadeh et al., 2024). A PPI is not fully described by whether two proteins make contact. Its binding affinity (BA) also matters because it measures how strongly the complex is favored over the unbound proteins (Siebenmorgen & Zacharias, 2020; Kastritis et al., 2011). Accurate BA estimates are useful in interaction analysis, therapeutic design, and protein engineering, where a change in affinity can alter biological function.

PPI binding affinity is commonly expressed through the Gibbs free-energy change, ΔG. For a dissociation reaction $PL \rightleftharpoons P + L$, the equilibrium dissociation constant is $K_d = [P][L]/[PL]$. In the datasets used here, K_d is converted to molar units and the regression target is computed at temperature T as

$$\Delta G = RT \ln(K_d), \tag{1}$$

where R is the gas constant. A smaller K_d indicates tighter binding and gives a more negative ΔG; a larger K_d indicates weaker binding. Thus, an error of a few kcal/mol can correspond to a large error in the inferred equilibrium constant because the relation is logarithmic.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Binding affinity can be measured by techniques such as isothermal titration calorimetry, surface plasmon resonance, and microscale thermophoresis (Day et al., 2002; Winiewska-Szajewska & Poznański, 2025). These experiments provide valuable reference data but require purified samples, suitable assay conditions, and specialized equipment. Molecular dynamics and free-energy perturbation offer computational alternatives, but their cost limits screening at scale (Hollingsworth & Dror, 2018; Hansson et al., 2002; Jorgensen & Thomas, 2008; Liu et al., 2012). Machine-learning models instead learn a mapping from protein representations to measured affinity. Their practical value depends on what input they require and whether a researcher can interpret the prediction.

Figure 1 uses the 1WR1 (Ohno et al., 2005) complex to relate reversible protein association to K_d, ΔG, and the energy difference between the unbound and bound states.

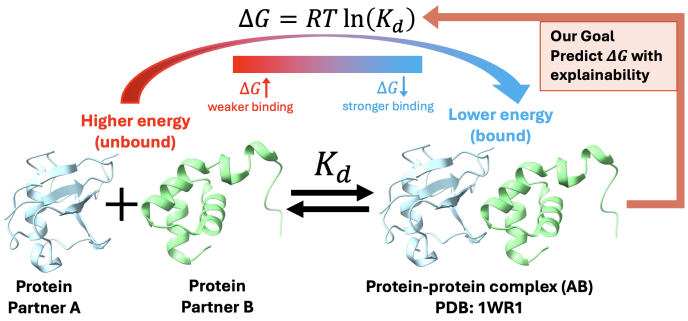

Figure 1: Conceptual view of protein–protein binding using the 1WR1 (Ohno et al., 2005) complex. Protein partners A and B reversibly associate to form complex AB; a smaller K_d and a more negative ΔG indicate stronger binding and a lower-energy bound state. The prediction task is to estimate ΔG while explaining the result.

1.2 Related Work and Current Limitations

Early structure-based methods estimate PPI affinity using statistical potentials or engineered descriptors. DFIRE learns an energy function from structural distributions, PRODIGY uses interfacial contacts, and PPI-Affinity applies support vector machines to protein complexes (Liu et al., 2004; Vangone & Bonvin, 2015; Xue et al., 2016; Romero-Molina et al., 2022). Neural predictors such as ProBAN, ProAffinity-GNN, and Multi_PPBA instead learn from spatial, graph, and multimodal representations (Bogdanova & Novoseletsky, 2024; Zhou et al., 2024; Mo et al., 2025). On Test Set 1, their reported MAE/PCC values range from 1.75/0.55 for ProBAN to 1.36/0.629 for Multi_PPBA. On Test Set 2, Multi_PPBA reports 1.20/0.782, compared with 1.36/0.722 for ProAffinity-GNN. Section 4 gives the full comparisons and cohort details.

Sequence-based methods reduce dependence on complex structures. ISLAND uses sequence-derived features, and later work has applied convolutional and Transformer models to ΔG prediction (Abbasi et al., 2020; Chen et al., 2024). Protein language models also provide residue representations that transfer to downstream biological tasks (Alley et al., 2019; Elnaggar et al., 2022). However, the strongest affinity predictors remain largely structure-dependent. Experimental complexes are not available for every protein pair, and errors in predicted chain placement or interface geometry can propagate into the affinity estimate (Jumper et al., 2021; Abramson et al., 2024).

Most neural affinity predictors also return a scalar without exposing how their models produce it. Post hoc attribution can rank influential features or residues, but it does not provide an executable form of the prediction (Samek et al., 2021). Sequence input addresses structural availability, not this lack of transparency. These limitations motivate a model that predicts from sequence, accepts structure as an additional input, and reports the mathematical relationships behind each estimate.

1.3 Our Approach

AffinitySym takes the amino acid sequences of both interacting protein partners as input and can incorporate a supplied complex structure as an additional input. Sequence-only inference supports cases in which the partner sequences are known but a reliable complex structure is unavailable; in

this setting, the model must infer the relevant bound state from sequence and cannot distinguish alternative conformations or binding arrangements of the same partners. When a structure is supplied, the model instead conditions its prediction on a particular chain assembly, interface, and residue geometry. The structure branch is therefore included to broaden the supported inputs and enable conformation-aware prediction, rather than solely to improve aggregate benchmark accuracy. Frozen ESM-3 (Hayes et al., 2025) and MINT (Ullanat et al., 2026) encoders represent the paired sequences. EGNN and OpenFold IPA branches process a residue-level k-nearest-neighbor graph derived from the supplied structure (Satorras et al., 2021; Ahdritz et al., 2024; Jumper et al., 2021). The fusion network combines the available representations and maps them to latent concept vectors. A symbolic decoder then generates 16 expert formulas with different expression budgets (Petersen et al., 2021). Each expert produces its own affinity estimate, and validation performance determines the top-K consensus used for the final prediction. Auxiliary three-class and 44-bin ordinal heads provide additional views of the predicted affinity range. The model also reports the learned formulas, expert-specific predictions, and residue-level contributions.

The contributions of this work are:

- **Two input modes in one trained model.** AffinitySym is the only method in our comparison that supports both sequence-only and sequence–structure inference with the same checkpoint. When a complex structure is available, the graph branch adds conformation- and interface-specific geometric information; when it is unavailable, the sequence path remains operational without retraining.
- **Symbolic and residue-level explanations.** The model returns multiple executable formulas, their individual predictions, a validation-selected consensus, and residue-level importance maps that link individual residues to the predicted ΔG, providing interpretable references for identifying binding-critical residues and guiding drug design.
- **Open and parameter-efficient implementation.** AffinitySym, ESM-3, and MINT are open source. Freezing the pretrained encoders limits optimization to approximately 32 million trainable parameters and avoids the cost of end-to-end backbone training.
- **State-of-the-art performance.** AffinitySym combines sequence-only affinity prediction with symbolic and residue-level explanations. Its best-observed checkpoint outperforms the listed methods on Test Set 1 and remains competitive on Test Set 2.

2 METHODOLOGY

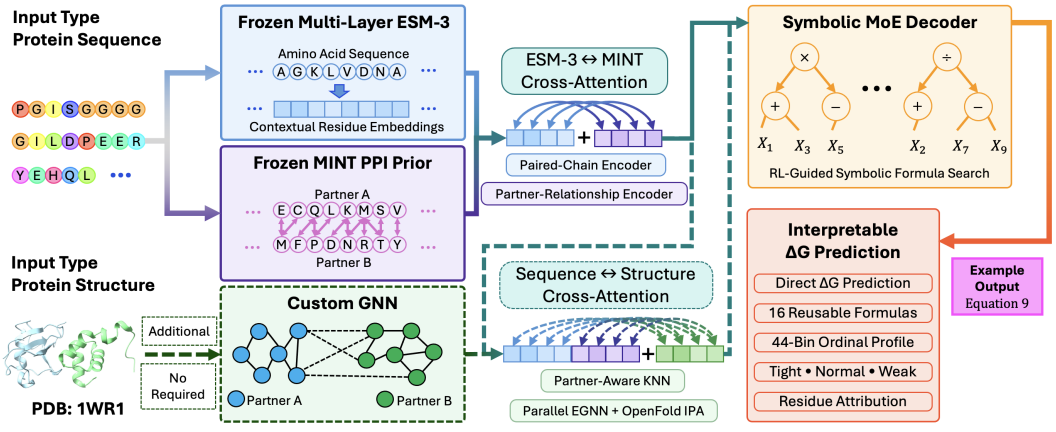

Figure 2: Model overview of AffinitySym. The displayed ΔG output denotes the scalar predictions of the symbolic experts; the active model does not use a separate direct-affinity head.

Figure 2 shows the two sequence encoders, the additional structure branch, and the symbolic mixture-of-experts decoder. The same checkpoint accepts either paired sequences or paired sequences with a complex structure. Appendix A gives the layer and loss details.

2.1 Problem Definition

For a receptor chain set $\mathcal{S}_R$, ligand chain set $\mathcal{S}_L$, residue graph $\mathcal{G}$, and structure-availability indicator $m_G \in \{0, 1\}$, AffinitySym learns

$$f_\theta(\mathcal{S}_R, \mathcal{S}_L, \mathcal{G}, m_G) \mapsto \left(\hat{y}, \{\hat{y}_e, \tau_e\}_{e=1}^{16}, \mathbf{p}_3, \mathbf{p}_{44}\right), \tag{2}$$

where θ denotes the trainable parameters, $\hat{y}$ is the reported consensus, τ_e and $\hat{y}_e$ are the formula and prediction of expert e, and $\mathbf{p}_3$ and $\mathbf{p}_{44}$ are auxiliary affinity distributions. The target $y = \Delta G$ is computed from measured K_d by Equation 1; K_d is not an input. The sequence-only path sets $m_G = 0$, whereas the sequence–structure path uses the graph of a supplied complex. Without that graph, the model cannot distinguish alternative binding arrangements formed by the same sequences.

