PACT: A SELF-EXPLAINING SCORING FUNCTION FOR PROTEIN–LIGAND BINDING AFFINITY WITH CONTEXT-DEPENDENT INTERACTION TERMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Structure-based affinity predictors return one number, and an explanation can be added afterward using post-hoc attributions or attention weights that are not expressed in affinity units and whose reliability is contested. Empirical scoring functions instead sum named physical terms but give each term one global weight, pricing a solvent-exposed hydrogen bond like a buried one. We introduce PACT, a predictor whose output is exactly decomposed into signed ligand-atom-residue contributions in pK over twelve interaction channels, with local chemistry and pocket context rescaling each contribution. Additivity determines how affinity is distributed over contacts, not whether that distribution can be trusted, so we evaluate the stability of the contributions under model retraining and contact removal. On a leakage-reduced split, PACT matches the accuracy of the strongest baselines we evaluated while producing faithful and reproducible contributions.

1 INTRODUCTION

Drug-target affinity (DTA) prediction has become an important part of early drug discovery. In structure-based prediction, a model estimates how strongly a ligand binds to a protein from their three-dimensional complex. Docking searches for ligand poses in the binding site, here we only score a given pose (Trott & Olson, 2010). A useful model should not only predict binding strength but also explain how each contact, residue or functional group contributes to the prediction. Chemists need the explanation to decide which parts of a ligand to keep in the next design cycle. Most models, however, are optimized for accuracy alone and treat explanation as a separate task, if at all.

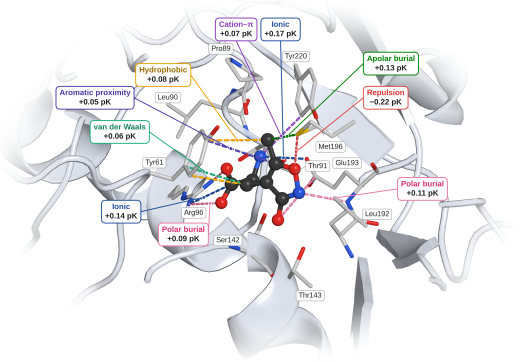


Figure 1: PACT scores a given pose. Each box is one interaction at one residue and its contribution in pK. With an intercept and a ligand term, these contributions sum to the prediction.

Explanations usually come from one of two sources. Post-hoc methods, such as gradients or SHAP, analyze a trained model after it has made its prediction. Gradients show which inputs the prediction is sensitive to, but sensitivity is not the same as contribution. Some attribution methods also fail basic sanity and stability checks (Adebayo et al., 2018; Ghorbani et al., 2019). The second source is attention. Attention weights can rank protein-ligand interactions, but they do not necessarily explain each interaction's effect on the prediction or provide contributions in affinity units (Jain & Wallace, 2019; Wiegreffe & Pinter, 2019). Some post-hoc methods, such as SHAP, produce additive contributions in the units of the explained output (Lundberg & Lee, 2017), but these assignments depend on the explanation procedure and reference choice. Rudin (2019) argues that models used for high-stakes decisions should be interpretable by design rather than explained afterward, because a post-hoc explanation can be wrong about what the model

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.

actually computes. PACT's contributions are instead the named chemical terms from which the prediction is computed; what sets them apart is not their units but that they are the model's own terms.

Explanations matter all the more because models can score well by learning dataset-level patterns instead of protein–ligand interactions (Durant et al., 2025; Volkov et al., 2022; Sieg et al., 2019). Leakage-reduced splits help limit these shortcuts (Li et al., 2026), but do not help in determining what contributes to an individual prediction.

Before neural networks, empirical scoring functions predicted binding affinity as a sum of named physical terms associated with interactions or desolvation (Böhm, 1994; Wang et al., 2002). Each term usually has a single global weight learned on a training set. Neural models can learn richer descriptions of local geometry (Moon et al., 2024; Meli et al., 2021). One global weight per term can be too coarse even within one complex: a hydrogen bond in a solvent-exposed region is not equivalent to one in a buried, preorganized pocket. Context-dependent weights can address this limitation, but they can affect the stability of the resulting contributions.

Our approach retains named chemical interaction terms while allowing bounded contextual adjustments to their contributions. An exact decomposition shows how the model computes its score, but not whether the assignment is reproducible, and it is not a unique physical decomposition. Affinity benchmarks mainly evaluate accuracy. They do not test whether interaction contributions are reproducible across retraining, consistent with interventions on the model, or chemically meaningful, which matters most when learned context can shift contributions without changing accuracy. We separate two properties (Jacovi & Goldberg, 2020): faithfulness (do the contributions describe what the model computes?) and chemical plausibility (do they match recognizable interactions?).

We introduce PACT (Protein-ligand Additive Contact Terms), a structure-based affinity model that re-weights named chemical interactions with bounded, context-dependent local corrections. Our contributions are: (i) a model with 4,784 parameters that decomposes each prediction exactly into an intercept, a ligand-level term and signed ligand-atom–residue contributions in affinity units; (ii) an evaluation protocol for intrinsic affinity explanations, covering reproducibility across retraining, agreement with contact removal and chemical agreement, applied to ten baselines on the leakage-reduced LP-PDBbind split (Li et al., 2026); (iii) evidence that PACT matches the strongest baselines in accuracy while giving the most reproducible contributions.

2 Related Work

We focus on structure-based DTA and group previous work by how explanations are obtained: after prediction, through attention, or from additive scoring terms. We also consider whether those terms have chemical meaning and use affinity units.

Post-hoc attribution and attention. Three-dimensional and equivariant networks (Schütt et al., 2017; Satorras et al., 2021) and interaction-graph networks (Jiang et al., 2021; Yang et al., 2023) lack a native interaction-level decomposition. Post-hoc attributions can be unstable or disagree across methods (Ghorbani et al., 2019; Krishna et al., 2024). LINKER and CheapNet use attention (Pham et al., 2026; Lim et al., 2025), which assigns weights to input features. However, a higher attention weight does not necessarily mean a greater effect on the prediction (Jain & Wallace, 2019), and these weights are not contributions in affinity units.

Additive models with unnamed terms. Additive and self-explaining models build explanations into the prediction rule (Agarwal et al., 2021; Bechler-Speicher et al., 2024; Alvarez-Melis & Jaakkola, 2018). For affinity prediction, Meli et al. (2021) sum learned contributions per ligand and binding-site atom, but do not assign them to predefined interaction classes. Concept supervision adds semantic structure (Koh et al., 2020), although concepts may encode information beyond their intended meanings (Mahinpei et al., 2021; Havasi et al., 2022).

Named chemical terms. Empirical scoring functions sum physical terms with globally fitted coefficients (Trott & Olson, 2010; Eberhardt et al., 2021), limiting adaptation to local context. PIGNet2 retains explicit interaction forms with learned parameters (Moon et al., 2024). CENsible uses a convolutional network to generate per-complex weights for named terms (Bhatt et al., 2024).

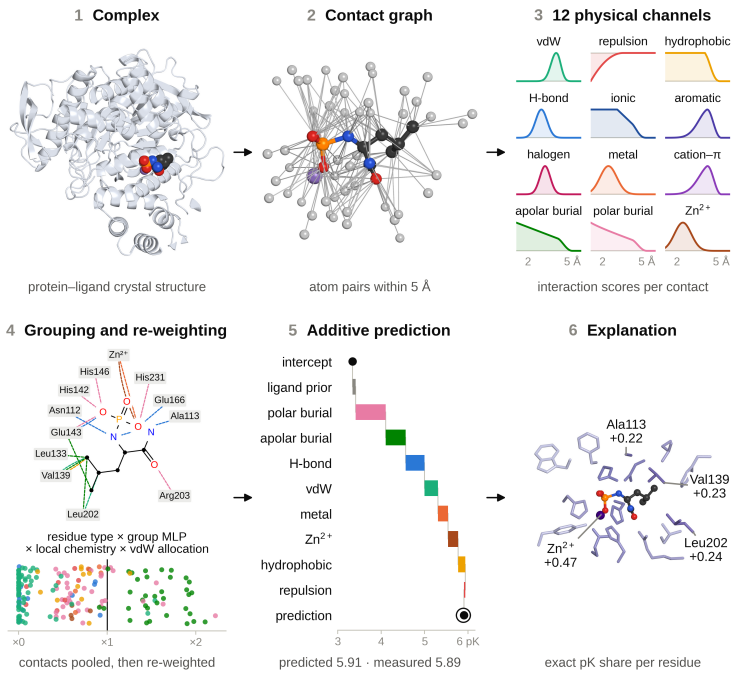


Figure 2: PACT on test complex 2TMN (thermolysin with a zinc-bound phosphoramidate inhibitor). Atom pairs within 5 Å (2) are scored by twelve interaction channels (3), pooled into ligand-atom–residue groups and re-weighted by bounded local corrections (4). The contributions add up to the prediction (5) and give each residue an exact share in pK (6).

Evaluation of explanations. General XAI research emphasizes faithfulness to the model and explanation stability (Jacovi & Goldberg, 2020; Nauta et al., 2023). For affinity prediction, we test whether residue rankings and signed contributions remain consistent across retraining and small changes in geometry, and whether assigned contributions match prediction changes after interventions. We also compare named interaction channels with ProLIF to assess chemical agreement.

3 METHODS

PACT writes each prediction as an exact sum of contributions from named interaction channels in ligand-atom–residue groups. We describe how these contributions are built, combined and trained.

3.1 CHEMICAL INTERACTION REPRESENTATION

PACT uses the supplied complex coordinates without structural optimization or protonation-state enumeration. The representation includes nearby metal ions and excludes water molecules.

For every ligand heavy atom and protein atom within 5 Å, we compute twelve interaction channels: van der Waals attraction, steric repulsion, hydrophobic, hydrogen bond, ionic, aromatic, halogen–acceptor, metal, cation–aromatic, apolar burial, polar burial and zinc. Each channel applies only to chemically eligible atom pairs and responds smoothly to distance; the hydrogen-bond channel also depends on donor direction. Contacts are pooled into ligand-atom–residue groups by retaining the strongest response per channel (minimum for signed repulsion). This limits repeated counting within a residue while preserving ligand-atom identity for fragment explanations. Before pooling, each contact is scaled by a bounded multiplier that depends on the residue type. The pooled value B_{gc} determines whether channel c can contribute in group g, and with which sign. Typing, basis functions, and pooling details are in Appendix A.1. Figure 2 shows the overall arrangement.

3.2 CONTEXTUAL CONTRIBUTIONS AND AFFINITY PREDICTION

Each label is $\mathrm{pK} = -\log_{10}(K/1\,\mathrm{M})$, where K is an experimentally measured binding constant (K_d, K_i or IC_{50}) recorded in PDBbind; a higher pK means tighter binding. Validation and test labels use only K_d or K_i (Appendix B.1). One pK unit is roughly 1.36 kcal mol^{-1} of binding free energy at room temperature, so contributions that add up in pK can be read as additive energies. This is an approximation: PDBbind does not record the measurement temperature, and IC_{50} also depends on how the experiment was run. The decomposition is still a modeling choice: solvation and entropy are not pairwise, so we decompose the model's estimate, not the thermodynamics of binding.

$$\widehat{y}(x) = b + G(x) + \sum_{g \in \mathcal{G}(x)} \sum_{c=1}^{12} \sigma_y w_c B_{gc} \ell_{gc} \alpha_{gc} (1 + q_{gc}), \quad w_c = \mathrm{softplus}(v_c). \tag{1}$$

All terms are reported in pK (σ_y converts from standardized training units), and larger values mean stronger binding. Steric repulsion has a negative basis and all other channels nonnegative ones. The channel names label model components, not independent physical energies. Each channel carries one unconstrained learned scalar v_c, so the global channel coefficient w_c is positive by construction and cannot flip a channel's sign. The local coefficient $\ell_{gc} \in [0, 2]$ adjusts strength using ligand-atom and residue chemistry. A small network computes a bounded update $1 + q_{gc} \in [0.75, 1.25]$ from all channels of the same group, so one interaction can modulate another, but no information passes between groups. Context can thus change how strong a contribution is, but not where it can occur. SEAL, an interpretable graph network for molecular property prediction, uses the same idea: it limits message passing between molecular fragments so that each fragment's contribution reflects that fragment and not the rest of the molecule (Musiał et al., 2025).

For van der Waals attraction, the groups sharing a ligand atom compete through a sparsemax allocation α_{gc} that can also leave part of the mass unassigned, so one atom cannot collect attraction from every nearby residue. Other channels use $\alpha_{gc} = 1$. All multipliers are nonnegative, so context can weaken or silence a contribution but never reverse its sign. Bounding the multipliers does not by itself make the explanations robust or reproducible; we test both in Section 5.

A separate ligand-level term G, bounded by $0.3\sigma_y$, captures size, flexibility, composition and charge, so these global effects are not attributed to individual contacts. Residue and fragment explanations are sums of the corresponding A_{gc}; together with b and G they add up exactly to the prediction. Appendix A.6 gives the parameterization of each factor.

3.3 TRAINING AND ATTRIBUTION REGULARIZATION

The loss combines a sequence-weighted squared error, a pairwise ranking term and two regularizers on the attributions, which make the explanation locally faithful and stable during training (Alvarez-Melis & Jaakkola, 2018):

$$L = L_\mathrm{aff} + 0.2 L_\mathrm{rank} + \lambda_\mathrm{local} L_\mathrm{local} + \lambda_\mathrm{transport} L_\mathrm{transport}. \tag{2}$$

Additivity alone does not stop the context factors from reweighting interactions arbitrarily. The local term makes each contribution's effective coefficient match the model's sensitivity to that contribution's basis, differentiating through the group update and the allocation only. The transport term keeps these coefficients stable when atoms are jittered by 0.05 Å, while still letting contributions follow their bases. Neither term uses measured interaction energies.

We standardize targets using training statistics and optimize with AdamW (learning rate 10^{-3}, batch size 32). Checkpoints are selected by validation sequence-balanced RMSE with early stopping. Formal regularizer definitions, weights, perturbation caching, and remaining settings are in Appendix A.7.

4 EXPERIMENTAL SETUP

We evaluate whether the structured predictor remains competitive and whether its contributions support reliable interpretation. The explanation experiments assess properties that do not follow

from additivity alone: reproducibility across retraining, agreement with model interventions, and correspondence with chemical interactions.

4.1 Datasets and predictive baselines

We use LP-PDBbind (Li et al., 2026) under its recommended protocol: the training split is filtered to the CL1 cleanliness level and the validation and test splits to the stricter CL2, and covalent complexes are excluded throughout. This yields 7,393 training, 958 validation, and 2,171 test complexes. BDB2020+, EGFR, and Mpro provide external evaluations. For the CASF-2016 column of Table 1, models are instead trained under the CleanSplit protocol (Graber et al., 2025) and evaluated on the 144-complex independent CASF-2016 subset. We report RMSE, Pearson correlation, and Spearman correlation on LP-PDBbind, and Pearson correlation on the independent CASF subset and external datasets.

Baselines are selected to cover both predictive alternatives and relevant explanation mechanisms. RF-Score provides a simple scoring reference (Ballester & Mitchell, 2010); SchNet, EGNN, IGN, and GIGN represent geometric neural predictors (Schütt et al., 2017; Satorras et al., 2021; Jiang et al., 2021; Yang et al., 2023). PIGNet2 and GRIPHIN incorporate chemical structure into neural scoring (Moon et al., 2024; Rose et al., 2026), while CENsible provides a direct comparison with context-dependent weights over named scoring terms (Bhatt et al., 2024). LINKER and CheapNet provide interaction-attention comparisons (Pham et al., 2026; Lim et al., 2025). Size-based predictors test how much performance can be obtained without modeling protein–ligand contacts.

