

Test Time Training for Sample-Efficient Practical Molecular Optimization

Stanford CS224N Custom Project

Aaron Lee

Department of Computer Science
Stanford University
aaroncl@stanford.edu

Ishvi Mathai

Department of Computer Science
Stanford University
ishvim@stanford.edu

Abstract

Test-time training (TTT) is a method to discover new state of the art for scientific problems. Unlike traditional inference that uses frozen language models, TTT allows the LLM to continue learning through reinforcement learning during test time. We applied this method to the Practical Molecular Optimization (PMO) benchmark by framing the molecular optimization as a conditional language modeling problem. We represent molecules as strings using SMILES and adapt ChemLactica-125m during test-time with LoRA. Our methods use entropic reward adaptation and variants of KL-controlled updates to improve diversity in generation and avoid mode collapse. We find that TTT can improve oracle efficiency on some PMO tasks, but is sensitive to over-exploitation, requiring updates that emphasize diversity of generation to improve search coverage.

1 Mentor: Nevin George

2 Introduction

In this paper, we investigate test-time training (TTT) for practical molecular optimization. Starting from a pretrained chemistry language model, ChemLactica, we perform lightweight online updates during search using low-rank adaptation (LoRA) and oracle feedback[1][2].

Oracles are black-box functions mapping molecules represented as Simplified Molecular Input Line Entry System (SMILES) to a scalar score $\in [0, 1]$ [3] [4], where maximizing that score provides the objective for our model’s optimization. Generally, for molecular optimization, the goal is, given a scoring oracle such as drug-likeness or target affinity, find molecules that maximize the score under a limited oracle budget.

We formalize this problem as molecular optimization under a fixed oracle budget: given a task-specific scoring oracle and a budget on the number of evaluations, the goal is to maximize our discovered score. The oracle is only observed at query time, so any update to the policy must occur during search, using the same limited feedback within the confines of TTT. The goal of this project is to study whether and under what task conditions these adaptive inference techniques improve budget-constrained molecular discovery. Our main contributions is an empirical comparison of novel approaches to PMO: an entropic test-time training method against our diversity-aware extension that modifies both candidate selection and adaptation signals (diversified prompts, screening, and novelty-shaped rewards)[5][6].

PMO defines 23 single-objective tasks under a fixed oracle budget (e.g. 10K queries) for sample-efficient discovery. Each task falls into one of seven categories: Bio (3): *JNK3*, *GSK3B*, *DRD2* (target affinity); Rediscovery of known molecules(3): *Celecoxib*, *Thiothixene*, *Troglitazone*; Similarity matching to target (2): *Albuterol*, *Mestranol*; Isomer generation under constraints (2): *Isomers_C7H8N2O2*, *Isomers_C9H10N2O2PF2Cl*; Median matching to a target distribution (2): *Median_1*, *Median_2*;

Multi-parameter Objectives (7): *Osimertinib*, *Fexofenadine*, *Ranolazine*, *Perindopril*, *Amlodipine*, *Sitagliptin*, *Zaleplon*; Hop techniques to identify new compounds (2): *Deco_Hop*, *Scaffold_Hop*. We use the same oracles and budget and report per-category and aggregate results [4].

3 Related Work

Previous papers on molecular optimization focuses on sample efficiency under a limited number of oracle calls. Works such as a genetic-guided GFlowNet (Genetic GFN) distilled a genetic algorithm via genetic search and reweighted replay training, iteratively evolving populations through crossover, mutation, and selection [7]; a genetic expert-guided learning (GEGL) trains a molecule-generating deep neural network using an additional genetic expert policy, which generates targets for imitation learning [8]; GenMol uses discrete diffusion with non-autoregressive models over fragments and fragment remasking for optimization [9]. These methods all improve exploration via explicit local search (mutations, crossover, fragment edits) and better coverage of chemical space.

Another line of research has explored test-time training as a paradigm shift from static, frozen language models. Early foundational work established that language representations need not remain static [10] [11], which has been extended to scientific discovery tasks through works such as TTT-Discover, which trains language model weights during inference on specific test problems [5]. Unlike evolutionary search methods such as AlphaEvolve which store past attempts in buffers and generate new prompts via heuristics while querying frozen LLMs, TTT-Discover argues that models better internalize learnings by adapting to problem-specific data distributions at test-time [12].

TTT-Discover introduces a key innovation that we use in our work: an entropic objective that exponentially upweights high-reward actions, focusing computational resources on the most promising solutions. This approach has achieved state-of-the-art results across diverse scientific domains, including Erdős’ Minimum Overlap Problem and generating GPU kernels that outperform human work. This method is effective for problems with continuous, verifiable objective functions, as its entropic objective requires gradient signals for probability distribution reshaping, a problem category that PMO belongs to. So we take inspiration from this approach that features aggressive search methods to quickly achieve one exceptional solution to best solve a singular problem, rather than generalize to other problems [5].

4 Approach

We treat PMO as budgeted black-box molecular optimization problem: a task oracle scores candidate molecules and the optimizer must then find high-scoring candidates under a fixed query budget [4]. We use a pretrained causal molecular language model, ChemLactica, over SMILES, conditioned on a prompt of previously discovered candidates, and then adapt the model online at test time with LoRA (reset per task), for our TTT approach. [1].

