CSRE-MOL: CONDITIONAL RESIDUAL EVIDENCE FOR MOLECULAR PROPERTY PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

A frozen molecular classifier supplies a global score, while labeled analogues expose an inference-time structural signal that may repeat what the classifier already knows. We ask a sharper fusion question: which part of retrieval evidence remains after accounting for the score being augmented and for neighborhood support? We present **Calibrated Structural Residual Evidence for Molecular Prediction (CSRE-Mol)**, a post-processor that estimates the predictable component of normalized retrieval evidence with a task-wise ridge projection and applies only the conditional residual to the frozen score. A positive-slope Platt map then separates score correction from probability calibration. On the existing development benchmark of GIN, GraphGPS, GraphMVP, and Uni-Mol across eight MoleculeNet classification datasets using three random seeds, CSRE-Mol raises mean ROC-AUC from 70.07% to 75.97%, mean average precision from 45.81% to 52.99%, and lowers mean NLL from 0.3933 to 0.3625. The predictable evidence component alone reaches 69.85% ROC-AUC, whereas the conditional residual reaches 75.97%; a matched conditional permutation falls to 70.48%. These controls support a conditional-innovation interpretation of the gain rather than treating retrieval as a second predictor. The evaluation uses an existing development split, and Uni-Mol/ToxCast identifies a useful boundary condition: conditional correction helps when the residual neighborhood signal is informative, but cannot recover ranking signal that is absent.

1 INTRODUCTION

Molecular property prediction often has two sources of task information. A learned encoder maps the query molecule to a global score, while labeled analogues provide a neighborhood signal at inference time. These sources are not automatically additive: a neighborhood may repeat evidence already encoded by the model, and sparse or imbalanced neighborhoods may make a local score appear more reliable than its support warrants. The central issue is therefore not how to attach another predictor, but which part of the auxiliary signal is actually incremental.

The closest existing operations each preserve a different object. Post-hoc calibration changes the mapping from a score to a probability while preserving its ranking (Kuleshov et al., 2018; Scalia et al., 2020). Retrieval and nearest-neighbor predictors treat the neighborhood as an independent local predictor (Khandelwal et al., 2020); direct fusion then adds that predictor to the global score. Stacking learns a supervised map from several predictors, while partialling-out removes a component explained by conditioning variables (Wolpert, 1992; Breiman, 1996). Structure-based similarity is useful when chemical and property similarity align, but activity cliffs and sparse neighborhoods make an unconditional local score unreliable (Rogers & Hahn, 2010; Maggiora, 2006; Stumpfe & Bajorath, 2012; van Tilborg et al., 2022). These lines leave a common fusion target underspecified: how much of the local signal remains after accounting for the global prediction and the amount of neighborhood support?

Our central principle is to make *conditional residual evidence*, rather than raw local evidence, the object being added to the model. For a frozen predictor with standardized score z_g and a structural evidence coordinate e, we fit a regularized best-linear approximation $\hat{m}(z_g, c)$ to normalized evidence using the global score and neighborhood support c. Only

$$r = \mathrm{clip}(\tilde{e} - \hat{m}(z_g, c), -6, 6)$$

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.

modifies the standardized global score. This changes the fusion problem from adding two predictions to testing whether the residual evidence contains label information conditional on the existing score and its support. The resulting predictions are concrete: the predictable component should not outperform the global score, conditional permutation should remove the gain, and increasing residualizer capacity should not be necessary if the conditional target is the right object. The claim is operational and testable; it is not a claim that the residual is statistically independent or causally novel.

We instantiate this principle as **Calibrated Structural Residual Evidence for Molecular Prediction (CSRE-Mol)**. Morgan and MACCS neighborhoods provide class-balanced evidence, a regularized conditional model predicts its score-explained component, and a task-wise monotone Platt map converts the corrected score to a probability. The backbone is frozen; training labels define retrieval statistics, validation features define the conditional correction, and validation labels define the final probability map. The formulation therefore separates structural correction from probability interpretation without introducing a second end-to-end molecular predictor.

The experiments directly test the diagnosis. Global calibration improves probability quality while leaving mean ROC-AUC essentially unchanged (70.07% before and after calibration); direct addition reaches 75.35% AUC, below CSRE-Mol's 75.97%; and a calibrated local-only predictor reaches 74.53%. The predictable component alone reaches 69.85% AUC, while the residual reaches 75.97%; conditionally permuting the residual reduces the mean to 70.48%. A linear stacker reaches 73.09%, and a quadratic conditional residualizer reaches 74.86%, so the completed capacity control does not displace the low-capacity target. These results are complemented by view, conditioning-coordinate, label-budget, and neighbor-count analyses. Across GIN, GraphGPS, GraphMVP, and Uni-Mol on eight MoleculeNet datasets using three random seeds, CSRE-Mol improves mean AP from 45.81% to 52.99%, lowers mean NLL from 0.3933 to 0.3625, and improves ROC-AUC in 31 of 32 backbone–dataset cells. The Uni-Mol/ToxCast result marks the boundary of the principle: a residual correction helps when the residual neighborhood signal is informative, but cannot recover ranking signal that is absent. The results establish the conditional-innovation effect within this development protocol and identify outer-audit and broader nonlinear-comparator settings as natural next tests.

2 Related Work

Learned molecular representations. Message-passing networks and pretrained chemical encoders compress molecular structure into a global task score (Gilmer et al., 2017; Wu et al., 2018; Yang et al., 2019; Ross et al., 2022; Chithrananda et al., 2020; Xiong et al., 2020; Duvenaud et al., 2015; Kearnes et al., 2016; Xu et al., 2019; Ying et al., 2021; Rampásek et al., 2022; Rong et al., 2020; Zhou et al., 2023; Liu et al., 2022; Schütt et al., 2018; Faber et al., 2017). Their design target is representation-level prediction from the query molecule. Once the encoder is fixed, however, labels from nearby molecules are an additional inference-time source rather than another parameterized view of the same query. Our setting asks how to use that source without counting the global representation twice.

Structure-based similarity. Extended-connectivity fingerprints and Tanimoto similarity make the neighborhood a local predictor (Rogers & Hahn, 2010). This is valuable when structural and property similarity coincide, but activity cliffs expose the failure of treating proximity as an unconditional property predictor (Maggiora, 2006; Stumpfe & Bajorath, 2012; van Tilborg et al., 2022). Sparse or class-imbalanced neighborhoods add a second failure mode: the same local score should not have the same weight under different support. CSRE-Mol treats the neighborhood coordinate as evidence whose incremental part depends on both the global score and support.

Uncertainty and probability calibration. Uncertainty-aware predictors estimate predictive spread or decompose it into aleatoric and epistemic components (Amini et al., 2020; Kendall & Gal, 2017; Soleimany et al., 2021). Calibration methods improve the relationship between a score and empirical frequency (Kuleshov et al., 2018; Abdar et al., 2021; Platt, 1999; Guo et al., 2017; Zadrozny & Elkan, 2002; Niculescu-Mizil & Caruana, 2005; Naeini et al., 2015), and molecular benchmarks show that this relationship can be poor even when ranking is useful (Scalia et al., 2020; Gal & Ghahramani, 2016; Lakshminarayanan et al., 2017; Hirschfeld et al., 2020; Gruich et al., 2023; Hüllermeier & Waegeman, 2021; Angelopoulos & Bates, 2023). Calibration preserves the ordering of a score; it

therefore cannot recover structural information absent from that ordering. CSRE-Mol uses calibration after, rather than instead of, the structural score correction.