2.2 Sequence and Structure Encoders

Frozen ESM-3 and MINT encoders provide complementary sequence representations (Hayes et al., 2025; Ullanat et al., 2026). ESM-3 produces contextual residue tokens, which a partner-aware encoder updates through chain-local and bidirectional cross-attention. For pooled partner vectors $\mathbf{r}, \mathbf{l} \in \mathbb{R}^{384}$, the order-invariant pair summary is

$$\mathbf{z}_E = \phi_E([\mathbf{r} + \mathbf{l}; |\mathbf{r} - \mathbf{l}|; \mathbf{r} \odot \mathbf{l}]), \qquad \mathbf{t}_{M,o} = \phi_M(\mathbf{m}_o), \quad o \in \{R \to L, L \to R\}. \tag{3}$$

Here, $[\cdot;\cdot]$ denotes concatenation, $\odot$ denotes element-wise multiplication, and ϕ_E and ϕ_M are trainable shared projections. The vectors $\mathbf{m}_{R\to L}$ and $\mathbf{m}_{L\to R}$ are MINT pair embeddings for the two partner orders. The shared map ϕ_M produces the two orientation tokens $\mathbf{t}_{M,o}$. Cross-attention between ESM-3 residue tokens and this two-token set, followed by symmetric pooling, yields the order-invariant sequence representation $\mathbf{z}_S \in \mathbb{R}^{384}$.

When structure is supplied, a partner-aware residue graph with 37-dimensional node and 10-dimensional edge features is processed in parallel by EGNN and OpenFold IPA branches (Satorras et al., 2021; Ahdritz et al., 2024; Jumper et al., 2021). Let $\bar{\mathbf{h}}_S$ and $\bar{\mathbf{h}}_G$ denote the pooled, cross-attended sequence and graph streams. The graph and fused representations are

$$\begin{aligned} \mathbf{H}_G &= \phi_G([\mathbf{H}_{\mathrm{EGNN}}; \mathbf{H}_{\mathrm{IPA}}]), \\ \mathbf{z} &= \mathrm{LN}\big(\mathbf{z}_S + \phi_F([\bar{\mathbf{h}}_S; \bar{\mathbf{h}}_G; \mathbf{q}_G])\big). \end{aligned} \tag{4}$$

Here, $\mathbf{H}_{\mathrm{EGNN}}$ and $\mathbf{H}_{\mathrm{IPA}}$ are the two node-level outputs, ϕ_G projects their concatenation to 384-dimensional graph tokens, and ϕ_F projects the fused inputs. The operator LN is layer normalization, and $\mathbf{q}_G \in \mathbb{R}^4$ records residue coverage, occupancy, temperature-factor quality, and the inter-partner edge fraction. If $m_G = 0$, the graph branch is skipped and $\mathbf{z} = \mathbf{z}_S$. Structure-derived interface labels add auxiliary supervision during training but are not formula inputs.

2.3 Symbolic Mixture-of-Experts Decoder

Sixteen learned queries compress the available ESM-3, MINT, fused, and structure tokens into $\mathbf{X} = [\mathbf{x}_0, \ldots, \mathbf{x}_{15}] \in \mathbb{R}^{16\times 32}$. The $\mathbf{x}_j$ are learned concepts rather than predefined biophysical quantities. A shared autoregressive decoder generates an expert formula τ_e with a budget from 3 to 18, using addition, subtraction, multiplication, protected division, and no-op tokens (Petersen et al., 2021). Restricting the grammar to four basic arithmetic operations reduces the combinatorial search space and search cost while favoring compact, readily interpretable formulas (Petersen et al., 2021). Its vector-valued evaluation is reduced to a scalar:

$$\begin{aligned} \widetilde{\mathbf{x}}_{e,j} &= \mathbf{w}_{e,j} \odot \mathbf{x}_j, & \mathbf{u}_e &= \mathrm{Eval}(\tau_e, \{\widetilde{\mathbf{x}}_{e,j}\}_{j=0}^{15}), \\ \boldsymbol{\pi}_e(\mathbf{z}) &= \mathrm{softmax}(\mathbf{c}_e + \mathbf{d}_e(\mathbf{z})), & \hat{y}_e &= s_e \boldsymbol{\pi}_e(\mathbf{z})^\top \mathbf{u}_e + b_e. \end{aligned} \tag{5}$$

The coefficient $\mathbf{w}_{e,j} \in \mathbb{R}^{32}$ reweights concept j for expert e, and Eval applies τ_e element-wise to produce $\mathbf{u}_e \in \mathbb{R}^{32}$. Learned logits $\mathbf{c}_e$ encode a static channel preference, while $\mathbf{d}_e(\mathbf{z})$ adds an input-dependent correction; their softmax gives the reduction weights $\boldsymbol{\pi}_e$. The learned s_e and b_e scale and shift the resulting scalar. The discrete syntax remains directly inspectable; the appendix specifies numerical guards and the concept diversity penalty.

2.4 Training and Reported Prediction

The model evaluates every training sample through both paths with shared parameters. The sequence-only pass samples formulas and receives the policy-gradient and pairwise-preference signals (Petersen et al., 2021; Shao et al., 2024; Rafailov et al., 2023); the structure pass replays those formulas on geometry-conditioned concepts. For each path $p \in \{\mathrm{seq}, \mathrm{str}\}$, direct Huber supervision is applied to up to four eligible experts with the lowest error:

$$\mathcal{L}^{p}_{\mathrm{top4}} = \frac{1}{N}\sum_{n=1}^{N} \omega^{(n)} \frac{1}{|\mathcal{T}^{(4)}_{n,p}|} \sum_{e \in \mathcal{T}^{(4)}_{n,p}} \mathrm{Huber}(\hat{y}_e^{(n,p)}, y^{(n)}). \tag{6}$$

Here, N is the mini-batch size, n indexes samples, $y^{(n)}$ is the measured target, and $\mathcal{T}^{(4)}_{n,p}$ contains the selected experts for sample n and path p. The factor $\omega^{(n)}$ is the normalized inverse frequency of the sample's Gaussian-smoothed integer affinity bin, computed from the training labels; it gives scarce affinity ranges more weight. Lower-weight regression and distillation terms supervise the other valid experts. A consistency loss uses detached sequence-only expert predictions as targets for the structure path. With the current weights,

$$\mathcal{L}_{\mathrm{dual}} = \mathcal{L}_{\mathrm{seq}} + 0.50\mathcal{L}_{\mathrm{str}} + 0.03\mathcal{L}_{\mathrm{cons}}. \tag{7}$$

The first two terms collect the task losses for their respective paths; $\mathcal{L}_{\mathrm{cons}}$ is a SmoothL1 loss from structure-path predictions to detached sequence-only predictions. The appendix defines the auxiliary classification, interface, diversity, and validity terms within $\mathcal{L}_{\mathrm{seq}}$ and $\mathcal{L}_{\mathrm{str}}$.

Checkpoint scoring does not use a validation-ranked top-K subset. Instead, checkpoint metrics use the valid-only mean from the full 16-expert candidate pool on the sequence-only validation path. Near-constant experts and experts that fail the positive-PCC validation gate are removed before averaging, subject to fallbacks that keep the estimator defined, and invalid predictions are omitted per sample. Thus, fewer than 16 experts may contribute even though the nominal logged value is $K = 16$. The primary checkpoint and early stopping use validation MAE. This checkpoint-scoring rule is distinct from final consensus construction, which ranks experts separately for each input mode using validation MAE and selects the consensus size only on validation data:

$$\hat{y}_K^{(n)} = \frac{\sum_{e \in \mathcal{R}_K} v_e^{(n)} \hat{y}_e^{(n)}}{\sum_{e \in \mathcal{R}_K} v_e^{(n)}}, \qquad K^* = \arg\min_K \mathrm{MAE}_{\mathrm{val}}(\hat{y}_K). \tag{8}$$

Here, $\mathcal{R}_K$ contains the K highest-ranked experts, $v_e^{(n)} \in \{0, 1\}$ marks whether expert e is valid for sample n, and $\hat{y}_K^{(n)}$ is their valid-only mean. The minimization chooses K^* from the allowed ensemble sizes. The ranking and K^* are fixed before test inference; neither test labels nor individual test samples affect them.

3 Experimental Setup

We constructed the training set from the protein–protein general index of PDBbind v2020 (Liu et al., 2017). The index contained 2,852 complexes, each with a corresponding cleaned PDB file. We retained records whose affinity field began with K_d, which removed 171 K_i records and 63 IC_{50} records. The parser accepted exact and approximate measurements written as `Kd=` or `Kd~` with units of fM, pM, nM, μM, mM, or M. We removed 14 lower-bounded and 16 upper-bounded measurements because they do not specify a single affinity value. This left 2,588 complexes. Each accepted K_d was converted to molar units and then to ΔG in kcal/mol using Equation 1, with $R = 0.0019858775\ \mathrm{kcal\,mol^{-1}\,K^{-1}}$ and $T = 298\ \mathrm{K}$.

The PDBbind index used above does not encode the receptor and ligand chain groups in the form required by our pipeline. We obtained these assignments from `ci4c01850_si_002.csv`, the supplementary table released with ProAffinity-GNN (Zhou et al., 2024). The table contains 2,283 PDB entries. Retaining its rows labeled K_d removed 214 entries and left 2,069 chain records, with one record per PDB. We parsed the first chain group in the `pairwise composition` field as the receptor and the second as the ligand. After joining these annotations to the 2,588 affinity records, 2,056 complexes had both chain groups; the remaining 532 were removed.

Our Test Set 1 and Test Set 2 follow the 82- and 166-entry benchmark cohorts used by ProAffinity-GNN and Multi_PPBA (Zhou et al., 2024; Mo et al., 2025). Test Set 1 contains 82 records from 82 PDB structures, of which 81 occurred in the 2,056-complex dataset. Test Set 2 contains 166 wild-type and mutant records derived from 26 base PDB structures (Jankauskaitė et al., 2019). We reduced each Test Set 2 identifier to its four-character PDB code, treating mutants and their corresponding wild-type structure as the same complex. One of these base structures occurred in the 2,056-complex dataset. Together, the benchmarks contain 108 unique PDB codes, 82 of which overlapped with this dataset. Removing them left 1,974 unique PDB complexes for model development. An 80:20 stratified split produced 1,578 training complexes and 396 validation complexes while maintaining the proportion of extreme-affinity cases ($\Delta G < -13$ or $\Delta G > -7$) in each set. Appendix C gives the filtering and split procedures in detail.