Each model is trained with six seeds and selected on validation data; Appendix B.1 lists seeds and cohorts, and Appendix E the reproduction protocol.

4.2 Explanation evaluation

We ask three questions of an intrinsic explanation: (1) does retraining reproduce the signed contributions? (2) do the contributions match the change in output when contacts are removed? (3) do the channels carry chemical information beyond their predefined bases? Residues are ranked by absolute contribution, while numerical comparisons keep the sign. Attention weights are never treated as contributions; where a baseline provides them, we use them only as an alternative ranking (Appendix D.3).

For reproducibility, we compare the five highest-ranked residues across retrained models (overlap and exact agreement) and the spread of the signed contributions: the first tells whether the ranking is stable, the second whether the values are. Small rigid displacements of the ligand probe sensitivity to geometry.

Interventions test whether the highest-ranked contacts matter to the model (Hama et al., 2023): we remove the contacts of the selected residues, recompute the score at fixed coordinates, and compare with removing the nearest or randomly chosen residues. Because the contributions are in pK, we can also compare each assigned contribution with the observed change in the prediction. Preservation tests, which keep only the selected contacts, are defined in Appendix D.5.

Chemical agreement is measured against ProLIF interaction annotations (Bouysset & Fiorucci, 2021), with distance and the untrained chemical bases as controls. ProLIF records whether an interaction is present, not how much it contributes.

4.3 Ablations

We report two sets of ablations. The coefficient ladder adds the multipliers one at a time: global channel weights only, then the per-group local coefficient, then the full model. The constraint study removes one restriction at a time: nonnegative channel weights, a sign-preserving local coefficient, and the rule that every learned factor multiplies a named basis (tested by adding free-form terms). Details are in Appendix C.3.

Table 1: Predictive performance. LP-PDBbind: mean ± SD over seeds on the 2,133 test complexes that every model except PIGNet2 could score (Appendices B.1 and E.1). CASF: Pearson r on the 144-complex independent subset, ± across CleanSplit folds. Appendix B.

	LP-PDBbind				CASF-2016 indep.
Model	RMSE ↓	r ↑	ρ ↑	CI ↑	r ↑
Ligand-size control	1.689	0.459	0.421	0.644	0.632
RF-Score	1.526 ± 0.002	0.579 ± 0.001	0.544 ± 0.001	0.691 ± 0.001	0.794 ± 0.005
SchNet	1.571 ± 0.041	0.529 ± 0.021	0.498 ± 0.019	0.674 ± 0.007	0.756 ± 0.040
EGNN	1.639 ± 0.078	0.498 ± 0.044	0.447 ± 0.044	0.656 ± 0.015	0.771 ± 0.054
IGN	$\underline{1.510} \pm 0.016$	0.580 ± 0.013	0.541 ± 0.013	0.689 ± 0.005	0.702 ± 0.016
GIGN	1.617 ± 0.059	0.532 ± 0.036	0.495 ± 0.040	0.673 ± 0.014	0.759 ± 0.014
PIGNet2	1.817 ± 0.054	0.484 ± 0.019	0.495 ± 0.023	0.672 ± 0.009	0.734 ± 0.018
GRIPHIN	1.659 ± 0.112	0.515 ± 0.045	0.471 ± 0.042	0.664 ± 0.015	X
CENsible	1.586 ± 0.011	$\underline{0.588} \pm 0.004$	0.548 ± 0.003	0.692 ± 0.001	0.770 ± 0.005
LINKER	1.608 ± 0.011	0.489 ± 0.009	0.484 ± 0.015	0.667 ± 0.005	X
CheapNet	1.518 ± 0.050	$\mathbf{0.592} \pm 0.021$	$\mathbf{0.557} \pm 0.022$	$\mathbf{0.696} \pm 0.009$	$\mathbf{0.831} \pm 0.003$
PACT	$\mathbf{1.504} \pm 0.005$	0.584 ± 0.001	$\underline{0.553} \pm 0.002$	$\underline{0.694} \pm 0.001$	$\underline{0.795} \pm 0.003$

5 RESULTS

Where repeated measurements are available, results report mean ± standard deviation across seeds. X denotes an unavailable result, – an undefined or inapplicable comparison. Bold and underline mark the best and second-best mean, ranked on unrounded values, within the comparison each table specifies, highlighting does not imply statistical significance.

5.1 PREDICTIVE PERFORMANCE

PACT reaches 1.504 ± 0.005 pK RMSE and a concordance index of 0.694 ± 0.001, matching IGN and CheapNet while retaining an explicit chemical decomposition (Table 1). On the independent CASF-2016 subset, PACT reaches $r = 0.795$, close to RF-Score (0.794) and behind CheapNet (0.831); ligand size alone already gives 0.632, so this column says little about generalization. PACT is also small, with 4,784 trained parameters against 10^5–10^8 for the neural baselines (Appendix E.3). Transfer to external sets is weaker. On BDB2020+ PACT reaches $r = 0.379$, lower than every baseline tested there except SchNet, but on Mpro it is second after RF-Score ($r = 0.582$; Appendix B).

5.2 REPRODUCIBILITY AND GEOMETRIC SENSITIVITY

Retrained PACT models select the same residues and assign them nearly the same values (Table 2): the top five residues are identical in 83.4% of seed pairs and overlap by 96.6% on average, and the signed contributions vary by 0.0062 pK across seeds. Each baseline row evaluates a model together with its explanation method (usually occlusion), not the architecture alone.

The explanations are also stable under small changes in geometry. After a rigid displacement of the ligand by 0.10 Å RMS, PACT keeps the same top five residues in $78.8 \pm 0.6\%$ of cases, and the prediction changes by 0.0270 ± 0.0004 pK on average. Pure translation gives almost the same result ($80.2 \pm 1.4\%$, 0.0271 ± 0.0003 pK), so adding rotation makes no difference.

5.3 INTERVENTION BEHAVIOR

Because PACT's contributions are in pK, they can be checked directly against the model: we remove a residue's contacts, recompute the score at fixed coordinates, and compare the change with the assigned contribution. Over removals of $k \in \{1, 3, 5, 8\}$ residues, PACT's contributions differ from the observed change by 0.191 ± 0.013 pK on average, with a regression slope of 0.9139 ± 0.0025 and the correct sign in all 999 interventions for every seed (Table 3). The next-best discrepancy is 1.7 times larger, and the next-best slope is 0.13 further from one.

Table 2: Explanation reproducibility and geometric sensitivity. Exact@5 and Overlap@5 are agreement across seed pairs; SD is the mean cross-seed standard deviation of signed residue contributions, in pK. Perturbation columns apply a 0.10 Å RMS rigid-body ligand displacement. Appendix C.

	Across retraining			Rigid-body perturbation	
Model	Exact@5 ↑	Overlap@5 ↑	SD ↓	$\lvert\Delta\widehat{y}\rvert$ ↓	Exact@5 ↑
RF-Score	0.104	0.592	0.0407	<u>0.0286</u>	0.203
SchNet	0.019	0.538	0.1499	0.1107	0.376
EGNN	0.019	0.537	0.1425	0.1416	0.291
IGN	0.061	0.629	0.1187	0.0664	0.685
GIGN	0.054	0.596	0.1473	0.1285	0.416
PIGNet2	<u>0.659</u>	<u>0.927</u>	<u>0.0212</u>	0.0907	<u>0.753</u>
CheapNet	0.036	0.593	0.1196	0.1676	0.345
PACT	**0.834**	**0.966**	**0.0062**	**0.0270** ± 0.0004	**0.788** ± 0.006

Table 3: Contribution–effect agreement under residue removal, pooled over $k \in \{1, 3, 5, 8\}$. Discrepancy is the mean absolute difference between assigned contribution and observed effect in pK, relative divides it by the preceding column, and sign agreement is the fraction sharing a sign. Some models are excluded for the reasons given in Appendix D.

Model	Mean $\lvert$effect$\rvert$ (pK)	Discrepancy (pK) ↓	Relative ↓	Slope (target 1)	Sign agr. ↑
RF-Score	0.461	<u>0.323</u>	0.701	0.410	0.890
CheapNet	1.370	0.328	0.239	<u>0.780</u>	0.973
EGNN	1.247	0.342	0.274	0.654	0.972
IGN	1.942	0.457	<u>0.235</u>	0.763	<u>0.999</u>
SchNet	1.123	0.634	0.565	0.460	0.955
GIGN	1.110	0.938	0.845	0.371	0.910
PIGNet2	1.133	1.591	1.404	0.490	0.922
PACT	1.114 ± 0.018	**0.191** ± 0.013	**0.172** ± 0.010	**0.9139** ± 0.0025	**1.000**

This is not because PACT's removal effects are small: at 1.114 ± 0.018 pK they are mid-range for the table. Relative to the effect size, PACT's error is $17.2 \pm 1.0\%$, against 23.5% for the next-best model. RF-Score's small discrepancy (0.323 pK) is 70% of its effect (0.461 pK), and LINKER's (0.0037 pK) comes with a top-five effect of only 0.0588 pK (Appendix D.1).

If anything, the comparison favors the baselines above the rule: they rank residues by their own measured deletion effects, so their agreement at $k = 1$ is exact by construction. PIGNet2 is the most informative comparison, because it is the only other model whose contributions are not derived from deletions. Its per-pair energies sum exactly to the prediction, yet removing pairs changes the prediction by a different amount, and its discrepancy is larger than the effect itself. An exact sum is therefore not enough: the contributions must also match what happens when contacts are removed.

Table 4: AUROC against ProLIF annotations on 62,552 ligand-atom–residue candidate pairs from 600 complexes. Distance ranks by proximity. Appendix D.7.

Interaction	PACT	PIGNet2	Distance
Aromatic	**0.972**	–	<u>0.637</u>
Cation–aromatic	**0.967**	–	<u>0.636</u>
Halogen	<u>0.9327</u>	–	**0.9334**
Hydrogen bond	0.844	<u>0.962</u>	**0.979**
Hydrophobic	<u>0.910</u>	**0.963**	0.659
Metal	**0.9996**	0.9746	<u>0.9982</u>
Van der Waals	<u>0.734</u>	0.701	**0.946**

Rankings differ much less between methods. Removing PACT's five highest-ranked residues changes the prediction 1.534 ± 0.007 times as much as removing the five nearest, and PACT beats this proximity baseline for every k. Ranking PACT's residues by occlusion instead gives almost the same ratio (1.524 ± 0.008), although it changes the discrepancy (Appendices D.5 and D.4). The methods thus differ in the values they assign, not in which residues they rank first.

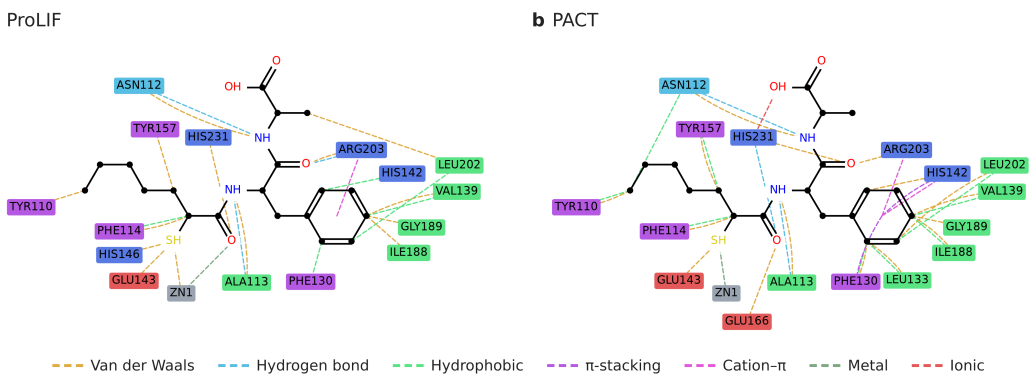


Figure 3: Thermolysin–thiol inhibitor contacts (PDB 1QF1, test set) from (a) ProLIF and (b) PACT.

5.4 CHEMICAL AGREEMENT

PACT's channels agree well with ProLIF annotations of the same interactions (Table 4; example in Figure 3), with AUROC above 0.9 for aromatic, cation–aromatic, halogen, hydrophobic and metal contacts. ProLIF is a standard detector, not ground truth, so this is only a consistency check.

Agreement depends on the interaction: distance alone ranks hydrogen bonds and van der Waals contacts better, and PIGNet2 does better on hydrogen bonds and hydrophobic contacts. Much of the agreement is already present in the untrained chemical bases (Appendix D.7). When poses change, however, six of eight channels track the recovery of their own interaction class better than the total score or the contact count (Appendix D.8).

5.5 ARCHITECTURAL ABLATIONS

Most of the accuracy gain comes from the per-group local coefficient, which lowers RMSE from 1.579 to 1.513 pK; the contact multiplier and group update add a further 0.005 pK (Table 5). The model with global weights only is the most reproducible, so accuracy and reproducibility trade off along the ladder.

Relaxing the constraints mainly costs reproducibility. Signed channel weights give the lowest RMSE we measured (1.4925 pK), but lower top-five agreement from 0.834 to 0.636 and more than double the cross-seed variability. Free-form terms outside the named channels lower agreement to 0.666 and 0.603 and do not improve RMSE. Letting the local coefficient reverse a basis lowers agreement to 0.792 without improving RMSE. Appendix C.3 reports the width ladders, component removals, and the fidelity evaluation of the attribution objectives.

Figure 4 shows how the local coefficient is applied. Each point is one contact's final contribution, colored by ℓ_{gc}; the line is the contribution without context (basis times channel weight). The coefficient varies between contacts at the same distance in the same channel, amplifying some and damping others, which a channel average would hide. Its typical size depends on the channel: the median is 1.32 for apolar burial and 0.74 for van der Waals attraction, but only 0.08 for hydrophobic and 0.09 for aromatic contacts, which the model effectively switches off. Appendix C.5 shows all twelve channels.

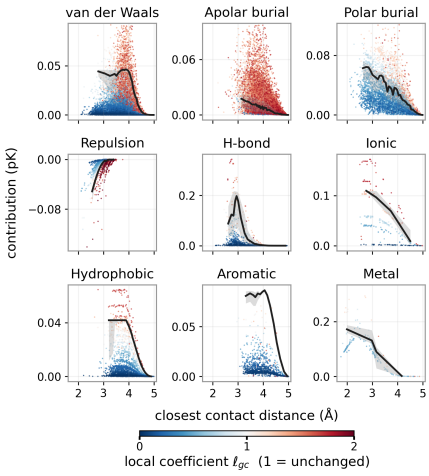


Figure 4: Contributions against contact distance.

Table 5: Ablations on all 2,171 test complexes, six seeds, so the PACT row differs slightly from Table 1 (2,133 complexes). RMSE is mean ± SD. Exact@5 and contribution SD are agreement across seed pairs. Each constraint row relaxes one setting, except the last, which combines two.

