PAIRANK: A MODEL-AGNOSTIC RELIABILITY PLUG-IN FOR PROTEIN–LIGAND BINDING AFFINITY PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Protein–ligand binding affinity predictors are widely used to support computational drug screening, but their performance can deteriorate on novel chemical scaffolds and other out-of-distribution (OOD) samples. Most predictors do not provide sample-level reliability information for identifying predictions with potentially large errors. Here, we introduce PAIRank (Protein–ligand Affinity Individual Reliability Ranking), a model-agnostic reliability plug-in for protein–ligand binding affinity prediction. PAIRank extracts joint protein–ligand embeddings from a frozen base affinity predictor, partitions the training embeddings into local regions and computes cluster-conditioned Mahalanobis distances from each evaluated sample to every region. Together, these distances describe how the sample is positioned relative to the training distribution. Pairwise supervision derived from base-model prediction errors then enables PAIRank to learn the relative error risk of different samples and assign a score for reliability ranking. These scores allow lower-risk predictions to be retained preferentially. We evaluated PAIRank under ligand-scaffold extrapolation on ChEMBL and on a temporally external BindingDB set using multiple affinity predictors and uncertainty-quantification methods. Experimental results showed that PAIRank reduced error in the retained subsets and exhibited competitive and relatively stable selective-retention performance across base models and data distributions. PAIRank requires neither modification nor retraining of the base predictor and it can serve as a lightweight post-processing plug-in for assessing the reliability of individual predictions from different affinity-prediction frameworks under distribution shift.

1 INTRODUCTION

Protein–ligand binding affinity prediction supports computational drug discovery by estimating binding strength for candidate comparison, prioritization and downstream analysis (Hughes et al., 2011; Bender & Cortes-Ciriano, 2021). Experimental measurements cannot cover the vast number of possible protein–ligand pairs, making computational prediction an important complement to experimental evaluation and virtual screening (Neves et al., 2018). By learning interaction patterns from protein sequences, ligand structures and joint protein–ligand information, these models have performed well across a range of affinity datasets.

Performance on random test splits, however, does not necessarily translate into reliable performance in practical applications (Wallach & Heifets, 2018; Guo et al., 2024). Most affinity predictors learn relationships among proteins, ligands and interaction patterns represented in the training data. They can interpolate when test samples resemble this distribution. Novel chemical scaffolds, protein families, binding modes, experimental sources or data collected at later time points instead require extrapolation to protein–ligand combinations that were poorly covered during training. The statistical associations learned during training may not hold in these out-of-distribution (OOD) regions, resulting in marked performance degradation (Ji et al., 2022).

Multiple factors can cause out-of-distribution (OOD) failures in protein–ligand affinity prediction. The estimated number of synthesizable or drug-like small molecules vastly exceeds the coverage of existing experimental databases, leaving novel candidate scaffolds with few similar training exam-

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.

ples (Reymond et al., 2012; Polishchuk et al., 2013). Protein families and their binding pockets also differ in sequence, structure and physicochemical properties. Interaction patterns learned from the training data may therefore not transfer to poorly represented protein–ligand combinations. The DrugOOD benchmark similarly documented marked degradation in affinity prediction across multiple OOD splits (Ji et al., 2022). For structure-based models, a single static structure cannot fully represent the conformational ensemble of a protein–ligand system, and conformational variation may affect the extraction of interaction information (Min et al., 2024). Differences among databases, assay conditions and label sources can also alter label distributions and noise patterns. Data leakage and redundancy in affinity benchmarks may further produce optimistic estimates of generalization performance (Graber et al., 2025).

Under these distribution shifts, affinity predictors still return a continuous point estimate for every protein–ligand pair, although some predictions may have large errors. Most affinity predictors provide no direct estimate of sample-level reliability, making high-error predictions difficult to distinguish from more reliable outputs. Distribution shift therefore creates two related challenges. It reduces aggregate predictive performance and leaves the reliability of individual predictions unclear, limiting the practical use of model outputs in drug screening.

Selective prediction assigns a reliability score to each protein–ligand prediction and allows a model to withhold high-risk outputs, retaining only predictions judged more reliable (Geifman & El-Yaniv, 2017). Ideally, stricter reliability thresholds retain fewer samples while reducing the average error of the retained subset. SelectiveNet jointly models prediction and sample selection while optimizing retained-sample risk at a target coverage (Geifman & El-Yaniv, 2019). Reliability-based filtering has also been examined in protein–ligand affinity analysis (Badkul et al., 2025).

Methods for estimating the reliability of individual protein–ligand predictions broadly include model-based uncertainty estimation, distributional-deviation analysis, calibration and error modelling (Mervin et al., 2021; Lanini et al., 2024; Rayka & Naghavi, 2025). Model-based methods estimate predictive variability by perturbing model parameters or the inference process. Monte Carlo dropout retains stochastic dropout during inference and obtains a predictive distribution from repeated forward passes. Deep ensembles instead train multiple affinity predictors independently and use predictive disagreement as an uncertainty signal (Gal & Ghahramani, 2016; Lakshminarayanan et al., 2017). These methods can capture variability arising from model parameters or training, but typically require repeated inference or multiple base models, increasing computational and storage costs.