Let $f_t(x)$ denote the oracle for task t , where x is a molecule represented as a SMILES string. The goal is to find high-scoring molecules under a limited oracle budget B :

$$\max_{x \in \mathcal{X}} f_t(x) \quad \text{subject to at most } B \text{ oracle evaluations.}$$

Our generator is a pretrained causal LM $p_\theta(x \mid c)$, Chemlactica-125m, where c is the prompt constructed from previously discovered molecules. The core question is whether adapting θ during search improves budget efficiency.

the entropic loss is:

1. Generate completions: $x_1, \dots, x_n$
2. Score them with oracle rewards: $r_1, \dots, r_n$
3. Compute weights:

$$w_i = \frac{\exp(\beta r_i)}{\sum_j \exp(\beta r_j)}$$

4. Compute reward-weighted negative log-likelihood:

$$\mathcal{L}_{\text{TTT}}(\theta) = \sum_i i = 1^n w_i, \ell_{\text{NLL}}(x_i; \theta)$$

4.1 Prescreening

We apply an RDKit-based chemistry prescreen to candidate SMILES before applying the reward: only molecules that parse successfully and satisfy Lipinski’s Rule of Five (molecular weight ≤ 500 Da, $\log P \leq 5$, ≤ 5 H-bond donors, ≤ 10 H-bond acceptors) are retained. We also exclude pan-assay interference (PAINS) substructures, known false positives, via RDKit’s FilterCatalog, so that the optimizer receives reward only from drug-like, non-promiscuous compounds.

4.2 TTT-Diverse

In our approach, we score a batch with the oracle, assign rewards, and minimize a weighted negative log-likelihood such that higher-reward candidates would get larger softmax weights [5]. A temperature β is adapted online via a KL budget to the pretrained model.

TTT-Diverse adds three mechanisms to this approach:

1. Diversified prompt memory: instead of just choosing the top 3 molecules to be in the prompt for the next step, we use a mix of top molecules, diversity anchors (molecules with largest Tanimoto distance to the current prompt), and recent improvements.
2. Screening before oracle calls based on (structural novelty, prompt distance, scaffold novelty, duplicate penalties)
3. novelty-shaped adaptation reward (oracle score plus novelty bonus and duplicate penalty)

TTT-Diverse aims to reduce local collapse and improve oracle efficiency without changing the benchmark’s scoring.

$$s(x) = \alpha \text{Novelty}(x) + \beta_s \text{PromptDistance}(x) + \gamma \text{ScaffoldNovelty}(x) + \delta \text{RecentAffinity}(x) - \eta \text{Duplicate}(x)$$

Here $\text{Novelty}(x)$ is the minimum Tanimoto distance (1 - Tanimoto similarity, the dissimilarity between two sets or binary vectors) between the candidate’s Morgan fingerprint (a fixed-length bit vector that summarizes local structure around each atom) and the fingerprints of all molecules currently in the replay/memory (how far x is from the closest already-seen molecule, so higher values mean more novel); $\text{PromptDistance}(x)$ is the mean distance to the current prompt molecules; $\text{ScaffoldNovelty}(x)$ indicates whether the Murcko scaffold (rings and linkers of the molecule) is new relative to memory; $\text{RecentAffinity}(x)$ rewards similarity to a recent improvement; $\text{Duplicate}(x)$ is 1 if the canonical SMILES of x was already in the oracle cache and 0 otherwise [13][7][8]. Only the top-ranked candidates after screening receive an oracle evaluation; the rest are not scored by the task oracle, saving budget.

Once oracle scores are observed, TTT-Diverse uses a shaped reward:

$$\tilde{r}(x) = f_t(x) + \lambda_{\text{nov}} \text{Novelty}(x) - \lambda_{\text{dup}} \text{Duplicate}(x),$$

where $f_t(x)$ represents the raw oracle score for molecule x [0,1]). This shaped reward is used for adaptation, but the raw oracle score is still used for benchmark reporting.

TTT traditionally adapts β from KL alone, but our approach also adapts β using duplicate rate (fraction of oracle calls that were cache hits), unique fraction (fraction of valid proposals that were unique SMILES), an entropy proxy over the rollout, and the screen rejection fraction (fraction of valid candidates not selected for oracle). This reduces exploitation when KL grows too large, duplicate rate increases, or screened pool becomes overly redundant. But when KL is low and generations are adequately diverse, β is decreased to allow stronger adaptation. This keeps the policy from mode collapsing and keeps search sufficiently exploratory.

5 Experiments

The main metric is normalized AUC@Top10 (average of top-10 mean oracle score at fixed checkpoints), with best score as a secondary metric; we also log validity rate, duplicate rate, unique

fraction, KL to the reference, β trajectories, and screening efficiency. Default setup: base model yerevann/chemlactica-125m, LoRA rank 64 and $\alpha = 128$, AdamW learning rate of 3×10^{-5} , 30 optimization steps, 16 evaluated candidates per step, an oracle budget of 10,000, and a checkpoint every 100; for the diversity-aware method we use screening enabled, pool factor 1.5, novelty-shaped reward, no replay buffer or mutation proposals, and report prescreening separately when enabled.

5.1 Data

We evaluate on the PMO benchmark, with 23 tasks covering biological activity, rediscovery, similarity, isomer generation, multi-parameter optimization, and scaffold/decorations hopping [4]. Each task provides a black-box oracle from SMILES to a scalar score. Inputs are prompt-augmented contexts of previously discovered molecules; outputs are generated SMILES, canonicalized and validated before oracle evaluation.

5.2 Evaluation method

The model receives a prompt of ChemLactica-style context of previously discovered molecules and proceeds to generate SMILES completions, which are then canonicalized, validated, and filtered to the task oracle. The oracle returns a task-specific scalar score in $[0,1]$. All reported metrics use these oracle scores and the canonical SMILES set. We proceed to use the following evaluation methods:

Normalized AUC@Top10: Let $(C_1, \dots, C_m)$ be the oracle-call checkpoints. Let $T_k(C_j)$ be the average score of the top ($k=10$) discovered molecules by checkpoint C_j . Then, we define:

$$\text{AUC@Top10} = \frac{1}{m} \sum_{j=1}^m T_{10}(C_j)$$

where m represents the number of checkpoints recorded. Thus, we reward methods that discover strong molecules early, which is highly appropriate and important for scientific discovery under a limited assay budget.

