AVID: ACTION-VALUE GUIDANCE AND IDENTITY-AWARE LEARNING FOR CANDIDATE-SET MOLECULAR DISCOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Molecular optimization rewards individual molecules, but a finite-budget discovery campaign seeks a growing set of distinct, high-quality candidates. Repeatedly generating a high-scoring molecule can maintain high per-output reward without adding a candidate. We introduce AVID, which couples action-value guidance during generation with identity-aware learning after completion. A frozen, archive-trained prefix-action guide biases token selection toward high-reward continuations. Once a molecule is complete, an identity ledger assigns within-run repeats zero weight in the direct policy loss while preserving cached rewards and the optimizer's other learning terms. Across all 23 tasks of the Practical Molecular Optimization (PMO) benchmark and five seeds per task, AVID acquires an average of 7,114.1 distinct candidates meeting fixed archive-derived reward thresholds within 32,768 generation attempts and at most 10,000 new oracle calls. This is 45.5% more than Augmented Memory, with higher task means on 18 of 23 tasks. Component studies distinguish the two interventions: adding guidance to identity weighting improves mean and top-10 reward at 512 newly scored identities on every task, while long-horizon experiments expose the trade-off between candidate quality and acquisition. Guidance can raise the qualifying fraction yet reduce distinct yield if new identities arrive more slowly. These results show that reward quality and qualifying-set growth provide complementary views of molecular optimization under finite budgets.

1 INTRODUCTION

Goal-directed molecular generation searches for molecules with favorable task-specific properties (Segler et al., 2018; Olivecrona et al., 2017; Popova et al., 2018; Blaschke et al., 2020a). Benchmark protocols measure reward improvement under limited property evaluations (Gao et al., 2022), but a discovery campaign also needs a growing set of distinct candidates. The difference becomes apparent when a high-scoring molecule is sampled again: its cached score remains high, while its contribution to the candidate set is zero. Reusing that score saves an oracle call, not the generation attempt, and leaves open how the repeated trajectory should influence learning. We therefore ask how many distinct qualifying molecules can be acquired within limited evaluation and generation budgets, considering both sampling and updates.

Prior work improves optimization through replay and diversity-aware memory (Blaschke et al., 2020b; Guo & Schwaller, 2024a), prefix-conditioned reward guidance (Yang & Klein, 2021; Deng & Raffel, 2023; Chakraborty et al., 2024), or diversity within sampled groups (Ye et al., 2026; Li et al., 2026). AVID couples two signals available at different stages of generation. Before completion, Q-Steer uses a frozen archive-trained model to predict terminal reward from a partial SMILES sequence and candidate next token; a normalized, bounded transformation biases token sampling. After completion, canonical identity reveals whether the molecule has appeared in the run. An identity ledger then gives within-run repeats zero weight in the direct policy objective while retaining cached rewards and other learning terms. The guide thus changes sampled actions, while the ledger selects completed trajectories for direct policy learning.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

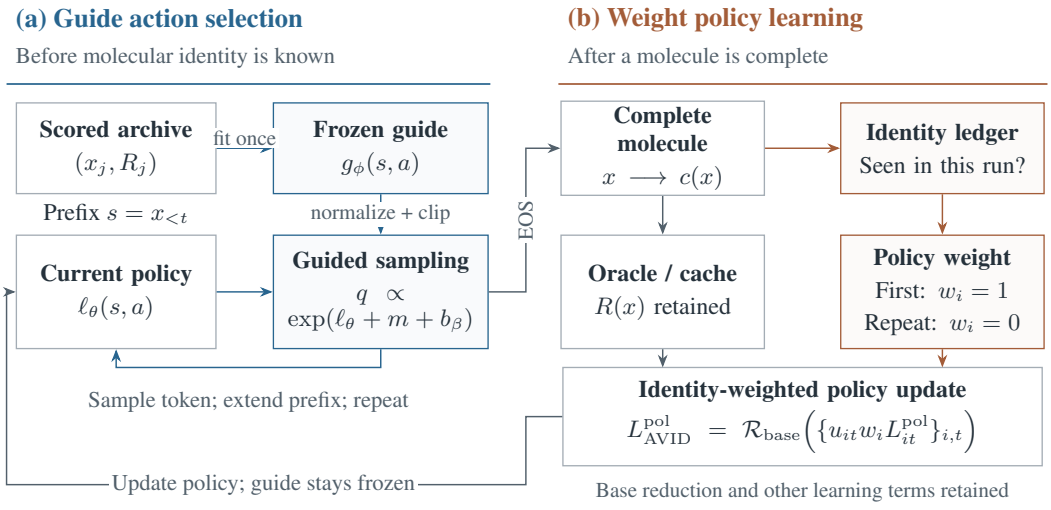

Figure 1: **AVID acts at two stages of molecular policy optimization.** Before completion, a frozen archive-trained guide supplies a normalized, bounded next-token preference. After completion, the canonical identity ledger gives within-run repeats zero weight in the direct policy term. Cached property rewards, the base reduction, and other native learning terms are retained; archive hits are not automatically treated as within-run repeats. The policy evolves while the guide remains frozen.

To evaluate this coupling, we separate new-identity acquisition N_t from the fraction f_t meeting a fixed reward criterion after t attempts, giving distinct qualifying yield $H_t = N_t f_t$. Across all 23 Practical Molecular Optimization (PMO) tasks and five paired seeds, AVID has higher task-mean yield than PPO on 22 tasks, REINVENT on 19, and Augmented Memory on 18, within 32,768 attempts and a 10,000-new-score ceiling. Adding Q-Steer to identity weighting improves mean and top-10 reward at 512 newly scored identities on every task. An independent long-horizon factorial nevertheless reveals the acquisition–qualification trade-off: guidance increases yield on Zaleplon and DRD2 but reduces it on Albuterol, where slower acquisition outweighs better qualification. This motivates our analysis of local discovery alignment and the distinction between identity selection and uniform loss scaling.

Contributions. **(1) Candidate-set objective.** We formulate cumulative marginal set utility under separate oracle and generation budgets. **(2) Two-stage method and analysis.** We couple prefix-action guidance with identity-weighted learning, derive a local discovery-alignment condition, and distinguish identity selection from uniform scaling. **(3) Task-wide and component evidence.** We evaluate 23 tasks and five seeds against three optimizers, with early ablations and a three-task factorial separating quality from acquisition.

2 Candidate-Set Molecular Optimization

For a valid generated string x, let $c(x)$ be its canonical isomeric identity and $R(x)$ its terminal reward. Let $\mathcal{C}_0$ contain the initial scored archive's identities and $\mathcal{C}_t$ also include all valid identities observed in t generation attempts. New candidates belong to $\mathcal{C}_t \setminus \mathcal{C}_0$; archive hits and repeats add none. Their total count N_t and qualifying count H_t are

$$N_t = |\mathcal{C}_t \setminus \mathcal{C}_0|, \qquad H_t = \sum_{c \in \mathcal{C}_t \setminus \mathcal{C}_0} \mathbf{1}\{R(c) \geq \tau\}, \tag{1}$$

where τ is the initial archive's 90th-percentile reward, fixed before optimization and shared by methods. To separate acquisition from qualification, we also report the qualifying fraction $f_t = H_t/N_t$ for $N_t > 0$. These quantities satisfy

$$H_t = N_t f_t. \tag{2}$$

2.1 MARGINAL CANDIDATE-SET UTILITY

Let $S \supseteq \mathcal{C}_0$ contain the available identities and $v : [0,1] \to [0,1]$ map reward to candidate utility. Accumulated utility $F(S)$ and the marginal contribution of output X are

$$F(S) = \sum_{c \in S \setminus \mathcal{C}_0} v(R(c)), \qquad \Delta(X; S) = \mathbf{1}\{c(X) \notin S\} v(R(X)), \tag{3}$$

with $\Delta = 0$ for invalid outputs. Archive hits and within-run repeats leave F unchanged regardless of reward. Choosing $v(r) = \mathbf{1}\{r \geq \tau\}$ counts each new qualifying identity once; cumulative marginal contributions then give H_t in Eq. (1). A different SMILES string for an already available molecule therefore adds no candidate, whereas a new low-reward identity increases N_t but not H_t. This separates the acquisition of a new molecule from its qualification under the task criterion.

2.2 ORACLE AND GENERATION BUDGETS

Every generated string consumes an attempt, including invalid outputs, archive hits, and repeats, whereas an oracle call is charged only when a valid identity requires an uncached score. Score reuse therefore saves evaluations but not generation attempts. We accordingly evaluate reward quality on the oracle-call axis and acquisition through N_t and H_t on the generation-attempt axis. Scaffold breadth, fingerprint dispersion, and chemical descriptors then characterize the acquired set separately from its size.

3 AVID: PREFIX GUIDANCE AND IDENTITY-WEIGHTED LEARNING

3.1 PREFIX-ACTION GUIDANCE BEFORE COMPLETION

Q-Steer supplies AVID's sampling guidance (Figure 1) from a scored archive $\mathcal{D} = \{(x_j, R_j)\}$. Each SMILES prefix $s = x_{j,<t}$ and observed next token $a = x_{j,t}$ are paired with terminal reward R_j, which the predictor $g_\phi(s, a)$ learns after label normalization fitted on training molecules. The target belongs to the completed molecule, so its score supervises multiple observed prefix–action pairs without additional oracle evaluations. Once trained, the guide remains frozen as the policy adapts online, providing a reusable reward preference rather than a value function updated from the changing policy's rollouts.

During generation, scores over the admissible actions $\mathcal{A}(s)$ at prefix s are standardized using mean $\bar{g}_{\mathcal{A}}(s)$ and variance $v_{\mathcal{A}}(s)$, scaled by β, and clipped. The resulting bonus is added to the current actor logits to obtain the guided distribution:

$$z_\phi(s, a) = \frac{g_\phi(s, a) - \bar{g}_{\mathcal{A}}(s)}{\sqrt{\max\{v_{\mathcal{A}}(s), \epsilon_g\}}}, \tag{4}$$

$$b_\beta(s, a) = \mathrm{clip}(\beta z_\phi(s, a), -b_{\max}, b_{\max}), \tag{5}$$

$$q_{\theta,\beta}(a \mid s) \propto \exp(\ell_\theta(s, a) + m(s, a) + b_\beta(s, a)) \,. \tag{6}$$

Here ℓ_θ are actor logits, m is the base sampling mask, ϵ_g is the variance floor, and $b_{\max}$ bounds the bonus. Because the bonus modifies logits rather than molecular scores, guidance changes sampling while preserving the task reward. Appendix A gives the architecture and training settings.

3.2 IDENTITY-WEIGHTED LEARNING AFTER COMPLETION

Once a molecule is complete, AVID canonicalizes valid outputs and checks their identities against the run-level ledger. This operation is distinct from reward caching: the cache determines whether an evaluation is needed, whereas the ledger determines whether the trajectory repeats a within-run identity. Only such repeats receive zero direct-policy weight; archive hits retain the base weight. For trajectory i, this rule is expressed as

$$d_i = \mathbf{1}\{\text{trajectory } i \text{ is a within-run repeat}\}, \qquad w_i = 1 - d_i. \tag{7}$$

Property rewards, including cached scores, remain unchanged. Because repetition is determined after completion, every active token in a trajectory receives the same identity weight; the rule does not

prevent a repeated molecule from being sampled. Let u_{it} be the base active-token mask and L_{it}^{pol} the direct policy term for token t in trajectory i. Applying w_i retains the original reduction:

$$\widetilde{L}_{it}^{\mathrm{pol}} = u_{it} w_i L_{it}^{\mathrm{pol}}, \qquad L_{\mathrm{AVID}}^{\mathrm{pol}} = \mathcal{R}_{\mathrm{base}}\left(\{\widetilde{L}_{it}^{\mathrm{pol}}\}_{i,t}\right), \tag{8}$$

where $\mathcal{R}_{\mathrm{base}}$ retains base aggregation and normalization. LSTM-PPO's critic, entropy, prior regularization, global gradient clipping, optimizer-state updates, and stopping rules are preserved. Repeats still affect auxiliary losses and batch statistics; zero direct-policy weight is not a separate repeat penalty. Appendix A specifies the reduction and accounting.

3.3 Analysis

When is reward guidance aligned with discovery? We first consider how guidance changes the expected utility of an action's completed output. For available identities S, let X follow action a at prefix s under a fixed continuation policy π. Its expected marginal discovery value is

$$D^{\pi}(s, a; S) = \mathbb{E}_{\pi}[\Delta(X; S) \mid s, a]. \tag{9}$$

By Eq. (3), this factors into new-identity probability and conditional utility:

$$D^{\pi}(s, a; S) = \Pr_{\pi}(c(X) \notin S \mid s, a)\mathbb{E}_{\pi}[v(R(X)) \mid c(X) \notin S, s, a]. \tag{10}$$

The first factor depends on which identities the run has already discovered, while the second measures the usefulness of a new output. A reward-only guide does not explicitly track the first factor, so high expected reward can coexist with low discovery value. To isolate the effect of guidance strength, hold s, S, the actor, and continuation policy fixed and vary only β. Away from clipping boundaries, write $h_{\beta}(a) = \partial b_{\beta}(s, a)/\partial\beta$. Then

$$\frac{\mathrm{d}}{\mathrm{d}\beta}\mathbb{E}_{a\sim q_{\theta,\beta}}[D^{\pi}(s, a; S)] = \mathrm{Cov}_{q_{\theta,\beta}}(D^{\pi}(s, a; S), h_{\beta}(a)). \tag{11}$$

Thus, increasing β locally improves expected marginal utility when discovery value and bonus sensitivity have positive covariance. The criterion connects reward guidance to the chance of acquiring a useful new identity, but holds the prefix and continuation fixed; whole-run effects also depend on subsequent sampling and policy updates.

