PROTO-NOOS: Rational Compute Allocation in Antibiotic Discovery via Mutual Information and Uncertainty Quantification

Abstract

Large-scale computational screening of compounds with tools designed to predict molecules with specific properties requires substantial computational time and resources. The initial search for promising candidates and the subsequent optimization of those candidates are distinct processes and should be treated as separate stages of the drug discovery workflow.

In this context, molecular search can be framed as an optimization problem: the goal is to obtain the best possible result at the lowest feasible computational cost. By using known antibiotics against a given target, it is possible to generalize how specific compounds interact with the receptor. When additional phenotypic data are incorporated, the performance of computational tools can be compared against experimental outcomes, allowing their behaviour to be estimated and evaluated using machine learning models.

In this study, we used mutual information and uncertainty quantification methods to analyze a dataset of de novo compounds generated by our PROTO-NOOS system. PROTO-NOOS is designed to generate labels consistent with state-of-the-art computational tools and public experimental data reported in previous studies.

We performed a benchmark evaluating EoE-5 and EoE-10 models across six random seeds for each method. The aim was to assess how these approaches perform on our problem and to determine their suitability for reducing computational cost while preserving predictive quality.

Our results indicate which strategies are most appropriate for users who want to simplify expensive computational workflows by replacing selected stages with lower-cost surrogate models. In addition, we analyzed how predictions differ between single models, EoE-5, and EoE-10 ensembles, and how uncertainty relates to the training data and the diversity of the evaluated compounds.

The results suggest that random forest-based strategies are useful when the primary goal is aggressive cost reduction, even if this comes with some loss of quality. In contrast, evidential models combined with MIST-QR provide a more conservative strategy, where computational savings are achieved while maintaining stronger control over uncertainty and prediction reliability.

This work is relevant for researchers developing or using modular, open-source drug design pipelines that combine multiple computational tools. The proposed approach can help determine where mutual information and uncertainty quantification can be used to reduce computational cost during large-scale compound screening without fully abandoning the reliability of more expensive downstream methods.

Introduction

Large-scale computational screening of chemical compounds with tools designed to predict molecules with desired biological or physicochemical properties requires substantial computational time and resources. The initial search for promising candidates and the subsequent optimization of those candidates are distinct stages of the drug discovery workflow and should be treated separately.

In this setting, molecular search can be framed as an optimization problem in which the objective is to obtain the best possible predictive outcome at the lowest feasible computational cost. For a defined biological target, prior information from known antibiotics and experimentally characterized compounds can be used to learn how specific molecular structures interact with the target receptor. When phenotypic data are also incorporated, computational predictions can be compared against experimental outcomes, allowing the behaviour and limitations of the pipeline to be assessed using machine learning models.

In this study, we investigated whether mutual information and uncertainty quantification can be used to support cost-aware decision-making in a modular drug design workflow. We focused on a dataset of de novo compounds generated by PROTO-NOOS, a pipeline designed to produce labels consistent with state-of-the-art computational tools and public experimental data from previous studies.

Conceptually, our goal was not to replace the full computational pipeline with a single model. Instead, we evaluated whether surrogate models can identify stages at which expensive downstream calculations may be skipped or delayed while maintaining acceptable predictive quality. This is relevant for researchers developing modular drug design systems composed of multiple open-source tools, where large-scale compound screening is often limited by computational cost.

Methods

We performed a benchmark of EoE-5 and EoE-10 models across ten predefined random seeds, training and evaluating each model variant independently. Mutual information was estimated using MIST, Mutual Information estimation via Supervised Training, implemented through the `grgera/mist` framework.

All models used UniMol as the molecular encoder. UniMol is a conformer-based molecular transformer that produces molecular representations from three-dimensional compound conformations. The encoder was connected to multi-task prediction heads.

Regression tasks, including docking score, binding affinity KD, growth ratio, inhibition, and Ki, were modelled using Normal-Inverse-Gamma, NIG, heads. These heads parameterize an evidential predictive distribution and return, for each compound, an estimate of predictive uncertainty, denoted as σ, without requiring Monte Carlo sampling during inference.

Binary classification tasks, including pipeline hit and binary inhibition, were modelled using standard sigmoid prediction heads.

The model was trained in three consecutive phases.

In the first phase, the experimental phase, the model was trained on experimental affinity and phenotypic data. The active heads in this phase were Ki, mean inhibition, and binary inhibition. The experimental phenotype data included the STOKES dataset.

In the second phase, the pipeline surrogate phase, the model was trained on pipeline-derived labels and de novo compounds. The active heads in this phase were Ki, docking score, growth ratio, and pipeline hit.

In the third phase, the calibration phase, the model was calibrated on validation data. The calibrated heads were pipeline hit and binary inhibition, and isotonic regression was used for probability calibration.

The training hyperparameters were as follows: a maximum of 60 epochs, early stopping with a patience of 8 epochs, batch size of 384, learning rate of 10^-3, weight decay of 10^-4, and 16 MC-dropout samples during training.

We compared three model families that differed in how predictive uncertainty was estimated and aggregated.

The first model family was Single NIG. This was a single model instance trained end-to-end with NIG heads. Aleatoric and epistemic uncertainty components were derived analytically from the parameters of the evidential distribution obtained in a single forward pass.

The second model family was EoE-5, an Ensemble of Ensembles with five members. This variant consisted of five independently trained model instances, each initialized with a different random seed. During inference, predictions and uncertainty estimates from all five members were aggregated. The variance between model members provided an additional source of epistemic uncertainty, complementing the NIG-based uncertainty estimates from individual ensemble members.

The third model family was EoE-10, an Ensemble of Ensembles with ten members. This variant followed the same protocol as EoE-5 but used ten independently trained members, allowing a richer estimate of model variance.

All experiments were performed using ten predefined random seeds: 42, 137, 256, 512, 1024, 2048, 4096, 8192, 16384, and 32768. These seeds controlled weight initialization, data shuffling, and dropout masks during training. The seed sets used in individual analyses were defined before evaluation to avoid post-hoc selection.

Pareto frontier analysis was performed using six seeds. Ranking quality evaluation was performed using eight seeds. Uncertainty quantification and calibration diagnostics were performed using all ten seeds.

For each model variant, namely Single NIG, EoE-5, and EoE-10, metrics were computed independently for each seed and then reported as mean ± standard deviation. The standard deviation was calculated using DDOF = 1.

Data Used in the Experiments

The dataset used in this study consisted of 3,327 experimentally labelled compounds and 13,531 unlabelled, computationally designed de novo molecules. The labelled data were obtained from four sources.

The Stokes et al. dataset contained 2,234 compounds with continuous measurements of *E. coli* growth inhibition, with 5.4% of compounds classified as active. ChEMBL contributed 605 compounds with biochemical activity data, including IC50, Ki, and KD values, collected from 74 unique assays against bacterial molecular targets. BindingDB contributed 424 compounds with Ki values for three target proteins, including UniProt identifiers B0BL08 and P0ABQ4. ChEMBL Decoys contributed 64 compounds and served as a negative control set composed exclusively of inactive molecules.

Continuous affinity labels, including Ki, IC50, and KD, were binarized using a threshold of 1.0 μM where applicable. Phenotype-derived hit labels were defined using a growth ratio threshold of 0.1.

The de novo molecules (n = 13,531) were generated with REINVENT across 19 generative branches. Their median QED was 0.765 ± 0.228. Raw outputs for these molecules were processed through the PROTO-NOOS system, which provided additional regression labels, including `docking_score`, `kd_pred_log10_uM`, and `growth_ratio`. These labels were used as training data for the second training phase, referred to as the pipeline surrogate phase.

Within-source duplicates were removed. Compounds appearing in more than one source were retained and explicitly annotated with source-overlap information.

Source	n	Hit Definition	Positive (n)	Prevalence
Stokes et al.	2,234	Inhibition binary, growth < 0.1	120	5.4%
ChEMBL	605	Pipeline hit, Ki ≤ 1 μM	404	66.8%
BindingDB	424	Pipeline hit, Ki ≤ 1 μM	301	71.0%
ChEMBL Decoys	64	Pipeline hit	0	0.0%
Combined	3,327	Pipeline hit	825	24.8%

Data Splitting

The labelled compounds were split chronologically into a training set (n = 1,898), a validation set (n = 1,005), and a locked test set (n = 424). Compounds with approximated labels were concentrated in the validation set.

Near-duplicate compounds were removed based on chemical similarity to the training set. Near duplicates were defined as compounds with Tanimoto similarity greater than 0.85, calculated using ECFP4 fingerprints. Out-of-distribution compounds were defined as

molecules with Tanimoto similarity below 0.30 relative to the training set. A threshold of 0.40 was used to identify structurally novel compounds for novelty analysis.

An additional shadow test set, consisting of 15% of the de novo molecules and with no overlap with the locked test set, was used for prospective evaluation.

Biological performance was evaluated on the following test sets:

Test Set	Biological Target	Metrics
`exp_affinity_test_lock`	Experimental binding affinity, Ki	Spearman rank correlation coefficient ρ, distance correlation, RMSE
`stokes_test_lock`	Phenotypic growth inhibition	Regression: Spearman rank correlation coefficient ρ, RMSE; binary classification: AUROC, ECE with 10 adaptive bins
`pipeline_test_plus_denovo_shadow`	Prospective pipeline-imitation performance	Pareto frontier, hit recovery, compute saved

Uncertainty Quantification Diagnostics

Uncertainty quantification diagnostics were computed using three groups of metrics.

First, we measured the association between predictive uncertainty and prediction error using the Spearman rank correlation coefficient ρ and distance correlation between σ and the absolute prediction error.

Second, we evaluated out-of-distribution detection using AUROC, where σ was used as the score for distinguishing in-distribution compounds from out-of-distribution compounds.

Third, we estimated the mutual information between σ and the absolute prediction error using the MIST-QR estimator. This analysis used 100 bootstrap repetitions and 1,000 Monte Carlo draws per compound. The results were additionally compared against the KSG estimator with k = 3.

