MOLE-PAIR: IMPROVING RELIABILITY OF MOLECULAR FOUNDATION MODELS WITH PREFERENCE OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Molecular foundation models pre-trained on large chemical corpora have become powerful tools for scientific discovery, yet their behavior can be unpredictable when inputs deviate from the pre-training distribution, which severely hinders deployment in safety-critical settings such as biomedical discovery and materials design. To mitigate this issue, we propose ***Mole**cular **P**reference-**A**ligned **I**nstance **R**anking* (Mole-PAIR), a plug-and-play module that can be attached to existing foundation models and improve their reliability w.r.t. out-of-distribution (OOD) data through cost-effective post-hoc preference optimization on frozen representations. The key idea of Mole-PAIR is to formulate OOD detection as a preference optimization problem whose objective aims at optimizing the model's "preferences" on in-distribution (ID) samples over OOD ones. Our analysis shows that this new objective adaptively prioritizes hard pairs throughout the training dynamics, which facilitates fast learning in practice. We demonstrate the practical efficacy of this approach with recently proposed molecular foundation models and challenging OOD benchmarks, achieving up to 9.7% improvement of AUROC and 48.5% reduction of FPR95 over powerful OOD detection approaches under shifts of scaffold, assay and molecular size.

1 INTRODUCTION

Artificial intelligence has enabled major advances in fields such as drug discovery (David et al., 2020), protein engineering (Wang et al., 2023b), and materials design (Sanchez-Lengeling et al., 2017). Among these critical domains, molecules represent a fundamental modality, forming the basic units through which design, optimization, and discovery are carried out. Foundation models, pre-trained on large chemical datasets containing massive numbers of diverse molecules, have shown strong potential to accelerate molecular design by predicting physicochemical and biological properties with improved accuracy (Beaini et al., 2023; Méndez-Lucio et al., 2024; Luo et al., 2023). Despite this promise, a central barrier to their deployment in industrial pipelines is the reliability of their predictions (Jiang et al., 2024; Wang et al., 2024). Without robust confidence estimation, these models may produce misleading outputs.

A key manifestation of this issue is *chemical hallucination*, analogous to the hallucinations observed in modern large language models (Xu et al., 2025). This problem is particularly severe for out-of-distribution (OOD) molecules (Liang et al., 2017), which deviate significantly from the training distribution. In such cases, models often generate incorrect predictions with high confidence, such as falsely assigning strong bioactivity to an inactive or toxic compound. These failures can have substantial consequences: confidently mispredicting the activity of a novel scaffold may lead to wasted investment in synthesis and testing (Ramsundar et al., 2017), while overlooking an activity cliff—where a small structural change causes a sharp loss of activity (Maggiora, 2006)—can misdirect optimization efforts in drug design (van Tilborg et al., 2022). However, despite extensive research on OOD detection for images (Hendrycks & Gimpel, 2016; Liang et al., 2017; Liu et al., 2020; Lee et al., 2018; Sun et al., 2022) and graph data (Li et al., 2022; Wu et al., 2024; Bao et al., 2024), how to equip large pre-trained molecular foundation models with the ability to recognize what they do not know remains largely unexplored.

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.

To fill this gap, we propose ***Mole**cular **P**reference-**A**ligned **I**nstance **R**anking* (Mole-PAIR), a simple yet effective post-hoc module for OOD detection with molecular foundation models. Mole-PAIR formulates OOD detection as a preference optimization problem: a lightweight detector head built on top of the frozen foundation model's embeddings is trained with a pairwise objective to "prefer" ID molecules over OOD ones.[1] This design is plug-and-play and model-agnostic, requiring neither model logits nor downstream labels, and can be attached to arbitrary off-the-shelf molecular foundation models at negligible cost.

To demonstrate its practical efficacy, we apply this approach to recently proposed molecular foundation models, MiniMol (Kläser et al., 2024) and Uni-Mol (Zhou et al., 2023; Lu et al., 2024), comparing it against powerful OOD detection methods across five datasets spanning two challenging OOD benchmarks (DrugOOD (Ji et al., 2022) and GOOD (Gui et al., 2022)). While competitor methods require additional downstream labels and training task-specific classifier heads—which our approach avoids entirely—Mole-PAIR achieves substantial performance gains, delivering up to a 9.7% improvement in AUROC and a 48.5% reduction in FPR95 across scaffold, assay, and molecular size shifts. **Our contributions are:**

(i) We propose Mole-PAIR, which formulates out-of-distribution detection as a preference optimization problem by optimizing the model's relative preference for ID over OOD samples. This yields a simple yet effective post-hoc objective that serves as a plug-and-play, model-agnostic framework to enhance the OOD detection capabilities of molecular foundation models.

(ii) We evaluate our method on prominent molecular foundation models (MiniMol and Uni-Mol) and challenging OOD benchmarks (DrugOOD and GOOD) involving diverse distribution shifts. The results consistently demonstrate that Mole-PAIR outperforms existing OOD detection baselines, improving AUROC by up to 9.7% and reducing FPR95 by up to 48.5%.

(iii) To provide theoretical and empirical justification for our method, we analyze the training dynamics and reveal that the proposed objective naturally and adaptively prioritizes hard pairs that are most informative for preference optimization, even in the absence of explicit guidance.

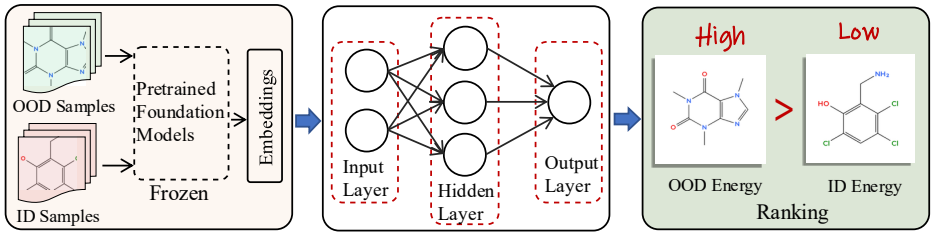
2 Related Work

General OOD Detection Out-of-distribution (OOD) detection has attracted widespread attention due to the growing need for reliable AI systems. Early approaches, such as One-Class SVM (Schölkopf et al., 1999), LOF (Breunig et al., 2000), and Isolation Forest (Liu et al., 2008), operate under Euclidean feature spaces and the independent and identically distributed (i.i.d.) assumption. In the deep learning era, popular post-hoc OOD methods include Maximum Softmax Probability (MSP) (Hendrycks & Gimpel, 2016), ODIN (Liang et al., 2017) via temperature scaling and input perturbation, Mahalanobis distance based on class-conditional Gaussian features (Lee et al., 2018), and Deep KNN (Sun et al., 2022), a nonparametric nearest-neighbor detector in deep feature space. However, these methods are largely tied to Euclidean embeddings and a standard softmax classifier head, and they are predominantly evaluated on vision benchmarks—limitations that hinder their transferability to chemistry-constrained molecular graphs and regression tasks.

OOD Detection for Molecules Because many real-world domains involve non-Euclidean relational data, graph-based models are widely used, sparking diverse graph OOD detection methods (Li et al., 2022; Guo et al., 2023). These approaches often leverage energy-based scoring (Wu et al., 2023), topological property analysis (Bao et al., 2024), or synthesis-based OOD generation (Wang et al., 2025). Nevertheless, because they are designed for general graphs, they frequently lack domain-specific chemical and molecular insights. Conversely, methods developed specifically for molecular OOD detection tend to be tightly coupled with specific architectures and training pipelines, such as reconstruction-similarity using molecular diffusion models (Shen et al., 2024). This tight coupling limits their adaptability to diverse off-the-shelf molecular foundation models, highlighting the need for a more flexible framework.

AI Reliability A key direction for improving AI reliability involves equipping pre-trained models with mechanisms to manage predictive uncertainty without requiring costly retraining. Existing post-hoc methods, such as Conformal Prediction and Selective Prediction, primarily focus on

[1] Training and test-time OOD molecules are drawn from different distributions (Section 3.1).



(a) Foundation Backbone (b) Lightweight Scoring MLP (c) Preference Learning

Figure 1: **Overview of Mole-PAIR.** (a) A frozen pre-trained molecular foundation model encodes ID molecules and auxiliary OOD molecules into embeddings. (b) A lightweight scoring MLP E_ϕ maps each embedding to a scalar OOD affinity and is the only trainable component. (c) The head is trained post hoc with the pairwise preference objective (Eqs. 6 and 7). At test time, $E_\phi(S)$ alone serves as the OOD indicator, without downstream labels or classifier logits.

uncertainty management for in-distribution (ID) data (Lei et al., 2018; Romano et al., 2020; Lin et al., 2023; Rong et al., 2020; Wang et al., 2024). Conformal prediction calibrates model scores to generate prediction sets that contain the true label with a user-specified frequency, providing distribution-free coverage guarantees for ID samples (Angelopoulos et al., 2022; Laghuvarapu et al., 2023; Arvidsson McShane et al., 2024). Selective prediction introduces a reject option that allows models to abstain on low-confidence inputs, thereby controlling error rates on accepted predictions (Geifman & El-Yaniv, 2017; 2019; Guo et al., 2017). Although these techniques can occasionally react to OOD inputs, their primary objective is risk control on ID data. In contrast, OOD detection explicitly aims to identify and flag samples drawn from a distribution distinct from the training data.

3 Methodology

In this section, we introduce ***Mole**cular **P**reference-**A**ligned **I**nstance **R**anking* (Mole-PAIR), a post-hoc preference optimization approach designed to enhance the OOD detection capability of molecular foundation models. Figure 1 illustrates the overall workflow of our model.

3.1 Out-of-Distribution Detection for Molecules

Problem Formulation. We consider an input molecule represented by its SMILES string $S \in \mathcal{S}$. Without loss of generality, we utilize MiniMol (2D) (Kläser et al., 2024) and Uni-Mol (3D) (Zhou et al., 2023; Lu et al., 2024) as representative model backbones to generate molecular embeddings. MiniMol converts the SMILES string into a 2D molecular graph representing atomic connectivity; this graph, complete with atom and bond features, is subsequently processed by a GNN to produce a final embedding capturing the topological structure of the molecule. In contrast, Uni-Mol first generates the molecule's 3D spatial structure from the SMILES string and employs a transformer-based architecture on these 3D coordinates to learn a representation encoding geometric shape and spatial properties.

For the data setup, we assume access to an in-distribution (ID) training dataset $D_{\text{in}}^{\text{train}}$ containing molecules with specific bioactive properties of interest. To support pairwise preference learning, we also leverage a set of auxiliary OOD samples denoted as $D_{\text{out}}^{\text{train}}$. Importantly, we do not assume prior access to the specific OOD distributions encountered at test time. In practice, $D_{\text{out}}^{\text{train}}$ can be constructed in a standard and realistic manner from readily available resources: it can either be sampled prior to deployment from large, generic public molecular databases (e.g., ZINC, PubChem, ChEMBL) to represent a broad chemical background (i.e., "what is not ID"), or be obtained by holding out a subset from the available validation pool as background molecules. In all cases, this auxiliary set is collected without targeting any particular test-time OOD category, reflecting real-world drug discovery settings where labeled bioactivity data is scarce while unlabeled molecular structures are abundant.

Formally, the training set is given by $D^{\text{train}} = D_{\text{in}}^{\text{train}} \cup D_{\text{out}}^{\text{train}}$, whereas at test time, the model is evaluated on $D^{\text{test}} = D_{\text{in}}^{\text{test}} \cup D_{\text{out}}^{\text{test}}$, where $D_{\text{out}}^{\text{test}}$ represents novel distribution shifts (e.g., novel scaffolds or assays) that are disjoint from the auxiliary training OOD samples, as standard in com-

mon molecular OOD benchmarks. We do not perform any sampling, splitting, or tuning using $D_{\text{out}}^{\text{test}}$, which is reserved strictly for final evaluation.