Configuration	RMSE ↓	Pearson r ↑	Exact@5 ↑	SD (pK) ↓
Coefficient ladder				
Global weights + allocation	1.5793 ± 0.0041	0.5383	**0.9138**	**0.00299**
+ local coefficient	1.5131 ± 0.0050	0.5798	0.8322	0.00613
PACT	1.5084 ± 0.0052	<u>0.5828</u>	0.8343	0.00620
Constraints relaxed				
Signed channel weights	**1.4925**±0.0050	**0.5860**	0.6365	0.01329
ℓ may reverse sign	1.5159 ± 0.0064	0.5745	0.7919	0.00716
+ free-form ligand term	1.5267 ± 0.0121	0.5627	0.6660	0.00977
+ two free-form terms	1.5189 ± 0.0068	0.5672	0.6028	0.01281

6 DISCUSSION AND LIMITATIONS

PACT's explanations are faithful to the model in three measurable ways: they add up exactly to the prediction, they survive retraining, and they match the effect of removing contacts. They are not an energy decomposition of binding: the channels are components of a model fitted to affinity, and solvation, entropy and induced fit are not pairwise. Much of the agreement with ProLIF is also already present in the untrained bases (Appendix D.7). The contributions should therefore be read as a chemically labeled account of what the model computes, not as measured interaction energies (Jacovi & Goldberg, 2020).

The constraint ablations point to a lesson beyond affinity prediction. Relaxed variants of PACT reach similar accuracy, yet their contributions change much more between retrainings. Accuracy alone therefore does not fix the explanation: the constraints choose among decompositions that fit the data about equally well. Self-explaining models should report reproducibility and intervention agreement alongside accuracy, because a lower loss can come from a less stable decomposition.

Several baselines were developed and tuned on other splits, and we run them with their published settings, so their accuracy here is probably below their best. Table 1 therefore compares architectures under one leakage-reduced protocol, not each method at its best. The explanation comparison is more robust: ranking each baseline by its own attribution method instead of occlusion leaves the conclusions unchanged (Appendix D.4). Path gradients rank slightly better than occlusion for the geometric networks when several residues are removed together, and attention ranks much worse for the two models that provide it, so using occlusion does not disadvantage the baselines.

Two limitations apply to the explanation results. Contact removal changes the model's input at fixed coordinates, so it ignores structural relaxation and does not simulate a real molecular edit. Low cross-seed variability is necessary but not sufficient, since a constant explanation would also have it (Yeh et al., 2019); we therefore interpret it together with the intervention and chemical-agreement results.

7 CONCLUSIONS

We introduced PACT, an affinity model in which every contact-level factor, including the bounded local corrections, rescales a named chemical basis. This makes PACT interpretable without losing accuracy: it matches the strongest baselines we evaluated on LP-PDBbind. Additivity alone does not make its contributions trustworthy, so we measured their reproducibility across retraining and agreement with interventions. Relaxed variants show why: they keep their accuracy but become much less reproducible. Being interpretable by design, rather than explained post hoc, did not exempt PACT's explanations from testing, nor should it exempt other self-explaining models.

AI USE STATEMENT

In this work, we used generative AI tools to improve the clarity and English of the manuscript, assist in drafting and revising portions of the text, identify and analyze existing literature, and implement parts of the code. We did not use generative AI tools for other tasks requiring disclosure; the remaining required disclosure tasks are not applicable to this work. We reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This study uses existing protein-ligand structure and affinity datasets and involves no human participants or personal data. The model is intended to support the faithful evaluation of affinity explanations, rather than deployment or drug generation.

REPRODUCIBILITY STATEMENT

We report the data splits, preprocessing, model configurations, baseline source revisions, and evaluation procedures in the paper and appendix. Experiments were run on an AMD EPYC 7763 CPU with two NVIDIA A4000 GPUs. The full code and instructions for reproducing the results are available at `https://anonymous.4open.science/r/PACT-SUP-CD5B`.

REFERENCES

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018.

Rishabh Agarwal, Levi Melnick, Nicholas Frosst, Xuezhou Zhang, Ben Lengerich, Rich Caruana, and Geoffrey E. Hinton. Neural additive models: Interpretable machine learning with neural nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

David Alvarez-Melis and Tommi S. Jaakkola. Towards robust interpretability with self-explaining neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Pascal Auffinger, Franklin A. Hays, Eric Westhof, and P. Shing Ho. Halogen bonds in biological molecules. *Proceedings of the National Academy of Sciences*, 101(48):16789–16794, 2004. doi: 10.1073/pnas.0407607101. URL `https://doi.org/10.1073/pnas.0407607101`.

Pedro J. Ballester and John B. O. Mitchell. A machine learning approach to predicting protein-ligand binding affinity with applications to molecular docking. *Bioinformatics*, 26(9):1169–1175, 2010. doi: 10.1093/bioinformatics/btq112.

Denis Baronas, Virginija Dudutienė, Vaida Paketurytė, Visvaldas Kairys, Alexey Smirnov, Vaida Juozapaitienė, Aivaras Vaškevičius, Elena Manakova, Saulius Gražulis, Asta Zubrienė, and Daumantas Matulis. Structure and mechanism of secondary sulfonamide binding to carbonic anhydrases. *European Biophysics Journal*, 50(7):993–1011, 2021. doi: 10.1007/s00249-021-01561-1. URL `https://doi.org/10.1007/s00249-021-01561-1`.

Maya Bechler-Speicher, Amir Globerson, and Ran Gilad-Bachrach. The intelligible and effective graph neural additive network. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Roshni Bhatt, David Ryan Koes, and Jacob D. Durrant. CENsible: Interpretable insights into small-molecule binding with context explanation networks. *Journal of Chemical Information and Modeling*, 64(12):4651–4660, 2024. doi: 10.1021/acs.jcim.4c00825.

Cédric Bouysset and Sébastien Fiorucci. ProLIF: a library to encode molecular interactions as fingerprints. *Journal of Cheminformatics*, 13(1):72, 2021. doi: 10.1186/s13321-021-00548-6.

Hans-Joachim Böhm. The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure. *Journal of Computer-Aided Molecular Design*, 8(3):243–256, 1994. doi: 10.1007/BF00126743. URL https://doi.org/10.1007/BF00126743.

Peter J. A. Cock, Tiago Antao, Jeffrey T. Chang, Brad A. Chapman, Cymon J. Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, and Michiel J. L. de Hoon. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, 2009. doi: 10.1093/bioinformatics/btp163.

Guy Durant, Fergus Boyles, Kristian Birchall, Brian Marsden, and Charlotte M. Deane. Robustly interrogating machine learning-based scoring functions: what are they learning? *Bioinformatics*, 41(2):btaf040, 2025. doi: 10.1093/bioinformatics/btaf040.

Jerome Eberhardt, Diogo Santos-Martins, Andreas F. Tillack, and Stefano Forli. AutoDock Vina 1.2.0: New docking methods, expanded force field, and python bindings. *Journal of Chemical Information and Modeling*, 61(8):3891–3898, 2021. doi: 10.1021/acs.jcim.1c00203.

David Errington, Constantin Schneider, Cédric Bouysset, and Frédéric A. Dreyer. Assessing interaction recovery of predicted protein-ligand poses. *Journal of Cheminformatics*, 17(1):76, 2025. doi: 10.1186/s13321-025-01011-6.

Justin P. Gallivan and Dennis A. Dougherty. Cation–π interactions in structural biology. *Proceedings of the National Academy of Sciences*, 96(17):9459–9464, 1999. doi: 10.1073/pnas.96.17.9459. URL https://doi.org/10.1073/pnas.96.17.9459.

Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 3681–3688, 2019. doi: 10.1609/aaai.v33i01.33013681.

David Graber, Peter Stockinger, Fabian Meyer, Siddhartha Mishra, Claus Horn, and Rebecca Buller. Resolving data bias improves generalization in binding affinity prediction. *Nature Machine Intelligence*, 7(10):1713–1725, 2025. doi: 10.1038/s42256-025-01124-5.

Naofumi Hama, Masayoshi Mase, and Art B. Owen. Deletion and insertion tests in regression models. *Journal of Machine Learning Research*, 24(290):1–38, 2023. URL http://jmlr.org/papers/v24/22-0560.html.

Marton Havasi, Sonali Parbhoo, and Finale Doshi-Velez. Addressing leakage in concept bottleneck models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 23386–23397, 2022.

Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of ACL*, pp. 4198–4205, 2020. doi: 10.18653/v1/2020.acl-main.386. URL https://aclanthology.org/2020.acl-main.386/.

Sarthak Jain and Byron C. Wallace. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3543–3556, 2019. doi: 10.18653/v1/N19-1357.

Dejun Jiang, Chang-Yu Hsieh, Zhenxing Wu, Yu Kang, Jike Wang, Ercheng Wang, Ben Liao, Chao Shen, Lei Xu, Jian Wu, Dongsheng Cao, and Tingjun Hou. InteractionGraphNet: A novel and efficient deep graph representation learning framework for accurate protein-ligand interaction predictions. *Journal of Medicinal Chemistry*, 64(24):18209–18232, 2021. doi: 10.1021/acs.jmedchem.1c01830.

Lena Kalinowsky, Julia Weber, Shantheya Balasupramaniam, Knut Baumann, and Ewgenij Proschak. A diverse benchmark based on 3D matched molecular pairs for validating scoring functions. *ACS Omega*, 3(5):5704–5714, 2018. doi: 10.1021/acsomega.7b01194. URL https://doi.org/10.1021/acsomega.7b01194.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International Conference on Machine Learning (ICML)*, pp. 5338–5348, 2020.

Tanja Kortemme, Alexandre V. Morozov, and David Baker. An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein–protein complexes. *Journal of Molecular Biology*, 326(4):1239–1259, 2003. doi: 10.1016/S0022-2836(03)00021-4. URL https://doi.org/10.1016/S0022-2836(03)00021-4.

Satyapriya Krishna, Tessa Han, Alex Gu, Steven Wu, Shahin Jabbari, and Himabindu Lakkaraju. The disagreement problem in explainable machine learning: A practitioner's perspective. *Transactions on Machine Learning Research*, 2024.

Greg Landrum et al. RDKit: Open-source cheminformatics, release 2026.03.3, 2026. URL https://www.rdkit.org.

Jie Li, Xingyi Guan, Oufan Zhang, Kunyang Sun, Yingze Wang, Dorian Bagni, and Teresa Head-Gordon. Leak proof PDBBind: A reorganized data set of protein–ligand complexes for more generalizable binding affinity prediction. *The Journal of Physical Chemistry B*, 130(2):730–740, 2026. doi: 10.1021/acs.jpcb.5c08598.

Hyukjun Lim, Sun Kim, and Sangseon Lee. CheapNet: Cross-attention on hierarchical representations for efficient protein-ligand binding affinity prediction. In *International Conference on Learning Representations (ICLR)*, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.

Anita Mahinpei, Justin Clark, Isaac Lage, Finale Doshi-Velez, and Weiwei Pan. Promises and pitfalls of black-box concept learning models. *arXiv preprint arXiv:2106.13314*, 2021.

Andre Martins and Ramon Astudillo. From softmax to sparsemax: A sparse model of attention and multi-label classification. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1614–1623, 2016. URL https://proceedings.mlr.press/v48/martins16.html.

Georgia B. McGaughey, Marc Gagné, and Anthony K. Rappé. π-stacking interactions: Alive and well in proteins. *Journal of Biological Chemistry*, 273(25):15458–15463, 1998. doi: 10.1074/jbc.273.25.15458. URL https://doi.org/10.1074/jbc.273.25.15458.

Rocco Meli, Andrew Anighoro, Mike J. Bodkin, Garrett M. Morris, and Philip C. Biggin. Learning protein-ligand binding affinity with atomic environment vectors. *Journal of Cheminformatics*, 13 (1):59, 2021. doi: 10.1186/s13321-021-00536-w.

Seokhyun Moon, Sang-Yeon Hwang, Jaechang Lim, and Woo Youn Kim. PIGNet2: a versatile deep learning-based protein–ligand interaction prediction model for binding affinity scoring and virtual screening. *Digital Discovery*, 3(2):287–299, 2024. doi: 10.1039/d3dd00149k.

Sebastian Musiał, Bartosz Zieliński, and Tomasz Danel. Fragment-wise interpretability in graph neural networks via molecule decomposition and contribution analysis. *arXiv preprint arXiv:2508.15015*, 2025.

Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI. *ACM Computing Surveys*, 55(13s):295:1–295:42, 2023. doi: 10.1145/3583558.

Phuc Pham, Viet Thanh Duy Nguyen, Kevin Song, Jake Chen, and Truong-Son Hy. LINKER: Learning interactions between functional groups and residues with chemical knowledge-enhanced reasoning and explainability. *Journal of Chemical Information and Modeling*, 66(14):8579–8593, 2026. doi: 10.1021/acs.jcim.6c00527.

Rodrigo Quiroga and Marcos A. Villarreal. Vinardo: A scoring function based on Autodock Vina improves scoring, docking, and virtual screening. *PLOS ONE*, 11(5):e0155183, 2016. doi: 10.1371/journal.pone.0155183. URL https://doi.org/10.1371/journal.pone.0155183.

Daniel Rose, Thomas Seidel, and Thierry Langer. GRIPHIN: grids of pharmacophore interaction fields for affinity prediction. *Journal of Cheminformatics*, 18(1):97, 2026. doi: 10.1186/s13321-026-01203-8.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019. doi: 10.1038/s42256-019-0048-x.

Diogo Santos-Martins, Stefano Forli, Maria João Ramos, and Arthur J. Olson. AutoDock4Zn: An improved AutoDock force field for small-molecule docking to zinc metalloproteins. *Journal of Chemical Information and Modeling*, 54(8):2371–2379, 2014. doi: 10.1021/ci500209e. URL https://doi.org/10.1021/ci500209e.

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, 2021.

Kristof T. Schütt, Pieter-Jan Kindermans, Huziel E. Sauceda, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller. SchNet: A continuous-filter convolutional neural network for modeling quantum interactions. In *Advances in Neural Information Processing Systems (NIPS)*, volume 30, pp. 991–1001, 2017.

A. Shrake and John A. Rupley. Environment and exposure to solvent of protein atoms. Lysozyme and insulin. *Journal of Molecular Biology*, 79(2):351–371, 1973. doi: 10.1016/0022-2836(73)90011-9. URL https://doi.org/10.1016/0022-2836(73)90011-9.

Jochen Sieg, Florian Flachsenberg, and Matthias Rarey. In need of bias control: Evaluating chemical data for machine learning in structure-based virtual screening. *Journal of Chemical Information and Modeling*, 59(3):947–961, 2019. doi: 10.1021/acs.jcim.8b00712.

Minyi Su, Qifan Yang, Yu Du, Guoqin Feng, Zhihai Liu, Yan Li, and Renxiao Wang. Comparative assessment of scoring functions: The CASF-2016 update. *Journal of Chemical Information and Modeling*, 59(2):895–913, 2019. doi: 10.1021/acs.jcim.8b00545.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 3319–3328, 2017.

Oleg Trott and Arthur J. Olson. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry*, 31(2):455–461, 2010. doi: 10.1002/jcc.21334.

Mikhail Volkov, Joseph-André Turk, Nicolas Drizard, Nicolas Martin, Brice Hoffmann, Yann Gaston-Mathé, and Didier Rognan. On the frustration to predict binding affinities from protein-ligand structures with deep neural networks. *Journal of Medicinal Chemistry*, 65(11):7946–7958, 2022. doi: 10.1021/acs.jmedchem.2c00487.

Renxiao Wang, Luhua Lai, and Shaomeng Wang. Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *Journal of Computer-Aided Molecular Design*, 16(1):11–26, 2002. doi: 10.1023/A:1016357811882. URL https://doi.org/10.1023/A:1016357811882.

Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In *Proceedings of EMNLP-IJCNLP*, pp. 11–20, 2019. doi: 10.18653/v1/D19-1002.