Selective prediction curves were computed by sorting compounds in ascending order according to σ̂ and measuring selective RMSE at 25 evenly spaced coverage levels from 0.10 to 1.00. Curves from individual seeds were interpolated onto a common grid and then averaged across seeds.

Threshold	Positive (n)	Prevalence

Ki ≤ 0.1 μM	402	36.8%
Ki ≤ 1.0 μM	710	65.0%
Ki ≤ 10 μM	836	76.5%

Dataset Grouping and Use Across Training Phases

The full dataset was divided into groups that served different roles during training and evaluation. Experimentally labelled compounds were used in Phase 1 to train the biological prediction heads, including Ki prediction, inhibition prediction, and binary inhibition classification. The same experimental subset was later used in Phase 3 for calibration with isotonic regression.

This experimental subset also formed the basis for biological evaluation on the `exp_affinity_test_lock` and `stokes_test_lock` test sets. For this reason, some analyses may appear to involve only approximately 3,200 compounds. However, this number refers only to compounds with experimental labels and does not represent the full amount of data used in training and benchmarking.

The remaining 13,531 de novo compounds did not have experimental ground truth, but they were not excluded from the study. These compounds were used in Phase 2 to train the pipeline surrogate model, because they had computational labels generated by the PROTO-NOOS pipeline. These labels included docking score, predicted KD, growth ratio, and checkpoint-level outputs from successive stages of the system.

Approximately 15% of the de novo pool was separated as a shadow test set for prospective evaluation. This shadow set was used to assess whether the model could recover pipeline-defined hits and reduce computational cost without running the full pipeline for every candidate compound. The de novo compounds were not included in `exp_affinity_test_lock` or `stokes_test_lock`, because they did not have laboratory measurements.

Therefore, the experimental subset was used for biological validation, whereas the larger de novo subset was used for pipeline-surrogate training and prospective pipeline-imitation evaluation.

Chemical Space and Dataset Structure

We analyzed the chemical space of the full dataset using UMAP and PCA projections, with compounds annotated by source and data split. The UMAP projection showed that the de novo compounds formed a broad and dense region of chemical space, while experimentally derived compounds from ChEMBL and BindingDB occupied smaller and more discrete regions. This pattern is consistent with the more target-specific nature of biochemical affinity datasets.

In contrast, the Stokes dataset showed a broader distribution, which is expected because it reflects phenotypic whole-cell growth inhibition rather than activity against a single molecular target. The PCA projection provided a complementary view. It showed weaker separation than UMAP, suggesting that some apparent UMAP clusters may reflect local neighborhood structure rather than large-scale linear separation in chemical space.

The separation between de novo and experimentally labelled compounds suggests that the de novo library expands the explored chemical space. However, this observation should not be interpreted as direct evidence that many de novo compounds would act as *E. coli* drugs through mechanisms unrelated to the selected receptor. Chemical-space position alone does not establish antibacterial activity or mechanism of action. A safer interpretation is that the de novo pool contains structurally novel candidates that may require additional phenotypic or mechanistic validation.

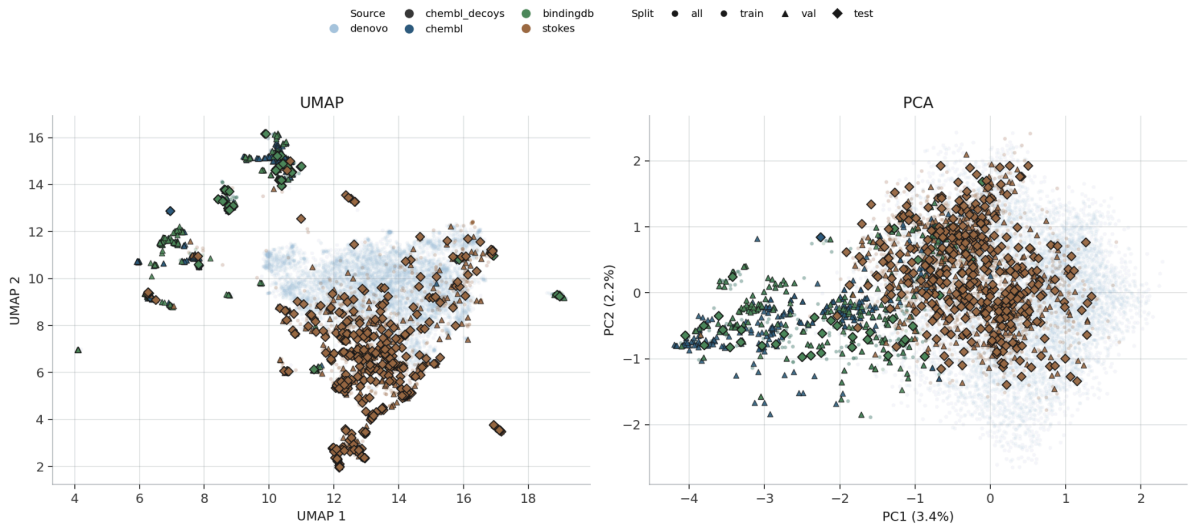


Figure 1. Chemical-space structure of experimental and de novo compounds. UMAP and PCA projections of the compound library, colored by data source and marked by split. De novo compounds occupy a broad region of chemical space, whereas ChEMBL and BindingDB compounds form more localized regions, consistent with their target-specific biochemical origin. The Stokes phenotypic dataset is more broadly distributed, reflecting whole-cell growth inhibition measurements rather than receptor-specific affinity labels. PCA shows weaker separation than UMAP, indicating that local clustering should not be overinterpreted as global linear separation.

Activity-Cliff and Novelty Analysis

A major challenge in chemical machine learning is the presence of activity cliffs, where structurally similar compounds show large differences in biological activity. This effect was also observed in our dataset. Several compounds had high nearest-neighbor similarity to the training set but still showed large differences in Ki, indicating that Tanimoto similarity alone is insufficient to guarantee easy predictability.

Activity cliffs were identified by comparing compounds to their nearest training-set neighbors and measuring the absolute difference in Ki on the log10 μM scale. The analysis showed that a small but non-negligible fraction of compounds formed activity-cliff pairs.

Group	n scored	Activity cliffs (n)	Cliff fraction	Mean nearest Tanimoto	Mean |ΔKi| (log10 μM)
Overall	686	9	1.3%	0.537	1.89
ChEMBL	436	6	1.4%	0.538	2.08
BindingDB	250	3	1.2%	0.535	1.56
Validation split	612	7	1.1%	0.511	2.02
Locked test split	74	2	2.7%	0.745	0.80

The scaffold and cluster diagnostics further showed that the de novo compounds had low overlap with the training set. Scaffold overlap was approximately 15%, while Butina cluster overlap was approximately 2.2%. This supports the interpretation that the de novo library introduces substantial novelty rather than merely reproducing the training distribution.

In contrast, the experimental test compounds were chemically closer to the training set. This distinction is important for evaluation: the locked experimental test sets measure biological generalization within experimentally characterized chemical space, whereas the de novo shadow set measures prospective behaviour in a more out-of-distribution screening scenario.

Figure 2. Activity-cliff analysis based on nearest training-set neighbors.
Scatter plot showing nearest-neighbor Tanimoto similarity to the training set versus absolute Ki difference on the log10 μM scale. Dashed lines mark the high-similarity and

large-activity-difference thresholds used to identify activity cliffs. Several compounds show substantial Ki differences despite high structural similarity to training compounds, demonstrating that molecular similarity alone is not sufficient for reliable activity prediction.

Pareto Analysis

Hit recovery was defined as the fraction of true full-pipeline hits retained by an early-exit strategy. Compute saved was defined as the fraction of expensive downstream pipeline stages that could be skipped.

For each seed, the best operating point was selected under a minimum compute-savings constraint of 55%. Hit recovery and compute saved were bootstrapped using 1,000 replicates to obtain confidence intervals.

Conformal prediction intervals with nominal 90% coverage, corresponding to $\alpha = 0.10$, were computed only for EoE conditions. These intervals were calculated using per-seed calibration sets.

Pairwise differences between EoE conditions and the Single NIG baseline were computed for each metric as:

$$\Delta = \mu_{EOE} - \mu_{Single}$$

where μ denotes the mean across seeds. No correction for multiple comparisons was applied. Effect sizes were reported together with within-condition standard deviations to support interpretation of the practical relevance of observed differences.

Inter-Stage Mutual Information Analysis

To evaluate how information propagated between stages of the pipeline surrogate, we computed pairwise mutual information between checkpoint-level outputs. Predictions were generated for biological heads, checkpoint outputs from A to H, random-forest baselines, affinity prediction, inhibition prediction, binary inhibition, pipeline growth prediction, and an experimental world-model component.

The originally intended world-model component was designed as a JEPA-like representation model. However, it was not retained as a central result because its performance was not sufficiently reliable in the low-data regime, despite being explicitly motivated by that setting.

The resulting inter-stage mutual information matrix was used to compare Single NIG, EoE-5, and EoE-10. The Single NIG model showed weaker and less structured mutual-information signals between stages. In contrast, EoE-5 produced a clearer and stronger inter-stage information structure. EoE-10 showed a similar pattern, supporting the interpretation that ensemble-based evidential models stabilize the estimation of information shared between pipeline stages relative to a single evidential model.

This result addresses whether EoE-style ensembling changes or stabilizes inter-stage mutual-information estimation compared with Single NIG. The observed matrices suggest

that ensemble aggregation can reveal stage-to-stage dependencies that are weak or unstable under a single-model evidential setup.

