PHARMAGENT: CONSTRAINT-AWARE SEARCH WITH FROZEN LANGUAGE MODELS FOR MOLECULAR OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Molecular optimization must improve target activity and satisfy developability constraints within limited evaluation budgets. Classical methods require tailored rules or training to incorporate chemical instructions and property feedback. Frozen language models can condition edits on this information, but need explicit constraint control and relevant experience. We therefore present PHARMAGENT, a constraint-aware molecular search method driven by adaptive external state. Its Lagrangian controller translates violations in accepted states into accumulated constraint pressure, keeping this history separate from current property measurements. Structure-indexed replay complements this feedback with relevant evaluated transitions that guide subsequent proposals. As a curriculum progressively activates constraints, candidates and the incumbent are compared under the same current objective, and the accepted state determines the next multiplier update. We derive an exact identity that characterizes how accepted-state violations accumulate in the controller's multipliers. Across five tasks with five independent runs, PHARMAGENT achieves a summed area under the target-score curves (AUC) of 3.9208 in target-only search, improving over MOLLEO by 37.3%. With online constraints, it achieves a property-adjusted AUC of 0.7076, improving over the strongest online baseline, ExLLM, by 53.8%. These results rank first among all evaluated methods in both target-only and constraint-aware search. The online comparison covers all five baseline frameworks. The full system leads every ablation variant in target quality, property-adjusted performance, and Pareto hypervolume. All five molecular cases reach feasible final states, documenting target gains and trade-offs. These results establish PHARMAGENT as a powerful framework for budget-limited molecular optimization, achieving leading aggregate performance through adaptive external search state and frozen language models.

1 INTRODUCTION

Molecular optimization searches for structures with improved target activity and suitable physicochemical, synthesizability, and absorption, distribution, metabolism, excretion, and toxicity (ADMET) properties (Bickerton et al., 2012; Gao & Coley, 2020). Under limited budgets, each evaluation should guide productive structural edits (Gao et al., 2022). Graph-GA's stochastic mutation and crossover can spend evaluations on unproductive edits (Jensen, 2019; Wang et al., 2025). REINVENT4 adapts its policy through reward-driven parameter updates (Loeffler et al., 2024), while latent-space Bayesian optimization depends on learned representations and valid decoding (Gómez-Bombarelli et al., 2018). These methods require translating chemical instructions and heterogeneous feedback into search rules or training signals (Ran et al., 2025). Large language models (LLMs) can instead condition chemically informed edits directly on task instructions and evaluated experience (Wang et al., 2025; Ran et al., 2025; Krüger et al., 2026). Freezing their parameters enables adaptation through context without retraining (Yang et al., 2024; M. Bran et al., 2024).

However, proposal quality alone leaves two feedback gaps (Figure 1(a)). First, target gains can introduce property violations (Gao & Coley, 2020). Persistent violations call for search priorities that adapt to violation history, beyond fixed penalty weights or post-hoc filtering. Second, multi-round interaction creates experience, but useful guidance requires edits relevant to the current molecular

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.

structure (Shinn et al., 2023; Liu et al., 2025). These gaps motivate dynamic constraint control and structure-aware experience retrieval.

To address both gaps, we propose PHARMAGENT, a constraint-aware molecular search method that adapts external state around frozen language models. Generator and Reviewer execute the proposal, evaluation, acceptance, and feedback loop with scientific tools and medicinal-chemistry retrieval-augmented generation (MedChem-RAG) (Figure 1(b)). Three mechanisms organize external state (Figure 1(c)). A Lagrangian controller accumulates accepted-state violations as historical pressure, separate from current measurements (Bertsekas, 1999). Chemical-aware contextual experience replay (Chemical-CER) retrieves structurally related evaluated edits, and a curriculum introduces constraints after target exploration. Candidates and the incumbent share the current objective, while stored attempts remain distinct from accepted states. Table 1 compares the evaluated configurations.

Table 1: Mechanisms in evaluated configurations (✓ enabled, ✗ not enabled). Baseline rows show original configurations. Online adaptations use fixed-penalty (FP) rewards or feasibility-first (FF) rewards, selection, or conversational guidance. Retrieval and cross-task reuse concern optimization attempts. Appendix D specifies protocols.

Method	Online constraints	Adaptive weights	Structure-based retrieval	Cross-task reuse
Graph-GA / REINVENT4	✗	✗	✗	✗
MOLLEO / ExLLM / SEISMO	✗	✗	✗	✗
Online adaptations	✓	✗	✗	✗
PHARMAGENT	✓	✓	✓	✓

We evaluate target-only search (E1) on practical molecular optimization (PMO) (Gao et al., 2022) and search with online constraints (E2) on ADMET-PMO, followed by ablations and molecular cases. Our contributions are:

- A unified molecular search framework that turns property feedback, structurally relevant transitions, and constraint timing into adaptive external decision state for frozen LLMs.
- An analytical characterization linking multiplier growth to cumulative accepted-state violations, explaining how the controller retains historical constraint pressure.
- The highest aggregate performance across five tasks and five runs, with a 37.3% gain in target-score area under the curve (AUC) in E1 and a 53.8% gain in ADMET-adjusted AUC (AA-AUC) over the strongest online baseline in E2. The full system leads every ablation, and all five cases reach feasibility.

2 RELATED WORK

Molecular generation with property feedback. Multi-property optimization must coordinate conflicting chemical objectives (Bickerton et al., 2012; Gao & Coley, 2020). InversionGNN uses learned structure–property knowledge and Pareto search (Niu et al., 2025). MolGen incorporates chemical feedback into molecular language modeling (Fang et al., 2024). Uncertainty-aware reinforcement learning (RL) guides diffusion models toward multiple molecular properties (Chen et al., 2025). These approaches connect generative mechanisms to scientific evaluations. PHARMAGENT adapts measured search state around a frozen model, with explicit online inequalities and a finite target-oracle budget.

Language-model search and memory. MOLLEO (Wang et al., 2025), ExLLM (Ran et al., 2025), and SEISMO (Krüger et al., 2026) organize search around populations, experience summaries, and interaction trajectories. Contextual experience replay (Liu et al., 2025) reuses agent interactions without parameter updates. ReAct (Yao et al., 2023) and ChemCrow (M. Bran et al., 2024) ground proposals in tool observations. PHARMAGENT integrates numerical constraint history and structurally relevant transitions directly into proposal-and-acceptance decisions.

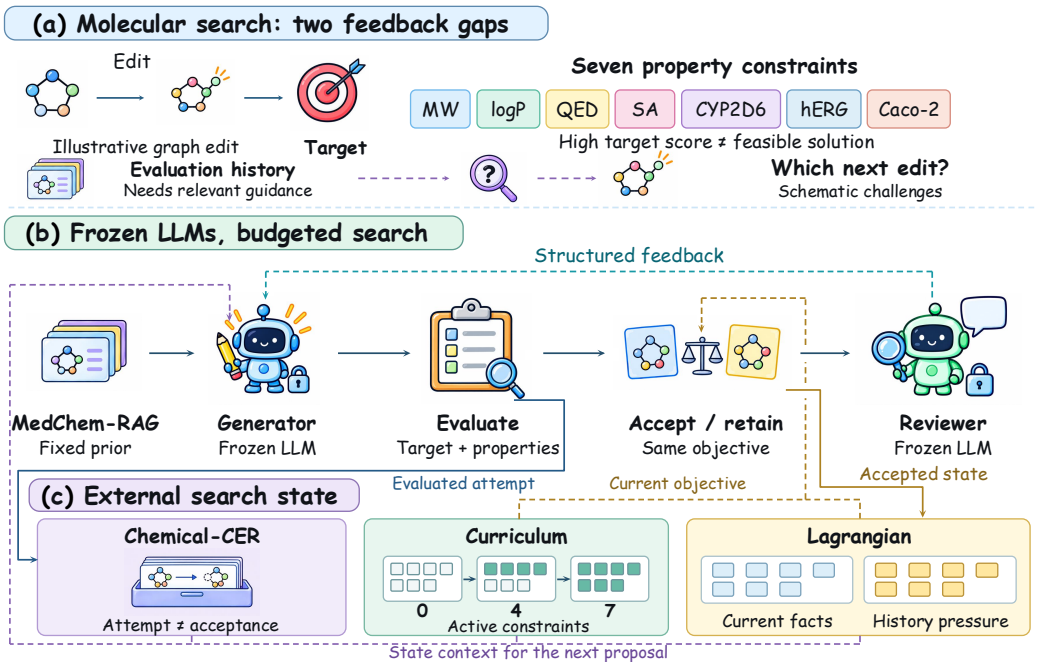

Figure 1: Motivation and the PHARMAGENT feedback loop. (a) Two feedback gaps in molecular search. High target scores need not imply feasibility, and past evaluations require relevant retrieval. (b) Frozen Generator–Reviewer workflow for proposal, evaluation, acceptance, and feedback. Candidates and the incumbent share the current objective. (c) External search state comprising Chemical-CER, curriculum, and Lagrangian control. Evaluated attempts enter replay, while accepted states update multipliers. Current measurements remain separate from historical pressure. Property labels denote molecular weight (MW), the logarithm of the octanol–water partition coefficient (logP), quantitative estimate of drug-likeness (QED), synthetic accessibility (SA), cytochrome P450 2D6 (CYP2D6), the human ether-à-go-go-related gene potassium channel (hERG), and Caco-2 cell permeability. Icons are schematic. Solid routes show flow or updates, whereas dashed routes show conditioning.

Constrained optimization and evaluation. Penalty and multiplier methods are classical (Bertsekas, 1999; Boyd & Vandenberghe, 2004). Our contribution connects these methods to frozen-LLM molecular search through accepted-state updates, shared-objective comparisons, and separate measurement/history feedback. Graph-GA (Jensen, 2019) and REINVENT4 (Loeffler et al., 2024) supply non-LLM controls. PMO measures target progress against oracle calls (Gao et al., 2022). ADMET-PMO adds shared RDKit and ADMET-AI evaluations (Landrum et al., 2024; Swanson et al., 2024), supporting both online constraints and post-hoc annotation.

3 PHARMAGENT: CONSTRAINT-AWARE MOLECULAR SEARCH

3.1 PROBLEM SETUP

Let $\mathcal{X}$ be the space of valid molecular structures, f a black-box target oracle, and g_i property evaluators with thresholds t_i. The budget B counts unique target evaluations, including initialization. Neither the evaluators nor the frozen LLM supply gradients. We seek a high-scoring molecule satisfying K property constraints:

$$\max_{x \in \mathcal{X}} f(x) \quad \text{s.t.} \quad \tilde{g}_i(x) \leq \tilde{t}_i, \quad i = 1, \ldots, K. \tag{1}$$

For an upper bound set $a_i = 1$, and for a lower bound set $a_i = -1$, with $\tilde{g}_i = a_i g_i$ and $\tilde{t}_i = a_i t_i$. Normalized violation is $v_i(x) = \max(0, (\tilde{g}_i(x) - \tilde{t}_i)/r_i)$ for fixed scales $r_i > 0$. Validity defines $\mathcal{X}$. Chemically valid structures remain eligible despite property violations.

3.2 SEARCH FRAMEWORK

The round-t state comprises incumbent $y^{(t)}$, active constraints $\mathcal{C}_t$, multipliers $\boldsymbol{\lambda}^{(t)}$, penalty $\rho^{(t)}$, archive $\mathcal{D}_t$, and Reviewer feedback R_t:

$$\begin{aligned} S_t &= (y^{(t)}, \mathcal{C}_t, \boldsymbol{\lambda}^{(t)}, \rho^{(t)}, \mathcal{D}_t, R_t), \\ \mathcal{P}_t &\sim q_\theta\big(\cdot \mid \mathrm{Encode}(S_t; \mathcal{I})\big), \qquad \theta \text{ fixed.} \end{aligned} \tag{2}$$

$\mathcal{P}_t$ is the proposal batch. $\mathcal{I}$ supplies molecular representations, property definitions, retrieval features, and fixed chemical context. Measurements adapt S_t while model parameters stay fixed.

The Generator proposes up to five molecular strings in the simplified molecular-input line-entry system (SMILES) (Weininger, 1988) per request. RDKit (Landrum et al., 2024) checks validity and novelty after canonicalization. RDKit and ADMET-AI v1.4.0 supply properties (Swanson et al., 2024). MedChem-RAG retrieves ChEMBL (Zdrazil et al., 2024) matched molecular pairs (MMPs) through fingerprint-based Facebook AI similarity search (FAISS) (Griffen et al., 2011; Dossetter et al., 2013; Johnson et al., 2021; Rogers & Hahn, 2010). Only Generator and Reviewer invoke the frozen LLM. The Reviewer returns structured priorities and modification categories. Replay supplies canonical structure pairs and numeric outcomes, excluding historical model prose (Appendix F).

3.3 LAGRANGIAN FEEDBACK AND ACCEPTANCE

The controller turns constraint violations into both a numerical acceptance criterion and feedback for subsequent proposals. Classical multiplier and quadratic-penalty terms (Bertsekas, 1976) define the current objective:

$$\mathcal{L}_{\mathcal{C}_t}(x, \boldsymbol{\lambda}, \rho) = f(x) - \sum_{i \in \mathcal{C}_t} \lambda_i v_i(x) - \frac{\rho}{2} \sum_{i \in \mathcal{C}_t} v_i(x)^2. \tag{3}$$

An empty active set recovers f. For incoming incumbent $y^{(t)}$ and current-objective round-best candidate $c^{(t)}$, acceptance is

$$\bar{x}^{(t)} = \begin{cases} c^{(t)}, & \mathcal{L}_{\mathcal{C}_t}(c^{(t)}, \boldsymbol{\lambda}^{(t)}, \rho^{(t)}) > \mathcal{L}_{\mathcal{C}_t}(y^{(t)}, \boldsymbol{\lambda}^{(t)}, \rho^{(t)}), \\ y^{(t)}, & \text{otherwise.} \end{cases} \tag{4}$$

Both sides use the same active constraints, multipliers, and penalty coefficient. Cached properties rescore the incumbent without another oracle call. This comparison isolates the candidate's improvement under the current objective from changes caused by the controller itself. A score stored before a constraint or multiplier update would conflate these two effects.

After a completed scored round, let $V_{\mathcal{C}_t}^{(t)} = \sum_{i \in \mathcal{C}_t} v_i(\bar{x}^{(t)})$ and update

$$\begin{aligned} \lambda_i^{(t+1)} &= \lambda_i^{(t)} + \rho^{(t)} v_i(\bar{x}^{(t)}), \quad i \in \mathcal{C}_t, \\ \rho^{(t+1)} &= \begin{cases} \min(\gamma \rho^{(t)}, \rho_{\max}), & V_{\mathcal{C}_t}^{(t)} > 0, \\ \rho^{(t)}, & V_{\mathcal{C}_t}^{(t)} = 0. \end{cases} \end{aligned} \tag{5}$$

Multipliers start at zero, making nonnegative projection inactive. Inactive multipliers remain unchanged. When a proposal is rejected, the retained incumbent drives the update. Thus the accumulated pressure reflects the search state that persists into the next round, while rejected attempts remain available to replay. Empty rounds skip updates. Active-set expansion resets ρ to $\rho_0 > 0$ before rescoring, retaining multipliers.

The current-status block reports each active property's value, threshold, direction, signed margin, and violation. The history block reports λ_i, accumulated from accepted-state violations. These blocks answer different questions. Measurements determine which constraints the current molecule satisfies, while multiplier history identifies constraints that have repeatedly resisted the search. A feasible molecule can retain a large multiplier because earlier accepted states violated that constraint. This pressure increases the cost of reintroducing the violation in future candidates. Conversely, zero pressure alone does not establish feasibility. The Reviewer receives each candidate's own measurements together with this history, allowing proposal guidance to use both present satisfaction and persistent constraint priorities.

