SCIENTIFIC DISCOVERY UNDER VALIDATION CONGESTION VIA MULTI-FIDELITY PAIRWISE RANKINGS

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern computational methods can now propose candidate molecules, materials, and other scientific designs at an unprecedented scale, creating a *validation congestion* where candidates are abundant, but experimental capacity to physically evaluate them remains scarce. Discovering novel scientific designs has therefore become increasingly dependent on *curation*: selecting a small set of promising designs for slow and costly experiments. Existing curation methods typically rely on data-driven regression models that predict absolute scores, but training these models requires substantial experimental data to begin with. Yet, useful curation signals do not have to take the form of absolute measurements, as scientific design discovery is often comparative in nature. Here, we propose that curation can instead be primarily driven by expert *pairwise rankings*, which are substantially easier to gather. The expertise can come from computational tools or human input of multiple levels of fidelity, ranging from empirical rules of thumb to agentic workflows and experienced scientists. We introduce PRISMS, a framework that uses pairwise rankings from one or more experts, potentially spanning multiple levels of expertise, to identify the most promising candidates without relying on data-hungry regressors. When experts differ in fidelity and cost, PRISMS escalates pairwise queries from lower- to higher-fidelity rankers based on a Fisher-information criterion. In iterative screening that selects designs from fixed drug discovery libraries, PRISMS achieves 50% top-10 discovery recall in $\sim 42\%$ fewer rounds than regression-only active learning, and in $\sim 15\%$ fewer rounds than the ranking-based method with no selective escalation. In optimization that generates new designs without restriction to a predefined library, PRISMS achieves $\sim 18.8\%$ higher hypervolume than the Bayesian optimization baseline. These results show that multi-fidelity expert pairwise rankings can provide an efficient alternative to regression-driven curation in scientific design discovery.

1 INTRODUCTION

Modern computational methods can generate or enumerate molecules, materials, and other scientific designs at increasingly unprecedented scales. Virtual screening libraries, for example, routinely contain 10^5–10^9 compounds (Gryniukova et al., 2023; Zhou et al., 2024; Luttens et al., 2025), while airfoil datasets may contain hundreds of thousands of feasible designs (Liu et al., 2024). Yet, the physical validation layer, i.e., *making* (synthesis or fabrication) and *testing* (experimental measurement and safety checks) remains slow and costly. A single experimental round in drug discovery, for example, typically requires one to three months to complete (Mas et al., 2026). As a result, scientific design discovery—the iterative process of proposing and validating scientific designs to identify high-performing ones (Figure 1)—is increasingly facing *validation congestion*, where many potentially valuable designs compete for limited experimental capacity. Under validation congestion, the challenge shifts from generating yet another design to allocating scarce experimental resources toward the most promising ones, a process we call *design curation*.

The dominant computational approach for curation is to train a regression model on previously validated designs and select new candidates based on their predicted property values or associated uncertainty estimates (Gryniukova et al., 2023; Zhou et al., 2024; Luttens et al., 2025). Such approaches can be effective when sufficient training labels (experimental measurements) are available. Under validation congestion, however, the labels needed to train reliable regression models are

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.

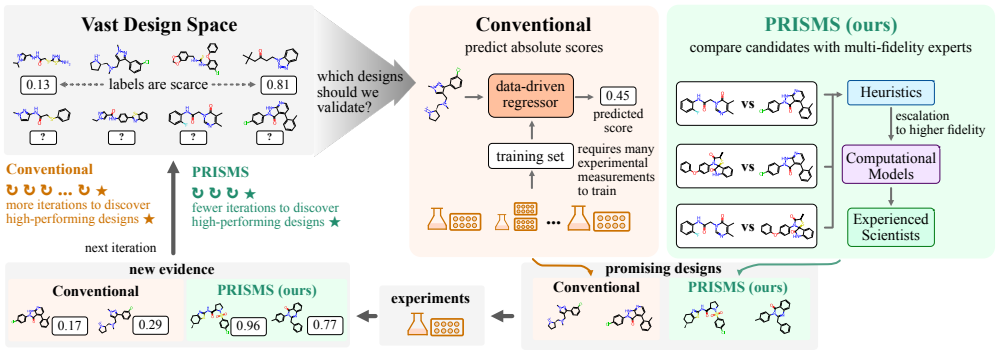

Figure 1: Accelerating scientific discovery through curation. Conventional curation methods predict absolute scores from scarce experimental data, while PRISMS curates designs using selectively queried multi-fidelity pairwise rankers. Better curation sends more promising designs for experiments at each round, reducing the number of design-experiment loops needed for discovery.

themselves expensive to acquire. This motivates complementary sources of information that can guide curation before sufficient absolute measurements have been collected.

When we reconsider the problem from a broader perspective, scientific design discovery is often comparative in nature. The goal is not to predict every design's exact property value, but to identify which designs are better: whether molecule A inhibits a target more strongly than molecule B, whether material A is stronger than material B, or whether airfoil A has a higher lift coefficient than airfoil B. Useful curation signals, therefore, do not have to take the form of absolute measurements. They may instead take the form of *pairwise rankings*: comparative judgments of whether one design is more promising than another. Such rankings can be obtained from heterogeneous sources of scientific knowledge, including empirical rules of thumb (Lipinski et al., 1997; Mizutani, 2010; George et al., 2020), computational models (Neese, 2012; Trott & Olson, 2010; Zeni et al., 2025), and domain experts, spanning different levels of fidelity and cost. Inexpensive heuristics may provide abundant but noisy rankings, while experienced scientists may provide more reliable judgments despite being unable to precisely quantify the absolute property values. Pairwise rankings thus provide a natural interface for integrating heterogeneous scientific knowledge and are often easier to elicit reliably than absolute property estimates (Thurstone, 1927; Stewart et al., 2005; Routh et al., 2023; Hoeijmakers et al., 2024).

To turn multi-fidelity pairwise rankings into curation decisions, we introduce Pairwise Ranking for Information-guided Scientific curation with Multi-fidelity Sources (PRISMS), a graph-based curation framework that aggregates pairwise comparisons across rankers and selectively queries higher-fidelity rankers. In PRISMS, nodes represent candidate designs and store either measured property values or current property estimates, while edges encode pairwise rankings accumulated across rankers and campaign rounds. The values of unmeasured designs are inferred via a regularized Bradley-Terry model (Bradley & Terry, 1952), anchored by available measurements and prior property estimates. Within each round, PRISMS queries rankers sequentially in increasing order of fidelity. Since higher-fidelity rankers often have limited availability, PRISMS selectively escalates comparisons based on a Fisher information-based criterion (Fisher, 1925), concentrating higher-fidelity queries on comparisons that can improve the property estimates the most. The resulting estimates are then used to curate the set of promising designs that merit experimental validation.

We evaluate PRISMS on 14 drug-discovery and three engineering datasets in two settings: (1) *screening*, which identifies top-performing designs from a fixed library, and (2) *optimization*, which generates new designs and searches for high-performing ones without a predefined library. In screening, PRISMS reaches 50% top-10 discovery recall in $\sim 42\%$ fewer iterations than conventional pool-based active learning. In optimization, it achieves $\sim 18.8\%$ higher hypervolume than GP-BO, a Bayesian optimization baseline. These results demonstrate that multi-fidelity pairwise rankings can effectively guide curation when experimental measurements are scarce both at initialization and throughout the campaign, and accelerate scientific design discovery by reducing the number of required experimental rounds.

2 BACKGROUND

Scientific discovery may involve either understanding existing systems or identifying new designs with desirable properties (Wang et al., 2023). The former includes discovering scientific relationships, mechanisms, or hypotheses. The latter, which we call scientific design discovery and focus on in this paper, concerns identifying new drugs, materials, or engineered objects that satisfy an objective. Scientific design discovery spans various domains from drug discovery to engineering (Reddy et al., 2007; Lam et al., 2018; Mathern et al., 2021). As computational candidate design generation becomes increasingly scalable while physical validation remains slow and costly (Mas et al., 2026), the bottleneck of scientific design discovery shifts to deciding which designs to validate, a process we call curation.

A common curation approach is to use surrogate models to predict candidate designs' property values, reducing reliance on experimental validation. When the underlying physics is sufficiently well understood and computationally tractable, physics-based models can provide such surrogates, as in molecular docking (Tuccinardi, 2009; Kontoyianni, 2017). In many scientific domains, however, such models are unavailable, motivating the widespread use of data-driven surrogates (Gryniukova et al., 2023; Zhou et al., 2024; Luttens et al., 2025; Ansari et al., 2022; Wijaya et al., 2025). For example, in virtual screening, uncertainty-aware surrogates such as Gaussian Processes (Williams & Rasmussen, 1995) can be combined with acquisition functions such as Upper Confidence Bound (UCB) (Auer et al., 2002) or Expected Improvement (EI) (Jones et al., 1998) to determine which designs to validate next. These approaches, however, rely primarily on absolute property measurements as supervision, which are themselves scarce under validation congestion.