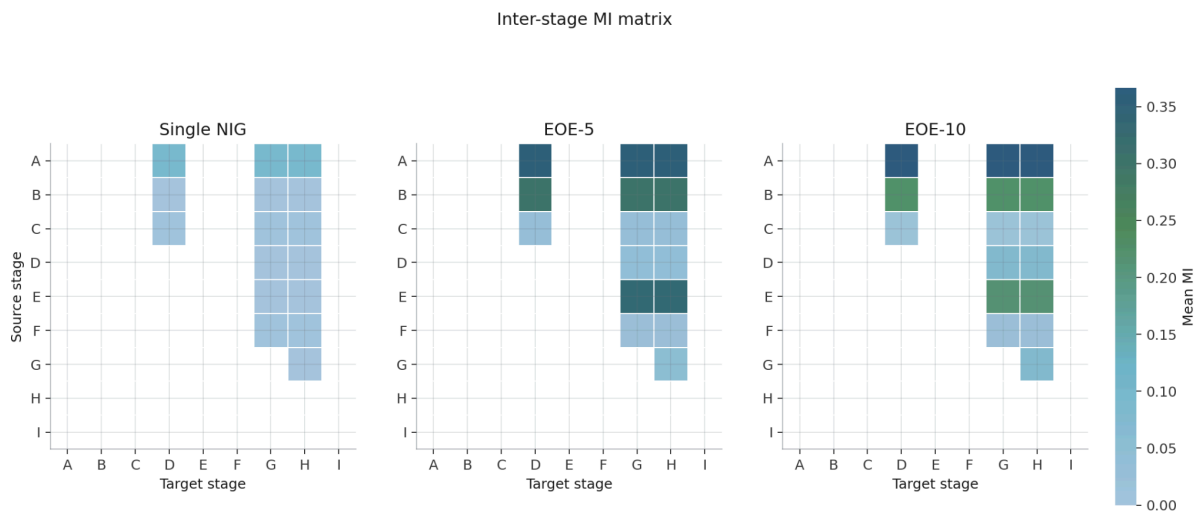


Figure 3. Inter-stage mutual-information matrices for Single NIG, EoE-5, and EoE-10. Pairwise mutual information between pipeline stages was estimated for Single NIG, EoE-5, and EoE-10. Single NIG shows weaker and less structured inter-stage signal, whereas EoE-5 and EoE-10 show stronger and more coherent dependencies between selected stages. This suggests that ensemble-based evidential models stabilize mutual-information estimation across the modular pipeline.

Constraints and Limitations

Several limitations affected the study design and interpretation.

First, the available data did not contain an ideal experimentally verified true-negative background for constructing a dataset that fully matches prospective real-world screening conditions. This limitation affected the analysis throughout the study, especially when evaluating hit recovery and pipeline-imitation behaviour.

Second, the datasets were heterogeneous. They combined biochemical affinity measurements, phenotypic growth-inhibition data, and computational labels generated by the PROTO-NOOS pipeline. These sources differ in assay type, biological resolution, and noise structure. As a result, model performance on one task should not be interpreted as equivalent to performance on another task.

Third, MIST-QR is a relatively new mutual-information estimation framework. Although it was suitable for the analyses performed here, its behaviour is still less established than more conventional estimators. For this reason, MIST-QR results were compared against KSG estimates where appropriate.

Fourth, the number of random seeds required for stable uncertainty estimates was not known before the analysis. This motivated the use of multiple seed regimes: six seeds for

Pareto frontier analysis, eight seeds for ranking quality evaluation, and ten seeds for uncertainty quantification and calibration diagnostics.

Finally, comparing Single NIG, EoE-5, and EoE-10 substantially increased the experimental workload and the number of outputs requiring interpretation. However, this comparison was necessary to determine whether ensemble-based evidential models provide practical advantages over a single evidential model in uncertainty estimation, mutual-information analysis, and cost-aware early-exit decisions.

Results

Early-Exit Strategies Preserve Hit Recovery While Reducing Pipeline Cost

We first evaluated whether early-exit strategies can reduce the computational cost of the full PROTO-NOOS pipeline while preserving recovery of pipeline-defined hits. The main trade-off was measured using two quantities: compute saved relative to the full pipeline and hit recovery relative to the full pipeline.

Figure 4 shows the Pareto frontier for the EoE-5 final setting. The left panel shows the relationship between compute saved and hit recovery. The right panel shows the corresponding threshold sweep, indicating what fraction of compounds continues to the full pipeline under each strategy.

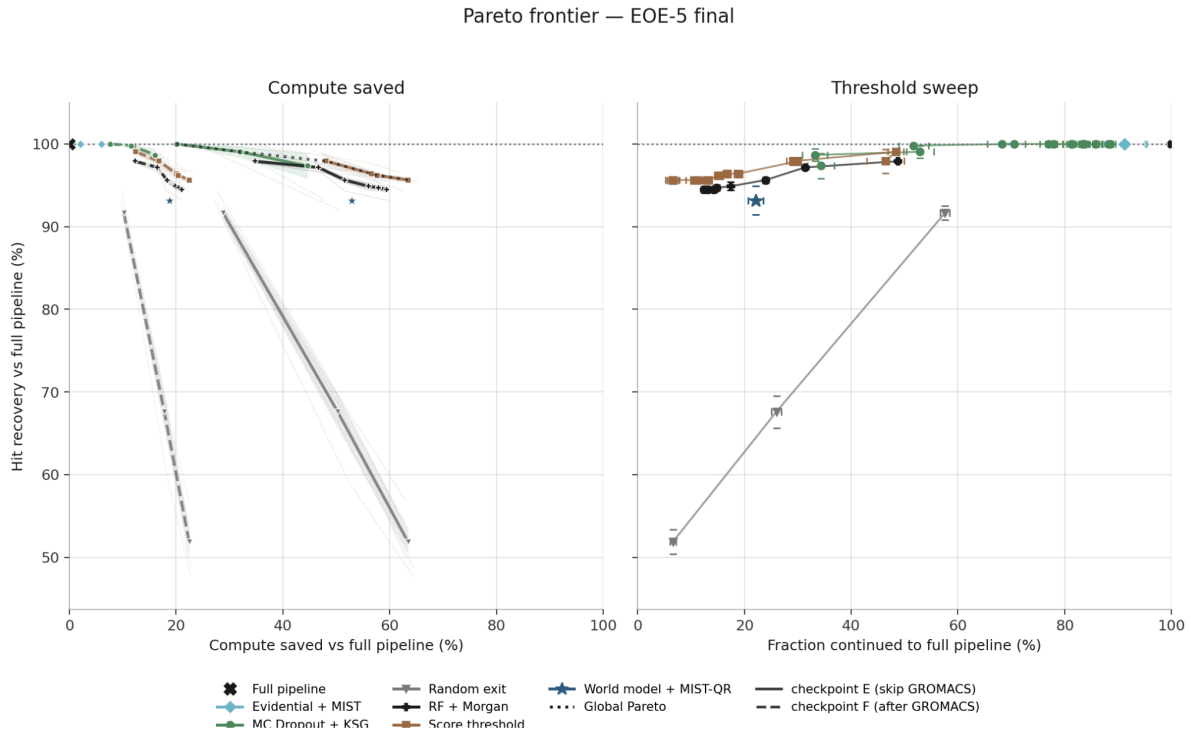


Figure 4. Pareto frontier for early-exit strategies under the EoE-5 final setting.
The left panel shows compute saved relative to the full pipeline versus hit recovery relative

to the full pipeline. The right panel shows the corresponding threshold sweep, where the x-axis represents the fraction of compounds continued to the full pipeline. Conservative uncertainty-aware strategies retain full hit recovery with modest compute savings, whereas more aggressive threshold-based strategies achieve larger savings but require stronger assumptions about acceptable hit loss and uncertainty reliability.

The full pipeline retained 100% of hits by definition and saved no compute. Among the early-exit methods, several strategies preserved complete hit recovery while reducing computational cost. Score thresholding and RF + Morgan achieved the largest compute savings while maintaining full hit recovery in this benchmark. However, these methods should be interpreted primarily as aggressive filtering strategies rather than uncertainty-aware decision policies.

Evidential + MIST achieved full hit recovery with a more conservative reduction in computational cost. Although its mean compute saving was smaller, this strategy was designed to provide stronger uncertainty control and therefore represents a safer operating point when the goal is to avoid losing potentially valuable compounds.

Method	**Compute Saved % ↑**	**Hit Recovery % ↑**	**AUC-ROC ↑**
Full Pipeline	0.0 ± 0.0	100.0 ± 0.0	1.000 ± 0.000
Random Exit	50.3 ± 1.5	97.0 ± 5.5	0.723 ± 0.029
Score Threshold	57.7 ± 0.7	100.0 ± 0.0	0.945 ± 0.004
RF + Morgan	57.9 ± 0.1	100.0 ± 0.0	0.951 ± 0.002
MC Dropout + KSG	14.9 ± 2.6	100.0 ± 0.0	0.994 ± 0.002
World Model + MIST-QR	52.9 ± 2.4	100.0 ± 0.0	0.936 ± 0.034
Evidential + MIST	6.0 ± 6.5	100.0 ± 0.0	1.000 ± 0.001

Overall, EoE-5 Evidential + MIST provides a conservative Pareto point with complete hit recovery and modest compute savings. In contrast, RF + Morgan and score-threshold strategies are more suitable for aggressive early-stage triage, where the primary goal is to rapidly remove clearly weak candidates. These methods are fast and stable, but they should not be treated as the main uncertainty-aware mechanism.

MC Dropout + KSG appears competitive in the Pareto analysis. However, its behaviour in selective prediction was less reliable, because it did not consistently rank uncertain compounds in the expected direction. For this reason, MC Dropout + KSG is best interpreted as a comparative baseline rather than the preferred deployment strategy.

The intended use of EoE-5 or EoE-10 is not to replace the full pipeline for final candidate validation. Instead, these models are best interpreted as early-exit controllers placed between stages of the pipeline. After lower-cost or intermediate-cost stages, such as docking, penetration or retention prediction, or Boltz/KD estimation, the model can estimate

whether a compound should be continued to more expensive downstream stages such as GROMACS and systems-level analysis.

