LLM-TO-FLOW DISTILLATION: TEACHING NATURAL LANGUAGE GOALS TO SCIENTIFIC GENERATORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative discovery in science requires coupling high-level scientific reasoning with the ability to generate samples respecting the structure of a design space. Yet these capabilities remain largely separated: foundation LLMs encode broad chemical and biological knowledge, while domain-specific flow- and diffusion-based generators capture the priors needed to produce valid sequences and geometries. We study how to align such scientific generators to high-level natural-language discovery goals, without requiring an explicit proxy reward, which is often unavailable or poorly captured by learned predictors in experimental sciences. We introduce **L**LM-to-**F**low **D**istillation (LFD), in which a frozen LLM acts as an adaptive teacher for a domain-specific generator. Given a goal such as *'generate peptides with promising anticancer activity'*, the teacher constructs a curriculum of intermediate sub-goals aware of the flow's current capabilities, and progressively directs them toward the original discovery goal. At each round, the LLM judges sampled candidates under the current objective. These judgments induce preferences that are iteratively distilled into the generator through DPO. This provides a dense learning signal even when the initial domain-specific model places negligible probability on designs satisfying the final goal, enabling adaptation toward poorly covered, out-of-distribution regions. Experiments on therapeutic peptides and small molecules structure design demonstrate that LFD enables a new form of generative control: aligning arbitrary domain-specific scientific generators to scientific discovery tasks expressed via high-level natural-language goals.

1 INTRODUCTION

Generative discovery in experimental sciences requires combining two complementary capabilities. The first is broad scientific knowledge and high-level reasoning about chemical and biological systems, e.g., which mechanism drives a disease and which molecules may act on it. The second is the ability to generate valid designs from low-level, domain-specific priors that can capture the structure of a design space, such as which amino-acid sequences fold and remain soluble, or which crystal geometries are plausible. Today, these capabilities are often split across two families of models: LLMs, which absorb the scientific literature yet cannot reliably produce valid designs in several domains (Sriram et al., 2024; Khastagir et al., 2026; Huang et al., 2026), and domain-specific generative models, such as flow (Lipman et al., 2023; 2024) and diffusion models (Sohl-Dickstein et al., 2015; Song & Ermon, 2019). Training the latter on scientific designs allows them to capture complex structural and geometric priors, but they lack both the high-level scientific knowledge and the natural language grounding required to interpret a scientist's discovery goal. An LLM's knowledge might be actionable in a discovery task in two key ways. It can *concretize* a goal, turning a high-level prompt such as *'generate peptides with promising anticancer activity'* into concrete measurable design requirements, e.g., engagement of the astrocytic protein whose aggregation drives the disease. And it can also *assess* generated designs via first-principles chemical and biological reasoning based on observable structural measurements.

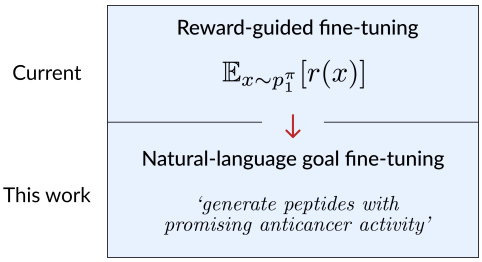


Figure 1: We extend the capabilities of current fine-tuning methods for specialized scientific generators (e.g., flow models), allowing them to align to natural-language goals.

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.

Over the last few years, a growing literature has proposed methods adapting pre-trained flow and diffusion models to downstream reward functions via fine-tuning, such as FlowGRPO, Adjoint Matching, DiffusionNFT, and Flow Density Control (Liu et al., 2025a; Domingo-Enrich et al., 2025; Zheng et al., 2026; De Santi et al., 2025). These methods assume that discovery objectives are expressed via an explicit reward function (Fig. 1). Nonetheless, proxy rewards in sciences are often unavailable or unreliable learned predictors. Moreover, discovery goals typically do not arrive in this form, but rather as high-level natural-language goals. This raises the central question of this work:

How can we align a domain-specific generator (e.g., a molecular flow model) directly to a natural-language discovery goal?

Answering this question would enable scientists to adapt specialized generators to natural-language goals without task-specific proxy rewards (Fig. 1). We cast the LLM as a *teacher* that distills high-level goal information into the domain-specific generator, the *student*, steering it toward the discovery region Ω_g specified by the natural-language goal g (Fig. 2). This raises two key challenges.

First, the teacher's knowledge must yield a usable signal for flow adaptation. Current LLMs are poor at estimating absolute scores over designs (Jacobs et al., 2024), so these cannot serve as reliable rewards. Their *relative* judgments are far more reliable, as we show across three scientific domains in Sec. 4. Second, the pre-trained generator may assign negligible probability to Ω_g (e.g., De Santi et al., 2026; Uehara et al., 2024b). Its samples then rarely satisfy g, so direct feedback under the goal cannot discriminate among them, thus yielding no useful learning signal. We address both challenges with **L**LM-to-**F**low **D**istillation (LFD), where the LLM acts as an adaptive teacher for the domain-specific generator. Rather than supervise the final goal g directly, it constructs a *generator-dependent* curriculum of intermediate sub-goals from the student's current capabilities, inferred through observable properties, and progressively tightens it toward Ω_g (Fig. 2, right). At each stage, the teacher assesses current samples under the sub-goal and distills the resulting judgements into the student via DPO. This provides dense supervision even when the initial generator places negligible probability on Ω_g, enabling adaptation to poorly covered, out-of-distribution discovery regions.