4 Results

4.1 Comparison with Existing Methods

Table 1 compares AffinitySym with existing methods on Test Set 1. The baselines operate under one fixed input mode, whereas the same AffinitySym checkpoint is evaluated with sequences alone and with sequences plus structure. AffinitySym is therefore the only compared method that supports both modes without training or selecting a separate model. At the best-observed checkpoint, a sequence-only formula reached an MAE of 1.2982 kcal/mol and a PCC of 0.7034. The corresponding structure-enabled formula gave an MAE of 1.3033 kcal/mol and a PCC of 0.6996. Both results compare favorably with Multi_PPBA, the strongest baseline in the table (MAE 1.36 kcal/mol; PCC 0.629).

The validation-selected consensuses remained close to the single-formula results. For each input mode, formulas are ranked and the consensus size K is chosen using validation MAE alone. This checkpoint selected six formulas with structure and seven without it. Once selected, the ranking and K are fixed for every test complex. The sequence-only top-7 consensus obtained an MAE/PCC of 1.2990/0.6806. Across its seven experts, the individual MAEs averaged 1.3909 ± 0.0798 kcal/mol (range: 1.2982–1.5526), and the PCCs averaged 0.6668 ± 0.0235 (range: 0.6312–0.7034). The structure-enabled top-6 consensus obtained an MAE/PCC of 1.3199/0.6710; five of the six selected experts were valid. Their individual MAEs averaged 1.4105 ± 0.0894 kcal/mol (range: 1.3033–1.5530), and their PCCs averaged 0.6676 ± 0.0218 (range: 0.6317–0.6996). These statistics measure variation among experts from one checkpoint, not variation across seeds.

Table 1: Comparison with existing protein–protein binding affinity prediction methods on Test Set 1. Unmarked results use the same 82-complex cohort. Symbol-marked results were reported on related filtered variants and are included for context. The best result is shown in bold, and the second-best result is underlined.

Method	Modality	Explainability	Pearson $r \uparrow$	MAE $\downarrow$
PRODIGY (Vangone & Bonvin, 2015)	Struct.	Contact-based equation	0.334	2.52
CP_PIE (Ravikant & Elber, 2010)	Struct.	Pairwise statistical potential	0.111	7.26
ISLAND (Abbasi et al., 2020)	Seq.	Not reported	0.217	2.15
PPI-Affinity (Romero-Molina et al., 2022)	Struct.	Not reported	0.436	1.78
ProAffinity-GNN (Zhou et al., 2024)	Seq. + Struct.	Not reported	0.620	1.49
ProBAN (Bogdanova & Novoseletsky, 2024)	Struct.	Not reported	0.55	1.75
PPBind-3D† (Anonymous, 2025)	Struct.	Not reported	0.559	1.65
PPI-Graphomer‡ (Xie et al., 2025)	Seq. + Struct.	Not reported	0.625	1.51
Multi_PPBA (Mo et al., 2025)	Seq. + Struct.	Attention analysis	0.629	1.36
AffinitySym (single formula)	Seq. + Struct.	Formula + residue attribution	<u>0.6996</u>	1.3033
AffinitySym (single formula)	Seq.	Formula + residue attribution	**0.7034**	**1.2982**
AffinitySym (top-6 consensus)	Seq. + Struct.	Formula + residue attribution	0.6710	1.3199
AffinitySym (top-7 consensus)	Seq.	Formula + residue attribution	0.6806	<u>1.2990</u>

Seq: sequence; Struct: structure. "Not reported" indicates that the cited work does not report an explicit equation, attention analysis, or sample-level attribution for its affinity prediction. AffinitySym results are reported to four decimal places; baseline results retain the precision used in the cited sources. †PPBind-3D used an 83-complex variant. ‡PPI-Graphomer used 87 complexes after its sequence-length filter. These two rows are not direct comparisons with the 82-complex cohort.

Test Set 2 is drawn from SKEMPI v2.0 (Jankauskaitė et al., 2019) and contains 26 wild-type and 140 mutant protein–protein complexes. AffinitySym was trained on PDBbind-derived affinity data rather than mutation-affinity measurements, so this experiment evaluates zero-shot transfer to mutation-containing complexes. The checkpoint reported in Table 1 obtained an MAE of 1.3754 kcal/mol, a PCC of 0.7149, and an SPCC of 0.8377 with structure input. With sequence alone, it obtained an MAE of 1.3677 kcal/mol, a PCC of 0.7267, and an SPCC of 0.8372. SPCC is useful in this setting because mutation studies often depend on ranking candidate variants by their predicted affinities, even when the absolute values contain a systematic offset. The SPCC values above 0.83 in both input modes indicate that this ordering was preserved well.

A separate ESM-3+MINT checkpoint produced a best-observed individual formula with an MAE of 0.9960 kcal/mol and a PCC of 0.8493 on Test Set 2. This peak was not retained by validation-selected consensus: its top-12 consensus obtained an MAE of 1.3910 kcal/mol, a PCC of 0.8262, and an SPCC of 0.8540. We therefore treat the individual formula as an attainable result from checkpoint analysis, not as the expected performance of the validation-selected predictor. For reference, Multi_PPBA and ProAffinity-GNN report MAE/PCC values of 1.20/0.782 and 1.36/0.722, respectively, on the same 166-complex cohort (Zhou et al., 2024; Mo et al., 2025). PPI-Affinity reports 1.40/0.780 on the earlier 177-complex cohort (26 wild type and 151 mutants), so that result provides context but is not a direct 166-complex comparison (Romero-Molina et al., 2022).

4.2 Multi-Seed Modality Ablation Study

Table 2 separates peak checkpoint performance from variation across training runs. For each run, we use the checkpoint with the lowest validation MAE under the mean of all valid experts and report the mean and sample standard deviation. Panel A pools Panels B and C (ten seeds total). Panel B contains five predefined seeds shared by all configurations, which supports a matched comparison, and Panel C contains five additional random seeds for each configuration.

Table 2: Multi-seed modality ablation on Test Sets 1 and 2. Panel A pools the seeds from Panels B and C. Results are reported as mean ± standard deviation for each seed group. The best result in each panel and metric is shown in bold, and the second-best result is underlined.

Configuration	Inference input	Test 1 MAE ↓	Test 1 PCC ↑	Test 2 MAE ↓	Test 2 PCC ↑
Panel A: Panels B + C pooled (ten seeds)					
Full (ESM-3 + MINT + GNN)	Seq.	1.6414 ± 0.0952	0.4904 ± 0.0661	1.5630 ± 0.2311	0.7137 ± 0.0874
Full (ESM-3 + MINT + GNN)	Seq. + Struct.	$\mathbf{1.6379 \pm 0.0960}$	0.4887 ± 0.0709	1.5650 ± 0.2459	0.7101 ± 0.0933
ESM-3 + MINT	Seq.	1.6557 ± 0.1066	$\mathbf{0.4961 \pm 0.0626}$	$\mathbf{1.4190 \pm 0.1625}$	$\mathbf{0.7333 \pm 0.0625}$
ESM-3 + GNN	Seq.	1.7187 ± 0.1354	0.4371 ± 0.0984	1.6334 ± 0.2055	0.5853 ± 0.1299
ESM-3 + GNN	Seq. + Struct.	1.7331 ± 0.1381	0.4150 ± 0.1181	1.6650 ± 0.2334	0.5535 ± 0.1625
ESM-3 only	Seq.	1.6765 ± 0.0740	0.4918 ± 0.0290	1.5886 ± 0.1797	0.6477 ± 0.1068
Panel B: Five matched predefined seeds					
Full (ESM-3 + MINT + GNN)	Seq.	1.6255 ± 0.1294	$\mathbf{0.5081 \pm 0.0871}$	1.5379 ± 0.2910	$\mathbf{0.7360 \pm 0.0991}$
Full (ESM-3 + MINT + GNN)	Seq. + Struct.	$\mathbf{1.6175 \pm 0.1328}$	0.5055 ± 0.0976	1.5401 ± 0.2937	0.7347 ± 0.0981
ESM-3 + MINT	Seq.	1.6745 ± 0.1254	0.4879 ± 0.0823	$\mathbf{1.4660 \pm 0.0834}$	0.7278 ± 0.0479
ESM-3 + GNN	Seq.	1.7466 ± 0.1171	0.4124 ± 0.0978	1.7114 ± 0.1652	0.5334 ± 0.1228
ESM-3 + GNN	Seq. + Struct.	1.7666 ± 0.1257	0.3760 ± 0.1317	1.7712 ± 0.1945	0.4725 ± 0.1597
ESM-3 only	Seq.	1.6640 ± 0.1021	0.4998 ± 0.0335	1.5496 ± 0.1838	0.6799 ± 0.1121
Panel C: Five additional random seeds					
Full (ESM-3 + MINT + GNN)	Seq.	1.6573 ± 0.0551	0.4727 ± 0.0382	1.5880 ± 0.1842	0.6915 ± 0.0782
Full (ESM-3 + MINT + GNN)	Seq. + Struct.	1.6584 ± 0.0457	0.4718 ± 0.0329	1.5899 ± 0.2197	0.6855 ± 0.0920
ESM-3 + MINT	Seq.	$\mathbf{1.6369 \pm 0.0948}$	$\mathbf{0.5044 \pm 0.0430}$	$\mathbf{1.3721 \pm 0.2168}$	$\mathbf{0.7388 \pm 0.0801}$
ESM-3 + GNN	Seq.	1.6907 ± 0.1600	0.4618 ± 0.1035	1.5553 ± 0.2291	0.6372 ± 0.1272
ESM-3 + GNN	Seq. + Struct.	1.6996 ± 0.1559	0.4540 ± 0.1010	1.5588 ± 0.2378	0.6344 ± 0.1324
ESM-3 only	Seq.	1.6891 ± 0.0390	0.4838 ± 0.0248	1.6276 ± 0.1873	0.6155 ± 0.1025

The ESM-3+MINT configuration remained close to the full model on Test Set 1 and was stronger on Test Set 2. Across all ten seeds, its sequence-only prediction gave an MAE/PCC of 1.6557 ± 0.1066/0.4961 ± 0.0626 on Test Set 1, compared with 1.6414 ± 0.0952/0.4904 ± 0.0661 for the full model. On Test Set 2, ESM-3+MINT improved the corresponding values from 1.5630 ± 0.2311/0.7137 ± 0.0874 to 1.4190 ± 0.1625/0.7333 ± 0.0625. In both five-seed panels, ESM-3+MINT retained a lower Test Set 2 MAE than the full model. Its PCC was similar in the matched panel and higher in the additional-seed panel.