Best Score:

$$\max_{x \in \mathcal{D}} f_t(x)$$

where $\mathcal{D}$ is the set of oracle-evaluated molecules. This measures whether the run found at least one exceptional molecule, which may be more important to scientists whose focus is on producing the single best molecule or discovering a usable molecule as quickly as possible.

5.2.1 Tertiary Metrics: Validity, Dedup hits

- Validity rate is the number of valid molecules generated as a fraction of all the molecules generated.
- Dedup hits are the number of duplicate molecules generated across rollouts.

5.3 Experimental details

We ran a series of TTT experiments and compared each configuration to a best-of-n baseline:

1. TTT with Chemlactica-125m with Top 3 Prompts with seeds 0,1
2. TTT with Chemlactica-125m with diversity prompt with seeds 0,1
3. TTT-Diverse with Chemlactica-125m with Prescreen with seeds 0,1,2
4. TTT-Diverse with Chemlactica-125m with seeds 0,1,2
5. TTT-Diverse with Chemlactica-1.3b with seed 0

We then report on the full PMO 23-task set, with the exception of our experiment with Chemlactica-1.3b. Due to budget constraints, we ran a more focused cross comparison ablation on 4 specific tasks to investigate how model scaling could potentially improve TTT-Diverse.

5.4 Results

Table 1: Summary of TTT wins across prompt configurations and metrics.

Prompt Type	Metric	Wins	Tasks Won by TTT against best-of-n
Top 3	AUC@Top10	7/23	jnk3, sitagliptin_mpo, thiothixene_rediscovery, albuterol_similarity, osimertinib_mpo, zaleplon_mpo, troglitazone_rediscovery
	Best Score	9/23	jnk3, gsk3b, thiothixene_rediscovery, albuterol_similarity, isomers_c7h8n2o2, osimertinib_mpo, median1, median2, troglitazone_rediscovery
Diversity	AUC@Top10	10/23	jnk3, thiothixene_rediscovery, albuterol_similarity, isomers_c7h8n2o2, deco_hop, qed, mestranol_similarity, median1, zaleplon_mpo, troglitazone_rediscovery
	Best Score	13/23	jnk3, drd2, thiothixene_rediscovery, albuterol_similarity, osimertinib_mpo, qed, isomers_c9h10n2o2pf2cl, median1, fexofenadine_mpo, ranolazine_mpo, amlodipine_mpo, zaleplon_mpo, troglitazone_rediscovery
TTT-Diverse	AUC@Top10	9/23	gsk3b, drd2, qed, troglitazone_rediscovery, mestranol_similarity, isomers_c7h8n2o2, median1, ranolazine_mpo, amlodipine_mpo
	Best Score	7/23	gsk3b, drd2, qed, troglitazone_rediscovery, mestranol_similarity, median1, ranolazine_mpo
TTT-Diverse + Prescreen	AUC@Top10	9/23	gsk3b, drd2, qed, troglitazone_rediscovery, albuterol_similarity, isomers_c7h8n2o2, median1, osimertinib_mpo, amlodipine_mpo
	Best Score	12/23	gsk3b, drd2, qed, albuterol_similarity, isomers_c7h8n2o2, isomers_c9h10n2o2pf2cl, median1, osimertinib_mpo, fexofenadine_mpo, ranolazine_mpo, amlodipine_mpo, zaleplon_mpo

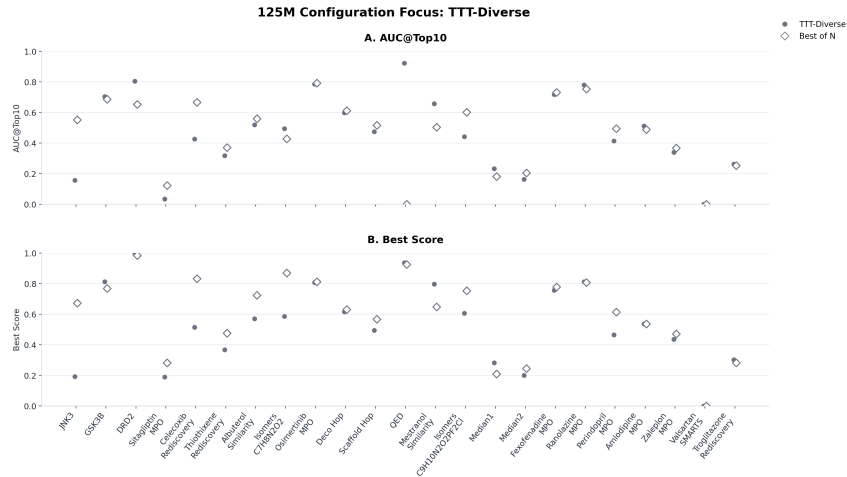


Figure 1: Task-level comparison of 125M TTT-Diverse against best-of-n across all 23 PMO tasks.

The top panel in Figure 1 reports normalized AUC@Top10 and the bottom panel reports best score for each task. Filled gray circles denote TTT-Diverse and hollow gray diamonds denote best-of-n. AUC@Top10 reflects budget efficiency, or how quickly a method accumulates strong candidates under the limited oracle budget, while best score captures the single strongest molecule found during the run. This figure shows that the 125M TTT-Diverse configuration is highly task-dependent: it outperforms best-of-n on several tasks such as drd2, gsk3b, and mestranol_similarity, but underperforms substantially on tasks such as jnk3, celecoxib_rediscovery, and several MPO objectives.