For the OOD detection task, we aim to design a detector g that determines whether a given molecule S is in-distribution (ID) or out-of-distribution (OOD):

$$g(S; \tau, E_\phi) = \begin{cases} 1 \ \text{(ID)}, & \text{if } E_\phi(S) < \tau, \\ 0 \ \text{(OOD)}, & \text{if } E_\phi(S) \geq \tau, \end{cases} \tag{1}$$

where $E_\phi(\cdot)$ is a learnable detector and τ is a decision threshold. The score $E_\phi(S)$ indicates the estimated OOD affinity of sample S. An ideal detector satisfies $E_\phi(S_{\text{in}}) < E_\phi(S_{\text{out}})$ for all $S_{\text{in}} \sim D_{\text{in}}$ and $S_{\text{out}} \sim D_{\text{out}}$.

OOD Detector for Foundation Models. Without loss of generality, we consider a molecular foundation model f_{Encoder} that maps any input molecule S to an embedding vector in a latent space:

$$h \ = \ f_{\text{Encoder}}(S) \in \mathbb{R}^p. \tag{2}$$

We construct the scoring function E_ϕ by combining a frozen, pre-trained f_{Encoder} with a trainable head g_{Head}:

$$E_\phi(S) \ = \ g_{\text{Head}}\big(f_{\text{Encoder}}(S); \phi\big), \tag{3}$$

where the trainable parameters ϕ comprise exclusively those of g_{Head}, leaving the parameters of f_{Encoder} frozen. This plug-and-play design confines training to the OOD scoring function (detector) and is compatible with any off-the-shelf molecular foundation model.

3.2 Preference Optimization Learning Objective

Achieving the condition $E_\phi(S_{\text{in}}) < E_\phi(S_{\text{out}})$ is non-trivial because it involves the pairwise ranking between arbitrary ID and OOD samples. A quantitative metric measuring how closely the model's ranking aligns with the ideal case is the Area Under the Receiver Operating Characteristic Curve (AUROC), which equals the probability that a randomly chosen ID sample is ranked before an OOD sample:

$$\text{AUROC} \ = \ \Pr\big(E_\phi(S_{\text{in}}) < E_\phi(S_{\text{out}})\big), \tag{4}$$

where $S_{\text{in}} \sim D_{\text{in}}^{\text{test}}$ and $S_{\text{out}} \sim D_{\text{out}}^{\text{test}}$. Consequently, the optimal learning objective for OOD detection should maximize the AUROC derived from the model's estimated OOD affinities.

Preference Optimization for OOD Detection. Directly optimizing AUROC is intractable due to its non-differentiability. To address this, we draw inspiration from direct preference optimization (DPO) (Rafailov et al., 2024), which has proven highly successful in post-training modern large language models (LLMs) (Wang et al., 2023a; Tunstall et al., 2023; Hong et al., 2024; Meng et al., 2024). The core principle of DPO is to learn directly from pairwise preferences via a simple logistic likelihood, bypassing the need to train a separate reward model or employ on-policy reinforcement learning.

We extend this principle to post-hoc preference optimization for OOD detection with molecular foundation models. Specifically, we replace the LLM policy score with the estimated OOD affinity $E_\phi(S)$ produced by the detector head on top of a frozen encoder, and optimize solely using pairwise ranking labels. Following the Bradley–Terry model (Bradley & Terry, 1952), we formulate a pairwise learning objective that enforces the target preference wherein ID samples are preferred over OOD ones:

$$P(S_{\text{in}} \succ S_{\text{out}}) \ = \ \sigma\big(\beta \cdot [E_\phi(S_{\text{out}}) - E_\phi(S_{\text{in}})]\big), \tag{5}$$

where $\sigma(u) = (1 + e^{-u})^{-1}$ is the logistic sigmoid function that maps the detector margin $\Delta E_\phi = E_\phi(S_{\text{out}}) - E_\phi(S_{\text{in}})$ to a probability in $[0, 1]$. A larger margin drives this probability closer to 1, signaling higher confidence that the ID sample is ranked ahead of the OOD sample. The temperature parameter $\beta > 0$ controls the sensitivity of the probability to the margin: a larger β increases the model's reaction to small differences, whereas a smaller β produces more gradual training signals. Maximizing the log-likelihood over pairs yields the **Mole-PAIR loss**:

$$\mathcal{L}_{\text{Mole-PAIR}}(\phi) = -\,\mathbb{E}_{S_{\text{in}} \sim D_{\text{in}},\, S_{\text{out}} \sim D_{\text{out}}} \left[\log \sigma\big(\beta \Delta E_\phi\big)\right]. \tag{6}$$

In practice, we construct balanced mini-batches of ID and OOD samples and compute margins within each batch to estimate the expectation efficiently. This focuses learning on relative ranking rather than absolute scores, aligning directly with AUROC optimization, as analyzed in Section 5.

Table 1: **Comparison of OOD detection performance and requirement for different components.** Left: AUROC and FPR95 averaged over the four shifts (assay shifts for EC50 and IC50, scaffold shifts for HIV and PCBA). Complete results with each shift type are presented in Appendix D. Right: whether a model family needs task labels (Labels), an auxiliary OOD set (Aux.), a task-specific classifier head (Head), or classifier logits (Logits). Best in **boldface**, second best underlined. All improvements are statistically significant through Wilcoxon signed-rank test under confidence level $p < 0.05$.

Method	MiniMol		Uni-Mol		Requirements			
	AUROC (↑)	FPR95 (↓)	AUROC (↑)	FPR95 (↓)	Labels	Aux.	Head	Logits
MSP	0.515	0.915	0.509	0.953				
ODIN	0.516	0.917	0.498	0.941	✓	×	✓	✓
Energy	0.514	0.916	0.513	0.951				
Mahalanobis	0.539	0.936	0.572	0.903				
LOF	0.543	0.936	0.559	0.930	✓	×	✓	×
KNN	0.523	0.935	0.577	0.917				
MSP-OE	0.728	0.790	0.691	0.829	✓	✓	✓	✓
ODIN-OE	0.721	0.778	0.682	0.820				
Mole-PAIR (Ours)	**0.776**	**0.664**	**0.725**	**0.753**	×	✓	×	×

Regularization. Because AUROC depends exclusively on score ordering and the loss function in Eq. 6 relies solely on margins, adding a constant offset to the detector output leaves the ranking invariant. Thus, the training objective is invariant to global translations $E_\phi(S) \mapsto E_\phi(S) + c$, which can otherwise induce scale drift and numerical instability (Grathwohl et al., 2020). To fix this gauge and stabilize the scale without affecting the optimal ranking, we introduce a lightweight ℓ_2 penalty on the detector outputs:

$$\mathcal{L}_{\text{total}}(\phi) = \mathcal{L}_{\text{Mole-PAIR}}(\phi) + \lambda \left(\mathbb{E}[E_\phi(S_{\text{in}})^2] + \mathbb{E}[E_\phi(S_{\text{out}})^2] \right), \tag{7}$$

where $\lambda > 0$. This penalty anchors the mean OOD affinity near zero and discourages unnecessarily large magnitudes, preventing loss saturation when the magnitude of $\beta \Delta E_\phi$ becomes excessively large in either direction. At test time, we use $E_\phi(S)$ as the OOD affinity and evaluate the model using threshold-free metrics, including AUROC on $(D_{\text{in}}^{\text{test}}, D_{\text{out}}^{\text{test}})$.

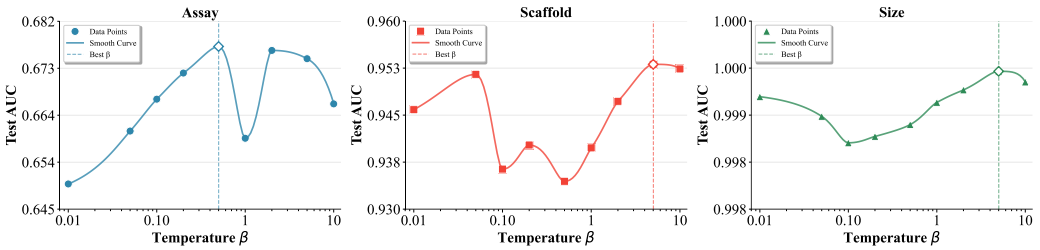
4 Experiments

Mole-PAIR can be applied as a unified framework for OOD detection across diverse tasks. In this section, we present empirical evidence to validate its effectiveness. Our primary goal is to demonstrate the practical efficacy of our approach in enhancing the OOD detection capabilities of existing molecular foundation models, rather than superseding state-of-the-art methods tailored for specialized molecular OOD detection tasks.

Datasets. We adopt molecular datasets from two recently proposed benchmarks, DrugOOD (Ji et al., 2022) and GOOD (Gui et al., 2022). To cover a spectrum of natural molecular distribution shifts, we evaluate performance across three properties used to partition ID and OOD data: assay type, molecular size, and molecular scaffold. For GOOD, we primarily consider covariate shifts. We strictly follow the official splits of both benchmarks, where OOD partitions comprise domains entirely absent from the ID training data. Training OOD samples are drawn from the OOD-validation split, whereas test OOD samples come from the OOD-test split, ensuring they remain strictly unseen during training. Further details regarding datasets and splits are provided in Appendix C.

Evaluation Metrics. Following standard practice, we evaluate OOD detection performance using: (1) the Area Under the Receiver Operating Characteristic Curve (AUROC), (2) the Area Under the Precision-Recall Curve (AUPR), and (3) the False Positive Rate at 95% True Positive Rate (FPR95).

Baselines. We compare Mole-PAIR against OOD detection methods that are similarly applicable to arbitrary encoder backbones, including MSP (Hendrycks & Gimpel, 2016), ODIN (Liang et al., 2017), Energy (Liu et al., 2020), Mahalanobis (Lee et al., 2018), LOF (Breunig et al., 2000), and KNN (Sun et al., 2022). For a fair comparison, all methods leverage identical, frozen embeddings



(a) Assay (b) Scaffold (c) Size

Figure 2: **Test AUROC on EC50 as a function of the temperature β ($\lambda = 0.01$).**

from the same pre-trained molecular foundation models (MiniMol (Kläser et al., 2024) and Uni-Mol (Zhou et al., 2023; Lu et al., 2024)). In accordance with their original designs, these baseline methods train a task-specific classifier using downstream molecular property labels, deriving OOD scores from the resulting classifier logits. In contrast, Mole-PAIR trains a label-free scoring head without requiring downstream labels. Additionally, we compare against two outlier exposure variants, MSP-OE and ODIN-OE (Hendrycks et al., 2019), which incorporate training OOD samples to fine-tune their scoring heads. Table 1 summarizes the component requirements for each method, with implementation details deferred to Appendices C.2 and C.3.

4.1 Overall Performance

Table 1 reports the out-of-distribution detection performance of Mole-PAIR alongside standard post-hoc detectors (MSP, ODIN, Energy, and Mahalanobis) and their outlier exposure counterparts (complete results for each shift type across all datasets are deferred to Appendix D.1). Traditional confidence-based methods (MSP and ODIN) exhibit unstable performance under complex distribution shifts, yielding near-random AUROC scores on challenging subsets because foundation model representations remain uncalibrated for OOD separation without dedicated alignment. Notably, Mole-PAIR achieves competitive or superior performance across the majority of evaluated benchmarks, attaining average AUROC scores of 0.776 on MiniMol and 0.725 on Uni-Mol while substantially reducing false positive rates (FPR95). While size-based distribution shifts yield high separability across most methods due to distinct atomic-count signatures, Mole-PAIR demonstrates consistent improvements on structurally intricate distribution shifts such as scaffold and assay splits. Additional comparisons and discussions are provided in Appendix D.

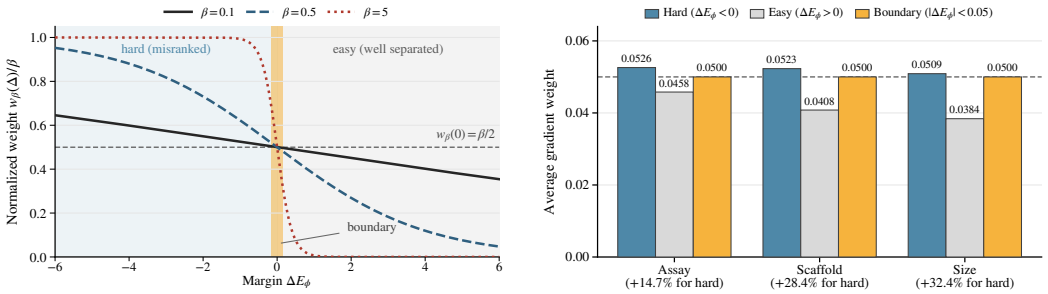
4.2 Ablation Study and Hyper-parameter Analysis

The effect of temperature parameter β. We evaluate the impact of the temperature parameter β in Eq. 6 across values ranging from 0.01 to 10.0 with a fixed $\lambda = 0.01$ (Figure 2). Scaffold shifts exhibit high sensitivity to β, peaking non-monotonically at $\beta = 5.0$ (AUROC ≈ 0.953), whereas Assay shifts show moderate sensitivity, peaking at $\beta = 0.5$. Size shifts remain near-optimal across the entire range due to clear geometric separation. This indicates that sharper preference signals (β) benefit complex structural shifts, whereas smoother penalties suffice for simpler tasks.