Our contributions.

- We formalize the problem of *aligning scientific generators to natural-language goals*: shifting the density of a pre-trained generative model (e.g., a molecular flow model) to sample designs from the discovery region Ω_g defined by prompt g, given access to LLM judgments (Sec. 2).
- We introduce LLM-to-Flow Distillation, which aligns domain-specific generators to natural-language discovery goals without employing a task-specific reward. A frontier LLM acts as an adaptive teacher, using first-principles scientific reasoning to construct intermediate sub-goals from the generator's current samples and steer it via iterative preference distillation (Sec. 3).
- We evaluate LFD on therapeutic peptide and molecular design tasks. Without accessing task-specific proxy rewards, LFD outperforms state-of-the-art reward-guided baselines on held-out activity and docking evaluations, showing advantages of language-guided adaptation: reduced proxy overoptimization, interpretable curricula, and successful adaptation to highly out-of-distribution goals where both the pre-trained generator and reward-guided methods fail (Sec. 5).

2 Challenge: Aligning Flow Models to Natural-Language Goals

We formalize the problem of aligning a domain-specific generative model to a natural-language discovery goal. We consider flow models as our running example while the discussion applies to other generators, e.g., discrete diffusion models. We first state the learning problem, then identify key challenges that distinguish it from standard reward-guided flow adaptation (Uehara et al., 2024a).

Let $\mathcal{X}$ be a design space, where each $x \in \mathcal{X}$ is a candidate design, e.g., a peptide sequence or a molecular structure, and let p_θ denote the terminal density on $\mathcal{X}$ induced by a flow model v_θ. We denote by θ_0 the velocity parameters of a pre-trained flow, e.g., via flow matching (Lipman et al., 2024), on a data distribution p_{data}. For a region $\Omega \subseteq \mathcal{X}$, we write $p_\theta(\Omega)$ for the probability the model assigns to it. Given a natural language discovery goal g, it induces on $\mathcal{X}$ the *discovery region*:

$$\Omega_g \;:=\; \{x \in \mathcal{X} \,:\, f_g(x) = 1\}, \tag{1}$$

where $f_g : \mathcal{X} \to \{0, 1\}$ indicates whether a design satisfies g. An overview is illustrated in Fig. 2 (right). As a running example, consider $g =$ *'generate peptides with promising anticancer activity'*.[1]

[1] Alternatively, Ω_g can be defined as an arbitrary level-set of a real-valued function f_g.

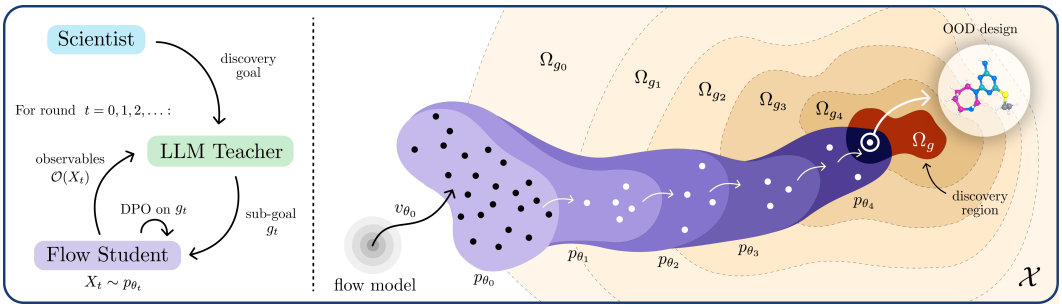


Figure 2: **(left):** LLM $\leftrightarrow$ Flow interaction loop in LFD. **(right):** LFD progressively redistributes model density toward the discovery region Ω_g induced by the natural-language discovery goal g.

We consider the practical setting where, besides the pre-trained flow v_{θ_0}, two auxiliary resources are available: a *frozen LLM* and a set of *observables*. We assume the LLM is accessible only via API queries, with no access to its parameters, gradients, or internal representations, and cannot fine-tune it, reflecting the standard deployment regime of frontier LLMs. The LLM observes a flow-generated design x via a set of computable observables $\mathcal{O}(x)$. For instance, for a 3D molecule, $\mathcal{O}(x)$ may include its SMILES representation and structural descriptors such as molecular weight, aromatic-ring count, and rotatable-bond count. This translates a generated design into a representation the LLM can reason about: bond angles and distances, for example, give it direct access to the geometry of a molecular structure. Analogous observables are available across scientific domains, including molecules, biological sequences, and materials. Crucially, these observables allow us to partially describe x, but do not evaluate whether it satisfies the discovery goal g via a scalar proxy reward for g.