Retrieval-augmented and kernel predictors. Retrieval can expose labeled examples that were not compressed into a frozen model, and nearest-neighbor predictors provide a clear precedent for inference-time augmentation (Khandelwal et al., 2020). Deep kernels and Gaussian processes instead model dependence through a learned representation or covariance structure (Wilson et al., 2016; Rasmussen & Williams, 2006; Ralaivola et al., 2005; Matheron, 1963; Cressie, 1993; Stein, 1999; Johnson et al., 2019). These lines specify how a second source is represented or fused; they do not make redundancy with the existing score the quantity to be estimated. Our conditional residual formulation targets that missing relation directly.

Post-hoc molecular refinement. Molecular uncertainty and active-learning systems often use model confidence to choose new measurements (Reker & Schneider, 2015; Settles, 2009; Graff et al., 2021; Soleimany et al., 2021). Those systems change the data available for later training. CSRE-Mol keeps the backbone and development protocol fixed and uses the existing labeled bank only as an inference-time reference. Its distinction from direct local prediction, unconditional score addition, and calibration-only refinement is the conditional target: the method estimates which part of local evidence remains after the global score and support have explained what they can.

Stacking and orthogonalization. Stacked generalization learns a supervised mapping from several predictors to an outcome (Wolpert, 1992; Breiman, 1996). Partialling-out and orthogonalization instead remove a component explained by nuisance covariates before estimating a target effect; double/debiased machine learning combines this idea with cross-fitting for semiparametric estimation (Chernozhukov et al., 2018). CSRE-Mol is closest in spirit to partialling-out an auxiliary evidence coordinate, but it is not a Neyman-orthogonal estimator or a double/debiased machine-learning procedure: its residualizer is a task-wise regularized best-linear approximation, and its validation labels enter only the subsequent Platt fit. The empirical question is whether this constrained fusion target isolates a useful residual relative to direct and supervised stacking.

Position of this work. The relevant distinction is between treating a neighborhood as a second predictor and treating it as a source of incremental evidence. Morgan, MACCS, weighted Tanimoto, ridge regression, and Platt scaling are established ingredients. The principle introduced here is to condition the local evidence on the score being augmented and on neighborhood support before it is allowed to change that score. The same formulation predicts the observed failure mode: a direction-preserving probability map cannot repair a residual that is uninformative for ranking.

3 METHOD

CSRE-Mol maps a frozen probability and train-derived structural arrays to a calibrated probability. All arrays have shape $[n, t]$, where n indexes molecules and t indexes tasks. The procedure has three stages.

Figure 1 gives the complete data flow. The frozen predictor supplies the standardized global logit z_g; the labeled bank supplies two class-contrast evidence views and a support coordinate c; conditional residualization removes the component predictable from (z_g, c); and the monotone probability map produces the final output. The backbone is never updated by this post-processor. The retrieval branch is therefore an inference-time evidence construction, not a second end-to-end molecular predictor.

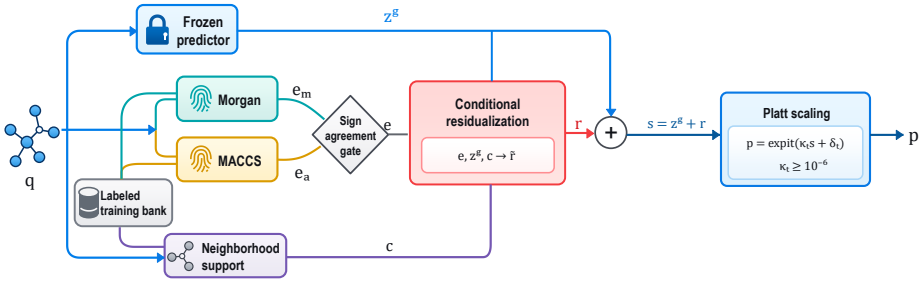


Figure 1: Overview of CSRE-Mol. A frozen global predictor yields z_g, while a labeled training bank produces Morgan and MACCS class-contrast evidence (e_M, e_A) and a separate support coordinate c. The agreement gate forms e; conditional residualization predicts its redundant component from (z_g, c) and applies the clipped residual r; positive-slope Platt scaling maps the corrected score s to the final probability p. Bank statistics are fit on the training data; conditional coefficients and feature scales are fit on validation features; and calibration uses validation labels only.

3.1 Two-view class-contrast signal

For task t, let $\mathcal{T}_{+t}$ and $\mathcal{T}_{-t}$ denote labeled positive and negative training molecules. We form Morgan radius-1 fingerprints with 1024 bits and MACCS fingerprints, following standard extended-connectivity and substructure-key constructions (Rogers & Hahn, 2010; Durant et al., 2002; Landrum, 2006). Each view uses task-specific weighted Tanimoto similarity, a fingerprint similarity measure with established justification in chemical similarity searching (Bajusz et al., 2015; Willett, 1998). If c_{+b} and c_{-b} are the counts of bit b in the positive and negative banks, with class sizes n_+ and n_-, the bit weight is

$$w_b = 1 + \min\left(\frac{\left|\log\frac{c_{+b}+2}{n_+-c_{+b}+2} - \log\frac{c_{-b}+2}{n_--c_{-b}+2}\right|}{1+\frac{1}{c_{+b}+2}+\frac{1}{n_+-c_{+b}+2}+\frac{1}{c_{-b}+2}+\frac{1}{n_--c_{-b}+2}}, 4\right). \tag{1}$$

The additive smoothing and cap keep the contrast finite for rare or absent bits; the equation is the exact weight used by the implementation and is consistent with frequency-aware fingerprint weighting of chemical similarity search (Rogers & Hahn, 2010; Willett, 1998). For query q and training molecule x, weighted Tanimoto similarity is

$$\mathrm{sim}_w(q,x) = \frac{\sum_b w_b q_b x_b}{\sum_b w_b q_b + \sum_b w_b x_b - \sum_b w_b q_b x_b}. \tag{2}$$

This weighted form reduces to ordinary Tanimoto when all $w_b = 1$ and is evaluated on the binary fingerprints generated by the cheminformatics toolkit (Bajusz et al., 2015; Landrum, 2006).

For each class and view, up to eight similarities are retained. The view evidence is a temperature-scaled log-mean-exp contrast,

$$e_v(q) = \mathrm{clip}\left(\log\frac{1}{|P_v|}\sum_{s\in P_v} e^{s/0.1} - \log\frac{1}{|N_v|}\sum_{s\in N_v} e^{s/0.1}, -4, 4\right), \tag{3}$$

where P_v and N_v are the positive and negative top-eight similarity sets. If either class is absent, the view contributes zero. The temperature and clipping constants are fixed before evaluation; they control the dynamic range of the evidence coordinate rather than introduce a learned predictor. The smooth top-k aggregation is an implementation choice layered on the established retrieval representation (Cover & Hart, 1967; Khandelwal et al., 2020). The aggregate is

$$e(q) = \mathbb{1}\{\mathrm{sign}(e_M) = \mathrm{sign}(e_A)\}\frac{e_M(q) + e_A(q)}{2}, \tag{4}$$

with M and A denoting Morgan and MACCS. The indicator implements the pre-specified agreement rule: if the two views disagree in sign, their aggregate is set to zero. Similarity search is therefore used as a retrieval signal, not as a second end-to-end molecular encoder (Cover & Hart, 1967; Khandelwal et al., 2020). The gate suppresses a conflicting raw aggregate; it does not assert that the final residual vanishes.

A separate support variable $c(q)$ averages two ordinary maximum similarities: for each of the Morgan radius-2 and MACCS views, we take the maximum similarity over the capped positive and negative support banks and then average the two view maxima. The support bank follows the archived deterministic policy: all training rows up to 20,000, otherwise a deterministic 4,096-row bank, with the per-task negative bank capped at $\min(N_-, \max(64, 4N_+))$. This variable is not the weighted radius-1 evidence coordinate.