Pairwise rankings provide an alternative source of supervision. Rather than requiring the absolute property value of a design, a ranker only determines whether one design is preferable to another. Pairwise ranking methods can recover latent scores, top-k partitions, or global orderings from such comparisons. Spectral MLE (Chen & Suh, 2015) estimates global scores under the Bradley-Terry model (Bradley & Terry, 1952), while HodgeRank (Jiang et al., 2011) uses Hodge decomposition to separate globally consistent ranking signals from cyclic inconsistencies. Mohajer et al. (2017) actively select comparisons near the top-k decision boundary to reduce the number of queries needed to identify the top-k partition. More recently, Qin et al. (2024) show that LLMs can serve as pairwise rankers, identifying top-ranked items through sequential pairwise rankings. For scientific design curation, pairwise rankings can also be used to improve estimates on an absolute property scale, allowing rankings to be combined with experimentally measured values. Projection (Yan et al., 2024) and RankRefine++ (Wijaya et al., 2026) follow this direction by using pairwise rankings to refine predictions from a learned regressor.

3 DESIGN CURATION WITH MULTI-FIDELITY PAIRWISE RANKINGS

We consider scientific discovery campaigns under validation congestion, where advances in in generative models and growth of large scientific design libraries make candidate designs abundant, while physical validation capacity remains limited. A campaign covers *multiple* rounds of selecting and experimentally validating a small batch of designs, and the goal of curation is to reduce the number of rounds needed to discover high-performing designs. Candidate designs come from a fixed library in iterative screening campaigns or are generated throughout an optimization campaign.

The overview of PRISMS is illustrated in Figure 2. At the start of round t, let $\mathbb{X}_{t-1}^V$ denote the set of experimentally validated designs from the previous rounds, and $\mathbb{X}_t^U$ the unmeasured designs available for curation. Each design x_i has a property value y_i that can be revealed through experimental validation, e.g., a drug candidate's bioactivity measured by an assay. Given a per-round validation budget b^{val}, curation selects a small batch $\mathbb{B}_t \subseteq \mathbb{X}_t^U$ with $|\mathbb{B}_t| \leq b^{\text{val}}$. The experimental validation layer then reveals the property values of the selected designs, and updates the set $\mathbb{X}_t^V = \mathbb{X}_{t-1}^V \cup \mathbb{B}_t$.

In each campaign round, PRISMS jointly estimates the property values of the unmeasured designs in $\mathbb{X}_t^U$ based on the measurements of validated designs in $\mathbb{X}_{t-1}^V$ and the pairwise rankings collected from multi-fidelity pairwise rankers. The property estimates therefore provide a common interface for combining multi-fidelity pairwise rankings with experimental measurements, which ultimately drives the curation process.

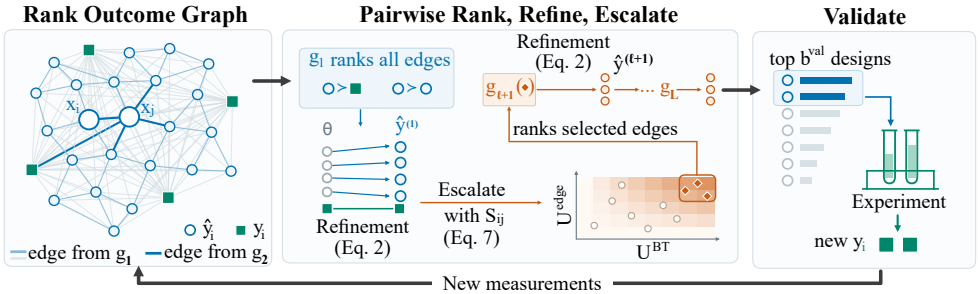

Figure 2: Overview of PRISMS. **(Left)** A rank outcome graph connects unmeasured designs (circles) to all experimentally validated designs (squares) and some unmeasured designs. **(Middle)** Starting with low-fidelity rankings on all eligible edges, a regularized Bradley-Terry model jointly refines property estimates of all unmeasured designs. Select pairs (edges) are escalated to higher-fidelity rankers based on a Fisher information criterion, and the estimates are further refined after exhausting the budget of a ranker g_ℓ. **(Right)** The highest-ranked designs are selected for experimental validation, subject to a budget. New measurements update the rank outcome graph for the next round.

3.1 Joint Property Estimation from Pairwise Rankings

We assume access to pairwise rankers $g_1, \ldots, g_L$ indexed in increasing level of fidelity. Each ranker takes two designs as input and predicts which has the higher property value. Lower-fidelity rankers are assumed to be cheaper to query, whereas higher-fidelity rankers provide more reliable rankings at a limited query budget. Their judgments may nevertheless be noisy and mutually inconsistent.

Within each campaign round, we query the rankers in L stages of increasing fidelity and store their outcomes in a graph where nodes represent designs and edges represent the pairwise ranking outcomes. An experimentally validated node carries its ground-truth property value y_i, while an unmeasured node carries its estimate $\hat{y}_i$. At the beginning of each round, we connect each unmeasured design to all experimentally validated designs and the k^U other unmeasured designs closest in prior property estimate θ_i. The connections remain fixed within a round and are recomputed at the start of the next round. Only nodes connected by edges are eligible to be ranked by the pairwise rankers.

Omitting the campaign-round index t for clarity, let $\mathbb{D}^{(\ell)} = \{r_{ij}^{(\ell)}\}$ denote the set of pairwise rankings obtained from g_ℓ, where $r_{ij}^{(\ell)} \in \{0, 1\}$ is the binary outcome and $r_{ij}^{(\ell)} = 1$ indicates that g_ℓ judges x_i to have a higher property value than x_j. Let z_i denote the property value assigned to design x_i during the optimization to estimate $\hat{y}_i$. Since experimentally validated nodes have their ground-truth values revealed, we keep their values fixed to $z_i = y_i$ for all $x_i \in \mathbb{X}^V$. Under the Bradley-Terry model (Bradley & Terry, 1952), the probability of $x_i \succ x_j$ is

$$p_{ij}(z) = \sigma\left((z_i - z_j)/\tau\right), \qquad \sigma(u) = \frac{1}{1 + \exp(-u)}, \tag{1}$$

where the temperature τ can be fitted from a small sample of measured designs (Wijaya et al., 2026). At stage ℓ, we jointly estimate the unmeasured values by minimizing

$$\mathcal{L}^{(\ell)}(z) = \sum_{s=1}^{\ell} \lambda_1^{(s)} \sum_{r_{ij}^{(s)} \in \mathbb{D}^{(s)}} \mathrm{BCE}\left(r_{ij}^{(s)}, p_{ij}(z)\right) + \lambda_2 \sum_{x_i \in \mathbb{X}^U} (z_i - \theta_i)^2, \tag{2}$$

where $\mathrm{BCE}(r, p) = -r \log p - (1 - r) \log(1 - p)$, $\lambda_1^{(\ell)}$ controls the contribution of ranker g_ℓ, and θ_i is a prior estimate of the property value with regularization strength $\lambda_2 \geq 0$. Minimizing this objective at stage ℓ yields an estimate $\hat{y}_i^{(\ell)}$ for each unmeasured design. After the final stage L, we select b^{val} unmeasured designs with the highest estimates $\hat{y}_i^{(L)}$ for experimental validation. Note that the prior estimates may be obtained from any estimator that predicts the numerical values of the property, for example, a random forest regressor trained on available measurements at the current round. Importantly, PRISMS uses θ_i only as a weak prior that is subsequently refined using the accumulated pairwise rankings.

3.2 INFORMATION-GUIDED ESCALATION TO HIGHER-FIDELITY PAIRWISE RANKER

We assume that the lowest-fidelity ranker g_1 can evaluate all eligible pairs in the comparison graph, producing the rankings $\mathbb{D}^{(1)}$. However, higher-fidelity rankers often have limited availability, such that at stage $\ell > 1$, ranker g_ℓ can evaluate at most $b^{(\ell)}$ pairs. To allocate this budget, we prioritize comparisons expected to reduce uncertainty in candidate property estimates. Our escalation criterion combines two measures, *outcome uncertainty* and *edge uncertainty*, defined below.

Outcome Uncertainty. Outcome uncertainty is the variance of the binary ranking outcome under the Bradley-Terry model. After stage ℓ, it is

$$U_{ij}^{\mathrm{BT}} = p_{ij}(\hat{y}^{(\ell)})(1 - p_{ij}(\hat{y}^{(\ell)})). \tag{3}$$

The outcome uncertainty is maximized when the two designs have equal estimated property values.