2.1 Formal Problem: Aligning a Flow Model to a Natural-Language Goal

Aligning a domain-specific generator to the discovery goal g corresponds to redistributing its density so that it better covers the discovery region $\Omega_g \subseteq \mathcal{X}$. We aim to tackle this problem by *fine-tuning* the pre-trained flow model v_{θ_0} into an updated model v_θ, as formally stated in the following Eq. 2.

Aligning a Flow Model to a Natural Language Goal

$$\text{update } \theta_0 \to \theta \text{ such that } p_\theta(\Omega_g) \gg p_{\theta_0}(\Omega_g), \tag{2}$$

where Ω_g denotes the discovery region for natural-language goal g

Solving Eq. (2) via reward-based adaptation methods such as FlowGRPO (Liu et al., 2025a), DiffusionNFT (Zheng et al., 2026), or Adjoint Matching (Domingo-Enrich et al., 2025), would require a scalar surrogate for f_g (e.g., Uehara et al., 2024a; 2025). In experimental sciences, e.g., chemistry and biology, such a surrogate is rarely at hand: a goal like *'generate peptides with promising anticancer activity'* is not an easily measurable property as it often requires long and expensive experiments, and a predictor fitted to existing data tends to be least reliable in poorly covered regions highly relevant for generative discovery tasks. We therefore study how current LLMs, which reason about g from first principles, can provide supervision for flow adaptation. We cast the LLM as a scientific expert and teacher: given g and candidate observables $\mathcal{O}(x)$, it assesses generated designs to supervise the flow student. This raises two key challenges.

Challenge 1: LLM scores of scientific designs can be miscalibrated

Absolute LLM scores may be miscalibrated with respect to the underlying scientific quantity, so one should first determine which form of feedback can reliably supervise model adaptation.

While LLMs are now widely used to *judge* text (Zheng et al., 2023; Liu et al., 2023), it is less clear whether they can provide reliable feedback in scientific domains. Given a set of observables $\mathcal{O}(x)$, one may ask an LLM to estimate an unknown scientific quantity of interest (e.g., score each peptide from 1 to 100 according to potential anticancer activity). In Sec. 4, we show across three scientific domains that, although these absolute scores can be far from the ground truth, their induced ordering remains informative, with pairwise accuracy consistently above chance. We therefore use this goal-conditioned ordinal signal to distill the teacher's knowledge into the scientific generator. Crucially, the usefulness of this signal depends on the current state of the generator relative to the goal g.

Challenge 2: Goals far from the initial generator can provide no learning signal

If the initial generator places negligible probability on the discovery region Ω_g, its samples rarely satisfy goal g; direct feedback under g is then non-discriminative and provides no learning signal.

This regime is common in generative discovery, where goals are often narrow and out-of-distribution relative to the pre-trained generator (De Santi et al., 2026; Uehara et al., 2024b). Its probability mass on the discovery region can therefore be negligible, i.e., $p_{\theta_0}(\Omega_g) \approx 0$. For instance, a pre-trained molecular flow model may rarely sample candidates relevant to a specific biological target. Direct feedback under g would then yield no positive examples to reward, causing adaptation to stall.

Together, these challenges suggest two requirements for tackling Problem (2): the teacher should give ordinal rather than cardinal feedback, and should judge candidates under objectives that distinguish among the student's current samples. We introduce LLM-to-Flow Distillation, an adaptation method that satisfies both through a generator-dependent curriculum of intermediate sub-goals.

3 Method: LLM-to-Flow Distillation

We introduce **L**LM-to-**F**low **D**istillation (LFD) (Alg. 1), which aligns a pre-trained flow v_{θ_0} to a natural-language goal g through an online LLM-generated curriculum. At each round, the teacher estimates the current generator capabilities from current observables, uses its scientific knowledge to propose an attainable sub-goal, and judges candidates under it. The resulting preferences distill the teacher's goal-related scientific knowledge into the flow as the curriculum tightens toward g.

Algorithm 1 **L**LM-to-**F**low **D**istillation (LFD)

Require: pre-trained flow v_{θ_0}, natural-language goal g, LLM teacher, observables $\mathcal{O}$, rounds T, batch size n.

$g_{-1} \leftarrow g$
for $t = 0, \dots, T-1$ **do**
  `# sample from current generator`
  $X_t = \{x_i\}_{i=1}^n \sim p_{\theta_t}$, observe $\mathcal{O}(X_t)$
  `# adaptive goal selection`
  $g_t \leftarrow \textsc{SetGoal}\big(g, g_{t-1}, \mathcal{O}(X_t)\big)$ ▷ Eq. (3)
  `# elicit ordinal LLM feedback`
  $\mathcal{D}_t \leftarrow \text{LLM-}\textsc{Judge}\big(X_t; g_t\big)$ ▷ Eq. (5)
  `# distill feedback into flow`
  $\theta_{t+1} \leftarrow \textsc{Minimize}\big(\mathcal{L}_t(\theta_t, \mathcal{D}_t)\big)$ ▷ Eq. (6)
end for
return v_{θ_T}