Why is identity weighting different from uniform scaling? For n fixed trajectories, let g_i be the summed direct-policy gradient before global clipping and optimizer transformation, $Z_{\mathrm{base}} > 0$ the base normalizer, and $\rho = n^{-1}\sum_i d_i$ the repeat fraction. Full, identity-weighted, and uniformly scaled gradients are

$$G_{\mathrm{full}} = Z_{\mathrm{base}}^{-1}\sum_i g_i, \quad G_{\mathrm{id}} = Z_{\mathrm{base}}^{-1}\sum_i (1 - d_i)g_i, \quad G_{\mathrm{scale}} = (1 - \rho)G_{\mathrm{full}}. \tag{12}$$

Their difference is

$$G_{\mathrm{id}} - G_{\mathrm{scale}} = -\frac{n}{Z_{\mathrm{base}}}\mathrm{Cov}_{\mathrm{emp}}(d_i, g_i). \tag{13}$$

Unlike uniform scaling, identity weighting subtracts a repeat-associated covariance term and can change gradient direction as well as magnitude. If the removed contribution is not collinear with the full gradient, no scalar rescaling reproduces the identity-weighted direction, even when it matches the total retained policy weight. The raw gradients coincide only when the covariance in Eq. (13) vanishes. This is an operator distinction before clipping and optimizer transformation; whether it improves discovery requires empirical controls. Appendix D.6 reports those diagnostics separately from complete-method comparisons.

4 Experimental Setup

We evaluate an LSTM–PPO implementation of AVID on all 23 PMO tasks (Gao et al., 2022) with five prescribed online seeds per task. The protocol separates reward improvement per oracle call from candidate acquisition per generation attempt, so that repeated and invalid outputs remain visible in the resource accounting. Appendix A specifies the architecture, hyperparameters, archives, and accounting rules used throughout these comparisons.

4.1 BENCHMARKS AND COMPARISON METHODS

We compare complete-method performance with standard PPO (Schulman et al., 2017), standard REINVENT (Blaschke et al., 2020a), and Augmented Memory (Guo & Schwaller, 2024a), evaluating each method with the task's PMO property score before method-specific shaping. To separate component effects, we also compare Base, Q-Steer only, identity weighting only (Mask-only), and AVID under a shared implementation; Base is therefore an ablation reference rather than a substitute for standard PPO. Matched subsets of random-count, uniform-weight, and reduction controls further probe how the amount and normalization of direct policy learning affect the results (Appendix D.6).

4.2 METRICS AND RESOURCE BUDGETS

We measure PMO sample efficiency by Top-10 AUC over 10,000 oracle calls, complemented by Top-1/100 AUC and terminal Top-1/10/100. Prespecified scientific terminations follow the official evaluator's `finish=True` convention, carrying terminal top-k values over the remaining AUC interval. The complementary discovery endpoints, H_{32768} and N_{32768}, use the archive-fixed threshold from Section 2 and a ceiling of 10,000 newly scored identities. If that ceiling is reached before 32,768 attempts, the run retains its terminal candidate set thereafter. Other prespecified stops contribute only observed attempt checkpoints; missing endpoints are neither imputed nor extrapolated.

To distinguish candidate quality from acquisition speed, we additionally compare rewards at a common new-identity count, using only pairs in which both runs attain that count. We characterize the acquired sets through absolute Bemis–Murcko scaffold counts (Bemis & Murcko, 1996), Morgan-fingerprint internal diversity (Rogers & Hahn, 2010), QED (Bickerton et al., 2012), synthetic-accessibility score (Ertl & Schuffenhauer, 2009), and structural alerts. These descriptors assess structural breadth and chemical properties separately from set size and PMO reward quality.

4.3 COMPONENT STUDIES AND MATCHED COMPARISONS

The early component study crosses guidance and identity weighting over all 23 tasks and five seeds, measuring quality at 512 newly scored identities. Within this design, AVID versus Mask-only isolates guidance with identity weighting present. Each comparison uses common task–seed blocks and shared archive conventions, endpoints, and method versions. Runs that do not reach 512 remain in attainment counts but provide no endpoint-quality estimate; unavailable runs are not replaced by different seeds or pooled with unmatched observations.

Long-horizon external comparisons instead match generation budgets. An independent 2×2 factorial on Albuterol, Zaleplon, and DRD2 crosses guidance and identity weighting over five seeds per condition to separate their effects on acquisition and qualification. This within-family design complements the end-to-end baselines by examining how the two interventions interact as identities accumulate. The additional operator controls have incomplete support, so they do not fully isolate selective weighting from nonselective alternatives.

4.4 AGGREGATION AND UNCERTAINTY

Task-level summaries report five-seed mean $\pm$ sample standard deviation when all five outcomes are available, and the canonical PMO aggregate sums the 23 seed-averaged task means. For incomplete long-horizon comparisons, aggregation is restricted to matched support rather than presented as a full-suite estimate. Factorial contrasts pair seeds within each task to preserve the experimental blocks; Appendix D and Appendix D.6 specify the corresponding uncertainty procedures and diagnostic controls.

5 RESULTS

5.1 HIGH-QUALITY CANDIDATE-SET EXPANSION

Table 1 compares distinct qualifying yield H_{32768} within 32,768 attempts and at most 10,000 oracle calls. AVID yields 7,114.1 high-scoring candidates per task on average, compared with 4,964.1 for REINVENT and 4,889.1 for Augmented Memory. Relative to Augmented Memory, this is a

Table 1: **High-scoring molecular discovery across 23 PMO tasks.** Entries report H_{32768} (↑): the number of distinct newly scored molecules meeting the task's offline-archive 90th-percentile reward threshold within 32,768 generation attempts and at most 10,000 oracle calls. A run reaching the oracle limit earlier retains its final candidate set. Values are mean ± sample standard deviation over five seeds; the last row averages task means equally. Bold marks the highest reported mean in each row.

Task	Standard PPO	Standard REINVENT	Augmented Memory	AVID
Albuterol similarity	2,137.4 ± 220.3	**9,890.4** ± 9.2	3,108.0 ± 587.0	7,170.0 ± 1,267.3
Amlodipine MPO	4,444.6 ± 2,354.2	2,828.4 ± 172.8	5,755.0 ± 1,438.7	**6,605.6** ± 290.8
Celecoxib rediscovery	1,092.0 ± 110.8	**6,356.8** ± 68.4	1,982.8 ± 189.6	2,869.4 ± 380.7
Deco hop	5,683.8 ± 727.4	3,678.6 ± 251.6	6,328.6 ± 469.4	**7,869.4** ± 357.6
DRD2	8,072.8 ± 1,718.0	6,827.2 ± 232.0	5,497.4 ± 1,083.5	**9,661.2** ± 27.8
Fexofenadine MPO	7,835.4 ± 1,687.0	7,423.8 ± 71.4	8,099.2 ± 732.4	**9,609.8** ± 43.2
GSK3B	1,570.4 ± 307.8	7,034.8 ± 92.8	2,907.0 ± 469.7	**9,774.8** ± 26.4
Isomers $C_9H_{10}N_2O_2PF_2Cl$	1,208.0 ± 519.3	6,307.6 ± 22.7	3,484.0 ± 540.5	**6,609.4** ± 1,307.2
Isomers $C_7H_8N_2O_2$	871.8 ± 398.1	**6,876.4** ± 687.6	1,849.6 ± 281.1	2,933.6 ± 447.3
JNK3	1,631.6 ± 418.1	6,090.8 ± 483.2	4,386.0 ± 547.2	**9,530.2** ± 49.8
Median molecules 1	1,712.6 ± 362.2	7,008.0 ± 44.4	2,634.0 ± 445.6	**9,716.6** ± 507.5
Median molecules 2	1,435.0 ± 376.1	3,344.2 ± 116.2	2,568.2 ± 199.6	**9,041.4** ± 357.0
Mestranol similarity	1,720.8 ± 487.0	**6,236.2** ± 76.1	4,359.8 ± 619.9	4,807.0 ± 590.3
Osimertinib MPO	7,805.0 ± 680.1	5,136.2 ± 130.7	**8,914.8** ± 550.6	8,483.6 ± 161.9
Perindopril MPO	4,522.2 ± 1,889.3	3,133.0 ± 43.6	**8,232.8** ± 140.6	7,448.6 ± 1,984.5
QED	6,872.4 ± 1,372.6	2,664.6 ± 80.9	5,357.6 ± 672.2	**7,282.6** ± 596.7
Ranolazine MPO	8,758.0 ± 839.1	6,291.4 ± 162.0	9,421.4 ± 44.9	**9,558.2** ± 43.3
Scaffold hop	7,008.8 ± 858.9	3,766.8 ± 153.7	**7,835.8** ± 235.4	6,620.6 ± 201.9
Sitagliptin MPO	497.0 ± 201.6	1,710.6 ± 78.4	**4,442.0** ± 2,260.9	3,336.0 ± 393.4
Thiothixene rediscovery	1,814.8 ± 647.7	4,572.4 ± 74.3	3,101.0 ± 377.8	**9,498.2** ± 512.9
Troglitazone rediscovery	3,212.0 ± 1,372.7	3,546.4 ± 150.4	5,161.8 ± 216.6	**6,486.4** ± 2,021.7
Valsartan SMARTS	504.8 ± 141.5	821.2 ± 26.5	**3,448.6** ± 392.1	2,067.2 ± 758.1
Zaleplon MPO	1,105.8 ± 660.5	2,628.4 ± 78.2	3,572.8 ± 815.8	**6,644.8** ± 98.2
23-task mean	3,544.2	4,964.1	4,889.1	**7,114.1**

45.5% increase in qualifying candidates, accompanied by more distinct identities (8,800.9 versus 6,051.1) and similar task-averaged qualifying fractions (82.20% versus 81.84%). The yield advantage therefore reflects expansion of the acquired set rather than a large change in the fraction passing the shared threshold.

This acquisition advantage complements the PMO quality comparison in Appendix Table 5, where Augmented Memory has a higher 23-task Top-10 AUC sum than AVID (16.1875 versus 15.4526; REINVENT: 13.9226). Optimizing the best few molecules and expanding a qualifying set thus need not induce the same ordering. Figure 2 illustrates this distinction through quality and acquisition trajectories on Albuterol and DRD2. The full panel in Appendix Figure 3 also includes Zaleplon, where AVID has lower Top-10 quality but higher yield than Augmented Memory. These examples explain the dynamics rather than establish a benchmark-wide ranking.

5.2 Guidance improves early quality across all 23 tasks

Adding Q-Steer to identity weighting produces a broad early quality gain. Relative to Mask-only, AVID increases the task-macro Mean@512 from 0.1973 to 0.2880 and Top-10@512 from 0.4684 to 0.5497, corresponding to gains of 0.0907 and 0.0813. The task-level mean difference is positive on all 23 PMO tasks for both measures. At the task–seed level, the direction is positive on 114/115 Mean@512 pairs and 108/115 Top-10@512 pairs. Complete per-task mean±SD values are given in Appendix Tables 6 and 7.

5.3 AVID increases fixed-budget distinct yield on most matched tasks

Across all 23 tasks and five paired seeds, AVID achieves higher task-mean distinct yield on 22 tasks relative to standard PPO, 19 relative to REINVENT, and 18 relative to Augmented Memory (Table 2,

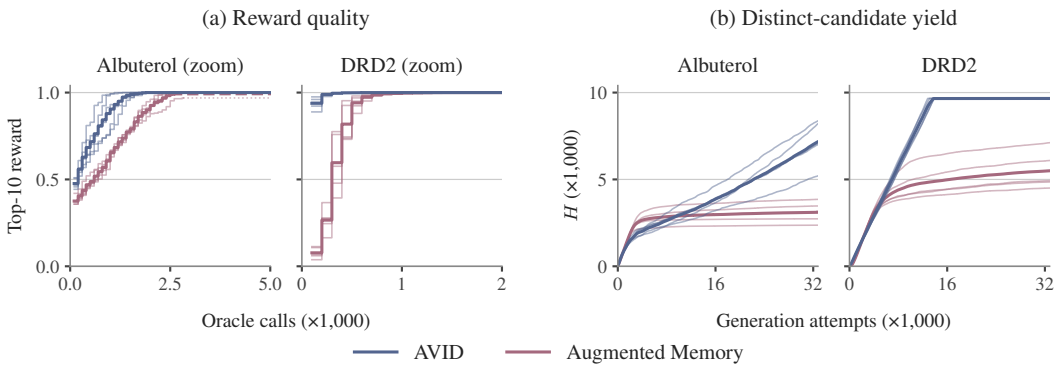

Figure 2: **Reward quality and distinct-candidate acquisition.** Thick/thin lines show the mean/individual trajectories of five matched seeds. **(a)** Best-ten reward versus oracle calls; Albuterol and DRD2 are zoomed to the first 5,000 and 2,000 calls, respectively. Dashed segments denote PMO terminal-value carry. **(b)** H_t counts distinct newly scored molecules meeting the task's fixed archive 90th-percentile reward threshold by attempt t, in thousands; repeats do not count. Full trajectories, including Zaleplon, appear in Appendix Figure 3.