The influence of regularization parameter λ. We further analyze the ℓ_2 regularization penalty λ with a fixed $\beta = 0.1$ (Appendix D.2, Figure 5). Assay shifts perform best under weak regularization, degrading as the penalty increases. Conversely, Scaffold shifts peak at $\lambda = 0.05$ before declining sharply, demonstrating that a modest penalty helps regularize harder topological boundaries. Size shifts remain robust across moderate values and degrade only under excessive regularization.

4.3 Further Discussions

Data and compute efficiency. Table 7 illustrates the effect of reducing the training OOD set on DrugOOD-EC50 (MiniMol) from an ID-to-OOD ratio of 1:1 down to 1:0.01 (i.e., from 2,000 down to 20 auxiliary molecules), keeping the ID set and all validation and test splits fixed. AUROC degrades gracefully, moving from 0.970 to 0.849 on Scaffold, from 1.000 to 0.974 on Size, and



(a) Predicted weight function (b) Measured weights on EC50 ($\beta = 0.1$)

Figure 3: **Hard-pair emphasis predicted by Proposition 5.1 and observed during training.** (a) The per-pair gradient weight $w_\beta(\Delta) = \beta\,\sigma(-\beta\Delta)$, normalized by β, as a function of the margin: misranked pairs receive the highest weight, boundary pairs sit at $\beta/2$, and well-separated pairs are nearly ignored, with larger β concentrating emphasis more sharply near the boundary. (b) Average per-pair gradient weight measured across the three EC50 shifts: hard pairs receive the largest weights and easy pairs the smallest, while boundary pairs align with the predicted level $w_\beta(0) = \beta/2$. Percentages below each shift denote the hard-over-easy advantage.

from 0.711 to 0.603 on Assay. Furthermore, with only 20 auxiliary molecules, Mole-PAIR still outperforms the best ID-only detector across all three shifts (Appendix D.3). Training the OOD detection head is computationally efficient, converging in approximately two minutes using around 1.6 GB of GPU memory on a single A100—comparable to the ID-only baselines (Appendix D.4).

Transfer across splits. To investigate whether Mole-PAIR can generalize across different split types, we train and validate it on the scaffold shift of one DrugOOD dataset and evaluate it on the size shift of another. The transferred models achieve 0.968 AUROC and 0.196 FPR95 for EC50-Scaffold→IC50-Size, and 0.984 AUROC and 0.069 FPR95 for IC50-Scaffold→EC50-Size, both outperforming the best ID-only detector trained directly on the target split. This suggests that Mole-PAIR acquires generalizable knowledge of background chemical distributions from the frozen backbones rather than overfitting to the specific OOD family observed during exposure. Details are provided in Appendix D.5.

5 Analysis and Justification

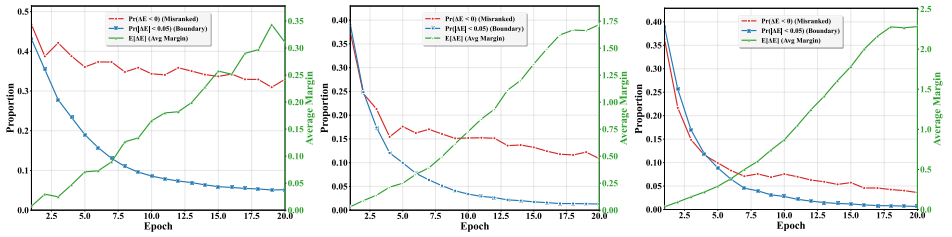
Having demonstrated that Mole-PAIR substantially improves OOD detection for molecular foundation models, we now provide complementary theoretical justification for the pairwise preference objective at its core. Out-of-distribution detection fundamentally evaluates ranking performance: the AUROC equals $\Pr\left(E_\phi(S_{\text{in}}) < E_\phi(S_{\text{out}})\right)$ (Fawcett, 2006; Cortes & Mohri, 2003), which measures how frequently in-distribution scores rank below out-of-distribution scores. Mole-PAIR directly optimizes this through pairwise ID–OOD comparisons via the logistic loss in Eq. 6 and the overall objective in Eq. 7. We first analyze which pairs the objective prioritizes during training, verify this behavior empirically, and subsequently characterize its global optimum.

5.1 Adaptive Learning by Prioritizing Hard Pairs

We begin by unpacking the behavior of our loss function in practice. For any ID–OOD pair $(S_{\text{in}}, S_{\text{out}})$, we define the margin between their ranking scores as:

$$\Delta E_\phi \;=\; E_\phi(S_{\text{out}}) - E_\phi(S_{\text{in}}). \tag{8}$$

The logistic term $\log\big(1 + \exp(-\beta\,\Delta E_\phi)\big)$ is small when the ordering is correct with an appropriate margin, large when the pair is misranked, and steepest near the decision boundary. Consequently, the loss translates the AUROC ranking objective into gradient updates that concentrate on misordered pairs or those near the boundary. To formalize this intuition, we examine the gradient, which naturally emphasizes misranked and borderline pairs. Detailed gradient and curvature derivations are provided in Appendix E.2.



(a) Assay (b) Scaffold (c) Size

Figure 4: **Training dynamics of Mole-PAIR across three distribution shifts.** Each panel plots the misranked-pair proportion $\Pr(\Delta E_\phi < 0)$ and boundary mass $\Pr(|\Delta E_\phi| < \varepsilon)$ with $\varepsilon = 0.05$ (left y-axis), alongside the average margin $\mathbb{E}[\Delta E_\phi]$ (right y-axis) over 20 epochs: hard and borderline pairs are corrected first.

Proposition 5.1 (Hard-pair emphasis)**.** *Let $d_\phi = \nabla_\phi E_\phi(S_{\text{out}}) - \nabla_\phi E_\phi(S_{\text{in}})$. A gradient step of size $\eta > 0$ on Eq. 6 modifies the margin by*

$$\delta(\Delta E_\phi) \;\approx\; \eta\,\beta\,\sigma(-\beta\,\Delta E_\phi)\,\|d_\phi\|_2^2 \;\geq\; 0, \tag{9}$$

where $\sigma(u) = (1 + e^{-u})^{-1}$. The weight $\sigma(-\beta\,\Delta E_\phi)$ decreases monotonically with ΔE_ϕ, meaning it is largest for misranked or borderline pairs and smallest for already well-separated pairs. A detailed proof is provided in Appendix E.3, and empirical verification is presented in Section 5.2.

Proposition 5.1 reveals the intuition behind Mole-PAIR: during training, pairs that degrade AUROC the most—namely, those misranked or near the boundary—receive the strongest gradient updates and are separated first. Once a pair is confidently ranked, it ceases to consume optimization effort, shifting the optimizer's attention toward remaining hard cases. This self-correcting behavior mirrors the nature of AUROC itself, which is influenced exclusively by misranked or borderline pairs.

5.2 Empirical Verification

Prioritization of difficult pairs. Across all shift types, we observe the predicted ordering *Hard* > *Boundary* > *Easy* (Figure 3), with the boundary mean matching $w_\beta(0) = 0.0500$ as required by Eq. 20. Gradient updates thus concentrate heavily on misranked or borderline pairs while dedicating minimal optimization budget to samples that are already well-separated, aligning precisely with the behavior predicted by Proposition 5.1. Because AUROC equals $\Pr\big(E_\phi(S_{\text{in}}) < E_\phi(S_{\text{out}})\big)$, this resource allocation first reduces the mass of misordered pairs before widening margins, yielding the exact training trajectory illustrated in Figure 4.

Shift-specific weight distribution. The hard-over-easy advantage increases progressively from Assay ($+14.7\%$) to Scaffold ($+28.4\%$) and Size ($+32.4\%$). In Assay splits, many misranked pairs exhibit only mildly negative margins and easy pairs reside close to the boundary, whereas Scaffold splits feature a more sharply separated easy-pair population, generating a larger gap. Size shifts yield the greatest separation because distinct atomic-count signatures push easy pairs far from the boundary. This ordering also elucidates the temperature sensitivity observed in Figure 2: because $w_\beta(\Delta)$ is bounded by β and equals $\beta/2$ at the boundary, β controls how sharply updates concentrate on the misranked region. Consequently, Scaffold benefits from the sharper emphasis of a larger β, Assay prefers a moderate temperature, and Size settles early regardless of β.

Self-paced learning behavior. Across all three shifts, the empirical dynamics in Figure 4 corroborate our theoretical analysis that Mole-PAIR prioritizes hard and borderline pairs. Specifically, both the misranked proportion $\Pr(\Delta E_\phi < 0)$ and the boundary mass $\Pr(|\Delta E_\phi| < \varepsilon)$ drop rapidly during the initial epochs, while the mean margin $\mathbb{E}[\Delta E_\phi]$ increases steadily throughout training. This behavior is directly explained by Eq. 9: gradient updates are weighted by $w_\beta(\Delta) = \beta\,\sigma(-\beta\Delta)$, which is largest for misranked or borderline pairs and vanishes for well-separated ones. Thus, the optimizer systematically corrects the pairs that most severely degrade AUROC first.

Shift-specific dynamics. The three splits exhibit distinct rates of separation, consistent with our ablation studies. For Size splits, the misranked and boundary proportions rapidly approach zero

while the average margin grows sharply, reflecting the strong geometric size cue in this split. Differently, with Scaffold splits the model learns steadily, requiring more epochs to push borderline pairs away from the decision boundary, which aligns with our β-sensitivity analysis showing that Scaffold favors a sharper preference signal. Assay is more challenging: we observe that the misranked proportion diminishes but plateaus at a higher level, the boundary mass decreases more gradually, and the margin grows slowly and remains comparatively small, consistent with Assay favoring a moderate β and weaker regularization λ.

5.3 Convergence to the Optimal Ranking

The preceding analysis demonstrates how our training objective dynamically prioritizes difficult instances. We now establish that this optimization process converges to a globally optimal solution. Specifically, we prove that given sufficient data and model capacity, the learned scoring function achieves the Bayes-optimal ranking. We define an ID-preference score as:

$$f(S) \triangleq -E_\phi(S), \tag{10}$$

such that larger values of f indicate higher "ID-likeness" (equivalently, lower OOD ranking scores). For any two samples S and S', the pairwise score margin is defined as:

$$z(S, S') \;\triangleq\; f(S) - f(S') \;=\; E_\phi(S') - E_\phi(S). \tag{11}$$

So $z > 0$ indicates that S is correctly ranked ahead of S' as in-distribution (i.e., $E_\phi(S) < E_\phi(S')$).

Lemma 5.2 (Local Pairwise Optimality)**.** *Let $\eta(S) = \Pr(\text{ID} \mid S)$ denote the true posterior probability that sample S is in-distribution. For any two samples S and S', consider the conditional pairwise risk for the margin $z = f(S) - f(S')$:*

$$r_\beta(z; S, S') = \eta(S)\big(1 - \eta(S')\big)\log\big(1 + e^{-\beta z}\big) + \eta(S')\big(1 - \eta(S)\big)\log\big(1 + e^{\beta z}\big). \tag{12}$$

Then, $r_\beta(z; S, S')$ is strictly convex in z and achieves its unique global minimum at:

$$z^\star = \beta^{-1}\log\frac{\eta(S)\,[1 - \eta(S')]}{\eta(S')\,[1 - \eta(S)]}, \tag{13}$$

whose sign matches that of $\eta(S) - \eta(S')$. A detailed proof is provided in Appendix E.4.