3.2 Conditional score correction

Let $p_g(q,t)$ be the frozen global probability. We use the conventional logit coordinate for an additive correction (Cox, 1958; Platt, 1999) and compute

$$\ell_g = \log \frac{\mathrm{clip}(p_g, 10^{-6}, 1-10^{-6})}{1-\mathrm{clip}(p_g, 10^{-6}, 1-10^{-6})}. \tag{5}$$

The clipping prevents numerical infinities at exactly zero or one and matches the implementation's input contract. On validation features, we store the median and IQR$/1.349$ scale, falling back to standard deviation and then one when necessary; this robust feature scaling is estimated without test labels (Scalia et al., 2020):

$$z_g = \frac{\ell_g - m_g}{s_g}, \qquad \tilde{e} = \frac{e - m_e}{s_e}. \tag{6}$$

For each task, we fit a three-column ridge model with unpenalized intercept (Hoerl & Kennard, 1970):

$$\hat{\theta} = \arg\min_{\theta_0,\theta_g,\theta_c} \sum_i (\tilde{e}_i - \theta_0 - \theta_g z_{g,i} - \theta_c c_i)^2 + 10(\theta_g^2 + \theta_c^2). \tag{7}$$

Writing $\hat{m}(z_g, c) = \hat{\theta}_0 + \hat{\theta}_g z_g + \hat{\theta}_c c$, this is a ridge-regularized best-linear approximation to the conditional mean. It estimates redundancy in the evidence coordinate without changing the frozen predictor; it is not an unrestricted estimate of $\mathbb{E}[e \mid z_g, c]$. The implementation adds a 10^{-10} diagonal stabilizer when solving the normal equations. The corrected score is

$$r_i = \mathrm{clip}(\tilde{e}_i - \theta_0 - \theta_g z_{g,i} - \theta_c c_i, -6, 6), \qquad s_i = z_{g,i} + r_i. \tag{8}$$

The clipping bound is applied after the conditional prediction and before calibration, matching the archived inference routine; it is a bounded residual operation rather than an additional learned module. The residual is defined relative to the specified conditioning model: it represents the evidence left after the score and support explain their predictable component.

3.3 Conditional innovation principle

The method changes the target of fusion. A retrieval score is first decomposed into the part predictable from the score being augmented and its support, and only the remainder is allowed to modify that score. This gives a falsifiable distinction between redundancy and incremental evidence.

Proposition 1 (Ideal proper-scoring-rule consequence)**.** *Let Z denote the frozen score, C the support coordinate, and R a residual evidence coordinate. Under the Bayes-optimal log-loss predictors,*

$$\mathcal{L}^*(Z, C) - \mathcal{L}^*(Z, C, R) = I(Y; R \mid Z, C) \geq 0. \tag{9}$$

Consequently, a residual with no conditional label information cannot reduce the ideal log loss.

This proposition is a population statement, not a finite-sample guarantee for the ridge implementation. It also does not by itself imply a higher ROC-AUC. A ranking gain requires an additional score-scale condition, such as an additive-logit model $\mathrm{logit}\, P(Y = 1 \mid Z, C, R) = g(Z, C) + \beta R$ with $\beta > 0$. The experiments therefore test the principle with matched controls: the predictable component alone, conditional permutation of the residual, direct addition, supervised stacking, and a quadratic residualizer. The quadratic control tests whether additional conditional capacity is useful beyond this low-capacity target.

3.4 MONOTONE PROBABILITY MAP

For each task, the final output is a monotone probability map in the spirit of Platt scaling (Platt, 1999; Guo et al., 2017):

$$p_i = \mathrm{expit}(\kappa_t s_i + \delta_t), \qquad \kappa_t \geq 10^{-6}. \tag{10}$$

It changes probability scale while retaining the ordering of s whenever the fitted slope is positive (Zadrozny & Elkan, 2002; Guo et al., 2017). The lower bound is an explicit implementation constraint, rather than a post-hoc sign correction. The unconstrained fit uses logistic regression with $C = 1$, LBFGS, maximum 500 iterations, tolerance 10^{-4}, and an unpenalized intercept. If its slope falls below the lower bound, the slope is fixed at 10^{-6} and the intercept is recomputed by Brent's root solver so that predicted and observed positive-label counts balance. If fewer than eight labels are known or only one class is present, the implementation uses $\mathrm{sigmoid}(\mathrm{clip}(s, -60, 60))$ as a fallback. The slope constraint preserves the direction of s in exact arithmetic, but the residual stage can still alter the ranking relative to the original global probability.

3.5 DATA-USE SEPARATION

Training labels determine the fingerprint weights and retrieval banks. On the validation split, feature values alone determine normalization and the conditional coefficients; no validation label enters the residual fit. Validation labels enter the Platt fit only, after the residual scores have been formed. Test labels are not consumed by CSRE-Mol and are used only for the final metrics.

4 EXPERIMENTAL SETUP

4.1 EVALUATION SETTING

We evaluate four frozen predictors: GIN, GraphGPS, GraphMVP, and Uni-Mol. The benchmark contains BACE, BBBP, SIDER, Tox21, ClinTox, HIV, ToxCast, and MUV from MoleculeNet (Wu et al., 2018). Each backbone–dataset pair uses the existing development split and is repeated with seeds $0, 1, 2$. The repetitions quantify training variability on this split; they are not independent resplits. The split, model versions, fingerprint radius, neighbor count, ridge strength, support rule, and calibration settings were selected during development and then held fixed for the rows reported here. Accordingly, the headline numbers characterize the evaluated development protocol; an outer audit would test transfer beyond this grid.

4.2 INFERENCE AND DATA-USE PROTOCOL

The post-processor receives global probabilities and retrieval features derived from a labeled training bank. Training labels determine fingerprint weights and the retrieval banks. Validation features alone determine robust scaling and the task-wise ridge coefficients; the residual fit has no validation-label input. Validation labels enter only the subsequent Platt fit, after the residual score has been computed. Test labels are used only to compute the reported metrics. The backbone parameters never change. For each task, Morgan and MACCS class-contrast evidence is combined by the pre-specified agreement gate, the conditional residual is added to the standardized global logit, and a positive-slope Platt map produces the final probability.

4.3 COMPARATORS AND DIAGNOSTICS

The frozen global predictor is the external reference. The main diagnosis uses global-only calibration, direct probability/logit addition, a calibrated local-only predictor, the predictable component $\hat{m}(z_g, c)$, conditional-permuted residuals, and support-scale controls. Completed matched alternatives include a supervised linear stacker, weighted kNN, a cross-fitted residual protocol, and a quadratic conditional residualizer. The cross-fitted row re-estimates the same residual principle under fold separation; it is a protocol check rather than a new architecture. The quadratic row changes only the evidence residualizer and uses the same calibration endpoint. These completed comparisons characterize the role of conditional residualization relative to the matched alternatives evaluated here.

4.4 Ablations and sensitivity

View ablations remove Morgan or MACCS while keeping the residual and calibration stages fixed. Coordinate ablations condition the residualizer on the global score, support, or both. We also record calibration variants, validation-label fractions, neighbor-count sensitivity, and Morgan-radius sensitivity. These records are reported in the appendix so that the main text can reserve its tables for the central diagnosis and matched comparisons. All variants use the same 96 backbone–dataset–seed cells unless a table states otherwise.

4.5 Measures and aggregation

ROC-AUC and average precision measure ranking. Negative log-likelihood, Brier score, and expected calibration error measure probability quality; these quantities are reported separately because calibration cannot change the ordering of a score (Gneiting & Raftery, 2007; Brier, 1950; Murphy, 1973). For multi-task datasets, ranking metrics are macro-averaged over tasks with both classes present, while probability metrics average over tasks with at least one known label. Headline values are equal-weight means over the 96 backbone–dataset–seed cells; no pooled micro ranking metric is used. The appendix gives paired dataset-cluster intervals, seed-level values, task-support counts, and the exact aggregation records.