Table 2: **Core effectiveness and ablation. Panel A:** fixed-attempt distinct yield against complete optimizer baselines. The AVID-higher column counts tasks whose matched-seed mean H_{32768} is larger for AVID. **Panel B:** early quality is task-averaged over the same 113 task–seed blocks attaining 512 new identities in all four conditions; attainment counts retain all 115 runs per condition. Appendix C.3 details the two excluded blocks. The independent three-task factorial reports 10k-new-identity attainment.

Panel A: fixed-budget candidate acquisition

Comparator	Matched H pairs	Tasks compared	AVID higher task mean	Tasks at 5/5
Standard PPO	115/115	23/23	**22/23**	23/23
Standard REINVENT	115/115	23/23	**19/23**	23/23
Augmented Memory	115/115	23/23	**18/23**	23/23

Panel B: component behavior

Condition	Guide	Identity wt.	Mean@512 ↑	Top-10@512 ↑	Reached 512	Reached 10k (3 tasks)
Base	–	–	0.1973	0.4674	115/115	5/15
Q-Steer only	✓	–	**0.2882**	**0.5507**	113/115	2/15
Mask only	–	✓	0.1972	0.4674	115/115	15/15
AVID	✓	✓	0.2881	0.5487	115/115	15/15

Panel A). Table 1 reports every task, including those favoring the comparator. Standard PPO yields 3,544.2 qualifying candidates per task on average, compared with AVID's 7,114.1.

The internal contrasts refine this picture. In the three-task long-horizon factorial, Base and Q-Steer-only reach the 10,000-new-identity ceiling in 5/15 and 2/15 runs, whereas Mask-only and AVID reach it in all 15. Across the broader AVID–Mask-only fixed-attempt comparison, 91/115 seed pairs are matched; among the supported task means, four favor Mask-only. Thus the guide's effect on long-horizon distinct yield is less uniform than its early quality effect, motivating a direct analysis of the acquisition–qualification trade-off below. Full signed task-level contrasts are reported in Appendix Table 8.

The reduction controls distinguish complete-method gains from the effect of identity selection alone. On Amlodipine, active-token normalization retains both Q-Steer and identity weighting but exceeds the frozen AVID reduction on all five seeds, with mean AVID-minus-active-token $H_{32768} = -1{,}536.8$. Together with the limited coverage of random-count and uniform-weight controls, this result supports reporting the long-horizon gains for the complete AVID configuration, without assigning them uniquely to identity selection. Appendix D.6 gives the matched diagnostic comparisons.

Table 3: **Matched candidate-set characterization.** Entries are task-first AVID-minus-comparator differences on each comparator's matched support. H is high-scoring distinct yield; Murcko is the absolute scaffold count; IntDiv is sampled Morgan-fingerprint internal diversity. Lower raw SA is favorable. Support differs by comparator, so rows should be interpreted within their own matched blocks.

Comparator	Support	$\Delta H_{32768}\uparrow$	Δ Murcko $\uparrow$	Δ IntDiv $\uparrow$	Δ QED $\uparrow$	Δ raw SA $\downarrow$	Δ alert-free $\uparrow$
Standard PPO	85/115	+3,630.8	+2,102.2	$+0.023$	-0.011	$+0.08$	$+0.101$
Standard REINVENT	73/115	+2,705.5	+21.2	-0.085	$+0.002$	-0.07	$+0.062$
Augmented Memory	40/115	+2,582.0	+931.9	-0.016	-0.020	-0.39	$+0.040$

Table 4: **Acquisition–qualification decomposition in the three-task factorial.** Values are five-seed means at 32,768 generation attempts under the 10k-new-identity ceiling. f is the mean of the per-seed qualifying fractions H/N. The last column is AVID minus Mask-only distinct yield.

Task	N Mask-only	N AVID	f Mask-only (%)	f AVID (%)	ΔH
Albuterol	9,175.2	6,704.8	88.96	96.34	$-1{,}707.6$
Zaleplon	10,000.0	10,000.0	48.99	71.70	$+2{,}271.6$
DRD2	10,000.0	10,000.0	89.66	95.88	$+621.4$

5.4 More identities do not imply uniformly greater normalized diversity

The additional candidates expand absolute structural breadth, but normalized dispersion depends on the comparator. On the matched descriptor supports in Table 3, AVID increases both distinct yield and the absolute number of Murcko scaffolds relative to PPO, REINVENT, and Augmented Memory. Internal fingerprint diversity also increases relative to PPO, but decreases relative to REINVENT and Augmented Memory. QED and raw synthetic-accessibility score move in mixed directions, while the alert-free fraction increases in all three comparison blocks.

Absolute breadth and normalized spread therefore describe different properties of the candidate set. In particular, acquiring many additional identities can increase the number of scaffolds while reducing average pairwise fingerprint dispersion. Detailed cluster coverage, largest-cluster fraction, and task-level descriptor values are reported in Appendix E–E.1. These computational descriptors characterize the generated set; they do not establish synthetic feasibility, biological activity, or safety.

5.5 Distinct yield depends on both acquisition and qualification

The controlled factorial exposes the two terms in $H_t = N_t f_t$ directly. Table 4 reports five-seed means of N_t, f_t, and the AVID–Mask-only change in H_t at 32,768 attempts. On Zaleplon and DRD2, both methods have reached the 10,000-new-identity ceiling, so AVID's higher qualifying fraction translates directly into higher yield. Albuterol exhibits the opposite outcome: AVID raises the mean qualifying fraction from 88.96% to 96.34%, but acquires 2,470.4 fewer identities on average, reducing yield by 1,707.6. This reversal shows why quality alone does not determine discovery; Appendix D reports the absolute factorial profiles and per-seed contrasts.

6 Related Work

Sample-efficient molecular optimization. Memory-assisted RL changes rewards as similarity or scaffold buckets saturate (Blaschke et al., 2020b); Augmented Memory combines replay, SMILES augmentation, and selective memory purging (Guo & Schwaller, 2024a). REINVENT 4 provides a transfer- and reinforcement-learning framework (Loeffler et al., 2024), while Saturn studies model capacity, augmented replay, efficiency, and diversity (Guo et al., 2026). AVID instead couples token-sampling guidance with exact-identity weighting of the direct policy term, retaining cached rewards and other optimizer operations.

Guided generation and alternative search. FUDGE predicts future attributes from prefixes (Yang & Klein, 2021); Reward-Augmented Decoding and Transfer Q^* guide decoding with reward or value estimates (Deng & Raffel, 2023; Chakraborty et al., 2024). Beam Enumeration exploits probable molecular substructures (Guo & Schwaller, 2024b). Graph policies, genetic search, Bayesian optimization, and generative flow networks offer alternative search strategies (You et al., 2018; Jensen, 2019; Korovina et al., 2020; Bengio et al., 2021). AVID links prefix guidance to a separate, history-dependent selection of completed trajectories for policy learning.

Set-aware objectives and candidate-set evaluation. SGRPO and SetPO assign group-level or leave-one-out diversity credit within sampled sets (Ye et al., 2026; Li et al., 2026), whereas AVID tracks exact identities across the run and suppresses repeats' direct policy terms rather than estimating an exact set-utility advantage. This distinction also affects evaluation: GuacaMol and MOSES assess generation quality and diversity (Brown et al., 2019; Polykovskiy et al., 2020), while PMO standardizes oracle-call efficiency (Gao et al., 2022). Our acquisition counts N_t and H_t complement these views, with canonical PMO Top-10 AUC reported separately. Appendix F expands the methodological comparison.

7 Discussion and Limitations

The broad comparisons support candidate-acquisition gains for complete AVID, while the early ablation isolates guidance's quality effect with identity weighting present. Their distinction is exposed by Albuterol, where better qualification is outweighed by slower acquisition. Candidate counts must also be distinguished from structural diversity: AVID increases qualified identities and absolute Murcko counts on matched supports, but fingerprint dispersion rises relative to PPO and falls relative to REINVENT and Augmented Memory. A larger set can span more scaffolds while remaining locally concentrated. These benchmark rewards and descriptors characterize computational optimization, not synthesizability, biological activity, or safety.

The two interventions use different information, which determines their scope. The frozen guide does not condition on the evolving discovered set, so accurate reward prediction need not remain aligned with discovery. The ledger tracks canonical identities rather than scaffolds, three-dimensional similarity, or learned novelty, and masks only within-run repeats while retaining archive hits' policy weight. It therefore does not implement the exact marginal utility in Eq. (3). Richer novelty-based credit requires separate validation, as does transfer beyond the evaluated LSTM–PPO setting, where probability accounting, reductions, and auxiliary losses must be checked for each optimizer.

Component attribution likewise requires distinguishing the implemented configuration from an individual operator. The gradient identity in Section 3 separates identity selection from uniform scaling, but does not establish selection as the unique source of long-horizon gains. Amlodipine's five-seed active-token variant retains guidance and identity weighting yet outperforms the frozen reduction, while random-count and uniform-weight controls remain incomplete. The evidence consequently supports complete AVID and the isolated early guidance effect, with stronger attribution to identity selection requiring additional matched controls.

8 Conclusion

AVID couples action-value guidance before molecular completion with identity-weighted policy learning afterward, targeting the expansion of a qualifying candidate set. Across 23 PMO tasks, it acquires 45.5% more qualifying candidates on average than Augmented Memory under the matched generation-attempt and new-score caps. Component studies identify a consistent early quality gain from guidance, while its long-horizon contribution depends on the balance between new-identity acquisition and the qualifying fraction. These results motivate evaluating molecular optimization through both reward quality and the growth of the useful candidate set.

AI USAGE

The research direction and central ideas were developed by the authors. Generative AI tools assisted with refining hypotheses and mathematical arguments, code and analysis support, interpretation of results, manuscript editing, and figure preparation. The authors reviewed and verified AI-assisted outputs against the cited literature, retained source and result artifacts, and executable analyses used in the paper. The authors take full responsibility for the final content, results, and conclusions.

REPRODUCIBILITY STATEMENT

The source package records task definitions, seeds, identity accounting, stopping rules, analysis scripts, and result provenance for the reported studies. Tables and figures are generated from versioned artifacts, and Appendix A specifies the evaluated implementation and budget conventions.

ETHICS STATEMENT

Molecular benchmark rewards are computational proxies rather than evidence of synthetic feasibility, clinical efficacy, or safety. Candidate identity is defined relative to a specified archive and run. Deployment requires expert review, objective constraints, and experimental validation.

REFERENCES

Guy W. Bemis and Mark A. Murcko. The properties of known drugs. 1. molecular frameworks. *Journal of Medicinal Chemistry*, 39(15):2887–2893, 1996. doi: 10.1021/jm9602928. URL https://doi.org/10.1021/jm9602928.

Emmanuel Bengio, Moksh Jain, Maksym Korablyov, Doina Precup, and Yoshua Bengio. Flow network based generative models for non-iterative diverse candidate generation. In *Advances in Neural Information Processing Systems*, volume 34, pp. 27381–27394, 2021. URL https://arxiv.org/abs/2106.04399.

G. Richard Bickerton, Gaia V. Paolini, Jeremy Besnard, Sorel Muresan, and Andrew L. Hopkins. Quantifying the chemical beauty of drugs. *Nature Chemistry*, 4:90–98, 2012. doi: 10.1038/nchem.1243. URL https://doi.org/10.1038/nchem.1243.

Thomas Blaschke, Josep Arús-Pous, Hongming Chen, Christian Margreitter, Christian Tyrchan, Ola Engkvist, Kostas Papadopoulos, and Atanas Patronov. REINVENT 2.0: An AI tool for de novo drug design. *Journal of Chemical Information and Modeling*, 60(12):5918–5922, 2020a. doi: 10.1021/acs.jcim.0c00915. URL https://doi.org/10.1021/acs.jcim.0c00915.

Thomas Blaschke, Ola Engkvist, Jürgen Bajorath, and Hongming Chen. Memory-assisted reinforcement learning for diverse molecular de novo design. *Journal of Cheminformatics*, 12:68, 2020b. doi: 10.1186/s13321-020-00473-0. URL https://doi.org/10.1186/s13321-020-00473-0.

Albert Bou, Morgan Thomas, Sebastian Dittert, Carles Navarro Ramírez, Maciej Majewski, Ye Wang, Shivam Patel, Gary Tresadern, Mazen Ahmad, Vincent Moens, Woody Sherman, Simone Sciabola, and Gianni De Fabritiis. ACEGEN: Reinforcement learning of generative chemical agents for drug discovery. *Journal of Chemical Information and Modeling*, 64(15):5900–5911, 2024. doi: 10.1021/acs.jcim.4c00895. URL https://doi.org/10.1021/acs.jcim.4c00895.

Nathan Brown, Marco Fiscato, Marwin H. S. Segler, and Alain C. Vaucher. GuacaMol: Benchmarking models for de novo molecular design. *Journal of Chemical Information and Modeling*, 59(3): 1096–1108, 2019. doi: 10.1021/acs.jcim.8b00839. URL https://doi.org/10.1021/acs.jcim.8b00839.

Souradip Chakraborty, Soumya Suvra Ghosal, Ming Yin, Dinesh Manocha, Mengdi Wang, Amrit Singh Bedi, and Furong Huang. Transfer Q star: Principled decoding for LLM alignment. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2405.20495.