Algorithm 1 PHARMAGENT: external-state search with a frozen proposal model

Require: Oracles $f, \{g_i\}$, budget B, round cap $T_{\max}$, run-local archive $\mathcal{D}$
1: Score seed pool within B. initialize incumbent x, seen set $\mathcal{S}$, spent calls b, $\boldsymbol{\lambda} = 0$, $\rho = \rho_0$
2: **for** $t = 0, \ldots, T_{\max} - 1$ while $b < B$ **do**
3: Set $\mathcal{C}_t$. on expansion reset ρ, retain $\boldsymbol{\lambda}$. rescore x
4: Retrieve eligible structure-indexed transitions. assemble facts, history, RAG, and feedback
5: Generate at most five candidates. verify validity and global novelty
6: If none remain, allow one repair request. if still empty, retain context and continue
7: Evaluate novel candidates within remaining B. update $b, \mathcal{S}$. select round-best c
8: Accept $\bar{x}$ by Eq. 4. invoke Reviewer on scored outcomes and accepted state
9: Update multipliers from $\bar{x}$ by Eq. 5. record evaluated transition (x, c)
10: Update replay and bounded recovery context. set $x \leftarrow \bar{x}$
11: **end for**
12: **return** Evaluations, accepted trajectory, dual history, and archive $\mathcal{D}$

3.4 STRUCTURE-INDEXED EXPERIENCE REPLAY

Chemical-CER retrieves structurally relevant attempts using source Morgan fingerprints (radius 2, 2048 bits) (Rogers & Hahn, 2010). A completed round records incoming incumbent $y^{(t)}$, round-best candidate $c^{(t)}$, target/ADMET changes, and the scalar change under the collection-time objective. Replay stores the evaluated attempt $(y^{(t)}, c^{(t)})$, and the acceptance rule tracks $(y^{(t)}, \bar{x}^{(t)})$. The stored scalar delta retains the collection-time objective.

Retrieve up to k eligible transitions by source-structure proximity:

$$\mathcal{E}_k^{(t)} = \mathrm{TopK}_{e \in \mathcal{D}_t^{\text{eligible}}} \left(-\|\mathbf{fp}_{y^{(t)}} - \mathbf{fp}_{\text{src}}(e)\|_2 \right). \tag{6}$$

Replay provides target-only information in Phases 1–2 and combined-objective information for all seven constraints in Phase 3, with phase-specific retrieval limits (Appendix G). Retrieved canonical source/destination pairs and finite numeric deltas condition the Generator alongside fixed RAG context. Each novel proposal is evaluated before acceptance. The run-local archive persists across tasks. The CER ablation evaluates the contribution of this replay component.

3.5 CONSTRAINT SCHEDULING

A fixed target-first curriculum (Hou, 2026) activates no ADMET constraints in rounds $[0, 20)$, MW/logP/QED/SA in $[20, 60)$, and all seven in $[60, \infty)$. Round indices determine the constraint-activation boundaries. Expansion adds measurements to evaluation and context, resetting quadratic pressure while retaining multipliers. Stagnation triggers bounded context recovery with the same activation schedule and acceptance rule (Appendix H). A continuous sigmoid-based ADMET score supports replay annotation and Pareto evaluation (Eq. A1). Hard feasibility requires all seven inequalities, and acceptance uses Eq. 3.

3.6 MULTIPLIER DYNAMICS

Index only completed updates after full activation, with $V^{(t)} = \sum_i v_i(\bar{x}^{(t)})$. $\boldsymbol{\lambda}^{(0)}$ includes prior-phase history. For any finite T, telescoping Eq. 5 gives the exact identity

$$\sum_{t=0}^{T-1} \rho^{(t)} V^{(t)} = \sum_{i=1}^{K} (\lambda_i^{(T)} - \lambda_i^{(0)}). \tag{7}$$

Each multiplier accumulates the ρ-weighted violations of accepted or retained states. When constraint i is satisfied, its increment is zero while earlier pressure persists. At fixed measurements and ρ, increasing λ_i by $\Delta\lambda_i$ lowers a candidate's score by $\Delta\lambda_i v_i(x)$. The accumulated history therefore raises the cost of reintroducing a violation. Appendix A derives the identity and states a conditional vanishing-violation result under bounded multipliers.

4 EXPERIMENTS

4.1 TASKS, METRICS, AND COMPARISON PROTOCOL

We evaluate glycogen synthase kinase 3 beta (GSK3β), c-Jun N-terminal kinase 3 (JNK3), dopamine receptor D2 (DRD2), Sitagliptin multi-property optimization (MPO), and Albuterol similarity, spanning predicted target activity, a multi-property objective, and structural similarity. These tasks test different routes to target improvement under the same seven-property screen. ADMET-PMO adds the seven shared property constraints illustrated in Figure 1(a), with thresholds in Table A4. Target AUC Top-10 measures target progress (Gao et al., 2022). AA-AUC averages the running best target score among molecules satisfying all seven constraints. Pareto hypervolume (PHV) measures dominated area in target–continuous-ADMET-score space (Zitzler et al., 2003). PHV uses both feasible and infeasible evaluated molecules. Definitions appear in Appendix E.

All main and ablation configurations use five independent runs indexed 0–4. All methods are allowed up to 1,000 unique target evaluations per task/run, including initialization. AUC uses a common 1,000-call horizon. SEISMO follows its original short-run protocol (Krüger et al., 2026). After termination, its final Top-10 and best-feasible scores are carried forward over the remaining horizon. All LLM methods use MiniMax-M3 (MiniMax, 2026). † denotes independent MOLLEO/ExLLM reimplementations, while SEISMO uses its released prompts and trajectory protocol. Table rankings compare five-run means. PHARMAGENT scores 26 seed molecules within budget. Defaults are $\rho_0 = 1$, $\gamma = 1.5$, $\rho_{\max} = 100$, and top-5 retrieval.

E1 compares five original baselines. E2 compares six online strategies because Graph-GA uses both fixed-penalty (FP) and feasibility-first (FF) search. FP keeps violation weights constant within each run. FF prioritizes constraint satisfaction over target score. All online baselines use the same seven constraints, ADMET-AI 1.4.0 predictions, and thresholds from their first evaluation. Original trajectories receive separate post-hoc annotation. Aggregate uncertainty averages the five fixed tasks within each run, then uses five run-level observations. We report unadjusted 95% intervals. E2–E1 differences use run-index matching, and cross-method differences use independent-group errors (Appendix B).

4.2 E1 TARGET-ONLY SEARCH CAPABILITY

Table 2: Standard PMO target AUC Top-10 (higher is better), averaged over five runs. Sum adds the five task means. All LLM methods use MiniMax-M3. † and SEISMO's original short-run protocol with terminal-score carry-forward are defined in Section 4.1. Best/second-best means are bold/underlined.

Method	GSK3β	JNK3	DRD2	Sitagliptin	Albuterol	Sum
Graph-GA	0.4600	0.1814	0.6939	0.2213	0.4423	1.9988
REINVENT4	0.3792	0.1742	0.6917	0.1513	0.3947	1.7911
MOLLEO†	0.5028	<u>0.3658</u>	0.8041	<u>0.2999</u>	0.8840	<u>2.8565</u>
ExLLM†	<u>0.5150</u>	0.3302	0.7981	0.2657	0.8860	2.7950
SEISMO	0.4850	0.2326	**0.9756**	0.1344	**0.9791**	2.8067
PHARMAGENT	**0.7564**	**0.9186**	<u>0.9494</u>	**0.3472**	<u>0.9492</u>	**3.9208**

E1 disables ADMET guidance and Lagrangian penalties, retaining target-only scoring and validity checks. PHARMAGENT achieves the highest target-AUC sum of 3.9208, surpassing MOLLEO† at 2.8565 by 37.3%. It leads GSK3β, JNK3, and Sitagliptin and ranks second on DRD2 and Albuterol. The aggregate target-AUC difference from each of the four original full-budget baselines has a positive 95% interval (Appendix B).

4.3 E2 SEARCH WITH ONLINE CONSTRAINTS

PHARMAGENT leads all six online strategies on every aggregate metric in Table 3. Its target-AUC sum of 3.6936 exceeds the strongest baseline, SEISMO FF (2.9319), by 26.0%.

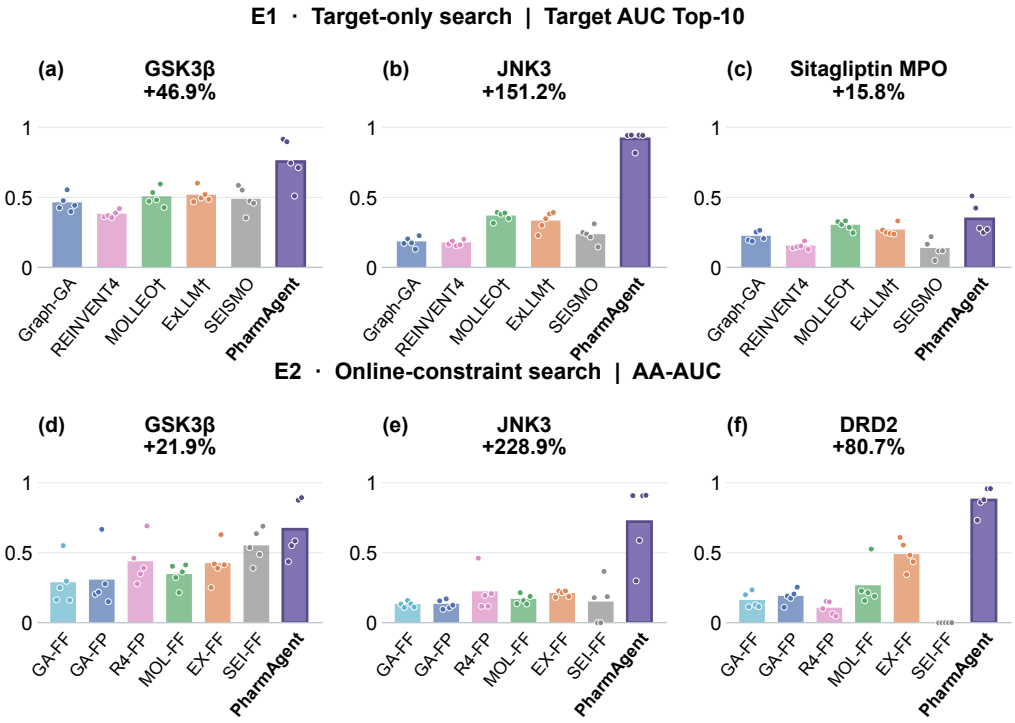

Figure 2: Task-level comparisons against (a–c) original baselines in E1 and (d–f) all six online strategies in E2. Bars show means, dots all five runs, and percentages gains over the strongest baseline in each panel. The selected tasks highlight leading means. Full task results appear in Appendix B. GA, R4, MOL, EX, and SEI abbreviate Graph-GA, REINVENT4, MOLLEO, ExLLM, and SEISMO. FF denotes feasibility-first and FP fixed-penalty search. SEISMO uses terminal-score carry-forward. † marks independent reimplementations.

Table 3: Online-constraint results across five tasks and five runs. Target AUC Sum adds task means. The other metrics average tasks and runs. All methods receive online constraints. FF denotes feasibility-first and FP fixed-penalty search. Best/second-best means are bold/underlined. Figure 3 reports uncertainty.

Method	Target AUC Sum	AA-AUC	PHV	Top-1	Top-10
Graph-GA (FF)	1.0675	0.2536	0.4181	0.4375	0.3219
Graph-GA (FP)	1.1809	0.2711	0.4630	0.4837	0.3575
REINVENT4 (FP)	1.8337	0.2859	0.5622	0.6221	0.4663
MOLLEO† (FF)	2.0465	0.4006	0.6430	<u>0.6788</u>	0.5864
ExLLM† (FF)	2.2732	<u>0.4602</u>	<u>0.6514</u>	0.6647	<u>0.6352</u>
SEISMO (FF)	<u>2.9319</u>	0.3838	0.5795	0.6405	0.5978
PHARMAGENT (E2)	**3.6936**	**0.7076**	**0.8248**	**0.8514**	**0.8305**

For AA-AUC and PHV, ExLLM FF is the strongest baseline. PHARMAGENT improves its 0.4602 AA-AUC to 0.7076 (53.8%) and its 0.6514 PHV to 0.8248 (26.6%). The AA-AUC difference is 0.2474, with an unadjusted 95% interval of $[0.1355, 0.3594]$.

PHARMAGENT leads on target-AUC sum (Figure 3(a)), AA-AUC (Figure 3(b)), and PHV (Figure 3(c)), with positive difference intervals against every online strategy on all three metrics. With shared online constraint measurements, PHARMAGENT leads target progress, feasible-solution progress, and Pareto coverage. Its target-AUC advantage extends to all five tasks, and it also leads AA-AUC on GSK3β, JNK3, and DRD2.

Impact of online constraint control. Post-hoc evaluation of E1's unchanged trajectories gives AA-AUC 0.5328, versus E2's 0.7076, a 32.8% improvement with difference interval

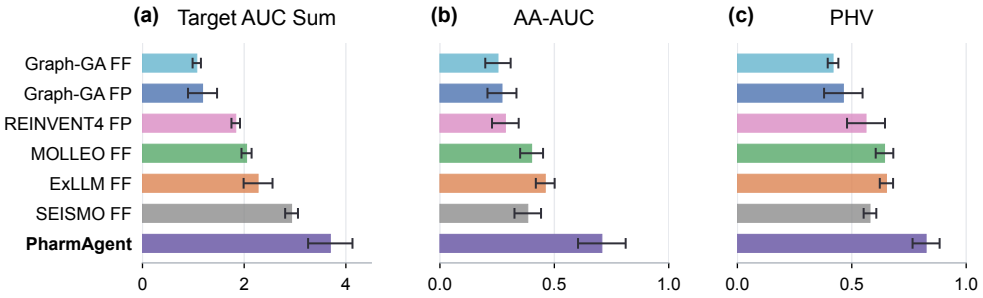

Figure 3: Online comparison across five tasks and five runs. (a–c) Target-AUC sum, AA-AUC, and PHV for six baseline strategies and PharmAgent E2. Feasibility-first (FF) search prioritizes constraint satisfaction. Fixed-penalty (FP) search penalizes total normalized violation with a weight fixed within each run. Bars show means with unadjusted 95% intervals after within-run task aggregation. All methods have the same budget allowance. SEISMO retains its short-run protocol and terminal-score carry-forward.