Distribution-based methods assess how a protein–ligand combination relates to the training distribution using distances, neighbourhood structure, local density or anomaly scores (Lanini et al., 2024). In affinity prediction, this relationship can be described through joint protein–ligand embeddings produced by a base model, as demonstrated by eMOSAIC (Badkul et al., 2025). Class-conditional Mahalanobis distance in deep embedding spaces has been used for OOD detection, demonstrating that distances between internal embeddings can signal atypical inputs (Lee et al., 2018). Unlike Euclidean distance, Mahalanobis distance accounts for differences in scale and correlations among embedding dimensions, making it suitable for characterizing deviations in joint protein–ligand embeddings. Calibration methods aim to produce confidence scores or prediction intervals with explicit statistical interpretations. Conformal prediction uses nonconformity scores on a calibration set to construct intervals with finite-sample coverage guarantees under exchangeability (Vovk et al., 2005; Alvarsson et al., 2021). Marginal coverage, however, does not ensure that interval widths correctly rank the error risk of individual protein–ligand samples. Distribution shifts among training, calibration and test data can also weaken empirical uncertainty discrimination (Rayka & Naghavi, 2025). Error-modelling methods instead learn the prediction errors of the base affinity model. RIO uses a Gaussian process to relate neural-network inputs and outputs to errors, thereby correcting base predictions and producing uncertainty estimates (Qiu et al., 2020).

To address this challenge, we propose PAIRank, a model-agnostic plug-in for ranking the reliability of individual protein–ligand binding affinity predictions. PAIRank uses joint protein–ligand embeddings from a trained model to characterize the training distribution and treats the deviation of an evaluated sample from this distribution as a potential risk signal (Lee et al., 2018; Badkul et al., 2025). Because protein families, ligand scaffolds and interaction patterns may occupy distinct local regions, PAIRank does not rely on a single global distance. Instead, it partitions the joint protein–ligand training embedding space into clusters and computes each sample's Mahalanobis distance to

every cluster. The resulting distance features describe how the sample deviates from different local regions of the training distribution.

Rather than directly regressing these distance features onto the absolute error of each prediction, PAIRank constructs pairwise supervision from base-model prediction errors. It then learns how protein–ligand samples should be ordered by prediction-error risk, following the broader principle that loss ordering can be more directly aligned with sample-selection decisions than exact loss reconstruction (Yoo & Kweon, 2019; Li et al., 2020). Once trained, PAIRank assigns a risk score to each prediction and preferentially retains lower-risk outputs at a specified retention rate. PAIRank operates as an independent post-processing module. It changes neither the architecture nor the parameters of the original predictor and requires neither model ensembles nor repeated stochastic inference. We evaluated its individual reliability ranking and selective-retention performance using multiple affinity predictors and uncertainty-quantification methods under ligand-scaffold extrapolation and temporally external testing.

Our contributions are threefold. (1) we formulate individual reliability assessment in protein–ligand affinity prediction as an error-risk ranking problem for selective retention. This formulation supports ordering predictions by risk and preferentially retaining reliable samples without requiring precise estimates of individual error magnitude. (2) we develop a model-agnostic reliability plug-in that requires no modification or retraining of the base affinity predictor and does not rely on model ensembles or repeated stochastic inference. It can therefore be attached to affinity predictors with different architectures. (3) we systematically evaluate PAIRank under ChEMBL ligand-scaffold extrapolation and on a temporally external BindingDB set using multiple affinity predictors and uncertainty-quantification methods. Across most evaluated settings, PAIRank reduces error in retained subsets and maintains competitive selective-retention performance across base models and data distributions.

2 METHODS

We assess the reliability of individual predictions from a pretrained protein–ligand affinity predictor. Given a protein–ligand sample x_i and its experimental affinity label y_i, the base model f_θ outputs $\hat{y}_i = f_\theta(x_i)$. PAIRank produces a risk score s_i for each protein–ligand prediction without updating the base-model parameters. By convention, a smaller s_i indicates lower prediction-error risk. The score is not interpreted as an error estimate in affinity units. Instead, it orders predictions so that a lower-risk subset can be retained at a specified coverage. This formulation is consistent with selective prediction and risk–coverage analysis (Geifman & El-Yaniv, 2017; Traub et al., 2024).

Figure 1 summarizes the PAIRank workflow, which comprises joint protein–ligand embedding extraction, cluster-conditioned Mahalanobis distance construction and error-risk ranking. During training, the frozen base model generates affinity predictions and joint embeddings. PAIRank fits local clusters to the training embeddings, converts the Mahalanobis distances to all clusters into a feature vector and uses absolute prediction errors to construct pairwise supervision for the risk-ranking head. During inference, the same frozen base model and fixed cluster statistics are applied to each new sample. The trained ranking head then assigns a risk score for selective retention.

PAIRank uses the joint embedding produced by the base affinity predictor. For each input x_i, we extract the hidden vector after protein and ligand information has been fused and immediately before the final prediction layer: $z_i = \psi_\theta(x_i) \in \mathbb{R}^d$. The mapping ψ_θ may be obtained from affinity predictors based on sequences, molecular graphs, three-dimensional complexes or multiple modalities. PAIRank requires only access to a sample-level joint protein–ligand embedding before the prediction layer. The resulting joint embedding serves as the post-processing input to PAIRank. The method assumes no specific encoder architecture, input modality or training loss and can therefore be used with different affinity predictors. The base-model parameters θ remain frozen throughout PAIRank training.