Edge Uncertainty. Edge uncertainty is the variance of the estimated property difference between two designs, connected by an edge, under a local Gaussian approximation. We calculate edge uncertainty by using the curvature of the objective in Equation 2 at the current estimates:

$$H^{(\ell)} = \nabla^2 \mathcal{L}^{(\ell)}(z)\Big|_{z=\hat{y}^{(\ell)}}, \tag{4}$$

where the Hessian is taken with respect to unmeasured values, holding measured values fixed. When $\lambda_2 > 0$, $H^{(\ell)}$ is positive definite, and its inverse gives the covariance of the local Gaussian approximation induced by the objective.

Let $\boldsymbol{v}_{ij}$ be an incidence vector restricted to the unmeasured variables, containing $+1$ for x_i if $x_i \in \mathbb{X}^U$, -1 for x_j if $x_j \in \mathbb{X}^U$, and zeros elsewhere. We define edge uncertainty as the variance of the property difference under the approximation,

$$U_{ij}^{\mathrm{edge}} = \boldsymbol{v}_{ij}^{\mathsf{T}} (H^{(\ell)})^{-1} \boldsymbol{v}_{ij}. \tag{5}$$

High edge uncertainty indicates that the property difference of a pair is poorly constrained under the local approximation, making the pair a potential target for more accurate ranking information.

Information-guided Escalation. For $\ell < L$, we evaluate the contribution of a new ranking from $g_{\ell+1}$ for pair (x_i, x_j) at the current estimates $\hat{y}^{(\ell)}$. The additional weighted Bradley-Terry term yields the following rank-one update to the local precision:

$$H' = H^{(\ell)} + \frac{\lambda_1^{(\ell+1)}}{\tau^2} U_{ij}^{\mathrm{BT}} \boldsymbol{v}_{ij} \boldsymbol{v}_{ij}^{\mathsf{T}}. \tag{6}$$

By the matrix determinant lemma, the resulting increase in log-determinant is

$$\begin{aligned} S_{ij} &= \log\det H' - \log\det H^{(\ell)} \\ &= \log\left(1 + \frac{\lambda_1^{(\ell+1)}}{\tau^2} U_{ij}^{\mathrm{BT}} U_{ij}^{\mathrm{edge}}\right). \end{aligned} \tag{7}$$

This local information-gain score, derived from the Bradley-Terry Fisher information, favors pairs with both high outcome uncertainty and high edge uncertainty. We select up to $b^{(\ell+1)}$ eligible pairs with the highest individual information-gain scores, all evaluated using the current precision matrix, and query ranker $g_{\ell+1}$ on these pairs. The returned rankings form $\mathbb{D}^{(\ell+1)}$, and the property estimates are updated by minimizing Equation 2. This process continues up to the last ranker at stage L, after which the final estimates are used to determine which designs are sent to the experimental validation layer.

3.3 RANKREFINE++ AS A SPECIAL CASE

RankRefine++ (Wijaya et al., 2026), the closest work to PRISMS, refines a pretrained regressor's predictions using pairwise comparisons against experimentally validated designs. Its per-design estimation objective is recovered as a special case of Equation 2 by considering only a single unmeasured design x_0, connected only to measured designs, and using only a single ranker g_1.

Let θ_0 be the regressor's prediction and σ^2_{re} its associated uncertainty. For $\lambda_1^{(1)} > 0$, setting $\frac{\lambda_2}{\lambda_1^{(1)}} = \frac{1}{2\sigma^2_{\text{re}}}$ makes Equation 2 proportional to the per-design MAP objective of RankRefine++, yielding the same estimate.

4 Experimental Results

4.1 Drug Discovery Campaign

Setup. We evaluate PRISMS and competing curation methods in simulated iterative screening campaigns that aim to identify the top-10 molecules in a fixed library using as few experimental rounds as possible. We use 10 ADMET tasks (Ainsley et al., 2026; MacDermott-Opeskin, 2026) and 4 binding-affinity tasks (Gorantla et al., 2024), detailed in Appendix Table 3, and each experiment is repeated over 30 random seeds. For each task-seed pair u, where $n^{(u)}_{\text{mol}}$ is the size of the molecular dataset used in u, we construct a library of $\min(n^{(u)}_{\text{mol}}, 1010)$ randomly sampled molecules, reveal the labels of 10 molecules to initialize $\mathbb{X}_0^V$, and treat the remainder as unlabeled. At each round, the curation selects 10 unlabeled molecules for experimental validation, simulated by revealing their labels.

Unless otherwise stated, we use two pairwise rankers of different fidelities. The lower-fidelity ranker g_1 is an ensemble of a rule-based classifier written by Claude Sonnet 4.6, a random forest (Breiman, 2001), and a support vector machine (Cortes & Vapnik, 1995); at each round t, the latter two are refitted on pairwise rankings derived from $\mathbb{X}_{t-1}^V$. On a separate 24-round drug-discovery campaign evaluation, g_1's pairwise accuracy increases from approximately 60% in round 1 to 68% by round 9 and remains stable afterwards. The higher-fidelity ranker g_2 is a simulated 75%-accurate oracle, implemented by flipping ground-truth rankings with probability 25%, with a budget of $b^{(2)} = 100$ queries per round. We initialize θ_i using a random forest regressor to build the initial unlabeled-unlabeled node connections, and update θ_i with a Bradley–Terry estimate obtained from g_1 rankings between the query and labeled nodes.

We compare PRISMS with seven baselines described in Section 2: random sampling, a regression-based Gaussian Process with Expected Improvement algorithm (GP-EI), a pairwise-ranking based GP-EI, HodgeRank, Active Top-k, Spectral MLE, and RankRefine++. All baselines except random sampling and GP-EI have access to the same multi-fidelity rankers and per-round query budgets as PRISMS. However, they allocate higher-fidelity queries to randomly selected pairs, without information-based escalation. We evaluate discovery efficiency and ranking quality using four metrics: (1) *discovery recall*: the fraction of true top-10 molecules cumulatively discovered at round t; (2) *NDCG*: the normalized discounted cumulative gain over the current candidate pool; (3) *top-k% F1*, the F1 score for identifying molecules belonging to the true top-$k\%$ of the library; and (4) *restricted mean time to 50% discovery recall*: the area under the Kaplan-Meier survival curve (Kaplan & Meier, 1958) for the time to discover five of the true top-10 molecules. We report discovery and ranking performance over the first 12 rounds, corresponding to approximately 1–3 years of experiments under the assumed duration of 1–3 months per round (Mas et al., 2026). The full 24-round campaigns are used only to estimate the restricted mean time to 50% discovery recall.

Results. Figure 3 summarizes the 12-round iterative screening campaigns across datasets and runs. PRISMS achieves the highest top-10 discovery recall curve throughout all rounds, improving area under the top-10 discovery recall curve (AUDC@10) by 25.1% over HodgeRank and 24.3% over RankRefine++. PRISMS also improves the top-10% F1 score, where PRISMS outperforms RankRefine++ by 31.2%. Because RankRefine++ is a special case of PRISMS (Section 3.3), the gains indicate the benefits of connecting unlabeled nodes, joint optimization, and information-based escalation under equal access to multi-fidelity pairwise rankers. More importantly, PRISMS achieves a 24-round restricted mean time to 50% discovery recall of 11.1 rounds, compared with conventional Bayesian approaches such as GP-EI (19.3 rounds) that do not use pairwise rankers. Under validation congestion, assuming 1–3 months per experimental round, this translates to an estimated 8–24 months of experimental cycle time saved. Finally, GP-EI Pairwise achieves $\sim 61\%$ higher AUDC@10 than GP-EI, highlighting the value of pairwise ranking supervisions when numerical labels are scarce. Detailed results are shown in Table 1, along with a scaling analysis in Section B.1 and a proof of concept for real-world applications with non-simulated rankers is shown in Appendix Section B.3.

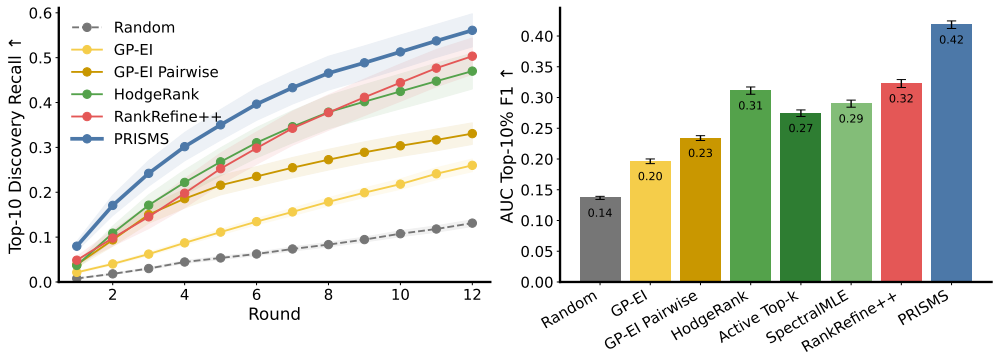

Figure 3: Aggregated results of 12-round scientific design discovery campaigns on 14 drug discovery datasets (ADMET and binding affinity). We report the mean and standard error over 30 random seeds. **Left.** Top-10 discovery recall over 12 rounds. **Right.** AUC of Top-10% F1 score.