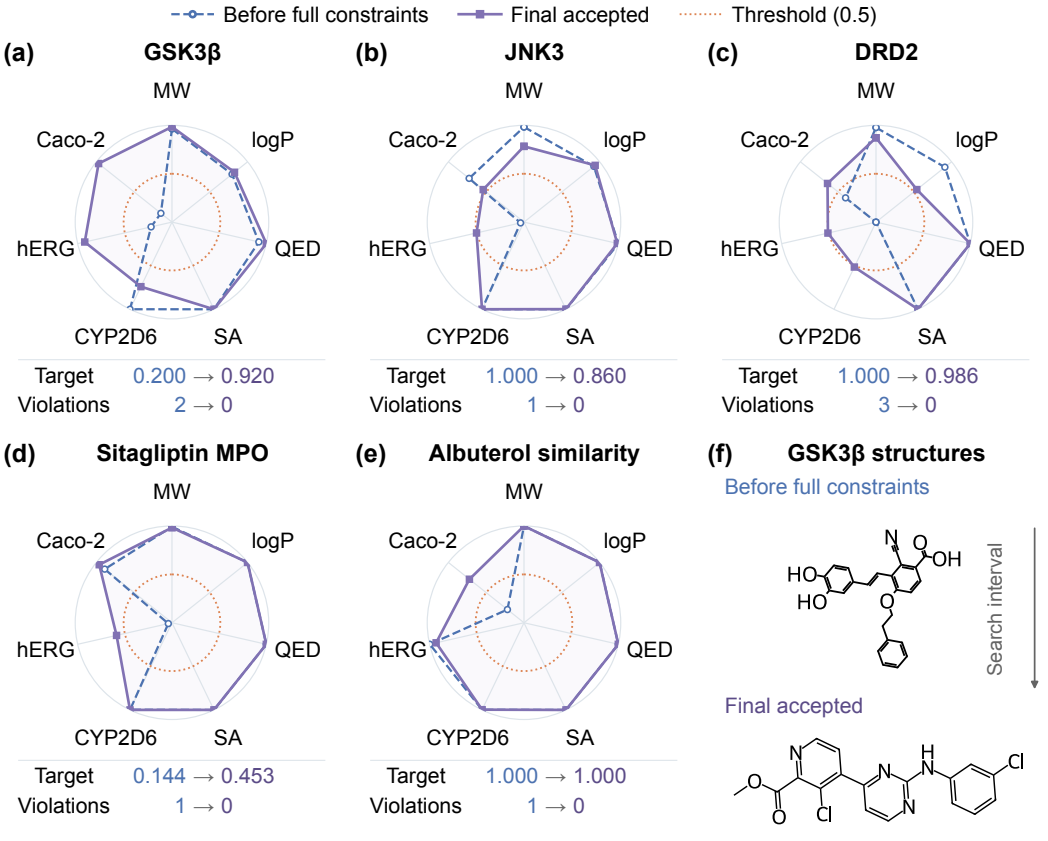

Figure 4: Five molecular cases (run 0, before full constraint activation at round 59 versus final accepted state). (a–e) Sigmoid desirability profiles under all seven requirements. Outward is favorable, and the 0.5 ring marks thresholds. CYP2D6 denotes inhibition. Target scores and violation counts accompany each profile. (f) Recorded GSK3β structures with an arrow spanning subsequent rounds. ADMET endpoints are predictions.

$[0.0511, 0.2987]$. All five run-level aggregate differences are positive. Figure 2(a–c) highlights target-search gains over the original baselines, while Figure 2(d–f) shows AA-AUC gains over their online adaptations. Appendix B provides budgeted progress in Figure A1(a) and all five E2–E1 task effects in Figure A1(b). Task-level AA-AUC intervals exclude zero for DRD2, Sitagliptin, and Albuterol. Online feedback thus turns strong target search into stronger feasible-solution search.

4.4 MECHANISM ABLATIONS

Table 4: ADMET-PMO ablations over five tasks and five runs. The full-system reference matches Table 3. Target AUC Sum adds task means. AA-AUC and PHV average tasks and runs. Best/second-best means are bold/underlined.

Variant	Target AUC Sum	AA-AUC	PHV
PHARMAGENT (full system)	**3.6936**	**0.7076**	**0.8248**
w/o Lagrangian	<u>3.6535</u>	0.4781	0.7815
w/o Chemical-CER	3.4008	0.5459	<u>0.7904</u>
w/o Curriculum	3.1981	<u>0.6260</u>	0.7295
w/o MedChem-RAG	2.9977	0.4146	0.6593
w/o Reviewer agent	3.3024	0.5566	0.7100
Single-agent baseline	3.2698	0.3000	0.6546

The full system outperforms all six variants on all three aggregate metrics (Table 4). Removing Lagrangian feedback reduces AA-AUC by 32.4%, compared with a 1.1% reduction in target-AUC sum. The loss is concentrated in feasible-solution progress. Removing CER reduces target-AUC sum by 7.9% and AA-AUC by 22.9%. Removing Curriculum reduces them by 13.4% and 11.5%, and PHV by 11.6%. MedChem-RAG removal produces the largest target-AUC reduction among component removals (18.8%) and lowers AA-AUC by 41.4%. Reviewer removal lowers AA-AUC by 21.3%. The full system reaches 0.7076 AA-AUC versus 0.3000 for the single-agent configuration. The distinct loss profiles reveal complementary contributions from constraint feedback, structural experience, scheduling, chemical priors, and review.

4.5 MOLECULAR CASES OF PROPERTY-VIOLATION RESOLUTION

Figure 4(a–e) compares property profiles before full constraint activation (round 59) and at termination, using run 0 uniformly across tasks. Figure 4(f) shows the corresponding GSK3β structures. PHARMAGENT resolves property violations in all five illustrated trajectories. Every earlier state violates at least one of the final seven requirements, and every final state satisfies all seven. Target scores increase for GSK3β (0.20 to 0.92) and Sitagliptin (0.1443 to 0.4533), remain at 1.00 for Albuterol, and decrease for JNK3 (1.00 to 0.86) and DRD2 (1.0000 to 0.9864). For GSK3β, the predicted hERG block score falls from 0.5754 to 0.3496 (bound 0.5), while Caco-2 increases from -5.5022 to -4.4525 (bound -5.15). Appendix C provides all structure pairs, raw properties, and the selection protocol.

5 DISCUSSION AND CONCLUSION

PHARMAGENT achieves leading aggregate performance through adaptive external search state and frozen language models. E1 leads aggregate target search, and E2 outperforms all six online strategies on all five aggregate metrics. The Lagrangian controller gives constraint history an explicit role in acceptance and proposal feedback, while Chemical-CER and the curriculum supply relevant measured experience and coordinate constraint activation. Removing Lagrangian feedback lowers AA-AUC by 32.4%. Every component removal lowers all three aggregate metrics, and all five molecular cases resolve their property violations. Together, these results establish adaptive constraint feedback and structured chemical experience as a powerful approach to molecular search with frozen models.

Limitations. Evaluation covers five tasks and predicted properties. Multiplier boundedness remains an assumption of the appendix analysis, and ablation uncertainty is unavailable. Appendix M summarizes the few task-level exceptions, validation needs, and computational-cost considerations.

AI USE STATEMENT

Generative AI tools were used solely for language editing and polishing, including improvements to grammar, wording, clarity, and readability. All scientific content, methodology, experiments, analyses, results, and conclusions were developed by the authors. All AI-assisted edits were reviewed and verified by the authors, who take full responsibility for the final manuscript.

ETHICS STATEMENT

This study evaluates computational molecular optimization using predicted properties. It does not establish experimental drug safety or clinical efficacy. Candidate structures and predicted ADMET measurements require independent validation before practical use. We discuss potential misuse and the limits of surrogate evaluation in Appendix J.

REPRODUCIBILITY STATEMENT

We have submitted an anonymized code package containing the complete source code for the PHARMAGENT system, in accordance with the submission requirements. Section 3 specifies the controller, acceptance rule, and experience retrieval. Appendix A derives the accumulation identity and states the assumptions for conditional feasibility. Appendices E, F, G, H, and I document the evaluation protocol, prompts, replay interface, runtime configuration, and hyperparameters. The experiments use five independent runs indexed 0–4. Provider-controlled sampling is described in Appendices H and M. Appendix B reports uncertainty and Appendix D specifies the online baseline adaptations. The accompanying scripts reproduce figures and interval tables from the recorded run-level results.

REFERENCES

Dimitri P. Bertsekas. On penalty and multiplier methods for constrained minimization. *SIAM Journal on Control and Optimization*, 14(2):216–235, 1976. doi: 10.1137/0314017.

Dimitri P Bertsekas. *Nonlinear Programming*. Athena Scientific, 2nd edition, 1999.

G Richard Bickerton, Gaia V Paolini, Jérémy Besnard, Sorel Muresan, and Andrew L Hopkins. Quantifying the chemical beauty of drugs. *Nature Chemistry*, 4:90–98, 2012.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

Lianghong Chen, Dongkyu Eugene Kim, Mike Domaratzki, and Pingzhao Hu. Uncertainty-aware multi-objective reinforcement learning-guided diffusion models for 3D de novo molecular design. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-4157.

Alexander G Dossetter, Edward J Griffen, and Andrew G Leach. Matched molecular pair analysis in drug discovery. *Drug Discovery Today*, 18:724–731, 2013.

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.

Yin Fang, Ningyu Zhang, Zhuo Chen, Lingbing Guo, Xiaohui Fan, and Huajun Chen. Domain-agnostic molecular generation with chemical feedback. In *International Conference on Learning Representations*, 2024.

Wenhao Gao and Connor W. Coley. The synthesizability of molecules proposed by generative models. *Journal of Chemical Information and Modeling*, 60(12):5714–5723, 2020. doi: 10.1021/acs.jcim.0c00174.

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 (NeurIPS), Datasets and Benchmarks Track*, 2022.

Rafael Gómez-Bombarelli, Jennifer N. Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D. Hirzel, Ryan P. Adams, and Alán Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Central Science*, 4(2):268–276, 2018. doi: 10.1021/acscentsci.7b00572.

Ed Griffen, Andrew G. Leach, Graeme R. Robb, and Daniel J. Warner. Matched molecular pairs as a medicinal chemistry tool. *Journal of Medicinal Chemistry*, 54(22):7739–7750, 2011. doi: 10.1021/jm200452d.

Junyu Hou. De novo molecular design enabled by direct preference optimization and curriculum learning. In *Machine Learning and Knowledge Discovery in Databases. Research Track. ECML PKDD 2025*, volume 16018 of *Lecture Notes in Computer Science*, pp. 95–111. Springer, 2026. doi: 10.1007/978-3-032-06106-5_6.

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

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2021. doi: 10.1109/TBDATA.2019.2921572.

Fabian P Krüger, Andrea Hunklinger, Adrian Wolny, Tim J Adler, Igor Tetko, and Santiago David Villalba. SEISMO: Increasing sample efficiency in molecular optimization with a trajectory-aware LLM agent. *arXiv preprint arXiv:2602.00663v1*, 2026.

Greg Landrum, Paolo Tosco, Brian Kelley, et al. RDKit: Open-source cheminformatics. Zenodo, 2024. URL `https://doi.org/10.5281/zenodo.10793672`. Version 2023.09.6.

Yitao Liu, Chenglei Si, Karthik R Narasimhan, and Shunyu Yao. Contextual experience replay for self-improvement of language agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14179–14198. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.694.

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(1):20, 2024.

Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D. White, and Philippe Schwaller. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 6(5):525–535, 2024. doi: 10.1038/s42256-024-00832-8.

MiniMax. OpenAI SDK. `https://platform.minimax.io/docs/api-reference/text-openai-api`, 2026. Accessed September 21, 2026.

Yifan Niu, Ziqi Gao, Tingyang Xu, Yang Liu, Yatao Bian, Yu Rong, Junzhou Huang, and Jia Li. InversionGNN: A dual path network for multi-property molecular optimization. In *International Conference on Learning Representations*, 2025.

Nian Ran, Yue Wang, Xiaoyuan Zhang, Zhongzheng Li, Qingsong Ran, Wenhao Li, and Richard Allmendinger. ExLLM: Experience-enhanced LLM optimization for molecular design and beyond. *arXiv preprint arXiv:2502.12845*, 2025.

David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50:742–754, 2010.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-0377.

Kyle Swanson, Parker Walther, Jeremy Leitz, Souhrid Mukherjee, Joseph C. Wu, Rabindra V. Shivnaraine, and James Zou. ADMET-AI: A machine learning ADMET platform for evaluation of large-scale chemical libraries. *Bioinformatics*, 40(7):btae416, 2024. doi: 10.1093/bioinformatics/btae416.

Haorui Wang, Marta Skreta, Cher-Tian Ser, Wenhao Gao, Lingkai Kong, Felix Strieth-Kalthoff, Chenru Duan, Yuchen Zhuang, Yue Yu, Yanqiao Zhu, Yuanqi Du, Alán Aspuru-Guzik, Kirill Neklyudov, and Chao Zhang. Efficient evolutionary search over chemical space with large language models. In *International Conference on Learning Representations (ICLR)*, 2025.

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.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V. Le, Denny Zhou, and Xinyun Chen. Large language models as optimizers. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Bb4VGOWELI.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

Barbara Zdrazil, Eloy Felix, Fiona Hunter, Emma J Manners, James Blackshaw, Sybilla Corbett, Marleen de Veij, Harris Ioannidis, David Mendez Lopez, Juan F Mosquera, Maria Paula Magarinos, Nicolas Bosc, Ricardo Arcila, Tevfik Kizilören, Anna Gaulton, A Patrícia Bento, Melissa F Adasme, Peter Monecke, Gregory A Landrum, and Andrew R Leach. The ChEMBL database in 2023: A drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic Acids Research*, 52(D1):D1180–D1192, 2024. doi: 10.1093/nar/gkad1004.

Eckart Zitzler, Lothar Thiele, Marco Laumanns, Carlos M. Fonseca, and Viviane Grunert da Fonseca. Performance assessment of multiobjective optimizers: An analysis and review. *IEEE Transactions on Evolutionary Computation*, 7(2):117–132, 2003.

APPENDIX

The appendices follow the main paper's questions. Appendix A derives multiplier accumulation and conditional feasibility. Appendix B reports run-level uncertainty and complete task results. Appendix C documents the molecular cases, and Appendix D details the online adaptations, task-level margins, and changes from the original protocols. Appendices E–H specify metrics, prompts, replay, and computational accounting. Appendix M collects the limitations and remaining validation questions.

A CUMULATIVE VIOLATION AND CONDITIONAL FEASIBILITY

Update sequence. Here t indexes completed updates after the full constraint set is active. Empty or failed rounds are excluded, and $\boldsymbol{\lambda}^{(0)} \geq 0$ retains earlier-phase history. The state $\bar{x}^{(t)}$ is the accepted candidate or retained incumbent from Eq. 4. Write $v_i^{(t)} = v_i(\bar{x}^{(t)}) \geq 0$ and $V^{(t)} = \sum_i v_i^{(t)}$.

Exact accumulation. Since $\rho^{(t)} > 0$ and violations are nonnegative, Eq. 5 keeps every multiplier nonnegative. Any nonnegative projection is therefore inactive. Telescoping the update for a single constraint gives

$$\lambda_i^{(T)} - \lambda_i^{(0)} = \sum_{t=0}^{T-1} \rho^{(t)} v_i^{(t)}.$$

Summing over i yields Eq. 7. These identities hold for every finite update sequence and follow algebraically from the rule, without assumptions on convexity or proposal quality.