4.6 Reproducibility scope

The TeX project documents the post-processor and archived numerical summaries, but does not include the original molecular inputs, backbone checkpoints, or full prediction cache. The method can be executed using compatible arrays; the complete benchmark cannot be recreated from this manuscript directory alone. The sequential validation protocol separates label-free residual fitting from label-dependent calibration, providing a clear data-use interpretation for the reported mechanism study.

5 Results

5.1 Main comparison

Across the 32 backbone–dataset cells, each evaluated using three random seeds, CSRE-Mol raises the equal-weight mean ROC-AUC from 70.07% for the frozen global predictor to 75.97%. Average precision increases from 45.81% to 52.99%, while mean NLL decreases from 0.3933 for the global + Platt reference to 0.3625. The method improves ROC-AUC in 31 of 32 backbone–dataset cells. These values are averages over the evaluated grid; the 96 repeated runs should not be interpreted as independent observations. The paired cluster analysis and seed records are given in the appendix.

Table 1 shows the per-backbone and per-dataset ROC-AUC values. The gains occur across GIN, GraphGPS, GraphMVP, and Uni-Mol, although the magnitude varies by task. Uni-Mol/ToxCast is the explicit ranking failure and is retained in the table rather than hidden by the aggregate.

Table 1: Main ROC-AUC results. Each frozen backbone is shown as a bold group with its Baseline and +CSRE-Mol settings on the eight datasets. Values are means over three seeds, shown to one decimal place; bold marks the best setting for each dataset within a backbone.

Model / setting		Tox21	ToxCast	SIDER	MUV	BACE	BBBP	ClinTox	HIV
GIN	Baseline	75.0	62.4	60.1	75.0	66.7	67.3	57.3	74.7
	+CSRE-Mol	**77.7**	**64.6**	**66.4**	**77.7**	**83.7**	**74.3**	**76.0**	**79.5**
GraphGPS	Baseline	73.9	61.1	57.1	67.2	76.5	67.1	66.6	77.1
	+CSRE-Mol	**77.1**	**64.3**	**65.2**	**74.1**	**85.9**	**74.9**	**82.1**	**80.2**
GraphMVP	Baseline	74.9	62.8	61.4	74.2	79.1	68.8	63.9	76.8
	+CSRE-Mol	**77.4**	**64.4**	**66.8**	**75.9**	**87.3**	**74.5**	**78.4**	**80.0**
Uni-Mol	Baseline	76.9	**68.2**	62.7	75.8	80.9	67.0	84.0	79.2
	+CSRE-Mol	**78.2**	67.3	**67.9**	**78.2**	**86.8**	**74.0**	**89.1**	**81.3**

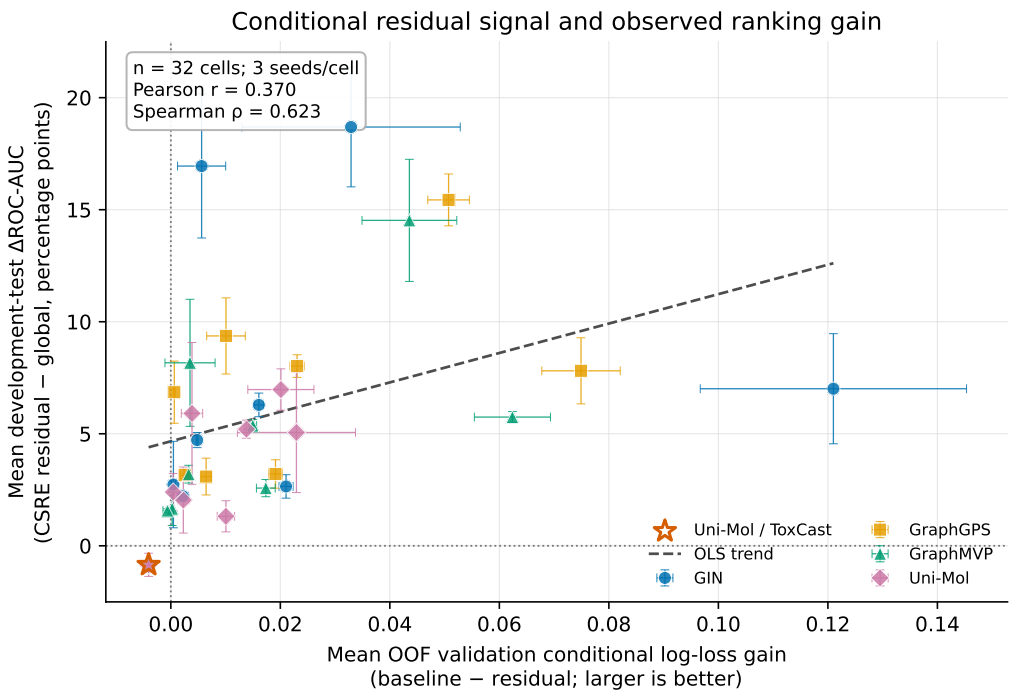


Figure 2: Conditional signal versus ranking gain on the development test split. The horizontal coordinate is the out-of-fold validation log-loss gain from adding the residual to (z_g, c); the vertical coordinate is the change in ROC-AUC on the development test split for CSRE-Mol relative to the frozen global predictor. Points are backbone–dataset means using three random seeds, with error bars showing seed standard deviations. The positive association is a mechanism diagnostic, not a claim of independent test data.

5.2 Conditional innovation diagnosis

The central test asks whether the useful retrieval signal is the part predictable from the score being augmented and neighborhood support. Table 2 compares the global score, direct evidence addition, the predictable component $\hat{m}(z_g, c)$, the conditional residual, and two controls. The predictable component alone is below the global reference (69.85% versus 70.06% ROC-AUC), whereas the residual reaches 75.97%. Conditional permutation reduces the mean to 70.48%, close to the global-only result. The same ordering is visible for AP and NLL. The standardized-support variant is close to the raw-support implementation, so the result is not explained by a particular support scale.

Table 2: Conditional-innovation diagnosis on the reused development split (96 cells). Here $\hat{m}$ is predictable evidence, $r = \tilde{e} - \hat{m}$, and permutation is within (z_g, c) strata. ROC-AUC/AP are percentages; $R^2 = 0.2554$, with the residual beating global in 93/96 cells and direct addition in 68/96.

Evidence object	$R^2(e \mid z_g, c)$	ROC-AUC ↑	AP ↑	NLL ↓	Brier ↓	ECE ↓
Global + Platt	–	70.0649	45.8133	0.3933	0.1245	0.0837
Direct evidence addition	–	75.3469	51.9716	0.3719	0.1146	0.0816
Predictable component $\hat{m}(z_g, c)$ only	0.2554	69.8486	45.0008	0.4012	0.1254	0.0863
Conditional-permuted residual	–	70.4836	46.8369	0.3856	0.1218	0.0800
Residual with standardized c	–	75.7705	52.9512	**0.3607**	**0.1127**	0.0804
Conditional residual r (CSRE-Mol)	–	**75.9728**	**52.9944**	0.3625	0.1131	**0.0801**

5.3 MATCHED FUSION ALTERNATIVES

Table 3 compares CSRE-Mol with completed controls using the same score and retrieval evidence. Direct addition and a supervised linear stacker reach 75.35% and 73.09% ROC-AUC; weighted kNN is a local-prediction control with poor NLL. Cross-fitted residualization matches the ranking but has slightly worse NLL. The quadratic residualizer reaches 74.86%, so the completed capacity control does not motivate a more flexible residual model. These comparisons support the conditional-innovation target relative to the matched alternatives evaluated in this study.

Table 3: Completed matched fusion and capacity controls on the same frozen predictions, evidence, split, calibration endpoint, and task-macro aggregation. The quadratic row predicts normalized evidence from (z_g, c) without labels in the residual fit; it is a capacity diagnostic.