Shoshana J. Wodak and Joël Janin. Analytical approximation to the accessible surface area of proteins. *Proceedings of the National Academy of Sciences*, 77(4):1736–1740, 1980. doi: 10.1073/pnas.77.4.1736. URL https://doi.org/10.1073/pnas.77.4.1736.

Ziduo Yang, Weihe Zhong, Qiujie Lv, Tiejun Dong, and Calvin Yu-Chian Chen. Geometric interaction graph neural network for predicting protein–ligand binding affinities from 3D structures (GIGN). *The Journal of Physical Chemistry Letters*, 14(8):2020–2033, 2023. doi: 10.1021/acs.jpclett.2c03906.

Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Sai Suggala, David I. Inouye, and Pradeep Ravikumar. On the (in)fidelity and sensitivity of explanations. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Mingyue Zheng, Bing Xiong, Cheng Luo, Shanshan Li, Xian Liu, Qianchen Shen, Jing Li, Weiliang Zhu, Xiaomin Luo, and Hualiang Jiang. Knowledge-based scoring functions in drug design: 3. A two-dimensional knowledge-based hydrogen-bonding potential for the prediction of protein-ligand interactions. *Journal of Chemical Information and Modeling*, 51(11):2994–3004, 2011. doi: 10.1021/ci2003939. URL https://doi.org/10.1021/ci2003939.

A MODEL SPECIFICATION AND CHEMICAL MOTIVATION

PACT first computes chemically defined interaction features and then learns how strongly they contribute to affinity. Distance functions describe candidate interactions, the predefined hydrogen-bond basis includes donor direction, and bounded contextual factors adjust interaction strength. We specify these steps below, followed by the training objectives. Interaction features are not themselves energies: their learned, scaled contributions are expressed in pK units.

A.1 MOLECULAR REPRESENTATION

The graph contains ligand heavy atoms, protein heavy atoms, and nearby metal ions. Whole residues within 6 Å of the ligand are retained, while scored ligand–protein contacts are restricted to 5 Å (Appendix A.2). Water molecules are excluded. We use the supplied coordinates and chemical assignments without optimizing the structure or enumerating alternative protonation states.

Each atom has twelve features: heavy-atom degree, hydrogen count, formal charge, ring membership, aromaticity, partial charge, hydrogen-bond donor and acceptor indicators, hydrophobicity, positive and negative ionizability, and halogen identity. Ligand features are obtained with RDKit (Landrum et al., 2026); protein features use residue templates and charge assignments for charged side chains.

Let i denote a ligand atom and j a protein atom or metal. Their center distance and surface separation are

$$d_{ij} = \|\mathbf{p}_i - \mathbf{p}_j\|_2, \qquad s_{ij} = d_{ij} - r_i - r_j, \tag{A.1}$$

where $\mathbf{p}_i$ is the coordinate vector and r_i is the van der Waals radius. Thus, $s_{ij} = 0$ corresponds to touching nominal atomic surfaces, while $s_{ij} < 0$ indicates overlap.

Chemical eligibility is determined before learning. For example, the hydrogen-bond channel requires a donor–acceptor pair, and the hydrophobic channel requires two hydrophobic atoms. Contextual adjustments multiply these bases, so they cannot create a nonzero contribution where the corresponding base is zero.

A.2 GRAPH CONSTRUCTION

Each complex is converted once into a graph, which is cached and reused for training and evaluation. All 7,393 training, 958 validation and 2,171 test complexes were converted without failure.

Input files. The ligand is read from the PDBbind SDF file with RDKit (Landrum et al., 2026) and the protein from the PDB file with Biopython (Cock et al., 2009), using the first model. Ligand hydrogens are kept while the molecule is sanitized and its Gasteiger partial charges are computed, and are then dropped, so every node is a heavy atom or a metal ion. Coordinates are used as supplied and only translated to the ligand centroid; no structure is minimized and no alternative protonation state is generated.

Ligand nodes. Each ligand heavy atom receives the twelve features of Appendix A.1. Degree, hydrogen count, formal charge, ring membership and aromaticity come from RDKit, the partial charge is the Gasteiger charge clipped to $[-2, 2]$, and the donor, acceptor, hydrophobe and ionizable flags come from RDKit's base pharmacophore definitions. The halogen flag marks Cl, Br and I. Bonds are kept with their RDKit bond order, and the six ligand descriptors used by G (Appendix A.6) are computed from the same molecule.

Protein nodes. Every heavy atom of a standard or modified amino acid becomes a protein node. Its features are looked up by residue and atom name in templates built with RDKit from G–X–G tripeptides, which fixes one protonation state per residue type: Asp and Glu carboxylates are deprotonated, Lys and Arg side chains are protonated, and histidine is neutral. Atoms without a template entry, such as those of modified residues, get zero features, except the C-terminal OXT oxygen, which has its own entry. Mg, Ca, Mn, Fe, Co, Ni, Cu and Zn ions become metal nodes with a fixed feature vector of charge $+2$. Water, cofactors and all other hetero groups are discarded.

Burial. Buried areas are computed before the pocket is cut. For each ligand atom, the Shrake–Rupley area in the isolated ligand minus its area in the presence of all protein atoms and metal ions within $10\,\text{Å}$ gives the buried area of Appendix A.3. Atoms flagged as donor or acceptor contribute polar area and all others apolar area. Discarded water and hetero groups do not occlude.

Pocket. The pocket keeps every residue with at least one heavy atom within $6\,\text{Å}$ of a ligand atom, with all of its heavy atoms, so side chains are never cut, and every metal ion within $3.6\,\text{Å}$ of the ligand.

Edges. Three edge sets are stored. Ligand edges join bonded atoms, and also any other ligand pairs closer than $2.5\,\text{Å}$. Protein edges join pocket atoms closer than $4.5\,\text{Å}$. PACT uses these two sets only to find covalent neighbors for the donor axes of Appendix A.4: ligand bonds, and protein edges shorter than $1.95\,\text{Å}$. Contact edges join each ligand atom to every pocket atom and metal ion within $5\,\text{Å}$; the twelve channels are evaluated on these contacts.

Groups and residue types. Contacts are grouped by ligand atom and residue into the ligand-atom–residue groups of Appendix A.6. Each metal ion is its own residue, so its contacts form separate groups. Each node also stores a residue type for the contact multiplier: one of the twenty standard amino acids, one class shared by modified residues and metal ions, or a ligand class. These are the 22 rows of the residue-type table in Equation A.6.

A.3 DISTANCE FUNCTIONS AND INTERACTION BASES

Most channels use a Gaussian response or a decreasing ramp:

$$\phi(v;\mu,w)=\exp\left[-\left(\frac{v-\mu}{w}\right)^2\right],\qquad R(v;a,b)=\operatorname{clip}\left(\frac{b-v}{b-a},0,1\right). \tag{A.2}$$

The Gaussian emphasizes a distance range, while the ramp gives full response at short distances and decreases linearly to zero. The parameters in Table S1 are fixed model settings; they are not measured interaction energies.

The attraction, repulsion, and hydrophobic functions follow the general design of empirical scoring functions such as AutoDock Vina (Trott & Olson, 2010). Gaussian attraction and quadratic overlap penalties offer inexpensive, nonsingular contact features. Our attraction center differs from Vina's original parameterization, and we do not include its second, broader attractive component. The analysis underlying Vinardo showed that numerous distant atom pairs can contribute substantially to such a broad term and motivated a revised short-range scoring function (Quiroga & Villarreal, 2016). These studies motivate the functional forms rather than establish our particular parameter values as optimal.

Every channel uses the same cutoff envelope:

$$E(d)=1-3t^2+2t^3,\qquad t=\operatorname{clip}\left(\frac{d-4\,\text{Å}}{1\,\text{Å}},0,1\right). \tag{A.3}$$

The envelope is one up to $4\,\text{Å}$ and reaches zero at $5\,\text{Å}$. For contact $e=(i,j)$ and channel c, the base is $b_{ec}=E(d_{ij})f_c(i,j)$, where f_c is the eligible response in Table S1. Ineligible responses are zero.

Figure S1 plots representative responses. The ionic channel is a short-range feature between oppositely ionizable atoms, rather than a Coulomb potential. Aromatic and cation–aromatic bases describe eligible atom proximity, following the structural roles established for π-stacking (McGaughey et al., 1998) and cation–π interactions (Gallivan & Dougherty, 1999), with orientation refined separately. The halogen response favors slightly negative surface separations, consistent with observations of short halogen–acceptor contacts (Auffinger et al., 2004). Since the active model does not apply a halogen-angle modifier, this channel is described as proximity rather than a complete halogen-bond potential.

Burial supplements pair distances with information about solvent exposure. For each ligand atom, we estimate solvent-accessible surface area in the isolated ligand and in the complex using the

Table S1: Interaction responses before the common cutoff and learned modifiers. Distance parameters are in Å. h_{ij} combines donor–acceptor eligibility with donor orientation. S_i^{ap} and S_i^{pol} are buried ligand-atom areas, and o_{ij} is the pairwise occlusion fraction. Metal contacts activate only the metal and, where applicable, zinc channels.

Channel	Eligible contacts	Response $f_c(i,j)$
van der Waals attraction	Ligand–protein atom pairs	$\phi(s_{ij}; 0.5, 0.5)$
Steric repulsion	Ligand–protein atom pairs	$-\max(-s_{ij}, 0)^2$
Hydrophobic	Two hydrophobic atoms	$R(s_{ij}; 0.5, 1.5)$
Hydrogen bond	Donor–acceptor pairs	$h_{ij}\phi(d_{ij}; 2.9, 0.55)$
Ionic	Oppositely ionizable atoms	$R(d_{ij}; 3.0, 5.0)$
Aromatic proximity	Two aromatic atoms	$\phi(d_{ij}; 4.2, 1.0)$
Halogen–acceptor proximity	Ligand Cl, Br, or I with a protein acceptor	$\phi(s_{ij}; -0.2, 0.5)$
Metal coordination	Eligible ligand atom with a metal	$\phi(d_{ij}; 2.4, 0.8)$
Cation–aromatic proximity	Cationic and aromatic atoms, in either orientation	$\phi(d_{ij}; 4.3, 1.1)$
Apolar burial	Ligand–protein atom pairs	$\frac{S_i^{\mathrm{ap}}}{20\,\text{Å}^2} o_{ij}$
Polar burial	Ligand–protein atom pairs	$\frac{S_i^{\mathrm{pol}}}{20\,\text{Å}^2} o_{ij}$
Zinc coordination	Eligible ligand atom with zinc	$\phi(d_{ij}; 2.4, 0.8)$

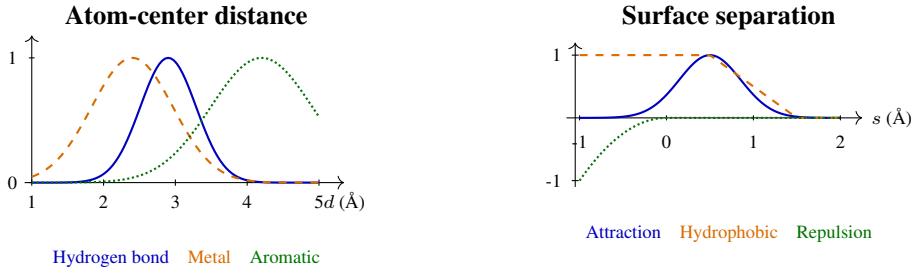


Figure S1: Representative interaction responses before eligibility, orientation, the common cutoff, and learned scaling. Curve heights are feature values rather than affinity contributions. Repulsion is negative because larger pK values indicate stronger binding.

Shrake–Rupley method (Shrake & Rupley, 1973), with a 1.4 Å probe, 128 surface points per atom, and a 10 Å protein shell:

$$S_i = \max\left(S_i^{\mathrm{free}} - S_i^{\mathrm{complex}}, 0\right), \qquad o_{ij} = \mathrm{clip}\left(\frac{h_{ij}^{\mathrm{cap}}}{2R_i}, 0, 1\right), \qquad h_{ij}^{\mathrm{cap}} = R_i - \frac{d_{ij}^2 + R_i^2 - R_j^2}{2d_{ij}}. \tag{A.4}$$

Here $R_i = r_i + 1.4\,\text{Å}$ is the probe-expanded radius. The buried area is assigned to polar or apolar features according to atom typing. The cap fraction associates it with individual contacts, following the motivation of pairwise accessible-area approximations (Wodak & Janin, 1980). Overlapping caps are not partitioned into disjoint areas, so the contact values are descriptors rather than an exact surface-area decomposition. Both burial channels have nonnegative bases; polar burial is not an unsatisfied-polar or desolvation penalty.

A.4 Hydrogen-bond direction and residue context

Distance alone cannot distinguish favorable and unfavorable hydrogen-bond arrangements. Orientation-dependent hydrogen-bond potentials have demonstrated the value of angular information for recognizing compatible structures (Kortemme et al., 2003; Zheng et al., 2011). We therefore include a fixed donor-direction factor in the hydrogen-bond basis.

For each atom, a local axis points opposite to the sum of unit vectors towards its covalent heavy-atom neighbors. Let $\mathbf{u}_{ij}$ point from ligand atom i to protein atom j, and let $\mathbf{v}_i$ be the reconstructed axis. Its directional response is

$$H_i(\mathbf{u}) = \mathrm{clip}\left(\frac{\mathbf{v}_i^\top \mathbf{u} - \cos 60^\circ}{\cos 60^\circ}, 0, 1\right), \qquad h_{ij} = \min(D_i A_j H_i(\mathbf{u}_{ij}) + A_i D_j H_j(-\mathbf{u}_{ij}), 1)\,, \tag{A.5}$$

where D_i and A_i indicate donor and acceptor eligibility. The response is one for alignment and zero at angles of 60° or greater. If an axis cannot be reconstructed, H is set to one and only distance is considered. The factor is only an approximation: it uses a heavy-atom axis in place of the donor–hydrogen direction, not an explicit donor–hydrogen–acceptor angle.

Let $r(e)$ identify the residue type of contact e. The contact-level context multiplier in PACT is

$$m_{ec} = 1 + 0.25 \tanh\big(t_{r(e),c}\big)\,, \tag{A.6}$$

where t is a learned 22×12 residue-type table. Consequently, $m_{ec} \in [0.75, 1.25]$. It changes the strength of an eligible interaction before pooling, but cannot activate an absent chemical basis or reverse its sign.

A.5 METAL COORDINATION AND ZINC

Metal eligibility includes ligand acceptors and nonpositively charged nitrogen, oxygen, or sulfur atoms. The latter condition covers potential coordinators missed by ordinary hydrogen-bond acceptor typing, although it can also admit unsuitable candidates. The broad response centered at $2.4\,\text{Å}$ therefore represents coordination proximity rather than a preferred bond length shared by all metals.

Specialized metal scoring illustrates the limitations of distance-only treatment. AutoDock4Zn improved docking and affinity estimation by explicitly modeling zinc coordination geometry (Santos-Martins et al., 2014). Binding can also involve protonation changes absent from a static contact graph. For carbonic-anhydrase inhibitors, experimental work identifies coupled protonation processes involving sulfonamide ligands and zinc-bound solvent (Baronas et al., 2021). A short zinc–ligand distance alone does not determine binding free energy.

The zinc base is the generic metal base restricted to zinc:

$$b_{e,\mathrm{Zn}} = b_{e,\mathrm{metal}}\, \mathbf{1}[Z_j = 30]. \tag{A.7}$$

Both channels can contribute to the same contact. The additional zinc term is nonnegative, providing an explicit zinc-specific adjustment while retaining shared metal information. These are overlapping model components, not two independent physical interactions.