4.2 Engineering Discovery Campaign

Setup. We evaluate PRISMS for iterative screening beyond drug discovery on three engineering datasets: 3D-printed structures (mechanical toughness) (Gongora et al., 2020), dye laser configurations (peak score), and battery redoxmers (absorption wavelength deviation) (Häse et al., 2021). We use 10 random seeds per dataset and otherwise follow the setup in Section 4.1.

Results. Figure 10 in the Appendix shows the aggregated results of the 12-round campaigns for the engineering datasets. Consistent with the findings in the drug discovery domain, PRISMS outperforms all evaluated baselines across all performance metrics.

4.3 Multi-Objective Drug Optimization Campaign

Setup. Beyond screening, we evaluate PRISMS in multi-objective molecular optimization targeting oracle-evaluated JNK3 activity (Li et al., 2018) and solubility (Swanson et al., 2024). Unlike fixed-library screening, candidates are generated throughout the campaign using molecular graph-based crossover and mutation operators (Jensen, 2019) in a genetic algorithm-style loop. Each curation method independently predicts both objectives for newly generated molecules to select the promising candidates.

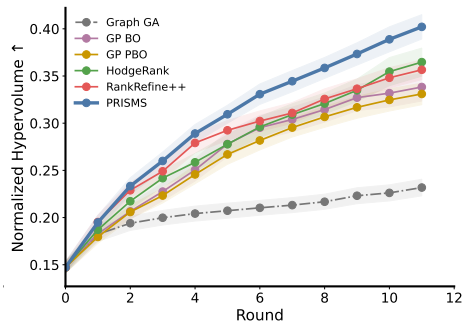

Figure 4: Normalized hypervolume over 12 campaign rounds and 30 random seeds for multi-objective optimization targeting JNK3 activity and solubility

We initialize each campaign with 10 randomly sampled molecules from ZINC 250K (Sterling & Irwin, 2015; Gómez-Bombarelli et al., 2018) and allow 10 designs per round for oracle evaluation. The baselines are RankRefine++, HodgeRank, Graph GA (Jensen, 2019), GP-BO (Tripp & Hernández-Lobato, 2024) and its preferential Bayesian optimization (González et al., 2017) extension, GP-PBO. Performance is measured using the hypervolume of the oracle-validated designs, normalized to the maximum in ZINC250K.

Results. Figure 4 shows the normalized hypervolume across the campaign rounds. PRISMS consistently outperforms all other baselines, especially in later rounds when pairwise evidences have accumulated. At the end of a 12-round campaign, PRISMS achieves about 14.3% higher hypervolume than HodgeRank and RankRefine++, suggesting that pairwise-ranking-based curation remains effective even when the candidate design set grows over time.

4.4 Ablation Study

Setup. We ablate the main components of PRISMS using a similar setup as described in Section 4.1, run for 24 rounds. Starting from the default two-fidelity configuration, we separately remove graph edges between two unlabeled nodes, edge uncertainty, and outcome uncertainty; replace information-guided escalation with random escalation; and randomize the Stage-1 initialization. We additionally vary the number of fidelity levels, comparing one-, two-, and three-fidelity configurations. For the three-fidelity setup, we set g_3 to 90% accuracy with a budget $b^{(3)} = 50$.

Results. Table 2 in the Appendix shows that PRISMS' discovery gains come from both access to additional pairwise rankings and how those rankings are propagated and acquired. Removing graph edges between unlabeled nodes, which prevents ranking information from propagating directly between candidate designs, reduces AUDC@10 by 7.6%. Within the escalation criterion (Section 3.2), removing edge or outcome uncertainty decreases AUDC@10 by 4.6% and 2.3%, respectively, while replacing both with random escalation yields a 7.7% reduction, showing that the two uncertainty terms provide complementary signals for allocating higher-fidelity queries. The number of fidelity stages has the largest effect: restricting PRISMS to a single stage lowers AUDC@10 by 16.5%, whereas introducing a third fidelity stage improves AUDC@10 by 11.4% over the default two-stage configuration. These results suggest that, under validation congestion, effectively exploiting external sources of knowledge can substantially improve discovery performance.

4.5 Sensitivity Analysis

Setup. We study the sensitivity of PRISMS to the accuracy, availability, and error structure of the higher-fidelity ranker, in a similar setup as described in Section 4.1. The pairwise accuracy and per-round query budget $b^{(2)}$ of g_2 are varied from 60% to 90% and from 10 to 200, respectively. We consider three noise structures: *uniform flip noise*, which flips pairwise outcomes with uniform probability; *difficulty-dependent noise*, where error probability follows the sigmoid of the difference in property values between the compared designs (designs with similar values are harder to rank accurately); and *structured noise*, which introduces scaffold-dependent biases. We report PRISMS' AUDC@10 and its difference to RankRefine++'s under matched ranker settings and budgets.

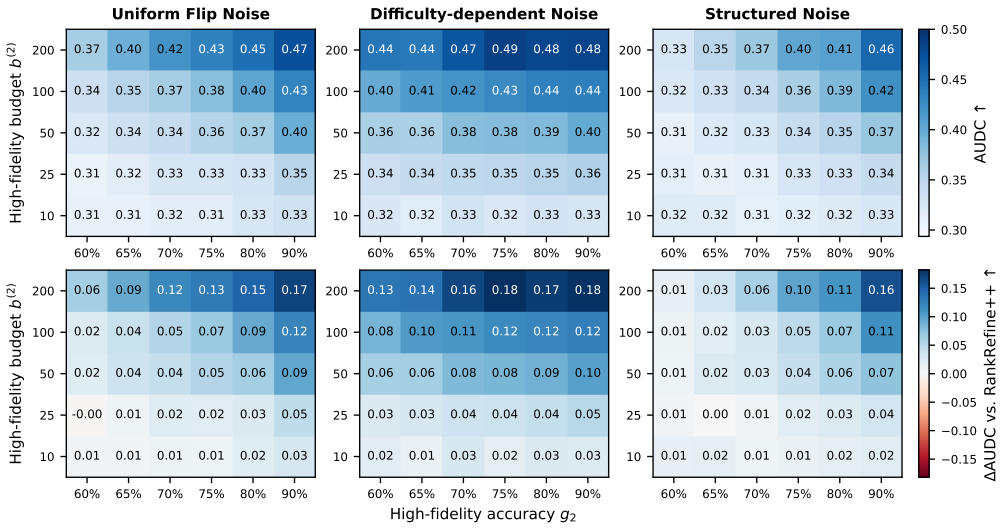

Figure 5: Sensitivity to higher-fidelity ranker accuracy and query budget under three ranking-noise structures. **(Top)**: PRISMS' AUDC@10. **(Bottom)**: AUDC@10 differences between PRISMS and RankRefine++ under matched settings.

Results. Figure 5 shows that PRISMS increasing the higher-fidelity query budget consistently improves AUDC@10 across all three noise models. Increasing ranker accuracy has a modest effect when the query budget is small, but yields increasingly larger gains as more higher-fidelity rankings become available. PRISMS also achieves higher AUDC@10 than RankRefine++ in nearly all

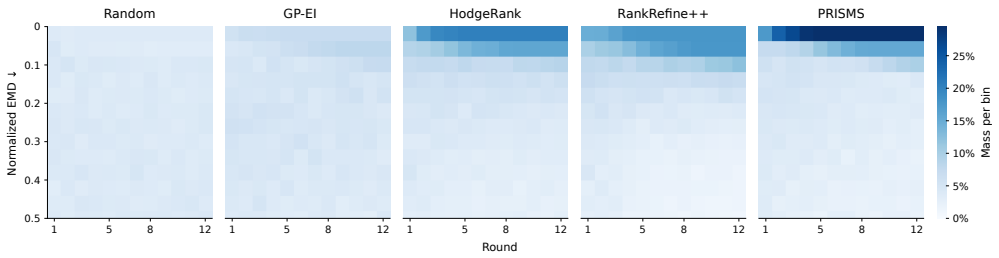

Figure 6: Recovery of true top-10 designs across campaign rounds. Heatmaps show the distribution of normalized Earth Mover's Distance (Rubner et al., 1998) for top-10 designs ($\mathrm{EMD}_{\mathrm{norm}}@10$, lower is better) over 12 rounds across 14 drug-discovery datasets and 30 random seeds. An $\mathrm{EMD}_{\mathrm{norm}}@10 = 0.0$ places all true top-10 designs in the top-10 positions, whereas $\mathrm{EMD}_{\mathrm{norm}}@10 = 1.0$ places them in the bottom-10 positions. Across rounds, PRISMS shifts true top-10 designs toward the highest ranks faster than competing methods.

evaluated settings, with differences generally widening as the budget increases. This indicates that PRISMS makes more effective use of additional ranking information than RankRefine++ Difficulty-dependent noise yields the strongest overall performance and largest gains over RankRefine++, supporting the relevance of Agentic PRISMS with an LLM as a ranker, whose errors exhibit a similar structure as discussed in Appendix Section B.3. More results for sensitivity analysis are available in Appendix B.2.