This result is consistent with the dual-path objective. The sequence-only path is the primary prediction path, while the structure-enabled path is assigned a weight of 0.5 during training. MINT can therefore supply much of the interaction information needed for sequence-only inference. The

graph branch supports structure-conditioned use cases rather than serving solely to improve average benchmark performance. It adds geometric evidence that complements the sequence features. Without structure, the model must infer the relevant bound state from the partner sequences and cannot distinguish alternative complexes formed by those sequences. With structure, the same sequence pair can be evaluated in the context of a supplied conformation, interface, and chain arrangement. The two paths consequently support different information settings: sequence-only screening when geometry is unavailable and conformation-aware analysis when it is provided.

4.3 EXPLAINABILITY

AffinitySym assigns formula-conditioned attribution scores to individual residues, connecting each affinity estimate to specific positions in both protein partners. We built a web demo that displays these scores alongside the predicted affinity, auxiliary probability distributions, and selected executable formula. The reported ΔG can therefore be traced through the equation to the residues that influence it. We illustrate this process with the protein complex 1WR1 (Ohno et al., 2005) from Test Set 1, using chain A as the receptor and chain B as the ligand.

For sequence-only input, the validation-selected top-7 consensus predicts $\Delta G = -6.525$ kcal/mol. The three-class head assigns the complex to the weak-binding regime ($\Delta G > -7$ kcal/mol) with 81.0% probability. The 44-bin ordinal head places probability over 0.5 kcal/mol affinity intervals; its expected value is -6.858 kcal/mol, and its most probable bin is centered at -6.75 kcal/mol. These classifier outputs are auxiliary references; the symbolic regression consensus remains the primary prediction.

Symbolic expert L2 ($e = 2$) predicts -6.5444 kcal/mol, close to the measured value of -6.58 kcal/mol. Its learned pre-reduction vector is

$$\begin{aligned} \mathbf{r}_{\mathrm{L2}} &= \left(1.0012\left(\mathbf{w}_{\mathrm{L2},14} \odot \mathbf{x}_{14}\right) - 1.0012\left(\mathbf{w}_{\mathrm{L2},7} \odot \mathbf{x}_{7}\right)\right) - 9.9405\,\mathbf{1}, \\ \hat{y}_{\mathrm{L2}} &= \boldsymbol{\pi}_{\mathrm{L2}}(\mathbf{z})^{\top}\mathbf{r}_{\mathrm{L2}}. \end{aligned} \tag{9}$$

where $\mathbf{x}_7$ and $\mathbf{x}_{14}$ are learned concept vectors, $\mathbf{w}$ contains their channel-wise coefficients, $\mathbf{1}$ broadcasts the bias across channels, and $\boldsymbol{\pi}$ gives the channel-reduction weights. The displayed scale and bias are specific to expert L2. Appendix B lists all 16 expert formulas produced for this sequence-only 1WR1 example. Although the model provides 16 latent concepts, Equation 9 selects only two of them. This sparsity makes the variables used by the expert explicit before their effects are traced to individual residues.

The sequence view in Figure 3 traces expert L2 back to the residues that influence its output. For expert e with prediction $\hat{y}_e$ and ESM-3 residue embedding $\mathbf{e}_i$, the signed score for residue i is computed by gradient times input,

$$s_i^{(e)} = \sum_d e_{i,d} \frac{\partial \hat{y}_e}{\partial e_{i,d}}. \tag{10}$$

The derivative follows the selected expert's symbolic prediction through its concept variables to the sequence representation. Summing over embedding channels gives one score per residue. We divide the scores by the 95th percentile of their absolute values and clip them to $[-1, 1]$ to limit the effect of outliers. A negative score lowers the predicted ΔG and is labeled favorable; a positive score raises ΔG and is labeled unfavorable. The median of the absolute normalized scores separates weak from strong contributions, producing four levels: strongly favorable, weakly favorable, weakly unfavorable, and strongly unfavorable. These labels describe relative effects under the selected formula and protein pair, not fixed biochemical properties.

This residue resolution matters for protein-targeted drug discovery. Selecting a protein identifies the macromolecular target, but molecular recognition depends on particular amino acids and the interface neighborhoods they form. The four-level map narrows the explanation from the two proteins in the complex to specific positions inside each chain. In Figure 3, strongly favorable and strongly unfavorable residues provide an initial set of positions for target refinement, mutational testing, and interface-focused design. Researchers can therefore prioritize amino acids for experimental validation instead of treating every residue in the selected protein as equally relevant. The scores remain model-derived hypotheses; they do not by themselves establish a causal binding mechanism or confirm a druggable site.

When the 1WR1 PDB structure is supplied, AffinitySym retains the prediction, formula, and sequence-attribution views and adds a structure-specific view. It computes gradient-times-input scores from the selected formula to each GNN residue node, normalizes these scores separately from the sequence scores, and maps the four contribution classes onto the three-dimensional complex. The three-dimensional view can show whether prioritized residues cluster at an interface or form a spatial neighborhood that is not apparent from the linear sequence. Keeping the two scales separate also shows what the sequence and structure paths contribute without conflating their attribution magnitudes.

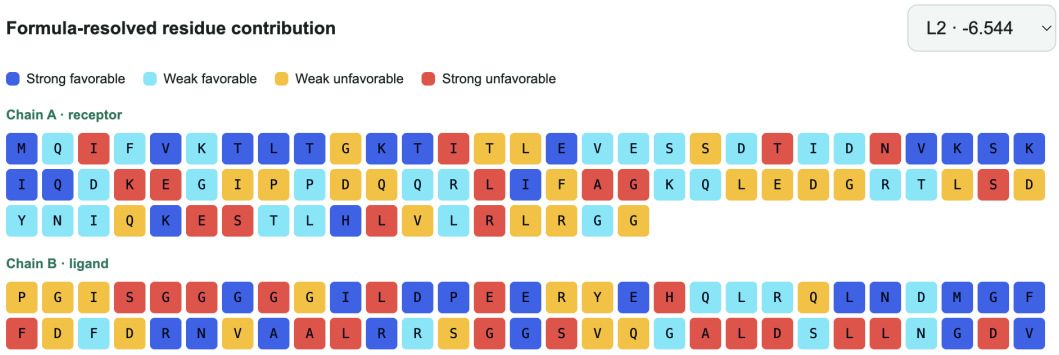

Figure 3: Screenshot from the AffinitySym website demo showing formula-resolved residue contributions for sequence-only expert L2 on 1WR1 (Ohno et al., 2005). The four colors distinguish strongly favorable, weakly favorable, weakly unfavorable, and strongly unfavorable contributions to the selected formula prediction, providing residue-level candidates for subsequent experimental validation.

4.4 Future Directions

We plan to extend AffinitySym in several directions. First, the learned variables can be linked to measurable physical properties, such as interface area, electrostatic interactions, solvation, and conformational change, by adding concept supervision and physical constraints. This would make the resulting formulas easier to interpret in biochemical terms. Second, we will study whether the formula families learned by each expert remain stable across protein families, mutations, datasets, and experimental conditions. Consistent relationships across these settings would help determine whether a universal equation for PPI binding affinity is achievable. On the modeling side, we plan to pretrain the autoregressive symbolic decoder on synthetic expression-recovery tasks before affinity-specific training. This could provide a better initialization for formula generation, reduce variation across random seeds, and support carefully expanded operator sets without requiring the PPI dataset to teach symbolic syntax from scratch. We will also evaluate the model on larger and more varied protein complexes, add calibrated uncertainty estimates, and compare residue-level contributions with experimental mutation measurements. These extensions retain the current sequence-first design while testing how broadly its formulas transfer.

5 Conclusion

We presented AffinitySym, a sequence-first model that predicts protein–protein binding affinity with symbolic formulas rather than explaining a black-box output after prediction. Frozen ESM-3 and MINT encoders represent the partner sequences, while an additional graph branch supplies complex geometry. A single checkpoint supports both sequence-only and sequence–structure inference and reports expert formulas, residue-level attributions, and a validation-selected consensus. At the best-observed Test Set 1 checkpoint, a sequence-only formula achieved an MAE of 1.2982 kcal/mol; the same checkpoint achieved 1.3677 kcal/mol on Test Set 2. As an open-source model that can run locally, AffinitySym offers a setting for neuro-symbolic research and helps protein scientists prioritize interactions and mutations for experimental follow-up.

AI USE STATEMENT

The idea of combining sequence-based binding affinity prediction with symbolic formulas originated with the authors. Generative AI tools were used only to assist with code implementation and debugging, discussion of experimental design and results, and drafting and editing manuscript text. The authors reviewed the AI-assisted work and made all final decisions and edits.

ETHICS STATEMENT

This study uses published protein-complex datasets and involves no human participants or personal data. Formula-conditioned residue scores are model-derived attributions, not measured per-residue binding energies or validated drug-target sites; the three-class and 44-bin outputs are auxiliary predictions. These outputs can guide experimental follow-up, but should not be used alone for drug-target selection or clinical decisions, particularly for complexes outside the benchmark distributions.

REPRODUCIBILITY STATEMENT

Sections 2 and 3 describe the model, evaluation, and data split; Appendices A and C give the implementation and dataset-construction details. The anonymous code release provides the training and preprocessing scripts, configurations, and checkpoint access instructions. The code seeds Python, NumPy, PyTorch, and data-loader workers and enables deterministic PyTorch algorithms to control stochastic behavior during training and evaluation. With the same seed and computing environment, repeated runs produced bit-for-bit identical training losses, validation and test metrics, and model weights. We have not established bitwise reproducibility across different GPUs, drivers, Python versions, or package versions. Processed PDBbind-derived data are not redistributed; the preprocessing scripts document their construction from the source data.

REFERENCES

Wajid Arshad Abbasi, Adiba Yaseen, Fahad Ul Hassan, Saiqa Andleeb, and Fayyaz Ul Amir Afsar Minhas. Island: in-silico proteins binding affinity prediction using sequence information. *BioData Mining*, 13:1–13, 2020.