Compounds predicted to be weak with low uncertainty could be stopped early. Compounds predicted to be promising, uncertain, or outside the applicability domain would continue to the full pipeline.

The most practical deployment point appears to be the B-light variant, where the gate is placed after Boltz/KD information is available but before GROMACS and downstream systems-level stages. At this point, the model has access to a stronger biological and structural signal than at the earliest checkpoints, while still retaining the ability to save a substantial fraction of downstream cost. Checkpoint A can be used for more aggressive high-throughput screening, but it is less appropriate as the default production setting.

Checkpoints explanation

THIS SECTION WILL BE UPDATED

Hit Recovery Relative to the Full Pipeline

We next compared early-exit strategies directly against the full pipeline across operating points. This analysis separates two questions: how much compute can be saved, and how much hit recovery is lost relative to the full pipeline.

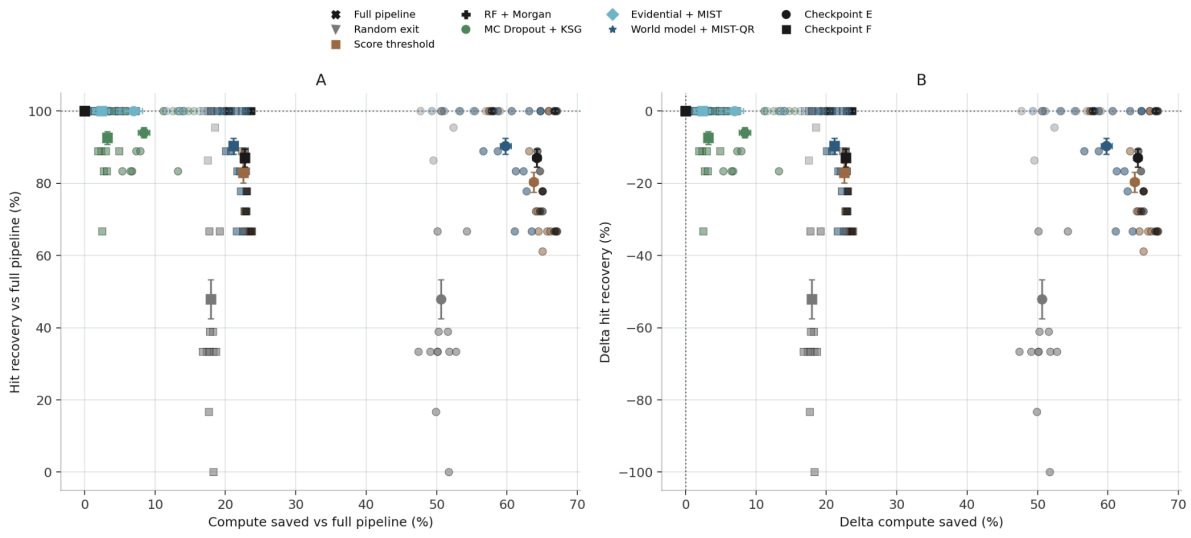


Figure 5. Hit recovery and compute savings relative to the full pipeline.
Panel A shows hit recovery as a function of compute saved relative to the full pipeline. Panel B shows the corresponding changes relative to the full-pipeline baseline. Several strategies achieve substantial compute savings, but their practical value depends on

whether hit recovery remains close to 100%. The plot highlights that compute reduction alone is not sufficient: early-exit strategies must be evaluated against the full-pipeline hit set.

This analysis shows that the apparent efficiency of a method depends strongly on the selected operating point. Some configurations achieve large compute savings but lose a meaningful fraction of hits. Others preserve hit recovery but save less compute. Therefore, the choice of method should depend on the intended use case.

For exploratory screening of very large chemical libraries, aggressive RF or score-threshold strategies may be useful as low-cost prefilters. For conservative production use, where false rejection of potentially valuable compounds is more costly, evidential uncertainty-aware strategies are more appropriate.

Biological Prediction Performance of Single NIG and EoE Models

We then compared Single NIG, EoE-5, and EoE-10 on biological prediction tasks. Performance was evaluated on experimental affinity and Stokes phenotypic inhibition endpoints.

Metric	Single NIG	EoE-5	EoE-10
Affinity Spearman ↑	0.131 ± 0.161	0.058 ± 0.115	0.069 ± 0.070
Affinity dCor ↑	0.322 ± 0.057	0.274 ± 0.028	0.261 ± 0.030
Affinity RMSE ↓	1.739 ± 0.036	1.742 ± 0.016	1.737 ± 0.012
Stokes Spearman ↑	0.119 ± 0.035	0.124 ± 0.014	0.131 ± 0.010
Stokes dCor ↑	0.281 ± 0.028	0.322 ± 0.014	0.333 ± 0.010
Stokes RMSE ↓	0.277 ± 0.006	0.267 ± 0.003	0.264 ± 0.002
Stokes AUC-ROC ↑	0.791 ± 0.033	0.828 ± 0.019	0.825 ± 0.024
Stokes ECE ↓	0.024 ± 0.010	0.043 ± 0.027	0.050 ± 0.030

The ensemble models did not clearly improve experimental affinity prediction. Affinity Spearman correlation remained low across all model families, and RMSE values were similar. This suggests that affinity prediction remained difficult under the available heterogeneous experimental data.

For the Stokes phenotypic endpoint, EoE-5 and EoE-10 improved distance correlation, RMSE, and AUC-ROC relative to Single NIG. However, calibration as measured by ECE was better for Single NIG. This indicates a trade-off: ensembles improved some aspects of predictive performance, but did not necessarily improve probability calibration.

AUC-ROC Alone Is Not Sufficient for Selecting an Early-Exit Strategy

We also compared AUC-ROC against hit recovery relative to the full pipeline. The purpose of this analysis was to distinguish global classification quality from practical early-exit utility.

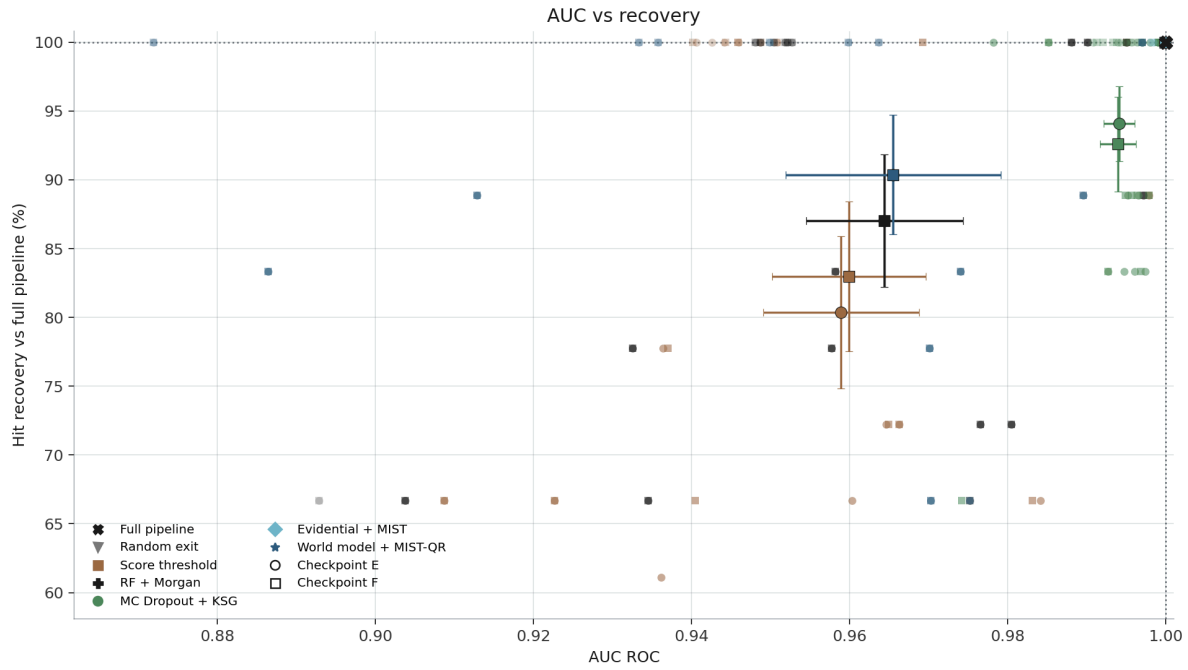


Figure 6. AUC-ROC versus hit recovery for early-exit strategies.
AUC-ROC was compared against hit recovery relative to the full pipeline. Several configurations achieved high AUC-ROC but differed substantially in hit recovery. This shows that global classification performance is not sufficient for selecting an early-exit strategy. For deployment, hit recovery at a selected compute budget is the more relevant criterion.

The plot shows that a high AUC-ROC is not sufficient for selecting a deployment strategy. Some configurations achieved AUC-ROC values close to 1.0 but still differed in hit recovery. This means that a method can separate active and inactive compounds well in a global ranking sense, while still losing important hits at a specific early-exit threshold.

Therefore, the main deployment metric should be hit recovery relative to the full pipeline at a predefined compute budget. AUC-ROC should be treated as a secondary diagnostic metric rather than the primary selection criterion.

Stability of Uncertainty Ranking Across Seed Counts

We evaluated whether increasing the number of random seeds stabilizes the relationship between predictive uncertainty and prediction error. For each checkpoint, we computed bootstrap estimates of Spearman correlation between σ and absolute error using increasing

seed subset sizes.

