# Machine Learning IN PROTO-NOOS:

## Rational Compute Allocation in Antibiotic Discovery via Mutual Information & Uncertainty Quantification

Konrad Gorzelańczyk[1○], Szymon Krawczyk[1], Florian Hołubowski[1], Maksymilian Korbik[1]

[1]Faculty of Computing and Telecommunications, Poznan University of Technology, Piotrowo 3a, 60-965 Poznań, Poland

○ Correspondence: sknwpl@proton.mail

PUT Drug Discovery x Ghostday2026

01 MOTIVATION

Multistage computational pipelines in drug discovery are costly because they often pass all compounds through the same stages, regardless of what is already known from earlier calculations. The motivation behind this project was to test whether mutual information and uncertainty quantification can indicate the point at which further computations can be safely stopped without losing the most important hits. Therefore, the goal was to develop an early-exit strategy for the PROTO-NOOS system, enabling faster exploration of larger chemical space under a limited computational budget.

***Is knowing when to stop computing as important as knowing what to compute?***

**779** CURATED INHIB. | **2,335** PHENOTYPIC CMPDS | **13,531** DE NOVO MOLECULES

02 METHOD

We benchmarked early-exit strategies comparing Single NIG, EOE-5, and EOE-10 surrogate models with uncertainty prediction. Models were evaluated on predefined seeds: 6 for Pareto frontier analysis, 8 for ranking quality, and 10 for UQ diagnostics. Results are reported as mean ± SD, DDOF = 1. Performance was assessed using hit recovery, compute saved, AUROC, RMSE, Spearman's ρ, distance correlation, and ECE. Mutual information was estimated with MIST-QR and compared against KSG. Result uncertainty was estimated using 1000 bootstrap replicates.

VISUALIZATION 1

FIG 1

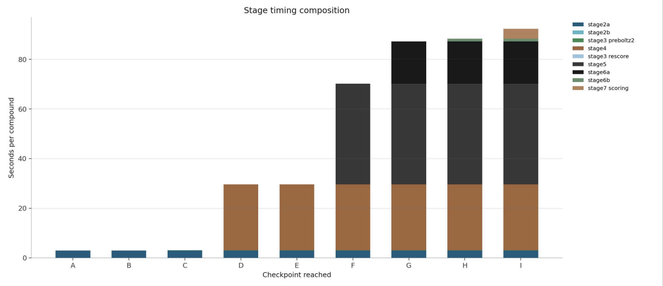



**Fig. 1.** Timings of the specific system stages of the drug discovery. Most pipeline cost concentrates in late stages (GROMACS, Boltz2 and Biotransformer 3)