Proposition 1 (Conditional vanishing violation)**.** *Consider an infinite sequence of completed updates under Eq. 5 after full constraint activation. Suppose $\rho^{(t)} \geq \rho_{\min} > 0$ for all t and, for each i, there is a finite constant M_i such that $\lambda_i^{(t)} \leq M_i$ for all t. Then $\sum_{t=0}^{\infty} V^{(t)} < \infty$ and $V^{(t)} \to 0$.*

Proof. For every finite T, the accumulation identity gives

$$0 \leq \sum_{t=0}^{T-1} \rho^{(t)} V^{(t)} = \sum_{i=1}^{K} (\lambda_i^{(T)} - \lambda_i^{(0)}) \leq \sum_{i=1}^{K} (M_i - \lambda_i^{(0)}).$$

The upper bound is independent of T. Nonnegativity and $\rho^{(t)} \geq \rho_{\min}$ therefore imply

$$\sum_{t=0}^{\infty} V^{(t)} \leq \frac{1}{\rho_{\min}} \sum_{i=1}^{K} (M_i - \lambda_i^{(0)}) < \infty.$$

A summable nonnegative sequence tends to zero. □

Interpretation of the assumptions. For $\gamma \geq 1$ and $\rho_{\max} \geq \rho_0 > 0$, the penalty update and active-set resets preserve $\rho^{(t)} \geq \rho_0$. The cap on ρ does not bound the multipliers, whose increments continue to accumulate. Uniform multiplier boundedness is an assumption of Proposition 1 and is not established for the implemented LLM search. The proposition describes a consequence of controlled cumulative violation, not an unconditional convergence guarantee for the search.

Limit points. Under the proposition's assumptions, suppose a subsequence $\bar{x}^{(t_m)}$ converges to $x^\star \in \mathcal{X}$ in a fixed metric and each v_i is continuous at $x^\star$. Then $0 \leq v_i(\bar{x}^{(t_m)}) \leq V^{(t_m)} \to 0$ implies $v_i(x^\star) = 0$ for every i. This establishes feasibility of such a limit point, without asserting that the molecular sequence has a limit point or converges.

B STATISTICAL REPORTING AND DETAILED RESULTS

All five tasks and five runs are retained. For a fixed-task aggregate, first average the five task values within each run index and then estimate uncertainty from the five resulting observations. Individual

molecules and the 25 task/run cells are not treated as independent repetitions. For values $z_1, \ldots, z_5$, intervals are $\bar{z} \pm t_{0.975,4} s_z/\sqrt{5}$, with $t_{0.975,4} = 2.776445$. For E2–E1 comparisons, z_j is the within-index difference. Index matching is not a claim of common LLM random draws. Cross-method differences use $\sqrt{s_A^2/5 + s_B^2/5}$ and the same conservative critical value. These intervals assume approximately normal run-level variation and are unadjusted descriptive intervals, not simultaneous claims over all comparisons. AUC sums are five times their corresponding fixed-task means, so their intervals scale by five and provide no additional independent evidence.

For aggregate E2–E1 AA-AUC, the difference is 0.1749 with matched interval $[0.0511, 0.2987]$. The independent-group sensitivity check gives $[0.0195, 0.3303]$. E2–E1 aggregate target AUC, Top-1, Top-10, and PHV intervals cross zero, which does not establish equivalence. Against original MOLLEO with post-hoc annotation, E2 has positive aggregate difference intervals for Top-1, target AUC, AA-AUC, and PHV. The final Top-10 interval crosses zero. Against all six online strategies, all 18 target-AUC, AA-AUC, and PHV difference intervals are positive (Figure A9(d–f)). These are separate, unadjusted comparisons. The online AA-AUC differences are 0.2474 versus ExLLM FF, 0.3070 versus MOLLEO FF, and 0.3238 versus SEISMO FF.

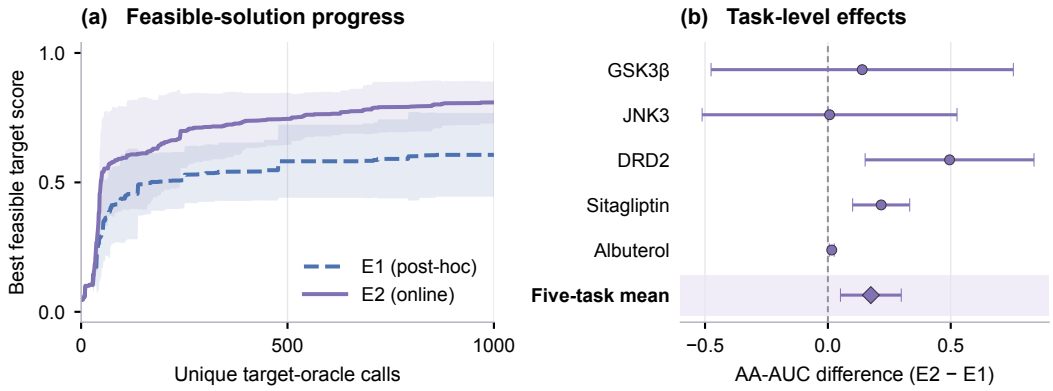

Figure A1: Online control improves feasible-solution search. (a) Best feasible target score versus unique oracle calls, including initialization. Lines average five task-averaged runs, with pointwise 95% intervals. (b) E2–E1 AA-AUC differences for every task and the fixed five-task mean, with run-index-matched 95% intervals ($n = 5$, unadjusted). Shared indices do not imply shared LLM draws. An independent-group aggregate check remains positive (Appendix B).

Online and original-protocol comparisons. Figures A2–A4 show every online strategy. Figures A5–A7 retain the original baselines and PharmAgent E1/E2. In each figure, panels (a–e) show GSK3β, JNK3, DRD2, Sitagliptin MPO, and Albuterol similarity, respectively, and panel (f) shows the fixed-task aggregate. All five runs are shown. Table A10 retains the post-hoc aggregate comparison. Relative to original MOLLEO, E2 improves target-AUC sum by 29.3%, AA-AUC by 70.7%, and PHV by 17.6%.

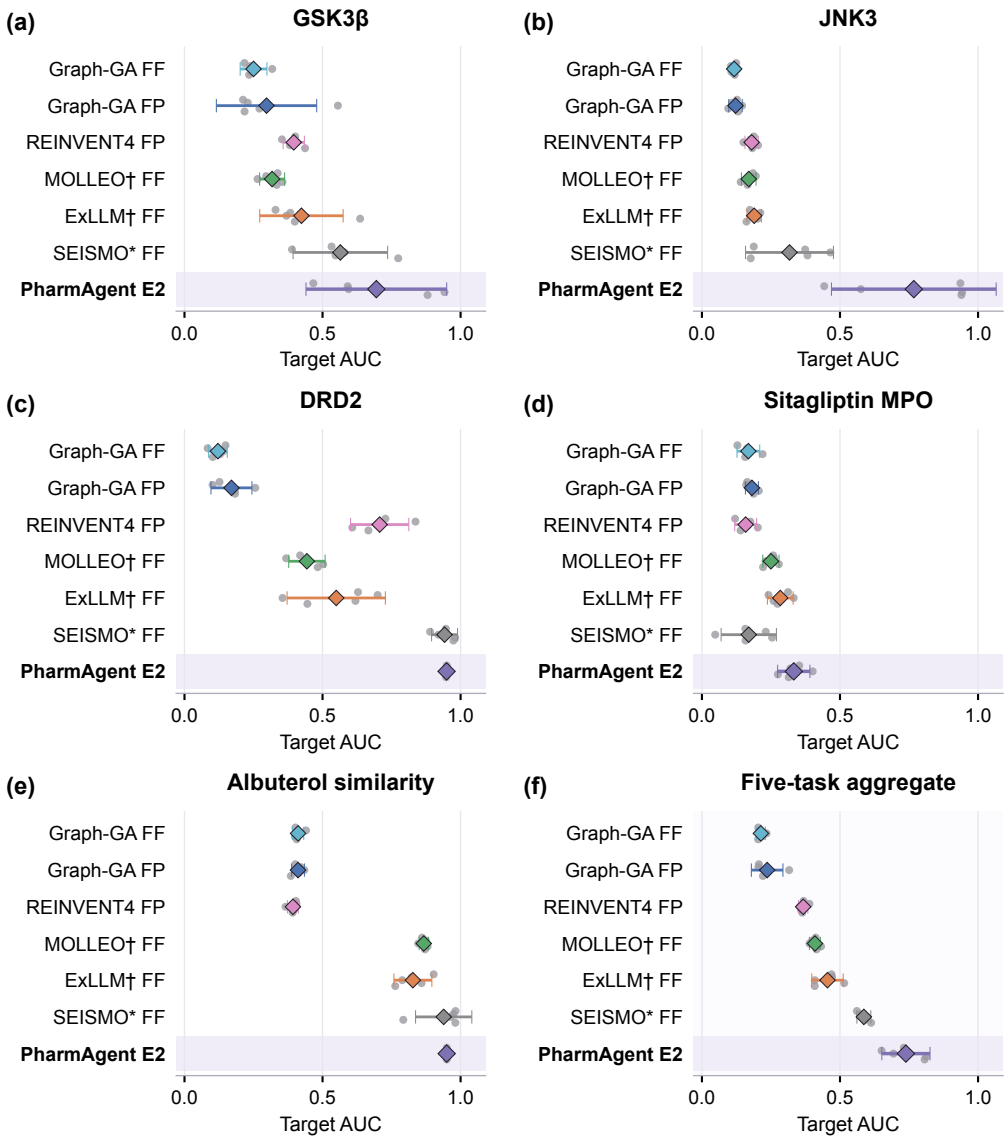

Figure A2: Target AUC across all six online baseline strategies and PharmAgent E2. Diamonds denote means, grey dots the five runs, and bars unadjusted 95% Student-t intervals. Panels (a–e) show the five tasks. Panel (f) averages tasks within each run before estimating uncertainty. Intervals are not clipped to $[0, 1]$. † denotes independent MiniMax-M3 reimplementations. *SEISMO follows its short-run protocol with terminal-score carry-forward. FF denotes feasibility-first and FP fixed-penalty search.

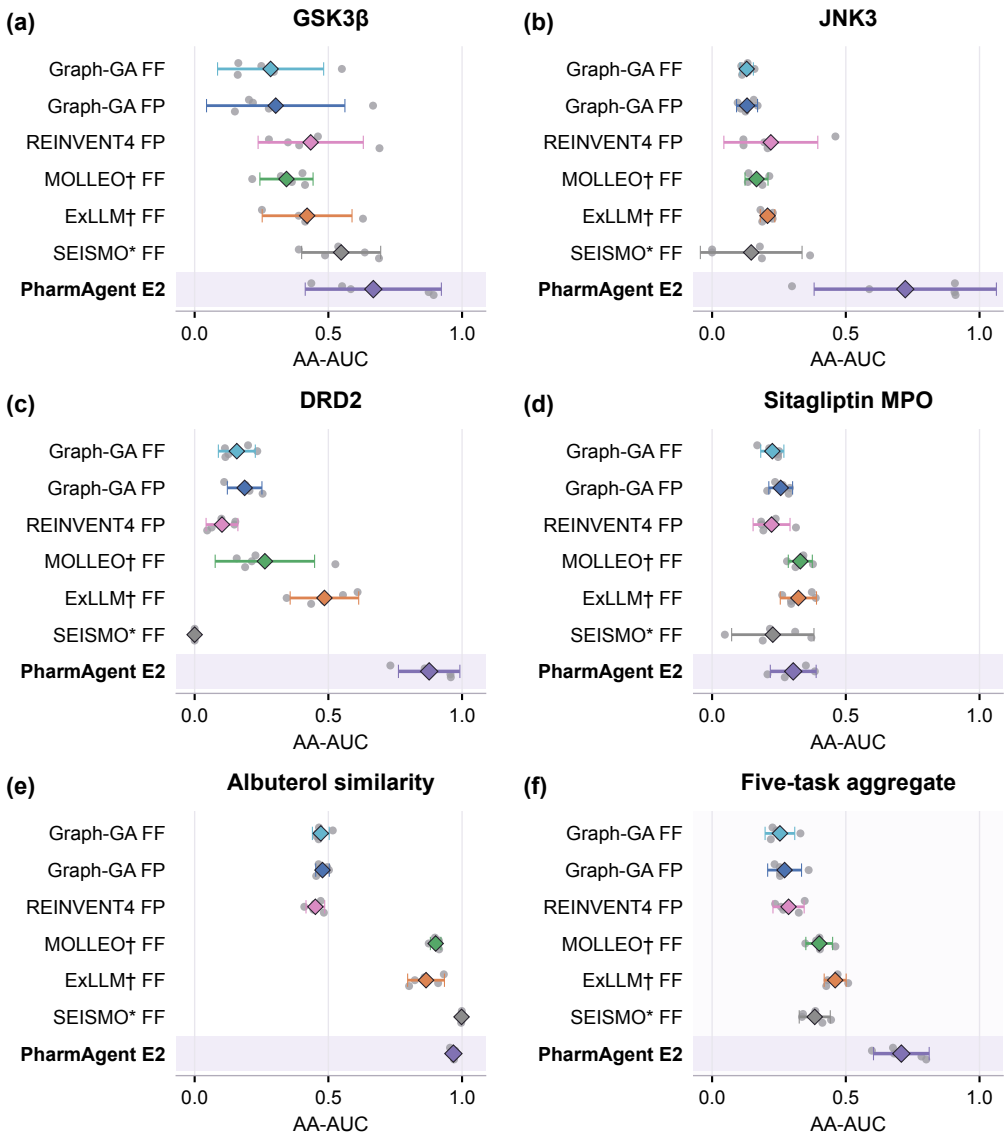

Figure A3: AA-AUC across all six online baseline strategies and PharmAgent E2. Diamonds denote means, grey dots the five runs, and bars unadjusted 95% Student-t intervals. Panels (a–e) show the five tasks. Panel (f) averages tasks within each run before estimating uncertainty. Intervals are not clipped to $[0, 1]$. † denotes independent MiniMax-M3 reimplementations. *SEISMO follows its short-run protocol with terminal-score carry-forward. FF denotes feasibility-first and FP fixed-penalty search.

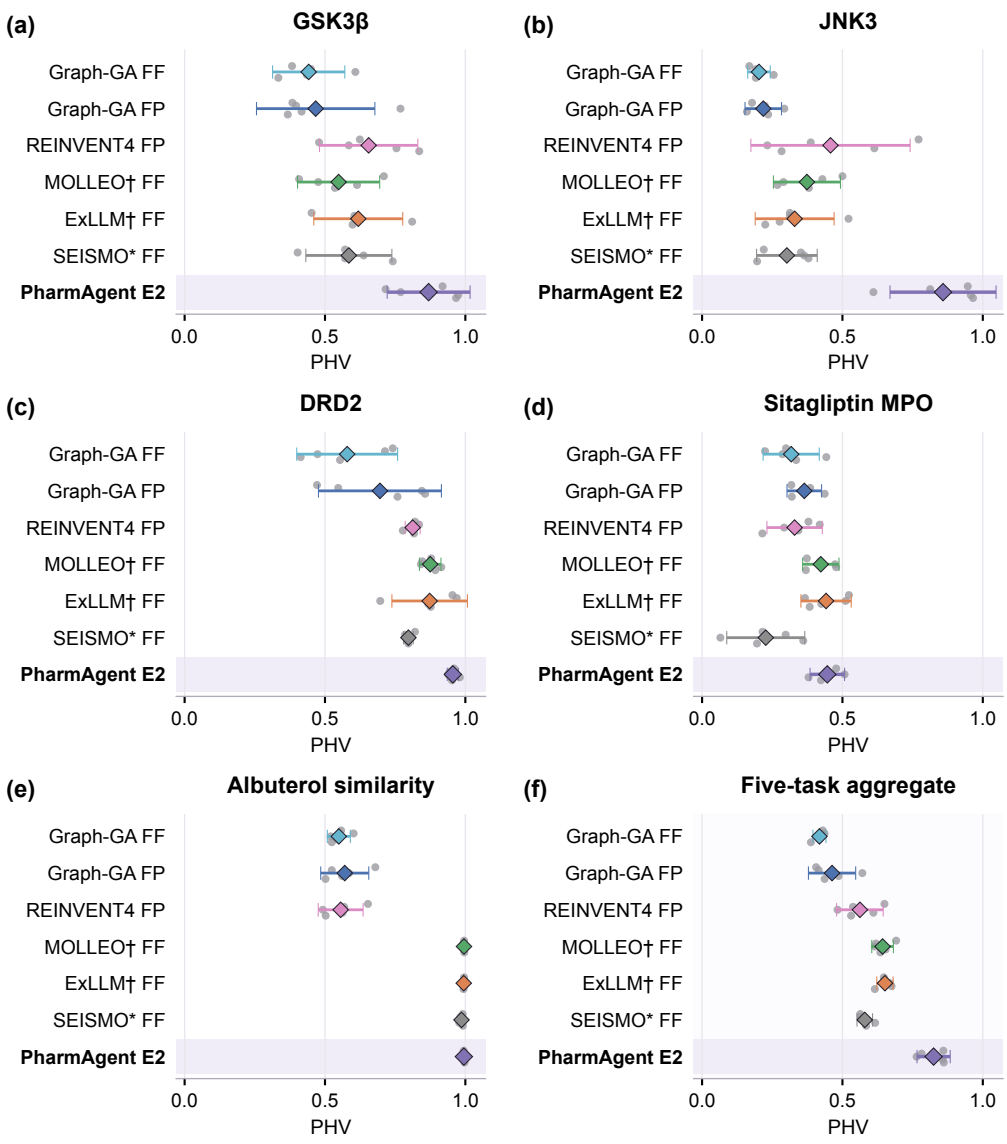

Figure A4: Pareto hypervolume across all six online baseline strategies and PharmAgent E2. Diamonds denote means, grey dots the five runs, and bars unadjusted 95% Student-t intervals. Panels (a–e) show the five tasks. Panel (f) averages tasks within each run before estimating uncertainty. Intervals are not clipped to $[0, 1]$. † denotes independent MiniMax-M3 reimplementations. *SEISMO follows its short-run protocol with terminal-score carry-forward. FF denotes feasibility-first and FP fixed-penalty search.