4.6 Recovery of Missed Top-10 Designs

Setup. We further examine the capability to recover high-performing designs that are not selected for experimental validation in earlier iterative screening rounds. Such early-round failures may arise especially when the initial labels or pairwise rankers are biased. We quantify the recovery using the Earth Mover's Distance (EMD) (Rubner et al., 1998), which measures how far the true top-k designs are from occupying the top-k positions at round t. For a fixed library of size N, let $r_1 < \cdots < r_k$ denote the positions occupied by the true top-k designs in the predicted ranking. The target is for these designs to occupy positions $1, \ldots, k$ in any order. We define the normalized EMD as

$$\mathrm{EMD}_{\mathrm{norm}}@k = \frac{1}{k(N-k)} \sum_{i=1}^{k} (r_i - i). \tag{8}$$

Results. Figure 6 shows that PRISMS progressively shifts true top-10 designs toward top-10 positions more quickly than competing methods, as suggested by the growing concentration of samples near $\mathrm{EMD}_{\mathrm{norm}}@10 = 0$. GP-EI remains close to random sampling, while HodgeRank and RankRefine++ reduce $\mathrm{EMD}_{\mathrm{norm}}@10$ but retain more spread distributions. Overall, PRISMS outperforms other methods at keeping missed high-performing designs near the top of the ranking, increasing their likelihood of selection in subsequent rounds.

5 Conclusion

Scientific design discovery increasingly faces *validation congestion*, where candidate designs are abundant but physical validation capacity remains scarce. We have introduced PRISMS to address this bottleneck by shifting *design curation* away from data-hungry regression toward multi-fidelity pairwise rankings, which provide useful supervision when experimental measurements are scarce and are often easier to elicit reliably. PRISMS leverages heterogeneous sources of scientific knowledge by integrating rankings from diverse experts and selectively escalating to higher-fidelity rankers. Across fixed-library screening and optimization, this approach improves both the rate and extent of discovery of high-performing designs. More broadly, our results suggest that accelerating scientific design discovery under validation congestion does not have to rely solely on acquiring more measurements, but can also come from making better use of external scientific knowledge that is already available.

AI USE STATEMENT

In this work, we used generative AI tools to assist us in implementing functions, identifying relevant literature, and improving readability of sentences. We have not used generative AI tools for generating synthetic data sets, helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, assisting translation, cleaning or reformatting datasets, supporting qualitative and thematic data analysis, or interpreting results. We have reviewed and verified all AI-assisted work including LLM-generated codes and text edits.

REFERENCES

Jon Ainsley, Good Andrew, Bourque Elyse, Vogeti Lakshminarayana, renato skerlj, Wang Tiansheng, Ledeboer Mark, W. Patrick Walters, Maria Castellanos, and Hugo MacDermott-Opeskin. Openadmet-expansionrx blind challenge data, July 2026. URL https://doi.org/10.5281/zenodo.21504733.

Navid Ansari, Hans-Peter Seidel, Nima Vahidi Ferdowsi, and Vahid Babaei. Autoinverse: Uncertainty aware inversion of neural networks. *Advances in Neural Information Processing Systems*, 35: 8675–8686, 2022.

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2):235–256, 2002.

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.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

Yuxin Chen and Changho Suh. Spectral mle: Top-k rank aggregation from pairwise comparisons. In *International conference on machine learning*, pp. 371–380. PMLR, 2015.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

Ronald Aylmer Fisher. Theory of statistical estimation. In *Mathematical proceedings of the Cambridge philosophical society*, volume 22, pp. 700–725. Cambridge University Press, 1925.

Janine George, David Waroquiers, Davide Di Stefano, Guido Petretto, Gian-Marco Rignanese, and Geoffroy Hautier. The limited predictive power of the pauling rules. *Angewandte Chemie*, 132 (19):7639–7645, 2020.

Rafael Gómez-Bombarelli, Jennifer N Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D Hirzel, Ryan P Adams, and Alán Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. *ACS central science*, 4(2):268, 2018.

Aldair E Gongora, Bowen Xu, Wyatt Perry, Chika Okoye, Patrick Riley, Kristofer G Reyes, Elise F Morgan, and Keith A Brown. A bayesian experimental autonomous researcher for mechanical design. *Science advances*, 6(15):eaaz1708, 2020.

Javier González, Zhenwen Dai, Andreas Damianou, and Neil D Lawrence. Preferential bayesian optimization. In *International Conference on Machine Learning*, pp. 1282–1291. PMLR, 2017.

Rohan Gorantla, Alzbeta Kubincova, Benjamin Suutari, Benjamin P Cossins, and Antonia SJS Mey. Benchmarking active learning protocols for ligand-binding affinity prediction. *Journal of Chemical Information and Modeling*, 64(6):1955–1965, 2024.

Anastasiia Gryniukova, Florian Kaiser, Iryna Myziuk, Diana Alieksieieva, Christoph Leberecht, Peter P Heym, Olga O Tarkhanova, Yurii S Moroz, Petro Borysko, and V Joachim Haupt. Ai-powered virtual screening of large compound libraries leads to the discovery of novel inhibitors of sirtuin-1. *Journal of Medicinal Chemistry*, 66(15):10241–10251, 2023.

Florian Häse, Matteo Aldeghi, Riley J. Hickman, Loïc M. Roch, Melodie Christensen, Elena Liles, Jason E. Hein, and Alán Aspuru-Guzik. Olympus: a benchmarking framework for noisy optimization and experiment planning. *Machine Learning: Science and Technology*, 2(3):035021, July 2021. ISSN 2632-2153. doi: 10.1088/2632-2153/abedc8.

Eva JI Hoeijmakers, Bibi Martens, Babs MF Hendriks, Casper Mihl, Razvan L Miclea, Walter H Backes, Joachim E Wildberger, Frank M Zijta, Hester A Gietema, Patricia J Nelemans, et al. How subjective ct image quality assessment becomes surprisingly reliable: pairwise comparisons instead of likert scale. *European Radiology*, 34(7):4494–4503, 2024.

Jan H Jensen. A graph-based genetic algorithm and generative model/monte carlo tree search for the exploration of chemical space. *Chemical science*, 10(12):3567–3572, 2019.

Xiaoye Jiang, Lek-Heng Lim, Yuan Yao, and Yinyu Ye. Statistical ranking and combinatorial hodge theory. *Mathematical Programming*, 127(1):203–244, 2011.

Donald R Jones, Matthias Schonlau, and William J Welch. Efficient global optimization of expensive black-box functions. *Journal of Global optimization*, 13(4):455–492, 1998.

Edward L Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481, 1958.

Maria Kontoyianni. Docking and virtual screening in drug discovery. In *Proteomics for drug discovery: Methods and protocols*, pp. 255–266. Springer, 2017.

Rémi Lam, Matthias Poloczek, Peter Frazier, and Karen E Willcox. Advances in bayesian optimization with applications in aerospace engineering. In *2018 AIAA Non-Deterministic Approaches Conference*, pp. 1656, 2018.

Yibo Li, Liangren Zhang, and Zhenming Liu. Multi-objective de novo drug design with conditional graph generative model. *Journal of Cheminformatics*, 10(1):33, Jul 2018. doi: 10.1186/s13321-018-0287-6.

Christopher A Lipinski, Franco Lombardo, Beryl W Dominy, and Paul J Feeney. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced drug delivery reviews*, 23(1-3):3–25, 1997.

Jian Liu, Jianyu Wu, Hairun Xie, Guoqing Zhang, Jing Wang, Wei Liu, Wanli Ouyang, Junjun Jiang, Xianming Liu, Shixiang Tang, et al. Afbench: A large-scale benchmark for airfoil design. *Advances in Neural Information Processing Systems*, 37:82757–82780, 2024.