Lemma 5.2 demonstrates that the optimal margin $z^\star$ for any pair aligns with the true underlying posteriors: it is positive when S has a higher ID probability than S' ($\eta(S) > \eta(S')$), zero when both samples share equal posterior probabilities so that no artificial preference is imposed on ambiguous pairs, and its magnitude $|z^\star|$ scales with the posterior gap. The temperature parameter β merely rescales this margin without altering the resulting ranking.

Proposition 5.3 (Global Convergence to the Bayes-Optimal Ranking)**.** *Define the expected pairwise risk of a scoring function f as:*

$$\mathcal{R}_\beta(f) \;=\; \mathbb{E}_{(S_{\text{in}}, S_{\text{out}})}\Big[\log\big(1 + e^{-\beta\,[\,f(S_{\text{in}}) - f(S_{\text{out}})\,]}\big)\Big], \tag{14}$$

where the expectation is taken over independent draws $S_{\text{in}} \sim D_{\text{in}}$ and $S_{\text{out}} \sim D_{\text{out}}$. For any sufficiently expressive function class, every global minimizer $f^\star$ of $\mathcal{R}_\beta$ induces the identical ranking order as $\eta(\cdot)$ for almost all pairs, thereby achieving the Bayes-optimal AUROC. A detailed proof is provided in Appendix E.5.

Proposition 5.3 extends the pairwise result of Lemma 5.2 to the entire data distribution. Because the global risk aggregates individual pairwise terms, any scorer that systematically misranks a subset of pairs can be improved by adjusting its margins toward the local optima of Lemma 5.2. Consequently, with sufficient model capacity and data, minimizing our objective recovers the exact ranking induced by $\eta(\cdot)$, which by definition maximizes the AUROC.

6 Conclusion

We study OOD detection for molecular foundation models and propose Mole-PAIR, a lightweight model-agnostic post-hoc module that equips off-the-shelf foundation models with OOD detection capability. Mole-PAIR formulates OOD detection as preference optimization with a pairwise loss that provably aligns with AUROC, and its effectiveness is demonstrated on five real-world datasets under three distribution shifts.

REFERENCES

Anastasios N Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

Anastasios N Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. *arXiv preprint arXiv:2208.02814*, 2022.

Staffan Arvidsson McShane, Ulf Norinder, Jonathan Alvarsson, Ernst Ahlberg, Lars Carlsson, and Ola Spjuth. Cpsign: conformal prediction for cheminformatics modeling. *Journal of Cheminformatics*, 16(1):75, 2024.

Tianyi Bao, Qitian Wu, Zetian Jiang, Yiting Chen, Jiawei Sun, and Junchi Yan. Graph out-of-distribution detection goes neighborhood shaping. In *Forty-first International Conference on Machine Learning*, 2024.

Dominique Beaini, Shenyang Huang, Joao Alex Cunha, Zhiyi Li, Gabriela Moisescu-Pareja, Oleksandr Dymov, Samuel Maddrell-Mander, Callum McLean, Frederik Wenkel, Luis Müller, et al. Towards foundational models for molecular learning on large-scale multi-task datasets. *arXiv preprint arXiv:2310.04292*, 2023.

Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. Lof: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pp. 93–104, 2000.

Corinna Cortes and Mehryar Mohri. Auc optimization vs. error rate minimization. *Advances in neural information processing systems*, 16, 2003.

Laurianne David, Amol Thakkar, Rocío Mercado, and Ola Engkvist. Molecular representations in ai-driven drug discovery: a review and practical guide. *Journal of cheminformatics*, 12(1):56, 2020.

Tom Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059. PMLR, 2016.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30, 2017.

Yonatan Geifman and Ran El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. In *International conference on machine learning*, pp. 2151–2159. PMLR, 2019.

Will Grathwohl, Kuan-Chieh Wang, Jörn-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one, 2020. URL https://arxiv.org/abs/1912.03263.

Shurui Gui, Xiner Li, Limei Wang, and Shuiwang Ji. Good: A graph out-of-distribution benchmark, 2022. URL https://arxiv.org/abs/2206.08452.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Yuxin Guo, Cheng Yang, Yuluo Chen, Jixi Liu, Chuan Shi, and Junping Du. A data-centric framework to endow graph neural networks with out-of-distribution detection ability. In *Proceedings of the 29th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 638–648, 2023.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure, 2019. URL https://arxiv.org/abs/1812.04606.

Jiwoo Hong, Noah Lee, and James Thorne. Orpo: Monolithic preference optimization without reference model. *arXiv preprint arXiv:2403.07691*, 2024.

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, 2022. URL https://arxiv.org/abs/2201.09637.

Shengli Jiang, Shiyi Qin, Reid C Van Lehn, Prasanna Balaprakash, and Victor M Zavala. Uncertainty quantification for molecular property predictions with graph neural architecture search. *Digital Discovery*, 3(8):1534–1553, 2024.

Thorsten Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 133–142, 2002.

Kerstin Kläser, Błażej Banaszewski, Samuel Maddrell-Mander, Callum McLean, Luis Müller, Ali Parviz, Shenyang Huang, and Andrew Fitzgibbon. MiniMol: A parameter-efficient foundation model for molecular learning, 2024. URL https://arxiv.org/abs/2404.14986.

Siddhartha Laghuvarapu, Zhen Lin, and Jimeng Sun. Codrug: Conformal drug property prediction with density estimation under covariate shift. *Advances in Neural Information Processing Systems*, 36:37728–37747, 2023.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113 (523):1094–1111, 2018.

Zenan Li, Qitian Wu, Fan Nie, and Junchi Yan. Graphde: A generative framework for debiased learning and out-of-distribution detection on graphs. *Advances in Neural Information Processing Systems*, 35:30277–30290, 2022.

Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 eighth ieee international conference on data mining*, pp. 413–422. IEEE, 2008.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33:21464–21475, 2020.

Yixin Liu, Kaize Ding, Huan Liu, and Shirui Pan. Good-d: On unsupervised graph out-of-distribution detection. In *Proceedings of the sixteenth ACM international conference on web search and data mining*, pp. 339–347, 2023.

Shuqi Lu, Zhifeng Gao, Di He, Linfeng Zhang, and Guolin Ke. Data-driven quantum chemical property prediction leveraging 3d conformations with uni-mol+. *Nature Communications*, 15(1): 7104, 2024.

Yizhen Luo, Kai Yang, Massimo Hong, Xing Yi Liu, and Zaiqing Nie. Molfm: A multimodal molecular foundation model. *arXiv preprint arXiv:2307.09484*, 2023.

Gerald M. Maggiora. On outliers and activity cliffswhy qsar often disappoints. *Journal of Chemical Information and Modeling*, 46(4):1535–1535, 2006. doi: 10.1021/ci060117s. URL https://doi.org/10.1021/ci060117s. PMID: 16859285.

Oscar Méndez-Lucio, Christos A Nicolaou, and Berton Earnshaw. Mole: a foundation model for molecular graphs using disentangled attention. *Nature Communications*, 15(1):9431, 2024.

Yu Meng, Mengzhou Xia, and Danqi Chen. Simpo: Simple preference optimization with a reference-free reward. *Advances in Neural Information Processing Systems*, 37:124198–124235, 2024.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2024. URL https://arxiv.org/abs/2305.18290.

Bharath Ramsundar, Bowen Liu, Zhenqin Wu, Andreas Verras, Matthew Tudor, Robert P. Sheridan, and Vijay Pande. Is multitask deep learning practical for pharma? *Journal of Chemical Information and Modeling*, 57(8):2068–2076, 2017. doi: 10.1021/acs.jcim.7b00146. URL https://doi.org/10.1021/acs.jcim.7b00146. PMID: 28692267.

Yaniv Romano, Matteo Sesia, and Emmanuel Candes. Classification with valid and adaptive coverage. *Advances in neural information processing systems*, 33:3581–3591, 2020.

Yu Rong, Yatao Bian, Tingyang Xu, Weiyang Xie, Ying Wei, Wenbing Huang, and Junzhou Huang. Self-supervised graph transformer on large-scale molecular data. *Advances in neural information processing systems*, 33:12559–12571, 2020.

Benjamin Sanchez-Lengeling, Carlos Outeiral, Gabriel L Guimaraes, and Alan Aspuru-Guzik. Optimizing distributions over molecular space. an objective-reinforced generative adversarial network for inverse-design chemistry (organic). 2017.

Bernhard Schölkopf, Robert C Williamson, Alex Smola, John Shawe-Taylor, and John Platt. Support vector method for novelty detection. *Advances in neural information processing systems*, 12, 1999.

Xu Shen, Yili Wang, Kaixiong Zhou, Shirui Pan, and Xin Wang. Optimizing ood detection in molecular graphs: A novel approach with diffusion models, 2024. URL https://arxiv.org/abs/2404.15625.

Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *International conference on machine learning*, pp. 20827–20840. PMLR, 2022.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro Von Werra, Clémentine Fourrier, Nathan Habib, et al. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944*, 2023.

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. doi: 10.1021/acs.jcim.2c01073. URL https://doi.org/10.1021/acs.jcim.2c01073. PMID: 36456532.

Chaoqi Wang, Yibo Jiang, Chenghao Yang, Han Liu, and Yuxin Chen. Beyond reverse kl: Generalizing direct preference optimization with diverse divergence constraints. *arXiv preprint arXiv:2309.16240*, 2023a.

Danny Wang, Ruihong Qiu, Guangdong Bai, and Zi Huang. Gold: Graph out-of-distribution detection via implicit adversarial latent generation. *arXiv preprint arXiv:2502.05780*, 2025.

Fangxin Wang, Yuqing Liu, Kay Liu, Yibo Wang, Sourav Medya, and Philip S Yu. Uncertainty in graph neural networks: A survey. *arXiv preprint arXiv:2403.07185*, 2024.

Yi Wang, Hui Tang, Lichao Huang, Lulu Pan, Lixiang Yang, Huanming Yang, Feng Mu, and Meng Yang. Self-play reinforcement learning guides protein engineering. *Nature Machine Intelligence*, 5(8):845–860, 2023b.

Qitian Wu, Yiting Chen, Chenxiao Yang, and Junchi Yan. Energy-based out-of-distribution detection for graph neural networks. *arXiv preprint arXiv:2302.02914*, 2023.

Shirley Wu, Kaidi Cao, Bruno Ribeiro, James Zou, and Jure Leskovec. Graphmetro: Mitigating complex graph distribution shifts via mixture of aligned experts, 2024. URL https://arxiv.org/abs/2312.04693.

Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. Hallucination is inevitable: An innate limitation of large language models, 2025. URL https://arxiv.org/abs/2401.11817.

Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. Uni-mol: A universal 3d molecular representation learning framework. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=6K2RM6wVqKu.

A Large Language Model Usage Statement

We acknowledge the use of a large language model (LLM) as an auxiliary tool in the preparation of this manuscript. The role of the LLM was strictly confined to language polishing and editing. Specifically, we utilized its capabilities to improve grammar, refine sentence structure, enhance clarity, and ensure stylistic consistency throughout the paper. This process helped us articulate our ideas more effectively and adhere to the high standards of academic writing.

It is crucial to emphasize that all core scientific aspects of this research were conducted entirely by the human authors without the assistance of any LLM. This includes, but is not limited to:

- The conceptualization of the research problem and the formulation of the main hypotheses.
- The design of the model architecture, algorithms, and the overall methodological framework.
- The collection, preprocessing, and curation of all datasets used in our experiments.
- The writing and implementation of all source code for experiments and data analysis.
- The execution of all experiments, the generation of results, and the subsequent data analysis.
- The interpretation of the results and the formulation of the scientific conclusions and future work.

The intellectual contributions, technical innovations, and scientific insights presented in this paper are exclusively the work of the authors. We critically reviewed and edited all text generated or modified by the LLM to ensure it accurately reflects our original thoughts and findings. The authors bear full responsibility for the final content, scientific accuracy, and all claims made in this work.