Binghong Chen, Tianzhe Wang, Chengtao Li, Hanjun Dai, and Le Song. Molecule optimization by explainable evolution. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=jHefDGsorp5.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. Plug and play language models: A simple approach to controlled text generation. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=H1edEyBKDS.

Haikang Deng and Colin Raffel. Reward-augmented decoding: Efficient controlled text generation with a unidirectional reward model. In *Proceedings of EMNLP*, pp. 11781–11791. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.721. URL https://aclanthology.org/2023.emnlp-main.721/.

Peter Ertl and Ansgar Schuffenhauer. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *Journal of Cheminformatics*, 1:8, 2009. doi: 10.1186/1758-2946-1-8. URL https://doi.org/10.1186/1758-2946-1-8.

Wenhao Gao, Tianfan Fu, Jimeng Sun, and Connor W. Coley. Sample efficiency matters: A benchmark for practical molecular optimization. In *Advances in Neural Information Processing Systems*, volume 35, pp. 21342–21357, 2022. URL https://arxiv.org/abs/2206.12411.

Rafael Gomez-Bombarelli, Jennifer N. Wei, David Duvenaud, Jose Miguel Hernandez-Lobato, Benjamin Sanchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D. Hirzel, Ryan P. Adams, and Alan Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Central Science*, 4(2):268–276, 2018. doi: 10.1021/acscentsci.7b00572. URL https://doi.org/10.1021/acscentsci.7b00572.

Jeff Guo and Philippe Schwaller. Augmented memory: Sample-efficient generative molecular design with reinforcement learning. *JACS Au*, 4(6):2160–2172, 2024a. doi: 10.1021/jacsau.4c00066. URL https://doi.org/10.1021/jacsau.4c00066.

Jeff Guo and Philippe Schwaller. Beam enumeration: Probabilistic explainability for sample efficient self-conditioned molecular design. In *International Conference on Learning Representations*, 2024b. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/ead542f13a38179d1b55b88610f959a1-Abstract-Conference.html.

Jeff Guo, Junwu Chen, Anthony GX-Chen, and Philippe Schwaller. Sample-efficient generative molecular design using memory manipulation. *Nature Machine Intelligence*, 8:449–460, 2026. doi: 10.1038/s42256-026-01200-4. URL https://www.nature.com/articles/s42256-026-01200-4.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rygGQyrFvH.

Emiel Hoogeboom, Victor Garcia Satorras, Clement Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3d. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 8867–8887, 2022. URL https://proceedings.mlr.press/v162/hoogeboom22a.html.

Jan H. Jensen. A graph-based genetic algorithm and generative model/monte carlo tree search for the exploration of chemical space. *Chemical Science*, 10:3567–3572, 2019. doi: 10.1039/C8SC05372C. URL https://doi.org/10.1039/C8SC05372C.

Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Junction tree variational autoencoder for molecular graph generation. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 2323–2332, 2018. URL https://proceedings.mlr.press/v80/jin18a.html.

Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Hierarchical generation of molecular graphs using structural motifs. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 4839–4848, 2020. URL https://proceedings.mlr.press/v119/jin20a.html.

Ksenia Korovina, Sailun Xu, Kirthevasan Kandasamy, Willie Neiswanger, Barnabas Poczos, Jeff Schneider, and Eric P. Xing. ChemBO: Bayesian optimization of small organic molecules with synthesizable recommendations. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, pp. 3393–3403, 2020. URL https://proceedings.mlr.press/v108/korovina20a.html.

Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. GeDi: Generative discriminator guided sequence generation. In *Findings of the Association for Computational Linguistics: EMNLP*, pp. 4929–4952, 2021. doi: 10.18653/v1/2021.findings-emnlp.424. URL https://aclanthology.org/2021.findings-emnlp.424/.

Mario Krenn, Florian Hase, AkshatKumar Nigam, Pascal Friederich, and Alan Aspuru-Guzik. Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation. *Machine Learning: Science and Technology*, 1(4):045024, 2020. doi: 10.1088/2632-2153/aba947. URL https://doi.org/10.1088/2632-2153/aba947.

Chenyi Li, Yuan Zhang, Bo Wang, Guoqing Ma, Wei Tang, Haoyang Huang, and Nan Duan. SetPO: Set-level policy optimization for diversity-preserving LLM reasoning. *arXiv preprint arXiv:2602.01062*, 2026. doi: 10.48550/arXiv.2602.01062. URL https://arxiv.org/abs/2602.01062. Version 1.

Long-Ji Lin. Self-improving reactive agents based on reinforcement learning, planning and teaching. *Machine Learning*, 8:293–321, 1992. doi: 10.1007/BF00992699. URL https://doi.org/10.1007/BF00992699.

Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. DExperts: Decoding-time controlled text generation with experts and anti-experts. In *Proceedings of ACL-IJCNLP*, pp. 6691–6706, 2021. doi: 10.18653/v1/2021.acl-long.522. URL https://aclanthology.org/2021.acl-long.522/.

Hannes H. Loeffler, Jiazhen He, Alessandro Tibo, Jon Paul Janet, Alexey Voronov, Lewis H. Mervin, and Ola Engkvist. REINVENT 4: Modern AI-driven generative molecule design. *Journal of Cheminformatics*, 16:20, 2024. doi: 10.1186/s13321-024-00812-5. URL https://link.springer.com/article/10.1186/s13321-024-00812-5.

Ximing Lu, Sean Welleck, Peter West, Liwei Jiang, Jungo Kasai, Daniel Khashabi, Ronan Le Bras, Lianhui Qin, Youngjae Yu, Rowan Zellers, Noah A. Smith, and Yejin Choi. NeuroLogic A*esque decoding: Constrained text generation with lookahead heuristics. In *Proceedings of NAACL-HLT*, pp. 780–799, 2022. doi: 10.18653/v1/2022.naacl-main.57. URL https://aclanthology.org/2022.naacl-main.57/.

Youzhi Luo, Keqiang Yan, and Shuiwang Ji. GraphDF: A discrete flow model for molecular graph generation. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 7192–7203, 2021. URL https://proceedings.mlr.press/v139/luo21a.html.

AkshatKumar Nigam, Pascal Friederich, Mario Krenn, and Alan Aspuru-Guzik. Augmenting genetic algorithms with deep neural networks for exploring the chemical space. *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=H1lmyRNFvr.

Marcus Olivecrona, Thomas Blaschke, Ola Engkvist, and Hongming Chen. Molecular de-novo design through deep reinforcement learning. *Journal of Cheminformatics*, 9:48, 2017. doi: 10.1186/s13321-017-0235-x. URL https://doi.org/10.1186/s13321-017-0235-x.

Daniil Polykovskiy, Alexander Zhebrak, Benjamin Sanchez-Lengeling, Sergey Golovanov, Oktai Tatanov, Stanislav Belyaev, Rauf Kurbanov, Aleksey Artamonov, Vladimir Aladinskiy, Mark Veselov, Artur Kadurin, Simon Johansson, Hongming Chen, Sergey Nikolenko, Alan Aspuru-Guzik, and Alex Zhavoronkov. Molecular sets (MOSES): A benchmarking platform for molecular generation models. *Frontiers in Pharmacology*, 11:565644, 2020. doi: 10.3389/fphar.2020.565644. URL https://doi.org/10.3389/fphar.2020.565644.

Mariya Popova, Olexandr Isayev, and Alexander Tropsha. Deep reinforcement learning for de novo drug design. *Science Advances*, 4(7):eaap7885, 2018. doi: 10.1126/sciadv.aap7885. URL https://doi.org/10.1126/sciadv.aap7885.

Kristina Preuer, Philipp Renz, Thomas Unterthiner, Sepp Hochreiter, and Gunter Klambauer. Frechet ChemNet distance: A metric for generative models for molecules in drug discovery. *Journal of Chemical Information and Modeling*, 58(9):1736–1741, 2018. doi: 10.1021/acs.jcim.8b00234. URL https://doi.org/10.1021/acs.jcim.8b00234.

David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 2010. doi: 10.1021/ci100050t. URL https://doi.org/10.1021/ci100050t.

John Schulman, Sergey Levine, Philipp Moritz, Michael I. Jordan, and Pieter Abbeel. Trust region policy optimization. In *Proceedings of the 32nd International Conference on Machine Learning*, pp. 1889–1897, 2015. URL https://proceedings.mlr.press/v37/schulman15.html.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

Marwin H. S. Segler, Thierry Kogej, Christian Tyrchan, and Mark P. Waller. Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS Central Science*, 4 (1):120–131, 2018. doi: 10.1021/acscentsci.7b00512. URL https://doi.org/10.1021/acscentsci.7b00512.

Chence Shi, Minkai Xu, Zhaocheng Zhu, Weinan Zhang, Ming Zhang, and Jian Tang. GraphAF: A flow-based autoregressive model for molecular graph generation. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=S1esMkHYPr.

Austin Tripp, Erik Daxberger, and Jose Miguel Hernandez-Lobato. Sample-efficient optimization in the latent space of deep generative models via weighted retraining. In *Advances in Neural Information Processing Systems*, volume 33, pp. 11259–11272, 2020. URL https://arxiv.org/abs/2006.09191.

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. doi: 10.1021/ci00057a005. URL https://doi.org/10.1021/ci00057a005.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. Neural text generation with unlikelihood training. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SJeYe0NtvH.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256, 1992. doi: 10.1007/BF00992696. URL https://doi.org/10.1007/BF00992696.

Yutong Xie, Chence Shi, Hao Zhou, Yuwei Yang, Weinan Zhang, Yong Yu, and Lei Li. MARS: Markov molecular sampling for multi-objective drug discovery. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=kHSu4ebxFXY.

Minkai Xu, Lantao Yu, Yang Song, Chence Shi, Stefano Ermon, and Jian Tang. GeoDiff: A geometric diffusion model for molecular conformation generation. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=PzcvxEMzvQC.

Kevin Yang and Dan Klein. FUDGE: Controlled text generation with future discriminators. In *Proceedings of NAACL-HLT*, pp. 3511–3535. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.naacl-main.276. URL https://aclanthology.org/2021.naacl-main.276/.

Xinwu Ye, He Cao, Hao Li, Bin Feng, Zijing Liu, Xiangru Tang, Yu Li, and Shenghua Gao. Pushing biomolecular utility-diversity frontiers with supergroup relative policy optimization. *arXiv preprint arXiv:2605.08659*, 2026. doi: 10.48550/arXiv.2605.08659. URL https://arxiv.org/abs/2605.08659. Version 3.

Jiaxuan You, Bowen Liu, Rex Ying, Vijay Pande, and Jure Leskovec. Graph convolutional policy network for goal-directed molecular graph generation. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/d60678e8f2ba9c540798ebbde31177e8-Abstract.html.

Zhenpeng Zhou, Steven Kearnes, Li Li, Richard N. Zare, and Patrick Riley. Optimization of molecules via deep reinforcement learning. *Scientific Reports*, 9:10752, 2019. doi: 10.1038/s41598-019-47148-x. URL https://doi.org/10.1038/s41598-019-47148-x.

The appendices provide implementation and evaluation details (A), derivations (B), task-wide results (C), component analyses (D), candidate-set characterization (E), and methodological context (F).

A Implementation and Evaluation Protocol

This appendix specifies the evaluated LSTM–PPO implementation, its identity accounting, and the rules used to compare runs. Architecture and training settings are listed below; the three-task studies use the target-matched archives described here, while Appendix C.3 specifies the archive convention of the broad early-quality study.

A.1 Policy, guide, and archives

The actor uses the LSTM ChEMBL prior in the ACEGEN-based implementation (Bou et al., 2024). The guide is trained offline and remains frozen during online optimization.

Component	Recorded setting
Actor	Three LSTM layers; hidden size 512; embedding size 256; no dropout
Rollout	32 trajectories; padded sequence horizon 200; float32
PPO	One epoch; no replay; learning rate 5×10^{-4}; clipping radius 0.5
Returns	Discount 0.999; GAE parameter 1.0
Auxiliary terms	Critic coefficient 0.25; entropy coefficient 0.01; prior-KL coefficient 0.001
Guide	Token embedding width 128; one-layer GRU of hidden width 128; action embedding/projection; MLP score head
Guide training	AdamW; learning rate 2×10^{-3}; weight decay 10^{-4}; batch size 512; eight epochs; gradient clipping at 5
Guide sampling	Strength $\beta = 0.5$; logit-bonus bound $b_{\max} = 5$; variance floor $\epsilon_g = 10^{-6}$
Online seeds	101, 202, 303, 404, 505

The target-matched guide uses 2,000 archived identities, split into 1,600 training and 400 validation identities before prefix expansion. Reward normalization is fitted on training identities. Checkpoint selection uses the saved validation quantile-ranking criterion with its MSE fallback. Albuterol uses the archive/guide associated with the earlier seed-11 diagnostic; Zaleplon and DRD2 use their seed-23 guides. These bindings select fixed guides, not the online random seed: five online seeds do not represent five independent guide fits.

Archive targets are matched to the online reward. Albuterol's original ECFP4-bit similarity labels were corrected to the transformed FCFP4-count target before the prospective screen; DRD2 and Zaleplon agreed within 10^{-5}. Results using older targets are not pooled with these studies. The offline archive is separate supervision, reused across online runs rather than counted as newly acquired candidates.

A.2 Identity accounting and the policy weight

The reward cache starts with the archive. A valid first-seen nonarchive canonical isomeric identity consumes one new oracle call. Later occurrences, including within the same batch, receive the cached reward and the category `run_repeat`. Archive hits, invalid outputs, and valid overflow outputs beyond the label limit remain distinct categories. Invalid outputs receive zero reward; overflow does not create extra charged labels.