Fusion method	ROC-AUC ↑	AP ↑	NLL ↓	Brier ↓	ECE ↓
Global + Platt	70.0649	45.8133	0.3933	0.1245	0.0837
Direct probability/logit addition	75.3469	51.9716	0.3719	0.1146	0.0816
Linear stacker $[z_g, e, c]$	73.0920	50.2127	0.3902	0.1205	0.0887
Weighted kNN	69.3529	49.6421	0.8039	0.1182	0.1017
Cross-fitted residual	75.9728	52.9944	0.3641	0.1141	0.0814
Quadratic conditional residualizer	74.8643	51.3043	0.3657	0.1144	0.0796
CSRE-Mol	**75.9728**	**52.9944**	**0.3625**	**0.1131**	**0.0801**

6 DISCUSSION

The result is a claim about the fusion target. Morgan, MACCS, Tanimoto similarity, ridge regression, and Platt scaling are established components; CSRE-Mol defines the auxiliary object as retrieval evidence after conditioning on score and support. The predictable-component, permutation, and quadratic controls behave as predicted, while Uni-Mol/ToxCast shows that calibration cannot create a missing ranking signal. This places CSRE-Mol between supervised stacking and partialling-out: it uses the latter's conditional target while retaining a simple inference-time correction.

The split and hyperparameters were selected during development, and validation molecules are reused sequentially: training labels build retrieval statistics, validation features fit the residualizer, and validation labels fit Platt scaling. The proposition in Section 3.3 describes the population scoring consequence, while the experiments characterize its finite-sample behavior across the evaluated grid. Transfer under distribution shift and richer conditional models are natural follow-up tests.

7 CONCLUSION

CSRE-Mol treats retrieval as conditional innovation: it removes the Morgan/MACCS evidence predictable from the frozen score and support, adds the residual, and calibrates the result. Direct addition, linear stacking, permutation, and the quadratic control support this target, while the Uni-Mol/ToxCast case clarifies when the principle is useful. The contribution is a reusable fusion target: retrieval becomes useful evidence only after the component already explained by the global score and support has been removed.

REFERENCES

Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U Rajendra Acharya, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76:243–297, 2021.

Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. Deep evidential regression. In *Advances in Neural Information Processing Systems*, volume 33, pp. 14927–14937, 2020.

Anastasios N Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023.

Dávid Bajusz, Anita Rácz, and Károly Héberger. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of Cheminformatics*, 7(1):20, 2015.

Leo Breiman. Stacked regressions. *Machine Learning*, 24:49–64, 1996. doi: 10.1007/BF00117832.

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Arie Dufour, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018. doi: 10.1111/ectj.12097.

Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. ChemBERTa: Large-scale self-supervised pretraining for molecular property prediction, 2020. arXiv:2010.09885. `https://arxiv.org/abs/2010.09885` (accessed 2026-07-11).

Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, 1967.

David R. Cox. The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, 20(2):215–232, 1958.

Noel A. C. Cressie. *Statistics for Spatial Data*. John Wiley & Sons, New York, revised edition, 1993.

Joseph L. Durant, David A. Leland, Douglas R. Henry, and James G. Nourse. Reoptimization of mdl keys for use in drug discovery. *Journal of Chemical Information and Computer Sciences*, 42(6): 1273–1280, 2002.

David Duvenaud, Dougal Maclaurin, Jorge Aguilera-Iparraguirre, Rafael Gomez-Bombarelli, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P. Adams. Convolutional networks on graphs for learning molecular fingerprints. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

Felix A Faber, Luke Hutchison, Bing Huang, Justin Gilmer, Samuel S Schoenholz, George E Dahl, Oriol Vinyals, Steven Kearnes, Patrick F Riley, and O Anatole von Lilienfeld. Prediction errors of molecular machine learning models lower than hybrid DFT error. *Journal of Chemical Theory and Computation*, 13(11):5255–5264, 2017.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. *Proceedings of The 33rd International Conference on Machine Learning*, 48:1050–1059, 2016.

Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1263–1272. PMLR, 2017.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

David E Graff, Eugene I Shakhnovich, and Connor W Coley. Accelerating high-throughput virtual screening through molecular pool-based active learning. *Chemical Science*, 12(22):7866–7881, 2021.

Cameron J Gruich, Varun Madhavan, Yixin Wang, and Bryan R Goldsmith. Clarifying trust of materials property predictions using neural networks with distribution-specific uncertainty quantification. *Machine Learning: Science and Technology*, 4(2):025019, 2023.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1321–1330, 2017.

Lior Hirschfeld, Kyle Swanson, Kevin Yang, Regina Barzilay, and Connor W Coley. Uncertainty quantification using neural networks for molecular property prediction. *Journal of Chemical Information and Modeling*, 60(8):3770–3780, 2020.

Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110(3):457–506, 2021.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2019.

Steven Kearnes, Kevin McCloskey, Marc Berndl, Vijay Pande, and Patrick Riley. Molecular graph convolutions: Moving beyond fingerprints. *Journal of Computer-Aided Molecular Design*, 30(8): 595–608, 2016.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Generalization through memorization: Nearest neighbor language models. In *International Conference on Learning Representations*, 2020.

Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 2796–2804, 2018.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Greg Landrum. RDKit: Open-source cheminformatics, 2006. `https://www.rdkit.org` (accessed 2026-07-11).

Shengchao Liu, Hongyu Wang, Weiyang Liu, et al. GraphMVP: Multi-view contrastive learning for graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 945–955, 2022.

Gerald M Maggiora. On outliers and activity cliffs—why QSAR often disappoints. *Journal of Chemical Information and Modeling*, 46(4):1535, 2006.

Georges Matheron. Principles of geostatistics. *Economic Geology*, 58(8):1246–1266, 1963.

Allan H. Murphy. A new vector partition of the probability score. *Journal of Applied Meteorology*, 12(4):595–600, 1973.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning into quantiles. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning*, pp. 625–632, 2005.

John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, pp. 61–74. MIT Press, 1999.

Liva Ralaivola, Sanjay J Swamidass, Hiroto Saigo, and Pierre Baldi. Graph kernels for chemical informatics. *Neural Networks*, 18(8):1093–1110, 2005.

Ladislav Rampásek, Mikhail Galkin, Vijay Prakash Dwivedi, Anh Tuan Luu, Thomas Wolf, and Dominique Beaini. Recipe for a general, powerful, scalable graph transformer. In *Advances in Neural Information Processing Systems*, volume 35, pp. 14524–14536, 2022.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA, 2006.

Daniel Reker and Gisbert Schneider. Active-learning strategies in computer-assisted drug discovery. *Drug Discovery Today*, 20(4):458–465, 2015.

David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 2010.

Yu Rong, Yatao Bian, Tingyang Xu, Weiyang Xie, Ying Wei, Wenbing Huang, and Junzhou Huang. Self-supervised graph transformer on large-scale molecular data. In *Advances in Neural Information Processing Systems*, volume 33, pp. 12559–12571, 2020.

Jerret Ross, Brian Belgodere, Vijil Chenthamarakshan, Inkit Padhi, Youssef Mroueh, and Payel Das. Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence*, 4(12):1256–1264, 2022.

Gabriele Scalia, Colin A Grambow, Barbara Pernici, Yi-Pei Li, and William H Green. Evaluating scalable uncertainty estimation methods for deep learning-based molecular property prediction. *Journal of Chemical Information and Modeling*, 60(6):2697–2717, 2020.

Kristof T Schütt, Huziel E Sauceda, Pieter-Jan Kindermans, Alexandre Tkatchenko, and Klaus-Robert Müller. SchNet – a deep learning architecture for molecules and materials. *The Journal of Chemical Physics*, 148(24):241722, 2018.

Burr Settles. Active learning literature survey. Technical Report Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009.