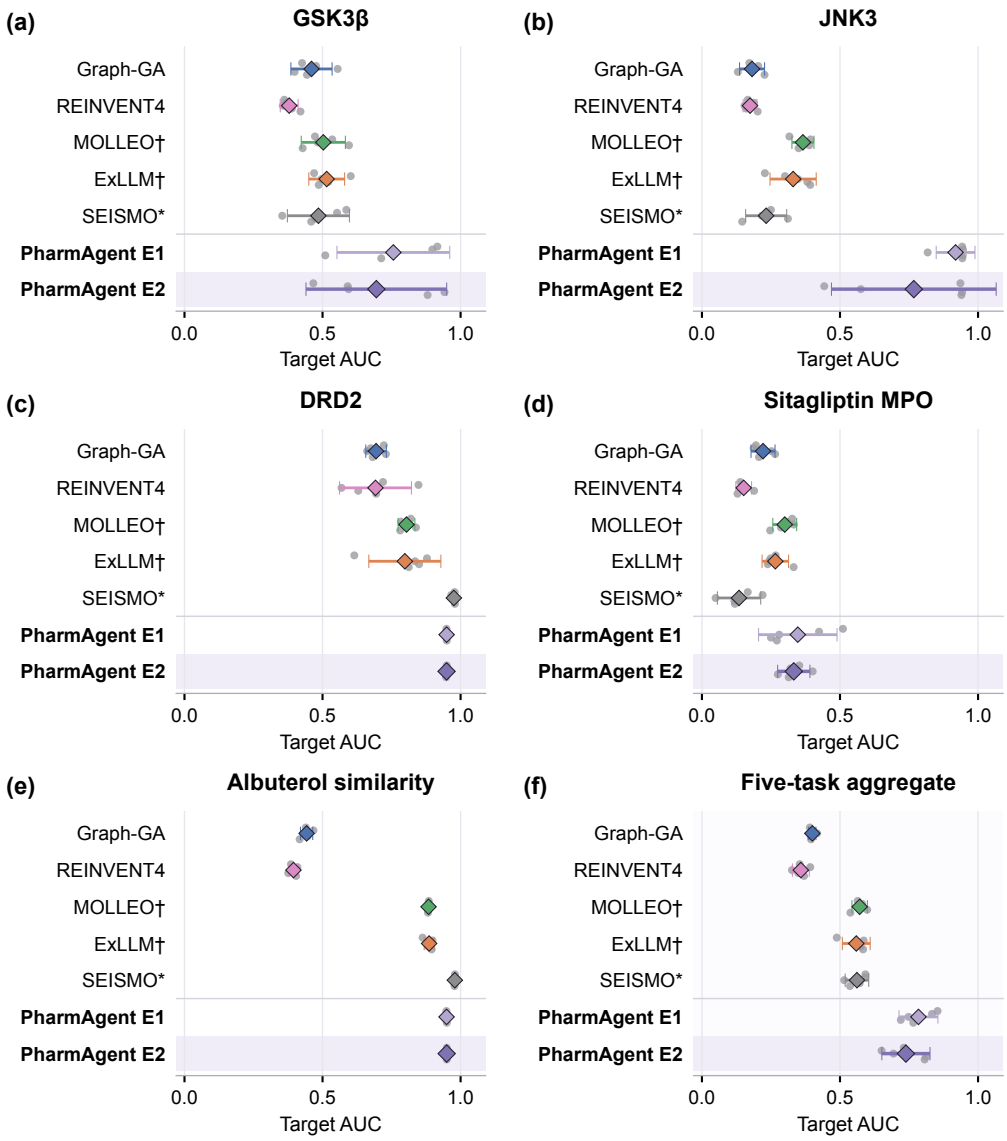

Figure A5: Target AUC for the five original baselines and PharmAgent E1/E2. Baselines and E1 receive post-hoc property annotation, while E2 uses online constraints. Diamonds denote means, grey dots the five runs, and bars unadjusted 95% Student-t intervals. Panels (a–e) show the five tasks. Panel (f) averages tasks within each run before estimating uncertainty. Intervals are not clipped to $[0, 1]$. † denotes independent MiniMax-M3 reimplementations. *SEISMO follows its short-run protocol with terminal-score carry-forward.

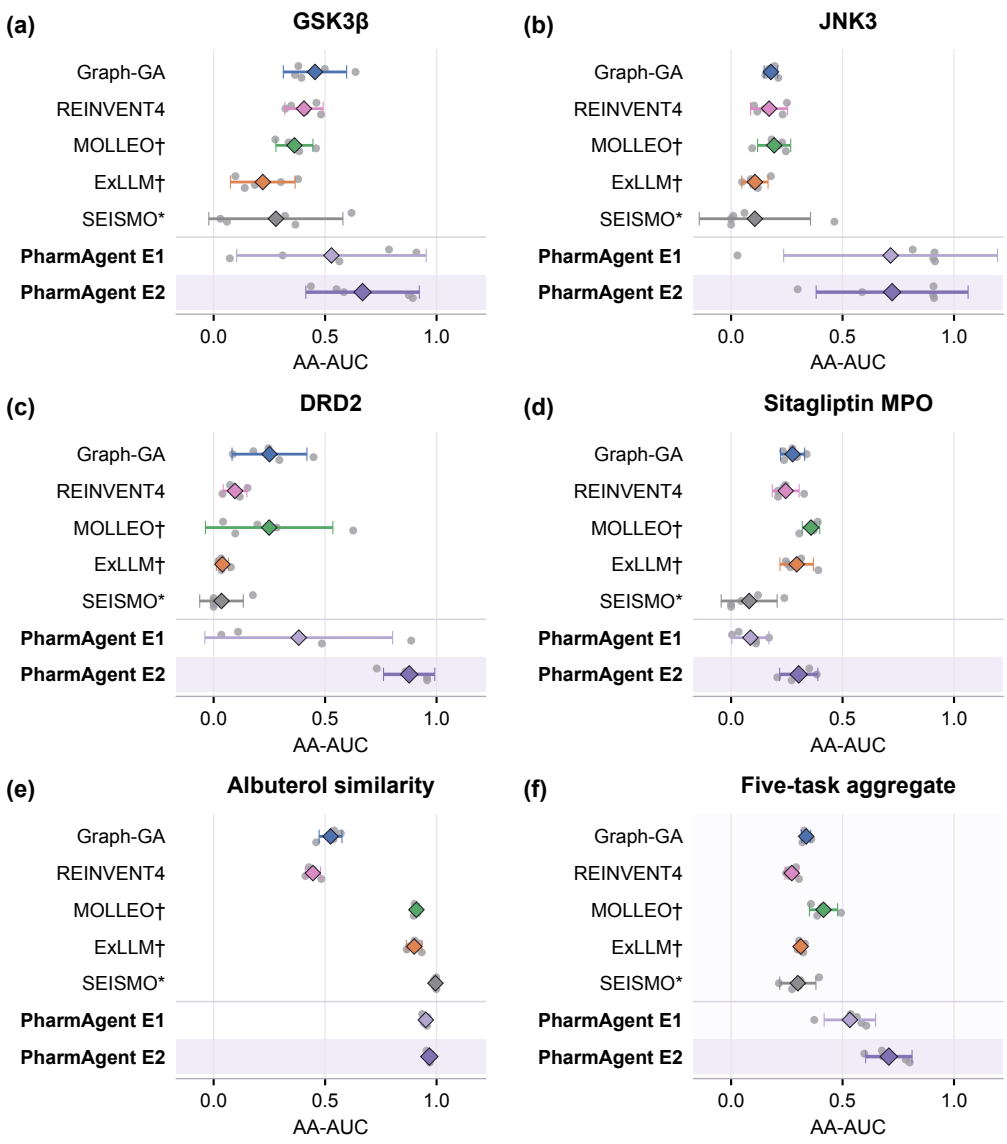

Figure A6: AA-AUC for the five original baselines and PharmAgent E1/E2. Baselines and E1 receive post-hoc property annotation, while E2 uses online constraints. Diamonds denote means, grey dots the five runs, and bars unadjusted 95% Student-t intervals. Panels (a–e) show the five tasks. Panel (f) averages tasks within each run before estimating uncertainty. Intervals are not clipped to $[0, 1]$. † denotes independent MiniMax-M3 reimplementations. *SEISMO follows its short-run protocol with terminal-score carry-forward.

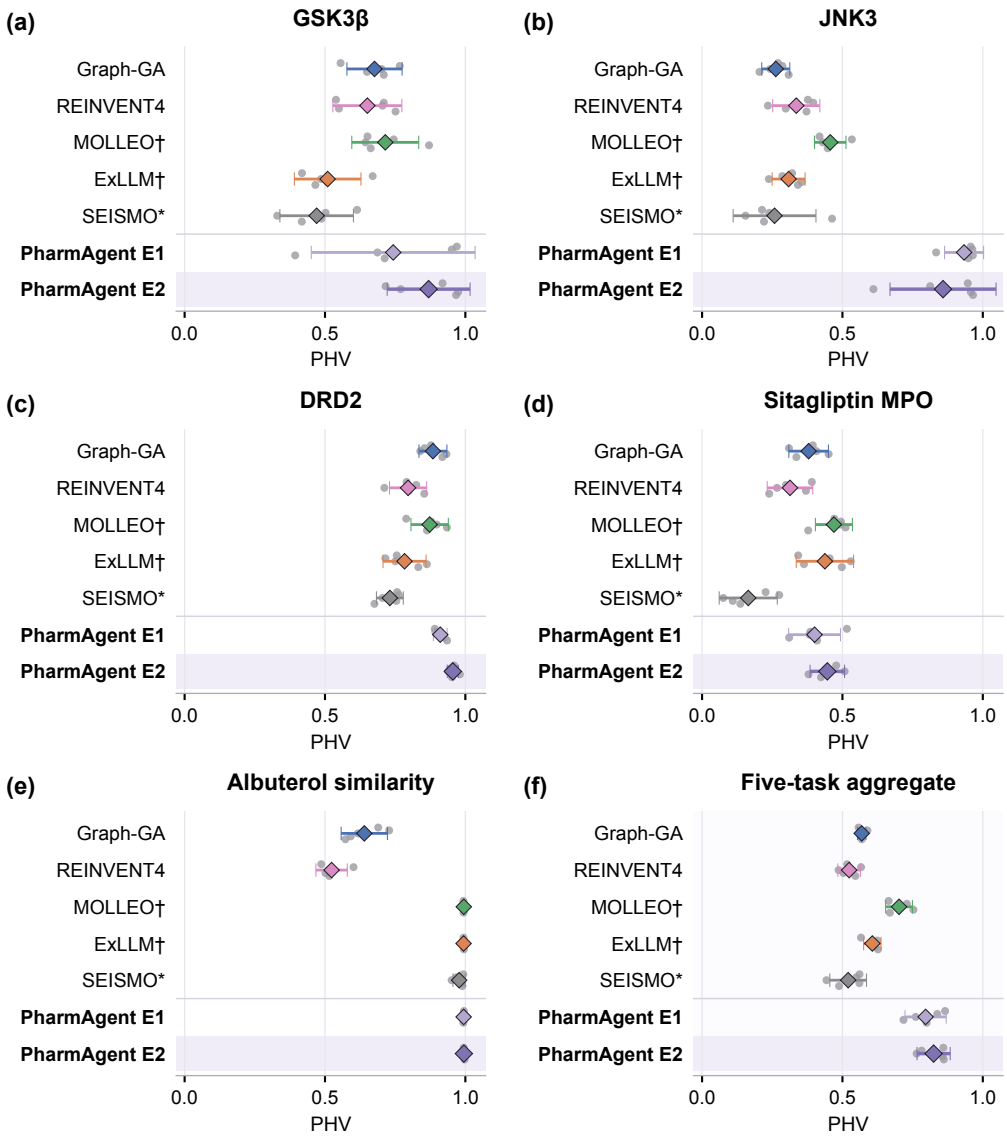

Figure A7: Pareto hypervolume for the five original baselines and PharmAgent E1/E2. Baselines and E1 receive post-hoc property annotation, while E2 uses online constraints. Diamonds denote means, grey dots the five runs, and bars unadjusted 95% Student-t intervals. Panels (a–e) show the five tasks. Panel (f) averages tasks within each run before estimating uncertainty. Intervals are not clipped to $[0, 1]$. † denotes independent MiniMax-M3 reimplementations. *SEISMO follows its short-run protocol with terminal-score carry-forward.

C Molecular Case Selection and Property Measurements

The five cases in Figure 4 use a fixed rule that compares the accepted state at round 59 with the final accepted state in run 0 for every task. Round 59 is the last round before all seven constraints are active under the prescribed schedule. The same run and stage endpoints are used for every task, independently of the magnitude of the target gain. A state can persist across many rounds. Its recorded evaluation call can therefore precede the displayed round. All plotted properties reuse stored measurements, and RDKit renders the corresponding structures.

For each raw property g_i, the radar radius is the existing desirability $d_i = \sigma(-5(\tilde{g}_i - \tilde{t}_i)/r_i)$, using the same directions, thresholds, and scales as Eq. A1. Thus a threshold maps to 0.5 and larger radii always indicate greater satisfaction. No normalization is fitted to the displayed cases and polygon area is not used as a metric. The sigmoid can saturate, so Table A1 retains raw values for inspection. The final accepted state and the historical best feasible molecule are distinct search outputs. In JNK3 run 0, for example, the final accepted target score is 0.86, while the historical best feasible score is 0.95. The full paired structures are shown in Figure A8(a–e), in the same task order.