The LLM first proposes a very broad initial sub-goal g_0, chosen to yield a useful learning signal on samples from the base generator while remaining a valid relaxation of g. As the student adapts, the LLM progressively tightens this objective toward the original goal. Formally, let g_t denote the sub-goal selected at round $t \in [T]$, and let $\Omega_{g_t} \subseteq \mathcal{X}$ be its corresponding goal region (see Definition 1). The resulting adaptive curriculum generated by LFD moves the generator through the following sequence of goal regions:

$$\Omega_{g_0} \longrightarrow \Omega_{g_1} \longrightarrow \cdots \longrightarrow \Omega_{g_{T-1}} \longrightarrow \Omega_g,$$

updating v_{θ_t} to $v_{\theta_{t+1}}$ at each round via LLM judgments. Conceptually, this resembles *annealing* methods, where a difficult optimization problem is approached through a sequence of easier intermediate objectives that are gradually tightened toward the final target (Kirkpatrick et al., 1983). Unlike classical annealing, LFD does not follow a fixed schedule: it adapts the intermediate objectives online based on the generator's current capabilities, using the LLM's scientific knowledge to advance them toward g.

Each LFD round has three steps: construct an admissible sub-goal sufficiently attainable by the current student, elicit teacher judgments under it, and update the student from the resulting preferences. We describe the core procedure below. Further implementation details are provided in Appendix B.

Sub-goal generation. At round t, the student generates a batch $X_t = \{x_i\}_{i=1}^n \sim p_{\theta_t}$, and computes the corresponding observables $\mathcal{O}(X_t)$, which allow the LLM to understand the region of design space currently covered by the student generator, i.e., its current generable set (De Santi et al., 2026). The LLM teacher is prompted to use this information to generate the next sub-goal:

$$g_t \leftarrow \textsc{SetGoal}(g, g_{t-1}, \mathcal{O}(X_t)), \qquad g_{-1} := g. \tag{3}$$

A proposed sub-goal is accepted only if the teacher identifies enough positive and negative candidates in X_t with respect to the current sub-goal, allowing discrimination between them for feedback.

From LLM judgments to induced preferences. Under the accepted sub-goal g_t, the LLM classifies each generated x as *positive*, *negative*, or *abstained*. This provides a cheap ordinal supervision

signal. Nonetheless, unlike poorly calibrated raw scores, induced preferences are generally more reliable, as we experimentally analyse in Sec. 4. Moreover, this preference elicitation method requires only one pass over the batch, avoiding the usual quadratic API cost of exhaustive pairwise duels.

For each candidate $x_i \in X_t$, let $\ell_t(x_i) \in \{+, -, \varnothing\}$ denote the LLM-JUDGE feedback, where $\varnothing$ denotes abstention. We define the positive and negative sets as

$$\mathcal{P}_t = \{x_i \in X_t : \ell_t(x_i) = +\}, \qquad \mathcal{N}_t = \{x_i \in X_t : \ell_t(x_i) = -\}. \tag{4}$$

Abstained candidates are ignored when constructing preferences; the remaining judgments induce the preference dataset:

$$\mathcal{D}_t = \{(x^+, x^-) : x^+ \in \mathcal{P}_t,\ x^- \in \mathcal{N}_t\}, \tag{5}$$

We use the induced preferences to update the student using distilled Direct Preference Optimization (dDPO) (Tunstall et al., 2024). Here, *distilled* refers to the fact that the preferences are acquired from the LLM teacher's judgements, while the student is updated via standard DPO (Rafailov et al., 2023). Unlike the fixed, offline preference data used in prior dDPO, LFD regenerates preferences from the current student's samples at every round and uses the current model p_{θ_t} as the reference:

$$\mathcal{L}_t(\theta, \mathcal{D}_t) = -\mathbb{E}_{(x^+,x^-)\sim\mathcal{D}_t} \log \sigma\left(\beta \left[\log \frac{p_\theta(x^+)}{p_{\theta_t}(x^+)} - \log \frac{p_\theta(x^-)}{p_{\theta_t}(x^-)}\right]\right). \tag{6}$$

Starting from current flow parameters θ_t, minimizing $\mathcal{L}_t$ yields θ_{t+1}. Using p_{θ_t} as the reference regularizes each update to the current generator, while resampling after every update keeps the teacher's judgments aligned with the student's evolving generable set. Since terminal log-densities are intractable, we optimize the corresponding denoising-loss surrogate (Wallace et al., 2024).

Each round is a KL-regularized step toward an implicit target. Suppose the teacher-induced preferences follow a Bradley–Terry model (Bradley & Terry, 1952), for an unobserved utility $f_{g_t} : \mathcal{X} \to \mathbb{R}$. Following the standard derivation of DPO (Rafailov et al., 2023) yields:

$$p_t^\star \propto p_{\theta_t} e^{f_{g_t}/\beta} \quad \Longleftrightarrow \quad p_t^\star = \arg\max_q \big\{\mathbb{E}_q[f_{g_t}] - \beta D_{\mathrm{KL}}(q \,\|\, p_{\theta_t})\big\}. \tag{7}$$

Thus, each sub-goal induces an implicit reward f_{g_t} and a KL-regularized target $p_t^\star$. Across rounds, these targets define local steps along the curriculum toward Ω_g. See Appendix C for the derivation.