Andreas Luttens, Israel Cabeza de Vaca, Leonard Sparring, José Brea, Antón Leandro Martínez, Nour Aldin Kahlous, Dmytro S Radchenko, Yurii S Moroz, María Isabel Loza, Ulf Norinder, et al. Rapid traversal of vast chemical space using machine learning-guided docking screens. *Nature Computational Science*, 5(4):301–312, 2025.

Hugo MacDermott-Opeskin. Openadmet's cyp challenge is underway, August 2026. URL https://doi.org/10.5281/zenodo.21987121.

Pablo Mas, Bruno Filoche-Rommé, Marc Bianciotto, and Rodolphe Vuilleumier. Back to the future of lead optimization: benchmarking compound prioritization strategies. *Digital Discovery*, 5(7): 2983–2997, 2026.

Alexandre Mathern, Olof Skogby Steinholtz, Anders Sjöberg, Magnus Önnheim, Kristine Ek, Rasmus Rempling, Emil Gustavsson, and Mats Jirstrand. Multi-objective constrained bayesian optimization for structural design. *Structural and Multidisciplinary Optimization*, 63(2):689–701, 2021.

Uichiro Mizutani. The hume-rothery rules for structurally complex alloy phases. In *Surface properties and engineering of complex intermetallics*, pp. 323–399. World Scientific, 2010.

Soheil Mohajer, Changho Suh, and Adel Elmahdy. Active learning for top-k rank aggregation from noisy comparisons. In *International Conference on Machine Learning*, pp. 2488–2497. PMLR, 2017.

F. Neese. The orca program system. *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 2(1):73–78, 2012. doi: 10.1002/wcms.81.

Zhen Qin, Rolf Jagerman, Kai Hui, Honglei Zhuang, Junru Wu, Le Yan, Jiaming Shen, Tianqi Liu, Jialu Liu, Donald Metzler, et al. Large language models are effective text rankers with pairwise ranking prompting. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 1504–1518, 2024.

A Srinivas Reddy, S Priyadarshini Pati, P Praveen Kumar, HN Pradeep, and G Narahari Sastry. Virtual screening in drug discovery-a computational perspective. *Current protein and peptide science*, 8 (4):329–351, 2007.

Jennifer Routh, Sharmini Julita Paramasivam, Peter Cockcroft, Sarah Wood, John Remnant, Cornélie Westermann, Alison Reid, Patricia Pawson, Sheena Warman, Vishna Devi Nadarajah, et al. Rating and ranking preparedness characteristics important for veterinary workplace clinical training: a novel application of pairwise comparisons and the elo algorithm. *Frontiers in Medicine*, 10: 1128058, 2023.

Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. A metric for distributions with applications to image databases. In *Sixth international conference on computer vision (IEEE Cat. No. 98CH36271)*, pp. 59–66. IEEE, 1998.

Teague Sterling and John J Irwin. Zinc 15–ligand discovery for everyone. *Journal of chemical information and modeling*, 55(11):2324–2337, 2015.

Neil Stewart, Gordon DA Brown, and Nick Chater. Absolute identification by relative judgment. *Psychological review*, 112(4):881, 2005.

Kyle Swanson, Parker Walther, Jeremy Leitz, Souhrid Mukherjee, Joseph C Wu, Rabindra V Shivnaraine, and James Zou. Admet-ai: a machine learning admet platform for evaluation of large-scale chemical libraries. *Bioinformatics*, 40(7):btae416, 2024.

LL Thurstone. A law of comparative judgment. *Psychological Review*, 34(4), 1927.

Austin Tripp and José Miguel Hernández-Lobato. Diagnosing and fixing common problems in bayesian optimization for molecule design. *arXiv preprint arXiv:2406.07709*, 2024.

Oleg Trott and Arthur J Olson. Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry*, 31(2):455–461, 2010.

Tiziano Tuccinardi. Docking-based virtual screening: recent developments. *Combinatorial chemistry & high throughput screening*, 12(3):303–314, 2009.

Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, et al. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60, 2023.

Kevin Tirta Wijaya, Navid Ansari, Hans-Peter Seidel, and Vahid Babaei. Trustworthy inverse molecular design via alignment with molecular dynamics. *Advanced Science*, 12(27):2416356, 2025.

Kevin Tirta Wijaya, Bing Xu Hu, Hans-peter Seidel, Wojciech Matusik, and Vahid Babaei. Bayesian post training enhancement of regression models with calibrated rankings. In *The Fourteenth International Conference on Learning Representations*, 2026.

Christopher Williams and Carl Rasmussen. Gaussian processes for regression. *Advances in neural information processing systems*, 8, 1995.

Le Yan, Zhen Qin, Honglei Zhuang, Rolf Jagerman, Xuanhui Wang, Michael Bendersky, and Harrie Oosterhuis. Consolidating ranking and relevance predictions of large language models through post-processing. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 410–423, 2024.

Claudio Zeni, Robert Pinsler, Daniel Zügner, Andrew Fowler, Matthew Horton, Xiang Fu, Zilong Wang, Aliaksandra Shysheya, Jonathan Crabbé, Shoko Ueda, et al. A generative model for inorganic materials design. *Nature*, 639(8055):624–632, 2025.

Guangfeng Zhou, Domnita-Valeria Rusnac, Hahnbeom Park, Daniele Canzani, Hai Minh Nguyen, Lance Stewart, Matthew F Bush, Phuong Tran Nguyen, Heike Wulff, Vladimir Yarov-Yarovoy, et al. An artificial intelligence accelerated virtual screening platform for drug discovery. *Nature Communications*, 15(1):7761, 2024.

A LIMITATIONS

Our study has several limitations that motivate future works:

- Most experiments in the paper are retrospective, using past experimental data, while the higher-fidelity rankers are primarily simulated using noisy oracles. This controlled setting allows us to compare curation methods under matched access to ranking information and systematically examine their sensitivity to ranker accuracy and error structure. This setting is therefore appropriate for comparing the performance of different curation methods. Real-world campaigns may nevertheless introduce challenges beyond those captured by our evaluation, including distribution shifts and domain-specific biases not represented by the tested noise models.
- Our agentic-ranker serves as a proof of concept that considers a single LLM-based ranker across three drug-discovery tasks and three random seeds. We designed this experiment as an exploratory study, motivated by the accessibility of LLMs and their potential to provide pairwise rankings from their broad knowledge. A systematic benchmark study across multiple LLMs would provide complementary insights about their reliability and suitability as ranking sources. Such evaluations must also account for task-specific restrictions: in our setup, Claude Fable declined biochemical ranking queries under its safety mechanisms.
- PRISMS assumes that rankers have a fixed ordering of fidelity, with higher-fidelity sources expected to provide more reliable rankings. This assumption allows us to study selective escalation in isolation from ranker reliability. In practice, ranker quality may depend on the selected pairs, scientific domain, and experimental measurements accumulated during the campaign. For example, an expert chemist may rank molecules more accurately within familiar scaffolds than unfamiliar ones. As experimental evidence accumulates, their accuracy on initially unfamiliar scaffolds may also improve. Extending PRISMS to accommodate such variations during a discovery campaign is an important direction for future work.

B MORE EXPERIMENTAL RESULTS

B.1 COMPUTATIONAL SCALING ANALYSIS

Setup. We evaluate the computational scalability of PRISMS on synthetic design libraries with sizes ranging from 10^3 to 10^5 candidates. For each library size, we measure the 1-round wall-clock runtime and peak resident set size (RSS) during execution of PRISMS with two levels of ranker fidelity. Wall-clock runtime is measured on Intel Xeon Sapphire Rapids 48-core/96-thread 2.10GHz.

Results. PRISMS scales to libraries of up to $N = 100{,}000$ with moderate memory growth. Wall-clock runtime and peak RSS increase from 1.45 s and 282 MB at $N = 1{,}000$ to 32.47 s and 356 MB at $N = 10{,}000$, and to $3{,}233$ s and 1.03 GB at $N = 100{,}000$. Even at the largest scale, the computational overhead is small relative to the physical validation layer, which requires days or months to complete, and is therefore unlikely to be the dominant bottleneck.

B.2 SENSITIVITY ANALYSIS

Figure 7 shows the effects of prior strength λ_2, query-query neighborhood size k^U, and ranker weights $\lambda_1^{(1)}$ and $\lambda_1^{(2)}$ to PRISMS' performance. The AUDC@10 is relatively stable when the prior regularization is weak ($\lambda_2 \leq 10^{-1}$), but declines as regularization becomes stronger, reaching approximately 0.33 AUDC@10 at $\lambda_2 = 100$. This pattern supports using regressor predictions as weak priors that pairwise rankings can refine in the label-scarce regimes.