B Impact Statement and Limitation Discussions

This work aims to improve the reliability of molecular foundation models under distribution shift, which can have positive societal impact in safety-critical scientific applications such as drug discovery, molecular screening, and materials design. By improving OOD detection, the proposed method can help reduce overconfident errors on unfamiliar molecules and support more reliable downstream decisions. At the same time, the method does not eliminate all risks. False positives and false negatives in OOD detection may still occur, and overreliance on automatic reliability scores could lead to inappropriate trust in model outputs. In high-stakes applications, the method should therefore be used as a supporting tool rather than as a replacement for experimental validation or expert judgment. More broadly, improved molecular modeling tools could also be misused in settings that require careful governance, so responsible release and deployment remain important.

C Experiment Details

C.1 Dataset

DrugOOD Dataset. DrugOOD (Ji et al., 2022) is a benchmark and automated dataset curator for OOD challenges in AI-aided drug discovery, built from the large-scale ChEMBL bioassay database.

Table 2: **Statistics of the DrugOOD datasets.** Pos and Neg denote the numbers of positive and negative data points, and $D^{\#}$ and $C^{\#}$ denote the numbers of domains and data points.

Data subset	$Pos^{\#}$	$Neg^{\#}$	Train		ID Val		ID Test		OOD Val		OOD Test	
			$D^{\#}$	$C^{\#}$	$D^{\#}$	$C^{\#}$	$D^{\#}$	$C^{\#}$	$D^{\#}$	$C^{\#}$	$D^{\#}$	$C^{\#}$
drugood-lbap-general-ec50-assay	92445	18000	1079	39333	1079	12849	1079	14086	1137	22095	2883	22082
drugood-lbap-general-ec50-scaffold	92445	18000	8677	23481	2450	4977	30611	37811	7659	22095	4844	22081
drugood-lbap-general-ec50-size	92445	18000	294	42697	238	14151	312	14531	4	19301	20	19765
drugood-lbap-general-ic50-assay	476865	91691	6917	201951	6917	65424	6917	73777	6207	113704	16814	113700
drugood-lbap-general-ic50-scaffold	476865	91691	43552	129740	13516	29174	142173	182220	29513	113723	15189	113699
drugood-lbap-general-ic50-size	476865	91691	290	217294	243	72349	311	72742	4	102544	22	103627

It focuses on the crucial task of drug-target binding affinity prediction, which is framed as a binary classification problem (active vs. inactive). For this, we focus on the Ligand-Based Affinity Prediction (LBAP) variants of DrugOOD under three domain shifts based on biochemistry knowledge:

- **Assay**: Samples are split by the experimental assay, simulating shifts in experimental environments. Assays with many samples are used for training, while those with fewer are used for testing.
- **Scaffold**: Samples are split by their molecular scaffold structure. The largest scaffolds are assigned to the training set and the smallest to the test set to maximize structural diversity.
- **Size**: Samples are split by the number of atoms in the molecule. Molecules with the largest atomic sizes are used for training and smaller ones for testing to ensure variability.

Specifically, we use the `drugood_lbap_general_[ec50, ic50]_(assay, scaffold, size)` subsets, which are the standard LBAP partitions provided in DrugOOD. These subsets cover complementary domain shifts under both `EC50` and `IC50` measurement settings, offering diverse and challenging scenarios for OOD evaluation. Deterministic splits are constructed using `data_seed=42`, with the following target sizes: `train_id=2000`, `train_ood=2000`, `val_id=600`, `val_ood=600`, `test_id=1000`, and `test_ood=1000`. Supervised training labels are read from the `cls_label` field in the JSON files. Statistics of these subsets are summarized in Table 2.

GOOD Datasets. The GOOD (Gui et al., 2022) benchmark provides molecular graph datasets designed for systematic OOD evaluation. We use three datasets:

- **GOOD-HIV**: A small-scale dataset adapted from MoleculeNet. Inputs are molecular graphs with atoms as nodes and chemical bonds as edges. The task is binary classification to predict whether a molecule inhibits HIV replication.
- **GOOD-PCBA**: A large-scale dataset with 128 bioassays, forming a multi-target binary classification task.
- **GOOD-ZINC**: A molecular property regression dataset derived from the ZINC database, where molecules contain up to 38 heavy atoms. The task is to predict constrained solubility.

For our experiments, we generate deterministic splits with sizes: `train_id/train_ood=5000`, `val_id/val_ood=1500`, and `test_id/test_ood=2000`. Labels follow the official benchmark protocol. For regression-like tasks such as `GOOD-ZINC`, we binarize the target via a fixed median split: samples with values greater than or equal to the median are assigned positive labels, and the rest are negative. Statistics of the datasets are summarized in Table 3.

Splitting, caching, and features. All dataset splits are generated deterministically using seed 42. We pre-compute molecular representations using foundation encoders and cache them for reuse. Feature extraction is performed with batch size 50, and the resulting representations are used both to train the supervised classifier head and as the basis for OOD scoring methods.

C.2 Training Details

We use pre-trained MiniMol (Kläser et al., 2024) (2D graph features) and Uni-Mol (Zhou et al., 2023; Lu et al., 2024) (3D conformational features) as frozen encoders, each yielding a fixed 512-

dimensional representation per molecule. Our Mole-PAIR module attaches a lightweight MLP scoring head with structure $512 \to 256 \to 128 \to 1$ and dropout 0.1. This head is the only trainable part. Mole-PAIR is trained without class labels, using only ID/OOD pairing signals. We optimize with AdamW, apply a StepLR scheduler (step size 10, $\gamma = 0.9$), and clip gradients at norm 1.0. The temperature is $\beta = 0.1$. Training runs for up to 500 epochs with batch size 512 for MiniMol and 256 for Uni-Mol.

C.3 BASELINES

We compare against several widely used supervised OOD detection baselines, all of which operate on the logits or penultimate features of a trained classifier head. Unless otherwise specified, the OOD score is defined such that larger values indicate a higher likelihood of being out-of-distribution. The classifier head is a three-layer MLP with hidden dimension 64 attached to the frozen 512-dimensional features, trained with CrossEntropyLoss for single-task classification, BCEWithLogitsLoss for multi-task classification, and MSELoss for regression prior to binarization. Continuous targets such as GOOD-ZINC are binarized by a median split (label 1 if the value is $\geq$ median, else 0). Optimization uses Adam with learning rate 0.01 and weight decay 5×10^{-4} for 500 epochs, with early stopping and best-validation checkpointing. The exposure-based variants MSP-OE and ODIN-OE use a capacity-matched $512 \to 256 \to 128 \to C$ classifier head.

MSP (Hendrycks & Gimpel, 2016) calculates the in-distribution (ID) score as the maximum class confidence from the classifier head's logits. For multi-class tasks, this is the maximum softmax probability, while for binary or single-output tasks it is $\max(p, 1-p)$ from the sigmoid output. In multi-task settings, the confidence is first computed for each task and then averaged to yield the final ID score. The reported OOD score is defined as $1 -$ ID score.

Energy (Liu et al., 2020) uses the energy score defined as $E(x) = -\log \sum_c \exp(z_c)$, where z are the logits. For binary or single-output tasks, we augment the logits as $[z, -z]$ before computing the log-sum-exp. In the multi-task case, the energy is computed per task and then averaged. Larger energy values correspond to higher OOD likelihood.

ODIN (Liang et al., 2017) is implemented in our setting directly on the pre-computed molecular feature representations that serve as inputs to the classifier head, rather than on raw molecules. A small perturbation is applied according to $x' = x - \varepsilon \cdot \text{sign}(\partial \mathcal{L}_{\text{CE}} / \partial x)$, with $\varepsilon = 0.0014$. Temperature scaling with $T = 1000$ is applied exactly once to the logits. The model is kept in evaluation mode with Batch Normalization layers frozen. The confidence score is computed in the same way as MSP after perturbation, and task-level confidences are averaged in the multi-task case. The OOD score is then reported as $1 -$ confidence.

OE-enabled classifier scores. MSP-OE and ODIN-OE add a uniform-output penalty on the auxiliary OOD samples. The ODIN-OE entry in Table 1 is a temperature-only frozen-feature adaptation: unlike the ID-only ODIN implementation above, it does not apply a feature perturbation. For GOOD, these classifier baselines use one fixed task adapter across methods and backbones (the native HIV label, a fixed PCBA assay column, and a fixed median-binarized ZINC target).

Mahalanobis (Lee et al., 2018) operates in the penultimate feature space of the classifier head. For single-task classification, it estimates class-conditional means and a shared precision matrix, where Ledoit–Wolf shrinkage is used preferentially and empirical covariance is used as a fallback. The OOD score for a test sample is its minimum Mahalanobis distance to any class mean. In the multi-task setting, we instead fit a single global mean and a shared precision matrix, and the OOD score is given by the Mahalanobis distance to this global mean.

KNN (Sun et al., 2022) also works in the penultimate feature space, standardized using the mean and standard deviation of the training set. The OOD score is defined as the average Euclidean distance to the $k = 50$ nearest neighbors in this space. A larger average distance indicates that the sample is more likely to be OOD.

LOF (Breunig et al., 2000) is applied in the same standardized feature space. A Local Outlier Factor (LOF) model is fitted in novelty detection mode with $n_{\text{neighbors}} = 20$. The OOD score is defined as $-\texttt{score_samples}(\cdot)$, so that higher values correspond to samples that deviate more strongly from the local density of their neighbors.

Table 3: **Statistics of the GOOD datasets.** Numbers of graphs in the training, ID validation, ID test, OOD validation, and OOD test sets.

Dataset	Shift	Train	ID validation	ID test	OOD validation	OOD test	Train	OOD validation	ID validation	ID test	OOD test
		Scaffold					Size				
GOOD-HIV	covariate	24682	4112	4112	4113	4108	26169	4112	4112	2773	3961
	concept	15209	3258	3258	9365	10037	14454	3096	3096	9956	10525
	no shift	24676	8225	8226	-	-	24676	8225	8226	-	-
		Scaffold					Size				
GOOD-PCBA	covariate	262764	43792	43792	44019	43562	269990	43792	43792	48430	31925
	concept	159158	34105	34105	90740	119821	150121	32168	32168	108267	115205
	no shift	262757	87586	87586	-	-	262757	87586	87586	-	-
		Scaffold					Size				
GOOD-ZINC	covariate	149674	24945	24945	24945	24946	161893	24945	24945	20270	17402
	concept	101867	21828	21828	43539	60393	89418	19161	19161	51409	70306
	no shift	149673	49891	49891	-	-	149673	49891	49891	-	-

Monte Carlo Dropout (Gal & Ghahramani, 2016) approximates Bayesian inference by maintaining stochastic dropout during the inference phase. In our implementation, the model is kept in evaluation mode to freeze Batch Normalization statistics, while dropout layers are explicitly set to training mode. We perform $T = 20$ stochastic forward passes for each test input. The raw logits from each pass are converted to probabilities—using softmax for multi-class tasks or sigmoid for binary tasks—and then averaged to obtain the expected predictive distribution. The OOD score is defined as the predictive entropy of this mean distribution. Higher entropy indicates higher uncertainty, corresponding to a higher likelihood of being OOD.

Conformal Prediction (Angelopoulos & Bates, 2021) adopts a standard split conformal prediction, utilizing the full validation set as the calibration set $\mathcal{C}$. The conformity score $s(x)$ is defined as the maximum predicted probability (averaged across tasks for multi-task settings). For a test input x, we compute the conformal p-value relative to the calibration scores: $p(x) = \frac{1+\sum_{s_i \in \mathcal{C}} \mathbb{I}(s_i \leq s(x))}{1+|\mathcal{C}|}$. The OOD score is reported as $1 - p(x)$, so that lower p-values indicate higher OOD likelihood.

D Additional Experimental Results

We supplement additional experimental results in this section to provide a comprehensive evaluation of Mole-PAIR. Section D.1 reports the complete AUROC, AUPR, and FPR95 results on all twelve dataset–shift cells. Section D.2 studies the sensitivity to the regularization strength, Section D.3 the impact of OOD data scarcity, Section D.4 the computational cost, and Section D.5 the generalization to unseen OOD domains. We further present a case study on objective–metric alignment in Section D.6, comparisons with domain-specific graph OOD detectors and uncertainty-based methods in Sections D.7 and D.8, and an ablation on backbone fine-tuning strategies in Table 12.