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

Gustaf Ahdritz, Nazim Bouatta, Christina Floristean, Sachin Kadyan, Qinghui Xia, William Gerecke, Timothy J. O'Donnell, Daniel Berenberg, Ian Fisk, Niccolo Zanichelli, Bo Zhang, Arkadiusz Nowaczynski, Bei Wang, Marta M. Stepniewska-Dziubinska, Shang Zhang, Adegoke Ojewole, Murat Efe Guney, Stella Biderman, Andrew M. Watkins, Stephen Ra, Pablo Ribalta Lorenzo, Lucas Nivon, Brian Weitzner, Yih-En Andrew Ban, Shiyang Chen, Minjia Zhang, Conglong Li, Shuaiwen Leon Song, Yuxiong He, Peter K. Sorger, Emad Mostaque, Zhao Zhang, Richard Bonneau, and Mohammed AlQuraishi. OpenFold: Retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization. *Nature Methods*, 21: 1514–1524, 2024. doi: 10.1038/s41592-024-02272-z. URL https://doi.org/10.1038/s41592-024-02272-z.

Sama Akbarzadeh, Özlem Coşkun, and Başak Günçer. Studying protein–protein interactions: Latest and most popular approaches. *Journal of structural biology*, 216(4):108118, 2024.

Ethan C Alley, Grigory Khimulya, Surojit Biswas, Mohammed AlQuraishi, and George M Church. Unified rational protein engineering with sequence-based deep representation learning. *Nature methods*, 16(12):1315–1322, 2019.

Anonymous. Enhancing PPB affinity prediction through data integration and feature alignment: Approaching structural model performance with sequences, 2025. URL https://openreview.net/forum?id=xNDydjYBmC. ICLR 2025 submission.

Elizaveta Alexandrovna Bogdanova and Valery Nikolaevich Novoseletsky. Proban: Neural network algorithm for predicting binding affinity in protein–protein complexes. *Proteins: Structure, Function, and Bioinformatics*, 2024.

Lingtao Chen, Kazi Fahim Ahmad Nasif, Bobin Deng, Shuteng Niu, and Chloe Yixin Xie. Predicting protein-protein binding affinity with deep learning: A comparative analysis of cnn and transformer models. In *2024 IEEE 36th International Conference on Tools with Artificial Intelligence (ICTAI)*, pp. 548–555. IEEE, 2024.

Yasmina SN Day, Cheryl L Baird, Rebecca L Rich, and David G Myszka. Direct comparison of binding equilibrium, thermodynamic, and rate constants determined by surface-and solution-based biophysical methods. *Protein Science*, 11(5):1017–1025, 2002. ISSN 0961-8368.

Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, et al. Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7112–7127, 2022. doi: 10.1109/TPAMI.2021.3095381.

Xin Geng. Label distribution learning. *IEEE Transactions on Knowledge and Data Engineering*, 28(7):1734–1748, 2016. doi: 10.1109/TKDE.2016.2545658. URL https://doi.org/10.1109/TKDE.2016.2545658.

Tomas Hansson, Chris Oostenbrink, and WilfredF van Gunsteren. Molecular dynamics simulations. *Current opinion in structural biology*, 12(2):190–196, 2002.

Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q Tran, Jonathan Deaton, Marius Wiggert, et al. Simulating 500 million years of evolution with a language model. *Science*, 387(6736):850–858, 2025.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. doi: 10.48550/arXiv.1503.02531. URL https://arxiv.org/abs/1503.02531.

Scott A Hollingsworth and Ron O Dror. Molecular dynamics simulation for all. *Neuron*, 99(6): 1129–1143, 2018.

Peter J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964. doi: 10.1214/aoms/1177703732. URL https://doi.org/10.1214/aoms/1177703732.

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4651–4664. PMLR, 2021. URL https://proceedings.mlr.press/v139/jaegle21a.html.

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.

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

William L Jorgensen and Laura L Thomas. Perspective on free-energy perturbation calculations for chemical equilibria. *Journal of chemical theory and computation*, 4(6):869–876, 2008.

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

Panagiotis L Kastritis, Iain H Moal, Howook Hwang, Zhiping Weng, Paul A Bates, Alexandre MJJ Bonvin, and Joël Janin. A structure-based benchmark for protein–protein binding affinity. *Protein Science*, 20(3):482–491, 2011.

Peng Liu, François Dehez, Wensheng Cai, and Christophe Chipot. A toolkit for the analysis of free-energy perturbation calculations. *Journal of chemical theory and computation*, 8(8):2606–2616, 2012.

Song Liu, Chi Zhang, Hongyi Zhou, and Yaoqi Zhou. A physical reference state unifies the structure-derived potential of mean force for protein folding and binding. *Proteins: Structure, Function, and Bioinformatics*, 56(1):93–101, 2004.

Zhihai Liu, Minyi Su, Li Han, Jie Liu, Qifan Yang, Yan Li, and Renxiao Wang. Forging the basis for developing protein–ligand interaction scoring functions. *Accounts of chemical research*, 50 (2):302–309, 2017.

Jisong Mo, Jingjing Guo, and Hongliang Duan. Multimodal learning of protein–protein interactions for accurate binding affinity prediction. *ACS omega*, 10(47):57137–57144, 2025.

Ayako Ohno, JunGoo Jee, Kenichiro Fujiwara, Takeshi Tenno, Natsuko Goda, Hidehito Tochio, Hideki Kobayashi, Hidekazu Hiroaki, and Masahiro Shirakawa. Structure of the uba domain of dsk2p in complex with ubiquitin: molecular determinants for ubiquitin recognition. *Structure*, 13 (4):521–532, 2005.

Brenden K. Petersen, Mikel Landajuela, T. Nathan Mundhenk, Claudio P. Santiago, Soo K. Kim, and Joanne T. Kim. Deep symbolic regression: Recovering mathematical expressions from data via risk-seeking policy gradients. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=m5Qsh0kBQG.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pp. 53728–53741, 2023. doi: 10.52202/075280-2338. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html.

DVS Ravikant and Ron Elber. Pie—efficient filters and coarse grained potentials for unbound protein–protein docking. *Proteins: Structure, Function, and Bioinformatics*, 78(2):400–419, 2010.

Sandra Romero-Molina, Yasser B Ruiz-Blanco, Joel Mieres-Perez, Mirja Harms, Jan Munch, Michael Ehrmann, and Elsa Sanchez-Garcia. Ppi-affinity: A web tool for the prediction and optimization of protein–peptide and protein–protein binding affinity. *Journal of proteome research*, 21(8):1829–1841, 2022.

Wojciech Samek, Grégoire Montavon, Sebastian Lapuschkin, Christopher J Anders, and Klaus-Robert Müller. Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE*, 109(3):247–278, 2021. ISSN 0018-9219.

Víctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E(n) equivariant graph neural networks. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9323–9332. PMLR, 2021. URL https://proceedings.mlr.press/v139/satorras21a.html.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. doi: 10.48550/arXiv.2402.03300. URL https://arxiv.org/abs/2402.03300.

Till Siebenmorgen and Martin Zacharias. Computational prediction of protein–protein binding affinities. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 10(3):e1448, 2020.

Varun Ullanat, Bowen Jing, Samuel Sledzieski, and Bonnie Berger. Learning the language of protein-protein interactions. *Nature Communications*, 17(1):1199, 2026.

Anna Vangone and Alexandre MJJ Bonvin. Contacts-based prediction of binding affinity in protein–protein complexes. *elife*, 4:e07454, 2015.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Maria Winiewska-Szajewska and Jarosław Poznański. Differential scanning fluorimetry followed by microscale thermophoresis and/or isothermal titration calorimetry as an efficient tool for ligand screening. *Biophysical Reviews*, 17(1):199–223, 2025.

Jun Xie, Youli Zhang, Ziyang Wang, Xiaocheng Jin, Xiaoli Lu, Shengxiang Ge, and Xiaoping Min. PPI-Graphomer: Enhanced protein–protein affinity prediction using pretrained and graph transformer models. *BMC Bioinformatics*, 26(1):116, 2025. doi: 10.1186/s12859-025-06123-2. URL https://doi.org/10.1186/s12859-025-06123-2.

Li C Xue, João Pglm Rodrigues, Panagiotis L Kastritis, Alexandre Mjj Bonvin, and Anna Vangone. Prodigy: a web server for predicting the binding affinity of protein–protein complexes. *Bioinformatics*, 32(23):3676–3678, 2016.

Zhiyuan Zhou, Yueming Yin, Hao Han, Yiping Jia, Jun Hong Koh, Adams Wai-Kin Kong, and Yuguang Mu. Proaffinity-gnn: A novel approach to structure-based protein–protein binding affinity prediction via a curated data set and graph neural networks. *Journal of chemical information and modeling*, 64(23):8796–8808, 2024.

APPENDIX

A ADDITIONAL METHODOLOGY DETAILS

A.1 PROBLEM DEFINITION

We study protein–protein binding affinity prediction for a pair of interacting protein partners. A partner may contain one chain or a group of chains. For sample n, let

$$\mathcal{I}^{(n)} = \left(\mathcal{S}_R^{(n)}, \mathcal{S}_L^{(n)}, \mathcal{G}^{(n)}, m_G^{(n)}\right), \tag{11}$$

where $\mathcal{S}_R = \{S_{R,1}, \ldots, S_{R,C_R}\}$ and $\mathcal{S}_L = \{S_{L,1}, \ldots, S_{L,C_L}\}$ are the receptor and ligand chain sets. Each sequence $S = (a_1, \ldots, a_L)$ is defined over the amino-acid vocabulary $\mathcal{V}$. The additional residue graph $\mathcal{G}$ is constructed when a complex structure is supplied, and $m_G \in \{0, 1\}$ records whether that structure is available. The dataset provides experimental dissociation constants K_d in molar units, which we convert to the Gibbs free-energy target $y = \Delta G \in \mathbb{R}$, measured in kcal/mol, using Equation 1.

AffinitySym learns

$$f_\theta : (\mathcal{S}_R, \mathcal{S}_L, \mathcal{G}, m_G) \longmapsto \left(\hat{y}, \{\hat{y}_e, \tau_e\}_{e=1}^{M}, \mathbf{p}_3, \mathbf{p}_{44}\right), \tag{12}$$

where $\hat{y}$ is the reported consensus affinity, $\hat{y}_e$ and τ_e are the prediction and expression of symbolic expert e, and $\mathbf{p}_3$ and $\mathbf{p}_{44}$ are auxiliary distributions over three affinity regimes and 44 ordered affinity intervals. The model has $M = 16$ symbolic experts, labeled $\mathrm{L1}, \ldots, \mathrm{L16}$ in its outputs; $e = 1, \ldots, 16$ indexes those same experts in the equations. Thus, L2 and expert $e = 2$ refer to the same component. The model supports two inference paths: a sequence-only path with $m_G = 0$, and a sequence–structure path that uses $\mathcal{G}$ when a complex structure is supplied. With $m_G = 0$, the prediction is conditioned only on the partner sequences and implicitly assumes the bound state relevant to the affinity label; it cannot distinguish different conformations or interfaces formed by the same sequences. With $m_G = 1$, $\mathcal{G}$ specifies a particular structural state, so the prediction can respond to changes in chain arrangement, interface contacts, and residue geometry. Structure thus supplies distinct geometric evidence when available, but it is neither a prerequisite for inference nor a feature expected to improve every aggregate benchmark result. Both paths return the same set of output types. The symbolic predictions are primary; the classification heads give secondary summaries of the predicted affinity range.