Development analyses also identified a substantial shift in zinc-binding chemotypes from other metals. Its contribution should not be interpreted as a uniquely recovered zinc bond energy.

A.6 GROUPING, CONTEXTUAL COEFFICIENTS, AND PREDICTION

A group g contains contacts between one ligand atom and one protein residue. We retain the strongest response within each channel, using the most negative response for steric repulsion. We distinguish the raw pooled bases z_{gc} from the bases B_{gc} after contact-level adjustment:

$$z_{gc} = \mathrm{pool}_{e \in g}(b_{ec}), \qquad B_{gc} = \mathrm{pool}_{e \in g}(m_{ec} b_{ec}), \tag{A.8}$$

where pool is maximum pooling except for repulsion, which uses minimum pooling. This reduces repeated counting within a residue without discarding ligand-atom identity.

The local chemistry vector $\mathbf{x}_g$ concatenates the ligand atom's features with a weighted average of the group's protein-atom features, weighted by the contact proximity weight

$$c_{ij} = E(d_{ij}) \exp\left(-\frac{d_{ij} - 4\,\text{Å}}{1.5\,\text{Å}}\right), \qquad C_g = \sum_{e \in g} c_{ij}. \tag{A.9}$$

The weight is one at the envelope plateau and decays with a $1.5\,\text{Å}$ length scale; a weighted average is unchanged by the common factor, so only the decay length affects $\mathbf{x}_g$. A linear map gives the

local chemistry multiplier, and a small neural network computes the group update from the group's raw interaction profile:

$$\ell_{gc} = 1 + \tanh(\mathbf{x}_g^\top \mathbf{u}_c), \qquad q_{gc} = 0.25 \tanh\big(\big[\mathrm{MLP}(\mathbf{z}_g)\big]_c\big). \tag{A.10}$$

Thus, ℓ_{gc} lies between zero and two, and the neural factor $1 + q_{gc}$ lies between 0.75 and 1.25.

A single MLP is shared by all twelve channels: a linear map takes the group's twelve raw pooled bases and a constant to a 24-dimensional state, one residual block of two two-layer SiLU networks follows, and a linear readout returns all twelve corrections. Its weight and bias start at zero, so training begins at the fixed-coefficient scoring function. It sees only the channels of its own group, so no information passes between groups. It holds $4{,}188$ of the $4{,}784$ parameters (Table S22).

Van der Waals contributions also compete among the groups of one ligand atom:

$$u_g = \log C_g + \mathbf{w}_z^\top \mathbf{z}_g + \mathbf{w}_x^\top \mathbf{x}_i + b_u, \tag{A.11}$$

with C_g the proximity mass of Equation A.9 and $\mathbf{x}_i$ the ligand atom's features. Sparsemax (Martins & Astudillo, 2016) on $\frac{1}{2}\mathbf{u}$, against a null entry fixed at $u_0 = 0.5$, returns weights $\alpha_{g,\mathrm{vdW}}$ that sum to one and are exactly zero on losing groups; the null may take some, leaving it unassigned. Other channels use $\alpha_{gc} = 1$. The neural update is local, but this allocation couples competing groups.

Targets are standardized as $y' = (y - \mu_y)/\sigma_y$ on the training set, and

$$a_{gc} = w_c B_{gc} \ell_{gc} \alpha_{gc} (1 + q_{gc}), \qquad w_c = \mathrm{softplus}(v_c), \qquad f(x) = b_0 + g(x) + \sum_{g,c} a_{gc}. \tag{A.12}$$

pK values follow as $\widehat{y} = \mu_y + \sigma_y f$, $A_{gc} = \sigma_y a_{gc}$, $b = \mu_y + \sigma_y b_0$ and $G = \sigma_y g$, the last two being the intercept and ligand correction of Equation 1. Softplus weights and nonnegative multipliers preserve each base's sign.

The ligand correction is computed from six scaled descriptors $\mathbf{v}$ — rotatable bonds$/10$, $\log(1 +$ heavy atoms$)/4$, heteroatom fraction, aromatic fraction, net charge$/3$, $|$net charge$|/3$ —

$$g(x) = \beta \tanh\big((\mathbf{w}_G^\top \mathbf{v} - c_0)/\beta\big), \qquad \beta = 0.3, \tag{A.13}$$

with c_0 fixed by the initialization fit. It is bounded by 0.3 standardized units and reported separately, which does not remove size dependence from the interaction contributions.

Residue and fragment explanations sum the matching A_{gc}. Atom-pair pictures split a group's contribution by absolute contextualized bases and proximity weights: the total is preserved, the parts are not independent removal effects. The ligand-atom–residue group is the primary unit of the decomposition.

A.7 Training objectives

The objective combines affinity regression, pairwise ranking and two attribution regularizers. All of them use standardized affinity units, so the training formulas stay separate from the pK values reported in explanations.

For a minibatch $\mathcal{B}$, let ω_n be inversely proportional to the number of training complexes that share example n's protein sequence. The affinity loss is a weighted mean squared error,

$$L_{\mathrm{aff}} = \frac{\sum_{n \in \mathcal{B}} \omega_n \big(f(x_n) - y'_n\big)^2}{\sum_{n \in \mathcal{B}} \omega_n}. \tag{A.14}$$

Computed from the training split and normalized inside each minibatch, they reduce the influence of sequences that appear often.

The ranking term pushes predictions to follow the observed order. Averaging over the ordered pairs (n, m) of the minibatch, those with $y'_n > y'_m$,

$$L_{\mathrm{rank}} = \operatorname*{mean}_{y'_n > y'_m} \log\big[1 + e^{-(f(x_n) - f(x_m))/T}\big], \qquad T = 0.25. \tag{A.15}$$

It is zero when no ordered pair exists, and uses every ordered pair, including pairs from different proteins, so it is not a matched-molecular-pair objective.

Table S2: Principal implementation and optimization settings. Validation sequence-balanced RMSE is the square root of the mean per-sequence mean squared error.

Setting	Value
Optimizer	AdamW (Loshchilov & Hutter, 2019)
Batch size	32
Initial learning rate	10^{-3}
Weight decay	5×10^{-4}; channel-weight logits excluded
Gradient-norm clipping	5
Maximum epochs	400
Minimum training epochs	25
Early-stopping patience	20 epochs
Learning-rate reduction	Factor 0.5 after 5 plateau epochs
Checkpoint selection	Validation sequence-balanced RMSE
Ranking temperature	0.25 standardized units
Transport perturbation SD	0.05 Å per coordinate
Active-base threshold	10^{-9}

The two attribution terms regularize the explanation during learning (Alvarez-Melis & Jaakkola, 2018). For a group–channel entry, the effective coefficient relative to the raw pooled base is

$$\theta_{gc}(x) = a_{gc}(x)/z_{gc}(x), \tag{A.16}$$

which absorbs the contact adjustment, the channel weight, the local chemistry factor, the allocation and the group update. An entry is *active* when $|z_{gc}| > 10^{-9}$; the rest are left out of both attribution losses.

The local term makes this coefficient agree with the prediction's sensitivity to the same pooled feature, averaged over the entries active across the minibatch:

$$L_{\text{local}} = \underset{(g,c)\ \text{active}}{\text{mean}} \left(\frac{\partial F(\mathbf{z})}{\partial z_{gc}} - \theta_{gc} \right)^2. \tag{A.17}$$

$F(\mathbf{z})$ reconstructs the standardized interaction readout at the current example, holding the contact-adjustment ratios B_{gc}/z_{gc}, the local chemistry factors and the distance priors fixed while the group update and the allocation are recomputed from $\mathbf{z}$. This is the derivative the implementation takes: it constrains contextual repricing inside the group readout, not the model's full coordinate sensitivity. Gradients flow through both the sensitivity and the effective coefficients.

The transport term compares nearby geometries. We build x' by perturbing every atomic coordinate with independent Gaussian noise of standard deviation 0.05 Å, keeping the cached contact connectivity and buried areas and recomputing the distance-dependent responses, hydrogen-bond direction included. The original coefficients then meet the perturbed bases:

$$L_{\text{transport}} = \underset{(g,c)\ \text{active in both}}{\text{mean}} \left[a_{gc}(x') - z_{gc}(x') \,\text{stopgrad}\big(\theta_{gc}(x)\big) \right]^2. \tag{A.18}$$

The average runs over entries active in both geometries and is zero if there are none. stopgrad makes the original coefficient a fixed target. The reference therefore lets a contribution change when its basis changes, and penalizes only the extra change from moving the effective coefficient, weighted by the perturbed basis. It does not ask explanations to stay constant under perturbation, and is not the rigid ligand translation used in evaluation.

The full objective is

$$L = L_{\text{aff}} + 0.2\, L_{\text{rank}} + 800\, L_{\text{local}} + 10{,}000\, L_{\text{transport}}. \tag{A.19}$$

The large multipliers offset the small numerical scale of the raw attribution losses, so they should not be read as importance. Neither attribution loss supervises experimental interaction energies or residue-removal effects. Table S2 lists the remaining implementation and optimization settings.

The architecture guarantees an explicit decomposition of the prediction, while the regularizers encourage consistency of its contextual coefficients. Reproducibility, chemical agreement, and intervention behavior remain empirical properties evaluated separately in the following sections.

Table S3: Baseline evaluation protocol: experimental role, explanation interface, and removal operator used in the intervention experiments. Checkpoints are selected on the corresponding validation split.

Model	Experimental role	Explanation	Removal operator
RF-Score	Statistical scoring control	Residue occlusion	Remove residue contact counts
SchNet	Continuous-filter 3D model	Residue occlusion	Remove protein atoms
EGNN	Equivariant geometric model	Residue occlusion	Remove ligand–protein edges
IGN	Interaction-graph model	Residue occlusion	Remove ligand–protein edges
GIGN	Geometric interaction graph	Residue occlusion	Remove ligand–protein edges
PIGNet2	Physics-informed scoring model	Native contributions	Remove residue contacts
CheapNet	Hierarchical interaction model	Residue occlusion	Remove intermolecular edges
CENsible	Context-weighted named terms	Residue occlusion	Remove residue atoms, rebuild terms and grid
PACT	Proposed structured predictor	Native contributions	Remove ligand–residue contact edges

B PREDICTIVE EVALUATION AND MOLECULAR CHANGES

This section evaluates whether the restrictions used to obtain an explicit interaction decomposition preserve useful affinity prediction. We first report performance on LP-PDBbind and external datasets, and then evaluate affinity differences between related molecules. The latter is a separate question: an inspectable decomposition of an absolute prediction does not imply accurate prediction of a chemical modification.

B.1 EVALUATION COHORTS AND BASELINE PROTOCOL

LP-PDBbind (Li et al., 2026) assigns 11,513 training, 2,422 validation, and 4,860 test complexes. We use its noncovalent complexes at clean level CL1 for training and CL2 for validation and test, which leaves 7,393 training, 958 validation, and 2,171 test complexes. CL1 removes complexes with clashing atoms (protein and ligand heavy atoms closer than 1.75 Å), ligands with low drug-likeness (QED below 0.2), and rare elements. CL2 also keeps only K_{d} or K_{i} values given as a single number (for example $K_{\mathrm{d}} = 5$ nM, not $K_{\mathrm{d}} < 5$ nM). The training set therefore also contains 2,921 IC_{50} values and 241 values given only as a bound or an estimate, which we use as recorded. The main predictive table uses the intersection of complexes for which all models in that comparison produced verified predictions ($n = 2{,}133$). Results from different subsets are not used for paired claims.

BDB2020+ and Mpro are external evaluations without the identifier and protein-sequence overlap identified for the other reference sets. EGFR is retained as a supplementary target-specific evaluation with known overlap. The independent CASF-2016 (Su et al., 2019) result uses a separate CleanSplit training protocol (Graber et al., 2025) and 144 complexes. It is not pooled with evaluations of LP-PDBbind-trained models on the complete CASF collection.

Table S3 summarizes the experimental role and available explanation interface of the baselines used in these comparisons. Exact upstream revisions, and any deviation from the published configuration of each baseline, are recorded in the released artifact manifest.

The baseline reproduction specification in Appendix E identifies the source revisions, preprocessing, hyperparameters, selection rules, and execution commands required to reconstruct each reported experiment.

B.2 EXTERNAL TRANSFER

Table S4 reports external Pearson correlation and concordance index. BDB2020+ and Mpro provide the overlap-free transfer evaluations. EGFR is reported separately because overlap prevents treating it as an independent generalization test. The two metrics agree on the ordering: PACT is second to last on BDB2020+ and second on Mpro under both.

Table S4: External transfer, mean ± sample SD across training seeds. BDB2020+ and Mpro are free of the identified sequence overlap; EGFR is a supplementary overlapping target-specific evaluation, and at $n = 23$ its concordance indices are not separated by their own seed spread.

	BDB2020+ ($n = 115$)		Mpro ($n = 40$)		EGFR ($n = 23$)	
Model	$r \uparrow$	CI ↑	$r \uparrow$	CI ↑	$r \uparrow$	CI ↑
RF-Score	0.461 ± 0.005	0.631 ± 0.005	**0.613** ± 0.010	**0.770** ± 0.010	$\underline{0.475} \pm 0.031$	$\underline{0.657} \pm 0.022$
SchNet	0.338 ± 0.057	0.593 ± 0.022	0.264 ± 0.025	0.591 ± 0.010	0.420 ± 0.084	0.622 ± 0.057
EGNN	$\underline{0.465} \pm 0.041$	$\underline{0.640} \pm 0.016$	0.459 ± 0.068	0.671 ± 0.034	**0.574** ± 0.100	**0.697** ± 0.054
GIGN	**0.479** ± 0.038	**0.648** ± 0.018	0.386 ± 0.115	0.636 ± 0.044	0.232 ± 0.279	0.587 ± 0.076
IGN	0.462 ± 0.025	0.632 ± 0.022	0.339 ± 0.135	0.641 ± 0.064	0.273 ± 0.216	0.577 ± 0.070
PACT	0.379 ± 0.004	0.594 ± 0.005	$\underline{0.582} \pm 0.014$	$\underline{0.705} \pm 0.007$	0.421 ± 0.002	0.636 ± 0.005

Table S5: External 3D-MMP evaluation: directional accuracy of PACT, the fraction of pairs with a nonzero observed difference whose sign PACT predicts correctly. N/N_{dir} gives the number of pairs and of pairs with a nonzero observed difference. Subsets overlap.

Subset	N/N_{dir}	Targets	Directional accuracy
All pairs	99/93	33	0.688
No training member	57/53	22	0.736
Equal heavy atoms	23/22	20	0.682
Gap ≥ 1 pK	39/39	30	0.718

PACT transfers substantial ranking signal to Mpro, reaching $r = 0.582$, but does not dominate the external datasets. On the 144-complex independent CASF-2016 subset, CheapNet leads at $r = 0.831 \pm 0.003$, followed by PACT at 0.795 ± 0.003, RF-Score at 0.794 ± 0.005, EGNN at 0.771 ± 0.054, CENsible at 0.770 ± 0.005, GIGN at 0.759 ± 0.014, SchNet at 0.756 ± 0.040, and IGN at 0.702 ± 0.016. Ligand size alone reaches $r = 0.632$, so PACT separates from the size control by 0.163 and from RF-Score by 0.001. The column is therefore a conventional reference point rather than evidence that the decomposition generalizes better. Fold SD describes training-fold variation rather than uncertainty over the 144 test complexes; PACT and the size control cover all five CleanSplit folds and the retrained baselines folds 0–2.