D.1 Complete Results on All Dataset–Shift Cells

Table 1 in the main text reports only the four shifts on which no method saturates. Tables 4, 5, and 6 report the complete AUROC, AUPR, and FPR95 results of all methods on all twelve dataset–shift cells, including the size shifts and GOOD-ZINC. On the size shifts, Mole-PAIR and the -OE baselines approach perfect separation, whereas the two GOOD-ZINC–Scaffold cells are the only configurations in which a distance-based ID-only detector (Mahalanobis or KNN) ranks first.

D.2 Sensitivity to the Regularization Strength

Figure 5 complements the temperature study in Figure 2 by varying the ℓ_2 regularization strength λ with the temperature fixed at $\beta = 0.1$. The corresponding discussion is given in Section 4.2.

D.3 Impact of OOD Data Scarcity

To validate the efficacy of Mole-PAIR under extreme OOD data scarcity, we manipulated the ID:OOD ratio in the training data. Specifically, we reduced the number of training OOD samples

Table 4: **Complete AUROC (↑) results on all twelve dataset–shift cases.** GOOD-ZINC is included here rather than in Table 1 because its continuous target has to be binarized to define in-distribution membership. The -OE variants and Mole-PAIR share the same auxiliary OOD set and exposure budget, while the rows above them use ID data only. Best in **boldface**, second best underlined.

		EC50			IC50			HIV		PCBA		ZINC	
		Scaffold	Size	Assay	Scaffold	Size	Assay	Scaffold	Size	Scaffold	Size	Scaffold	Size
MiniMol	MSP	0.677	0.449	0.420	0.574	0.515	0.600	0.408	0.194	0.632	0.697	0.359	0.398
	ODIN	0.633	0.450	0.437	0.575	0.516	0.614	0.390	0.336	0.623	0.691	0.360	0.398
	Energy	0.685	0.527	0.455	0.562	0.573	0.569	0.388	0.167	0.642	0.735	0.359	0.398
	Mahalanobis	0.660	0.831	0.572	0.620	0.751	0.516	0.503	0.918	0.564	0.728	**0.638**	0.593
	LOF	0.665	0.838	0.537	0.625	0.745	0.564	0.508	0.889	0.564	0.687	0.544	0.688
	KNN	0.671	0.855	0.569	0.655	0.765	0.502	0.562	0.921	0.459	0.527	0.636	0.676
	MSP-OE	0.949	0.991	0.685	0.965	0.991	0.634	0.750	0.994	0.842	0.990	0.526	0.992
	ODIN-OE	0.947	0.991	0.675	0.966	0.991	0.622	0.749	0.992	0.837	0.992	0.524	0.992
	Mole-PAIR (Ours)	**0.970**	**1.000**	**0.743**	**0.983**	**0.999**	**0.660**	**0.777**	**1.000**	**0.924**	**1.000**	0.614	**1.000**
Uni-Mol	MSP	0.694	0.722	0.547	0.585	0.637	0.549	0.465	0.218	0.476	0.350	0.352	0.369
	ODIN	0.621	0.656	0.558	0.546	0.591	0.538	0.451	0.278	0.445	0.353	0.352	0.369
	Energy	0.691	0.652	0.547	0.542	0.598	0.551	0.476	0.243	0.476	0.350	0.352	0.369
	Mahalanobis	0.724	0.784	0.591	0.699	0.723	0.562	0.567	0.826	0.568	0.690	0.643	0.643
	LOF	0.748	0.855	0.558	0.717	0.759	0.547	0.594	0.860	0.538	0.607	0.577	0.621
	KNN	0.704	0.747	0.599	0.682	0.660	0.575	0.580	0.826	0.553	0.668	**0.645**	0.686
	MSP-OE	0.952	0.995	0.629	0.951	0.996	0.615	0.711	0.994	0.807	0.986	0.503	0.986
	ODIN-OE	0.939	0.995	0.624	0.952	0.997	0.594	0.712	0.994	0.799	0.987	0.492	0.986
	Mole-PAIR (Ours)	**0.976**	**1.000**	**0.658**	**0.984**	**1.000**	**0.640**	**0.728**	**1.000**	**0.875**	**1.000**	0.549	**1.000**

Table 5: **Complete AUPR (↑) results on all twelve dataset–shift cases.** Same layout as Table 4. Best in **boldface**, second best underlined.

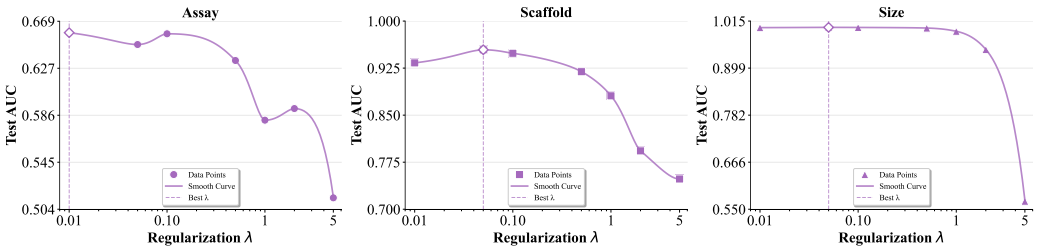
		EC50			IC50			HIV		PCBA		ZINC	
		Scaffold	Size	Assay	Scaffold	Size	Assay	Scaffold	Size	Scaffold	Size	Scaffold	Size
MiniMol	MSP	0.627	0.453	0.445	0.538	0.494	0.580	0.438	0.353	0.601	0.654	0.416	0.417
	ODIN	0.597	0.453	0.454	0.714	0.495	0.590	0.430	0.430	0.591	0.657	0.416	0.417
	Energy	0.625	0.541	0.480	0.557	0.593	0.550	0.423	0.344	0.615	0.700	0.416	0.417
	Mahalanobis	0.696	0.836	0.577	0.656	0.765	0.520	0.497	0.906	0.559	0.713	**0.637**	0.548
	LOF	0.690	0.847	0.545	0.652	0.753	0.561	0.499	0.877	0.575	0.696	0.548	0.661
	KNN	0.700	0.853	0.571	0.679	0.776	0.504	0.538	0.898	0.485	0.519	**0.637**	0.647
	MSP-OE	0.933	0.981	0.657	0.948	0.981	0.599	0.713	0.980	0.822	0.977	0.538	0.975
	ODIN-OE	0.937	0.982	0.648	0.956	0.981	0.598	0.714	0.973	0.823	0.977	0.537	0.973
	Mole-PAIR (Ours)	**0.975**	**1.000**	**0.698**	**0.984**	**0.999**	**0.653**	**0.740**	**1.000**	**0.924**	**1.000**	0.587	**1.000**
Uni-Mol	MSP	0.692	0.697	0.555	0.604	0.642	0.562	0.489	0.353	0.504	0.455	0.437	0.455
	ODIN	0.620	0.652	0.561	0.571	0.625	0.545	0.474	0.387	0.481	0.456	0.437	0.454
	Energy	0.681	0.635	0.553	0.561	0.593	0.560	0.492	0.376	0.504	0.455	0.437	0.455
	Mahalanobis	0.717	0.726	0.590	0.667	0.683	0.568	0.547	0.801	0.558	0.689	**0.694**	0.685
	LOF	0.763	0.849	0.569	0.714	0.757	0.555	0.576	0.862	0.546	0.639	0.610	0.633
	KNN	0.696	0.691	0.595	0.654	0.615	0.581	0.550	0.806	0.552	0.684	0.680	0.704
	MSP-OE	0.932	0.994	0.592	0.911	0.993	0.593	0.677	0.981	0.792	0.965	0.510	0.955
	ODIN-OE	0.930	0.993	0.602	0.919	0.993	0.572	0.682	0.981	0.788	0.970	0.502	0.963
	Mole-PAIR (Ours)	**0.972**	**1.000**	**0.641**	**0.980**	**1.000**	**0.648**	**0.708**	**1.000**	**0.876**	**1.000**	0.563	**1.000**

while keeping the number of ID samples, as well as the validation and test splits constant. We varied the training ID:OOD ratio from 1:1 to 1:0.01 to simulate increasingly scarce OOD availability. The frozen foundation model used here is MiniMol. The AUROC results for Mole-PAIR are shown below:

We observed that performance degradation was moderate even when training OOD samples became extremely scarce. For example, at the 1:0.01 ratio, there are only 20 training OOD samples. How-

Table 6: **Complete FPR95 (↓) results on all twelve dataset–shift cases**, where lower is better. Same layout as Table 4. Best in **boldface**, second best underlined.

		EC50			IC50			HIV		PCBA		ZINC	
		Scaffold	Size	Assay	Scaffold	Size	Assay	Scaffold	Size	Scaffold	Size	Scaffold	Size
MiniMol	MSP	0.725	0.833	0.972	0.759	0.743	0.903	0.953	0.983	0.833	0.698	0.985	0.946
	ODIN	0.734	0.832	0.950	0.563	0.741	0.900	0.962	0.819	0.856	0.763	0.985	0.947
	Energy	0.716	0.726	0.973	0.781	0.662	0.905	0.955	0.985	0.832	0.697	0.985	0.946
	Mahalanobis	0.898	0.633	0.929	0.909	0.786	0.946	0.947	0.315	0.922	0.752	**0.905**	0.874
	LOF	0.893	0.667	0.933	0.909	0.803	0.934	0.943	0.458	0.934	0.804	0.933	0.797
	KNN	0.870	0.550	0.930	0.898	0.782	0.942	0.906	0.280	0.960	0.902	0.906	0.801
	MSP-OE	0.188	0.002	0.859	0.126	0.006	0.916	0.814	**0.000**	0.569	0.004	0.918	**0.000**
	ODIN-OE	0.200	0.003	0.856	0.151	0.009	0.903	0.815	**0.000**	0.536	**0.000**	0.917	**0.000**
	Mole-PAIR (Ours)	**0.178**	**0.000**	**0.823**	**0.084**	**0.004**	**0.861**	**0.624**	0.001	**0.348**	0.001	0.925	**0.000**
Uni-Mol	MSP	0.848	0.733	0.957	0.926	0.861	0.945	0.954	0.985	0.954	0.925	0.992	0.996
	ODIN	0.816	0.798	0.913	0.945	0.911	0.943	0.947	0.977	0.962	0.925	0.993	0.997
	Energy	0.849	0.740	0.948	0.935	0.872	0.949	0.952	0.982	0.954	0.925	0.993	0.996
	Mahalanobis	0.831	0.605	0.921	0.818	0.718	0.936	0.920	0.514	0.836	0.855	0.965	0.938
	LOF	0.844	0.554	0.935	0.832	0.746	0.942	0.894	0.567	0.948	0.926	**0.945**	0.919
	KNN	0.855	0.707	0.914	0.833	0.806	0.929	0.885	0.563	0.941	0.895	0.961	0.899
	MSP-OE	**0.147**	0.006	0.929	0.327	0.002	0.915	0.846	**0.000**	0.627	0.002	**0.945**	0.003
	ODIN-OE	0.291	0.008	0.912	0.270	0.003	0.921	0.831	**0.000**	0.616	0.014	0.947	0.019
	Mole-PAIR (Ours)	0.178	**0.000**	**0.869**	**0.139**	**0.000**	**0.890**	**0.736**	**0.000**	**0.515**	**0.000**	0.949	**0.000**



(a) Assay (b) Scaffold (c) Size

Figure 5: **Test AUROC sensitivity to the ℓ_2 regularization λ with $\beta = 0.1$.** Performance varies with regularization strength: **Assay** performs best with weak regularization, **Scaffold** benefits from a modest amount of regularization, and **Size** remains robust until regularization becomes excessive.

ever, even under such severe data scarcity, our model still outperformed the strongest ID-only detector of Table 4 on all three shifts. These results serve as strong empirical evidence that Mole-PAIR is robust to limited and imbalanced OOD availability, remaining a cost-effective post-hoc remedy even with very few OOD samples.

D.4 Computational Cost

We evaluated the computational efficiency of Mole-PAIR in terms of training time to convergence and peak GPU memory usage. All experiments were conducted on a single NVIDIA A100 GPU. The results are summarized in Table 8.