LFD relies on the ability of an LLM to reliably provide accurate ordinal signal across scientific designs that one can then utilize for model adaptation. We analyze empirically this capability next.

4 How Reliable is LLM Scientific Feedback?

We test whether, and in what form, frontier LLMs can provide useful feedback on scientific designs given only a set of *observables*. We evaluate three API-accessible LLMs, `Gemini 3.7 Flash`, `GPT 5.6 Terra`, and `Claude Opus 5`, on three tasks from distinct domains, judging (i) peptides from PepDFM (Chen et al., 2025b) by mean Kyte–Doolittle hydropathy; (ii) crystal structures from MatterGen (Zeni et al., 2025) by energy above the convex hull after MatterSim relaxation (Yang et al., 2024); and (iii) QM9 small molecules (Ramakrishnan et al., 2014) by MolLogP. The ground-truth value is hidden from the LLMs and used only for evaluation.

We find LLMs unreliable at estimating raw scores on a 1–100 scale[2], and equally unreliable at estimating differences between designs (Figure 3, top). Their judgments may nevertheless carry useful information about relative quality, so we compare four forms of ordinal feedback: independent 1–100 scoring, listwise ranking, batch classification as positive, negative, or abstain, and direct pairwise comparison. Scores and rankings induce preferences through their ordering, while batch classification treats every positive candidate as preferred to every negative one. As shown in Figure 3 (bottom), all four recover pairwise accuracy above chance, despite the unreliable magnitudes of cardinal scores. Duels scale quadratically in the number of samples, while the other methods scale linearly and match or exceed their accuracy, so we use batch classification for preference supervision in LFD. Section 5 shows these preferences suffice to adapt generators toward natural-language goals.

Insight 0. LLM scientific judgements tend to be more reliable in relative than absolute terms.

[2]For the sake of comparison, the ground truth scores are normalized in 1-100.

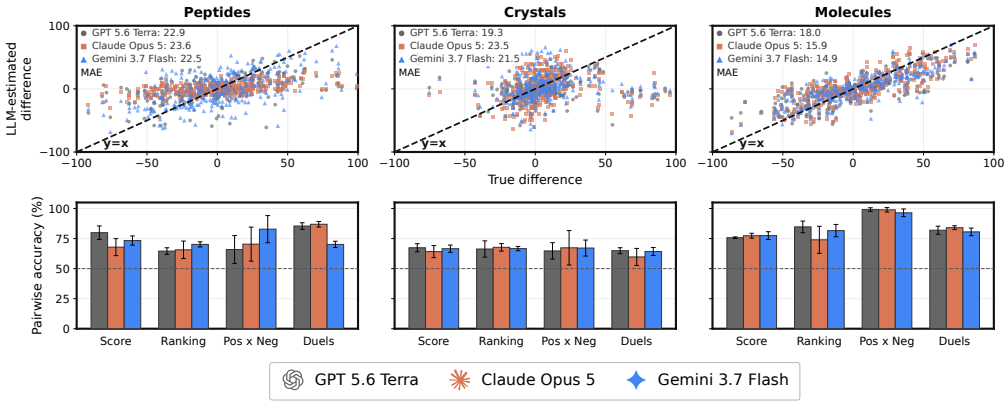


Figure 3: **(top)** LLM estimated scores tend to be far from the ground truth ($y = x$ line). **(bottom)** Pairwise accuracy of four elicitation mechanisms can provide a learning signal (dashed line at 50%).

5 Experimental Evaluation of LLM-to-Flow Distillation

We assess the ability of LFD to align arbitrary scientific generators to natural-language discovery goals, without access to task-specific proxy rewards. All experiments use `Gemini 3.1 Pro` to construct the curriculum and `Gemini 3.7 Flash` for LLM judgments. The experiments cover both discrete and continuous flows, over peptide and molecular design tasks. Details are in Apx. B.

Baselines and evaluation. We compare LFD via *in silico* experiments, against classifier Guidance (Dhariwal & Nichol, 2021), DPO (Wallace et al., 2024), and recent state-of-the-art methods, including Flow-GRPO (Liu et al., 2025a) and DiffusionNFT (Zheng et al., 2026). Crucially, while these baselines optimize given proxy reward functions, LFD neither observes nor utilizes them.

5.1 Case Study I: Therapeutic Peptide Design

Therapeutic peptides are short amino-acid sequences whose biological effects can arise through target binding, pathway modulation, or interactions with cellular membranes (Wang et al., 2022). We employ PepDFM (Chen et al., 2025b), a discrete flow model over peptide sequences, as the domain-specific generator, and adapt it via LFD to the following two natural-language discovery goals.