A.2 PARTNER-AWARE SEQUENCE REPRESENTATION

AffinitySym uses frozen ESM-3 and MINT encoders (Hayes et al., 2025; Ullanat et al., 2026). Freezing these backbones allows their representations to be precomputed and keeps optimization within the task-specific network.

ESM-3 path. The receptor and ligand chains are joined with explicit chain boundaries and truncated by allocating the residue budget across both partners. ESM-3 produces a residue representation $\mathbf{E} \in \mathbb{R}^{L \times 1536}$. Under the current setting, representations from an intermediate layer and the final layer are mixed by learned token-wise attention. A nonlinear adapter maps each residue through $1536 \to 768 \to 384$ dimensions. The projected tokens then pass through shared chain-local Transformer blocks followed by bidirectional receptor-to-ligand and ligand-to-receptor cross-attention (Vaswani et al., 2017).

For pooled partner vectors $\mathbf{r}, \mathbf{l} \in \mathbb{R}^{384}$, the paired sequence representation is

$$\mathbf{z}_E = \phi_E\left([\mathbf{r} + \mathbf{l}; |\mathbf{r} - \mathbf{l}|; \mathbf{r} \odot \mathbf{l}]\right), \tag{13}$$

where $[\cdot;\cdot]$ denotes concatenation, $\odot$ denotes element-wise multiplication, and ϕ_E is a shared projection. This construction is unchanged when the two partner labels are exchanged.

MINT path. MINT encodes the two partner orientations, receptor followed by ligand and ligand followed by receptor. Each orientation yields a 2560-dimensional pair vector. The same projection ϕ_M is applied to both orientations, and their projected values are averaged:

$$\mathbf{t}_{M,o} = \phi_M(\mathbf{m}_o), \qquad \mathbf{z}_M = \frac{1}{2}\left(\mathbf{t}_{M,R\to L} + \mathbf{t}_{M,L\to R}\right). \tag{14}$$

The current projection uses a gated residual adapter followed by a map to the shared 384-dimensional model space. The pooled vector $\mathbf{z}_M$ is computed for fusion modes that require one MINT vector. The active bidirectional cross-attention mode instead retains the two projected tokens $\mathbf{t}_{M,R\to L}$ and $\mathbf{t}_{M,L\to R}$; $\mathbf{z}_M$ is not an input to this active fusion path.

The ESM-3 residue tokens and the two projected MINT orientation tokens interact through two layers of bidirectional cross-attention. The ESM stream is mask-averaged over residues, the MINT stream is averaged over its two orientation tokens, and their fused output is added to $\mathbf{z}_E$ through a residual connection. Sharing ϕ_M across orientations and pooling the orientation tokens symmetrically make the resulting $\mathbf{z}_S \in \mathbb{R}^{384}$ invariant to swapping the partner labels.

A.3 Additional Structural Representation

When a complex structure is supplied, AffinitySym builds a residue-level k-nearest-neighbor graph $\mathcal{G} = (\mathcal{V}_G, \mathcal{E}_G)$. Only residues from the selected receptor and ligand chains are included. Each node contains a 21-dimensional amino-acid encoding, backbone-angle terms, and 14 residue or structure descriptors, for 37 features in total. The descriptors include hydrophobicity, charge, molecular weight, hydrogen-bond counts, side-chain volume, flexibility, occupancy, temperature factor, elastic entropy, and rotatable-bond information. Each edge has ten features that record distance, chain membership, bounded sequence separation, and indicators for electrostatic, hydrogen-bond, hydrophobic, aromatic, cation–π, and disulfide interactions. The neighbor search reserves connections within and between the two protein partners when such neighbors are available.

Node features are projected to 256 dimensions and processed in parallel by an equivariant graph neural network (EGNN) (Satorras et al., 2021) and OpenFold invariant point attention (IPA) (Ahdritz et al., 2024; Jumper et al., 2021). The EGNN uses C_α coordinates and edge attributes, while IPA also uses the local backbone frame formed by the N, C_α, and C atoms. Their node outputs are concatenated and projected to 384-dimensional structure tokens:

$$\mathbf{H}_G = \phi_G\left([\mathbf{H}_{\mathrm{EGNN}}; \mathbf{H}_{\mathrm{IPA}}]\right). \tag{15}$$

If the structure is absent, this branch is skipped rather than replaced by coordinates inferred from sequence.

A.4 Multimodal Fusion and Interface Supervision

The current fusion module applies bidirectional cross-attention (Vaswani et al., 2017) between the sequence tokens and $\mathbf{H}_G$. Masked mean pooling produces one vector for each stream. These vectors are joined with four structure-quality measurements: residue coverage, mean occupancy, normalized temperature-factor quality, and the fraction of inter-partner edges. A residual update gives

$$\mathbf{z} = \mathrm{LN}\left(\mathbf{z}_S + \phi_F\left([\mathrm{pool}(\widetilde{\mathbf{H}}_S); \mathrm{pool}(\widetilde{\mathbf{H}}_G); \mathbf{q}_G]\right)\right), \tag{16}$$

where $\widetilde{\mathbf{H}}_S$ and $\widetilde{\mathbf{H}}_G$ are the cross-attended token streams and $\mathbf{q}_G \in \mathbb{R}^4$ contains the quality measurements. Without a structure, $\mathbf{z} = \mathbf{z}_S$.

Structure-derived interface labels provide auxiliary supervision during the sequence–structure pass. Separate tasks identify interface residues, inter-partner contacts, distance intervals, and coarse interaction types. The sequence-interface head predicts these labels from sequence tokens through receptor–ligand cross-attention; target edge attributes are not passed into its input. If ℓ_i is the predicted interface logit for residue i, the active guidance step rescales its projected structure token as

$$\mathbf{h}_{G,i} \leftarrow [1 + \gamma \tanh(\ell_i)]\, \mathbf{h}_{G,i}, \qquad \gamma = 1. \tag{17}$$

The rescaled tokens then enter sequence–structure cross-attention and masked mean pooling. Thus, the predicted interface scores modulate the structure tokens rather than acting as a separate pooling softmax. These interface losses shape the representation but do not appear in the exported symbolic formulas.

A.5 Latent Concepts and Symbolic Experts

The symbolic branch retains token-level information instead of using only the pooled vector $\mathbf{z}$. Its memory contains the ESM-3 tokens, MINT orientation tokens, and the fused vector. The sequence–

structure path also includes structure tokens. Sixteen learned queries attend to this memory and then attend to one another, following the fixed-query compression used in Perceiver-style models (Jaegle et al., 2021). Two query blocks produce

$$\mathbf{X} = [\mathbf{x}_0, \ldots, \mathbf{x}_{15}] \in \mathbb{R}^{16\times 32}. \tag{18}$$

The vectors $\mathbf{x}_j$ are learned latent variables, not predefined physical quantities. To reduce redundancy among them, we penalize pairwise cosine similarity. For a mini-batch of size B, define

$$\overline{\mathbf{x}}_{n,j} = \frac{\mathbf{x}_{n,j}}{\max(\|\mathbf{x}_{n,j}\|_2, \epsilon_{\text{norm}})}, \tag{19}$$

where $\epsilon_{\text{norm}} = 10^{-8}$ prevents division by zero. With $J = 16$ concepts, the diversity loss for either inference path is

$$\mathcal{L}_{\text{div}} = \frac{1}{BJ(J-1)} \sum_{n=1}^{B} \sum_{j=0}^{J-1} \sum_{\substack{k=0 \\ k\neq j}}^{J-1} \left(\overline{\mathbf{x}}_{n,j}^{\top} \overline{\mathbf{x}}_{n,k}\right)^2. \tag{20}$$

The sum excludes diagonal entries. Squaring penalizes concepts that are aligned or anti-aligned, encouraging the queries to encode different directions without constraining their magnitudes.

Each symbolic expert has its own coefficient vector for every latent variable. For expert e and variable j,

$$\widetilde{\mathbf{x}}_{e,j} = \mathbf{w}_{e,j} \odot \mathbf{x}_j, \qquad \mathbf{w}_{e,j} \in \mathbb{R}^{32}. \tag{21}$$

A two-layer autoregressive decoder generates a prefix expression τ_e conditioned on the expert prompt and the full concept memory (Petersen et al., 2021). The experts use different expression budgets, currently ranging from 3 to 18. The active grammar contains addition, subtraction, multiplication, protected division, and a no-operation token. Restricting the grammar to four basic arithmetic operations reduces the combinatorial search space and search cost while favoring compact, readily interpretable formulas (Petersen et al., 2021). Operators act element-wise on the 32 channels, so evaluating τ_e yields a vector $\mathbf{u}_e \in \mathbb{R}^{32}$. Protected division lower-bounds the magnitude of each denominator by $\epsilon_{\text{div}} = 0.05$ while preserving its sign.

Static expert parameters and an input-dependent projection of $\mathbf{z}$ define channel weights

$$\boldsymbol{\pi}_e(\mathbf{z}) = \text{softmax}\left(\mathbf{c}_e + \mathbf{d}_e(\mathbf{z})\right). \tag{22}$$

The scalar prediction of expert e is

$$\hat{y}_e = s_e\, \boldsymbol{\pi}_e(\mathbf{z})^{\top} \mathbf{u}_e + b_e, \tag{23}$$

where s_e and b_e are learned expert-specific scale and bias terms. The exported expression records the discrete formula and its learned coefficients. Internal numerical guards prevent non-finite evaluation; the current model does not append a final clamp or a validation-derived offset to the exported formula.

We use a symbolic mixture of experts rather than a single dense regressor or a tree-based surrogate for two reasons. First, each expression is constrained to contain only the latent concepts selected by its syntax, yielding sparse concept-level dependencies that can be traced directly to the prediction; a tree may instead distribute its decisions across many variables and paths. Second, binding affinity is continuous, and algebraic expressions represent continuous relationships directly rather than through piecewise regions. Multiple experts with different expression budgets provide alternative sparse functional forms, while validation-selected consensus combines their predictions without hiding the component equations.