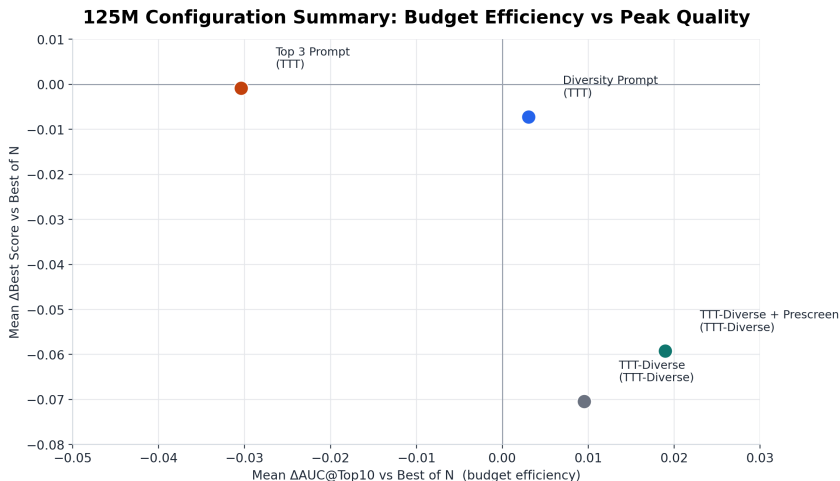


Figure 2: Configuration-level tradeoff between budget efficiency and peak discovery quality for 125M runs.

Each point in Figure 2 summarizes one configuration by its mean performance difference relative to best-of-n across the 23 PMO tasks. The horizontal axis is the mean $\Delta\text{AUC@Top10}$, which measures budget efficiency, and the vertical axis is the mean $\Delta\text{Best Score}$, which measures peak-quality improvement. Points to the right indicate configurations that outperform best-of-n on average in budget efficiency, while points above zero indicate stronger best-score performance. The plot shows that TTT-Diverse + Prescreen achieves the strongest average budget efficiency gains, whereas Top 3 Prompt TTT is the most competitive configuration in best score. This highlights a central tradeoff in our experiments: diversity-aware adaptive search improves early, budgeted discovery more reliably than it improves the final best molecule found.

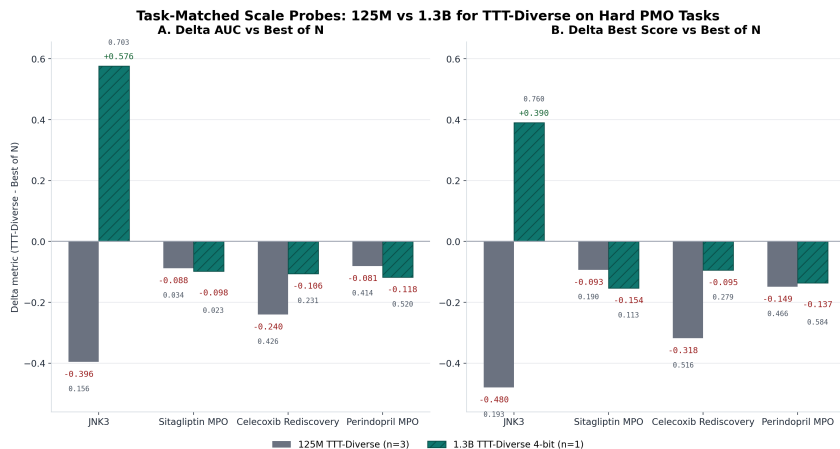


Figure 3: Task-matched scale probes comparing TTT-Diverse at 125M and 1.3B on four different hard PMO tasks.

Given constraints regarding compute and budget, we ran a useful ablation study testing a bigger model, ChemLactica-1.3b, with our TTT-Diverse approach, analyzing the four tasks that TTT-Diverse with ChemLactica-125m performed especially poorly on compared to best-of-n (see Figure 3). We find that the bigger model improves performance, to even beat best-of-n, in the jnk3 task, but only shows marginal improvements for the other 3 tasks which still underperform compared to best-of-n. This suggests that bigger models with TTT can lead to performance gains on specific tasks, but scaling alone is insufficient to solve these molecular optimization problems.

6 Analysis

Comparing AUC@Top10 scores across our different experiment configurations shows that performance depends heavily on task structure. We organize our analysis by task type to understand when test-time training outperforms baseline sampling:

Binding affinity tasks The Top 3 Prompt strategy shows high variance: it excels on jnk3 but underperforms on drd2. TTT-Diverse with Prescreen does well on drd2 and gsk3b. This suggests that more diverse search helps avoid local optima in protein-ligand optimization, while jnk3 has a more concentrated optimum that rewards aggressive exploitation.

Drug rediscovery tasks All TTT methods underperform relative to best-of-n on celecoxib_rediscovery, while performance on thiothixene_rediscovery and troglitazone_rediscovery remains comparable across both approaches. This suggests that when the target molecule has well-established structural features, sampling beats adaptive refinement, which may drift from the known solution.

Similarity tasks. Diversity Prompt and TTT-Diverse without prescreen notably outperform on mestranol, while TTT-Diverse with prescreen underperforms. This implies that aggressive filtering removes structurally similar candidates that test-time training could otherwise refine toward the target.

Isomer enumeration tasks The diversity prompt methods show clear advantages on both isomer tasks (isomers_c7h8n2o2, isomers_c9h10n2o2pf2c1). while Top 3 Prompt fails badly on both, showing that exploitation cannot handle the complex space of constitutional isomers.