Only `run_repeat` receives zero direct-policy weight. New identities, archive hits, invalid outputs, and overflow retain weight one. Tags stay aligned with trajectories after buffer permutation. The evaluated reduction uses the padded tensor size $[32, 200]$, giving $Z_{\text{base}} = 6{,}400$; active-token and retained-token reductions are separate controls. Repeats remain in advantage normalization, critic fitting, regularizers, global clipping, and optimizer history. The four component conditions share the wrapper that preserves every sampled trajectory instead of applying native duplicate removal or reordering.

A.3 WORKED IDENTITY-LEDGER EXAMPLE

Suppose the archive contains identity C, while A and B are initially absent from both the archive and the run. Process the illustrative outputs below in order, before exhausting the label budget. A repeated A may be written as a different SMILES string; identity, rather than its spelling, determines the category.

Output	Ledger category	Reward source	Set increment	w_i
First A	New	New evaluation	$v(R(\mathrm{A}))$	1
A again	Within-run repeat	Cached	0	0
First B	New	New evaluation	$v(R(\mathrm{B}))$	1
C	Archive hit	Archive cache	0	1

The last column is the trajectory-level identity weight applied to the direct policy objective. It does not replace the active-token mask or the remaining optimizer terms. The archive hit C has zero marginal set increment but retains weight one, whereas only the within-run repeat receives zero direct-policy weight.

A.4 SAMPLING PROBABILITIES AND OPTIMIZER STEPS

The selected action's stored log probability comes from the guided behavior distribution, whereas the numerator in the clipped direct objective is the base actor:

$$r_\theta(s, a) = \frac{\pi_\theta(a \mid s)}{q_{\theta_{\mathrm{old}},\beta}(a \mid s)}. \tag{14}$$

The ratio therefore need not equal one before an update. In the no-update check, 8.43% of 427 active guided tokens lie outside $[0.5, 1.5]$, compared with zero of 386 unguided tokens. The reported method uses the ratio above; the direct-gradient analysis in Appendix B.4 uses the same numerator and denominator.

The optimizer performs four Adam steps on the first trained rollout and one on each later trained rollout. A cap-triggering batch can be scored without a further update. An all-repeat batch retains auxiliary learning operations. In the earlier pairwise study, control and treatment use fresh processes on the same A100-PCIE-40GB, with 16 CPU cores, 64 GB host memory, and four numerical threads; this is a matching rule, not a wall-clock speed comparison.

A.5 Optimization procedure

Algorithm 1 inserts the identity weight from Eq. 8 into the evaluated LSTM–PPO update while retaining its sampling, stopping, and auxiliary learning rules.

Algorithm 1: AVID iteration

Inputs: actor and critic initialized by the native configuration; frozen scorer g_ϕ; archive-initialized reward cache and identity ledger; the original generation, label, and optimizer settings.
Repeat while the native search guard permits another rollout:

1. Sample the batch using Eq. 6. Retain each sampled action's behavior log probability and active-token mask. Count every attempted output, including invalid and repeated strings.
2. For each output in the original ledger order, distinguish eligible new identities, within-run repeats, archive hits, invalid outputs, and budget overflow; perform step 3 before processing the next output.
3. Evaluate only eligible new identities and update the cache and ledger, making later same-batch occurrences repeats. Retrieve scores for repeats and archive hits; preserve native invalid/overflow treatment. Set $d_i = 1$ only for `run_repeat`.
4. Apply the original stopping and overflow checks. If they terminate this rollout before learning, record the terminal state and stop; do not add an update merely because a batch was sampled.
5. Follow the native optimizer loop, including its batch statistics, advantage normalization, data ordering and masks. Before the base reduction, multiply each trajectory's direct policy terms by its aligned identity weight as in Eq. 8.
6. Retain critic, entropy and prior-KL terms, global clipping, and Adam state and stepping. An all-repeat batch does not skip these operations; one rollout does not necessarily mean one Adam step.
7. Record charged identities, attempts, categories, and stopping state; continue with the updated actor and unchanged scorer.

Output: the evaluated discovery set, scores, and budget-indexed records. No post-stop discoveries are imputed.

A.6 Endpoints, area conventions, and eligibility

Let $r_1, \ldots, r_n$ be rewards in charged-identity order. Mean reward is $n^{-1} \sum_i r_i$; endpoint Top-10 averages the best $\min(10, n)$ rewards. The component analyses use a discrete label-axis area with ten fixed slots. If S_k sums the best $\min(10, k)$ rewards among the first k charged identities, then

$$A_{10}^{\text{fixed}}(B) = \frac{1}{B} \sum_{k=1}^{B} \frac{S_k}{10}. \tag{15}$$

Unfilled slots contribute zero. This quantity requires the ordered prefix through B and is distinct from endpoint Top-10 and the canonical PMO evaluator. A shorter run is neither padded nor renormalized to represent an attained label budget.

Canonical PMO admission uses the original generated string: it must be nonempty and parseable, and one RDKit canonical serialization must match the recorded buffer key. Calls retain their original order and rewards. Runtime-charged empty proposals fail this equivalence check; affected runs remain unavailable for canonical PMO AUC rather than being repaired by deleting calls. Prespecified scientific terminations use PMO's `finish=True` terminal-value convention. This differs from imputing an infrastructure-interrupted trajectory.

Distinct yield counts newly acquired canonical identities above the task's fixed archive 90th-percentile threshold. Its fixed-attempt endpoint uses the observed prefix through 32,768 attempts, or the final set after an earlier 10,000-new-label stop. Empty proposals retain their charged budget and attempt positions but contribute no molecule. This explains why eligibility for distinct yield, canonical PMO quality, and structural descriptors can differ.

Common-label comparisons use each arm's first n charged identities and require both arms to reach n. Paired summaries retain the prescribed seeds; partial support is explicitly identified. Full-suite averages weight task means equally. The original three-task intervals use complete five-seed pairs, and reported intervals are preserved without new resampling. Behavior log probabilities, recurrent resets, ledger order, and charged-identity sequences are checked against saved records; synthetic checks cover counting, fixed-slot area, and eligibility.

B Derivations Supporting the Method

We derive the marginal-set identity, the local guidance condition, and the fixed-batch gradient comparison used in Section 3. The final subsection makes the signed effect of identity weighting explicit for the evaluated clipped objective.

B.1 Marginal utility, history, and stopping

Assume a deterministic reward for each valid canonical identity and a utility $v(R) \in [0, 1]$. Invalid outputs contribute zero. For an identity set $S \supseteq \mathcal{C}_0$ containing the initial archive, set $F(S) = \sum_{c \in S \setminus \mathcal{C}_0} v(R(c))$. For any valid molecule X,

$$F(S \cup \{c(X)\}) - F(S) = \mathbf{1}\{c(X) \notin S\} v(R(X)). \tag{16}$$

Proof. If the identity already belongs to S, the set is unchanged. Otherwise exactly one utility term is added. Summing successive increments telescopes to the final set utility. Conditional on a fixed generation history containing S,

$$\mathbb{E}[v(R(X))] - \mathbb{E}[\Delta(X; S)] = \mathbb{E}[\mathbf{1}\{c(X) \in S\} v(R(X))] \leq \Pr(c(X) \in S). \tag{17}$$

Thus the expected utility assigned to seen identities is precisely the gap between property utility and marginal discovery utility.

For attempt cap T and new-label cap B, let $\sigma = \min(T, \inf\{t : N_t = B\})$ and evaluate $J_{T,B} = \mathbb{E}[F(S_\sigma)]$. After stopping, the set stays fixed and increments are zero. Equal-attempt evaluation therefore carries forward an actually completed discovery set; it does not simulate additional generation. The historical-archive parsing rule determines the operational $\mathcal{C}_0$ in the exploratory panel.

Conditioning on the event $c(X) \notin S$ proves Eq. 10; when its probability is zero the product is defined as zero. This is the expected marginal utility of the current molecule. A longer-horizon value also includes future discoveries and depends on the evolving policy, optimizer state, and remaining budgets. A local change in this conditional expectation alone gives no general guarantee about the entire learning trajectory.

B.2 Clipped guidance: variational form and local sensitivity

At a fixed prefix, write $p_m(a) = \mathrm{softmax}(\ell_\theta + m)_a$, including the implementation's finite sampling-mask offsets. Let z_a be the fixed normalized guide score and $b_\beta(a) = \mathrm{clip}(\beta z_a, -b_{\max}, b_{\max})$. Equation 6 becomes

$$q_\beta(a) = p_m(a) e^{b_\beta(a)} / Z_\beta, \qquad Z_\beta = \sum_a p_m(a) e^{b_\beta(a)}. \tag{18}$$

For any distribution q on this support,

$$\mathbb{E}_q[b_\beta] - \mathrm{KL}(q \| p_m) = \log Z_\beta - \mathrm{KL}(q \| q_\beta). \tag{19}$$

Nonnegativity of KL identifies q_β as the unique maximizer of the local bonus-minus-KL objective. The bonus is the clipped guide score, rather than an exact discovery-value function.

Away from clipping boundaries, $\partial_\beta b_\beta(a) = h_\beta(a) = z_a \mathbf{1}\{|\beta z_a| < b_{\max}\}$, so

$$\partial_\beta \log q_\beta(a) = h_\beta(a) - \mathbb{E}_{q_\beta}[h_\beta]. \tag{20}$$

Holding continuation values $d_a = D^\pi(s, a; S)$ fixed, multiplying by $q_\beta(a) d_a$ and summing proves Eq. 11. At clipping boundaries the corresponding one-sided derivatives apply.

The continuation values are held fixed in this derivative. Changing guidance throughout a sequence or updating the actor changes additional terms; Eq. 11 describes the local decision specified in the main text.

B.3 IDENTITY WEIGHTING AND MATCHED REDUCTIONS

Fix a batch of n trajectories, its parameters, detached advantages and repeat indicators. Let g_i denote the sum of active-token direct-policy ascent gradients for trajectory i, before global clipping and optimizer transformation, and let $Z_{\text{base}} > 0$ be the base optimizer's reduction normalizer.

$$G_{\text{full}} = Z_{\text{base}}^{-1} \sum_i g_i, \qquad G_{\text{id}} = Z_{\text{base}}^{-1} \sum_i (1 - d_i) g_i, \qquad G_{\text{repeat}} = Z_{\text{base}}^{-1} \sum_i d_i g_i. \tag{21}$$

The decomposition $G_{\text{full}} = G_{\text{id}} + G_{\text{repeat}}$ is exact. With $\rho = n^{-1} \sum_i d_i$,

$$G_{\text{id}} - (1 - \rho) G_{\text{full}} = Z_{\text{base}}^{-1} \sum_i (\rho - d_i) g_i \tag{22}$$

$$= -\frac{n}{Z_{\text{base}}} \operatorname{Cov}_{\text{emp}}(d_i, g_i). \tag{23}$$

Here $\operatorname{Cov}_{\text{emp}}(d, g) = n^{-1} \sum_i (d_i - \rho) g_i$ is a vector covariance. A uniformly random mask that removes exactly $n\rho$ trajectories retains each trajectory with probability $1 - \rho$, giving $\mathbb{E}[G_{\text{random}} \mid \text{batch}] = (1 - \rho) G_{\text{full}}$. This expectation matches the raw direct gradient, not its variance or the result of nonlinear gradient clipping and Adam.

These identities compare raw direct-policy gradients on the same batch. Auxiliary losses, full-batch advantage normalization, global clipping, and Adam remain part of the implementation. Moreover, an on-policy discovery gradient can contain a baseline contribution even for an outcome with zero marginal utility. Zeroing a repeat's direct PPO term therefore differs from assigning it zero discovery reward; the identities do not identify that weight as an unbiased discovery-gradient estimator.

B.4 SIGNED DIRECT-POLICY EFFECTS UNDER CLIPPING

This calculation resolves the signed direct-policy terms affected by the evaluated identity weight. Let A_{it} be the detached native advantage at an active token, and write $r_{it} = \exp(\ell_{it} - \ell_{it}^{\text{old}}) > 0$, where ℓ_{it} is the numerator action log probability and ℓ_{it}^{old} is the stored behavior log probability. In the current implementation, these come from the base actor and guided behavior, respectively (Eq. 14). For clipping radius $0 < \epsilon < 1$, the ascent-form direct surrogate is

$$s(r, A) = \min\{rA, \operatorname{clip}(r, 1 - \epsilon, 1 + \epsilon) A\}. \tag{24}$$

Hold the advantages, stored probabilities, repeat tags, active masks, and base normalizer fixed. Away from the sign-relevant clipping boundary, differentiation with respect to the numerator log probability gives

$$\kappa(r, A) = \frac{\partial s}{\partial \ell} = \begin{cases} rA, & A > 0 \text{ and } r < 1 + \epsilon, \\ rA, & A < 0 \text{ and } r > 1 - \epsilon, \\ 0, & \text{otherwise.} \end{cases} \tag{25}$$

Derivation. For $A > 0$, $s(r, A) = A \min(r, 1 + \epsilon)$; for $A < 0$, $s(r, A) = A \max(r, 1 - \epsilon)$. Apply $\partial r / \partial \ell = r$. The sign reversal inside the minimum explains why the two advantages have different clipping boundaries. When $A = 0$, the objective is identically zero. At $r = 1 + \epsilon$ for positive advantage, or $r = 1 - \epsilon$ for negative advantage, the one-sided derivatives differ; an implementation's automatic-differentiation convention must be inspected rather than assigning a universal derivative at the kink.