Ava P Soleimany, Alexander Amini, Samuel Goldman, Daniela Rus, Sangeeta N Bhatia, and Connor W Coley. Evidential deep learning for guided molecular property prediction and discovery. *ACS Central Science*, 7(8):1356–1367, 2021.

Michael L. Stein. *Interpolation of Spatial Data: Some Theory for Kriging*. Springer, New York, 1999.

Dagmar Stumpfe and Jürgen Bajorath. Exploring activity cliffs in medicinal chemistry. *Journal of Medicinal Chemistry*, 55(7):2932–2942, 2012.

Derek van Tilborg, Alisa Alenicheva, and Francesca Grisoni. Exposing the limitations of molecular machine learning with activity cliffs. *Journal of Chemical Information and Modeling*, 62(23): 5938–5951, 2022.

Peter Willett. Chemical similarity searching. *Journal of Chemical Information and Computer Sciences*, 38(6):983–996, 1998.

Andrew Gordon Wilson, Zhiting Hu, Ruslan Salakhutdinov, and Eric P Xing. Deep kernel learning. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pp. 370–378. PMLR, 2016.

David H. Wolpert. Stacked generalization. *Neural Networks*, 5(2):241–259, 1992. doi: 10.1016/S0893-6080(05)80023-1.

Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, et al. MoleculeNet: A benchmark for molecular machine learning. *Chemical Science*, 9(2):513–530, 2018.

Zhaoping Xiong, Dingyan Wang, Xiaohong Liu, et al. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *Journal of Medicinal Chemistry*, 63 (16):8749–8760, 2020.

Keyulu Xu, Weihuan Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2019.

Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, et al. Analyzing learned molecular representations for property prediction. *Journal of Chemical Information and Modeling*, 59(8): 3370–3388, 2019.

Chengxuan Ying, Tianle Cai, Shengding Luo, Shuxin Zheng, Guolin Ke, Di He, Yanming Shen, and Tie-Yan Liu. Do transformers really perform bad for graph representation? In *Advances in Neural Information Processing Systems*, volume 34, pp. 28877–28888, 2021.

Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 694–699, 2002.

Geng Zhou et al. Uni-Mol: A universal 3d molecular representation learning framework. In *International Conference on Learning Representations*, 2023.

A STATISTICAL REPORTING AND TASK SUPPORT

This appendix expands the main comparison with the seed-level records, the paired aggregate test, and the task-level label support used by the current CSRE-Mol version. The evaluation still uses the existing development split and seeds $0, 1, 2$; these tables quantify variability within that protocol and do not constitute an untouched external audit.

A.1 AGGREGATION AND UNCERTAINTY CONVENTIONS

For a multi-task dataset, ROC-AUC and average precision are first computed separately for each task and then averaged over tasks with at least one positive and one negative test label. NLL, Brier score, and ECE are averaged over tasks with at least one known test label. We call these task-macro metrics. The headline number is then an equal-weight average over the resulting 96 backbone–dataset–seed cells. We do not report a pooled micro ROC-AUC or micro average precision; consequently, the headline should not be interpreted as a molecule-weighted micro average.

The paired interval in Table A2 is a percentile bootstrap over eight dataset clusters. Within each cluster, the three seeds and four backbones remain paired and are averaged before resampling. The exact sign-flip test uses the same eight paired cluster summaries. This provides an overall test for the evaluated grid, while Table A5 exposes the seed variation in each cell. The cluster interval is not a separate confidence interval for each of the 32 cells.

Table A1 reports validation-side label support used to fit the post-processor, whereas Table A4 reports test-side support used for evaluation. These are different splits by design and are not expected to match.

Table A1: Validation-side label support and evaluation coverage. Each dataset is evaluated with four frozen backbones and three seeds on the existing development split. "Validation known labels" and "Validation positive labels" count the task entries in the validation split used to fit the post-processor; they are not test-side counts. Each dataset contributes 12 backbone–seed cells.

Dataset	Tasks	Validation known labels	Validation positive labels	Backbones	Seeds	Cells
BACE	1	151	130	4	3	12
BBBP	1	204	81	4	3	12
SIDER	27	3861	2285	4	3	12
TOX21	12	7144	687	4	3	12
CLINTOX	2	296	149	4	3	12
HIV	1	4113	81	4	3	12
TOXCAST	617	123683	12291	4	3	12
MUV	17	24478	52	4	3	12

Table A2: Paired aggregate statistics for Full CSRE-Mol versus the frozen Global-only reference. Full and Global-only are equal-weight means over 96 backbone–dataset–seed cells. The interval is a 95% percentile bootstrap over eight dataset clusters after averaging the three seeds and four backbones within each dataset; the backbones and seeds remain paired. The sign-flip test and Holm correction are computed within the five-metric global-comparison family. ROC-AUC and average precision are percentages; NLL, Brier, and ECE are on native scales. Differences use percentage points for the first two metrics and native units otherwise; negative differences indicate improvement for NLL, Brier, and ECE.

Metric	Full	Global-only	Full−Global	95% cluster CI	Sign-flip p	Holm p
ROC-AUC	75.9728	70.0654	+5.9073	[3.4755, 8.7339]	0.0078	0.0391
Average precision	52.9944	45.8147	+7.1797	[5.0024, 9.4262]	0.0078	0.0391
NLL	0.3625	0.5088	-0.1464	[-0.2752, -0.0365]	0.0391	0.1172
Brier	0.1131	0.1426	-0.0295	[-0.0571, -0.0056]	0.0625	0.1172
ECE	0.0801	0.1267	-0.0466	[-0.0854, -0.0125]	0.0469	0.1172

Table A3: Dataset-cluster view of the ROC-AUC improvement. Each row averages the four backbones after averaging the three seeds within one dataset. The two delta columns agree up to rounding; the last column is the independently stored cluster summary used by the bootstrap/sign-flip calculation. Wins/4 and Losses/4 count backbone-level means (not the 12 individual seed cells).

Dataset	Full−Global (pp)	Wins/4	Losses/4	Stored delta (pp)
Bace	10.098	4	0	10.098
BBBP	6.881	4	0	6.881
Sider	6.221	4	0	6.221
Tox21	2.437	4	0	2.437
ClinTox	13.427	4	0	13.427
HIV	3.265	4	0	3.265
ToxCast	1.524	3	1	1.524
MUV	3.406	4	0	3.406

Table A4: Task-level label support for the Full model. Counts are taken from the test-side diagnostic records (the label mask is identical across backbones and seeds). AUC-valid tasks have at least one positive and one negative test label. Positive-count summaries are minimum–median–maximum over tasks; "low-pos" counts tasks with 1–5 positives. These test-side counts are intentionally distinct from the validation-side counts in Table A1.

Dataset	Tasks	AUC-valid	Known labels (sum)	Test positives (min–median–max)	Zero-positive	Low-pos (1–5)	Known/task min	Known/task max
Bace	1	1	152	81–81.0–81	0	0	152	152
BBBP	1	1	204	108–108.0–108	0	0	204	204
Sider	27	27	3861	3–99.0–138	0	1	143	143
Tox21	12	12	7069	22–45.0–118	0	0	471	709
ClinTox	2	2	296	10–74.5–139	0	0	148	148
HIV	1	1	4113	130–130.0–130	0	0	4113	4113
ToxCast	617	610	141809	0–13.0–300	7	131	1	792
MUV	17	15	25342	0–3.0–7	2	11	1305	1709

A.2 Seed-level main results

Table A5 reports all three seed values for every backbone–dataset cell, together with the sample standard deviation and the paired difference from the same frozen Global-only predictor. The machine-readable archive preserves the full numerical precision; the printed table is rounded for readability.

B Mechanism and ablation records

The main text keeps the central diagnosis visible and moves the implementation audits here. These tables use the same development protocol unless their captions specify a sensitivity or validation-sample change. Only completed diagnostics are reported as result tables.