Multi-parameter optimization tasks MPO task results vary by constraint complexity. On simpler tasks like osimertinib_mpo and ranolazine_mpo, all methods perform relatively similarly (osimertinib_mpo ranolazine_mpo, with TTT-Diverse without prescreen achieving the best ranolazine_mpo score. However, on the highly constrained sitagliptin_mpo, all methods dramatically underperform. This seems to suggest that balancing multiple objectives with strict constraints may require different reward shaping.

Scaffold modifications and median tasks scaffold_hop and deco_hop tasks uniformly favor best-of-n across all methods. These structural transformation tasks seem to need broader sampling diversity than test-time training provides. Median tasks (median1, median2) show minimal differentiation, suggesting these intermediate-difficulty objectives don’t benefit much from adaptive refinement.

Unconstrained optimization. The most striking result appears on the unconstrained qed task, where both TTT-Diverse variants achieve high scores (0.9142 and 0.9230) compared to baseline best-of-n’s 0.0000.

Overall, TTT-Diverse methods perform well on continuous optimization with moderate structural constraints (binding affinity, QED, certain isomer tasks), while TTT with Top 3 Prompting succeeds only on concentrated optima. Simple best-of-n sampling is competitive or superior on rediscovery, scaffold modification, and highly constrained MPO tasks.

7 Conclusion

While TTT-Diverse adds computational overhead compared to frozen sampling, this cost is negligible relative to real-world oracle evaluation time. In practical drug discovery, a single *in vitro* assay may take days and cost hundreds of dollars, while our LoRA updates complete in minutes on a single GPU. Thus, the oracle savings achieved by TTT-Diverse on tasks like drd2 (288 calls vs. 447 for baseline) or qed (246 calls vs. 114 baseline, but vastly higher scores) could justify the additional computation.

7.1 Future Works

It is important to note that the best score may be the best metric to achieve the most ideal molecule for this optimization task; however, third-party experts in previous studies also noted that the model may be overfitting in biology since a solution with a high score may not necessarily generate biologically valid data [5]. Future work should include some focused wet-lab verification of such discovery results.

Another promising area to expand on is the scaling up the backbone model for TTT. With a 125M model, some adaptation capacity could likely be consumed by recovering low-level issues such as invalid generations, generally poor local exploration, or weak structural priors of certain categories of molecules. But, with a larger compute budget stretched across a longer time-frame, having a base model that does already have stronger chemical priors could allow TTT to hypothetically spend fewer updates configuring plausible molecules and more on solving for the specific optimization problem at hand. We briefly explored this in Figure 3, where we observed in different types of settings that TTT-Diverse could be able to see drastic improvements in molecular matching compared to best-of-n.

Team contributions (Required for multi-person team)

Both the paper and code were split equally among Ishvi and Aaron, involving a great deal of co-working and pair programming. Aaron discovered the initial inspiration paper, TTT-Discover, the problem scope of Practical Molecular Optimization, and ran many of the experiments on TTT-Diverse [5] [4]. Ishvi ran many of the baseline experiments with best-of-n and TTT-Discover, ablation studies, and performed much of the analysis aggregation which were incredibly important to shift and align our approach more closely with the problem at hand (because raw scientific discovery and molecular optimization did not necessarily respond equally well to the same rewards).

References

- [1] Philipp Guevorguian, Menua Bedrosian, Tigran Fahradyan, Gayane Chilingaryan, Hrant Khachatrian, and Armen Aghajanyan. Small molecule optimization with large language models, 2024.
- [2] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.
- [3] David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36, 1988.
- [4] Wenhao Gao, Tianfan Fu, Jimeng Sun, and Connor W. Coley. Sample efficiency matters: A benchmark for practical molecular optimization, 2022.
- [5] Mert Yuksekgonul, Daniel Kocejka, Xinhao Li, Federico Bianchi, Jed McCaleb, Xiaolong Wang, Jan Kautz, Yejin Choi, James Zou, Carlos Guestrin, and Yu Sun. Learning to discover at test time, 2026.
- [6] Yu Sun, Xinhao Li, Karan Dalal, Jiarui Xu, Arjun Vikram, Genghan Zhang, Yann Dubois, Xinlei Chen, Xiaolong Wang, Sanmi Koyejo, Tatsunori Hashimoto, and Carlos Guestrin. Learning to (learn at test time): Rnns with expressive hidden states, 2025.
- [7] Hyeonah Kim, Minsu Kim, Sanghyeok Choi, and Jinkyoo Park. Genetic-guided gflownets for sample efficient molecular optimization. In *Advances in Neural Information Processing Systems*, 2024.
- [8] Sungsoo Ahn, Junsu Kim, Hankook Lee, and Jinwoo Shin. Guiding deep molecular optimization with genetic exploration. In *Advances in Neural Information Processing Systems*, 2020.
- [9] Seul Lee, Karsten Kreis, Sriram Prasad Veccham, Meng Liu, Danny Reidenbach, Yuxing Peng, Saeed Gopal Paliwal, Weili Nie, and Arash Vahdat. Genmol: A drug discovery generalist with discrete diffusion. In *International Conference on Machine Learning*, 2025.