> **GOAL 1:** *Generate peptides with promising anticancer activity while maintaining low toxicity toward healthy cells.*

For this goal, the LLM teacher observes only the amino-acid sequence, together with some generic physicochemical descriptors, including charge, hydrophobicity, aromaticity, amphipathicity, and coarse bands for generic developability properties such as solubility, non-hemolysis, and non-fouling. The baselines: Guidance, DPO, and DiffusionNFT, directly optimize the anticancer ACP predictor (Ahmed et al., 2021). Nevertheless, LFD achieves the highest (53%) held-out AntiCP2-alternate predictor (Agrawal et al., 2021) (see Table 1), at the lowest marginal increase in toxicity (42%), measured as with the HemoPI2 random-forest classifier (Rathore et al., 2025).

> **Insight 1.** LFD allows to align domain-specific scientific generators to high-level, natural-language discovery goals, without utilizing any proxy reward.

LFD exposes an interpretable curriculum and candidate judgments during adaptation. Fig. 5 shows the teacher progressively refining the goal from positive charge to hydrophobic residues and aromatic interfacial anchors. Fig. 4 shows the LLM rejecting a peptide because the prolonged hydrophobic stretch `MLIVFVL` violates the current goal while reporting unavailable in vitro hemolysis and aggregation measurements as sources of uncertainty. Together, these outputs provide an interpretable interface for human oversight: scientists can inspect the curriculum and candidate judgments and *intervene* if needed.

LLM interpretable traces

Peptide (−/−):
`MLKGRWPDEQKKNHLMNYNFMLIVFVLTFNISFQSSNI`
`Charge index = +2.1; hydrophobicity = 47.4%; lower-third non-hemolysis band.`

LLM rationale: *Violates the operative goal (g_3) by possessing a prolonged, contiguous hydrophobic tract (MLIVFVL) [...]*

LLM uncertainty: *In vitro hemolytic threshold and aggregation state in physiological buffer.*

Figure 4: Rejected peptide LLM trace. LLM reports rationale and possible sources of uncertainty.

Table 1: LFD achieves the strongest held-out activity on both peptide tasks while keeping low hemolysis, and showing the largest distance from the pre-trained amino-acid distribution (Dai et al., 2025).

Method	Therapeutic Peptides (CPP)			Therapeutic Peptides (Anti-Cancer)		
	pLM4CPPs (%) ↑	HemoPI2 (%) ↓	Distance from prior ↑	AntiCP2-alt. (%) ↑	HemoPI2 (%) ↓	Distance from prior ↑
Pre-trained	8.81 ± 3.20	40.84 ± 1.96	0.06 ± 0.04	26.43 ± 2.91	**40.85** ± 1.89	0.00 ± 0.00
DPO	10.39 ± 3.19	40.49 ± 1.92	0.07 ± 0.04	25.29 ± 3.19	42.77 ± 1.81	0.08 ± 0.04
DIFFUSIONNFT	57.04 ± 6.22	38.87 ± 1.62	0.29 ± 0.03	33.26 ± 2.76	45.61 ± 2.39	0.16 ± 0.03
GUIDANCE	8.49 ± 3.30	43.45 ± 1.85	0.06 ± 0.04	35.94 ± 2.33	47.25 ± 1.81	0.29 ± 0.03
LFD	**69.06** ± 6.36	**37.02** ± 1.69	**0.34** ± 0.03	**53.54** ± 3.32	42.48 ± 2.04	**0.31** ± 0.03

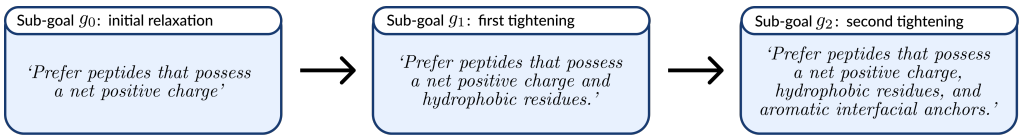


Figure 5: LFD generates online a sequence of scientist-interpretable intermediate sub-goals.

Insight 2. LFD generates scientist-interpretable sub-goals and decision-making traces, allowing human scientists to understand and steer the generative model adaptation process.

Fine-tuning generators against learned proxies is often prone to *reward hacking*: proxy errors can boost scores without matching gains under an independent evaluator (Gao et al., 2023; Yoshizawa et al., 2025). GUIDANCE shows this, saturating the used ACP proxy at 100%, but reaching only 35.94% on the held-out AntiCP2-alt evaluation. LFD instead never optimizes either anticancer predictor. Using LLM judgments, it achieves the best held-out score, suggesting less overoptimization.

Insight 3. LFD mitigates reward hacking in out-of-distribution discovery by supervising adaptation via first-principles scientific reasoning rather than direct optimization of a proxy reward.

We next test whether the gains over reward-guided methods extend to cell-penetrating peptides.