PAIRank locates each sample relative to the training distribution by computing Mahalanobis distance features in the joint protein–ligand embedding space. Mahalanobis distance accounts for the scale and correlation of embedding dimensions. It has been used for anomaly detection, OOD detection and distance-based uncertainty estimation in deep embeddings, including cluster-conditioned residual modelling for protein–ligand affinity prediction (Mahalanobis, 1936; Lee et al., 2018; Rippel

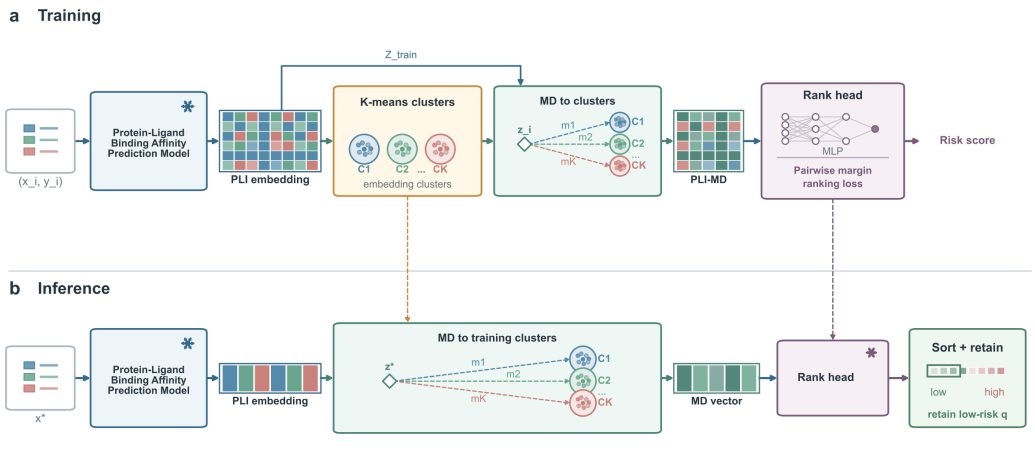

Figure 1: Overview of PAIRank. (a) During training, a frozen affinity predictor generates joint protein–ligand embeddings, which are partitioned by K-means and converted into cluster-conditioned Mahalanobis distance vectors for pairwise error-risk ranking. (b) During inference, the fitted clusters and trained ranking head are reused to score new samples and retain the lower-risk subset. Solid arrows denote forward computation, and dashed arrows denote components fixed and reused during inference.

et al., 2021; Badkul et al., 2025). Here, these distances quantify the deviation of a new protein–ligand sample from local regions represented by the training embeddings.

Joint protein–ligand embeddings may follow a multimodal distribution. Different protein families, ligand scaffolds or binding modes may occupy distinct local regions, making a single global distance potentially inadequate. PAIRank therefore normalizes the training embeddings before applying K-means clustering in the normalized joint protein–ligand embedding space. For each training cluster k, we estimate the centre μ_k and covariance matrix Σ_k. The Mahalanobis distance from a sample z to that cluster is $m_k(z) = \sqrt{(z-\mu_k)^\top \Sigma_k^{-1}(z-\mu_k)}, \quad k = 1, \ldots, K$. The resulting cluster-conditioned Mahalanobis distance vector is $m(z) = [m_1(z), m_2(z), \ldots, m_K(z)] \in \mathbb{R}^K$.

PAIRank maps the cluster-conditioned Mahalanobis features to a risk score for candidate retention. Training supervision is derived from the absolute prediction error of each training sample: $e_i = |y_i - \hat{y}_i|$. Rather than learning absolute error magnitudes directly, PAIRank focuses on relative error risk because selective retention at a specified coverage depends primarily on risk order rather than exact calibration of every error value.

PAIRank uses a lightweight multilayer perceptron g_ϕ as the ranking head. The head maps the Mahalanobis distance features to a scalar risk score: $s_i = g_\phi(m(z_i))$. During training, pairs of protein–ligand samples (i, j) are drawn from the training set, and their errors define the relative risk ordering. To reduce supervision noise from nearly equal errors, only pairs satisfying $|e_i - e_j| > \delta$ are retained. Their pairwise labels are $t_{ij} = \mathrm{sign}(e_i - e_j)$. When $t_{ij} = 1$, sample i has a larger prediction error than sample j, so the ranking head should assign $s_i > s_j$. The desired order is reversed when $t_{ij} = -1$. We train the model using a pairwise margin-ranking loss (Yoo & Kweon, 2019): $\mathcal{L}_{\mathrm{rank}} = \frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} \max\{0, \gamma - t_{ij}(s_i - s_j)\}$, where $\mathcal{P}$ is the retained set of training pairs, δ is the minimum error difference for pair inclusion and γ is the ranking margin. This objective encourages higher-error protein–ligand samples to receive higher risk scores, allowing the ranking head to order samples by error risk from the Mahalanobis distance features.

During inference, PAIRank applies the same post-processing sequence used during training. For a new protein–ligand sample, the frozen base model first returns an affinity prediction and a joint

embedding. The fixed normalizer, cluster centres and covariance matrices transform this embedding into cluster-conditioned Mahalanobis features. The trained ranking head then maps these features to a risk score. For a candidate batch, PAIRank sorts samples by increasing risk score and retains the lower-risk subset at a specified retention rate. Hereafter, coverage and retention rate both denote the proportion of samples retained. The retention rate can be selected according to prior experience, the experimental budget or the capacity for downstream validation.

3 Experiments and results

3.1 Datasets and scaffold splits

The complete ChEMBL31 (Mendez et al., 2019) and BindingDB (Gilson et al., 2016) filtering, overlap-control, temporal-split, sample-count and label-scale procedures are reported in Supplementary Tables S1 and S2. The main experiments use the resulting ChEMBL scaffold splits (Bemis & Murcko, 1996) and temporally external BindingDB set.