Table A1: Raw computed or predicted properties for the ten states in Figure 4. All cases use run 0. Bold values violate the displayed threshold. CYP denotes CYP2D6 inhibition. MW is in Da and Caco-2 follows the predictor scale. R59 means the accepted state at round 59. Final means the final accepted state.

Task	State	MW	logP	QED	SA	CYP	hERG	Caco-2
Bound		≤ 500	≤ 5	≥ 0.3	≤ 6	≤ 0.5	≤ 0.5	≥ -5.15
GSK3β	R59	401.418	4.4596	0.3982	2.4764	0.0553	**0.5754**	**-5.5022**
GSK3β	Final	375.215	4.3756	0.6825	2.3569	0.4371	0.3496	-4.4525
JNK3	R59	378.407	4.0013	0.7018	2.7675	0.0638	**0.6968**	-4.9562
JNK3	Final	461.497	3.8926	0.4802	3.4456	0.0391	0.4993	-5.1188
DRD2	R59	386.511	4.0715	0.6044	2.9680	**0.9720**	**0.9889**	**-5.2292**
DRD2	Final	442.506	4.9369	0.5675	3.1650	0.4963	0.4977	-5.0343
Sitagliptin	R59	353.321	2.4220	0.7800	2.2937	0.1481	**0.6927**	-4.7337
Sitagliptin	Final	377.750	2.1280	0.8235	3.0148	0.0292	0.4788	-4.5046
Albuterol	R59	225.288	0.9159	0.6010	2.6694	0.1282	0.1020	**-5.4044**
Albuterol	Final	237.343	2.3761	0.7533	2.6696	0.1501	0.3435	-4.9597

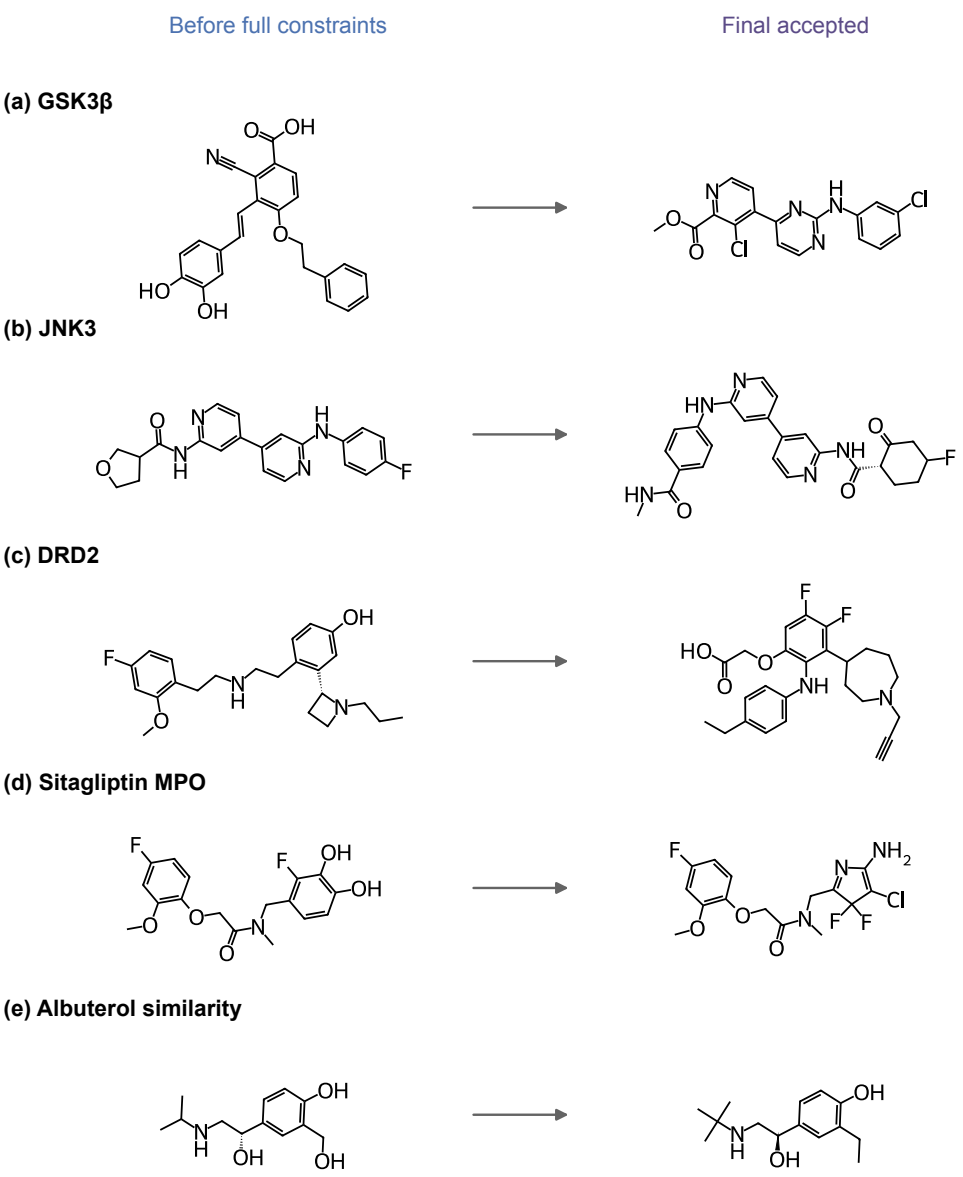

Figure A8: All five recorded accepted-state structure pairs from E2 underlying Figure 4. Panels (a–e) show GSK3β, JNK3, DRD2, Sitagliptin MPO, and Albuterol similarity. Every task uses run 0, comparing the state before full constraint activation (round 59) with the final accepted state. All structures use a common bond-length scale, atom-label size, and line width. Structures are drawn from original canonical SMILES. Target scores and violation counts appear in Figure 4. Raw properties are reported in Table A1.

D Online Baseline Adaptations and Detailed Comparisons

D.1 Shared Comparison Design

The six online strategies span all five baseline frameworks, with five tasks and five formal runs each (150 task/run cells). PharmAgent E2 contributes 25 reference cells. Protocols were fixed before observing the formal evaluation results. All seven constraints, their thresholds, and normalized violation scales match E2. Each unique molecule receives the same property prediction procedure. Graph-GA, REINVENT4, MOLLEO, and ExLLM reach 1,000 unique target evaluations per cell. SEISMO retains its original short-run protocol.

Figure A9(a–c) shows aggregate performance with uncertainty for every online strategy. Figure A9(d–f) shows the corresponding PharmAgent E2 advantage over each baseline. All 18 difference intervals are positive.

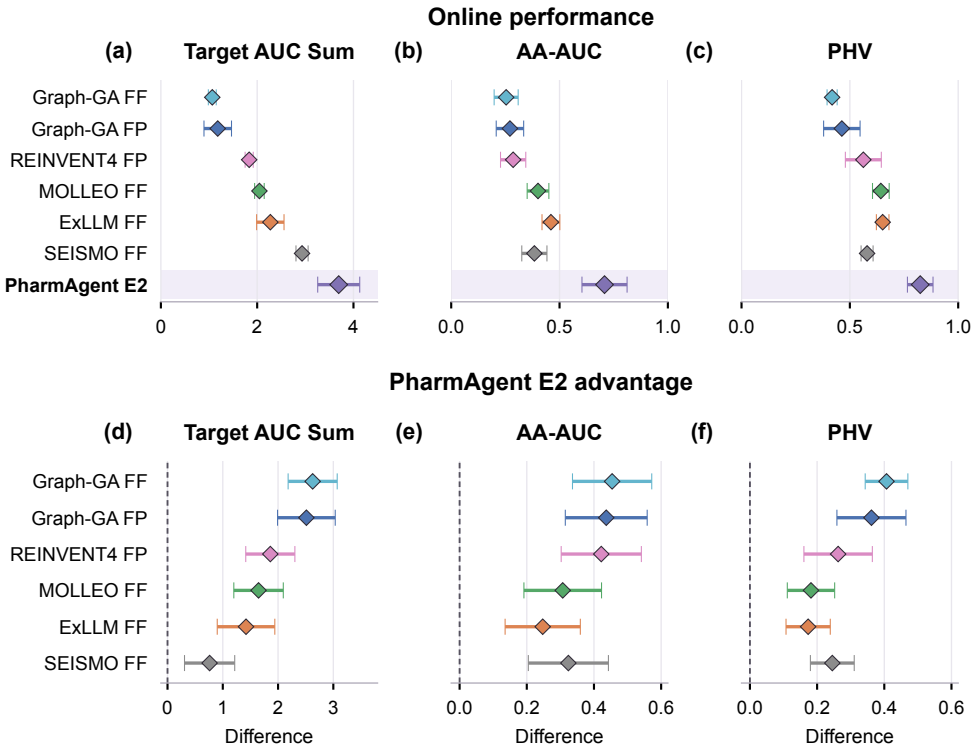

Figure A9: Online performance and comparative gains across five tasks and five runs. Panels (a–c) show five-run means with unadjusted 95% Student-t intervals. Panels (d–f) show PharmAgent E2 minus each baseline with unadjusted 95% independent-group intervals. Target AUC uses five-task sums, while AA-AUC and PHV use task means. Each interval uses five run-level aggregates per method. The zero line in (d–f) denotes equal aggregate performance. All six baseline strategies are retained. FF denotes feasibility-first and FP fixed-penalty search. SEISMO retains its short-run protocol and terminal-score carry-forward.

D.2 Optimizer and Language-Model Adaptations

Non-LLM adaptations. Graph-GA uses FF and FP, and REINVENT4 uses FP. All seven constraints enter rewards from the first evaluation. FP fixes penalty weight w within each run and uses reward $\sigma(4(f(x) - w\sum_i v_i(x) - 0.5))$. Graph-GA's FF reward is $0.5 + 0.5f(x)$ if all seven constraints pass, and $0.5\exp(-\sum_i v_i(x))$ otherwise. Metrics use raw target and property records, separately from optimizer rewards.

For each optimizer, $w \in \{0.1, 1, 10\}$ is selected by mean AA-AUC over all five tasks and two separate development runs (indices 100/101, 500 calls each). Ties within 10^{-12} use PHV then

smaller weight. Graph-GA selects 10 and REINVENT4 selects 1. All development outcomes are retained and excluded from formal means. Graph-GA uses population 120, offspring 70, mutation rate 0.067, and ZINC initialization. REINVENT4 uses its original prior, batch size 64, $\sigma = 128$, learning rate 10^{-4}, and an identical Murcko-scaffold diversity filter. All 75 non-LLM formal cells reach 1,000 unique evaluations. Invalid proposals and cache hits consume no new target calls. The 60 development runs consume 30,000 additional target evaluations. The 75 formal runs consume 75,000. These controls use no LLM calls, while target and ADMET measurements run on the shared server.

Table A2: Independent development runs for selecting fixed-penalty weights. Each row averages five tasks and two runs. Formal outcomes do not enter selection.

Optimizer	Weight	AA-AUC	PHV	Selected
Graph-GA	0.1	0.206613	0.349602	No
Graph-GA	1	0.209172	0.294164	No
Graph-GA	10	0.209456	0.290444	Yes
REINVENT4	0.1	0.257339	0.489099	No
REINVENT4	1	0.286928	0.542494	Yes
REINVENT4	10	0.262841	0.506246	No

LLM feasibility-first adaptations. MOLLEO, ExLLM, and SEISMO retain MiniMax-M3, their task descriptions, and response formats. Prompts add the common constraint objective and relevant parents' or historical molecules' property values, thresholds, normalized violations, and feasibility. These adaptations use no PharmAgent trajectories, Chemical-CER, adaptive multipliers, or curriculum. For MOLLEO and ExLLM population selection, feasible candidates precede infeasible candidates. Feasible candidates are ordered by decreasing raw target score and infeasible candidates by increasing total normalized positive violation. Canonical SMILES breaks ties. MOLLEO retains roulette parent sampling using a positive feasibility reward and uses this lexicographic ordering for survivors. Its population size is 120, offspring count 70, and mutation rate 0.067. ExLLM retains uniform parent sampling, a population of 50, two proposals per call, and experience injection probability 0.5. Its high/low evidence groups use the same feasibility-first order. Crossover and mutation probabilities are 0.8 and 0.2. SEISMO receives the same feasibility-first priority through its conversational objective and measured constraint feedback, retaining its original proposal and decision process. The hosted model uses temperature 0.7 and at most 2,048 completion tokens per request. Run indices label repeated runs rather than deterministic hosted-model seeds. Functional pilots are excluded from all reported means.

D.3 Budget and Measurement Accounting

SEISMO uses 51 generation transitions, the last an unscored summary, and at most 50 target-evaluation events. Repeated target evaluations count as events, while property predictions are cached by canonical molecule. After termination, the final Top-10 and historical best feasible scores are carried forward to the common 1,000-call horizon without further evaluations or molecules. Table A3 reports unique evaluation counts and evaluation events.

Table A3: Target-evaluation counts across 25 task/run cells per strategy. Ranges are per cell. FF denotes feasibility-first and FP fixed-penalty search. Repeated SEISMO evaluations add events without adding unique molecules. All AUCs use the same 1,000-call horizon.

Method	Unique evaluations	Evaluation events	Total unique
Graph-GA (FF)	1,000	1,000	25,000
Graph-GA (FP)	1,000	1,000	25,000
REINVENT4 (FP)	1,000	1,000	25,000
MOLLEO (FF)	1,000	1,000	25,000
ExLLM (FF)	1,000	1,000	25,000
SEISMO (FF)	6–47	37–50	955
PharmAgent (E2)	1,000	1,000	25,000

D.4 Task-Level Margins Against Every Online Strategy

Figure A10(a) shows positive target-AUC differences in all 30 method–task comparisons. The smallest is 0.0073 against SEISMO on DRD2. Figure A10(b) shows 27 positive AA-AUC differences. The exceptions are MOLLEO and ExLLM on Sitagliptin and SEISMO on Albuterol. Figure A10(c) shows 29 positive PHV differences, with a 0.0004 reversal against MOLLEO on Albuterol. Each cell subtracts two five-run task means. The counts summarize these task-level mean comparisons. Figures A2–A4 provide all runs and score intervals.

(a) Target AUC

	GSK3β	JNK3	DRD2	Sitagliptin MPO	Albuterol similarity
Graph-GA FF	+0.4447	+0.6507	+0.8281	+0.1645	+0.5382
Graph-GA FP	+0.3977	+0.6458	+0.7791	+0.1516	+0.5385
REINVENT4 FP	+0.2991	+0.5874	+0.2423	+0.1743	+0.5569
MOLLEO FF	+0.3775	+0.5979	+0.5059	+0.0828	+0.0830
ExLLM FF	+0.2712	+0.5782	+0.3997	+0.0490	+0.1224
SEISMO FF	+0.1303	+0.4503	+0.0073	+0.1630	+0.0109

(b) AA-AUC

	GSK3β	JNK3	DRD2	Sitagliptin MPO	Albuterol similarity
Graph-GA FF	+0.3840	+0.5930	+0.7192	+0.0782	+0.4958
Graph-GA FP	+0.3649	+0.5916	+0.6899	+0.0468	+0.4895
REINVENT4 FP	+0.2341	+0.5027	+0.7747	+0.0810	+0.5162
MOLLEO FF	+0.3247	+0.5562	+0.6143	-0.0264	+0.0663
ExLLM FF	+0.2475	+0.5150	+0.3916	-0.0193	+0.1023
SEISMO FF	+0.1202	+0.5761	+0.8769	+0.0764	-0.0305

(c) PHV

	GSK3β	JNK3	DRD2	Sitagliptin MPO	Albuterol similarity
Graph-GA FF	+0.4272	+0.6556	+0.3764	+0.1287	+0.4454
Graph-GA FP	+0.4024	+0.6404	+0.2597	+0.0820	+0.4243
REINVENT4 FP	+0.2136	+0.4008	+0.1431	+0.1165	+0.4389
MOLLEO FF	+0.3206	+0.4848	+0.0806	+0.0231	-0.0004
ExLLM FF	+0.2506	+0.5286	+0.0826	+0.0044	+0.0005
SEISMO FF	+0.2842	+0.5562	+0.1583	+0.2192	+0.0084

PharmAgent E2 − baseline (signed fourth-root color scale)

−1.00 −0.10 −0.01 0.00 0.01 0.10 1.00

Figure A10: PharmAgent E2 minus every online baseline on each task. Panels (a–c) show target AUC, AA-AUC, and PHV. Each cell is a difference between five-run means, calculated before rounding. Positive values favor PharmAgent. A shared signed color scale uses purple for positive differences and orange for negative differences. Cell fills and the color bar follow the same signed fourth-root scale, which makes small signed differences visible. Printed values are rounded to four decimals. All 90 comparisons are retained. Target AUC is reported per task. Colors encode mean differences, with uncertainty reported in Figures A2–A4. FF denotes feasibility-first and FP fixed-penalty search.