B.3 EXTERNALLY CURATED MATCHED MOLECULAR PAIRS

The 3D-MMP benchmark of Kalinowsky et al. (2018) contains 99 molecular pairs, 178 crystal structures, and 33 target clusters. We use the supplied structures and affinities and orient each difference as the right-hand compound minus the left-hand compound. Predictions are averaged over six seeds before computing differences.

Training membership is determined from the graph cache actually used for training. Of the 99 pairs, 42 contain at least one training structure; the "no training member" subset contains the remaining 57 pairs.

As shown in Table S5, PACT predicts the direction of 0.736 of the nonzero differences among pairs without a training member, and 0.688 over all pairs.

For the historical subset in which both compounds bind below $1\ \mu$M, PACT predicts $19/25$ directions correctly at a minimum observed gap of 0.5 pK and $10/13$ at a gap of 1.0 pK. Table S6 places these rates beside the scoring functions evaluated on the same two cohorts by Kalinowsky et al. (2018), whose published sizes of 25 and 13 pairs are reproduced exactly here. PACT matches the highest published point estimate for structures scored before geometry optimization on both subsets. Both datasets are small enough that one pair moves the first column by four points and the second by almost eight, and this is a historical comparison rather than a common-training one: the published functions were not retrained on our split.

Table S6: Directional accuracy (%) on the historical 3D-MMP subsets in which both compounds bind below 1 μM, ordered by the first column. Published values are the better of the two columns reported for each scoring function by Kalinowsky et al. (2018), which report no error metric for these cohorts. Nothing is highlighted: the published functions were not retrained on the split used here, and the cohorts are too small to separate the leading entries.

Method	Gap ≥ 0.5 pK ($n = 25$)	Gap ≥ 1.0 pK ($n = 13$)
Affinity dG	76.0	76.9
ASE	76.0	76.9
Consensus	76.0	69.2
Alpha HB	68.0	76.9
GBVI/WSA dG	68.0	76.9
London dG	68.0	69.2
ChemScore	68.0	61.5
ChemPLP	60.0	61.5
ASP	56.0	53.9
GOLDScore	52.0	61.5
PACT	76.0	76.9

C EXPLANATION REPRODUCIBILITY AND CONTEXTUAL BEHAVIOR

This section asks whether independently trained models reproduce the same explanations, whether apparent top-k agreement survives inspection of the selection boundary, and whether the learned contextual coefficients materially change the decomposition.

All cross-seed statistics intersect complexes and residues across checkpoints before comparison. Residues are ranked by absolute contribution, while numerical variability retains the signed contribution in pK.

C.1 REPRODUCIBILITY ACROSS RETRAINING

Across the six PACT checkpoints, mean absolute-contribution rank correlation is high and the exact top-five set is preserved in 83.4% of seed-pair comparisons. Mean signed residue-contribution SD is 0.00620 pK. Exact agreement is stricter than overlap and full-vector rank correlation; all three should therefore be reported together.

Only 9.45×10^{-5} of residue contributions are exactly zero, so the agreement is not caused by a predominantly zero attribution vector. Contribution magnitude and cross-seed SD correlate at $r = 0.681$, although relative variability decreases with magnitude.

The retraining columns of Table 2 use the full test subsets. Baseline rows use one seed unless a spread is reported. PACT and PIGNet2 compare native contributions; the other populated rows compare occlusion explanations, and GRIPHIN and CENsible have no native residue-level decomposition to compare across seeds.

C.2 SELECTION-BOUNDARY SENSITIVITY

The gap between the kth and $(k + 1)$th mean absolute contribution is compared with the root-sum-square of their signed cross-seed standard deviations. This quantity is a descriptive noise reference and not a confidence interval because it does not model covariance between the two residues.

Table S7 shows that high aggregate agreement can coexist with an uncertain selection boundary. At $k = 5$, the fifth–sixth contribution gap is below the noise reference in 35.1% of eligible complexes. The top-five set is therefore useful as a reproducibility summary but should not always be interpreted as a sharp distinction between residue ranks.

Table S7: Selection-boundary sensitivity. The final column reports the fraction of complexes in which the contribution gap at the selection boundary is smaller than the root-sum-square SD reference.

k	Complexes	Median gap (pK)	Gap below noise reference
1	2,171	0.1140	12.1%
2	2,171	0.0543	21.7%
3	2,170	0.0346	24.8%
5	2,169	0.0184	35.1%
8	2,158	0.0109	41.1%
10	2,112	0.0090	40.1%

Table S8: Complete ablations on all 2,171 test complexes, six seeds. On the 2,133-complex cohort of Table 1, PACT reaches 1.5043 ± 0.0053 pK.

Configuration	RMSE ↓	r ↑	Exact@5 ↑	Δ Exact@5	res.
Global weights + allocation	1.5793 ± 0.0041	0.5383	0.9138	—	—
+ local coefficient	1.5131 ± 0.0050	0.5798	0.8322	—	—
PACT	1.5084 ± 0.0052	0.5828	0.8343	—	—
Signed channel weights	1.4925 ± 0.0050	0.5860	0.6365	-0.1978	yes
ℓ may reverse sign	1.5159 ± 0.0064	0.5745	0.7919	-0.0424	yes
+ free-form ligand term	1.5267 ± 0.0121	0.5627	0.6660	-0.1683	yes
+ two free-form terms	1.5189 ± 0.0068	0.5672	0.6028	-0.2315	yes
− ligand correction G	1.5276 ± 0.0065	0.5671	0.9161	$+0.0818$	yes
− allocation	1.5268 ± 0.0062	0.5657	0.8372	$+0.0029$	no
− ranking term	1.5182 ± 0.0048	0.5753	0.8553	$+0.0210$	yes
− both attribution terms	1.5001 ± 0.0067	0.5895	0.8345	$+0.0002$	no

C.3 COEFFICIENT LADDER AND CONSTRAINT ABLATIONS

All configurations share the chemical bases, optimisation protocol and six-seed evaluation. Unless stated, each differs from PACT by exactly one setting. The ladder cells are:

- **Global weights + allocation**: $m_{ec} = \ell_{gc} = 1$ and $q_{gc} = 0$, leaving twelve learned global channel coefficients. Input-dependent allocation and the ligand correction are retained.
- **+ local coefficient**: restores the per-group local chemistry coefficient ℓ_{gc} while retaining no contact-level context.
- **PACT**: adds the residue-conditioned contact multiplier and the local within-group update.

Exact@5 intervals. Exact@5 averages a per-complex indicator over seed pairs, and the pairs within one complex are correlated because they share seeds and describe the same pocket. We therefore report a cluster bootstrap over complexes (2,000 resamples), pairing each configuration against PACT on the same resample. The interval treats the trained seeds as fixed and so does not express seed sampling; differences below roughly 0.03 should be read as provisional. All configurations are compared at equal seed count, since Exact@5 rises mechanically when fewer pairs are available.

Nonnegative channel coefficients. The softplus on w_c fixes each channel's sign by its basis function. Removing it gives the lowest RMSE of any configuration we trained, 1.4925 pK, and the largest loss of reproducibility among bounded cells, Exact@5 0.6365 (-0.1978), with cross-seed contribution SD rising from 0.0062 to 0.0133 pK.

Free-form neural terms. Every learned factor in PACT multiplies a named basis, so the reported contributions reconstruct the prediction exactly. We add a neural term on the ligand descriptors whose output is added to the prediction and belongs to no channel. Its output layer is zero-initialised, so the configuration begins identical to PACT and the comparison isolates the addition. Exact@5

Table S9: Use of the contextual coefficient ranges. Values are averaged across six seeds on 600 sequence-distinct complexes. "Live" is the fraction of evaluated entries with a nonzero basis for the channel. The 90% span is $p_{95} - p_5$ divided by the permitted range.

	Contact multiplier $[0.75, 1.25]$				Local coefficient $[0, 2]$			
Channel	Live	Mean	SD	90% span	Live	Mean	SD	90% span
Apolar burial	.687	1.034	.123	.667	.698	1.345	.285	.457
Aromatic	.060	1.010	.075	.438	.052	.224	.300	.429
Cation–aromatic	.022	.971	.089	.587	.024	.222	.293	.475
Halogen	.002	1.020	.094	.625	.004	1.361	.218	.353
Hydrogen bond	.025	.995	.120	.820	.057	.335	.411	.578
Hydrophobic	.132	.990	.098	.717	.164	.286	.369	.519
Ionic	.012	.968	.100	.609	.012	.804	.569	.828
Metal	.002	1.024	.000	.000	.006	.752	.250	.413
Polar burial	.279	.997	.116	.605	.253	.684	.352	.553
Repulsion	.030	1.024	.097	.509	.066	1.176	.561	.867
van der Waals	.986	1.012	.139	.773	.982	.809	.465	.746
Zinc	.002	1.036	.000	.000	.004	.822	.163	.199

falls to 0.6660 (-0.1683), and with a second free-form term to 0.6028 (-0.2315) against an additive expectation of -0.1778, so the two compound rather than add. Neither improves RMSE. About 10% of the prediction's standard deviation moves outside the named decomposition; the term's mean is larger, but that part is a constant the intercept would otherwise carry, and only its varying part is unattributable.

Permitted widths. Both bounds were fixed during development on the validation split, the test-set comparisons below are reported for completeness. For the within-group update, single-contact infidelity is lowest at the ± 0.25 we use and higher on both sides: it rises by 0.0017 at ± 0.10 and by 0.0171 at ± 4. The large degradation appears where the factor may become negative. At ± 1 the factor spans $[0, 2]$ and may extinguish a contact without inverting it, and the change is not resolved; at ± 4 it may invert, and fidelity degrades sharply. For the local coefficient, tightening to $[0.5, 1.5]$ costs 0.016 pK of RMSE, while widening to $[-1, 3]$ costs 0.042 of Exact@5 without a resolved accuracy gain, so our $[0, 2]$ lies between the two failure modes.

Attribution objectives. The attribution terms target agreement between reported contributions and the model's response to removing a contact, so cross-seed agreement does not test them. Measured on single-contact infidelity over six seeds, with interventions held fixed across configurations and differences paired by seed, removing both terms degrades infidelity by 0.0085 and removing the transport term alone by 0.0021. The two are super-additive: removing both costs more than three times the sum of removing each alone, so they are complementary and retained together. Removing both, however, improves RMSE by 0.008 pK.

Component removals. The ligand correction, the allocation and the ranking term each improve predictive accuracy significantly, by 0.019, 0.018 and 0.010 pK respectively. Removing the ligand correction raises Exact@5 to 0.9161, so our configuration trades some reproducibility for accuracy.

C.4 Use of the permitted coefficient ranges

Coefficient usage is measured on 600 test complexes, one per exact protein sequence, across six training seeds. Statistics for a channel include only entries where its chemical basis is nonzero. The contact multiplier has permitted range $[0.75, 1.25]$, while the per-group local coefficient has range $[0, 2]$.

Table S9 shows that the permitted ranges are used substantially for common channels. For example, the middle 90% of the contact multiplier spans 77.3% of its permitted range for van der Waals and 82.0% for hydrogen bonds. The local coefficient spans 74.6% of its range for van der Waals and 82.8% for ionic interactions. The metal and zinc contact multipliers are effectively constant in this test, their contextual variation instead enters through the local coefficient.

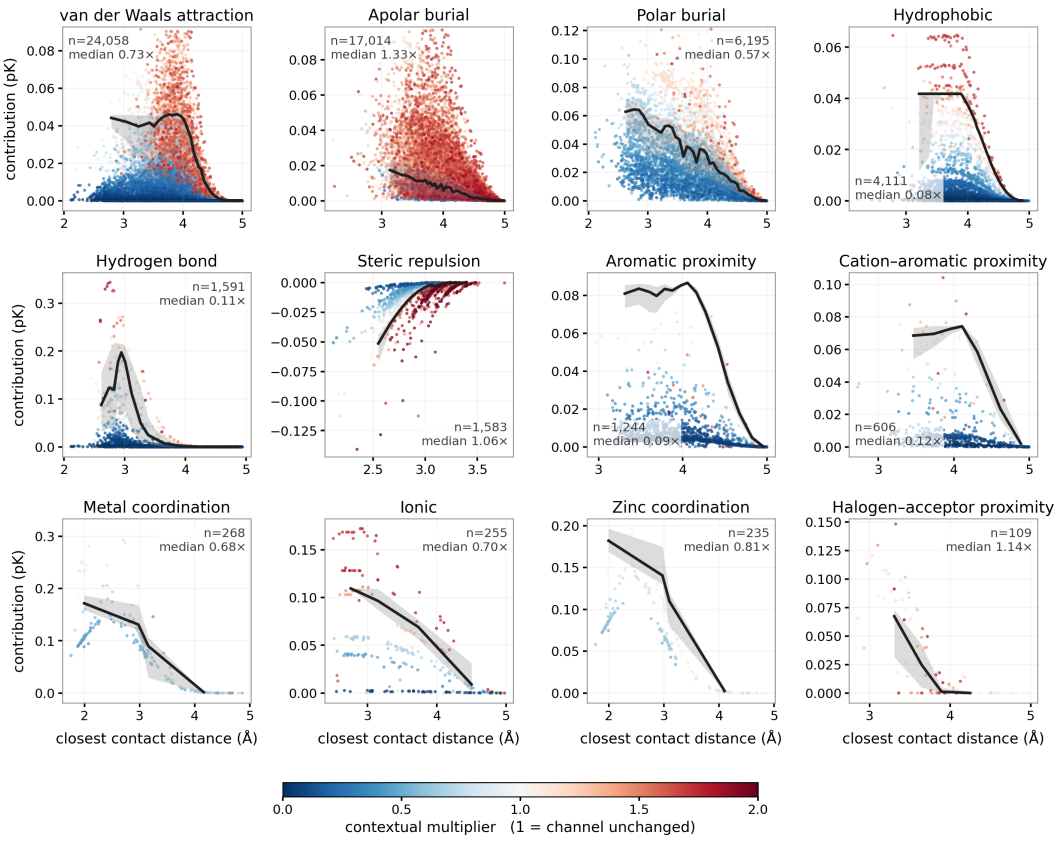


Figure S2: Contextual corrections are applied per contact and differ by channel. The dark line is the uncorrected profile (geometric basis times the dataset-wide channel weight), a median over equal-count bins with the interquartile range shaded; each point is one contact, coloured by its contextual multiplier, where one leaves the term unchanged. Vertical scales are independent and clipped to the 0.2–99.8 percentile. One checkpoint, 250 test complexes, 57,269 contacts. Figure 4 shows nine of these channels in the main text.

C.5 Where the contextual correction is applied

The range-use figures are averages, and an average cannot tell apart three cases: a correction applied uniformly across a complex, one applied to only a few contacts, and one that damps some contacts while amplifying others in the same channel. Figure S2 separates them by plotting every contact.

Each point is one (ligand atom, residue) contact at the value the model adds to the prediction, coloured by its contextual multiplier; the line is the same quantity before that multiplier is applied, which is the geometric basis of Table S1 times the single learned weight for that channel. The multiplier is bounded to $[0, 2]$ and zero-initialized, so an untrained model applies exactly one everywhere and training begins at the constant-coefficient configuration; any departure from one is learned. It is read from the forward pass rather than recovered by division, because the correction enters both before pooling and again through the interaction graph, so the ratio of the realized to the uncorrected term is not the multiplier. The line is a pooled group basis rather than the per-contact response f_c, so the Gaussian channels appear flatter than their definitions: a group takes the largest response among its contacts, and a second contact a few tenths of an Å further away holds that maximum near its peak.