3.2 PAIRank as a plug-in across base affinity models

We applied PAIRank as a post-processing plug-in to four protein–ligand affinity predictors with different architectures: TrustAffinity, MGraphDTA, BACPI and MMSG-DTA (Badkul et al., 2025; Yang et al., 2022; Li et al., 2022; Xu et al., 2025). These models span joint protein–ligand embedding fusion, molecular-graph modelling, protein–ligand interaction modelling and protein-language-model-enhanced prediction. We repeated the evaluation across three scaffold splits generated with different random seeds. Within each split, we averaged results across 10 risk-ranking heads before aggregating them across splits. We fixed the retention rate at 15% to assess individual reliability under stringent candidate filtering while limiting variability at lower retention rates. For all four models, the retained subsets had lower MAE and RMSE and higher Pearson and Spearman correlations than the complete test sets.

As shown in Fig. 2a–d, retaining the lowest-risk 15% of samples on the ChEMBL scaffold test sets reduced MAE and RMSE for all four base models, while both correlation metrics increased. TrustAffinity achieved a retained MAE of 0.515 and RMSE of 0.705, reductions of 0.212 and 0.311 relative to the complete test set. Its Pearson and Spearman correlations increased to 0.840 and 0.839, respectively. Retained MAE values for MGraphDTA, BACPI and MMSG-DTA were 0.646, 0.685 and 0.584, corresponding to reductions of 0.094, 0.104 and 0.242. The Spearman correlation between risk score and absolute error was 0.086 for BACPI and 0.221 for MMSG-DTA. Thus, the error-risk signal captured by the Mahalanobis features varied among the joint protein–ligand embeddings supplied by different base models.

The corresponding results on the temporally external BindingDB set are reported in Supplementary Fig. S1. Retained error decreased for TrustAffinity, MGraphDTA and MMSG-DTA, whereas BACPI showed little benefit, indicating model-dependent transfer of the risk signal.

TrustAffinity had the strongest baseline performance among the four models on this dataset and was therefore selected for the activity-classification and screening analyses. As shown in Fig. 2e–i, PAIRank increased the area under the receiver operating characteristic curve (AUROC) from 0.888 to 0.939 at a $1\,\mathrm{nM}$ activity threshold. Because AUROC summarizes ranking across all positive–negative pairs without specifically emphasizing the top of a screening list, we also evaluated the Boltzmann-enhanced discrimination of receiver operating characteristic (BEDROC) metric (Truchon & Bayly, 2007). BEDROC assigns exponentially decreasing weights to active compounds appearing later in the ranking, and larger values of α place greater emphasis on early retrieval. At $\alpha = 20$, PAIRank increased BEDROC20 from 0.623 to 0.681. The PAIRank ROC curve lay above the TrustAffinity baseline in the low-false-positive-rate region (Fig. 2e). The cumulative-recall curve also showed that PAIRank retrieved more active samples when the same fraction of highly ranked candidates was inspected (Fig. 2f). At both the $0.1\,\mathrm{nM}$ and $1\,\mathrm{nM}$ thresholds, its enrichment-factor curves remained higher near the front of the ranking (Fig. 2g,h). These results show that PAIRank used individual risk information to concentrate active candidates nearer the top of the affinity-ranked list under a limited screening budget.

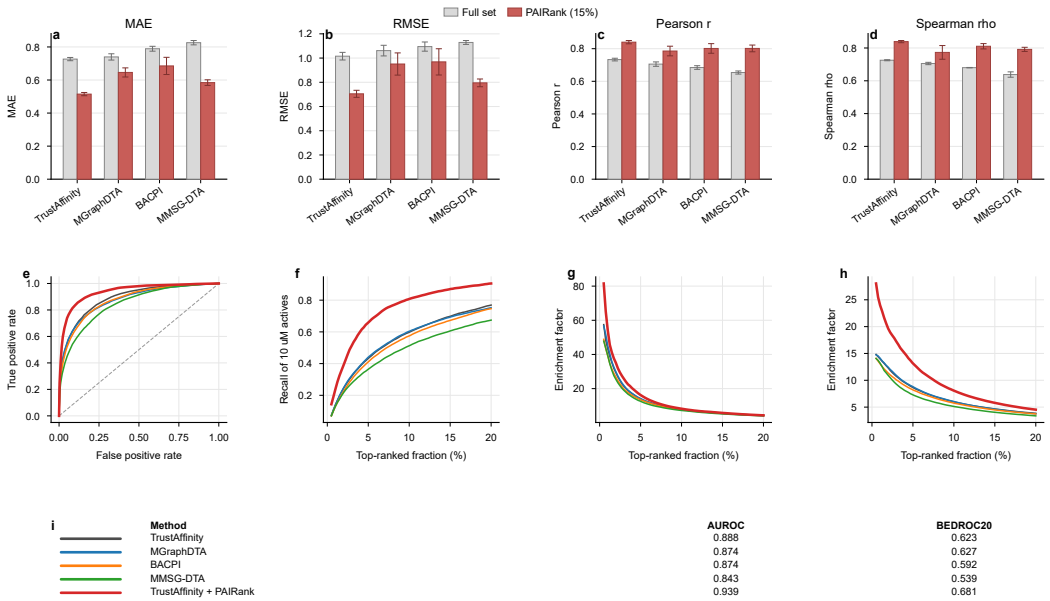

Figure 2: PAIRank selective-retention and screening performance on the ChEMBL ligand-scaffold extrapolation test sets. (a–d) MAE, RMSE, Pearson correlation and Spearman rank correlation for the complete test sets and the 15% lowest-risk subsets across four base models. (e) ROC curve at the 1 nM threshold. (f) Cumulative recall at 1 nM. (g,h) Enrichment-factor curves at 0.1 nM and 1 nM, respectively. (i) Method key with mean AUROC and BEDROC20. Bars and curves show means across three splits and 10 risk-ranking heads per split; error bars show standard deviations.