Observations indicate that Mole-PAIR maintains a highly competitive computational profile. In terms of memory, Mole-PAIR requires only 1.60 GB, which is comparable to the 1.52 GB footprint of standard uncertainty baselines, ensuring it fits easily within standard hardware constraints. Regarding training speed, while the pairwise preference optimization introduces a marginal overhead compared to simple scalar-output methods like MSP or Energy, Mole-PAIR remains more efficient than retrieval-based methods such as KNN. This slight increase in training time is negligible in practice, confirming that Mole-PAIR achieves significant robustness improvements without compromising computational efficiency or scalability.

Table 7: **Robustness of Mole-PAIR to auxiliary OOD scarcity.** AUROC on DrugOOD-EC50 (MiniMol) as the training ID:OOD ratio decreases.

ID:OOD Ratio	**1:1**	**1:0.1**	**1:0.02**	**1:0.01**
DrugOOD-EC50-Scaffold	0.970 ± 0.000	0.924 ± 0.001	0.898 ± 0.003	0.849 ± 0.005
DrugOOD-EC50-Size	1.000 ± 0.000	0.993 ± 0.001	0.991 ± 0.001	0.974 ± 0.020
DrugOOD-EC50-Assay	0.711 ± 0.002	0.638 ± 0.005	0.621 ± 0.004	0.603 ± 0.002

Table 8: **Computational efficiency** measured by training time (s) and peak GPU memory (GB), where lower is better. Best in **boldface**, second best underlined.

Method	Training Time (s) ↓	GPU Memory (GB) ↓
MSP	<u>108.09</u>	**1.52**
ODIN	108.35	<u>1.53</u>
Energy	**107.12**	**1.52**
Mahalanobis	108.31	**1.52**
LOF	108.43	**1.52**
KNN	117.47	**1.52**
Mole-PAIR (Ours)	112.61	1.60

D.5 Generalization to Unseen OOD Domains

To fully evaluate the generalization ability of Mole-PAIR, we train and validate the model on DrugOOD-EC50-Scaffold and test it on DrugOOD-IC50-Size. These challenging settings investigate whether the method relies on seeing specific OOD patterns during training. The results are shown in Table 9.

Table 9: **Cross-domain generalization of Mole-PAIR** across assay types and splitting strategies (AUROC ↑, FPR95 ↓).

Train	**Test**	**AUROC**	**FPR95**
EC50-Scaf.	IC50-Size	$0.968_{\pm 0.001}$	$0.196_{\pm 0.003}$
EC50-Size	EC50-Size	$1.000_{\pm 0.000}$	$0.000_{\pm 0.000}$
IC50-Scaf.	EC50-Size	$0.984_{\pm 0.001}$	$0.069_{\pm 0.003}$
IC50-Size	IC50-Size	$0.999_{\pm 0.000}$	$0.004_{\pm 0.001}$

Despite the significant structural discrepancies, Mole-PAIR achieves high AUROC scores, demonstrating strong robustness. The model performs consistently not only on matched datasets but also on unseen distributions involving different assays and splitting criteria. In particular, successful transfer between scaffold-based and size-based shifts suggests that the detector learns signals that generalize beyond the auxiliary OOD pattern used during training.

D.6 Case Study on Objective–Metric Alignment

Figure 6 compares the OOD affinity scores produced by the detector head trained with a pairwise ranking objective (Joachims, 2002) and with two pointwise objectives, BCE and MSE, on IC50-Scaffold under identical splits and head capacity. The pairwise objective yields globally separated ID and OOD score distributions, which is exactly what AUROC rewards, whereas the pointwise objectives, which fit each score to an absolute target, produce heavily overlapping distributions.

D.7 Comparison with Domain-Specific Graph OOD Methods

The ID-only baselines in Table 1 are generic post-hoc detectors. A more demanding question is how a lightweight ranking head compares against detectors that were purpose-built for graph- and molecule-level OOD detection. GraphDE (Li et al., 2022) infers a latent environment variable

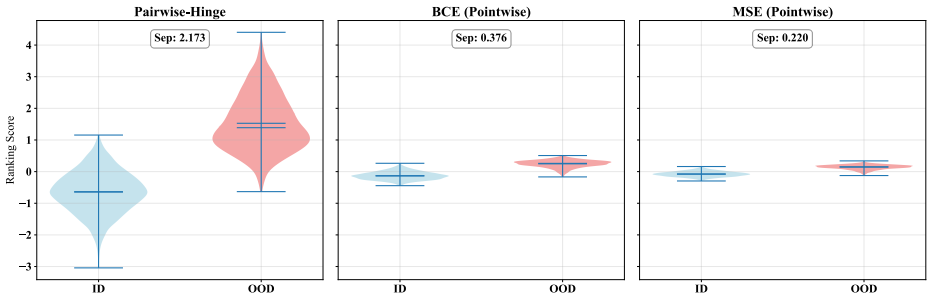


Figure 6: **A case study on objective–metric alignment.** The figure plots the estimated OOD affinity scores yielded by the model trained with different objectives for ID and OOD samples on the IC50-Scaffold task. The Pairwise-Hinge loss (Joachims, 2002) produces a globally separated score distribution between ID and OOD, aligning with AUROC, whereas the two pointwise objectives yield heavily overlapping scores due to their per-sample calibration loss.

from a generative model of the graph, GOOD-D (Liu et al., 2023) learns unsupervised hierarchical contrastive representations over augmented views, and PGR-MOOD (Shen et al., 2024) scores a molecule by its similarity to a diffusion-model reconstruction. Table 10 reports AUROC on four DrugOOD shifts for which results of all three competitors are available.

Table 10: **Comparison with domain-specific graph and molecular OOD detectors on DrugOOD** (AUROC ↑). Best in **boldface**, second best underlined.

Method	DrugOOD-EC50		DrugOOD-IC50	
	Scaffold	Size	Scaffold	Size
GraphDE	0.686±0.010	0.796±0.012	0.692±0.011	0.787±0.010
GOOD-D	0.825±0.013	0.925±0.013	0.854±0.012	0.916±0.011
PGR-MOOD	0.875±0.013	0.977±0.015	0.916±0.013	0.938±0.015
Mole-PAIR (Ours)	**0.970**±0.000	**1.000**±0.000	**0.983**±0.000	**0.999**±0.000

Averaged over the four shifts, Mole-PAIR reaches 0.988 AUROC against 0.927 for the strongest competitor PGR-MOOD, 0.880 for GOOD-D, and 0.740 for GraphDE. On scaffold shifts, which require recognizing unseen molecular skeletons, Mole-PAIR improves over PGR-MOOD by $+0.095$ on EC50 and $+0.067$ on IC50. Even on the easier size splits, it retains gains of $+0.023$ and $+0.061$, respectively. Mole-PAIR is also markedly more stable: its standard deviation stays below 0.001 across all four settings against 0.010–0.015 for the three competitors, as expected when only a small head is trained on frozen features.

We attribute this gap to the objective rather than to the representation. All three competitors estimate an intermediate quantity that is harder than the one the metric evaluates—an environment posterior, a reconstruction likelihood, or a distance in a contrastively learned space—and each is only loosely coupled to the relative ordering of ID and OOD molecules, whereas Mole-PAIR optimizes that ordering directly through Eq. 6, whose optimum induces the Bayes-optimal ranking (Sec. 5). The same holds for the training cost: these methods require a dedicated encoder and a complete training run per dataset, while Mole-PAIR reuses a frozen foundation model and trains only a light-weight head, which converges in under two minutes within 1.6 GB of GPU memory (Appendix D.4).

D.8 Comparison with Uncertainty-Based Methods

To further validate the reliability of Mole-PAIR in high-stakes settings, we compared it against established uncertainty estimation baselines, including Monte Carlo (MC) Dropout and Conformal Prediction, alongside the standard Maximum Softmax Probability (MSP). As shown in Table 11, while MC Dropout and Conformal Prediction provide improvements over the MSP baseline, Mole-PAIR significantly outperforms all uncertainty-based methods.

We attribute this performance gap to the limitations of standard uncertainty quantification under severe distribution shifts. Traditional methods rely on the intrinsic variance of the classifier, which often fails when the model produces high-confidence false predictions on OOD samples. In contrast, Mole-PAIR actively learns a discriminatory scoring function via preference optimization. By relying on learned relative margins rather than unreliable absolute confidence scores, Mole-PAIR demonstrates superior robustness in identifying out-of-distribution molecular scaffolds.

Table 11: **Comparison with uncertainty-based methods** on DrugOOD-EC50-Scaffold and GOOD-HIV-Scaffold (AUROC ↑). Best in **boldface**, second best underlined.

Method	DrugOOD-EC50-Scaffold	GOOD-HIV-Scaffold
MSP (Baseline)	0.677 ± 0.093	0.408 ± 0.054
MC Dropout	0.728 ± 0.015	0.538 ± 0.055
Conformal Prediction	$\underline{0.803 \pm 0.007}$	$\underline{0.576 \pm 0.074}$
Mole-PAIR (Ours)	$\mathbf{0.970 \pm 0.000}$	$\mathbf{0.777 \pm 0.003}$

D.9 Ablation Study on Fine-tuning Strategies

To investigate the impact of encoder adaptability on OOD detection performance, we evaluated three different fine-tuning strategies for the Uni-Mol backbone: full fine-tuning, partial fine-tuning (updating only the last two transformer layers), and the frozen encoder. The quantitative results are summarized in Table 12.

Contrary to the common intuition that fine-tuning improves downstream performance, we observed that the **frozen Uni-Mol backbone** consistently outperformed both partial and full fine-tuning strategies. Specifically, the frozen encoder achieved the highest AUROC scores of 0.965 and 0.650 on the DrugOOD-EC50-Scaffold and DrugOOD-EC50-Assay benchmarks, respectively. We posit that full or partial fine-tuning may introduce representation bias, potentially overfitting to the ID distribution and undermining the generalization capability required for effective OOD detection. Furthermore, keeping the encoder frozen offers a lightweight, plug-and-play solution that is computationally efficient and preserves the inherent adaptability of the foundation model.

Table 12: **Ablation on fine-tuning strategies for the Uni-Mol backbone** (AUROC ↑). Best in **boldface**, second best underlined.

Fine-tuning Strategy	DrugOOD-EC50-Scaffold	DrugOOD-EC50-Assay
Full fine-tuning (Uni-Mol)	0.833	0.583
Partial fine-tuning (Uni-Mol)	$\underline{0.943}$	$\underline{0.623}$
Frozen (Uni-Mol)	**0.965**	**0.650**

E Derivations and Proofs

E.1 Notation and Setup

Let $E_\phi : \mathcal{X} \to \mathbb{R}$ be a differentiable scoring head parameterized by ϕ, with the encoder kept frozen. For any ID–OOD pair $(S_{\text{in}}, S_{\text{out}})$, we define the margin as

$$\Delta E_\phi \;\triangleq\; E_\phi(S_{\text{out}}) - E_\phi(S_{\text{in}}), \tag{15}$$

which is identical to Eq. 8 in the main text. The per-pair logistic term is given by

$$\ell(\Delta E_\phi) \;=\; \log\big(1 + \exp(-\beta\, \Delta E_\phi)\big), \tag{16}$$

and the overall objective corresponds to the pairwise expectation

$$\mathcal{L}(\phi) \;=\; \mathbb{E}_{S_{\text{in}} \sim D_{\text{in}},\, S_{\text{out}} \sim D_{\text{out}}}\big[\ell(\Delta E_\phi)\big], \tag{17}$$

which coincides with Eq. 6 (and its population expectation form in Eq. 14 under the change of variable $f = -E_\phi$ used in the main text). When appropriate, we also consider the total objective regularized by a lightweight ℓ_2 gauge-fixing term:

$$\mathcal{L}_{\text{total}}(\phi) \;=\; \mathcal{L}(\phi) \;+\; \lambda \left(\mathbb{E}_{S_{\text{in}} \sim D_{\text{in}}}[E_\phi(S_{\text{in}})^2] + \mathbb{E}_{S_{\text{out}} \sim D_{\text{out}}}[E_\phi(S_{\text{out}})^2]\right), \tag{18}$$

matching Eq. 7 in the main text.