1 / 5

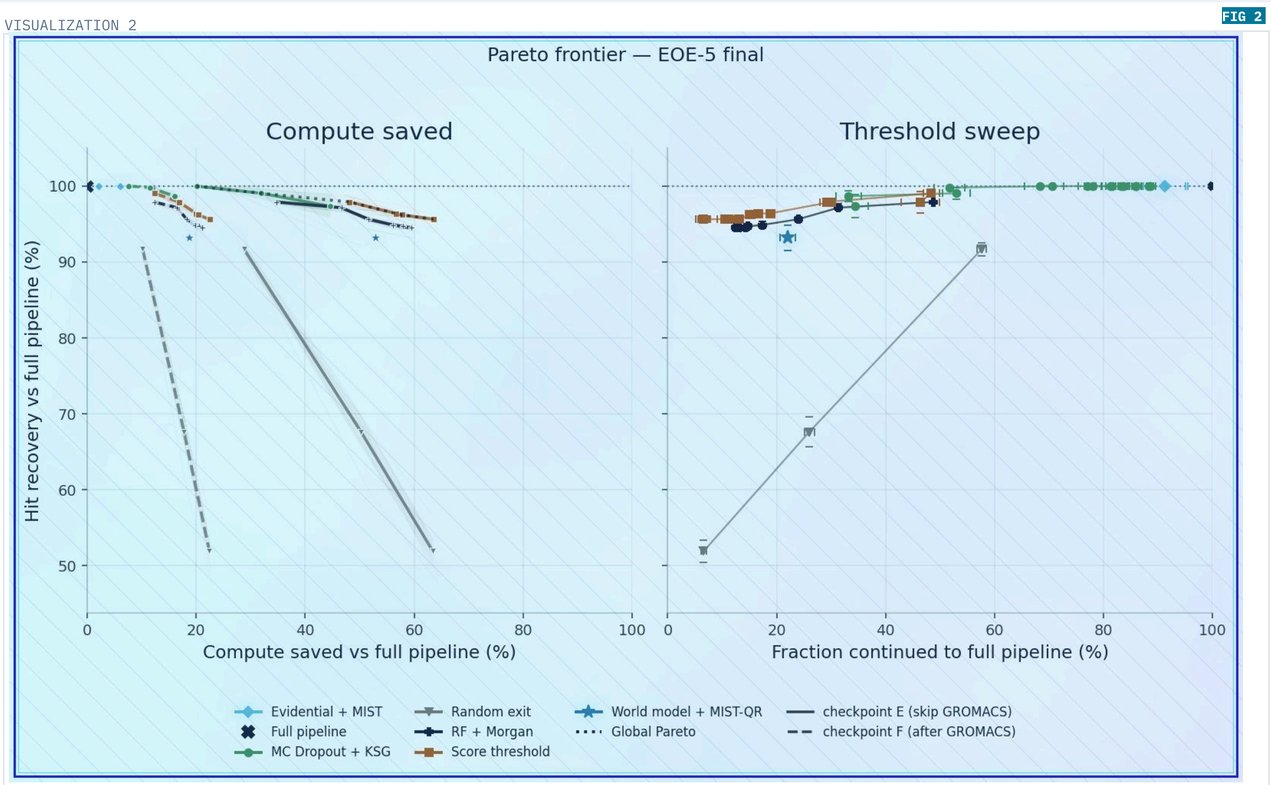
03 MODELS COMPARED

| |
|---|
| Single NIG |
| EOE-5 |
| EOE-10 |

Three model families compared: Single NIG (one forward pass), EoE-5 and EoE-10 (5- and 10-member ensembles of independently trained NIG models). All use UniMol encoder.

VISUALIZATION 2

FIG 2



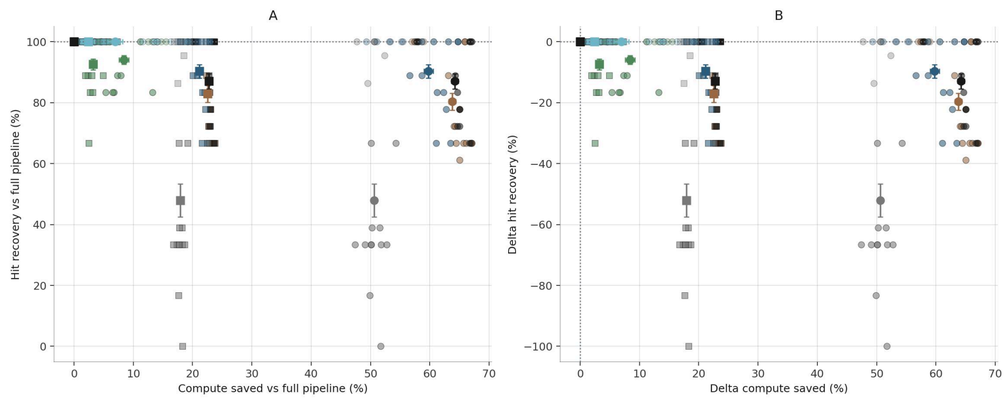




**Fig. 2.** Pareto frontier of early-exit strategies, showing the trade-off between compute saved and hit recovery relative to the full pipeline.

**Fig. 3.** Comparison of early-exit methods across checkpoints, shown as absolute performance (A) and as change relative to the full pipeline (B).

2 / 5

VISUALIZATION 3

FIG 4

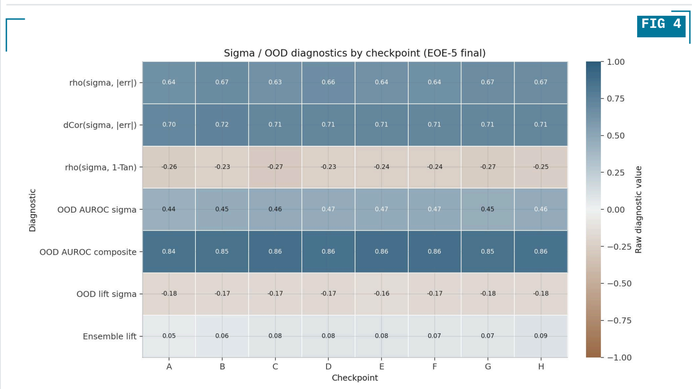



**Fig. 4.** EOE-5 uncertainty σ correlates consistently with prediction error across checkpoints A–H, but σ alone fails as an OOD detector. A composite score combining σ with a Tanimoto-based applicability guard restores strong OOD separation.