Table A6: Progressive construction of CSRE-Mol. All stages use the same 96 frozen backbone–dataset–seed cells and the same positive-slope Platt endpoint. The rows add the structural views, agreement gate, conditional residualization, and neighborhood-support coordinate in sequence. Entries are mean ROC-AUC (%).

Stage / setting	ROC-AUC (%; mean over four backbones and three seeds)							
	Tox21	ToxCast	SIDER	MUV	BACE	BBBP	ClinTox	HIV
B0: global + calibration	75.2	63.6	60.3	73.1	75.8	67.5	68.0	77.0
B1: + Morgan direct evidence	77.1	64.5	65.6	77.9	83.9	73.5	80.2	79.3
B2: + MACCS, no gate	77.8	65.4	66.1	76.5	83.8	73.6	79.3	79.6
B3: + agreement gate	77.8	65.3	66.1	76.5	83.6	73.7	80.0	79.7
B4: conditional residual on global score	77.4	65.0	66.5	76.4	85.9	74.4	81.7	79.7
B5: + neighborhood support (Full)	77.6	65.2	66.6	76.5	85.9	74.4	81.4	80.3

Table A5: Seed-level ROC-AUC records for all 32 backbone–dataset cells. Values are percentages; $s0$–$s2$ are the three repeated seeds, followed by the sample mean ± sample SD. The final column is the paired mean difference in percentage points. Values are rounded for typesetting; the exact machine-readable records are retained in the accompanying CSV/JSON archive.

Backbone	Dataset	Full CSRE-Mol				Global-only				Full−Global
		$s0$	$s1$	$s2$	mean ± SD	$s0$	$s1$	$s2$	mean ± SD	pp
gin	Bace	83.533	81.603	85.881	83.67 ± 2.14	62.893	66.215	71.066	66.72 ± 4.11	+16.95
graphgps	Bace	85.637	86.594	85.516	85.92 ± 0.59	74.978	79.151	75.517	76.55 ± 2.27	+9.37
graphmvp	Bace	87.376	86.455	88.089	87.31 ± 0.82	78.752	75.709	82.959	79.14 ± 3.64	+8.17
unimol	Bace	87.498	86.628	86.420	86.85 ± 0.57	83.446	77.065	82.307	80.94 ± 3.40	+5.91
gin	BBBP	74.277	74.489	74.238	74.33 ± 0.14	65.066	70.120	66.831	67.34 ± 2.57	+7.00
graphgps	BBBP	74.778	74.441	75.395	74.87 ± 0.48	67.303	65.017	68.866	67.06 ± 1.94	+7.81
graphmvp	BBBP	74.470	74.614	74.441	74.51 ± 0.09	68.702	69.126	68.461	68.76 ± 0.34	+5.75
unimol	BBBP	73.794	73.987	74.171	73.98 ± 0.19	65.885	67.038	68.109	67.01 ± 1.11	+6.97
gin	Sider	66.826	66.278	65.986	66.36 ± 0.43	60.800	60.334	59.090	60.07 ± 0.88	+6.29
graphgps	Sider	64.969	65.187	65.356	65.17 ± 0.19	57.086	56.606	57.754	57.15 ± 0.58	+8.02
graphmvp	Sider	67.097	66.807	66.481	66.80 ± 0.31	62.035	61.228	61.016	61.43 ± 0.54	+5.37
unimol	Sider	67.457	68.483	67.839	67.93 ± 0.52	62.258	63.681	62.230	62.72 ± 0.83	+5.20
gin	Tox21	77.802	77.331	77.829	77.65 ± 0.28	75.376	74.083	75.554	75.00 ± 0.80	+2.65
graphgps	Tox21	76.600	76.964	77.822	77.13 ± 0.63	72.898	73.541	75.332	73.92 ± 1.26	+3.21
graphmvp	Tox21	77.351	77.298	77.673	77.44 ± 0.20	74.860	75.058	74.677	74.86 ± 0.19	+2.58
unimol	Tox21	78.349	78.468	77.707	78.17 ± 0.41	77.794	76.989	75.789	76.86 ± 1.01	+1.32
gin	ClinTox	78.980	67.521	81.617	76.04 ± 7.49	59.232	46.853	65.965	57.35 ± 9.69	+18.69
graphgps	ClinTox	81.018	83.155	82.035	82.07 ± 1.07	64.366	68.809	66.721	66.63 ± 2.22	+15.44
graphmvp	ClinTox	85.179	75.612	74.334	78.37 ± 5.93	73.744	59.029	58.779	63.85 ± 8.57	+14.52
unimol	ClinTox	87.519	89.182	90.459	89.05 ± 1.47	79.694	86.709	85.584	84.00 ± 3.77	+5.06
gin	HIV	79.285	79.624	79.462	79.46 ± 0.17	74.302	74.792	75.123	74.74 ± 0.41	+4.72
graphgps	HIV	79.623	80.729	80.319	80.22 ± 0.56	76.370	78.519	76.470	77.12 ± 1.21	+3.10
graphmvp	HIV	80.050	80.381	79.660	80.03 ± 0.36	77.278	77.125	76.101	76.83 ± 0.64	+3.20
unimol	HIV	81.566	79.832	82.477	81.29 ± 1.34	80.729	76.148	80.871	79.25 ± 2.69	+2.04
gin	ToxCast	64.543	64.458	64.869	64.62 ± 0.22	62.397	62.146	62.693	62.41 ± 0.27	+2.21
graphgps	ToxCast	64.194	64.723	64.014	64.31 ± 0.37	61.045	61.824	60.567	61.15 ± 0.63	+3.16
graphmvp	ToxCast	64.625	63.969	64.529	64.37 ± 0.35	62.977	62.480	62.954	62.80 ± 0.28	+1.57
unimol	ToxCast	67.252	67.049	67.712	67.34 ± 0.34	68.191	67.343	69.029	68.19 ± 0.84	-0.85
gin	MUV	76.312	79.084	77.767	77.72 ± 1.39	75.663	76.003	73.301	74.99 ± 1.47	+2.73
graphgps	MUV	73.931	74.102	74.164	74.07 ± 0.12	66.821	68.757	66.093	67.22 ± 1.38	+6.84
graphmvp	MUV	75.870	75.445	76.346	75.89 ± 0.45	74.317	72.995	75.383	74.23 ± 1.20	+1.66
unimol	MUV	79.513	78.089	76.921	78.17 ± 1.30	77.741	76.021	73.579	75.78 ± 2.09	+2.39

Table A7: Component ablation. Rows add structural evidence, residual correction, agreement gating, probability calibration, and the positive-slope constraint to the frozen-backbone baseline. Values are mean ROC-AUC (%) over the four backbones; abbreviations are defined in the column headers.

Base model	Struct. evidence	Residual corr.	Agree. gate	Prob. calib.	Pos. slope	Tox21	ToxCast	SIDER	MUV	BACE	BBBP	ClinTox	HIV
Mean over four frozen backbones	×	×	×	×	–	75.2	63.6	60.3	73.1	75.8	67.5	68.0	77.0
	✓	×	✓	✓	✓	77.8	65.3	66.1	76.5	83.6	73.7	80.0	79.7
	✓	✓	✓	×	–	77.6	65.2	66.6	76.5	85.9	74.4	81.3	80.3
	✓	✓	✓	✓	×	77.6	64.2	66.1	76.5	85.9	74.4	81.4	80.3
	✓	✓	✓	✓	✓	77.6	65.2	66.6	76.5	85.9	74.4	81.4	80.3

Table A8: Evidence-view design comparison. Residual correction and monotone calibration are fixed; rows compare Morgan-only, MACCS-only, two-view fusion, and the agreement-gated CSRE-Mol evidence. Values are mean ROC-AUC (%) over four frozen backbones; – is not applicable.