D.5 Changes from Original to Online Search

Figure A11 complements the direct comparison by showing how each framework's outcomes change after online constraint handling is introduced. Each baseline adaptation is compared with that framework's original run configuration. PharmAgent E2 is compared with E1. The original trajectories receive the same post-hoc property evaluation, enabling comparison of AA-AUC and PHV on a common scale.

The three panels separate changes in target progress (Figure A11(a)), feasible-solution progress (Figure A11(b)), and two-objective coverage (Figure A11(c)). ExLLM's AA-AUC increases from 0.3121 to 0.4602, with a difference interval of $[0.1033, 0.1929]$, while its target-AUC sum decreases by 0.5218. PharmAgent's AA-AUC increases from 0.5328 to 0.7076. Its independent-group difference interval is $[0.0195, 0.3303]$, consistent with the run-index-matched comparison in the main paper. Its target-AUC-sum change of -0.2272 has interval $[-0.7880, 0.3336]$.

The Graph-GA adaptations lower the three aggregate means under the tested configurations. REINVENT4 and SEISMO increase all three means, although their difference intervals cross zero. MOLLEO's changes in AA-AUC and PHV also have intervals spanning zero. These framework-specific responses show that access to constraint measurements and effective use of those measurements are distinct experimental questions. The direct E2 comparison in Table 3 evaluates the resulting online search methods.

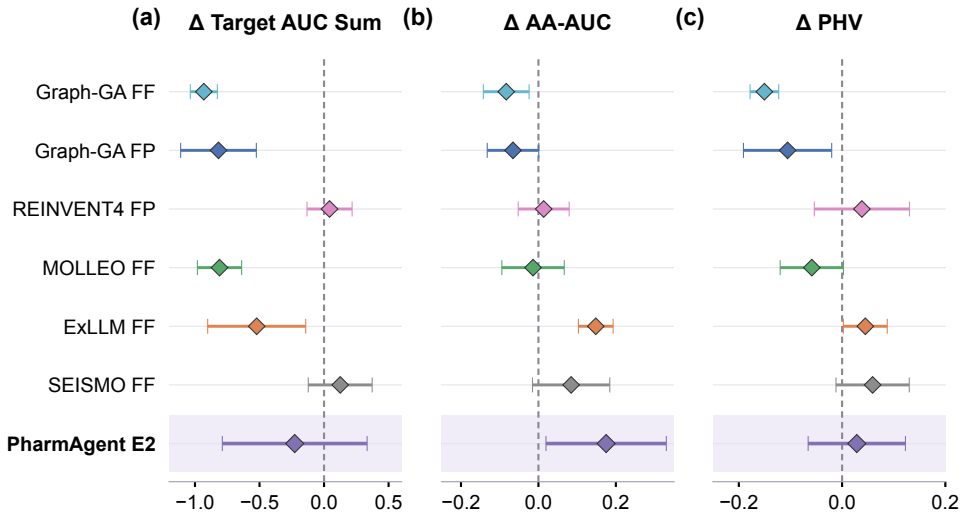

Figure A11: Online minus original aggregate performance within each framework. Panels (a–c) show changes in target-AUC sum, AA-AUC, and PHV. Diamonds denote mean differences and bars unadjusted 95% independent-group intervals over five run-level task aggregates per configuration. Positive changes favor online search. Graph-GA's original reference is shared by its two adaptations. PharmAgent compares E2 with E1, using the independent-group sensitivity analysis. Each contrast compares complete original and online configurations. FF denotes feasibility-first and FP fixed-penalty search. SEISMO retains its short-run protocol in both conditions.

All five tasks and all five runs are retained for every contrast. Target-AUC changes use task sums, whereas AA-AUC and PHV use task means. The uncertainty calculation follows Appendix B and does not assume shared hosted-model random draws. The 21 displayed intervals are individual descriptive intervals without multiplicity adjustment. Both Graph-GA contrasts share an original reference, so they are not independent replications.

E ADMET-PMO Protocol Details

QED and SA follow Bickerton et al. (2012) and Ertl & Schuffenhauer (2009). Learned endpoints use ADMET-AI (Swanson et al., 2024).

Table A4: ADMET constraints in the ADMET-PMO benchmark. The same seven constraints are used for all five tasks evaluated in this paper.

Constraint	Property	Type	Threshold	Tool
Lipinski MW	Molecular weight	$\leq$	500 Da	RDKit
Lipinski logP	Lipophilicity	$\leq$	5.0	RDKit
QED	Drug-likeness	$\geq$	0.3	RDKit
SA Score	Synth. accessibility	$\leq$	6.0	RDKit
CYP2D6 inhibition	Metabolic safety	$\leq$	0.5	ADMET-AI
hERG blocking	Cardiac safety	$\leq$	0.5 probability	ADMET-AI
Caco-2	Permeability	$\geq$	-5.15	ADMET-AI

E.1 Consistent Evaluation Protocol

Online ADMET-PMO measurements and post-hoc baseline annotations follow the same evaluation procedure. Each molecule is represented by its RDKit-canonical SMILES (Weininger, 1988; Landrum et al., 2024) and evaluated individually. Predictions are stored in a cache of up to 10,000 molecules, with the least recently used entries removed when capacity is reached. This procedure gives consistent predictions across input batch sizes, molecule orderings, and prior cache contents. Cached results are copied before reuse. The evaluation protocol fixes backend versions, endpoint mappings, constraints, normalization scales, and failure handling. Online measurements and post-hoc baseline annotations are checked for consistency with this protocol before results are aggregated.

E.2 Property-Specific Normalization Scales

Constraint violations are normalized by property-specific scales so that the sigmoid transition zone is comparable across properties with vastly different natural ranges:

Table A5: Property-specific normalization scales used in the ADMET score and Lagrangian penalty terms.

Property	Scale r_i	Steepness s_i	Justification
MW	150 Da	5.0	Std. dev. in drug-like space
logP	2.0	5.0	Std. dev. in drug-like space
QED	0.2	5.0	Range $[0, 1]$, std ≈ 0.2
SA Score	2.0	5.0	Range $[1, 10]$, std ≈ 2
CYP2D6 inh.	0.3	5.0	Probability scale $[0, 1]$
hERG blocking	0.3	5.0	Probability scale $[0, 1]$
Caco-2 perm.	1.0	5.0	Log permeability units

E.3 Continuous ADMET Score

For full-constraint evaluation, the ADMET Analyst computes a continuous score by applying a sigmoid to each normalized distance from its threshold:

$$\text{ADMET}_{\text{score}}(x) = \frac{1}{\sum_i w_i} \sum_{i=1}^{K} w_i \cdot \sigma\left(\frac{s_i \cdot (\tilde{t}_i - \tilde{g}_i(x))}{r_i} \right), \tag{A1}$$

where $\sigma(\cdot)$ is the logistic sigmoid, s_i is the steepness, r_i is a property-specific normalization scale (Table A5), and w_i are configurable per-constraint weights. The direction-normalized margin treats upper- and lower-bound requirements consistently. This continuous score supports experience annotation and Pareto evaluation. Hard feasibility requires all seven inequalities, and candidate acceptance uses (3). The proposal model does not use gradients of the sigmoid score. Prompt exposure follows the active constraints. Phase 1 and Standard PMO use target-only Generator and Reviewer templates.

E.4 Formal Metric Definitions

Let $\{(x_j, s_j)\}_{j=1}^{N}$ denote scored molecules, sorted by s_j descending. Let $\text{pass}(x) = \mathbb{1}[\forall i\colon \tilde{g}_i(x) \leq \tilde{t}_i]$.

Pass rate among the top k molecules (PR@k). PR@$k = \frac{1}{k}\sum_{j=1}^{k} \text{pass}(x_j)$.

AA-AUC. Let $s^*_{\text{pass}}(n)$ be the best score among passing molecules after n oracle calls. AA-AUC $= \frac{1}{B}\sum_{n=1}^{B} s^*_{\text{pass}}(n)$.

PHV. Using reference point $\mathbf{r} = (0, 0)$: PHV $= \text{vol}\big(\bigcup_{j\in\mathcal{P}}[\mathbf{r}, (s_j, a_j)]\big)$, where $\mathcal{P}$ is the Pareto front of (s_j, a_j) pairs and a_j is the ADMET score of molecule j.

F Prompt Templates

F.1 Generator Prompts

The prompts request structured JavaScript Object Notation (JSON) responses.

```
You are PharmAgent-Generator, an expert medicinal chemist
AI. You optimize molecules by proposing specific structural
modifications with detailed pharmaceutical reasoning.
Rules: 1. Output ONLY valid JSON. 2. Every proposed SMILES
must be chemically valid. 3. Each modification must have
explicit medicinal chemistry rationale. 4. Consider only
the explicitly active declared constraints. 5. Output a
JSON object with key "candidates" containing an array.
```

Target-only and constrained branches. The quoted system prompt is used in constrained phases. Standard PMO and curriculum Phase 1 instead use an explicit target-only system prompt that instructs the Generator to optimize only the stated target and omits all ADMET fields. Both user-prompt branches receive the current SMILES and target score, ChEMBL matched-pair context from MedChem-RAG, previous Reviewer feedback, structured round-memory records, and the top-5 Chemical-CER experiences. Only the constrained branch adds the current active-constraint ADMET summary and two explicitly separated blocks containing current measurements (value, threshold, direction, signed margin, and normalized violation) and historical multiplier pressure λ_i. The multiplier block records optimization history, while current measurements determine pass/fail status. Pan-assay interference compounds (PAINS), Brenk, and Veber alerts appear in neither Generator branch and are never optimization guidance. If evaluated, they are post-hoc analyses outside the formal online state. Every Generator request asks for at most five candidates.

Additional proposal instructions. When a restart is scheduled, the base prompt receives one proposal-strategy suffix (scaffold-hop, ring/linker re-architecture, or bioisosteric replacement). The suffix explicitly preserves the incumbent, supplied constraints, RAG evidence, and run history. If and only if the primary response yields no valid molecules absent from the evaluated set, a second and final Generator request appends standardized rejection or duplication reasons and candidate indices and up to 20 previously evaluated canonical SMILES to avoid ordered by Morgan-Tanimoto similarity to the incumbent. The second request preserves the base objective, active constraints, Verifier, and five-candidate cap. Cross-round and cross-agent reuse is restricted to structured fields comprising canonical SMILES, finite scores and deltas, active-constraint facts, validated modification types, and normalized Reviewer priorities. Raw model prose is not forwarded as replay context. RAG exposes canonical structure pairs and finite target-score deltas. CER exposes canonical pairs and finite numerical deltas.

Phase-3 context-only repair anchors. After five consecutive non-improving rounds within the phase in Phase 3, each Generator request already permitted in that activated round receives the current archive of at most two infeasible Pareto anchors. This applies to the primary request and, when no valid novel candidate is produced, the single repair request. The same archive is supplied to the Reviewer when the round produces scored candidates. The highest-target and lowest-total-normalized-violation representatives are selected among observations that strictly outperform the best feasible target score. Each anchor is accompanied by its SMILES and per-constraint value, threshold, direction, signed margin, and normalized violation. The Generator is instructed to edit only the incumbent. The Verifier does not enforce structural similarity between the incumbent and a

proposal. The Reviewer receives the accepted state (the improving candidate or retained incumbent) as the incumbent for the next round. Anchors provide measured examples for proposal revision. Acceptance, multiplier updates, and Chemical-CER recording follow their existing rules. CER stores the transition to the best evaluated candidate of the round rather than to the accepted iterate. Thresholds, metrics, the Verifier, transition rules, candidate and request caps, and the oracle budget remain fixed. Through their influence on proposals, anchors can affect accepted states, multiplier updates, replay records, and the sequence of model requests, network attempts, and oracle evaluations. They add bounded input tokens to those existing prompts.

F.2 Reviewer Prompts

```
You are PharmAgent-Reviewer, a senior medicinal chemistry
advisor.  You analyze optimization results and provide
strategic guidance.

Rules:  1. Output ONLY valid JSON. 2. Be specific in
suggestions:  name exact structural modifications.  3.
Consider both activity and ADMET properties.  4. Recommend
termination if optimization has stagnated or goals are met.
```

The quoted prompt is used only in constrained phases. The target-only branch omits ADMET fields and directs the Reviewer not to use secondary properties. For constrained rounds, every displayed candidate is paired with that candidate's own per-constraint value, threshold, direction, signed margin, and normalized violation. The incumbent's values are never substituted for a candidate's values, and historical λ_i pressure remains a separate block. When anchors are active, the Reviewer receives the same selected anchors as the Generator and uses the accepted state as the incumbent for the next round. In formal runs with the 1,000-call budget, early termination is disabled, so the Reviewer is called at most once after a round scores at least one novel candidate and its termination recommendation cannot truncate the oracle budget.

G Chemical-CER Algorithm Details

Scores recorded with each transition. For a record collected in round τ, let J_τ denote that round's scalar evaluation rule, with all its parameters fixed. The record stores

$$\Delta_{\text{target}} = f(x_{\text{tgt}}) - f(x_{\text{src}}), \qquad \Delta_{\text{combined}} = J_\tau(x_{\text{tgt}}) - J_\tau(x_{\text{src}}).$$

For the constrained full system, $J_\tau = \mathcal{L}_{\mathcal{C}_\tau}(\cdot, \boldsymbol{\lambda}^{(\tau)}, \rho^{(\tau)})$. The target-only objective gives $J_\tau = f$. The stored ADMET change describes the annotation used at collection time. An entry remains an evaluated transition even when the candidate was rejected. Eligibility for later retrieval is a separate operation.

Eligibility before similarity. The retrieval interface distinguishes positive target-score change from positive collection-objective change. Let $h_t(e)$ be the retrieval eligibility rule, determined by the improvement criterion and the permitted origin phases. Then $\mathcal{D}_t^{\text{eligible}} = \{e \in \mathcal{D} : h_t(e) = 1\}$. Replay uses target-only information in Phases 1–2 and combined-objective information for all seven constraints in Phase 3. Phase 1 exposes records originating before round 20 and Phase 2 before round 60. Phase 3 has no origin-round cutoff. Retrieval uses completed records from earlier tasks and rounds of the current task. It excludes attempts from the ongoing round and incomplete records from failed rounds. Historical task text and modification prose are not replayed. The predicate is applied to recorded outcomes before source-structure distances are ranked. Historical scalar changes retain their collection-time objective and multiplier values during retrieval.