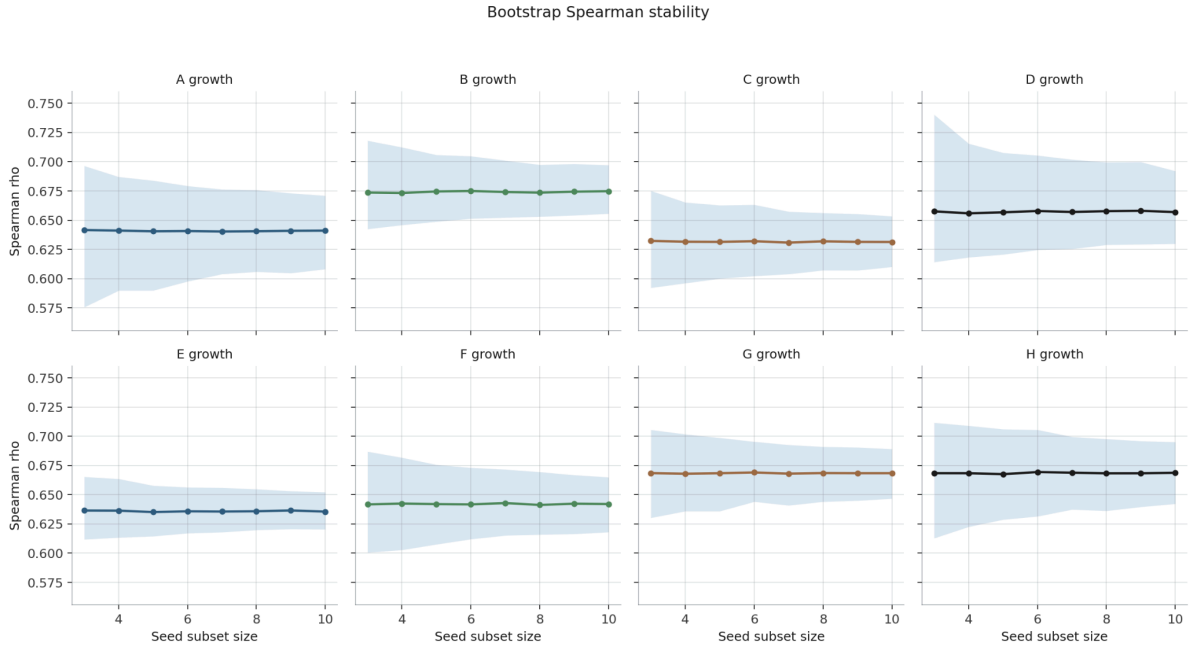


Figure 7. Bootstrap stability of the association between predictive uncertainty and absolute error.
Spearman correlation between σ and absolute prediction error was estimated for increasing seed subset sizes. Across checkpoints, the central estimate remains stable when moving from 3 to 10 seeds, while confidence intervals become narrower. This indicates that additional seeds reduce uncertainty in the estimate rather than changing the direction of the result.

Across checkpoints, the correlation estimates remained consistently positive and generally fell between approximately 0.6 and 0.7. Increasing the number of seeds did not change the direction of the result. Instead, it reduced the width of the confidence intervals.

Checkpoint	n = 3 seeds	n = 6 seeds	n = 8 seeds	n = 10 seeds
A	0.642 [0.576, 0.696], w = 0.121	0.641 [0.598, 0.679], w = 0.082	0.641 [0.606, 0.676], w = 0.070	0.641 [0.608, 0.671], w = 0.063
B	0.674 [0.643, 0.718], w = 0.075	0.675 [0.651, 0.705], w = 0.053	0.674 [0.653, 0.697], w = 0.044	0.675 [0.656, 0.697], w = 0.041
E	0.636 [0.612, 0.665], w = 0.053	0.636 [0.617, 0.656], w = 0.039	0.636 [0.620, 0.655], w = 0.035	0.636 [0.620, 0.652], w = 0.032
G	0.669 [0.630, 0.706], w = 0.075	0.669 [0.644, 0.695], w = 0.051	0.669 [0.644, 0.691], w = 0.047	0.668 [0.647, 0.689], w = 0.042

H	0.668 [0.613, 0.712], w = 0.099	0.669 [0.632, 0.706], w = 0.074	0.668 [0.636, 0.698], w = 0.061	0.669 [0.642, 0.695], w = 0.053

n seeds	Mean CI width	Reduction vs n = 3
3	~0.085	—
6	~0.057	−33%
8	~0.051	−40%
10	~0.046	−46%

This analysis supports the use of ten seeds for uncertainty quantification diagnostics. The main qualitative conclusion is already visible with fewer seeds, but ten seeds provide narrower confidence intervals and more stable reporting.

Calibration of Binary Phenotypic Predictions

We additionally evaluated calibration quality for the binary Stokes inhibition endpoint. Calibration was assessed using expected calibration error, mean absolute calibration error, and Brier score. Single NIG showed the lowest calibration error, whereas EoE-5 and EoE-10 showed slightly worse pooled calibration despite improving selected ranking and regression metrics for the Stokes endpoint.

Method	ECE ↓ mean ± std	ECE ↓ pooled	MACE ↓ pooled	Brier Score ↓ pooled
Single NIG	0.024 ± 0.010	0.015	0.015	0.057
EoE-5	0.043 ± 0.027	0.026	0.026	0.057
EoE-10	0.050 ± 0.030	0.035	0.035	0.059

These results indicate that ensembling did not automatically improve probability calibration. In this benchmark, Single NIG provided the best calibrated binary probabilities, while EoE models were more useful for stabilizing uncertainty-based ranking and inter-stage mutual information estimates.

Selective Prediction Using Predictive Uncertainty

We next evaluated whether predictive uncertainty could be used for selective prediction. In this analysis, compounds were sorted by increasing uncertainty, and RMSE was calculated for progressively larger retained subsets. A useful uncertainty estimate should produce lower error at low coverage, because only the most confident predictions are retained.

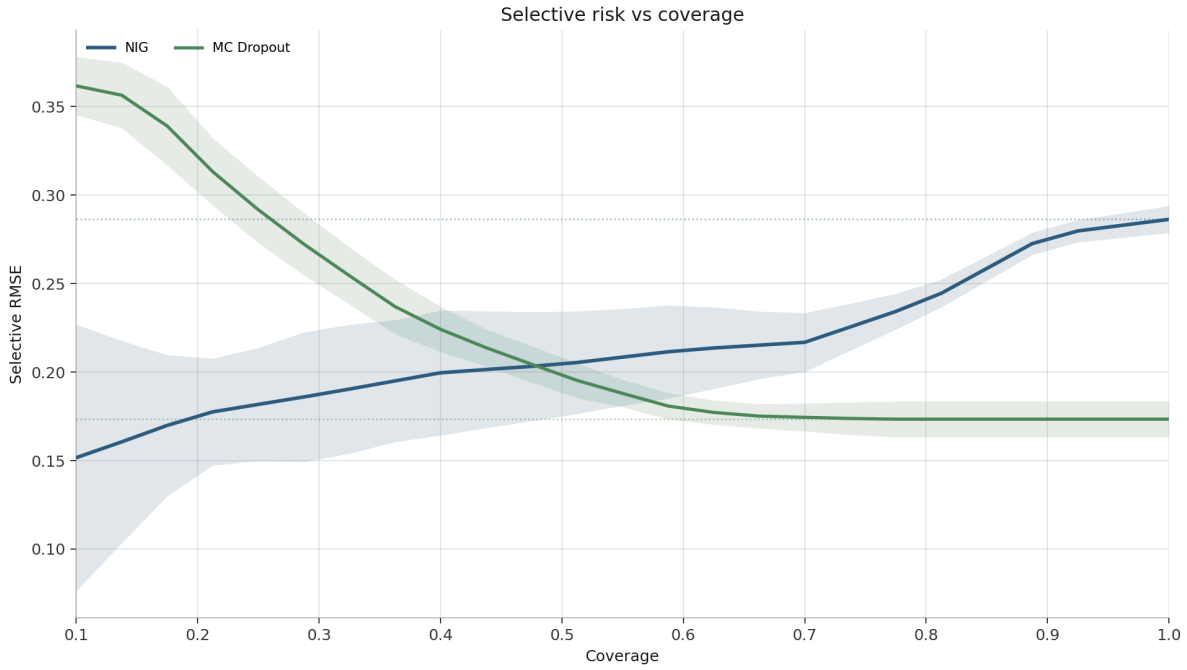


Figure 8. Selective risk versus coverage for uncertainty-based selective prediction. Selective RMSE was calculated after sorting compounds by increasing predictive uncertainty. NIG uncertainty produced the expected selective-prediction behaviour: retaining only low-uncertainty compounds reduced RMSE, and error increased as coverage approached the full dataset. MC Dropout showed the opposite pattern, with high selective RMSE at low coverage, indicating an unreliable or inverted uncertainty ranking.

The NIG model followed the expected pattern. At low coverage, where only the lowest-uncertainty predictions were retained, selective RMSE was substantially lower. As coverage increased and more uncertain compounds were included, RMSE increased toward the full-dataset error level.

MC Dropout showed the opposite behaviour. Selective RMSE was highest at low coverage and decreased as more compounds were included. This indicates that MC Dropout did not provide a useful uncertainty ranking in this setting. In practical terms, it would reject many compounds that are not necessarily the highest-risk predictions, making it unsuitable as the primary uncertainty mechanism for early-exit control.

This result supports the use of NIG-based uncertainty for selective prediction and cost-aware pipeline decisions.

Sigma Tracks Prediction Risk but Is Not a Direct Error Estimator

We further analyzed the relationship between NIG uncertainty, denoted σ, and absolute prediction error. The relationship between σ and growth-ratio error showed a clear and reproducible structure, but it was not a simple deterministic mapping. Low σ values were generally associated with lower error, whereas high σ values were associated with higher error. However, some compounds still showed high error despite low predicted uncertainty.

This means that NIG σ is useful as a ranking-based proxy for prediction risk, but it should not be interpreted as a directly calibrated estimate of absolute error. In other words, σ can help decide which compounds are safer to stop early and which should continue to the full pipeline, but it should not be used as a literal numerical prediction of the expected error.