3.3 Comparison of four uncertainty methods

Using TrustAffinity as the base predictor, we compared four uncertainty methods: PAIRank, split conformal prediction (CP), RIO and MC dropout. The comparison assessed whether each method could rank protein–ligand samples by prediction-error risk and form lower-error subsets at specified retention rates. We also report R_S, the Spearman rank correlation between the risk score and the true absolute prediction error. A larger R_S indicates that high-error protein–ligand samples are placed nearer the high-risk end of the ordering. Retained MAE measures the mean absolute error of the retained subset, whereas retained Spearman measures agreement between predicted and measured affinity rankings within that subset. AUSE-MAE is the area between the MAE curve obtained by removing samples according to estimated risk and the oracle curve obtained by removing samples according to true error. A lower AUSE-MAE indicates a risk ordering closer to the oracle. Together, these quantities provide task-aligned, multi-metric evaluation of risk ordering across retention rates (Lanini et al., 2024; Traub et al., 2024; Rayka & Naghavi, 2025). Figure 3 brings these four quantities together to assess retention-rate behaviour, retained-subset performance and risk-order quality.

PAIRank showed generally stable retention behaviour on both datasets (Fig. 3). On ChEMBL, its retained MAE and retained Spearman curves changed smoothly as the retention rate decreased. An informative risk ordering should progressively lower retained MAE and generally increase retained Spearman as higher-risk predictions are removed. The PAIRank curves broadly followed this pattern. PAIRank achieved a ChEMBL AUSE-MAE of 0.291, lower than those of CP, RIO and MC dropout. Its R_S was 0.198, close to that of CP (0.199), indicating similar agreement between the estimated-risk and absolute-error rankings. Retained Spearman also remained high as higher-risk predictions were removed.

On the external BindingDB set, all methods were affected to different degrees by the larger distribution shift. PAIRank achieved an AUSE-MAE of 0.615, higher than those of RIO (0.464) and MC dropout (0.571). However, it maintained a comparatively high retained Spearman correlation at low retention rates, and its retention curves remained generally stable. A higher retained Spearman indi-

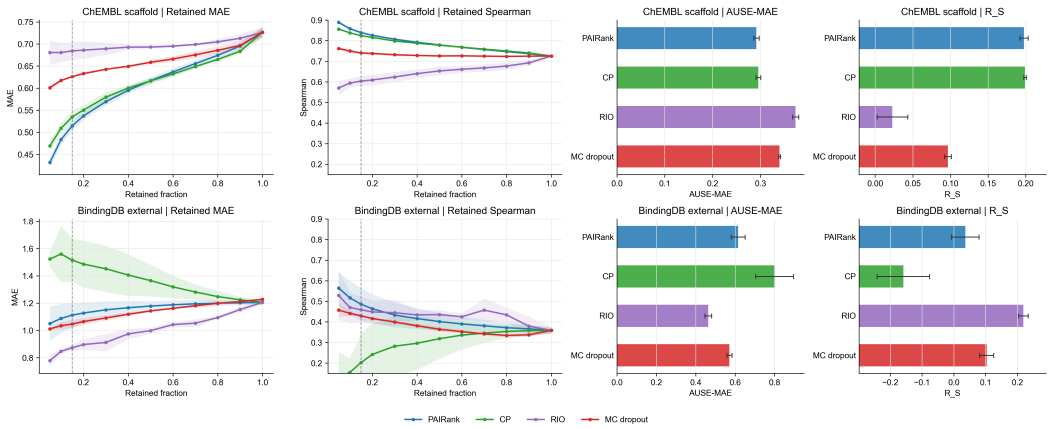

Figure 3: Selective-retention and risk-ranking comparison of four uncertainty methods on TrustAffinity. The upper row shows the ChEMBL ligand-scaffold extrapolation test sets, and the lower row shows the temporally external BindingDB set. From left to right, the panels report retained MAE, retained Spearman correlation, AUSE-MAE and R_S. Vertical dashed lines mark 15% retention. Curves and bar heights denote means; shading and error bars denote standard deviations.

cates closer agreement between predicted and measured affinity rankings after risk-based filtering, supporting candidate prioritization within the retained set.

PAIRank showed relatively consistent retention behaviour across the two datasets. Its risk ordering progressively reduced retained error and achieved the lowest AUSE-MAE on ChEMBL. On BindingDB, its low-retention subsets maintained comparatively high retained Spearman correlations and stable curves. The benefits of CP, RIO and MC dropout were more concentrated in particular metrics or datasets. These findings support the relative consistency of PAIRank's risk stratification across the evaluated distributions rather than universal superiority over all uncertainty methods.

Implementation-choice sensitivity was assessed under a fixed ChEMBL protocol by varying the local partitioning scheme and risk-ranking head; the complete analysis protocol and numerical results are provided in Supplementary Tables S4 and S5.

3.4 Pairwise ranking versus regression-based error modelling

We next examined the contribution of the ranking objective in PAIRank. Using the same base models and Mahalanobis-distance features, we compared the pairwise risk head with an eMOSAIC-style regression risk head trained to predict continuous scores from absolute prediction errors (Badkul et al., 2025). Retention curves on ChEMBL and BindingDB were used to compare error control and affinity-ranking consistency across retention rates.