GOAL 2: *Generate peptides with promising mammalian-cell-penetrating activity while minimizing toxicity to mammalian cells.*

Cell-penetrating peptides (CPPs) cross cellular membranes, enabling intracellular cargo delivery and direct targeting (Heitz et al., 2009). Here, the LLM teacher supervises PepDFM adaptation from sequences and observables without access to any proxy reward or to the ESM2-based pLM4CPPs evaluator (Kumar et al., 2025). The baselines optimize a classifier CPP proxy trained on CellPPD (Gautam et al., 2013). Table 1 shows LFD achieves the highest pLM4CPPs score, rising from 8.81 to 69.06 versus 57.04 for the best baseline, DIFFUSIONNFT. Its HemoPI2 toxicity (Rathore et al., 2025) also falls from 40.84 to 37.02, yielding the lowest predicted hemolysis among compared methods.

5.2 CASE STUDY II: DE NOVO DESIGN OF SMALL MOLECULES

In molecular design, target-specific requirements may lie far outside the training distribution and be hard to encode as explicit reward functions, while molecules must remain chemically valid (Guo & Schwaller, 2024; De Santi et al., 2026). We utilize FlowMol3 (Dunn & Koes, 2026), a multimodal flow over topology and 3D geometry, pre-trained on drug-like GEOM-Drugs dataset (Axelrod & Gomez-Bombarelli, 2022).

GOAL 3: *Generate promising drug-like small molecules for binding the dopamine D2 receptor, balancing plausible receptor recognition with moderate lipophilicity and chemically viable structures.*

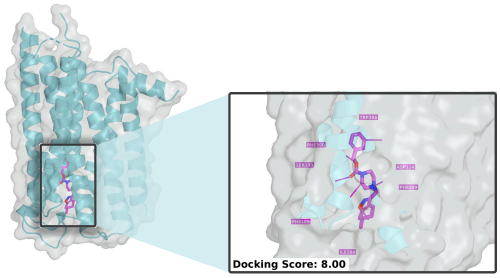
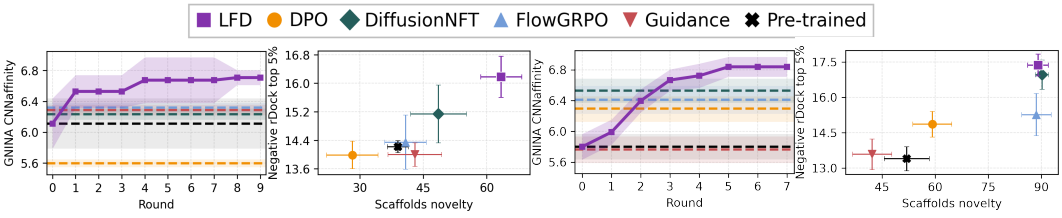


Figure 6: LFD-generated molecule docking to D2 Dopamine receptor, showing multiple close interactions (pink labels) below 3.5 Å (GNINA).

The LLM teacher supervises generator adaptation from candidate observables such as molecular weight, hydrogen-bond counts, topological polar surface area, and rotatable-bond counts, without

Table 2: LFD outperforms reward-guided baselines on both D2 and GSK3β docking tasks, under all three held-out evaluators, despite never accessing the proxy docking scores utilized by the baselines.

	Small Molecules (D2 Receptor)			Small Molecules (Kinase-3 Beta)		
Method	**GNINA ↑**	**Vina ↓**	**rDock ↓**	**GNINA ↑**	**Vina ↓**	**rDock ↓**
Pre-trained	6.11 ± 0.33	−7.52 ± 0.39	−7.64 ± 1.33	5.80 ± 0.17	−7.32 ± 0.19	−8.51 ± 0.55
DPO	5.60 ± 0.05	−6.81 ± 0.08	−8.11 ± 0.36	6.29 ± 0.17	−7.60 ± 0.20	−8.77 ± 0.74
FLOW-GRPO	6.32 ± 0.06	−7.83 ± 0.09	−8.15 ± 0.27	6.41 ± 0.18	−8.06 ± 0.25	−9.31 ± 0.60
DIFFUSIONNFT	6.23 ± 0.07	−7.90 ± 0.23	−9.39 ± 0.26	6.53 ± 0.16	−8.68 ± 0.22	−11.71 ± 0.50
GUIDANCE	6.29 ± 0.07	−7.65 ± 0.21	−7.92 ± 0.25	5.76 ± 0.17	−7.12 ± 0.22	−8.03 ± 0.60
LFD	**6.54** ± 0.09	**−8.48** ± 0.16	**−9.89** ± 0.38	**6.77** ± 0.11	**−8.84** ± 0.16	**−12.87** ± 0.57



(a) Dopamine docking (b) Novel scaffolds (c) Kinase-3β binding (d) Novel scaffolds

Figure 7: LFD outperforms reward-guided methods on held-out D2 and GSK3β docking scores without proxy rewards (a,c), yielding the best top-5% rDock scores and high scaffold novelty (b,d).

access to docking scores. The baselines instead optimize a task-specific D2 receptor proxy reward (Olivecrona et al., 2017). We evaluate docking quality with GNINA CNNaffinity (McNutt et al., 2021), Vina (Trott & Olson, 2010), and rDock (Ruiz-Carmona et al., 2014). Table 2 (left) shows LFD has the best mean on all three: GNINA CNNaffinity rises from 6.11 to 6.54 pK, above FLOW-GRPO (6.32). Vina improves from −7.52 to −8.48 kcal/mol, versus −7.90 for DIFFUSIONNFT; rDock improves from −7.64 to −9.89, versus −9.39 of the best baseline. LFD retains high validity (98%) (RDKit) and PoseBuster conformer quality (89.5%) (Buttenschoen et al., 2024). Fig. (6), shows a docking pose with multiple interactions under 3.5 Å for a LFD-generated molecule (DeLano et al., 2002). Fig. (7a) reports how LFD surpasses the baselines GNINA score over just one round, and Fig. (7b) shows that LFD achieves the best trade-off between top-samples quality and diversity. Overall, LFD achieves strong docking results without utilizing any proxy docking reward.