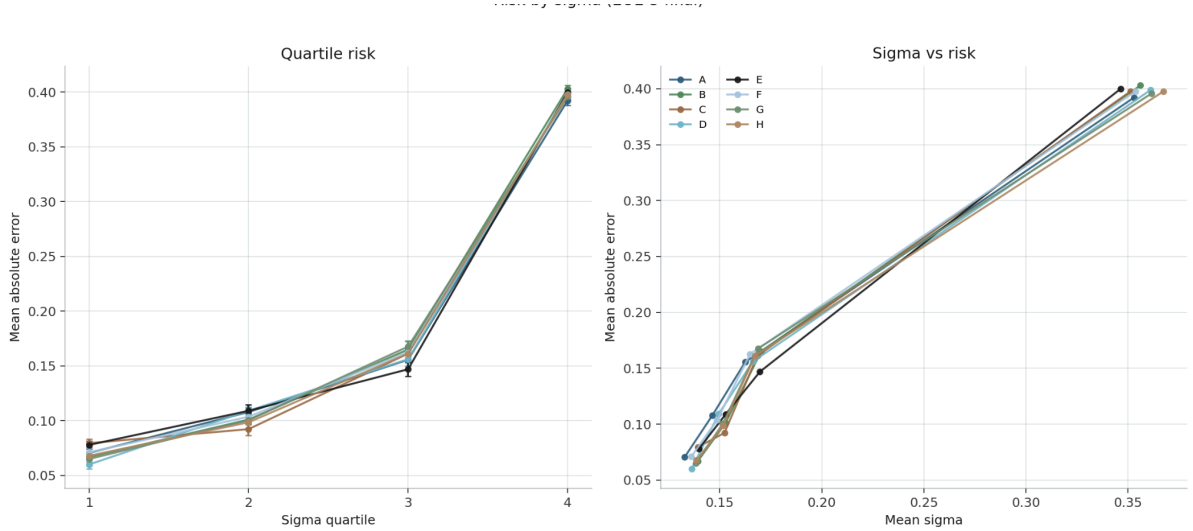


Figure 9. Prediction error increases across NIG sigma quartiles.
Compounds were grouped into quartiles according to NIG σ. Mean absolute growth-ratio error increased consistently from the lowest to the highest σ quartile across checkpoints, indicating that NIG uncertainty provides a useful ranking proxy for prediction risk. The relationship was consistent across stages, supporting the use of σ for selective prediction and early-exit policies.

The quartile analysis showed a strong and consistent increase in mean absolute error across σ quartiles. This confirms that, although σ is not a perfectly calibrated error estimate, it provides a practically useful risk ranking. Across checkpoints, higher σ corresponded to higher mean absolute error.

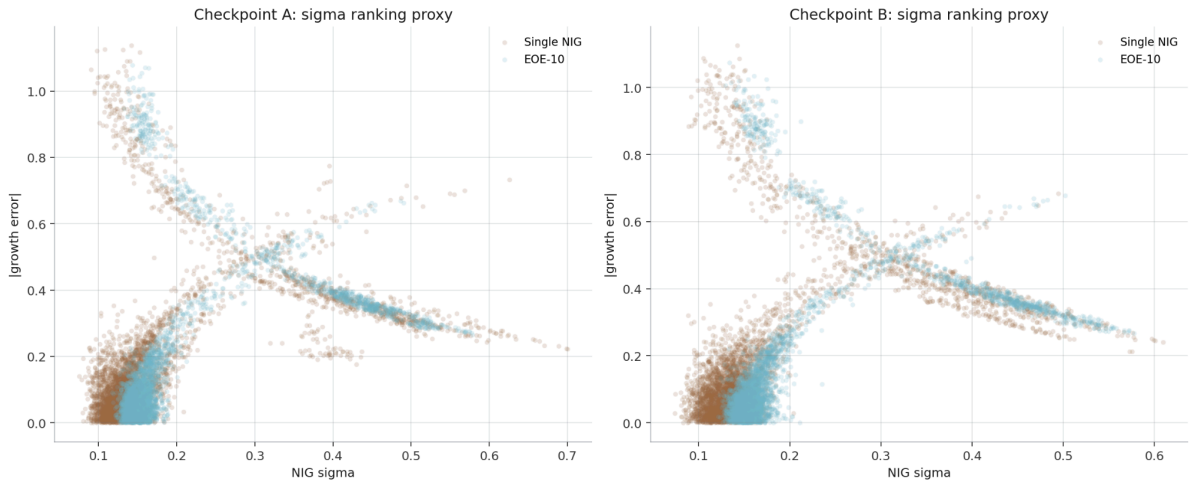


Figure 10. Relationship between NIG uncertainty and absolute growth-ratio prediction error at checkpoints A and B.

Scatter plots show the relationship between NIG σ and absolute growth-ratio prediction error for Single NIG and EoE-10. The relationship has a repeated structure but is not strictly monotonic at the individual-compound level. EoE-10 stabilizes the distribution relative to Single NIG, supporting its use in uncertainty-aware selective prediction and early-exit control.

The scatter plots for checkpoints A and B further show that EoE-10 produces a more stable uncertainty-error structure than Single NIG. This supports the use of EoE models when the goal is not only point prediction, but also robust uncertainty ranking across pipeline stages.

Discussion

The main finding of this study is that uncertainty-aware early-exit strategies can reduce the cost of a modular drug design pipeline while preserving recovery of pipeline-defined hits. The strongest practical result is not that a surrogate model can replace the full pipeline. It cannot. Rather, the result shows that surrogate models can act as decision controllers between pipeline stages, allowing some compounds to be stopped early while uncertain or promising compounds continue to more expensive downstream calculations.

Our initial expectation was that the evaluated methods would differ more dramatically, with some strategies approaching complete failure and others clearly dominating the benchmark. Instead, the observed differences were more nuanced. Several methods achieved high or complete hit recovery, and the main distinction between them was the amount of compute saved and the reliability of their uncertainty estimates. This suggests that the benchmark is partly saturated: many strategies can recover the available hits under the current class distribution, so the practical comparison shifts toward cost, uncertainty ranking, and deployment risk.

The Pareto analysis showed that aggressive strategies such as RF + Morgan and score thresholding can save substantially more compute than evidential uncertainty-based approaches. These methods are attractive for early-stage library triage, where the goal is to quickly remove obviously weak candidates. However, they should not be interpreted as uncertainty-aware models. Their value is mainly operational: they are simple, fast, and effective as low-cost filters.

In contrast, Evidential + MIST-QR represents a more conservative operating point. It provides lower compute savings but maintains complete hit recovery and offers a principled uncertainty signal. This makes it more appropriate for decision points where false rejection of potentially useful compounds is costly. In a production pipeline, this model would be most useful as a gate after intermediate biological or structural information is available, for example after docking and Boltz/KD prediction, but before GROMACS and systems-level analysis.

The selective prediction results support this interpretation. NIG uncertainty behaved as expected: when the most uncertain compounds were excluded, prediction error decreased. This confirms that NIG σ can be used as a ranking proxy for prediction risk. MC Dropout did not show the same behaviour. Its selective risk was highest at low coverage, which indicates

that its uncertainty ranking was not reliable in this setting. Therefore, MC Dropout should remain a comparison baseline rather than the main deployment mechanism.

The σ-error analyses also show an important limitation. NIG σ is useful for ranking risk, but it is not a direct calibrated estimator of absolute error. Some compounds still had high error despite low uncertainty. This matters for deployment: σ should be used to define conservative decision policies, not to guarantee that individual predictions are correct. A realistic early-exit system should therefore combine uncertainty with additional applicability-domain checks, such as chemical similarity to the training set or explicit OOD detection.

The inter-stage mutual information analysis further supports the value of ensemble-based evidential models. Single NIG produced weaker and less structured MI signals between pipeline stages, whereas EoE-5 and EoE-10 produced clearer inter-stage dependencies. This suggests that EoE-style ensembling stabilizes the estimation of information flow between stages. In practice, this helps identify where an early-exit gate is likely to be meaningful, because it shows which stages contain predictive information about later pipeline outputs.

A major caveat is that the benchmark uses an enriched dataset rather than a fully realistic drug-discovery screening distribution. In real screening, the true hit rate may be extremely low, for example close to 0.1% or lower depending on the target, library, and assay. In contrast, the current benchmark contains a much higher fraction of candidate-like or active compounds. This makes it easier for methods to saturate hit recovery and may reduce the apparent difficulty of the task. Therefore, the results should not be interpreted as proof that the system can recover rare true positives from a massive pool of negatives under fully realistic screening conditions.

Instead, the correct interpretation is narrower: under an enriched benchmark built from experimental data and PROTO-NOOS-generated labels, uncertainty-aware surrogate models can reduce pipeline cost while preserving the available hit set. This is still valuable, but it should be described as an approximate and enriched benchmark rather than a final simulation of real-world drug discovery.

Several additional limitations should be noted. First, the study was constrained by the limited availability of large experimental datasets with verified true-positive and true-negative labels for the PROTO-NOOS setting. We did not have laboratory validation showing the actual hit rate of PROTO-NOOS over a large number of screened compounds. Second, the data were heterogeneous, combining biochemical affinity data, phenotypic growth inhibition data, and pipeline-generated computational labels. These sources differ in biological meaning, noise level, and assay resolution. Third, the models were trained under practical time and methodological constraints, which limited the extent of hyperparameter exploration and architecture comparison.

Finally, this study did not include a broad comparison against all established uncertainty quantification methods or external drug-discovery benchmarks. The results therefore support a specific claim about the tested PROTO-NOOS setting, rather than a universal claim about all drug-design pipelines.

Conclusions

This study shows that uncertainty-aware surrogate models can support early-exit decisions in a modular drug-design pipeline, but the role of each model family differs.

Single NIG provided the strongest per-prediction uncertainty diagnostics. It achieved the best association between predictive uncertainty and absolute error, the lowest calibration error, and the highest KSG-estimated mutual information between σ and prediction error. Therefore, Single NIG is the preferred option when the main objective is direct uncertainty-based ranking of individual predictions.