On ChEMBL, pairwise ranking showed a consistent advantage across the four base models. Along the retention curves, it generally yielded lower MAE and RMSE than regression and higher Pearson and Spearman correlations at every retention rate below full retention. The differences widened as the retention rate decreased and were most pronounced for the ranking metrics in the lowest-retention region. Its clearest benefit emerged under the most stringent retention setting, where distinguishing lower-risk protein–ligand predictions was particularly important (Fig. 4).

The corresponding comparison on BindingDB is reported in Supplementary Fig. S3. Pairwise ranking showed model-dependent effects on MAE and RMSE but generally higher Pearson and Spearman correlations across most retention rates.

3.5 Retrospective enrichment of high-error candidates and an individual case study

We retrospectively assessed whether pairwise risk scores identified larger observed errors among high-affinity protein–ligand candidates removed by the pairwise gate. At 15% coverage and a 0.1%

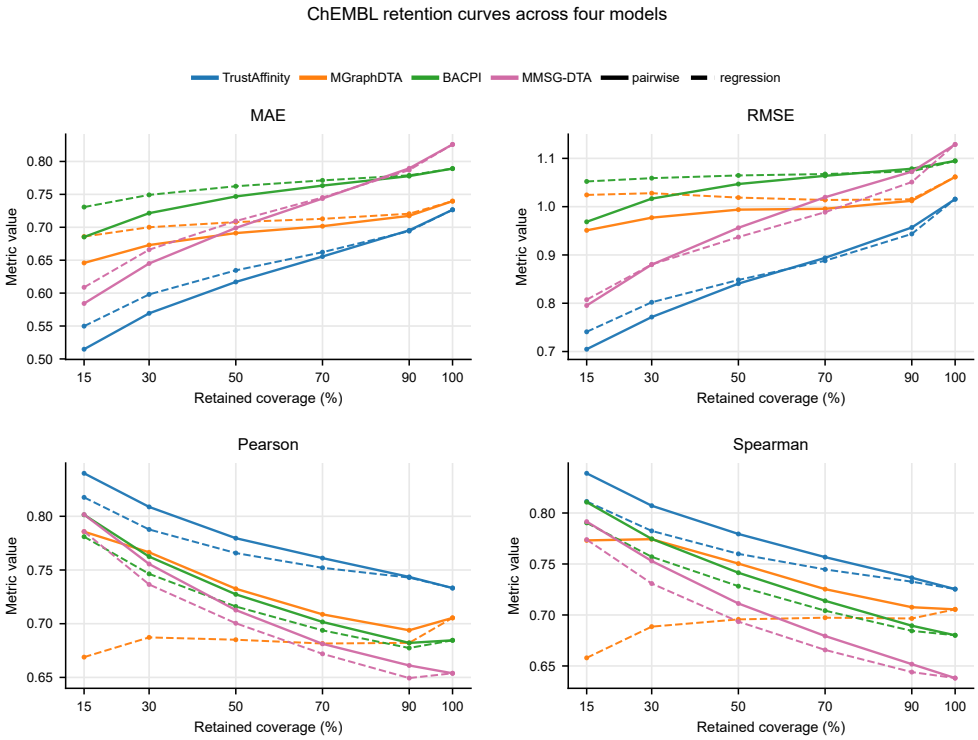

Figure 4: Retention curves for pairwise ranking and regression-based error modelling on the ChEMBL ligand-scaffold extrapolation test sets. Colours distinguish the four base affinity models; solid lines denote pairwise ranking and dashed lines denote error regression. The four panels show MAE, RMSE, Pearson correlation and Spearman rank correlation. Curves represent averages across three data splits and 10 risk-ranking heads per split.

screening budget, we constructed a top 0.1% affinity shortlist for each target–seed run before applying the gate. Across seeds 0, 11 and 42, the gate removed 150 candidates (51, 48 and 51, respectively). We defined the 20% with the largest absolute errors as the high-error set and calculated cumulative recall after ranking candidates by decreasing pairwise risk.

The resulting high-error set contained 30 candidates, and its cumulative-recall curve remained generally above the random baseline (Fig. 5). The enrichment AUROC was 0.618, indicating that higher pairwise risk was associated with larger observed errors among removed high-affinity candidates. Pairwise gating therefore provided sample-level reliability information beyond affinity ranking and helped identify high-affinity candidates with elevated error risk.

For target `P34972|PF00001` under seed 0, affinity-only ranking selected candidate `147355`. Its predicted and measured model-scale affinities were -0.350 and -3.926, yielding an absolute error of 3.576. Its pairwise risk percentile was 0.967, placing it near the high-risk end of the candidate pool. Pairwise gating instead selected candidate `149324`, with predicted and measured values of -0.485 and -0.146, an absolute error of 0.339 and a risk percentile of 0.093. The measured affinities differed by 3.780 units, and the replacement's absolute error was 3.237 lower. Thus, the gate excluded a candidate favoured by affinity-only ranking but associated with a large observed error. Together, the enrichment and case analyses support using affinity as the primary ranking signal and pairwise risk as a secondary reliability filter for candidates with potentially high prediction errors.