- [10] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013.
- [11] Ben Krause, Emmanuel Kahembwe, Iain Murray, and Steve Renals. Dynamic evaluation of neural sequence models, 2017.
- [12] Alexander Novikov, Ng  n V  , Marvin Eisenberger, Emilien Dupont, Po-Sen Huang, Adam Zsolt Wagner, Sergey Shirobokov, Borislav Kozlovskii, Francisco J. R. Ruiz, Abbas Mehrabian, M. Pawan Kumar, Abigail See, Swarat Chaudhuri, George Holland, Alex Davies, Sebastian Nowozin, Pushmeet Kohli, and Matej Balog. Alphaevolve: A coding agent for scientific and algorithmic discovery, 2025.
- [13] Marcus Olivecrona, Thomas Blaschke, Ola Engkvist, and Hongming Chen. Molecular de novo design through deep reinforcement learning, 2017.

A Appendix

Table 2: AUC@Top10 comparison

Task	Top 3 Prompt			Diversity Prompt			TTT-Diverse + Prescreen			TTT-Diverse		
	TTT	best-of-n	Δ	TTT	best-of-n	Δ	TTT	best-of-n	Δ	TTT	best-of-n	Δ
jnk3	0.8624	0.6219	+0.2405	0.7642	0.7450	+0.0192	0.2012	0.5817	-0.3804	0.1564	0.5520	-0.3956
gsk3b	0.6289	0.6774	-0.0485	0.6391	0.7459	-0.1068	0.7681	0.6701	+0.0980	0.7042	0.6866	+0.0176
drd2	0.9011	0.9854	-0.0843	0.5459	0.8107	-0.2648	0.9251	0.6540	+0.2711	0.8055	0.6531	+0.1524
sitagliptin_mpo	0.1886	0.1148	+0.0738	0.0780	0.1120	-0.0340	0.0252	0.1233	-0.0981	0.0338	0.1218	-0.0880
celecoxib_rediscovery	0.3826	0.4634	-0.0808	0.4004	0.5939	-0.1936	0.4014	0.6667	-0.2654	0.4263	0.6667	-0.2404
thiothixene_rediscovery	0.3841	0.3180	+0.0661	0.3509	0.3451	+0.0058	0.3276	0.3719	-0.0443	0.3177	0.3713	-0.0536
albuterol_similarity	0.5594	0.5355	+0.0239	0.5496	0.5113	+0.0383	0.6482	0.5606	+0.0876	0.5202	0.5594	-0.0392
isomers_c7h8n2o2	0.2539	0.6862	-0.4323	0.4559	0.3878	+0.0682	0.6457	0.4275	+0.2182	0.4947	0.4278	+0.0669
osimertinib_mpo	0.4078	0.3394	+0.0684	0.7951	0.8050	-0.0099	0.7874	0.7933	-0.0059	0.7860	0.7929	-0.0069
deco_hop	0.6182	0.6279	-0.0097	0.6083	0.6074	+0.0009	0.5917	0.6126	-0.0209	0.5981	0.6126	-0.0145
scaffold_hop	0.4942	0.5129	-0.0187	0.4840	0.5028	-0.0188	0.4719	0.5167	-0.0448	0.4743	0.5167	-0.0424
qed	0.0000	0.0000	+0.0000	0.4636	0.0000	+0.4636	0.9142	0.0000	+0.9142	0.9230	0.0000	+0.9230
mestranol_similarity	0.5766	0.6734	-0.0968	0.7444	0.5651	+0.1793	0.4348	0.5049	-0.0701	0.6579	0.5036	+0.1543
isomers_c9h10n2o2pf2cl	0.2896	0.5532	-0.2636	0.5845	0.6223	-0.0378	0.5252	0.6013	-0.0761	0.4413	0.6013	-0.1599
median1	0.2202	0.2253	-0.0051	0.2067	0.1996	+0.0071	0.1963	0.1807	+0.0156	0.2319	0.1807	+0.0512
median2	0.1968	0.2006	-0.0038	0.1754	0.1963	-0.0209	0.1628	0.2054	-0.0426	0.1625	0.2054	-0.0429
fexofenadine_mpo	0.7159	0.7326	-0.0167	0.7151	0.7191	-0.0039	0.6960	0.7314	-0.0354	0.7183	0.7314	-0.0131
ranolazine_mpo	0.6821	0.8049	-0.1228	0.7280	0.7395	-0.0115	0.7316	0.7536	-0.0220	0.7805	0.7536	+0.0269
perindopril_mpo	0.4465	0.4621	-0.0156	0.4258	0.4841	-0.0583	0.4365	0.4953	-0.0587	0.4139	0.4949	-0.0811
amlodipine_mpo	0.5174	0.5220	-0.0046	0.4916	0.4983	-0.0068	0.5188	0.4888	+0.0300	0.5118	0.4891	+0.0227
zaleplon_mpo	0.3865	0.3737	+0.0128	0.4475	0.4145	+0.0330	0.3306	0.3680	-0.0374	0.3402	0.3680	-0.0278
valsartan_smarts	0.0000	0.0000	+0.0000	0.0000	0.0000	+0.0000	0.0000	0.0000	+0.0000	0.0000	0.0000	+0.0000
trogliatzone_rediscovery	0.2836	0.2651	+0.0185	0.2585	0.2368	+0.0218	0.2554	0.2523	+0.0030	0.2621	0.2524	+0.0098