EoE-5 and EoE-10 were more informative at the pipeline-analysis level. In particular, ensemble-based evidential models produced clearer inter-stage mutual-information structure than Single NIG, suggesting that ensembling can stabilize the analysis of information flow between pipeline stages. EoE-5 is therefore best interpreted as a practical conservative deployment candidate, not as the best per-prediction uncertainty model.

RF + Morgan and score-threshold strategies achieved larger compute savings and are useful as aggressive low-cost filters. However, they should not be treated as principled uncertainty models.

NIG uncertainty is useful for ranking prediction risk, but it should not be interpreted as a directly calibrated estimate of absolute error. In addition, σ alone is insufficient for OOD detection and should be combined with an applicability-domain guard, such as nearest-neighbor Tanimoto similarity.

The main limitation is that the benchmark is enriched and does not yet reproduce the extremely low true-positive rate expected in fully realistic drug-discovery screening. Future validation should include larger experimental datasets, external benchmarks, and laboratory testing of PROTO-NOOS-selected compounds.

Claims vs evidence

THIS SECTION WILL BE UPDATED

Code and Data Availability

THIS SECTION WILL BE UPDATED

[Code availability: repository link, commit hash, license, and environment specification file to be provided. Data availability: ChEMBL and BindingDB data are publicly available; de novo compound labels are generated outputs of PROTO-NOOS. Model weights and inference

outputs: availability to be specified. All train/validation/test splits and evaluation CSV files required for result reproduction will be deposited.]

References

THIS SECTION WILL BE UPDATED

1. Stokes, J. M., Yang, K., Swanson, K., Jin, W., Cubillos-Ruiz, A., Donghia, N. M., ... & Collins, J. J. (2020). A deep learning approach to antibiotic discovery. Cell, 180(4), 688–702.

2. Mendez, D., Gaulton, A., Bento, A. P., Chambers, J., De Veij, M., Félix, E., ... & Leach, A. R. (2024). ChEMBL database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. Nucleic Acids Research, 52(D1), D1180–D1187.

3. Gilson, M. K., Liu, T., Baitaluk, M., Nicola, G., Hwang, L., & Chong, J. (2024). BindingDB in 2024: A FAIR knowledgebase of protein–small molecule binding data. Nucleic Acids Research, 53(D1), D1633–D1641.

4. Blaschke, T., Arus-Pous, J., Chen, H., Margreitter, C., Tyrchan, C., Engkvist, O., ... & Patronov, A. (2020). REINVENT 2.0: An AI tool for de novo drug design. Journal of Chemical Information and Modeling, 60(12), 5918–5922.

5. Zhou, G., Gao, Z., Ding, Q., Zheng, H., Xu, H., Wei, Z., ... & Ke, G. (2023). Uni-Mol: A universal 3D molecular representation learning framework. OpenReview (ICLR 2023).

6. Amini, A., Schwarting, W., Soleimany, A., & Rus, D. (2020). Deep evidential regression. Advances in Neural Information Processing Systems, 33, 14927–14937.

7. Gretton, A., & others (2023). MIST: Mutual information via supervised training. arXiv:2511.18945.

8. Rogers, D., & Hahn, M. (2010). Extended-connectivity fingerprints. Journal of Chemical Information and Modeling, 50(5), 742–754.

9. McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform manifold approximation and projection for dimension reduction. Journal of Open Source Software, 3(29), 861.

10. Zhao, Y., Nasrullah, Z., & Li, Z. (2019). PyOD: A Python toolbox for scalable outlier detection. Journal of Machine Learning Research, 20(96), 1–7.

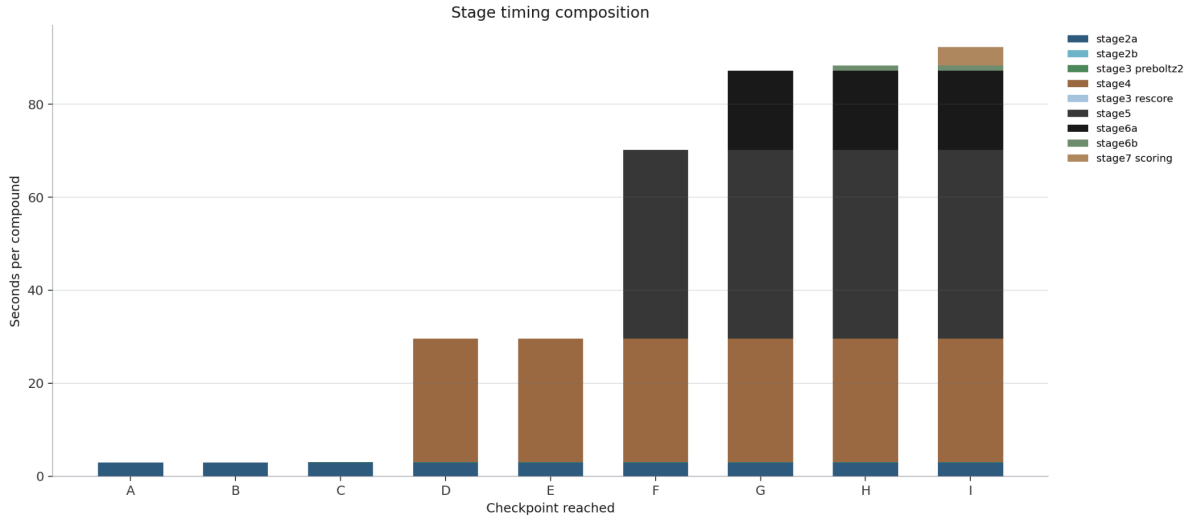
11. Blank, J., & Deb, K. (2020). pymoo: Multi-objective optimization in Python. IEEE Access, 8, 89497–89509.

12. Shafer, G., & Vovk, V. (2008). A tutorial on conformal prediction. Journal of Machine Learning Research, 9, 371–421.

Supplementary Information

Supplementary Runtime Analysis

To quantify where computational cost accumulates in the PROTO-NOOS workflow, we analyzed the mean runtime contribution of each pipeline stage. The description of each stage follows the PROTO-NOOS system specification.



Supplementary Figure S1. Stage-wise runtime composition of the PROTO-NOOS pipeline.
Runtime contribution of individual stages in the PROTO-NOOS pipeline, reported as seconds per compound. Most of the computational cost is concentrated in later stages, particularly Boltz2/KD prediction, GROMACS-based molecular dynamics, and systemic activity modelling. This uneven runtime distribution explains why early-exit gates placed before late-stage calculations can save substantial compute even when only a subset of compounds is stopped early.

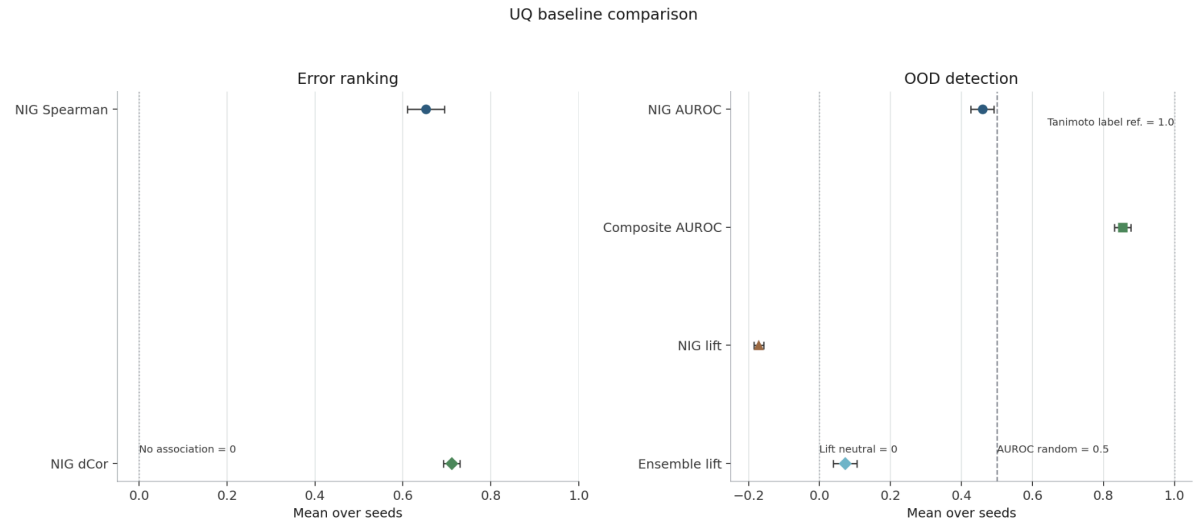
Stage	Description	Mean (s)	Median (s)	P90 (s)	% of Total
2a	Docking, initial scoring	2.90	1.70	5.18	3.1%
2b	Docking rescore	0.076	0.07	0.10	0.1%
3, pre-Boltz	Pre-Boltz2 processing	0.070	0.06	0.08	0.1%

3, rescore	Docking rescore pass	0.063	0.05	0.08	0.1%
4	Boltz2 / KD prediction	26.53	26.91	27.68	28.7%
5	MD / growth-ratio calculation, GROMACS	40.53	38.42	41.47	43.9%
6a	BioTransformer / systemic activity	17.07	11.93	24.47	18.5%
6b	Systemic scoring	1.11	1.10	1.14	1.2%
7	Final scoring	4.00	4.00	4.00	4.3%
Total	Full pipeline	~92.4	—	—	100%

This analysis supports the placement of early-exit gates before the most expensive downstream stages. In particular, gates placed after intermediate structural or affinity information, but before GROMACS and systemic modelling, have the highest practical value.

Supplementary UQ Baseline Comparison

We evaluated whether NIG uncertainty can be trusted as a proxy for prediction risk and whether it is sufficient for out-of-distribution detection.