4 Discussion

PAIRank is intended to complement rather than replace a base affinity predictor. As a post-processing plug-in, it provides a relative reliability ranking after the base model has produced point

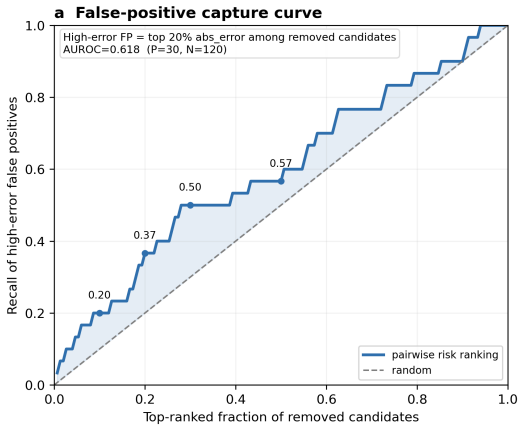

Figure 5: Retrospective enrichment of the 30 high-error candidates among 150 candidates removed by the pairwise gate. The horizontal axis shows the inspected fraction after ranking by decreasing pairwise risk, and the vertical axis shows cumulative high-error recall. The dashed line denotes random ranking; enrichment AUROC was 0.618.

estimates for individual protein–ligand pairs. Across multiple base models, PAIRank combined cluster-conditioned Mahalanobis features derived from joint protein–ligand embeddings with error-risk ranking. It thereby identified retained subsets with generally lower error and more stable affinity ordering. This objective more closely reflects virtual-screening decisions than reducing average error across the complete test set alone. Because experimental budgets often permit validation of only a small set of candidates with high predicted affinity, reliability information is valuable for identifying elevated error risk within these shortlists.

The advantage of pairwise ranking over regression-based error modelling should be interpreted within this task boundary. For candidate retention, risk–coverage analysis and prioritization of lower-risk candidates, relative ordering by error risk is better aligned with the decision objective than fitting the absolute error of every protein–ligand sample. Pairwise ranking generally performed better on ChEMBL and retained an advantage in most BindingDB settings, but it was not superior for every model or external condition.

The method has two clear limitations. First, PAIRank produces a relative risk score rather than a calibrated error in affinity units or a prediction interval. Interpretation on an error scale requires additional calibration and revalidation on external data. Second, PAIRank depends on the joint protein–ligand embedding supplied by the base model. Its benefit will be limited if this embedding does not distinguish distributional deviations associated with prediction error.

Future work should test risk-ranking stability on target-cold, protein-family-cold and additional temporal datasets. It should also examine how risk-score calibration relates to experimental budgets, allowing retention rates to correspond to more explicit error tolerances. Combining PAIRank with calibration or prediction-interval methods may provide interpretable uncertainty estimates alongside candidate rankings.

5 Conclusion

We developed PAIRank to provide an additional reliability ranking for individual protein–ligand affinity predictions without modifying the underlying affinity predictor. Across ligand-scaffold extrapolation on ChEMBL and temporally external testing on BindingDB, PAIRank preferentially retained lower-error predictions for most base models and showed competitive selective-retention performance. The advantage of pairwise risk ranking over regression-based error modelling was most evident under stringent, low-retention settings. This pattern is consistent with the alignment between relative risk modelling and candidate-retention decisions. PAIRank offers a lightweight, model-agnostic reliability plug-in for existing affinity predictors and can serve as an auxiliary tool for candidate retention and individual reliability ranking in virtual screening.

REFERENCES

Jonathan Alvarsson, Ulf Norinder, Ola Spjuth, Lars Carlsson, and Scott Boyer. Predicting with confidence: using conformal prediction in drug discovery. *Journal of Pharmaceutical Sciences*, 110(1):42–49, 2021. doi: 10.1016/j.xphs.2020.09.055.

Amitesh Badkul, Li Xie, Shuo Zhang, and Lei Xie. Multimodal out-of-distribution individual uncertainty quantification enhances binding affinity prediction for polypharmacology. *Nature Machine Intelligence*, 7(12):1985–1995, 2025. doi: 10.1038/s42256-025-01151-2.

Guy W. Bemis and Mark A. Murcko. The properties of known drugs. 1. molecular frameworks. *Journal of Medicinal Chemistry*, 39(15):2887–2893, 1996. doi: 10.1021/jm9602928.

Andreas Bender and Isidro Cortes-Ciriano. Artificial intelligence in drug discovery: what is realistic, what are illusions? part 1: Ways to make an impact, and why we are not there yet. *Drug Discovery Today*, 26(2):511–524, 2021. doi: 10.1016/j.drudis.2020.12.009.

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

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems Workshop / arXiv preprint arXiv:1705.08500*, 2017.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2151–2159. PMLR, 2019.

Michael K. Gilson, Tiqing Liu, Michael Baitaluk, George Nicola, Linda Hwang, and Jenny Chong. Bindingdb in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Research*, 44(D1):D1045–D1053, 2016. doi: 10.1093/nar/gkv1072.

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.

Jeff Guo et al. Scaffold splits overestimate virtual screening performance. In *Machine Learning in Structural Biology*, pp. 73–88. Springer, 2024. doi: 10.1007/978-3-031-72359-9_5.

John P. Hughes, Sian Rees, S. B. Kalindjian, and Karen L. Philpott. Principles of early drug discovery. *British Journal of Pharmacology*, 162(6):1239–1249, 2011. doi: 10.1111/j.1476-5381.2010.01127.x.

Yuanfeng Ji, Lu Zhang, Jiaxiang Wu, Bingzhe Wu, Long-Kai Huang, Tingyang Xu, Yu Rong, Lanqing Li, Jie Ren, Ding Xue, Houtim Lai, Shaoyong Xu, Jing Feng, Wei Liu, Ping Luo, Shuigeng Zhou, Junzhou Huang, Peilin Zhao, and Yatao Bian. DrugOOD: Out-of-distribution (OOD) dataset curator and benchmark for AI-aided drug discovery – a focus on affinity prediction problems with noise annotations. *arXiv preprint arXiv:2201.09637*, 2022. doi: 10.48550/arXiv.2201.09637.