Table 3: Best Score comparison

Task	Top 3 prompts			Diversity Prompt			TTT-Diverse + Prescreen			TTT-Diverse		
	TTT	best-of-n	Δ	TTT	best-of-n	Δ	TTT	best-of-n	Δ	TTT	best-of-n	Δ
jnk3	0.9250	0.7550	+0.1700	0.8950	0.8900	+0.0050	0.4500	0.7433	-0.2933	0.1933	0.6733	-0.4800
gsk3b	0.8650	0.7800	+0.0850	0.7350	0.8250	-0.0900	0.9300	0.7300	+0.2000	0.8133	0.7700	+0.0433
drd2	0.9774	0.9887	-0.0113	0.9964	0.9811	+0.0152	0.9953	0.9847	+0.0106	0.9976	0.9847	+0.0128
sitagliptin_mpo	0.2953	0.3306	-0.0353	0.1587	0.2890	-0.1303	0.0866	0.2830	-0.1965	0.1896	0.2830	-0.0934
celecoxib_rediscovery	0.4584	0.8214	-0.3630	0.6000	0.7500	-0.1500	0.4846	0.8333	-0.3487	0.5157	0.8333	-0.3177
thiothixene_rediscovery	0.5261	0.3678	+0.1583	0.4422	0.4276	+0.0146	0.3968	0.4765	-0.0797	0.3677	0.4765	-0.1087
albuterol_similarity	0.8175	0.6364	+0.1811	0.7151	0.6162	+0.0988	0.7056	0.7246	-0.0189	0.5711	0.7246	-0.1535
isomers_c7h8n2o2	0.7798	0.7201	+0.0597	0.8228	0.8970	-0.0742	0.8289	0.8709	-0.0420	0.5860	0.8709	-0.2849
osimertinib_mpo	0.7818	0.7532	+0.0286	0.8264	0.8231	+0.0033	0.8042	0.8128	-0.0087	0.8078	0.8128	-0.0051
deco_hop	0.6349	0.6485	-0.0136	0.6209	0.6243	-0.0035	0.6099	0.6322	-0.0223	0.6155	0.6322	-0.0167
scaffold_hop	0.5260	0.5508	-0.0248	0.5275	0.5526	-0.0251	0.4906	0.5678	-0.0773	0.4955	0.5678	-0.0723
qed	0.9210	0.9216	-0.0006	0.9357	0.9270	+0.0087	0.9245	0.9256	-0.0012	0.9374	0.9256	+0.0117
mestranol_similarity	0.7293	0.7401	-0.0108	0.7602	0.7663	-0.0061	0.5349	0.6493	-0.1144	0.7975	0.6493	+0.1482
isomers_c9h10n2o2pf2cl	0.6780	0.7739	-0.0959	0.8226	0.6888	+0.1338	0.6734	0.7533	-0.0799	0.6065	0.7533	-0.1468
median1	0.2831	0.2452	+0.0379	0.2680	0.2291	+0.0389	0.2318	0.2091	+0.0227	0.2818	0.2091	+0.0727
median2	0.2332	0.2256	+0.0076	0.1922	0.2491	-0.0569	0.1897	0.2452	-0.0555	0.2005	0.2452	-0.0446
fexofenadine_mpo	0.7567	0.7993	-0.0426	0.7807	0.7678	+0.0129	0.7251	0.7794	-0.0543	0.7567	0.7794	-0.0227
ranolazine_mpo	0.7671	0.8605	-0.0934	0.8093	0.8030	+0.0063	0.7845	0.8087	-0.0242	0.8133	0.8087	+0.0046
perindopril_mpo	0.4969	0.5239	-0.0270	0.5043	0.5727	-0.0685	0.4778	0.6147	-0.1369	0.4656	0.6147	-0.1491
amlodipine_mpo	0.5423	0.5936	-0.0513	0.5428	0.5418	+0.0010	0.5481	0.5372	+0.0109	0.5363	0.5372	-0.0009
zaleplon_mpo	0.4264	0.4432	-0.0168	0.5370	0.4883	+0.0487	0.4051	0.4724	-0.0673	0.4353	0.4724	-0.0370
valsartan_smarts	0.0000	0.0000	+0.0000	0.0000	0.0000	+0.0000	0.0000	0.0000	+0.0000	0.0000	0.0000	+0.0000
trogliatzone_rediscovery	0.3389	0.3002	+0.0387	0.3141	0.2630	+0.0511	0.2964	0.2823	+0.0141	0.3028	0.2823	+0.0205

Table 4: Validity Rate comparison

Task	Top 3 Prompt		Diversity Prompt		TTT-Diverse + Prescreen		TTT-Diverse	
	TTT	best-of-n	TTT	best-of-n	TTT	best-of-n	TTT	best-of-n
jnk3	0.9833	0.9698	0.9792	0.8094	0.9153	0.7229	0.8630	0.7361
gsk3b	0.9979	0.9823	0.9771	0.9667	0.8870	0.9458	0.9250	0.9694
drd2	0.9646	0.9802	0.9729	0.9865	0.9852	0.9708	0.9634	0.9708
sitagliptin_mpo	0.9906	0.9427	0.9521	0.8854	0.9106	0.9076	0.9588	0.9076
celecoxib_rediscovery	1.0000	0.9979	0.9990	0.9990	0.9782	0.9986	0.9907	0.9986
thiothixene_rediscovery	0.9844	0.9927	0.9771	0.9948	0.9884	0.9924	0.9782	0.9924
albuterol_similarity	1.0000	1.0000	0.9990	0.9990	0.9958	0.9986	0.9991	0.9986
isomers_c7h8n2o2	0.9989	0.9989	0.9990	1.0000	0.9727	0.9972	0.9861	0.9979
osimertinib_mpo	0.9938	0.9989	0.9958	1.0000	0.9755	0.9965	0.9880	0.9965
deco_hop	1.0000	0.9989	0.9510	1.0000	0.9949	0.9979	0.9903	0.9979
scaffold_hop	0.9917	0.9979	0.9906	1.0000	0.9806	0.9993	0.9259	0.9993
qed	1.0000	0.9989	0.9990	0.9990	0.9968	0.9986	0.9972	0.9986
mestranol_similarity	1.0000	1.0000	1.0000	0.9979	0.9986	0.9979	0.9963	0.9979
isomers_c9h10n2o2pf2cl	0.9989	0.9969	0.9833	0.9990	0.9903	0.9521	0.9861	0.9521
median1	1.0000	1.0000	0.9990	0.9969	0.9852	0.9910	0.9907	0.9910
median2	0.9688	0.9625	0.9667	0.9427	0.9657	0.9493	0.9273	0.9493
fexofenadine_mpo	1.0000	1.0000	1.0000	0.9990	0.9917	0.9972	0.9741	0.9972
ranolazine_mpo	1.0000	0.9823	0.9927	0.9812	0.9991	0.9840	0.9833	0.9840
perindopril_mpo	0.9958	0.9969	0.9917	0.9969	0.9944	0.9972	0.9815	0.9972
amlodipine_mpo	0.9979	0.9979	1.0000	1.0000	0.9912	0.9993	0.9949	0.9993
zaleplon_mpo	0.9834	0.9896	0.9927	0.9906	0.9921	0.9910	0.9787	0.9910
valsartan_smarts	0.9854	0.8781	0.9938	0.9365	0.8843	0.8125	0.8912	0.9535
troglitazone_rediscovery	0.9938	0.9761	0.9854	0.9844	0.9884	0.9861	0.9852	0.9861