Three properties are visible that the tabulated averages do not show. The correction is applied per contact rather than per complex, so contacts at the same distance in the same channel receive different multipliers. It is bidirectional within a single channel, damping some contacts and amplifying others, which a channel-level mean cancels. And its typical magnitude is channel-specific: apolar

Table S10: Change in residue prioritization after freezing contextual multipliers. Means and SDs use six trained seeds on 600 sequence-distinct complexes.

Removed factors	Top-1 changes	Top-5 set changes	Top-5 Jaccard	Residue Spearman
Contact/group + local	0.462 ± 0.016	0.826 ± 0.024	0.634	0.8775
Contact adjustment + group update	0.175 ± 0.032	0.387 ± 0.039	0.866	0.9805
Local coefficient only	0.388 ± 0.012	0.783 ± 0.027	0.669	0.8980

Table S11: Concentration of explanation mass over ligand-atom–residue contacts, The last three columns give the smallest fraction of contacts carrying 50%, 80% and 90% of the absolute mass; n is the mean number of contacts per complex.

Model	n	Entropy ↓	Eff. support ↓	50% ↓	80% ↓	90% ↓
PIGNet2	132.3	0.815	0.419	0.097	0.281	0.444
CheapNet occlusion	101.4	0.906	0.655	0.192	0.452	0.605
PACT	101.3	0.853	0.517	0.155	0.353	0.469

burial carries a median multiplier of 1.32 and van der Waals attraction 0.74, while hydrophobic contacts are reduced to 0.08 and aromatic proximity to 0.09, the difference between a channel the model keeps and one it effectively switches off for most contacts.

This establishes that the contextual pathway is active and non-uniform. It does not establish that the correction improves predictions, which is what the fixed-coefficient ladder in Appendix C.3 measures, nor that the resulting reranking is chemically sensible, which is Appendix D.7.

C.6 Does context alter residue prioritization?

We compare each trained model's contribution ledger with component counterfactuals: global weights and allocation alone, local chemistry alone, or contextual adjustment alone. Here contextual adjustment includes both the contact multiplier and the within-group update; allocation is held fixed. The same trained weights and the same forward pass supply both ledgers, so this diagnostic measures how much the contextual factors move the displayed decomposition. It does not measure the accuracy of a separately retrained ablation.

Table S10 reports the effect of freezing the contextual multipliers. Removing both coefficient pathways changes the highest-ranked residue in 46.2% of complexes and changes the top-five set in 82.6%. Most of this reranking is associated with the local coefficient. These results show that context is active and materially affects which residues receive contribution mass.

C.7 Sparsity and concentration of explanation mass

Exact zeros are not a useful sparsity measure, because almost every contact receives a nonzero contribution. We measure concentration instead. Over the n_x ligand-atom–residue contacts of complex x, write $p_u = |A_u| / \sum_j |A_j|$ and $H = -\sum_u p_u \log p_u$. We report normalized entropy $H/\log n_x$, effective support e^H/n_x, and the smallest fraction of contacts carrying 50%, 80% and 90% of the absolute mass. Lower means more concentrated. Each statistic is computed per complex and then averaged.

The contact is PACT's native group. PIGNet2's interaction pairs are summed to it, and CheapNet's is obtained by occluding the intermolecular edges that join one ligand atom to one residue.

PACT carries 90% of its mass in 46.9% of contacts, against PIGNet2's 44.4% and CheapNet's 60.5%. The comparison with CheapNet is the clean one, since the two read the same contact graph and average 101.3 and 101.4 contacts. PIGNet2 has 132.3 such contacts per complex, which lowers every fraction it reports, so its lower fractions are not directly comparable with PACT's.

Concentration is not a virtue on its own, so we make no claim of superior sparsity: a model that attributes less, or over more contacts, scores better on every column. These figures are read beside the attributed mass, 2.62 pK for PACT against 6.41 for PIGNet2 and 5.22 for CheapNet, and beside

Table S12: Sensitivity to geometry and protonation edits. Rigid-body rows use 250 complexes and mean ± SD across six checkpoints: translation alone uses three unit directions, translation with rotation uses four draws, following the protocol of the baselines scored with it.

Perturbation	n	Mean $\lvert\Delta\widehat{y}\rvert$ (pK)	Exact@5	Overlap@5
Translation 0.10 Å	250	0.0271 ± 0.0003	0.802 ± 0.014	0.960 ± 0.003
Translation 0.25 Å	250	0.0716 ± 0.0011	0.556 ± 0.010	0.906 ± 0.003
Translation + rotation 0.10 Å	250	0.0270 ± 0.0004	0.788 ± 0.006	0.957 ± 0.001
Translation + rotation 0.25 Å	250	0.0736 ± 0.0016	0.536 ± 0.007	0.901 ± 0.002
Deprotonate one acid	82	0.0667	0.780	0.954
Deprotonate all acids	82	0.1152	0.720	0.939
Neutralize cations	125	0.0674	0.776	0.954
Protonate bases	40	0.0867	0.425	0.875

the share of the prediction the attribution accounts for, which is 0.949 for PACT and 0.942 for PIGNet2 and undefined for occlusion explanations, CheapNet's included, since they provide no additive decomposition of which to take a share.

C.8 GEOMETRY AND MOLECULAR PREPARATION

We use two rigid-body perturbations. The first translates and rotates the ligand about its centroid with the protein fixed, splitting the displacement as $\mathrm{RMS}^2 = \mathrm{RMS}_{\mathrm{rot}}^2 + \|t\|^2$ with half assigned to rotation; PIGNet2, CheapNet, CENsible and GRIPHIN are scored with it. The second translates the ligand alone along three unit directions at the same amplitude and rebuilds the contact graph, so that a displacement can create or break a contact; RF-Score, SchNet, EGNN, IGN and GIGN are scored with it. Pure translation is in principle the gentler perturbation, so PACT is scored with both, on all six checkpoints, using the baselines' own implementation of each. The two agree within the seed spread (0.0270 ± 0.0004 pK and 0.0271 ± 0.0003 pK), so the choice of perturbation does not matter at this amplitude, and the main-table value can be compared with either group.

At 0.10 Å the prediction moves by under 0.03 pK and 95.7 to 96.0% of the top-five residues are retained; the two perturbations separate only at 0.25 Å, where adding rotation lowers Exact@5 from 0.556 to 0.536. Whole-complex rigid transformations change the prediction by at most about 1.4×10^{-6} pK. Base protonation edits move exact selection more than broad ranking, Exact@5 0.425 against Overlap@5 0.875; these are controlled input edits rather than samples from a validated distribution of protonation states.

The baseline rows use their own cohorts and are not paired with each other. PIGNet2 changes by 0.0907 pK with Exact@5 0.753 over four perturbations on 600 complexes, and by 0.0861 pK and 0.805 on the 71 complexes it shares with the common 250-complex cohort. CheapNet changes by 0.1676 pK over four perturbations on 249 complexes, rebuilding pockets and graphs through its input pipeline. CENsible changes by 0.0453 pK on the same cohort. GRIPHIN changes by 0.0127 pK over four draws on 250 complexes and six seeds, which is not comparable with the other rows because its GRAIL interaction field is reused rather than recomputed, so its ligand moves within a fixed protein field while the other rows re-derive the protein representation around the displaced ligand. LINKER has no entry because its input contains no coordinates: it uses a ligand graph, checkmol functional groups and ESM sequence embeddings, so a rigid-body displacement cannot change its prediction.

Retraining is not the only source of variation. For shared structures processed through two molecular-preparation pipelines, the median absolute prediction difference is 0.0374 pK and the 95th percentile is 1.270 pK, against a mean seed SD of 0.0662 pK on identical graphs.

D INTERVENTIONS AND CHEMICAL AGREEMENT

This section evaluates three distinct properties of an explanation. Selection fidelity asks whether highly ranked residues strongly affect the predictor. Quantitative fidelity asks whether the assigned

Table S13: Joint residue-removal diagnostics at $k = 5$. Both PACT rows use six checkpoints and 250 complexes, and differ only in how residues are ranked and what is claimed. CheapNet uses five seeds and 249 complexes of the shared 250-complex cohort. The one-complex coverage difference prevents a fully paired comparison.

Model	Top/random ↑	Top/nearest ↑	Contribution–effect discrepancy (pK) ↓
CheapNet occlusion	2.195 ± 0.106	1.534 ± 0.020	0.395 ± 0.061
PACT occlusion	2.349 ± 0.012	1.524 ± 0.008	0.130 ± 0.010
PACT	2.365 ± 0.011	1.534 ± 0.007	0.231 ± 0.015

pK contribution matches the recomputed change. Chemical agreement asks whether named channels track recognizable chemical interactions. None of these interventions simulates structural relaxation or an experimental molecular modification.

D.1 LINKER RESIDUE-REMOVAL PROTOCOL

LINKER occlusion zeros the selected residue masks and cached stage-1 logits, then recomputes the affinity head at batch size one. Stage 1 is held fixed. Of 250 requested complexes, 240 have sufficient usable residue features; 179 satisfy the stricter requirement that top, random, and nearest-residue plans each remove exactly five representable rows. On this matched subset, Top/random is 1.814, Top/nearest is 1.962, and joint-removal discrepancy is 0.0037 pK. The top-five removal effect is only 0.0588 pK, against 1.338 ± 0.022 pK for PACT at the same k, so the small absolute discrepancy should be interpreted alongside this small response.

A separate checkpoint audit found batch-composition dependence: changing from recorded batch-16 inference to batch-one inference shifted predictions by 0.217 pK on average and up to 1.358 pK on 2,141 complexes. The removal evaluation therefore holds batching fixed at one. Six retraining prediction sets are available for predictive evaluation, but their affinity-head weights are unavailable for cross-seed occlusion comparisons.

D.2 CENSIBLE RESIDUE-REMOVAL PROTOCOL

CENsible cannot be occluded in memory. Its input is a voxel grid together with 349 scoring terms computed by an external `smina` call, so deleting a residue means deleting its atoms from the protonated receptor and rebuilding both: the terms are recomputed by the same `smina` invocation the dataset build uses, and the grid by the direct `.pdb`/`.mol2` read that `censible/inference/inference.py` takes. Bond records are rewritten, so that surviving atoms keep their perceived bonds and therefore their gnina types.

The cohort is the 249 complexes of the shared 250 that CENsible's own reader retains, scored from one checkpoint: 4,634 single-residue deletions for the attribution and 1,245 joint removals for the five-residue sets. Recomputed baseline predictions reproduce the recorded ones to at most 1.05×10^{-3} pK, and the re-derived scoring terms are bit-identical to the build's, so the reported changes are the deletions and not the re-featurisation. At $k = 1$ the discrepancy is 0.000 pK, the agreement the occlusion rule builds in.

Top-five removal moves the prediction by 1.795 pK, against 1.338 ± 0.022 pK for PACT at the same k, so unlike LINKER's the discrepancy of 0.743 pK accompanies a large response rather than a small one.

D.3 RESIDUE SELECTION AND JOINT REMOVAL

For PACT, residue removal deletes its ligand–protein contact edges. The controls remove the same number of residues with the same operator. We rank native explanations by absolute contribution but keep the sign when computing agreement. We also test an occlusion version of PACT on the same cohort and nulls, ranked by individual deletion effects.

Table S14: Selection fidelity under ranking rules other than occlusion, on each model's own cohort and removal operator; only the rule ordering residues changes.

Model	Ranking	k=1	k=3	k=5	k=8
IGN	occlusion	**2.510**	2.020	1.585	1.260
	path gradient	2.385	**2.152**	**1.732**	**1.356**
GIGN	occlusion	**2.513**	1.646	1.220	1.064
	path gradient	1.947	1.694	**1.500**	**1.279**
EGNN	occlusion	**2.379**	1.778	1.400	1.205
	path gradient	1.973	1.754	**1.473**	**1.305**
CheapNet	occlusion	**1.779**	**1.746**	**1.489**	**1.232**
	attention	0.783	0.891	0.896	0.858
LINKER	occlusion	**2.812**	**2.325**	**1.962**	**1.653**
	attention	0.777	0.621	0.594	0.633

Table S13 shows similar selection ratios for all three rows. The main difference is numerical agreement: PACT occlusion is better at $k = 5$, but native contributions are better by $k = 8$ because separate deletion effects do not add well for larger sets.

Removal is architecture-specific: models remove atoms, edges, counts or rebuilt features. These results therefore compare complete model–explanation–intervention procedures. Appendix D.4 also compares each model's native ranking.

D.4 RANKING RULES OTHER THAN OCCLUSION

The selection experiment reaches its ranking through one function, which orders residues by the magnitude of their assigned score. Replacing that score with a different attribution, while holding the cohort, the removal operator, the proximity null and the re-forward fixed, isolates the ranking rule. We re-rank each model by its own native attribution: a path-gradient attribution (Sundararajan et al., 2017) for the geometric networks, with its endpoints matched to the intervention the removal operator performs, and attention mass in the ligand-to-protein direction, mapped onto residues through the model's own pooling, for the attention models. Both alternatives reproduce their occlusion baselines before being compared, and the residue universe is identical under either rule.

Occlusion wins at $k = 1$ for every model, but that column is close to a tautology: the occlusion score for a residue is its own single-residue removal effect, which is exactly the quantity $k = 1$ measures, so ordering by it maximises that numerator by construction. The comparison becomes informative once removal is joint and the sum of single effects is no longer the joint effect. There the alternative is the better ranking for the three edge-mode models, leading at $k = 5$ by 0.147 (IGN), 0.281 (GIGN) and 0.073 (EGNN), and at $k = 8$ by 0.096, 0.215 and 0.101. At $k = 3$ the difference is unresolved for GIGN and EGNN. SchNet is excluded from that reading, because its evaluation removes atoms while its attribution is over edges, so the two rows attribute different interventions.

Attention does much worse than occlusion. For both attention models it is below occlusion at every k and below one throughout, so the residues it selects move the prediction less than the nearest residues do.

D.5 SELECTION FIDELITY ACROSS REMOVAL SIZES AND NULL MODELS

A selection ratio depends on its reference, so we compare the top residues with random, nearest and next-ranked residues. Occlusion explanations rank by deletion effect, so their agreement at $k = 1$ is built in. Cohorts and removal operators also differ between models.

Table S15 shows that PACT's independent ranking beats the nearest-residue baseline at every k. PIGNet2 is below this baseline at $k = 1$ and above it from $k = 3$. PACT occlusion gains at $k = 1$, but the gap from its native ranking is within the seed spread by $k = 5$.

Table S15: Selection fidelity at k residues: mean absolute prediction change for the top k residues over that for the k nearest. Operators differ by row (Appendix D.3).

Model	Unit	Claim	k=1	k=3	k=5	k=8
IGN	edge	occlusion-derived	2.510	2.020	1.585	1.260
GIGN	edge	occlusion-derived	2.513	1.646	1.220	1.064
EGNN	edge	occlusion-derived	2.379	1.778	1.400	1.205
SchNet	atom	occlusion-derived	1.997	1.831	1.802	1.609
RF-Score	count	occlusion-derived	2.728	1.587	1.293	1.138
CheapNet	residue	occlusion-derived	1.779	1.746	1.489	1.232
PACT	edge	occlusion-derived	1.865	1.758	1.524	1.276
PIGNet2	residue	independent	0.861	1.083	1.259	1.339
PACT	edge	independent	1.816	1.749	1.534	1.290

Table S16: Bias-free deletion and preservation fidelity for PACT, one checkpoint on 250 complexes: the attributable part of the prediction surviving removal, over the same part before removal. Delete removes the selected residues, preserve removes all others. "Next" takes the $k\%$ ranked immediately below the selected set, "random" draws $k\%$ uniformly with five draws per complex. Every complex is scored at every fraction.