A.6 Dual-Path Training and Expert Supervision

Each training sample is evaluated twice by the same model parameters. The sequence-only pass omits the graph branch, while the sequence–structure pass includes the supplied residue graph. The two passes compute separate activations. The sequence-only decoder samples one expression $\tau_e^{(n)}$ for each expert. The sequence–structure pass replays those expression tokens on its own geometry-conditioned concepts and computes separate decoder activations for them. Formula-policy training uses the sequence-only pass; both passes receive supervised regression and auxiliary losses. During

inference, the shared decoder generates formulas separately from the concepts of the selected input mode.

The current regression objective supervises up to four eligible experts with the lowest absolute error on each path. If no expert passes the validity filters for a sample, the selection code falls back to the available formulas so that the loss remains defined. Let $\mathcal{T}_{n,p}^{(4)}$ denote this set for sample n and path $p \in \{\mathrm{seq}, \mathrm{str}\}$, where str denotes the sequence–structure path. The path-specific regression loss uses the Huber objective (Huber, 1964):

$$\mathcal{L}_{\mathrm{top4}}^{p} = \frac{1}{N} \sum_{n=1}^{N} \omega^{(n)} \frac{1}{|\mathcal{T}_{n,p}^{(4)}|} \sum_{e \in \mathcal{T}_{n,p}^{(4)}} \mathrm{Huber}\left(\hat{y}_e^{(n,p)}, y^{(n)}\right). \tag{24}$$

This top-4 loss spreads direct supervision across several formulas while retaining expert specialization. The sample weight $\omega^{(n)}$ is obtained by label-distribution smoothing on the training set. For integer affinity bin b, let h_b be its count and $\widetilde{h}_b = (G_{\sigma=2} * h)_b$ its Gaussian-smoothed count. We use

$$r_b = (\widetilde{h}_b + 10^{-4})^{-1}, \qquad \omega^{(n)} = \frac{r_{\lfloor y^{(n)} \rfloor}}{|\mathcal{B}|^{-1} \sum_{b \in \mathcal{B}} r_b}, \tag{25}$$

where $\mathcal{B}$ spans the occupied integer-bin range. A second, lower-weight Huber term supervises every valid expert against the measured ΔG and uses the same $\omega^{(n)}$ to improve calibration in sparsely sampled affinity ranges. The knowledge-distillation target for each nonselected expert is the detached mean prediction of the selected experts (Hinton et al., 2015).

For formula generation on the sequence-only path, a GRPO-style policy loss uses negative Huber error as the reward and a leave-one-out expert baseline (Petersen et al., 2021; Shao et al., 2024). A pairwise preference term favors expressions with higher rewards (Rafailov et al., 2023). Invalid expressions, repeated variables, identical subtraction or division subexpressions, algebraic cancellations, low-sensitivity formulas, and batch-constant outputs are excluded or penalized. An expert-usage penalty limits persistent concentration on a small subset of experts during the first part of training.

The auxiliary 44-bin head divides $[-22, 0]$ kcal/mol into intervals of width 0.5 kcal/mol. Its target is a Gaussian-smoothed distribution over bin centers rather than a one-hot class (Geng, 2016), so an error of one bin receives a smaller penalty than an error several bins away. The loss is the Kullback–Leibler divergence between this target and the predicted distribution. Its density multiplier is the attenuated factor $0.7 + 0.3\,\mathrm{clip}(\omega^{(n)}, 0.5, 2.0)$. The three-class head uses weighted cross-entropy for tight ($\Delta G < -13$), normal ($-13 \le \Delta G \le -7$), and weak ($\Delta G > -7$) binding. Samples near the two regime boundaries place more weight on the fine ordinal task, while samples farther from a boundary place more weight on the three-class task.

For each path, the task losses are

$$\mathcal{L}_{\mathrm{seq}} = \mathcal{L}_{\mathrm{top4}}^{\mathrm{seq}} + \mathcal{L}_{\mathrm{GRPO}}^{\mathrm{seq}} + \lambda_{\mathrm{DPO}}\mathcal{L}_{\mathrm{DPO}}^{\mathrm{seq}} + \lambda_{\mathrm{all}}\mathcal{L}_{\mathrm{all}}^{\mathrm{seq}} + \lambda_{\mathrm{KD}}\mathcal{L}_{\mathrm{KD}}^{\mathrm{seq}} + \lambda_{44}\mathcal{L}_{44}^{\mathrm{seq}} + \lambda_{3}\mathcal{L}_{3}^{\mathrm{seq}} + \lambda_{\mathrm{div}}\mathcal{L}_{\mathrm{div}}^{\mathrm{seq}}, \tag{26}$$

$$\mathcal{L}_{\mathrm{str}} = \mathcal{L}_{\mathrm{top4}}^{\mathrm{str}} + \lambda_{\mathrm{all}}\mathcal{L}_{\mathrm{all}}^{\mathrm{str}} + \lambda_{\mathrm{KD}}\mathcal{L}_{\mathrm{KD}}^{\mathrm{str}} + \lambda_{44}\mathcal{L}_{44}^{\mathrm{str}} + \lambda_{3}\mathcal{L}_{3}^{\mathrm{str}} + \mathcal{L}_{\mathrm{interface}} + \lambda_{\mathrm{div}}\mathcal{L}_{\mathrm{div}}^{\mathrm{str}}. \tag{27}$$

Here, $\mathcal{L}_{\mathrm{all}}$ is the all-valid-expert calibration loss, and $\lambda_{\mathrm{div}} = 0.002$ weights the concept-diversity loss defined above. The current model does not use a direct-affinity anchor, a symbolic residual on such an anchor, or a separate structure-to-sequence distillation head.

The two paths are linked at the expert level. Let Ω contain expert predictions that are valid in both passes. The sequence-only prediction is the detached teacher in

$$\mathcal{L}_{\mathrm{cons}} = \frac{1}{|\Omega|} \sum_{(n,e) \in \Omega} \mathrm{SmoothL1}\left(\hat{y}_e^{(n,\mathrm{str})}, \mathrm{sg}\left[\hat{y}_e^{(n,\mathrm{seq})}\right]\right), \tag{28}$$

where sg stops gradients through the consistency target. The full objective under the current configuration is

$$\mathcal{L}_{\mathrm{dual}} = \mathcal{L}_{\mathrm{seq}} + 0.50\,\mathcal{L}_{\mathrm{str}} + 0.03\,\mathcal{L}_{\mathrm{cons}}. \tag{29}$$

The sequence branch remains trainable through $\mathcal{L}_{\mathrm{seq}}$; detachment applies only to its role as the consistency target. This sequence-led objective keeps the sequence path directly supervised and trains the structure path to use geometric information during formula evaluation.

A.7 Validation-Selected Consensus and Reported Outputs

Formula generation is greedy during validation and inference. During training, checkpoint scoring uses the full expert candidate pool rather than a validation-ranked top-K subset. This pool is filtered before averaging. Experts with validation prediction standard deviation below 0.02 are removed when at least one nonconstant expert remains; otherwise, the full pool is restored before applying the PCC gate. The gate then retains experts with positive validation PCC. If none passes, the implementation keeps the expert with the highest finite PCC, or the first otherwise eligible expert when no PCC is finite. Let $g_e \in \{0, 1\}$ mark this global eligibility, and let $v_e^{(n)}$ indicate whether expert e returns a valid prediction for sample n. Define $m^{(n)} = \sum_{e=1}^{16} g_e v_e^{(n)}$. With valid-only averaging enabled, the checkpoint prediction is

$$\hat{y}_{\text{all}}^{(n)} = \begin{cases} \dfrac{\sum_{e=1}^{16} g_e v_e^{(n)} \hat{y}_e^{(n)}}{m^{(n)}}, & m^{(n)} > 0, \\ \dfrac{1}{16} \displaystyle\sum_{e=1}^{16} \hat{y}_e^{(n)}, & m^{(n)} = 0. \end{cases} \tag{30}$$

This criterion is computed independently for the sequence-only and sequence–structure validation passes. The sequence-only mean supplies the primary validation metrics used for checkpoint selection and early stopping, whereas the sequence–structure mean is recorded as the alternate validation result. The logged $K = 16$ denotes the nominal candidate capacity, not the number of predictions contributing to every sample; global filtering and sample-level invalidity can reduce the effective count. The primary `best` checkpoint and early stopping follow validation MAE. Additional retained slots track validation MSE, PCC, RMSE, SPCC, and MAPE, while later archive slots preserve displaced or qualifying MAE-ranked states. The checkpoint suffix therefore identifies its validation selection criterion.

Final test ensemble construction uses validation-ranked consensus. For each checkpoint and input mode, the 16 experts are ranked by validation MAE with an added penalty for invalid predictions. Experts must also have positive validation PCC to enter the consensus; if none passes this gate, the expert with the highest finite validation PCC is retained (or the first eligible expert if no PCC is finite). For a candidate ensemble size K, the valid-only consensus is

$$\hat{y}_K^{(n)} = \frac{\sum_{e \in \mathcal{R}_K} v_e^{(n)} \hat{y}_e^{(n)}}{\sum_{e \in \mathcal{R}_K} v_e^{(n)}}, \qquad \sum_{e \in \mathcal{R}_K} v_e^{(n)} > 0, \tag{31}$$

where $\mathcal{R}_K$ contains the K highest-ranked experts. If none of these experts is valid for a sample, the implementation uses the mean of all valid experts, or the raw all-expert mean if none is valid. The selected size is

$$K^* = \arg \min_{K \in \{K_{\min}, \ldots, K_{\max}\}} \text{MAE}_{\text{val}} \left(\hat{y}_K \right). \tag{32}$$

The ranking and K^* are fixed before test prediction. Test labels and individual test proteins cannot alter either choice. AffinitySym reports the selected consensus, its variation across the included experts, the individual formulas and their predictions, both auxiliary probability distributions, and residue-level attribution scores. For a selected expert, the website computes signed gradient-times-input scores with respect to ESM-3 residue embeddings and, when present, GNN node inputs. Formula-resolved maps restrict the decomposition to latent variables that appear in that expert's expression and backpropagate each variable's local contribution to the residue inputs. The displayed favorable and unfavorable categories summarize these normalized model sensitivities; they are not measured per-residue binding energies. The eligible-pool criterion determines which epochs are retained, while the final consensus for each checkpoint remains validation-ranked.