3 / 5

04 Pareto Frontier

Evidential + MIST-QR occupies a conservative Pareto point: full hit recovery at ~6% compute saved. More aggressive thresholds (Score threshold, RF + Morgan) save ~58% compute but sacrifice uncertainty reliability in selective prediction.

VISUALIZATION 4

FIG 5

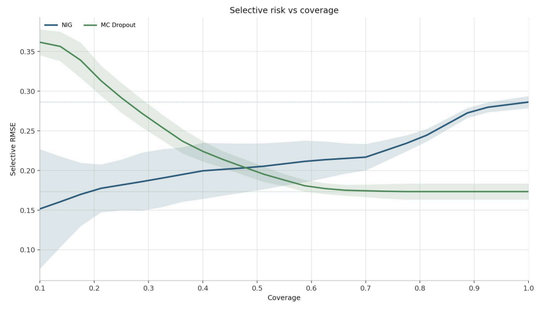



**Fig. 5.** Selective RMSE vs. coverage for NIG and MC Dropout uncertainty estimators. Lower selective RMSE at reduced coverage indicates correct uncertainty ranking.

4 / 5

01 **TABLE**

| Method(EoE-5) BEST **BOLDED** | 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(Cheap.) | **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 (Conservative)** | 6.0±6.5 | **100.0±0.0** | **1.000±0.001** |

01 **TABLE**

Table 1: Best values are bolded. All metrics are averaged over some amount of seeds unless specified otherwise all of the benchmarks were done on the same dataset with same amount of seeds. Averaged over 6 seeds (Pareto), 8 seeds (ranking), 10 seeds (UQ diagnostics) per pre-registered protocol.

02 **TABLE**

| Method BEST **BOLDED** | Single NIG | EOE-5 | EOE-10 |
|---|---|---|---|
| Spearman ρ(σ, \|ε\|) (↑) | **0.452±0.098** | 0.409±0.053 | 0.444±0.050 |
| dCor(σ, \|ε\|) (↑) | **0.514±0.076** | 0.468±0.046 | 0.511±0.042 |
| OOD AUROC σ (↑) | **0.576±0.043** | 0.569±0.045 | 0.570±0.037 |
| ECE, 10 adaptive bins(↓) | **0.030±0.011** | 0.042±0.023 | 0.049±0.027 |
| MI(σ, \|ε\|) KSG (↑) | **0.289±0.087** | 0.186±0.039 | 0.207±0.053 |

02 **TABLE**

Table 2: Comparison of metrics from SINGLE NIG, EOE-5, EOE-10. Best are bolded. ECE computed per seed (10 adaptive quantile bins, n≈349 per bin), reported as mean±SD across 10 seeds. Pooled-prediction ECE shown in Supplementary.

05 OOD / UQ DIAGNOSTICS

NIG uncertainty is useful for ranking prediction risk. However, σ alone gives poor OOD detection. Combining σ with chemical novelty through a Tanimoto guard increases OOD AUROC to 0.84–0.86, showing that model uncertainty and chemical applicability are complementary signals.