	Delete ↓			Preserve ↑		
Fraction	Top	Next	Random	Top	Next	Random
0.05	0.802	0.878	0.943	0.293	0.216	0.105
0.10	0.701	0.837	0.901	0.436	0.279	0.174
0.20	0.546	0.799	0.816	0.609	0.332	0.299
0.30	0.422	0.789	0.733	0.725	0.348	0.406

We also compare top, next-ranked and random fractions of the pocket. Since removal cannot change the intercept or ligand correction, we report the remaining attributable part $A(x) = \sum_r A_r(x)$:

$$\rho_f^{\mathrm{del}}(x) = \frac{A(x \setminus S_f)}{A(x)}, \qquad \rho_f^{\mathrm{pre}}(x) = \frac{A(x \setminus (\mathcal{R}_x \setminus S_f))}{A(x)},$$

where S_f contains the largest $|A_r|$ values. Lower deletion and higher preservation scores are better. The denominator is safe because $A(x)$ is at least 0.52 pK on this cohort.

Table S16 shows that the ordering matters at every fraction. For example, deleting the top 5% removes 19.8% of attributable prediction, compared with 12.2% for next-ranked and 5.7% for random residues. PIGNet2 is absent because its recorded runs do not provide the attributable part required by this ratio.

Selection and numerical calibration are different. Table 3 compares assigned contributions with removal effects, pooled over $k \in \{1, 3, 5, 8\}$, on the cohorts and removal operators of Table S15. PACT has the lowest discrepancy and a slope closest to one. PIGNet2 shows that an exact additive decomposition can still disagree with removal effects. CENsible and LINKER are absent there because their removal evaluations cover $k \in \{1, 5\}$ only and cannot be pooled on this grid; their values at those two sizes are in Appendices D.2 and D.1, and GRIPHIN is absent for the reason given in Appendix D.3.

D.6 Quantitative contribution–effect agreement

For single-contact evaluation, each complex contributes up to eight interventions: the four contacts with the largest absolute assigned contributions and four sampled without replacement from the rest, giving 2,000 interventions per seed over the 250-complex cohort. The observed effect is $e_i = \widehat{y}(x) - \widehat{y}(x \setminus \{i\})$ and the displayed contribution is a_i. Instantiating infidelity (Yeh et al., 2019) for contact removal with the displayed contribution as the assigned effect, our normalized estimator is specific to the sampled interventions:

$$\mathrm{Infid} = \frac{\sum_i (a_i - e_i)^2}{\sum_i e_i^2}.$$

Table S17: Quantitative agreement with contact removal. Single-contact values are mean ± sample SD across six seeds. The top-five row uses joint residue removal for six checkpoints and should not be compared directly with the single-contact row.

Intervention	Slope (target 1)	Normalized infidelity ↓	Discrepancy (pK) ↓	Sign agreement ↑
Single contact	0.9627 ± 0.0018	0.0499 ± 0.0018	0.01113 ± 0.00014	0.807
Top-five residues	0.8598 ± 0.0043	0.0357 ± 0.0030	0.231 ± 0.015	–

Table S18: Average precision against ProLIF annotations, averaged over six seeds on one candidate universe. "Raw" is the predefined channel basis, "fixed" adds the global channel coefficient and allocation, "contact" retains contact adjustment and the within-group update. These are component counterfactuals, not retrained architectures. Ionic AP is undefined, as the cohort has no ionic positives.

Interaction	n_+	Raw	Fixed	Local	Contact	Distance	Random
Aromatic	511	**0.2283**	0.2282	0.1725	0.2133	0.0104	0.0082
Cation–aromatic	135	**0.1282**	0.1276	0.0733	0.0912	0.0029	0.0019
Halogen	43	0.4631	0.4631	**0.4672**	0.3293	0.0049	0.0005
Hydrogen bond	1,623	0.5687	0.5687	0.5285	**0.5783**	0.4085	0.0257
Hydrophobic	2,831	**0.3600**	0.3600	0.2525	0.3237	0.0601	0.0448
Metal	103	**0.7908**	0.7908	0.7738	0.7903	0.4377	0.0018
van der Waals	10,107	0.3065	0.3562	0.2977	0.3450	**0.7967**	0.1602

The calibration slope is from an ordinary least-squares regression of e_i on a_i with an intercept, and calibration requires a slope near one with an intercept near zero. Discrepancy is $\frac{1}{N}\sum_i |a_i - e_i|$ in pK, and sign agreement is computed over interventions where both quantities are nonzero.

Table S17 shows close numerical agreement for individual contact interventions. It does not imply that every weak effect has a reliable sign, nor that removing a contact at fixed coordinates represents an experimental energetic perturbation. For an occlusion explanation the agreement with the corresponding individual removal is built in; joint removals stay informative because the sum of individual effects need not equal the jointly recomputed one.

D.7 Chemical agreement and the predefined bases

We evaluate 62,191 ligand-atom–residue pairs from 600 complexes and six seeds. ProLIF 2.2.1 (Bouysset & Fiorucci, 2021) provides binary interaction labels. All variants use the same candidates; "raw" uses only the predefined chemical basis.

Table S18 shows that the predefined representation provides much of the chemical agreement. Contact context improves hydrogen-bond and van der Waals scores, while the raw basis is better in four other classes. Distance alone is strongest for van der Waals interactions. ProLIF labels interaction presence, not energy, so this test shows chemical recognition but not physical accuracy.

D.8 Interaction recovery across altered poses

Static annotation agreement partly rewards chemical rules already present in the input. The pose-recovery experiment, motivated by interaction-based pose assessment (Errington et al., 2025), asks whether a named channel responds when its own interaction pattern changes. We score 22,487 redocked poses from 285 target groups and compare within-target pose pairs whose RMSD differs by at most 0.5 Å and whose reference-interaction recovery differs by at least 0.2. The aggregate PACT score prefers the pose with greater recovery in 62.5% of 17,633 eligible comparisons, against 62.4% for contact count, so the total score alone separates little from that control. The class-specific test asks the same question of each named channel.

Six of eight named channels prefer greater recovery more often than either the total score or contact count (Table S19), led by metal at 0.836 and ionic at 0.762. RMSD matching is only approximate: differences up to 0.5 Å remain and RMSD itself reaches preference 0.652 over the whole pose-pair set, so the analysis reduces geometric confounding without eliminating it. Multiple pairs can share poses and targets, so pair counts are not independent sample sizes.

Table S19: Preference for greater class-specific interaction recovery among approximately RMSD-matched pose pairs. The named channel exceeds both controls for hydrogen-bond, hydrophobic, ionic, metal, cation–aromatic, and aromatic recovery.

Recovered interaction	Named channel	Total score	Contact count
van der Waals	0.524	0.608	0.649
Metal	**0.836**	0.695	0.577
Ionic	**0.762**	0.489	0.524
Hydrophobic	**0.622**	0.533	0.589
Hydrogen bond	**0.655**	0.607	0.509
Halogen	0.596	0.648	0.583
Cation–aromatic	**0.658**	0.574	0.584
Aromatic	**0.549**	0.549	0.504

Table S20: Five largest residue-level contributions for 1c7e, seed 0. Residue identifiers come from the verified graph-to-PDB mapping and all lie in chain A. Removal deletes all ligand-contact edges involving the selected group while holding coordinates fixed.

Residue	Assigned contribution (pK)	Individual removal effect (pK)
Tyr98	0.651	0.396
Trp60	0.545	0.433
Gly61	0.337	0.322
Tyr100	0.287	0.286
Thr59	0.252	0.180

D.9 Worked molecular explanation

The complex 1c7e illustrates the distinction between an assigned contribution and a recomputed intervention effect. For the evaluated checkpoint, predicted affinity is 6.91 pK against an observed 6.30 pK. That error of 0.61 pK is well below the 1.506 pK test-set RMSE of the same checkpoint without being negligible, so the example is neither a best case nor one in which the explanation describes a badly wrong prediction.

Table S20 spans the whole range of agreement inside one complex. Tyr100 is assigned 0.287 pK and its removal moves the prediction by the same amount to within 0.001 pK; Gly61 agrees to 0.015 pK; Tyr98 is assigned 0.651 pK against a removal effect of 0.396 pK. The gap is contextual redistribution: deleting a group recomputes the contextual terms of the groups that remain, so part of the deleted group's assigned value is taken up by them rather than lost. Neither quantity is an experimental residue binding free energy.

E Baseline reproduction and model size

This section specifies the information needed to reconstruct the baseline experiments: where each implementation was obtained, what was changed relative to it, and how large each model is. The PACT experiments are identified by their saved configuration and checkpoint manifests. All models were trained on two NVIDIA RTX A4000 GPUs.

E.1 Baseline source code and retraining

Baselines were retrained from their released implementations, at the revisions of Table S21, on the noncovalent LP-PDBbind cohort of Appendix B.1 (7,393 training, 958 validation, 2,171 test), which supplied membership, labels, and the validation set wherever the source procedure requires one. Architectures and configurations were kept as released, except where noted below, and repeated runs use five explicit seeds. Complexes that a model's input processing rejects are dropped from that model's cohort rather than replaced.

Table S21: Baselines code sources: the first row is for RF-Score, SchNet, EGNN, IGN and GIGN. Model definitions were used as released in every case, any modifications are described in the text.

Source	Revision
guaguabujianle/GIGN	`ef514d326202d71b0674ae20e95456454bf10c6e`
mseok/PIGNet2	`6cf5a26789c021fa01dcb49696ffa0cd1f2ab44b`
durrantlab/censible	`9038756cf0e9759b462e38f448c647a751725c73`
molinfo-vienna/GRIPHIN	`3a12278d614c854c22a05e3361a8b19080a2f0be`
HySonLab/LINKER	`d7a6e9e2eac6dd7c861fd65548773b72a1275d81`
hyukjunlim/CheapNet	`de9e5b3c30d6e9f707250e370d8fad076d3e571e`

RF-Score, SchNet, EGNN, IGN and GIGN. All five come from the GIGN benchmark repository with its files unchanged – RF-Score (Ballester & Mitchell, 2010), SchNet (Schütt et al., 2017), EGNN (Satorras et al., 2021), IGN (Jiang et al., 2021) and GIGN (Yang et al., 2023) – using the supplied features, model classes, interaction layers, graph builders and readouts; RF-Score's random state is set per replicate. Three adaptations concern input preparation and execution rather than architecture: a local data adapter using a whole-residue 6 Å pocket shell and deposited ligand SDF files, which differs from the source preprocessing and aligns the contacted-residue population for the explanation comparisons; compatibility shims restoring renamed DGL and NumPy symbols; and SchNet's radius-graph adapter, which keeps the nearest neighbors when the cap is exceeded and caches the graph at fixed coordinates.

PIGNet2. A local adapter supplies the split and labels; atom features, physical energy terms, interaction graph and loss come from the pinned release (Moon et al., 2024). Labels are converted once from pK to kcal/mol at 298.15 K, instead of with the fixed factor used upstream. Its native interaction-term contributions serve as the explanation, converted to the common pK scale. Our runs differ from the published PIGNet2 in six ways. (i) The release contains two model classes. We use `PIGNet` with its released configuration, whose van der Waals term is a Lennard-Jones potential; the published training scripts use `PIGNetMorse`, which uses a Morse potential. (ii) Weights use the PyTorch default initialization, not the Xavier initialization of the release. (iii) Training starts from scratch, not from the published weights, and uses crystal poses only. The pose augmentation, redocking, cross-docking and random-negative data of the published system are not used. (iv) Training stops early on validation RMSE, after at most 400 epochs, while the release trains for 5,000 epochs without a validation set. Each seed is scored alone, not in the four-model ensemble the authors recommend. (v) Pockets are the whole-residue 6 Å shell used for the other baselines, and the release's protonation step is skipped. The release uses a 5 Å shell and protonates both the protein and the ligand. (vi) Ligands are read from the SDF files only. Complexes without a processed input are dropped, almost all because RDKit cannot sanitize the SDF ligand: 1,022 training, 122 validation and 309 test complexes. This leaves 6,371 of 7,393 training, 836 of 958 validation and 1,862 of 2,171 test complexes.

CENsible. Its scoring-term preparation, voxel representation, coefficient network and training procedure were used on the shared cohort, through the source CUDA path with upstream files unchanged (Bhatt et al., 2024). Two preparation steps do not run as released and were bridged at call time: scoring-term extraction redirects to a null device the output it then parses, so that redirection was removed while its parsing, concurrency and output format were left intact; and atom typing requires a binary that cannot be installed here, so the same typing was produced through the gridding library the model already depends on, with the resulting voxel grids verified identical to those from the read path used at inference. Upstream seeds no randomness, so explicit seeding was added, the only departure affecting the training loop. Upstream performs no validation-based checkpoint selection and that protocol was retained, so the reported model is the final epoch.

GRIPHIN, LINKER and CheapNet. Each keeps its own preprocessing, training configuration and model definition, with cohort-restricted input files that change no selection logic (Rose et al., 2026; Pham et al., 2026; Lim et al., 2025). GRIPHIN and LINKER hardcode seed 42, overridden with seeds 0–4. GRIPHIN's released configuration was used with its data-split key set to the leak-proof split; it matches the hyperparameters recorded in the authors' published LP-PDBbind

Table S22: Parameter counts for the configurations evaluated here. RF-Score is a random forest, so its forest size is reported instead. LINKER's count covers its two trained stages, excluding the frozen protein language model run once during preparation. Counts indicate scale only.

Model	Parameters	Counted from
PACT	4,784	trained checkpoint
SchNet	281,395	trained checkpoint
PIGNet2	567,816	trained checkpoint
GIGN	626,177	trained checkpoint
CheapNet	1,341,849	trained checkpoint
EGNN	1,591,553	trained checkpoint
IGN	1,667,046	trained checkpoint
CENsible	4,453,462	trained checkpoint
LINKER	67,368,769	trained checkpoints, both stages
GRIPHIN	193,805,067	pinned source, training configuration
RF-Score	500 trees, 1,703,546 nodes	trained forest

checkpoint, and that checkpoint, scored here without retraining, reproduces its published correlation to within 0.001, which verifies the preparation pipeline. LINKER resolves its functional-group table and molecule-typing executable against locations absent from the distributed layout and filters every split against an absolute path on the authors' machine; these three references were redirected at run time, leaving the forward pass, loss, optimizer, schedule and featurizer untouched. Its protein representation is produced once by a frozen protein language model and is not trained, and its published second stage selects the epoch on the test set, which was reproduced rather than corrected. CheapNet's cached-graph loader predates a change in its tensor library's deserialization default; the stored graph classes were allowlisted rather than the check disabled.

E.2 RECONSTRUCTING PREDICTIONS AND EXPLANATIONS

Reconstruction follows the released manifest, which records each model's source revision, preprocessing, hyperparameters, seeds, checkpoint-selection rule and execution commands. Two points are specific to the explanation evaluation. Batch composition is held fixed between intact and intervened evaluations, because at least one baseline's predictions depend on it (Appendix D.1). Attention values are used as selection scores only, and are never converted into contributions.

E.3 MODEL SIZE

Table S22 reports the number of parameters of each model as reproduced here. Counts were taken from a trained checkpoint where one was retained, and otherwise by instantiating the pinned published architecture under the configuration used for training; they therefore describe the configurations evaluated in this paper rather than values quoted from the original publications. A count includes every component required to produce a prediction.