For each smooth token, its unmasked direct contribution is $u_{it} \kappa(r_{it}, A_{it}) \nabla_\theta \ell_{it} / Z_{\text{base}}$. Consequently, identity weighting changes the direct gradient by

$$G_{\text{id}} - G_{\text{full}} = -\frac{1}{Z_{\text{base}}} \sum_{i,t} d_i u_{it} \kappa(r_{it}, A_{it}) \nabla_\theta \ell_{it}. \tag{26}$$

This refines the earlier repeat-gradient decomposition by specifying which signed terms are active after PPO clipping. It is still a statement before global gradient clipping and the optimizer transformation.

Three distinct effects of removing a repeat. For a repeated active token with $r = 1$ and $A = +1$, the unmasked scalar coefficient is $+1/Z_{\text{base}}$ and the weighted coefficient is zero: the positive direct signal is removed. With $A = -1$ at the same ratio, the coefficient changes from $-1/Z_{\text{base}}$ to zero: a negative direct signal is removed instead. With $A > 0$ and $r > 1 + \epsilon$, or $A < 0$ and $r < 1 - \epsilon$, the derivative was already zero. These cases distinguish the sign of the removed direct signal without tying the analysis to a particular tensor size. A positive cached property reward does not determine the sign of a normalized advantage.

A negative active coefficient discourages the numerator action likelihood when considered in isolation. Removing it is therefore not equivalent to applying a negative likelihood penalty to the repeated molecule. Nor does a coefficient alone determine the full parameter update: token gradients can reinforce or cancel, parameters are shared, and other losses, global clipping and Adam remain. The numerator distribution also need not be the sampling distribution in the current implementation. The complete online effect is determined by the resulting vector update together with the remaining losses and optimizer state.

The scalar derivative was checked against centered finite differences at 100 smooth combinations of ratio, advantage, and clipping radius, with maximum absolute discrepancy 4.94×10^{-10}. Padding, clipping boundaries, zero advantage, and signed aggregation were checked separately.

C Task-Wide Results and Matched Support

C.1 Fixed-attempt discovery coverage

Table 1 contains all 23 tasks and five paired seeds for each complete optimizer. Each endpoint uses the observed prefix through 32,768 attempts or the final set after an earlier 10,000-label stop. The task-level AVID-minus-comparator means follow directly from that table, so they are not tabulated again here. The component and structural-descriptor comparisons below retain their own matched support.

C.2 Canonical PMO reward quality

Table 5 reports canonical PMO Top-10 AUC (Gao et al., 2022). Augmented Memory exceeds AVID on the complete 23-task sum, complementing the distinct-yield comparison in the main text. Standard PPO has no complete aggregate: two runs contain charged empty proposals that fail exact PMO-buffer admission. Their budget and attempt positions remain unchanged in the distinct-yield accounting (Appendix A).

C.3 Early quality across all 23 tasks

Matched four-condition early-quality summary. Table 2, Panel B, compares quality on the identical 113 task–seed blocks in which all four conditions attained 512 newly scored identities. Two Q-Steer-only runs did not attain this endpoint: $C_7H_8N_2O_2$ (seed 505) reached 470 new scores after 10,272 generation attempts, with 9,785 cache hits and 17 invalid outputs; Median molecules 1 (seed 202) reached 478 new scores after 10,272 attempts, with 9,788 cache hits and six invalid outputs. Their 512-label quality is undefined, not zero. These two task–seed blocks are excluded from the quality summary of *every* condition, not just Q-Steer-only; their original outcomes remain in the attainment counts.

For each condition, we first average the attained seed values within each task and then average the 23 task means equally. Thus the two affected tasks contribute four matched seeds each and the other 21 tasks contribute five each. This is a conditional quality comparison among jointly attained endpoints, not an estimate over all planned runs. Bold values in Panel B mark the largest descriptive mean, not statistical significance. The complete 115-run Base, Mask-only, and AVID comparisons elsewhere in this appendix and Section 5.2 retain their original five-seed support; their means therefore need not equal the matched-subset values in Panel B. The complete 23-task early panel crosses Q-Steer and repeat-objective masking on 23 PMO tasks, five paired online seeds, and 512 newly scored identities. AVID enables both components; its comparison to masking alone isolates Q-Steer within the combined method. Table 7 reports every task mean contributing to the task-macro result.

Table 5: **Canonical PMO Top-10 AUC.** Numeric cells are five-seed mean ± sample standard deviation. Italic $n/5$ entries denote incomplete seed coverage and are not converted into available-seed means. Prespecified scientific terminations use the official PMO `finish=True` convention. The 23-task sum is reported only when the complete task set is available for that method.

Task	Standard PPO	Standard REINVENT	Augmented Memory	AVID
Albuterol similarity	0.9610 ± 0.0053	0.7935 ± 0.0129	0.9201 ± 0.0149	0.9658 ± 0.0093
Amlodipine MPO	0.6456 ± 0.0373	0.5584 ± 0.0058	0.7049 ± 0.0577	0.5638 ± 0.0060
Celecoxib rediscovery	0.8594 ± 0.0045	0.7097 ± 0.0216	0.7951 ± 0.0059	0.8683 ± 0.0035
Deco hop	0.6483 ± 0.0212	0.6098 ± 0.0013	0.6765 ± 0.0044	0.6234 ± 0.0036
DRD2	0.9840 ± 0.0038	0.9708 ± 0.0054	0.9714 ± 0.0037	0.9942 ± 0.0003
Fexofenadine MPO	0.8396 ± 0.0421	0.7277 ± 0.0042	0.8746 ± 0.0113	0.7425 ± 0.0183
GSK3B	0.9120 ± 0.0111	0.7981 ± 0.0167	0.8706 ± 0.0022	0.9116 ± 0.0092
Isomers $C_9H_{10}N_2O_2PF_2Cl$	*4/5*	0.6976 ± 0.0139	0.8355 ± 0.0200	0.7882 ± 0.0116
Isomers $C_7H_8N_2O_2$	*3/5*	0.8575 ± 0.0261	0.9618 ± 0.0047	0.9243 ± 0.0212
JNK3	0.6728 ± 0.0776	0.5403 ± 0.1186	0.8006 ± 0.0382	0.5746 ± 0.0187
Median molecules 1	*3/5*	0.2780 ± 0.0135	0.3513 ± 0.0089	0.3643 ± 0.0130
Median molecules 2	0.3140 ± 0.0133	0.2388 ± 0.0121	0.2976 ± 0.0054	0.3246 ± 0.0215
Mestranol similarity	0.9042 ± 0.0689	0.7770 ± 0.0190	0.7809 ± 0.0589	0.9459 ± 0.0132
Osimertinib MPO	0.8262 ± 0.0096	0.8064 ± 0.0033	0.8654 ± 0.0050	0.8179 ± 0.0048
Perindopril MPO	0.5824 ± 0.0371	0.4881 ± 0.0038	0.6052 ± 0.0140	0.5261 ± 0.0127
QED	0.9413 ± 0.0004	0.9381 ± 0.0008	0.9415 ± 0.0006	0.9411 ± 0.0006
Ranolazine MPO	0.8091 ± 0.0127	0.7496 ± 0.0026	0.8094 ± 0.0070	0.7549 ± 0.0100
Scaffold hop	0.8374 ± 0.0415	0.7929 ± 0.0025	0.8478 ± 0.0026	0.8003 ± 0.0028
Sitagliptin MPO	0.4237 ± 0.0646	0.3280 ± 0.0117	0.5689 ± 0.0470	0.3842 ± 0.0286
Thiothixene rediscovery	0.6976 ± 0.1237	0.4259 ± 0.0103	0.5797 ± 0.0095	0.5410 ± 0.0142
Troglitazone rediscovery	0.7639 ± 0.0400	0.3617 ± 0.0280	0.5405 ± 0.0289	0.6050 ± 0.0486
Valsartan SMARTS	0.0275 ± 0.0007	0.0218 ± 0.0012	0.0273 ± 0.0011	0.0268 ± 0.0010
Zaleplon MPO	0.5152 ± 0.0269	0.4527 ± 0.0092	0.5611 ± 0.0196	0.4641 ± 0.0084
23-task sum	*pending*	13.9226	16.1875	15.4526

All 115 primary pairs complete. Across individual pairs, 114/115 mean-reward and 108/115 Top-10 differences are positive. The sole nonpositive mean-reward difference is -0.000042 on valsartan SMARTS. The mean additional attempts required to acquire 512 labels is 14.2, with median zero; this attempt difference is reported rather than interpreted as extra oracle use.

The complete four-condition cells contain 113 task–seed blocks because two guide-only runs reach the frozen exploratory attempt ceiling. Pooling those blocks with equal block weights, rather than the equal task weights of Table 2, the guide's mean-reward effect is $+0.09125$ without masking and $+0.09127$ with masking; their difference-in-differences is $+0.00002$. The corresponding Top-10 interaction is -0.00150. These values localize the early effect to guidance and do not imply that masking should help before repetition accumulates.

This expansion uses the same fixed reward cache and identity lookup within every paired block. A separate parser audit finds 29–46 archive strings per 2,048-row task archive that the current RDKit build cannot canonicalize; those rows retain their reward metadata but cannot enter canonical repeat lookup. Therefore this panel supports a paired Q-Steer effect under the implemented lookup, not an exact claim that every historical archive row participates in the identity ledger. The longer three-task causal panel instead uses the verified target-matched 2,000-identity archives described in Appendix A.

C.4 Guidance increments at the fixed attempt budget

The AVID–Mask-only comparison has 91 matched task–seed pairs. Table 8 reports every task and its supported seed count, including the four negative available-task means. This is a component comparison, separate from the complete optimizer comparison in Table 1.

Table 6: **Complete 23-task early comparison: Top-10 reward at 512 new labels.** Five-seed mean ± sample standard deviation.

Task	Base	Q-Steer alone	Mask only	AVID
albuterol similarity	0.4803 ± 0.0769	0.7298 ± 0.0772	0.4822 ± 0.0809	0.7211 ± 0.0730
amlodipine mpo	0.5006 ± 0.0107	0.5160 ± 0.0144	0.5006 ± 0.0107	0.5160 ± 0.0144
$C_7H_8N_2O_2$	0.3928 ± 0.1175	— (4/5)	0.3928 ± 0.1175	0.6010 ± 0.0674
celecoxib rediscovery	0.3609 ± 0.0257	0.7344 ± 0.0637	0.3609 ± 0.0257	0.7354 ± 0.0415
deco hop	0.5892 ± 0.0068	0.6089 ± 0.0022	0.5892 ± 0.0068	0.6089 ± 0.0022
drd2	0.9662 ± 0.0244	0.9995 ± 0.0006	0.9662 ± 0.0244	0.9992 ± 0.0009
fexofenadine mpo	0.6668 ± 0.0089	0.6694 ± 0.0033	0.6668 ± 0.0089	0.6694 ± 0.0033
gsk3b	0.5476 ± 0.0472	0.7684 ± 0.0208	0.5476 ± 0.0472	0.7684 ± 0.0208
$C_9H_{10}N_2O_2PF_2Cl$	0.4848 ± 0.0357	0.5942 ± 0.0214	0.4869 ± 0.0434	0.5893 ± 0.0290
jnk3	0.2564 ± 0.0325	0.3414 ± 0.0245	0.2564 ± 0.0325	0.3420 ± 0.0248
median molecules 1	0.1946 ± 0.0166	— (4/5)	0.1946 ± 0.0166	0.2895 ± 0.0194
median molecules 2	0.1897 ± 0.0080	0.2164 ± 0.0067	0.1900 ± 0.0086	0.2164 ± 0.0067
mestranol similarity	0.4047 ± 0.0376	0.5563 ± 0.0260	0.4047 ± 0.0376	0.5711 ± 0.0274
osimertinib mpo	0.7770 ± 0.0057	0.7837 ± 0.0048	0.7780 ± 0.0048	0.7837 ± 0.0048
perindopril mpo	0.4258 ± 0.0176	0.4579 ± 0.0085	0.4258 ± 0.0176	0.4579 ± 0.0085
qed	0.9284 ± 0.0048	0.9359 ± 0.0034	0.9288 ± 0.0041	0.9359 ± 0.0033
ranolazine mpo	0.6780 ± 0.0103	0.6942 ± 0.0101	0.6780 ± 0.0103	0.6937 ± 0.0096
scaffold hop	0.7539 ± 0.0073	0.7753 ± 0.0045	0.7539 ± 0.0073	0.7759 ± 0.0039
sitagliptin mpo	0.2036 ± 0.0322	0.2440 ± 0.0190	0.1983 ± 0.0345	0.2440 ± 0.0190
thiothixene rediscovery	0.3234 ± 0.0187	0.4043 ± 0.0106	0.3234 ± 0.0187	0.4065 ± 0.0124
troglitazone rediscovery	0.2483 ± 0.0198	0.3044 ± 0.0291	0.2483 ± 0.0198	0.3034 ± 0.0295
valsartan smarts	0.0116 ± 0.0022	0.0172 ± 0.0032	0.0116 ± 0.0022	0.0172 ± 0.0032
zaleplon mpo	0.3903 ± 0.0109	0.3963 ± 0.0148	0.3893 ± 0.0095	0.3973 ± 0.0144
Task macro	0.4685	—	0.4684	0.5497
Budget-complete runs	115/115	113/115	115/115	115/115

Table 7: All 23 task-level Q-Steer effects at 512 newly scored molecules. Deltas are guide-on minus guide-off with repeat masking in both arms; each row averages five paired seeds.