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.

Jessica Lanini, Minh Tam Davide Huynh, Gaetano Scebba, Nadine Schneider, and Raquel Rodríguez-Pérez. UNIQUE: A framework for uncertainty quantification benchmarking. *Journal of Chemical Information and Modeling*, 64(22):8379–8386, 2024. doi: 10.1021/acs.jcim.4c01578.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Min Li, Zhangli Lu, Yifan Wu, and Yaohang Li. BACPI: a bi-directional attention neural network for compound–protein interaction and binding affinity prediction. *Bioinformatics*, 38(7):1995–2002, 2022. doi: 10.1093/bioinformatics/btac035.

Minghan Li, Xialei Liu, Joost van de Weijer, and Bogdan Raducanu. Learning to rank for active learning: A listwise approach. *arXiv preprint arXiv:2008.00078*, 2020.

Prasanta Chandra Mahalanobis. On the generalized distance in statistics. *Sankhya*, 2(1):49–55, 1936.

David Mendez, Anna Gaulton, A. Patrícia Bento, Jon Chambers, Marleen De Veij, Eloy Félix, María Paula Magariños, Juan F. Mosquera, Prudence Mutowo, Michał Nowotka, María Gordillo-Marañón, Fiona Hunter, Luisa Junco, Grace Mugumbate, Milagros Rodriguez-Lopez, Francis Atkinson, Nicolas Bosc, Chris J. Radoux, Aldo Segura-Cabrera, Anne Hersey, and Andrew R. Leach. ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Research*, 47(D1): D930–D940, 2019. doi: 10.1093/nar/gky1075.

Lewis H. Mervin, Simon Johansson, Elizaveta Semenova, Katie A. Giblin, and Ola Engkvist. Uncertainty quantification in drug design. *Drug Discovery Today*, 26(2):474–489, 2021. doi: 10.1016/j.drudis.2020.08.021.

Yaosen Min, Ye Wei, Peizhuo Wang, Xiaoting Wang, Li Han, Nian Wu, Stefan Bauer, Shuxin Zheng, Yu Shi, Yingheng Wang, Ji Wu, Dan Zhao, and Jianyang Zeng. From static to dynamic structures: Improving binding affinity prediction with graph-based deep learning. *Advanced Science*, 11(40): e2405404, 2024. doi: 10.1002/advs.202405404.

Bruno J. Neves, Rodolpho C. Braga, Cleber C. Melo-Filho, Jeremias T. Moreira-Filho, Eugene N. Muratov, and Carolina Horta Andrade. QSAR-based virtual screening: Advances and applications in drug discovery. *Frontiers in Pharmacology*, 9:1275, 2018. doi: 10.3389/fphar.2018.01275.

Pavel G. Polishchuk, Timur I. Madzhidov, and Alexandre Varnek. Estimation of the size of drug-like chemical space based on GDB-17 data. *Journal of Computer-Aided Molecular Design*, 27 (8):675–679, 2013. doi: 10.1007/s10822-013-9672-4.

Xin Qiu, Elliot Meyerson, and Risto Miikkulainen. Quantifying point-prediction uncertainty in neural networks via residual estimation with an I/O kernel. In *International Conference on Learning Representations*, 2020. arXiv:1906.00588.

Milad Rayka and Seyed Shahab Naghavi. Uncertainty quantification enables reliable deep learning for protein–ligand binding affinity prediction. *Scientific Reports*, 2025. doi: 10.1038/s41598-025-27167-7.

Jean-Louis Reymond, Lars Ruddigkeit, Lorenz Blum, and Ruud van Deursen. Exploring chemical space for drug discovery using the chemical universe database. *ACS Chemical Neuroscience*, 3 (9):649–657, 2012. doi: 10.1021/cn3000422.

Oliver Rippel, Patrick Mertens, and Dorit Merhof. Modeling the distribution of normal data in pre-trained deep features for anomaly detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 672–680, 2021.

Jeremias Traub, Till J. Bungert, Carsten T. Lüth, Michael Baumgartner, Klaus H. Maier-Hein, Lena Maier-Hein, and Paul F. Jäger. Overcoming common flaws in the evaluation of selective classification systems. In *Advances in Neural Information Processing Systems*, volume 37, 2024. arXiv:2407.01032.

Jean-François Truchon and Christopher I. Bayly. Evaluating virtual screening methods: Good and bad metrics for the "early recognition" problem. *Journal of Chemical Information and Modeling*, 47(2):488–508, 2007. doi: 10.1021/ci600426e.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005. doi: 10.1007/b106715.

Izhar Wallach and Abraham Heifets. Most ligand-based classification benchmarks reward memorization rather than generalization. *Journal of Chemical Information and Modeling*, 58(5):916–932, 2018. doi: 10.1021/acs.jcim.7b00403.

Jiahao Xu, Lei Ci, Bo Zhu, Guanhua Zhang, Linhua Jiang, Shixin Ye-Lehmann, and Wei Long. MMSG-DTA: A multimodal, multiscale model based on sequence and graph modalities for drug-target affinity prediction. *Journal of Chemical Information and Modeling*, 65(2), 2025. doi: 10.1021/acs.jcim.4c01828.

Ziduo Yang, Weihe Zhong, Lu Zhao, and Calvin Yu-Chian Chen. MGraphDTA: deep multiscale graph neural network for explainable drug–target binding affinity prediction. *Chemical Science*, 13(3):816–833, 2022. doi: 10.1039/D1SC05180F.

Donggeun Yoo and In So Kweon. Learning loss for active learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 93–102, 2019.