Increasing the low-fidelity weight $\lambda_1^{(1)}$ beyond 10^{-1} reduces AUDC@10 from approximately 0.41 to 0.32. In contrast, increasing the high-fidelity weight $\lambda_1^{(2)}$ initially improves performance, with the highest observed mean near $\lambda_1^{(2)} = 10$, before performance declines at 100. These results support assigning greater weights, but not full confidence, in higher-fidelity rankers with higher accuracy.

Smaller query-query neighborhoods k^U generally yield higher AUDC@10 than larger neighborhoods, with performance approximately plateauing for $k^U \geq 10$. Thus, denser connectivity does not necessarily improve discovery.

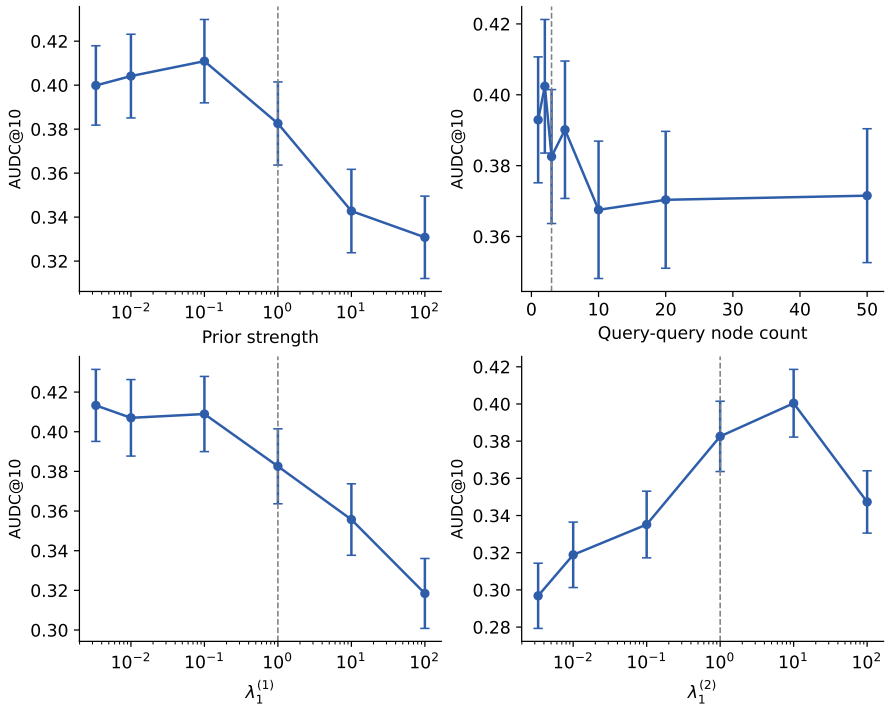

Figure 7: Sensitivity of PRISMS to prior strength λ_2, query-query neighborhood size k^U, and low- and high-fidelity ranker weights $\lambda_1^{(1)}$ and $\lambda_1^{(2)}$. Each parameter is varied while the others remain at their default values, indicated by vertical dashed lines. Points show mean AUDC@10.

B.3 Agentic PRISMS with LLM as Higher-Fidelity Ranker

Setup. We investigate whether PRISMS can operate with an LLM-based agent as the higher-fidelity ranker, replacing the simulated oracle used in the previous experiments. We evaluate three drug-discovery tasks (Ainsley et al., 2026; MacDermott-Opeskin, 2026; Gorantla et al., 2024) targeting kinetic solubility, CYP inhibition, and dopamine D2 receptor activity. Each campaign runs for 12 rounds and is repeated over three random seeds. For the agentic setting, we use Claude Opus 5.5 as the higher-fidelity pairwise ranker, keeping the ensemble classifiers as the lower-fidelity ranker.

Results. Agentic PRISMS achieves an area under the discovery recall curve (AUC) of 0.105, compared with 0.162 for PRISMS using the 75%-accurate simulated oracle. Despite the gap to the oracle setting, Agentic PRISMS outperforms the oracle-based RankRefine++ and HodgeRank baselines, which achieve AUCs of 0.093 and 0.092, respectively. These results suggest the feasibility of Agentic PRISMS for real-world scientific design curation.

B.4 Characterization of LLM as a Pairwise Ranker

Setup. We characterize the error structure of the LLM-based higher-fidelity ranker (Claude Opus 5.5 High effort) by examining how its pairwise accuracy varies with the difficulty of a comparison. For each ranked pair (x_i, x_j), we define the standardized property gap as $d_{ij} = |y_i - y_j|/s_y$, where s_y is the standard deviation of the property values in the corresponding dataset. We bin pairwise comparisons by d_{ij} and compute the empirical pairwise ranking accuracy within each bin. We report results separately for the D2R, kinetic-solubility, and CYP-inhibition datasets, along with the pooled accuracy across all three datasets and three random seeds.

Results. Figure 8 shows a dependence of LLM ranking accuracy on the standardized property gap. For pairs with similar property values, the LLM operates near chance level, whereas its accuracy generally increases as the property difference grows, reaching approximately 0.65–0.66 for the largest gaps in the pooled results. The same trend is visible across the individual datasets. These

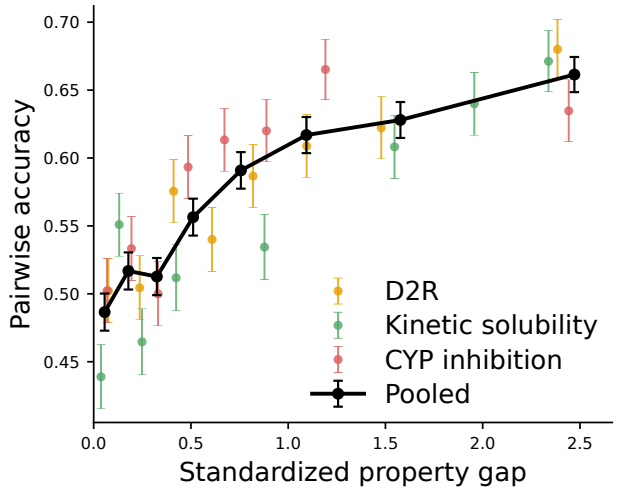

Figure 8: Pairwise ranking accuracy of the LLM-based ranker as a function of the standardized true property gap $d_{ij} = |y_i - y_j|/s_y$, where s_y is the dataset-level standard deviation of the property values. Accuracy generally increases with property gap, indicating a difficulty-dependent error structure in which pairs with similar property values are more difficult for the LLM to rank correctly.

results indicate that LLM ranking errors are *difficulty-dependent*, where comparisons between similar candidates are substantially harder than comparisons between candidates with clearly separated property values.

B.5 Drug Discovery Campaign

Table 1: Experimental results for iterative screening in drug discovery datasets, repeated over 30 random seeds.

Method	12-round AUC				24-round RMST
	Recall	NDCG	Top-1% F1	Top-5% F1	to 50% Recall
Random Sampling	0.077 (0.005)	0.334 (0.057)	0.120 (0.015)	0.134 (0.005)	23.3
GP-EI	0.141 (0.015)	0.398 (0.057)	0.192 (0.027)	0.206 (0.010)	19.3
GP-EI Pairwise	0.227 (0.026)	0.416 (0.067)	0.313 (0.043)	0.260 (0.013)	17.9
HodgeRank	0.306 (0.042)	0.490 (0.055)	0.372 (0.055)	0.326 (0.024)	13.5
Active top-k	0.306 (0.043)	0.487 (0.055)	0.248 (0.022)	0.307 (0.017)	13.6
Spectral MLE	0.292 (0.046)	0.494 (0.056)	0.354 (0.059)	0.312 (0.022)	13.7
RankRefine++	0.308 (0.053)	0.489 (0.052)	0.353 (0.061)	0.326 (0.029)	13.1
PRISMS	0.383 (0.050)	0.541 (0.054)	0.421 (0.051)	0.421 (0.024)	11.1

Table 1 shows the detailed results for the iterative screening campaigns in drug discovery described in Section 4.1. The AUC metrics summarize the first 12 rounds. To estimate the restricted mean time to 50% discovery recall (RMST), all methods were evaluated for 24 rounds. We calculate RMST as the area under the Kaplan–Meier survival curve from round 0 to round 24, with campaigns that have not reached 50% recall by round 24 censored at that horizon.

Figure 9 shows the dataset-wise relative discovery performance of PRISMS over RankRefine++ and HodgeRank. Against RankRefine++ and HodgeRank, PRISMS improves AUDC@10 on all drug discovery datasets, with statistically significant results under Wilcoxon signed-rank tests. The results indicate that the aggregate improvements of PRISMS in the main paper are broadly applicable across the drug discovery datasets.