Supplementary Figure S2. UQ baseline comparison for error ranking and OOD detection.
The left panel shows that NIG uncertainty has a strong positive association with prediction error, measured by Spearman correlation and distance correlation. The right panel shows that NIG uncertainty alone is not sufficient for OOD detection. A composite score combining NIG uncertainty with nearest-neighbor Tanimoto novelty provides stronger OOD detection, indicating that model uncertainty and chemical applicability-domain information are complementary.

NIG uncertainty showed a clear positive relationship with prediction error. Across checkpoints, both Spearman correlation and distance correlation were consistently above 0.5, indicating that higher uncertainty was associated with larger absolute prediction error.

However, NIG uncertainty alone was not sufficient for reliable OOD detection. Its AUROC was close to, or below, the random reference level. OOD lift was also weak or negative. In contrast, the composite OOD score, which combined predictive uncertainty with chemical novelty measured by nearest-neighbor Tanimoto similarity, achieved stronger OOD discrimination.

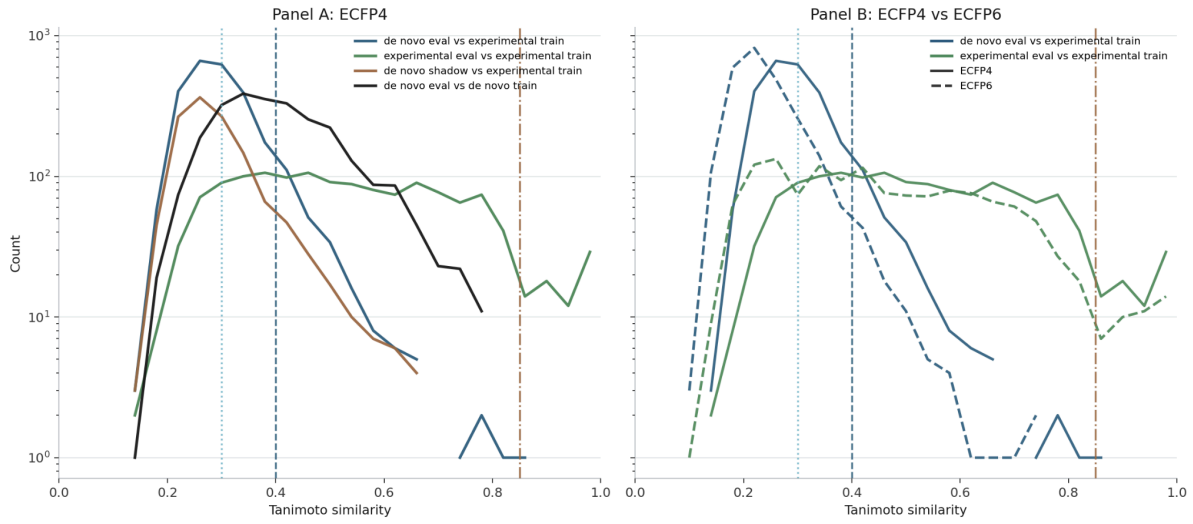
This result shows that predictive uncertainty and chemical applicability-domain information are complementary. NIG uncertainty can rank prediction risk, but it should not replace explicit chemical novelty or OOD checks.

Checkpoint	NIG Spearman ρ ↑	NIG dCor ↑
A, early post-docking	0.641 ± 0.055	0.703 ± 0.019

B	0.674 ± 0.036	0.724 ± 0.021
C	0.631 ± 0.039	0.706 ± 0.020
D	0.657 ± 0.055	0.711 ± 0.022
E, mid-pipeline	0.636 ± 0.027	0.713 ± 0.013
F	0.642 ± 0.041	0.711 ± 0.018
G	0.668 ± 0.037	0.706 ± 0.018
H, late pre-final	0.668 ± 0.047	0.713 ± 0.018

Supplementary Chemical Similarity Analysis

We analyzed nearest-neighbor Tanimoto similarity distributions to compare experimental evaluation compounds, de novo evaluation compounds, and de novo shadow compounds against the training set.



Supplementary Figure S3. Similarity distributions between evaluation compounds and training compounds.

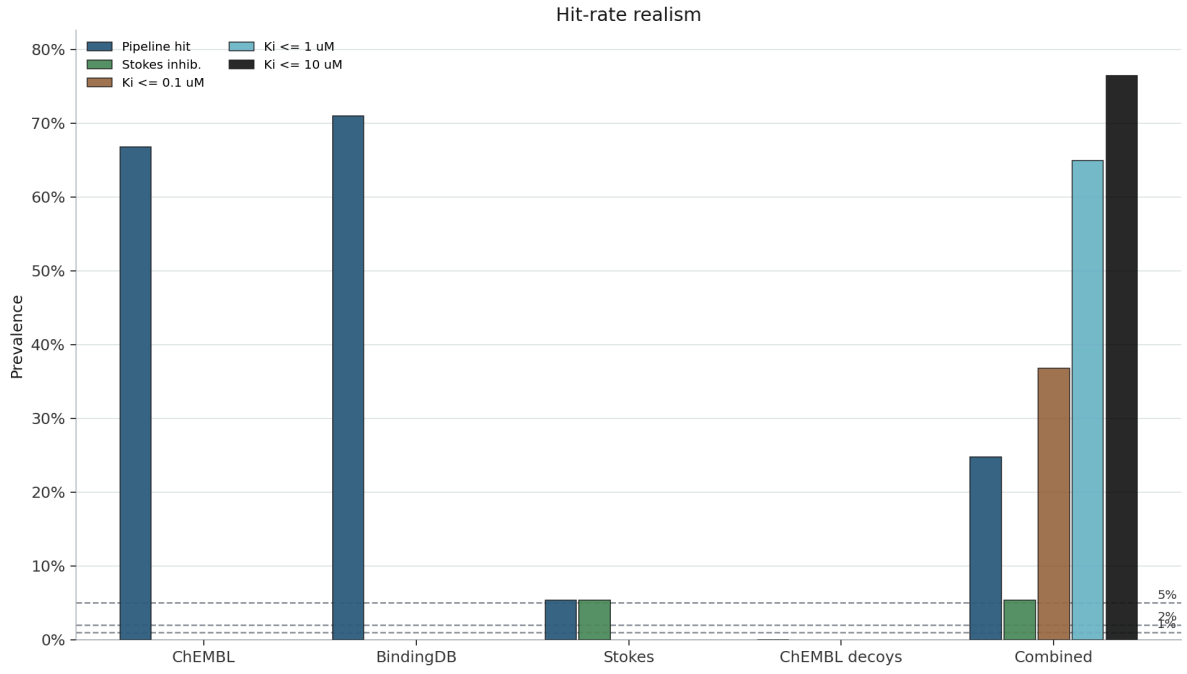
Panel A shows nearest-neighbor Tanimoto similarity distributions using ECFP4 fingerprints. Panel B compares ECFP4 and ECFP6. De novo evaluation compounds are shifted toward lower similarity relative to experimental evaluation compounds, indicating that the de novo benchmark is more out-of-distribution. ECFP6 shifts distributions leftward, as expected, because radius-3 fingerprints require a stricter match of the local atomic environment.

The experimental evaluation compounds showed higher similarity to the experimental training set than the de novo evaluation compounds. This indicates that the de novo benchmark is chemically more challenging and more out-of-distribution.

The comparison between ECFP4 and ECFP6 provides an additional control. ECFP6 fingerprints shifted similarity distributions toward lower values, which is expected because radius-3 fingerprints encode a larger atomic neighborhood and therefore require stronger local structural agreement. The consistency between the ECFP4 and ECFP6 trends supports the conclusion that the de novo set is less similar to the experimental training distribution.

Supplementary Hit-Rate Realism Analysis

We analyzed the prevalence of positive labels across data sources and activity definitions to assess how closely the benchmark resembles realistic drug-discovery screening.



Supplementary Figure S4. Hit-rate realism across experimental and combined data sources.
Prevalence of positive labels across ChEMBL, BindingDB, Stokes, ChEMBL Decoys, and the combined dataset. Dashed horizontal lines indicate low-hit-rate reference levels. The benchmark is enriched relative to realistic large-scale screening, where the true hit rate is often much lower. The positive-label categories are not mutually exclusive, so combined values should not be interpreted as summing to 100%.

The benchmark contains a substantially higher prevalence of active or candidate-like compounds than expected in fully realistic large-scale drug discovery screening. ChEMBL and BindingDB are strongly enriched because they contain compounds already associated with biochemical activity measurements. The Stokes dataset has a lower active fraction and better reflects a phenotypic screening context, but it is still not equivalent to a prospective ultra-low-hit-rate screen.

This matters for interpretation. Because the benchmark is enriched, several methods can reach saturated or near-saturated hit recovery. Therefore, the benchmark is suitable for evaluating early-exit behaviour under controlled enriched conditions, but it should not be presented as a complete simulation of real-world screening where true positives may be extremely rare.

This analysis belongs either in the Supplementary Information or in the Methods subsection describing dataset limitations. In the main Discussion, it should be used as a limitation: the current benchmark demonstrates cost reduction under enriched conditions, not final prospective discovery performance under a realistic ultra-low hit rate.

Software Used in the Experiments

The following software packages and libraries were used in the experiments.

Software	Role in the study
`mist-statinf` / MIST-QR	Mutual information estimation between uncertainty and absolute error
`MAPIE`	Conformal prediction, including `SplitConformalRegressor` and `ResidualNormalisedScore`
`PyOD`	OOD detection baselines, including autoencoder-based detection
`Weights & Biases`	Experiment tracking and logging
`umap-learn`	Chemical-space visualization

`RDKit`	Molecular fingerprints, ECFP4/ECFP6, Tanimoto similarity, SMILES handling, InChIKey generation
`SciPy`	Spearman correlation and numerical utilities
Custom dCor implementation	Distance correlation analysis
`pymoo`	Non-dominated sorting for Pareto frontier analysis
`NetworkX`	Graph-based runtime and pipeline analyses