Task	ΔMean reward	ΔTop-10	Positive mean seeds
Albuterol similarity	+0.103970	+0.238967	5/5
Amlodipine MPO	+0.067800	+0.015453	5/5
Celecoxib rediscovery	+0.134658	+0.374495	5/5
Deco hop	+0.026212	+0.019679	5/5
DRD2	+0.265796	+0.032958	5/5
Fexofenadine MPO	+0.067963	+0.002560	5/5
GSK3β	+0.183449	+0.220800	5/5
Isomers $C_9H_{10}N_2O_2PF_2Cl$	+0.070447	+0.102388	5/5
Isomers $C_7H_8N_2O_2$	+0.048427	+0.208140	5/5
JNK3	+0.053797	+0.085600	5/5
Median molecules 1	+0.084155	+0.094912	5/5
Median molecules 2	+0.020711	+0.026385	5/5
Mestranol similarity	+0.086584	+0.166417	5/5
Osimertinib MPO	+0.258323	+0.005788	5/5
Perindopril MPO	+0.073601	+0.032076	5/5
QED	+0.073262	+0.007094	5/5
Ranolazine MPO	+0.174715	+0.015650	5/5
Scaffold hop	+0.152778	+0.022018	5/5
Sitagliptin MPO	+0.004452	+0.045780	5/5
Thiothixene rediscovery	+0.069267	+0.083143	5/5
Troglitazone rediscovery	+0.037603	+0.055048	5/5
Valsartan SMARTS	+0.000201	+0.005640	4/5
Zaleplon MPO	+0.028677	+0.007966	5/5

Table 8: Task-level AVID-minus-Mask-only differences in high-scoring distinct yield at the fixed attempt budget. Each mean uses the matched seeds counted in the adjacent $n/5$ column. Positive values favor AVID; negative values favor Mask-only. All 23 tasks are listed, with missing entries left unestimated. This partially covered component contrast is reported task by task, without a full-suite aggregate.

Task	Paired seeds	ΔH_{32768} (AVID–Mask)
Albuterol similarity	3/5	$-1,715.7$
Amlodipine MPO	1/5	$+803.0$
Celecoxib rediscovery	0/5	–
Deco hop	4/5	$+5,239.0$
DRD2	0/5	–
Fexofenadine MPO	4/5	$+539.75$
GSK3B	5/5	$+2,078.4$
Isomers $C_9H_{10}N_2O_2PF_2Cl$	4/5	$-1,758.25$
Isomers $C_7H_8N_2O_2$	0/5	–
JNK3	5/5	$+4,630.0$
Median molecules 1	5/5	$+238.4$
Median molecules 2	5/5	$+1,409.0$
Mestranol similarity	5/5	$+376.4$
Osimertinib MPO	5/5	$+1,572.2$
Perindopril MPO	5/5	$+1,504.0$
QED	5/5	$+1,271.0$
Ranolazine MPO	5/5	$+1,272.0$
Scaffold hop	5/5	$+1,886.4$
Sitagliptin MPO	5/5	-935.2
Thiothixene rediscovery	5/5	$+1,113.4$
Troglitazone rediscovery	5/5	$-1,279.0$
Valsartan SMARTS	5/5	$+251.0$
Zaleplon MPO	5/5	$+1,683.6$

D Component Behavior and Robustness

D.1 All primary paired seed differences

Table 9 contains every primary combination-minus-mask effect, without excluding a seed or replacing the primary attempt checkpoint by an equal-label endpoint. Five seeds vary online optimization conditional on a fixed task archive and guide. Their identifiers coincide with those of the earlier panel, but each factorial arm is a fresh process and the panels are not pooled.

Table 9: All primary high-score-discovery differences at 32,768 attempts, subject to the 10k-label ceiling.

Task	Seed 101	Seed 202	Seed 303	Seed 404	Seed 505
Albuterol	-3659	-1234	-1970	+41	-1716
Zaleplon	+1962	+1954	+2194	+2485	+2763
DRD2	+842	+691	+599	+605	+370

D.2 Realized discovery and generation cost

In Table 10, `q0`/`q1` denotes guidance off/on and `m0`/`m1` denotes masking off/on. Thus `q1_m1` is AVID, `q1_m0` is Q-Steer alone, `q0_m1` is masking alone, and `q0_m0` is base. The source data retain these condition identifiers; "combination" denotes AVID. New identities at the primary checkpoint and attempts at the actual terminal measure different points on a trajectory. Runs that reach the attempt limit before the label budget remain included.

Table 10: Complete factorial conditions: budget attainment, primary-checkpoint new identities, and actual terminal attempts. Counts other than completion are five-seed means.

Task	Condition	Completion	New at primary cap	Terminal attempts
Albuterol	q0_m0	0/5	2499.0	200032.0
Albuterol	q0_m1	5/5	9175.2	34048.0
Albuterol	q1_m0	0/5	994.2	200032.0
Albuterol	q1_m1	5/5	6704.8	46905.6
Zaleplon	q0_m0	0/5	6447.6	200032.0
Zaleplon	q0_m1	5/5	10000.0	10489.6
Zaleplon	q1_m0	0/5	1936.6	200032.0
Zaleplon	q1_m1	5/5	10000.0	22995.2
DRD2	q0_m0	5/5	10000.0	13580.8
DRD2	q0_m1	5/5	10000.0	11123.2
DRD2	q1_m0	2/5	4044.8	152544.0
DRD2	q1_m1	5/5	10000.0	15769.6

D.3 Common labels, intervals, and missingness

Common-label analysis uses each arm's first n charged identities. All 15 combination/mask pairs attain 10k; other contrasts may not. Each endpoint uses its eligible pairs, without mixing unattained label budgets into an attained-budget estimate.

A budget-capped process is carried forward at attempt checkpoints only after the label ceiling is reached. No infrastructure-truncated trajectory is treated as a stopped successful budget. This capped comparison does not force all arms to expend the same number of oracle labels. At equal 10k labels, the high-score differences are $+758.4$ $[623.4, 871.6]$ on albuterol, $+2,271.6$ $[2,005.2, 2,547.2]$ on zaleplon, and $+621.4$ $[480.0, 751.4]$ on DRD2. For the latter two tasks, equal-label and primary per-seed differences coincide; their separately generated original Monte Carlo bootstrap endpoints can differ slightly. We preserve both frozen outputs rather than resampling for display consistency.

The four conditions use a prespecified cyclic execution order, and the guide-off arm retains the guide-evaluation path. Architecture, loss reduction, and update accounting follow Appendix A. Each arm starts in a fresh process; these runs are not pooled with the earlier paired study.

D.4 ACQUISITION–QUALIFICATION DECOMPOSITION

The following descriptive analysis uses the completed three-task factorial's existing five-seed records. It introduces no oracle calls, new thresholds, or significance tests. At the original 32,768-attempt checkpoint under the 10k-label ceiling, write $H = Nf$, where f is the high-score fraction among newly discovered identities. For each seed, AVID (A) and masking alone (M) satisfy

$$H_A - H_M = (N_A - N_M)\frac{f_A + f_M}{2} + (f_A - f_M)\frac{N_A + N_M}{2}. \tag{27}$$

Expanding the products proves the identity. Table 11 averages the per-seed terms; it does not multiply separately averaged quantities. The two terms provide arithmetic attribution, not causal mediation.

Table 11: **Distinct yield and high-score fraction explain different regimes.** AVID minus masking alone at the original primary checkpoint. All entries average five paired seeds.

Task	ΔH	Identity-count term	High-score-fraction term
Albuterol	-1707.6	-2295.9	$+588.3$
Zaleplon	$+2271.6$	0.0	$+2271.6$
DRD2	$+621.4$	0.0	$+621.4$

On albuterol, the mean per-seed high-score fraction rises from 88.96% to 96.34%, while mean new-identity count falls from 9175.2 to 6704.8. On zaleplon and DRD2, both masked conditions already reach 10k identities before the checkpoint, so the differences arise entirely from the fraction qualifying. These fractions do not replace the separately reported mean reward and Top-10 measures.

Table 12: New identities per generation attempt over the original active windows. Entries are **Q-Steer alone / AVID**. A window requires both arms to remain observed through its end; no active continuation is inferred after reaching the label cap.

Task	Attempts 1–2048	Attempts 2049–8192	Attempts 8193–32768
Albuterol	$0.41963/0.58711$	$0.00843/0.19089$	$0.00338/0.17617$
Zaleplon	$0.79414/0.90957$	$0.03828/0.54935$	unavailable
DRD2	$0.74336/0.85273$	$0.09111/0.59368$	unavailable

Every reported active-window comparison uses all five paired seeds. Zaleplon and DRD2 have no eligible pairs in the last window because AVID has already completed its label budget. These observed trajectories quantify acquisition dynamics; they do not directly estimate the counterfactual action values in Eq. 11.

D.5 QUALITY AND ACQUISITION TRAJECTORIES

The main-text trajectory figure uses a compact two-task layout. Figure 3 retains all three original tasks and all 30 runs, without changing the data or the seed aggregation. In particular, Zaleplon retains the lower AVID Top-10 trajectory alongside its higher distinct-candidate yield relative to Augmented Memory. This comparison distinguishes reward quality from candidate acquisition; the illustrative tasks are not a substitute for the full benchmark table.

D.6 TARGETED QUALITY AND REDUCTION CONTROLS

The following diagnostics retain their fixed earlier matched subset rather than substituting later benchmark coverage. Table 13 shows why eligibility must be assessed separately for each endpoint: Amlodipine has five discovery pairs but only four Mean@10k pairs, and C7 has two discovery pairs and three area pairs. The interrupted C7 seed101 run had already reached the reported attempt and 1,500-label checkpoints; no unattained 10k outcome is inferred. The active-token control in Table 14 retains guidance and identity weighting but changes the direct-loss denominator. AVID exceeds the internal base in all five Amlodipine pairs and falls below this reduction variant in all five. Thus the complete-method comparison does not establish the frozen reduction as optimal or attribute its benefit uniquely to identity selection. Changes in direct-loss scale can also change its balance with

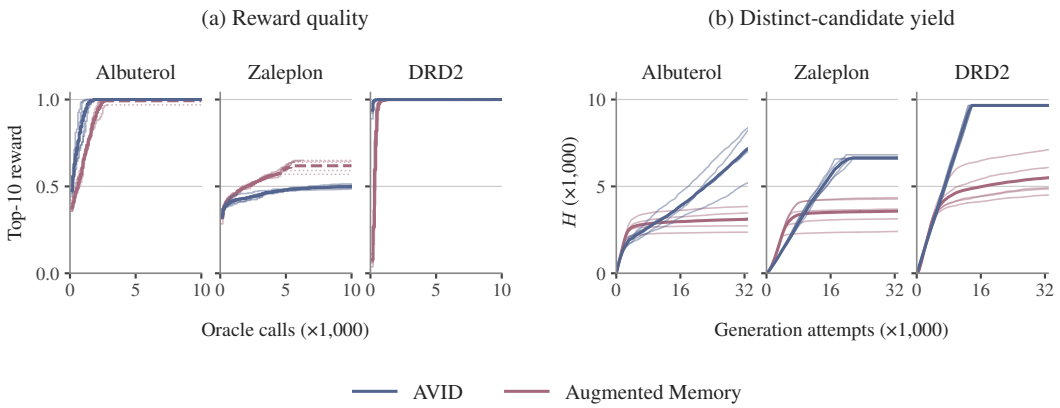

Figure 3: **Full three-task quality and acquisition trajectories.** Albuterol, Zaleplon, and DRD2, with five matched seeds for AVID and Augmented Memory. Thick lines show means; thin lines show every seed. **(a)** Best-ten reward versus oracle calls; dashed extensions indicate terminal-value carry under PMO's `finish=True` convention, not additional evaluations. **(b)** H_t is the cumulative number of newly scored, distinct molecules whose reward meets the task's fixed archive 90th-percentile threshold by generation attempt t, shown in thousands. A run reaching its 10,000-new-label cap retains its observed stopped set. No new discoveries are interpolated or extrapolated.

Table 13: **Additional standard-PPO comparisons, with support per endpoint.** Each difference is AVID minus standard PPO, averaged within the listed task and seed set. The target seeds are 101/202/303/404/505. C7 denotes the C7H8N2O2 isomer task. A_{10} is the fixed-ten label-axis area, not endpoint Top-10. Discovery uses 32,768 attempts and a 10k-label ceiling; quality uses the attained label endpoint. These are descriptive task results, with no confidence interval or full23 aggregate.

Task	Readout	Eligible seeds	PPO	AVID	Difference
Amlodipine	H_{32768}	All five	4,444.6	6,605.6	+2,161.0
	$A_{10}(1500)$	All five	0.5012	0.5140	+0.0127
	Mean@10k	202/303/404/505	0.5240	0.4161	-0.1079
C7	H_{32768}	101/202	566.0	2,749.0	+2,183.0
	$A_{10}(1500)$	101/202/303	0.6099	0.6023	-0.0076

Table 14: **Five-seed Amlodipine reduction comparison.** Each entry is a paired difference in H_{32768}, AVID minus the named comparator, from the same new A100 campaign. Both rows reuse the same five AVID runs; they are not independent replications of AVID. Active-token retains the identity mask and changes the direct-objective denominator. All seeds and both directions are retained; no statistical significance is claimed.