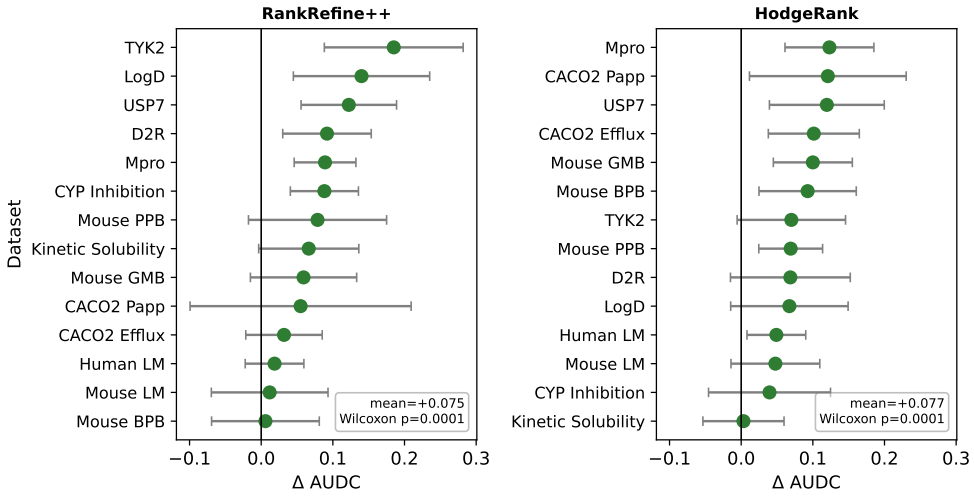

Figure 9: Paired difference in area under the top-10 discovery recall curve for RankRefine++ (left) and HodgeRank (right). Positive values indicate higher discovery performance for PRISMS.

B.6 Engineering Discovery Campaign

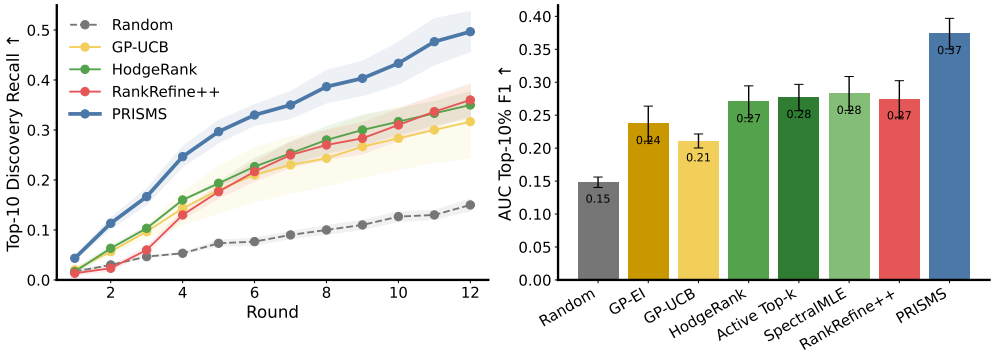

Figure 10: Aggregated results of 12-round scientific design discovery campaigns on three engineering datasets. We report the mean and standard error over 10 random seeds. **Left.** Top-10 discovery recall over 12 rounds. **Right.** AUC of Top-10% F1 score.

The detailed results for engineering discovery campaign are shown in Figure 10.

B.7 Ablation Study

The detailed results for the ablation study are shown in Table 2.

C Implementation Details

C.1 Graph construction

At the beginning of each round, each unmeasured design is connected to all measured designs and its k^U nearest unmeasured neighbors according to the absolute difference in prior estimates. The topology remains fixed throughout the round. Comparisons are stored based on the edge and ranker identity.

Table 2: Ablation of PRISMS components and fidelity stages. We report the 24-round mean area under the discovery recall curve (AUDC) across 14 drug-discovery datasets, repeated for 10 random seeds, with standard error of the mean. Relative change is computed with respect to the default two-stage PRISMS configuration.

Method	AUDC@10 ↑	Relative change
PRISMS 2 stage	0.527 ± 0.049	—
No unlabeled-unlabeled edges	0.487 ± 0.048	-7.6%
No edge uncertainty	0.502 ± 0.047	-4.6%
No outcome uncertainty	0.514 ± 0.048	-2.3%
Random escalation	0.486 ± 0.050	-7.7%
No lateral edges + random escalation	0.481 ± 0.051	-8.7%
Random Stage-1 initialization	0.519 ± 0.047	-1.4%
PRISMS - 1 stage	0.440 ± 0.051	-16.5%
PRISMS - 3 stage	0.586 ± 0.045	$+11.4\%$

C.2 Optimization.

Experimentally validated property values are held fixed, and the objective in Equation 2 is optimized over unmeasured values using L-BFGS-B with an analytic gradient. The first optimization is initialized from the supplied prior estimates. After higher-fidelity comparisons are added, optimization is initialized from the preceding $(\ell - 1)$ stage's solution. The implementation defaults to at most 300 iterations, with function and gradient tolerances of 10^{-12} and 10^{-8}, respectively.

C.3 Sparse Hessian construction.

After optimization, we assemble the analytic Hessian of the objective at the current estimates. Each observed comparison from ranker g_s contributes

$$w_{ij}^{(s)} \boldsymbol{v}_{ij} \boldsymbol{v}_{ij}^{\mathsf{T}}, \qquad w_{ij}^{(s)} = \frac{\lambda_1^{(s)}}{\tau^2} p_{ij}(\hat{y}^{(\ell)}) \left(1 - p_{ij}(\hat{y}^{(\ell)})\right). \tag{9}$$

For a comparison between two unmeasured designs, this adds $w_{ij}^{(s)}$ to both diagonal entries and $-w_{ij}^{(s)}$ to the corresponding off-diagonal entries. Comparisons against experimentally validated designs contribute only to the diagonal of the unmeasured designs. The prior contributes $2\lambda_2 I$.

The Hessian is stored in compressed sparse column format. Contributions from multiple rankers to the same matrix entry are summed. Its off-diagonal sparsity follows the comparison graph among unmeasured designs.

C.4 Computing edge uncertainty.

We evaluate $\boldsymbol{v}_{ij}^{\mathsf{T}} (H^{(\ell)})^{-1} \boldsymbol{v}_{ij}$ without explicitly forming the inverse. A sparse LU factorization of $H^{(\ell)}$ is computed once before scoring the eligible comparisons. For each pair, we reuse this factorization to solve

$$H^{(\ell)} \mathbf{u}_{ij} = \boldsymbol{v}_{ij}, \tag{10}$$

and compute

$$U_{ij}^{\text{edge}} = \boldsymbol{v}_{ij}^{\mathsf{T}} \mathbf{u}_{ij}. \tag{11}$$

This evaluates the quadratic form through direct linear solves without storing a dense covariance matrix.

C.5 LLM-generated Rule-based Ranker

For the cheap ranker ensemble, the rule-based ranker is generated using Claude Sonnet 4.6. An example of the LLM-generated rule-based ranker is provided below.

```python
def predict_expansionrx_lm_clint(
    train_features: dict,
    test_features: dict,
) -> int:
    """Predict liver microsomal intrinsic clearance (CLint) comparison."""

    if train_features is None or test_features is None:
        return _random_binary()

    def clint_score(f):
        score = 0.0
        score += f['aromatic_rings'] * 0.45
        score += f['logp'] * 0.35
        score += f['num_heteroatoms'] * 0.2
        score += f['rotatable_bonds'] * 0.12
        score -= f['fraction_csp3'] * 0.3
        return score

    if clint_score(train_features) > clint_score(test_features):
        return 1
    else:
        return 0
```

C.6 DATASET

Table 3: Details of datasets used in the experiments.

Dataset Name	Dataset Size	Target Property
Caco-2 efflux	3777	Caco-2 efflux ratio
Caco-2 Papp	3773	Caco-2 permeability
Human LM	4541	microsomal clearance
Kinetic solubility	7298	kinetic solubility
LogD	7309	LogD
Mouse LM	5692	microsomal clearance
Mouse BPB	1426	brain fraction unbound
Mouse GMB	431	mouse gastrocnemius muscle binding
Mouse PPB	1756	plasma fraction unbound
Octant CYP3A4	1084	CYP3A4 inhibition potency
D2R	2502	binding affinity ($\mathrm{p}K_i$)
Mpro	665	binding affinity (pIC_{50})
TYK2	9997	binding affinity ($\mathrm{p}K_i$)
USP7	4535	binding affinity (pIC_{50})
3D printed structure	600	mechanical toughness
Dye lasers	3458	peak score
Redoxmer	1408	absorption wavelength deviation