B Complete Symbolic Expert Set for 1WR1

The sequence-only 1WR1 inference discussed in Section 4.3 produced the following 16 formulas from checkpoint `checkpoint_KNN_trial_1346757343_2.pt`. Define $\mathbf{q}_{e,j} = \mathbf{w}_{e,j} \odot \mathbf{x}_j$ for expert e and latent variable j, and define the pre-reduction vector $\mathbf{r}_e = s_e \mathbf{u}_e + b_e \mathbf{1}$. The scalar

prediction is $\hat{y}_e = \boldsymbol{\pi}_e(\mathbf{z})^\top \mathbf{r}_e$. The equations below expand $\mathbf{r}_e$ with each expert's output scale factored for readability; scalar constants are broadcast across the 32 channels.

$$
\begin{aligned}
\mathbf{r}_{\mathrm{L1}} &= 1.0641\,(\mathbf{q}_{\mathrm{L1},14} - \mathbf{q}_{\mathrm{L1},7}) - 10.0132,\\
\mathbf{r}_{\mathrm{L2}} &= 1.0012\,(\mathbf{q}_{\mathrm{L2},14} - \mathbf{q}_{\mathrm{L2},7}) - 9.9405,\\
\mathbf{r}_{\mathrm{L3}} &= 1.0080\,(-\mathbf{q}_{\mathrm{L3},1} + \mathbf{q}_{\mathrm{L3},14} - \mathbf{q}_{\mathrm{L3},4}) - 9.9813,\\
\mathbf{r}_{\mathrm{L4}} &= 1.0228\,(-\mathbf{q}_{\mathrm{L4},1} + \mathbf{q}_{\mathrm{L4},14} - \mathbf{q}_{\mathrm{L4},4}) - 9.9598,\\
\mathbf{r}_{\mathrm{L5}} &= 1.0281\,(-\mathbf{q}_{\mathrm{L5},1} + \mathbf{q}_{\mathrm{L5},14} - \mathbf{q}_{\mathrm{L5},4}) - 9.9930,\\
\mathbf{r}_{\mathrm{L6}} &= 0.9998\,(-\mathbf{q}_{\mathrm{L6},1} + \mathbf{q}_{\mathrm{L6},14} - \mathbf{q}_{\mathrm{L6},4}) - 9.9648,\\
\mathbf{r}_{\mathrm{L7}} &= 1.0395\,(-\mathbf{q}_{\mathrm{L7},1} + \mathbf{q}_{\mathrm{L7},14} - \mathbf{q}_{\mathrm{L7},4}) - 9.9925,\\
\mathbf{r}_{\mathrm{L8}} &= 1.0014\,(-\mathbf{q}_{\mathrm{L8},1} + \mathbf{q}_{\mathrm{L8},14} - \mathbf{q}_{\mathrm{L8},4}) - 9.9943.
\end{aligned}
$$

$$
\begin{aligned}
\mathbf{r}_{\mathrm{L9}} &= 0.9979\,[-\mathbf{q}_{\mathrm{L9},15} \odot (\mathbf{q}_{\mathrm{L9},0} - \mathbf{q}_{\mathrm{L9},6}) + \mathbf{q}_{\mathrm{L9},14} - \mathbf{q}_{\mathrm{L9},4}] - 9.9547,\\
\mathbf{r}_{\mathrm{L10}} &= 1.0095\,[-\mathbf{q}_{\mathrm{L10},15} \odot (\mathbf{q}_{\mathrm{L10},0} - \mathbf{q}_{\mathrm{L10},6}) + \mathbf{q}_{\mathrm{L10},14} - \mathbf{q}_{\mathrm{L10},4}] - 9.9447,\\
\mathbf{r}_{\mathrm{L11}} &= 1.0638\,[-\mathbf{q}_{\mathrm{L11},15} \odot (\mathbf{q}_{\mathrm{L11},0} - \mathbf{q}_{\mathrm{L11},6}) + \mathbf{q}_{\mathrm{L11},14} - \mathbf{q}_{\mathrm{L11},4}] - 9.9862,\\
\mathbf{r}_{\mathrm{L12}} &= 0.9996\,[-\mathbf{q}_{\mathrm{L12},15} \odot (\mathbf{q}_{\mathrm{L12},0} - \mathbf{q}_{\mathrm{L12},6}) + \mathbf{q}_{\mathrm{L12},14} - \mathbf{q}_{\mathrm{L12},4}] - 9.9525,\\
\mathbf{r}_{\mathrm{L13}} &= 1.0478\,[-\mathbf{q}_{\mathrm{L13},15} \odot (\mathbf{q}_{\mathrm{L13},0} - \mathbf{q}_{\mathrm{L13},6}) + \mathbf{q}_{\mathrm{L13},14} - \mathbf{q}_{\mathrm{L13},4}] - 10.0137,\\
\mathbf{r}_{\mathrm{L14}} &= 0.9829\,[-\mathbf{q}_{\mathrm{L14},15} \odot (\mathbf{q}_{\mathrm{L14},0} - \mathbf{q}_{\mathrm{L14},6}) + \mathbf{q}_{\mathrm{L14},14} - \mathbf{q}_{\mathrm{L14},4}] - 9.9715,\\
\mathbf{r}_{\mathrm{L15}} &= 1.0261\,[-\mathbf{q}_{\mathrm{L15},15} \odot (\mathbf{q}_{\mathrm{L15},0} - \mathbf{q}_{\mathrm{L15},6}) + \mathbf{q}_{\mathrm{L15},14} - \mathbf{q}_{\mathrm{L15},4}] - 10.0234,\\
\mathbf{r}_{\mathrm{L16}} &= 1.0583\,[-\mathbf{q}_{\mathrm{L16},15} \odot (\mathbf{q}_{\mathrm{L16},0} - \mathbf{q}_{\mathrm{L16},6}) + \mathbf{q}_{\mathrm{L16},14} - \mathbf{q}_{\mathrm{L16},4}] - 10.0376.
\end{aligned}
$$

C Dataset Construction and Benchmark Provenance

C.1 PDBbind Affinity Records

The starting point was the 2,852-entry protein–protein general index from PDBbind v2020 (Liu et al., 2017), restricted to entries with a corresponding cleaned PDB structure. The preprocessing script reads the affinity field in each index row. It discards 171 K_i and 63 IC_{50} records, then rejects 14 `Kd<` and 16 `Kd>` records because a bound does not define a single regression target. The remaining 2,588 entries have `Kd=` or `Kd~` measurements. The parser accepts fM, pM, nM, μM, mM, and M, converts each measurement to molar units, and computes ΔG with Equation 1 at 298 K. An approximate measurement marked `Kd~` is treated as its stated value; the approximation marker is not retained as a separate uncertainty estimate.

C.2 Partner Chains and Benchmark Exclusion

The ProAffinity-GNN supplementary table `ci4c01850_si_002.csv` provides the partner-chain annotations (Zhou et al., 2024). Of its 2,283 entries, 2,069 are labeled K_d. For each PDB code, we retain one chain annotation and parse the first two groups in `pairwise composition` as receptor and ligand. PDB codes are matched without regard to case. Requiring both chain groups leaves 2,056 of the 2,588 affinity records.

The local Test Set 1 labels come from the 82-entry K_d subset of the PDBbind benchmark used by PPI-Affinity (Romero-Molina et al., 2022); the ProAffinity-GNN supplement supplies the corresponding pKa table `ci4c01850_si_004.csv` (Zhou et al., 2024). For the local test file, pKa is converted to $K_d = 10^{-\mathrm{pKa}}$ M and then to ΔG at 298 K. Test Set 2 comes from a SKEMPI v2.0 cohort first used by PPI-Affinity and subsequently deduplicated to 26 wild-type and 140 mutant records by ProAffinity-GNN; its supplementary `ci4c01850_si_005.csv` contains the affinity labels used here (Romero-Molina et al., 2022; Zhou et al., 2024; Jankauskaitė et al., 2019). Multi_PPBA evaluated on these same 82- and 166-entry cohorts (Mo et al., 2025). Our Test Set 1/2 naming follows Multi_PPBA; ProAffinity-GNN calls the 82-entry PDBbind cohort its second test set and lists the SKEMPI cohort separately. The test files are external benchmarks, not samples assigned by our train–validation split.

Before splitting, we exclude training candidates that share a PDB code with either benchmark. For Test Set 2, we take the first four characters of each wild-type or mutant identifier as its base PDB code. Test Set 1 contributes 82 unique codes, 81 of which match the 2,056 candidates; Test Set 2 contributes 26 base codes, one of which matches. There are 108 distinct test codes in their union and 82 matching training candidates. Removing these candidates leaves 1,974 complexes. This is an exact-PDB exclusion, not a sequence-homology split.

C.3 Training and Validation Split

For each training seed, the 1,974 remaining complexes are shuffled and divided into extreme ($\Delta G < -13$ or $\Delta G > -7$ kcal/mol) and central affinity groups. The first 80% of each shuffled group, rounded down to an integer count, is assigned to training; the remainder is assigned to validation. The groups are recombined and shuffled again. This procedure yields 1,578 training and 396 validation complexes. Each PDB code appears once in the development set, so a complex cannot occur in both partitions. The reported multi-seed experiments use a fresh seeded split for each run rather than one fixed validation partition. Runtime input validation removes four records after the seeded split. The inherited source tables contain three protein–nucleic-acid complexes—`4WB2`, `3SWP`, and `4R8I`—whose chain annotations are correct but whose ligand chains lack the protein N/C_α/C backbone required by AffinitySym; they therefore fall outside our PPI scope. `1TZN` is instead a PPI with the unusual case-sensitive partner IDs `A` and `a`, the only such collision in the candidate set. Our parser normalized both to `A`, so we initially grouped this failure with the other input exclusions. We identified the distinction after completing the multi-seed and ablation experiments and retained one consistent set of reported runs rather than rerunning every configuration for a single additional complex. This omission does not change the splitting or evaluation protocol. The four exclusions leave 1,970 usable complexes; their training–validation allocation varies by seed. Accordingly, the effective training/validation counts can be 1,574/396, 1,575/395, 1,576/394, 1,577/393, or 1,578/392, with every run following the same preprocessing and splitting protocol.