Comparator	101	202	303	404	505	Mean
Internal base	+1,121	+1,864	+1,708	+1,314	+757	+1,352.8
Active-token	-1,914	-1,290	-1,531	-1,373	-1,576	-1,536.8

auxiliary terms and subsequent sampling. Other Amlodipine Mask-only/random-count/retained-token comparisons in this subset contain only seed404, while uniform weighting and combined-method transfer have no qualified pairs in the extraction. They do not provide additional five-seed operator comparisons.

E CANDIDATE-SET CHARACTERIZATION

Table 15 gives the exact task–seed support behind Table 3. A 0/5 cell is unavailable, not a measured zero. Every available pair uses the same task, seed, archive, threshold, and fixed-attempt endpoint on both sides; 85 PPO, 73 REINVENT, and 40 AM pairs yielded 293 distinct runs because AVID records can be reused across comparator blocks. These are partial external comparisons, whereas the independent 512-label experiment covers all 23 tasks and five seeds.

Table 15: Matched task/seed chemical descriptor support, $n/5$ per complete optimizer. Missing seeds are never set to zero or replaced by a different attempt.

Task	AVID–PPO	AVID–REINVENT	AVID–AM
Albuterol similarity	5/5	2/5	5/5
Amlodipine MPO	5/5	0/5	5/5
Celecoxib rediscovery	1/5	0/5	2/5
Deco hop	5/5	4/5	5/5
DRD2	5/5	0/5	5/5
Fexofenadine MPO	5/5	5/5	5/5
GSK3B	5/5	5/5	5/5
Isomers $C_9H_{10}N_2O_2PF_2Cl$	5/5	4/5	3/5
Isomers $C_7H_8N_2O_2$	3/5	0/5	5/5
JNK3	3/5	5/5	0/5
Median molecules 1	2/5	5/5	0/5
Median molecules 2	5/5	5/5	0/5
Mestranol similarity	5/5	5/5	0/5
Osimertinib MPO	5/5	5/5	0/5
Perindopril MPO	5/5	5/5	0/5
QED	5/5	5/5	0/5
Ranolazine MPO	5/5	5/5	0/5
Scaffold hop	5/5	5/5	0/5
Sitagliptin MPO	4/5	5/5	0/5
Thiothixene rediscovery	2/5	3/5	0/5
Troglitazone rediscovery	0/5	0/5	0/5
Valsartan SMARTS	0/5	0/5	0/5
Zaleplon MPO	0/5	0/5	0/5

The input set is formed from ordered charged labels paired with the corresponding first-seen generation attempts. For a run reaching attempt 32,768, only charged molecules acquired by that attempt enter; if its 10,000-new-label cap is reached earlier, the stopped set enters without extrapolation. Qualification requires a nonblank canonical SMILES and reward at least the task's frozen archive-defined threshold. The reconstructed qualified-set size matches the fixed-attempt count for each included run. Murcko scaffolds are canonical scaffold SMILES; acyclic molecules share an explicit empty-scaffold class. QED is RDKit's 0–1 descriptor; SA is the untransformed RDKit Contrib score, where lower is easier; PAINS A/B/C and BRENK are filter-catalog alerts. Ratios with an empty qualifying set would be undefined, not zero. The present matched pairs all had nonempty qualifying sets. Attempts/H and charged labels/H use the same fixed endpoint and therefore do not import post-checkpoint labels. All descriptors are post-hoc measurements, not additional oracle evaluations or proof of experimental safety.

E.1 SAMPLED FINGERPRINT DIVERSITY ON THE SAME FROZEN SUPPORT

The diversity extension uses exactly the above 85/73/40 matched pairs and 293 authenticated runs, not the broader fixed-attempt comparison support in Table 1. The qualified canonical identities are sorted; runs with more than 512 qualifying molecules use a fixed SHA256-derived per-run seed to select 512 without replacement, while smaller sets are used in full. Of 293 runs, 288 are sampled and five are full-set. RDKit Morgan radius-2, 2,048-bit fingerprints define the mean pairwise $1 - \mathrm{Tanimoto}$ distance. Butina clustering uses Tanimoto distance threshold 0.4; cluster coverage is cluster count/sample size and largest-cluster fraction is largest cluster size/sample size. These are

sample-based structural descriptors, not full-run cluster counts. The three tables retain all 23 tasks, including 0/5 missing support, and report task-level AVID-minus-comparator means over only the paired seeds shown. No molecule-level independence or full-23 confidence interval is inferred.

AVID versus standard PPO (85 pairs).

Task	Paired seeds	Δ diversity ↑	Δ cluster coverage ↑	Δ largest cluster ↓
Albuterol similarity	5/5	−0.067	−0.111	−0.022
Amlodipine MPO	5/5	+0.087	+0.459	−0.109
Celecoxib rediscovery	1/5	−0.063	−0.145	−0.045
Deco hop	5/5	+0.128	+0.480	−0.216
DRD2	5/5	+0.038	+0.163	−0.038
Fexofenadine MPO	5/5	+0.001	+0.196	−0.118
GSK3B	5/5	−0.004	+0.114	−0.271
Isomers $C_9H_{10}N_2O_2PF_2Cl$	5/5	+0.041	+0.212	−0.048
Isomers $C_7H_8N_2O_2$	3/5	+0.024	+0.083	−0.013
JNK3	3/5	−0.006	+0.101	−0.168
Median molecules 1	2/5	−0.069	+0.014	−0.019
Median molecules 2	5/5	−0.068	−0.049	+0.142
Mestranol similarity	5/5	+0.036	+0.077	−0.162
Osimertinib MPO	5/5	+0.015	+0.158	−0.045
Perindopril MPO	5/5	+0.097	+0.379	−0.184
QED	5/5	+0.021	+0.096	−0.013
Ranolazine MPO	5/5	+0.087	+0.388	−0.259
Scaffold hop	5/5	+0.156	+0.494	−0.151
Sitagliptin MPO	4/5	+0.058	+0.140	−0.149
Thiothixene rediscovery	2/5	−0.060	−0.025	+0.007
Troglitazone rediscovery	0/5	–	–	–
Valsartan SMARTS	0/5	–	–	–
Zaleplon MPO	0/5	–	–	–

AVID versus standard REINVENT (73 pairs).

Task	Paired seeds	Δ diversity ↑	Δ cluster coverage ↑	Δ largest cluster ↓
Albuterol similarity	2/5	−0.143	−0.398	+0.057
Amlodipine MPO	0/5	–	–	–
Celecoxib rediscovery	0/5	–	–	–
Deco hop	4/5	−0.054	−0.100	+0.013
DRD2	0/5	–	–	–
Fexofenadine MPO	5/5	−0.096	−0.168	+0.027
GSK3B	5/5	−0.132	−0.305	+0.011
Isomers $C_9H_{10}N_2O_2PF_2Cl$	4/5	−0.043	−0.155	+0.007
Isomers $C_7H_8N_2O_2$	0/5	–	–	–
JNK3	5/5	−0.095	−0.224	+0.014
Median molecules 1	5/5	−0.112	−0.146	+0.014
Median molecules 2	5/5	−0.162	−0.449	+0.263
Mestranol similarity	5/5	−0.024	−0.106	−0.070
Osimertinib MPO	5/5	−0.053	−0.084	+0.016
Perindopril MPO	5/5	−0.083	−0.232	+0.067
QED	5/5	−0.060	−0.180	+0.009
Ranolazine MPO	5/5	−0.101	−0.188	+0.043
Scaffold hop	5/5	+0.002	−0.014	+0.004
Sitagliptin MPO	5/5	−0.050	−0.370	+0.037
Thiothixene rediscovery	3/5	−0.154	−0.413	+0.069
Troglitazone rediscovery	0/5	–	–	–
Valsartan SMARTS	0/5	–	–	–
Zaleplon MPO	0/5	–	–	–

AVID versus Augmented Memory (40 pairs).

Task	Paired seeds	Δ diversity ↑	Δ cluster coverage ↑	Δ largest cluster ↓
Albuterol similarity	5/5	−0.087	−0.235	+0.050
Amlodipine MPO	5/5	+0.081	+0.248	−0.108
Celecoxib rediscovery	2/5	−0.058	−0.216	+0.148
Deco hop	5/5	+0.003	+0.069	−0.029
DRD2	5/5	−0.0003	+0.027	−0.062
Fexofenadine MPO	5/5	−0.012	+0.026	−0.094
GSK3B	5/5	−0.084	−0.164	−0.008
Isomers $C_9H_{10}N_2O_2PF_2Cl$	3/5	+0.022	+0.006	−0.005
Isomers $C_7H_8N_2O_2$	5/5	−0.008	−0.014	+0.001
JNK3	0/5	–	–	–
Median molecules 1	0/5	–	–	–
Median molecules 2	0/5	–	–	–
Mestranol similarity	0/5	–	–	–
Osimertinib MPO	0/5	–	–	–
Perindopril MPO	0/5	–	–	–
QED	0/5	–	–	–
Ranolazine MPO	0/5	–	–	–
Scaffold hop	0/5	–	–	–
Sitagliptin MPO	0/5	–	–	–
Thiothixene rediscovery	0/5	–	–	–
Troglitazone rediscovery	0/5	–	–	–
Valsartan SMARTS	0/5	–	–	–
Zaleplon MPO	0/5	–	–	–

F Relationship to Prior Methods

This section locates AVID's intervention relative to existing methods. The distinctions concern the evaluated sampling and learning operations; literature coverage does not imply that every cited method is an experimental baseline.

Memory, replay, and policy learning. REINFORCE, trust-region optimization, and PPO provide reward-based update rules (Williams, 1992; Schulman et al., 2015; 2017), while experience replay reuses past observations (Lin, 1992). Molecular generators can first learn chemical distributions and then specialize to focused libraries (Segler et al., 2018); molecular RL and ReLeaSE adapt them using property rewards (Olivecrona et al., 2017; Popova et al., 2018). REINVENT and ACEGEN supply implementation frameworks (Blaschke et al., 2020a; Loeffler et al., 2024; Bou et al., 2024). Memory-assisted RL changes rewards through structural memory, whereas Augmented Memory and Saturn study scored-molecule reuse and augmented replay (Blaschke et al., 2020b; Guo & Schwaller, 2024a; Guo et al., 2026). AVID retains cached rewards and changes only the direct policy weight of within-run repeats; it does not discard those trajectories from every learning operation.

Generation-time guidance. PPLM, GeDi, and DExperts introduce attribute gradients or auxiliary model distributions (Dathathri et al., 2020; Krause et al., 2021; Liu et al., 2021). FUDGE predicts future attributes from prefixes; RAD and Transfer Q^* use reward or value signals during decoding (Yang & Klein, 2021; Deng & Raffel, 2023; Chakraborty et al., 2024). NeuroLogic A*esque decoding adds lookahead heuristics (Lu et al., 2022), and Beam Enumeration uses probable molecular substructures for self-conditioned optimization (Guo & Schwaller, 2024b). AVID combines generation-time guidance with a separate post-completion learning weight. This also differs from changing decoding alone (Holtzman et al., 2020) or adding an unlikelihood penalty (Welleck et al., 2020): zeroing a direct term can remove a positive, negative, or already-clipped contribution (Appendix B.4).

Sampled sets and cumulative identities. SGRPO and SetPO assign diversity-aware credit within sampled groups using group-relative or leave-one-out contributions (Ye et al., 2026; Li et al., 2026). AVID's ledger instead records exact identities accumulated across the run. A qualifying new analogue can increase H_t while adding little structural diversity; a diverse batch of already acquired identities adds no new candidates. The implemented weight tests within-run repetition rather than assigning the exact threshold-qualified marginal utility, and archive hits retain weight one. These distinctions identify the objective and operator being compared without treating identity weighting as an exact set-policy-gradient estimator.

Representation and search choices. SMILES and SELFIES define string representations (Weininger, 1988; Krenn et al., 2020); latent, junction-tree, motif, and graph-flow models change how molecules are assembled (Gomez-Bombarelli et al., 2018; Jin et al., 2018; 2020; Shi et al., 2020; Luo et al., 2021). Equivariant diffusion and GeoDiff address molecular geometry (Hoogeboom et al., 2022; Xu et al., 2022). Alternative search procedures include graph policies and molecular editing (You et al., 2018; Zhou et al., 2019), genetic and rationale-based search (Jensen, 2019; Nigam et al., 2020; Chen et al., 2021), Bayesian optimization and weighted retraining (Korovina et al., 2020; Tripp et al., 2020), and reward-directed sampling through GFlowNets or MARS (Bengio et al., 2021; Xie et al., 2021). AVID is evaluated with a SMILES-based sequential policy; exact-identity acquisition is an evaluation axis distinct from these architectural choices.

Complementary evaluation objects. GuacaMol and MOSES assess generation quality and diversity, while PMO standardizes oracle-call efficiency (Brown et al., 2019; Polykovskiy et al., 2020; Gao et al., 2022). QED, synthetic-accessibility scores, fingerprints, scaffolds, and Fréchet ChemNet Distance characterize different chemical or distributional properties (Bickerton et al., 2012; Ertl & Schuffenhauer, 2009; Rogers & Hahn, 2010; Bemis & Murcko, 1996; Preuer et al., 2018). Our budgeted acquisition results are reported alongside reward and structural descriptors, rather than interpreted as prospective biological activity or synthetic success.