VISUALIZATION 5

FIG 6

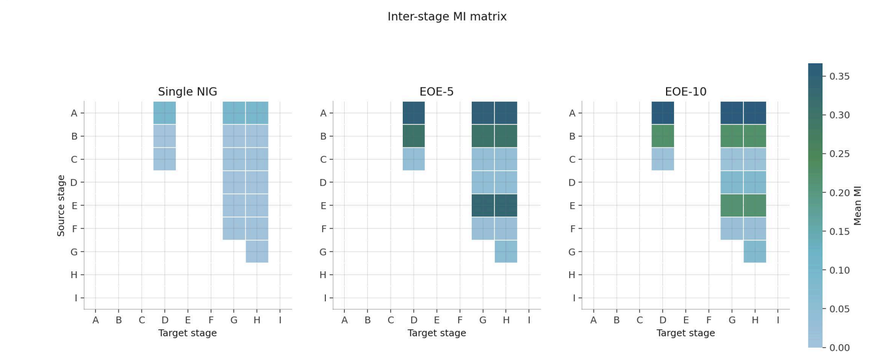


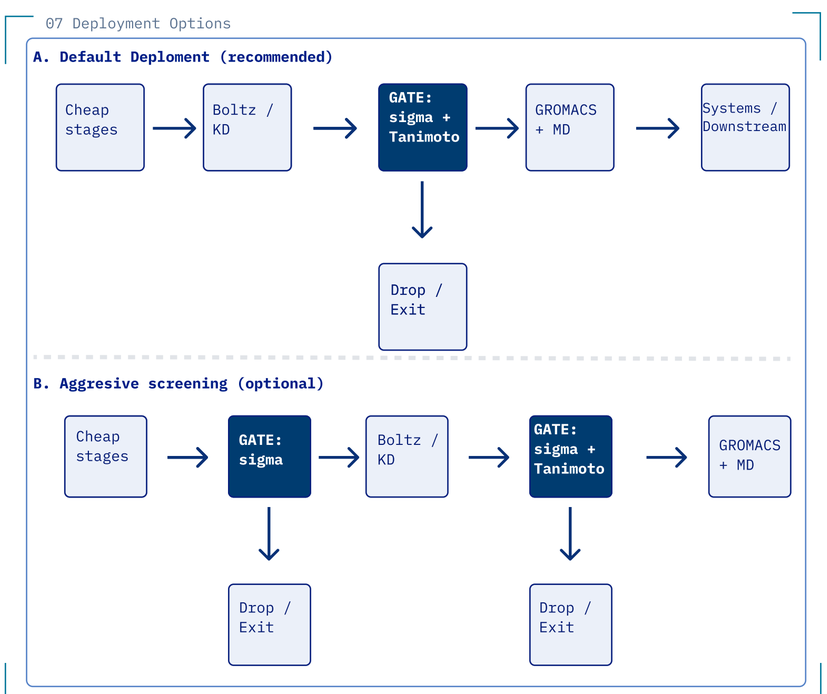

**Fig. 6.** Inter-stage mutual information matrix (MIST-QR) across pipeline checkpoints A–I for Single NIG, EOE-5, and EOE-10. Higher values indicate stronger predictive coupling between stages.

5 / 5

06 MI estimates

EOE-5 produces substantially higher inter-stage MI estimates relative to Single NIG, particularly at checkpoints D and E–H. This indicates that ensemble aggregation stabilises mutual information estimation across pipeline stages, rather than simply averaging individual NIG heads. EOE-10 confirms this direction with reduced variance. Single NIG shows near-zero MI for most stage pairs, suggesting that a single forward pass is insufficient to capture inter-stage predictive dependencies reliably.

07 Deployment Options




08 Deployment Steps

1. Choose the model: Use Single NIG for production gating. It is cheapest to train and gave the best per-prediction UQ in our benchmark. EOE-5/10 are mainly for diagnostic analysis of inter-stage information flow, not default deployment.
2. Place the gate: Default deployment uses one main gate before Boltz/KD and before GROMACS. Optional aggressive screening can add an earlier gate before Boltz/KD.
3. Set the threshold: Start with the upper-quantile σ threshold on a held-out set. In our conservative setting, this kept 100% hit recovery with 6% compute saved. Relaxing the threshold can save more compute, but increases risk.
4. Add OOD guard: σ alone is insufficient for OOD detection. Use a composite score: OOD_score = α · O_norm + (1−α) · (1 − max Tanimoto to train), with α ≈ 0.5. Normalize both terms to [0,1] before combining.
5. Validate before deployment: On your held-out set, verify hit recovery, compute saved, and uncertainty calibration before using the gate in production.

09 CONCLUSIONS

## Mutual information + uncertainty = principled compute allocation.

01 All UQ-based methods preserve 100% hit recovery. The difference between methods is not in hit recovery, but in the trade-off between compute saved and UQ reliability, measured by AUROC and selective risk. This reframes the method ranking beyond the classical "hit recovery vs cost" view.

02 Per-prediction UQ: Single NIG is sufficient. It achieves the best Spearman's ρ(σ, |ε|) = 0.452, dCor = 0.514, ECE = 0.030, and MI(σ; |ε|) = 0.289 nats. The ensemble does not improve uncertainty quality for individual molecules. For selective prediction and early-exit gating, Single NIG is the appropriate choice.

03 σ alone is not sufficient for OOD detection. AUROC ~0.57 across all three models is close to a random baseline. OOD detection requires a composite score combining σ with a chemical applicability guard, such as Tanimoto < 0.30. Model uncertainty and chemical novelty are complementary signals, not interchangeable ones.

04 MIST-QR reveals inter-stage dependencies that are not visible in a single model. Single NIG shows near-zero MI between stages, whereas EOE-5 and EOE-10 provide a meaningful signal. The ensemble not only averages predictions, but also stabilizes MI estimation.

**OOD AUROC 0.57 (σ ALONE INSUFFICIENT)** | **ECE 0.030 (SINGLE NIG)** | **I(σ; |ε|) = 0.29 NATS** | **6% to 58% COMPUTE SAVED**

PREPRINT x REPOSITORY

PREPRINT

https://komputerowe-projektowanie-lekow.github.io/webpage/

CODE

https://github.com/