Evidence components			ROC-AUC (%; mean over four backbones and three seeds)							
Morgan	MACCS	Agreement gate	Tox21	ToxCast	SIDER	MUV	BACE	BBBP	ClinTox	HIV
✓	×	–	76.6	64.2	66.0	77.9	86.1	73.7	81.1	80.2
×	✓	–	77.1	65.0	65.5	72.1	83.7	72.2	76.2	78.7
✓	✓	×	77.6	65.2	66.6	76.5	86.2	74.2	80.5	80.2
✓	✓	✓	77.6	65.2	66.6	76.5	85.9	74.4	81.4	80.3

Table A9: Matched ablation of conditional evidence residualization (M2). All rows use the same 96 frozen backbone–dataset–seed cells, the same two-view structural evidence, and the same positive-slope Platt endpoint whenever a structural score is present. The conditional ridge model is changed only in the indicated row. ROC-AUC and AP are percentages; NLL, Brier, and ECE use their native scales.

Setting	Conditional residualization			Metrics				
	Residual	Global cond.	Support cond.	ROC-AUC ↑	AP ↑	NLL ↓	Brier ↓	ECE ↓
Global + calibration	×	–	–	70.0649	45.8133	0.3933	0.1245	0.0837
Direct evidence addition	×	–	–	75.3469	51.9716	0.3719	0.1146	0.0816
Residual, global only	✓	✓	×	75.8513	52.7100	0.3636	0.1133	0.0802
Residual, support only	✓	×	✓	75.4242	52.1900	0.3702	0.1145	0.0815
Residual, global + support (Full)	✓	✓	✓	75.9728	52.9944	0.3625	0.1131	0.0801

Table A10: Validation-sample fraction ablation for Full CSRE-Mol. For each fraction, a fixed random subset of validation molecules is used to fit the robust global/local scales and task-wise ridge residualizer; the same subset's labels fit the positive-slope Platt map. Training retrieval banks and test features/labels are fixed, and all rows are evaluated on the same test records. The 10%, 25%, and 50% rows average five masks; 100% uses all validation molecules. ROC-AUC and AP are percentages; NLL, Brier, and ECE use their native scales. Because the residualizer is refit on a different feature subset, ROC-AUC/AP need not be identical; monotonic Platt preserves ranking only conditional on a fixed corrected score.

Validation fraction	Full CSRE-Mol metrics (mean over 96 cells)				
	ROC-AUC (%)	AP (%)	NLL	Brier	ECE
10%	75.9695	52.9364	0.593447	0.190657	0.199985
25%	75.9728	52.9944	0.429640	0.137867	0.117809
50%	75.9728	52.9944	0.382089	0.121233	0.091757
100%	75.9728	52.9944	0.362479	0.113087	0.080145

Table A11: Matched calibration ablation on the fixed Full structural score. All rows use the same 96 frozen backbone–dataset–seed cells and differ only in the final probability map. ROC-AUC and AP are percentages; NLL, Brier, and ECE are averaged on their native scales. The positive-slope endpoint has the best mean NLL and ranking, while the fixed-slope endpoint has slightly lower Brier and ECE.

Calibration variant	ROC-AUC (%)	AP (%)	NLL	Brier	ECE	Interpretation
Full positive-slope	**75.973**	**52.994**	**0.3625**	0.1131	0.0801	$\kappa_t \geq 10^{-6}$
Without calibration	75.968	52.898	0.8679	0.2890	0.3747	raw corrected score
Unconstrained slope	75.788	52.897	0.3627	0.1131	0.0804	negative slopes allowed
Fixed slope $\kappa_t = 1$	75.943	52.989	0.3643	**0.1100**	**0.0778**	intercept only

Capacity limitation. Beyond the completed quadratic residualizer reported in Table 2, broader nonlinear residualizer and supervised-combiner controls (e.g., shallow GBDT or interaction stacking) were not completed; these comparisons therefore remain open.

C Detailed sensitivity records

The following tables give backbone-level top-k and Morgan-radius sensitivity records, averaged over the three seeds.

Table A12: Top-k sensitivity by backbone (ROC-AUC percentages, averaged over three seeds). Bold marks the best value within each backbone.

Backbone	k	Tox21	ToxCast	Sider	MUV	Bace	BBBP	ClinTox	HIV
GIN	2	77.6	64.1	65.5	76.5	82.7	74.4	75.4	78.6
	4	**77.7**	64.4	65.9	77.2	83.4	**74.6**	75.8	79.1
	8	77.7	64.6	66.4	77.7	83.7	74.3	**76.0**	79.5
	16	77.4	**64.6**	66.6	**77.8**	82.3	74.1	74.5	**79.6**
	32	77.1	64.5	**66.7**	77.6	**83.9**	73.5	73.6	79.5
GraphGPS	2	76.9	63.7	64.4	72.7	85.3	74.9	**82.4**	79.1
	4	77.1	64.1	64.8	73.6	85.8	**75.2**	82.1	79.8
	8	**77.1**	64.3	65.2	**74.1**	**85.9**	74.9	82.1	80.2
	16	76.8	**64.4**	**65.4**	74.0	85.2	74.7	80.6	80.7
	32	76.5	64.2	65.3	73.9	85.6	74.2	79.9	**80.9**
GraphMVP	2	77.3	63.8	65.9	74.9	86.5	74.5	78.2	79.2
	4	**77.5**	64.1	66.3	75.5	87.1	**74.8**	78.4	79.7
	8	77.4	64.4	66.8	**75.9**	**87.3**	74.5	**78.4**	80.0
	16	77.2	**64.4**	67.1	75.9	86.7	74.2	77.0	**80.2**
	32	76.8	64.3	**67.2**	75.7	85.6	73.4	76.3	80.1
Uni-Mol	2	78.2	66.8	66.8	77.1	86.2	73.8	**89.9**	80.9
	4	**78.2**	67.0	67.4	77.8	**87.0**	**74.1**	89.2	81.1
	8	78.2	67.3	67.9	78.2	86.8	74.0	88.9	81.3
	16	77.8	**67.3**	68.1	78.3	86.6	73.9	88.4	81.5
	32	77.4	67.1	**68.3**	**78.3**	84.5	73.0	87.7	**81.7**

Table A13: Morgan-radius sensitivity by backbone (ROC-AUC percentages, averaged over three seeds). Bold marks the best radius within each backbone.

Backbone	r	Tox21	ToxCast	Sider	MUV	Bace	BBBP	ClinTox	HIV
GIN	1	**77.7**	**64.6**	66.4	**77.7**	83.7	74.3	**76.0**	**79.5**
	2	77.5	64.5	66.4	76.9	**84.4**	**74.5**	75.7	78.6
	3	77.0	64.3	**66.5**	76.8	83.3	74.3	74.2	76.9
GraphGPS	1	**77.1**	**64.3**	**65.2**	**74.1**	**85.9**	74.9	82.1	**80.2**
	2	76.9	64.2	64.9	73.0	85.8	**75.0**	**82.1**	79.5
	3	76.4	63.9	65.0	72.9	84.9	74.6	81.4	77.7
GraphMVP	1	**77.4**	**64.4**	66.8	**75.9**	**87.3**	**74.5**	**78.4**	**80.0**
	2	77.3	64.2	67.0	75.0	87.0	74.4	77.9	79.2
	3	76.7	64.0	**67.1**	74.9	86.2	74.2	77.0	77.6
Uni-Mol	1	78.2	**67.3**	67.9	**78.2**	**86.8**	74.0	**88.9**	**81.3**
	2	**78.2**	67.2	68.0	77.1	86.6	**74.1**	88.8	80.8
	3	77.6	67.1	**68.2**	76.4	85.5	73.7	88.6	80.0