The database stores fingerprints as a flat NumPy array and computes all distances in a single vectorized operation, avoiding the need for an external index (such as FAISS) at the typical scale of $<$1,000 entries. The database persists across tasks through an optional serialization mechanism. When the number of stored experiences exceeds the configured maximum (1,000 by default), the oldest entries are evicted in first-in, first-out (FIFO) order.

Algorithm A1 Chemical-CER: structure-aware experience retrieval

Require: Experience database $\mathcal{D} = \{e_1, \ldots, e_N\}$, query molecule x, integer k, eligibility predicate h_t
Ensure: Retrieved experiences $\mathcal{E}_k$
1: $\mathbf{fp}_x \leftarrow \textsc{MorganFP}(x, \text{radius} = 2, \text{bits} = 2048)$
2: $d_j \leftarrow \|\mathbf{fp}_x - \mathbf{fp}_{\text{src}}(e_j)\|_2$ for all $j = 1, \ldots, N$ {Vectorized}
3: $\mathcal{S} \leftarrow \{e_j \in \mathcal{D} : h_t(e_j) = 1\}$ {Apply configured eligibility}
4: $\mathcal{E}_k \leftarrow$ first $\min(k, |\mathcal{S}|)$ elements of $\mathcal{S}$ after stable sorting by ascending d_j
5: Format canonical structure pairs and finite numeric deltas for proposal conditioning
6: **return** $\mathcal{E}_k$

H Computational Resources and Reproducibility

Model and scientific evaluation. Generator and Reviewer access the provider-hosted MiniMax-M3 model through MiniMax's OpenAI-compatible application programming interface (API) in China (MiniMax, 2026). Provider-side model versioning and sampling are addressed in Appendix M. Run indices 0–4 identify repetitions and seed local random operations where applicable. ADMET measurements use ADMET-AI v1.4.0 with RDKit (Swanson et al., 2024; Landrum et al., 2024). Each uncached canonical molecule is evaluated individually and reused through the property cache.

Requests, rounds, and recovery. The formal target budget is 1,000 unique calls per task/run, including the 26-member seed pool. Each primary Generator request proposes at most five candidates. A response with no valid novel molecules can trigger one final repair request with the same cap. A scored round invokes at most one Reviewer. Both E1 and E2 use a cap of 60,100 Hypertext Transfer Protocol (HTTP) attempts per task/run. Observed round counts range from 458 to 3,592 for E1 and 538 to 6,616 for E2. A round may contain no novel molecule, so rounds, HTTP attempts, and unique oracle evaluations are different quantities. All E1/E2 cells reach 1,000 unique target calls. No performance-based early stopping is used. The reported experiments use the request and round limits specified here.

Bounded novelty repair and proposal-context restarts preserve the incumbent and target budget. After five consecutive non-improving rounds within the phase in Phase 3, up to two previously scored near-feasible anchors may enter existing prompts. They supply context without directly replacing the incumbent or changing the acceptance rule. The archive is evaluated separately from Chemical-CER, whose record remains the incoming-incumbent-to-round-best transition. Stage changes occur only at rounds 20 and 60, and the penalty reset retains multipliers. Appendix F describes the structured fields passed through recovery and feedback.

Recorded cost. Table A6 reports original provider usage across five tasks and five runs per arm. Token totals measure recorded model usage. E2 uses 2.408 times E1's recorded total tokens. Post-hoc E1 annotations reuse the 25,000 molecules from its existing generation runs.

Table A6: Recorded E1/E2 compute over 25 runs per arm. Oracle calls count unique target evaluations. HTTP attempts include retries. Round counts include rounds without new molecules.

Quantity	E1	E2
Unique target evaluations	25,000	25,000
Recorded optimization rounds	38,013	49,854
Successful HTTP calls	83,653	108,055
HTTP attempts	83,751	108,168
Input tokens	119,050,374	301,475,520
Output tokens	55,388,732	118,552,476
Total tokens	174,439,106	420,027,996

The non-LLM online controls use 75,000 formal and 30,000 development target evaluations, with no LLM calls. MOLLEO FF and ExLLM FF contribute 25,000 unique formal target evaluations each. SEISMO FF evaluation counts appear in Table A3. Oracle evaluation totals and LLM token/time costs are accounted for separately. Graph-GA and REINVENT4 use the same server-side

target/ADMET measurement environment. REINVENT4 optimization can execute locally through that measurement service. Both optimizers therefore obtain measurements from the same evaluation service. Missing baseline time/token measurements are left unreported. Appendix M discusses computational-cost comparability.

Finite controller observations. All 25 E2 runs terminate with accepted molecules passing the seven final constraints. During Phase 3, the mean run-level fraction of feasible accepted rounds is 95.62%. This includes repeated incumbents during rounds without new molecules. It differs from the 29.13% feasible fraction among evaluated molecules. The largest observed multiplier is 1,815.5313. Both E1 and E2 find an initially feasible seed at call 1. AA-AUC and best-feasible-score curves therefore track improvement beyond initial feasibility. These curves report finite-budget performance. Trajectory analyses include every task and run, distinguishing violations of active constraints from violations under the full set of seven constraints.

I HYPERPARAMETER CONFIGURATION

Table A7 records the default controller, retrieval, scheduling, and recovery settings.

Table A7: Hyperparameters for Lagrangian, Chemical-CER, and Curriculum modules.

State / operation	Parameter	Default
Constraint state	Initial penalty ρ_0	1.0
	Growth factor γ	1.5
	Maximum penalty $\rho_{\max}$	100
Experience state	Retrieval count k	5
	Archive capacity	1000
	Fingerprint radius / bits	2 / 2048
Schedule state	Phase 1 end R_1	20
	Phase 2 end R_2	60
Context recovery	Stagnation trigger within each phase	5

The initial ρ_0 sets the quadratic penalty when constraints activate or the active set expands. The growth factor γ increases this penalty after a violated accepted state, subject to the cap $\rho_{\max}$. Feasible accepted states leave it unchanged. The retrieval count limits the number of examples exposed to the Generator, and archive capacity limits retained experience. The curriculum changes only at its fixed round boundaries. The stagnation trigger controls proposal context and does not activate constraints.

J BROADER IMPACT

PHARMAGENT is a computational tool for molecular optimization and does not directly produce physical molecules. Its outputs are SMILES strings that must be validated by medicinal chemists and tested in wet-lab assays before entering any drug development pipeline. By incorporating ADMET constraints, the framework is designed to steer optimization toward molecules with better safety profiles, which could reduce the number of toxic candidates that proceed to animal testing.

A potential risk of any molecular generation system is the production of molecules with harmful properties (toxins, controlled substances, environmental pollutants). The implemented screen penalizes predicted hERG inhibition. It does not include genotoxicity endpoints or establish safety against these broader risks. Nonetheless, the system is not a substitute for regulatory toxicology assessment, and users should apply additional safety checks before any experimental synthesis.

The ADMET-PMO benchmark itself poses minimal risk, as it consists of computational oracle functions applied to virtual molecules. Releasing the benchmark and evaluation code promotes transparency and reproducibility in the field.

K Comparison of Evaluated Baselines

Table A8 compares the original five baseline implementations along six axes. The online adaptations in Table 3 enable online measurements and constraint-aware selection or feedback. Their protocol is specified in Appendix D. Its original-method roster matches Tables 2 and A10. The comparison describes the configurations evaluated in this study.

Table A8: Features enabled in the evaluated five-baseline protocol. ✓ indicates enabled features and ✗ indicates features not enabled. The symbol ❍ denotes within-task memo or trajectory reuse without fingerprint-based experience retrieval. Online ADMET columns concern the optimization loop. All five baselines separately receive the same post-hoc ADMET annotation. † denotes our independent MiniMax-M3 reimplementations.

System	Multi-agent	Online ADMET evaluation	Online ADMET constraints	Memory retrieval	Cross-task transfer	Five-task PMO
Graph-GA	✗	✗	✗	✗	✗	✓
REINVENT4	✗	✗	✗	✗	✗	✓
MOLLEO-MiniMax-M3†	✗	✗	✗	✗	✗	✓
ExLLM-MiniMax-M3†	✗	✗	✗	❍	✗	✓
SEISMO-MiniMax-M3	✗	✗	✗	❍	✗	✓
PharmAgent	✓	✓	✓	✓	✓	✓

Graph-GA and REINVENT4 optimize the target-only objective in our baseline runs. REINVENT4's broader configurable scoring capabilities are not enabled as online ADMET constraints in this experiment. MOLLEO-MiniMax-M3† and ExLLM-MiniMax-M3† are independent reimplementations. SEISMO-MiniMax-M3 preserves the released prompts and original short-run PMO protocol with a MiniMax-M3 backend (Krüger et al., 2026, Section 5.1 and Appendix A). The protocol permits at most 50 oracle evaluations. Our runs produce 27–48 unique target evaluations after molecular canonicalization. After optimization ends, the final Top-10 score is held constant through the common 1,000-call AUC horizon, following the original paper's reporting convention. This carry-forward is used only for metric integration and creates no additional oracle evaluations. Post-hoc annotation evaluates the original generated trajectories through the common property protocol in Table A10. PharmAgent uses property measurements during search to guide subsequent decisions. Experience retrieval and cross-task transfer refer specifically to retrieval and reuse of stored optimization attempts. The main results quantify the performance advantage.

Table A9: State and memory mechanisms in the five baseline frameworks and Chemical-CER. Framework names omit the LLM backend. Implementation details are specified in the experimental setup. Within-task state includes populations, learned policy parameters, and explicit history. Structure-aware denotes fingerprint-based retrieval of optimization experiences. Cross-task denotes reuse of an optimization experience archive in the evaluated configurations.

Method	Within-task state	Structure-aware	Cross-task
Graph-GA	Evolutionary population	✗	✗
REINVENT4	Learned policy parameters	✗	✗
MOLLEO	Population / parent scores	✗	✗
ExLLM	Experience memo	✗	✗
SEISMO	Interaction trajectory	✗	✗
Chemical-CER (ours)	Retrieved transitions	✓	✓

Complete aggregate results. Table A10 reports the original-protocol property comparison. Table A11 gives final target scores and uncertainty for both original and online configurations.

Table A10: Original baselines with post-hoc property annotation, compared with online PharmAgent E2. Results average five tasks and five runs. The 5-Task Sum adds task-level AUC means. All metrics are higher-is-better. Baselines receive post-hoc ADMET annotation. PHARMAGENT uses online constraints. Backend and evaluation protocols follow Section 4.1. Best/second-best means are bold/underlined.

	Target Score			Property-Aware Quality	
Method	Top-1	AUC Top-10	5-Task Sum	AA-AUC	PHV
Graph-GA	0.6029	0.3998	1.9988	0.3367	0.5686
REINVENT4	0.5864	0.3582	1.7911	0.2723	0.5238
MOLLEO†	<u>0.7797</u>	<u>0.5713</u>	<u>2.8565</u>	<u>0.4145</u>	<u>0.7016</u>
ExLLM†	0.7364	0.5590	2.7950	0.3121	0.6065
SEISMO	0.6053	0.5613	2.8067	0.2994	0.5204
PHARMAGENT	**0.8514**	**0.7387**	**3.6936**	**0.7076**	**0.8248**

Table A11: Final target-score summaries for all configurations across five fixed tasks and five runs. Means ± unadjusted 95% interval half-widths. The lower block contains online adaptations. FF denotes feasibility-first and FP fixed-penalty search. Best/second-best means are bold/underlined across all rows. *SEISMO follows its original short-run protocol.

Method	Top-1	Top-10
PharmAgent E1	$\mathbf{0.8641} \pm 0.0645$	$\mathbf{0.8606} \pm 0.0634$
PharmAgent E2	$\underline{0.8514} \pm 0.0561$	$\underline{0.8305} \pm 0.0751$
Graph-GA	0.6029 ± 0.0194	0.5498 ± 0.0157
REINVENT4	0.5864 ± 0.0469	0.4613 ± 0.0264
MOLLEO†	0.7797 ± 0.0401	0.7484 ± 0.0406
ExLLM†	0.7364 ± 0.0565	0.7184 ± 0.0561
SEISMO*	0.6053 ± 0.0512	0.5733 ± 0.0432
Graph-GA (FF)	0.4375 ± 0.0141	0.3219 ± 0.0257
Graph-GA (FP)	0.4837 ± 0.0852	0.3575 ± 0.0981
REINVENT4 (FP)	0.6221 ± 0.0901	0.4663 ± 0.0197
MOLLEO† (FF)	0.6788 ± 0.0517	0.5864 ± 0.0290
ExLLM† (FF)	0.6647 ± 0.0321	0.6352 ± 0.0320
SEISMO* (FF)	0.6405 ± 0.0338	0.5978 ± 0.0261

L SEED MOLECULE POOL

For seed selection (Section 4), we use a pool of 26 diverse molecules covering simple aromatics, heterocycles, approved drugs, bicyclic scaffolds, aliphatic chains, and common drug motifs:

```
c1ccccc1  c1ccncc1  c1ccoc1  c1ccsc1  c1cc[nH]c1
C1CCNCC1  C1CCOCC1  C1CCSCC1
CC(=O)Oc1ccccc1C(=O)O   CC(C)Cc1ccc(cc1)C(C)C(=O)O
CC(=O)Nc1ccc(O)cc1      CC(C)NCC(O)c1ccc(O)c(O)c1
c1ccc2ncccc2c1  c1ccc2c(c1)[nH]c1ccccc12  c1ccc2c(c1)cc1ccccc1n2
CCCCCCCC(=O)O  CC(C)(C)C(=O)O  C1CCC(CC1)O  CC(C)(C)O
O=C(O)c1ccccc1O  c1ccc(cc1)O  NC(=O)c1ccccc1
c1ccc(cc1)S(=O)(=O)N  CC(=O)c1ccccc1  C1COCCN1  C1CNCCN1
```

At the start of each run, all 26 molecules are scored by the target oracle (at most 26 calls, <3% of the 1,000-call budget). The highest-scoring molecule initializes optimization, reducing dependence on a single arbitrary seed at low oracle cost.

M LIMITATIONS AND REMAINING VALIDATION

PHARMAGENT leads all five E2 aggregate metrics, with higher means in 30/30 target-AUC, 27/30 AA-AUC, and 29/30 PHV task-level comparisons (Figure A10). The few online exceptions concern Sitagliptin and Albuterol, including a PHV reversal of 0.0004. In E1, SEISMO leads DRD2 and Albuterol. E2's target-AUC sum is 3.6936 versus E1's 3.9208, and their target-AUC and PHV difference intervals cross zero.

Evaluation covers five tasks and predicted properties. Experimental safety validation and cross-domain transfer remain open. All methods receive the same oracle allowance, with actual evaluations and computation varying by protocol. SEISMO retains short-run optimization and terminal-score carry-forward for AUC integration. Online adaptations were designed after the original PharmAgent runs and fixed before formal outcomes. Hosted-model versions and sampling are provider-controlled rather than fixed by local run labels. We plan to extend evaluation to open-weight language models as computational resources permit.

Proposition 1 assumes uniformly bounded multipliers, which is not established for the search and cannot be verified from finite runs. The analysis does not guarantee finite-budget feasibility, optimality, acceleration, or stationarity.

Intervals are unadjusted. Molecular cases describe fixed endpoints rather than isolated scheduling effects, and original-to-online contrasts compare complete configurations.