Table 5: Oracle Calls comparison

Task	Top 3 prompt		Diversity prompt		TTT-Diverse + Prescreen		TTT-Diverse	
	TTT	best-of-n	TTT	best-of-n	TTT	best-of-n	TTT	best-of-n
jnk3	314	463	388.5	381.5	476.3	304.3	469.7	348.7
gsk3b	279.5	433	455.5	436.5	416	438	418	447
drd2	175.5	171.5	224.5	179.5	324.7	211.7	288.7	211.7
sitagliptin_mpo	449.5	447.5	448	407.5	434	423.7	477	423.7
celecoxib_rediscovery	474	357	479.5	452	480	374.7	480	374.7
thiothixene_rediscovery	468.5	476	469	477.5	480	476	480	476
albuterol_similarity	373.5	409	438	393	397	353	458.3	353
isomers_c7h8n2o2	86	202	113	160	405.3	157	406.7	161.3
osimertinib_mpo	156.5	197	375	259.5	346.7	265.7	375	266
deco_hop	466	456.5	434.5	411	455	399	458.7	399
scaffold_hop	475	477	468	471	469.7	470	476.7	470
qed	40	38.5	95.5	52.5	217	47	246	47
mestranol_similarity	347.5	300	335.5	432	479.7	442.3	451.7	442.7
isomers_c9h10n2o2pf2cl	160	172.5	257.5	238	402.3	226.3	339.7	232.3
median1	463.5	467	474	477	475.3	474.7	480	474.7
median2	464	462	462	452.5	480	455.7	479.3	455.7
fexofenadine_mpo	330.5	397	369	397.5	435.3	393.7	446	393.7
ranolazine_mpo	452	387	469.5	382.5	454.7	406	451	406
perindopril_mpo	476	477	473	474.5	480	475.3	455.3	475.3
amlodipine_mpo	472.5	472.5	472	473.5	468.3	474.3	480	474.3
zaleplon_mpo	467.5	471.5	445.5	472	395	470.7	339	470.7
valsartan_smarts	226.5	402	259.5	395.5	418	363.3	472	376.7
troglitazone_rediscovery	476.5	468	473	472	480	473	480	473

Table 6: Deduplication Hits

Task	Top 3 prompt		Diverse prompt		TTT-Diverse + Prescreen		TTT-Diverse	
	TTT	best-of-n	TTT	best-of-n	TTT	best-of-n	TTT	best-of-n
jnk3	158	2.5	49	5	0	21	0	3.3
gsk3b	199.5	38.5	9.5	24.5	21.3	13.7	24	16.3
drd2	287.5	299	105	141.5	68.3	118.3	96	118.3
sitagliptin_mpo	26	5	4	5	27.3	3.7	1.7	3.7
celecoxib_rediscovery	6	122	0	24.5	0	40	0	40
thiothixene_rediscovery	4	0.5	0	0	0	0.3	0	0.3
albuterol_similarity	106.5	71	23.5	57.5	56	67	19	67
isomers_c7h8n2o2	393.5	277.5	125.5	141	50	120.7	51.3	118.7
osimertinib_mpo	320.5	282.5	43.5	88.5	48.7	95.3	57.3	95.3
deco_hop	14	23	9.5	33	12.3	36.7	17.3	36.7
scaffold_hop	1	2	5	4	10.3	4.3	0	4.3
qed	440	441	100	130	108.7	114.3	67.3	114.3
mestranol_similarity	132.5	180	59.5	26.5	0.3	22.3	8.3	22.3
isomers_c9h10n2o2pf2cl	319.5	306	117.5	102.5	35.7	105.7	56	99.3
median1	16.5	13	5.5	1.5	0.3	1	0	1
median2	1	0	2	0	0	0	0	0
fexofenadine_mpo	149.5	83	55	42.5	31.7	43.3	17.7	43.3
ranolazine_mpo	28	84.5	5	32	16	27.7	14.3	27.7
perindopril_mpo	2	1.5	2	3	0	2.3	17.7	2.3
amlodipine_mpo	6.5	6.5	4.5	4	3	3.7	0	3.7
zaleplon_mpo	4.5	3.5	22.5	2	19.7	1.7	77.3	1.7
valsartan_smarts	246.5	19.5	94.5	47.5	45.7	24	7.3	53
trogliatzone_rediscovery	0.5	0.5	0	0.5	0	0.3	0	0.3