E.2 Full Gradient and Curvature Calculations

Gradient with respect to the margin. For $z = \Delta E_\phi$, we have

$$\ell'(z) \;=\; \frac{\partial}{\partial z}\log\bigl(1+e^{-\beta z}\bigr) \;=\; -\frac{\beta}{1+e^{\beta z}} \;=\; -\beta\,\sigma(-\beta z),$$

$$\ell''(z) \;=\; \frac{\partial}{\partial z}\ell'(z) \;=\; \beta^2\,\sigma(\beta z)\,\sigma(-\beta z) \;>\; 0,$$

confirming that $\ell(z)$ is strictly convex in z.

Gradient with respect to the parameters. Applying the chain rule yields

$$\begin{aligned}\nabla_\phi \ell(\Delta E_\phi) &= \ell'(\Delta E_\phi)\,\nabla_\phi \Delta E_\phi \\ &= -\beta\,\sigma(-\beta \Delta E_\phi)\,\bigl(\nabla_\phi E_\phi(S_{\text{out}}) - \nabla_\phi E_\phi(S_{\text{in}})\bigr).\end{aligned} \tag{19}$$

Defining the direction term as

$$d_\phi \;\triangleq\; \nabla_\phi E_\phi(S_{\text{out}}) - \nabla_\phi E_\phi(S_{\text{in}}),$$

and the weight function as

$$w_\beta(t) \;\triangleq\; \beta\,\sigma(-\beta t) \in (0,\beta), \tag{20}$$

we obtain $\nabla_\phi \ell(\Delta E_\phi) = -w_\beta(\Delta E_\phi)\,d_\phi$ and

$$\nabla_\phi \mathcal{L}(\phi) \;=\; \mathbb{E}\bigl[-w_\beta(\Delta E_\phi)\,d_\phi\bigr].$$

Furthermore, since

$$w'_\beta(t) \;=\; -\beta^2\sigma(\beta t)\sigma(-\beta t) \;<\; 0, \quad w_\beta(0) = \frac{\beta}{2}, \quad \lim_{t\to-\infty} w_\beta(t) = \beta, \quad \lim_{t\to+\infty} w_\beta(t) = 0,$$

misranked ($t < 0$) or borderline ($t \approx 0$) pairs receive larger weights, whereas already well-separated pairs ($t \gg 0$) are nearly ignored.

Per-example backpropagation signals. Differentiating Eq. 16 with respect to the scalar ranking scores gives

$$\frac{\partial \ell}{\partial E_\phi(S_{\text{out}})} \;=\; -\beta\,\sigma(-\beta\Delta E_\phi), \qquad \frac{\partial \ell}{\partial E_\phi(S_{\text{in}})} \;=\; +\beta\,\sigma(-\beta\Delta E_\phi),$$

which are equal in magnitude and opposite in sign. If E_ϕ is linear over a frozen representation $h(\cdot)$ such that $E_\phi(S) = \langle w, h(S)\rangle + b$, then d_ϕ simplifies to $h(S_{\text{out}}) - h(S_{\text{in}})$, reducing the gradient update to a standard contrastive form.

E.3 Proof of Proposition 5.1

We demonstrate that a gradient update increases the margin by an amount proportional to $w_\beta(\Delta E_\phi)\|d_\phi\|_2^2$.

First-order calculation. A gradient descent step with step size $\eta > 0$ yields

$$\phi^+ \;=\; \phi - \eta\,\nabla_\phi \ell(\Delta E_\phi) \;=\; \phi + \eta\,w_\beta(\Delta E_\phi)\,d_\phi.$$

By a first-order Taylor expansion,

$$\begin{aligned}\delta(\Delta E_\phi) &\triangleq \Delta E_\phi(\phi^+) - \Delta E_\phi(\phi) \\ &\approx \nabla_\phi \Delta E_\phi(\phi)^\top(\phi^+ - \phi) \\ &= d_\phi^\top\bigl(\eta\,w_\beta(\Delta E_\phi)\,d_\phi\bigr) \\ &= \eta\,w_\beta(\Delta E_\phi)\,\|d_\phi\|_2^2 \;\geq\; 0,\end{aligned} \tag{21}$$

which recovers Eq. 9 in the main text with $w_\beta(\cdot) = \beta\sigma(-\beta\cdot)$.

Second-order control (sufficient condition for strict increase). If $\nabla_\phi \Delta E_\phi$ is L-Lipschitz (equivalently, the operator norm of the Hessian of ΔE_ϕ is bounded by L on the segment between ϕ and ϕ^+), Taylor's theorem implies

$$\begin{aligned}\delta(\Delta E_\phi) &\geq \eta\, w_\beta(\Delta E_\phi)\, \|d_\phi\|_2^2 \\ &\quad - \frac{L}{2} \left\|\eta\, w_\beta(\Delta E_\phi)\, d_\phi\right\|_2^2 \\ &= \eta\, w_\beta(\Delta E_\phi)\, \|d_\phi\|_2^2 \left(1 - \frac{L\eta}{2}\, w_\beta(\Delta E_\phi)\right).\end{aligned}$$

Consequently, choosing $\eta < 2/\big(L\, w_\beta(\Delta E_\phi)\big)$ for any pair guarantees $\delta(\Delta E_\phi) > 0$.

E.4 Proof of Lemma 5.2 (Local Pairwise Optimality)

Let $f = -E_\phi$ so that larger values of f represent higher "ID-likeness", consistent with the main text. For any two samples S and S', let $z = f(S) - f(S') = E_\phi(S') - E_\phi(S)$ (Eq. 11). Denoting $\eta(S) = \Pr(\mathrm{ID} \mid S)$, we define

$$\alpha \;\triangleq\; \eta(S)\big(1 - \eta(S')\big), \qquad \alpha' \;\triangleq\; \eta(S')\big(1 - \eta(S)\big).$$

The conditional pairwise risk (Eq. 12) can then be expressed as

$$r_\beta(z; S, S') \;=\; \alpha\, \log(1 + e^{-\beta z}) \;+\; \alpha'\, \log(1 + e^{\beta z}).$$

Strict convexity. Using the property $\ell''(z) > 0$ established in Section E.2, we obtain

$$\frac{\partial^2 r_\beta}{\partial z^2} \;=\; (\alpha + \alpha')\, \beta^2\, \sigma(\beta z)\sigma(-\beta z) \;>\; 0,$$

provided that $\alpha + \alpha' > 0$. Thus, $r_\beta(z; S, S')$ is strictly convex in z for any non-degenerate pair.

Stationary point. Differentiating with respect to z and setting the derivative to zero yields

$$\frac{\partial r_\beta}{\partial z} = -\alpha\, \beta\, \sigma(-\beta z) \;+\; \alpha' \beta\, \sigma(\beta z) \;=\; 0 \quad\Longleftrightarrow\quad \alpha\, \sigma(-\beta z) \;=\; \alpha'\, \sigma(\beta z).$$

Using the identity $\sigma(\beta z)/\sigma(-\beta z) = e^{\beta z}$, we find

$$e^{\beta z^\star} \;=\; \frac{\alpha}{\alpha'} \;=\; \frac{\eta(S)\,[1 - \eta(S')]}{\eta(S')\,[1 - \eta(S)]} \quad\Longrightarrow\quad z^\star \;=\; \frac{1}{\beta} \log \frac{\eta(S)\,[1 - \eta(S')]}{\eta(S')\,[1 - \eta(S)]},$$

which matches Eq. 13. Because the mapping $u \mapsto \log(u/(1-u))$ is strictly increasing on $(0, 1)$, it follows that $\mathrm{sign}(z^\star) = \mathrm{sign}\big(\eta(S) - \eta(S')\big)$.

E.5 Proof of Proposition 5.3 (Global Bayes-Optimal Ranking)

We now prove that over a sufficiently expressive function class, any global minimizer of the population risk

$$\mathcal{R}_\beta(f) \;=\; \mathbb{E}_{(S_{\mathrm{in}}, S_{\mathrm{out}})}\big[\log\big(1 + e^{-\beta\,[\,f(S_{\mathrm{in}}) - f(S_{\mathrm{out}})\,]}\big)\big]$$

induces the exact same ranking order as $\eta(\cdot)$ for almost all pairs, thereby achieving the Bayes-optimal AUROC.

A canonical minimizer realizing point-wise optima simultaneously. Define

$$f^\star(S) \;\triangleq\; \frac{1}{\beta}\, \log \frac{\eta(S)}{1 - \eta(S)} \;\; (+ \text{ any additive constant}). \tag{22}$$

For any pair (S, S'), the margin $z^\star = f^\star(S) - f^\star(S')$ equals the two-point optimum established in Lemma 5.2. Consequently, $f^\star$ attains the minimal possible pairwise value for every term in the population risk, establishing that $f^\star$ minimizes $\mathcal{R}_\beta$. Furthermore, if g is any alternative function satisfying $\mathcal{R}_\beta(g) = \mathcal{R}_\beta(f^\star)$, then for almost every pair (S, S') we must have $g(S) - g(S') = f^\star(S) - f^\star(S')$, implying $g - f^\star$ is almost everywhere constant. Thus, the set of global minimizers is precisely $\{\, f^\star + c \mid c \in \mathbb{R} \,\}$, ensuring that all global minimizers induce the same ranking order as η.

Remarks. (i) Any strictly monotonic transformation $h \circ \eta$ preserves the ranking and achieves the Bayes-optimal AUROC. However, it does not minimize $\mathcal{R}_\beta$ unless $h(u) = \beta^{-1} \log\big(u/(1-u)\big)$ up to an additive constant, as only this form aligns all pairwise margins with $z^\star$. (ii) An additive constant leaves pairwise differences invariant, thereby affecting neither AUROC nor $\mathcal{R}_\beta$, while the ℓ_2 penalty in Eq. 18 fixes this gauge degree of freedom without altering the induced ranking.

E.6 AUROC Connection and Temperature Limit

The AUROC can be expressed as $\Pr\big(E_\phi(S_{\text{in}}) < E_\phi(S_{\text{out}})\big)$ (equivalently, $\Pr(\Delta E_\phi > 0)$), representing the ranking probability smoothed by $\ell(\cdot)$ in Eq. 16. As $\beta \to \infty$,

$$\ell(z) \;=\; \log(1 + e^{-\beta z}) \;\longrightarrow\; \begin{cases} 0, & z > 0, \\ \log 2, & z = 0, \\ +\infty, & z < 0, \end{cases}$$

showing that the loss converges to a hard 0–1 ranking penalty that heavily penalizes misordered pairs ($z < 0$).

E.7 Additional Implementation-Facing Identities

Per-pair gradient magnitude. The magnitude of the per-pair backpropagation signal is given by

$$\big\|\nabla_\phi \ell(\Delta E_\phi)\big\| \;=\; w_\beta(\Delta E_\phi)\, \|d_\phi\|,$$

where $w_\beta(t) = \beta\, \sigma(-\beta t)$ is monotonically decreasing in t. As $\Delta E_\phi \to -\infty$ (severe misranking), $w_\beta(\Delta E_\phi) \to \beta$, maximizing the gradient magnitude at $\beta\, \|d_\phi\|$. At the boundary ($\Delta E_\phi = 0$), the magnitude is $\frac{\beta}{2}\, \|d_\phi\|$, and for well-separated pairs ($\Delta E_\phi \gg 0$), it vanishes. Thus, the per-pair gradient slope is maximized for the most severely misranked instances, whereas the curvature $\ell''(z) = \beta^2 \sigma(\beta z)\sigma(-\beta z)$ peaks at $\Delta E_\phi = 0$. Both perspectives confirm that optimization effort concentrates on instances that most degrade AUROC.

Invariance to shifts and the role of ℓ_2 regularization. For any constant c, shifting E_ϕ to $E_\phi + c$ leaves ΔE_ϕ (and consequently $\mathcal{L}(\phi)$ and AUROC) completely unchanged. The ℓ_2 term in Eq. 18 removes this scale ambiguity by selecting the unique representative with minimal squared energy norm while preserving the ranking order.

Effect of the temperature parameter β. A larger β sharpens $w_\beta(t)$ toward a hard 0–1 ranking objective, accelerating the correction of borderline pairs but increasing variance if margins become excessively large due to gradient saturation on well-separated pairs. Conversely, a smaller β yields smoother update signals that are more numerically stable. In all regimes, Proposition 5.3 guarantees convergence to the same Bayes-optimal ranking at the population optimum.