Insight 4. LFD achieves strong docking performance without using proxy rewards, outperforming state-of-the-art reward-guided baselines under physics- and learned-scoring evaluation.

We next target GSK3β to test whether the gains over reward-guided methods extend beyond D2.

GOAL 4: *Generate promising drug-like small molecules for binding the ATP pocket of glycogen synthase kinase-3 beta (GSK3β), while maintaining reasonable aqueous solubility and chemically plausible structures.*

Via the same observables as in the previous task, the LLM supervises adaptation without observing docking scores. Baselines instead optimize a proxy task-specific TDC GSK3β classifier (Huang et al., 2021). Table 2 (right) shows the results given by held-out evaluations: LFD achieves the best results on all three evaluators over all four baselines, improving GNINA CNNaffinity from 5.80 to 6.77 pK, Vina from −7.32 to −8.84 kcal/mol, and rDock from −8.51 to −12.87. LFD preserves high RDKit validity (96%) and PoseBuster conformer quality (84%). Fig. (8) illustrates a docking pose of an LFD-generated molecule with multiple interactions under 3.5 Å (right, pink labels). Fig. (7c) reports how LFD surpasses all baselines evals after 3 rounds. Fig. (7d) shows that LFD achieves the best top-samples quality over baselines, and near-optimal diversity. Across both protein targets, LFD consistently surpasses state-of-the-art reward-guided methods without access to docking rewards.

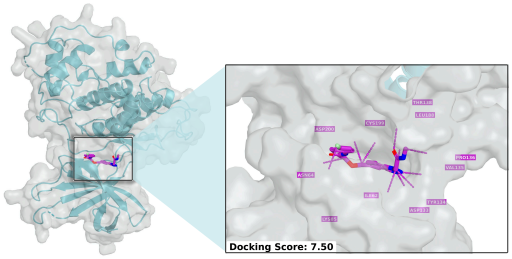


Figure 8: LFD-generated molecule docking to ATP pocket of GSK3β, showing multiple close interactions (pink labels) below 3.5 Å (GNINA).

GOAL 5: *Generate neutral, connected small molecules for a kinase-focused discovery campaign that contain a six-membered aza-aromatic ring bearing a free primary amino group on a ring carbon adjacent to a pyridine-like ring nitrogen, and a distinct fully saturated ring containing a free secondary amine.*

This goal imposes constraints on ring structures and saturation, atom adjacency, and functional-group placement, that would require complex task-specific reward logic. LFD instead adapts directly the flow generator from the natural-language description. The hit in Fig. (9) satisfies the full, goal description.

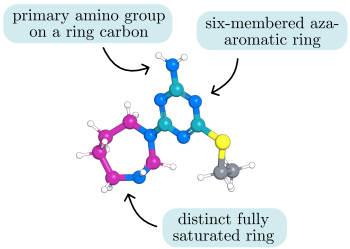


Figure 9: LFD-generated 3D molecule satisfying Goal 5.

Insight 5. LFD allows to specify complex constraints on generated designs directly as a natural language prompt.

While the pre-trained model yields no hits over 38,500 samples, LFD-adapted FlowMol3 produces the hit in Fig. (9) in a 512-sample batch at round 6 under the third sub-goal g_2. All reward-guided baselines fail despite receiving an LLM-generated binary reward implementing the full goal, since positive examples are initially too sparse for this reward to provide useful learning signal. The curriculum of LFD, which is adaptive to the current generator coverage, provides informative feedback under intermediate sub-goals, enabling progress toward an initially hard-to-sample region.

Insight 6. Through its adaptive LLM-generated sub-goal curriculum, LFD allows to adapt flow models to OOD discovery goals where state-of-the-art reward-guided methods might fail.

6 Related Works

Reward-guided and distributional adaptation of flow and diffusion models. Expected-reward maximization is a standard formulation for adapting pre-trained diffusion and flow models, including policy-gradient methods such as DDPO (Black et al., 2024) and Flow-GRPO (Liu et al., 2025a), reward-weighted flow matching (Fan et al., 2025), stochastic-control formulations such as Adjoint Matching (Domingo-Enrich et al., 2025), and forward-process or distillation-based methods such as DiffusionNFT (Zheng et al., 2026) and iterative reward distillation (Su et al., 2026). Crucially, these methods rely on a given reward function. Instead, LFD allows to align flow and diffusion models to a natural-language goal. It uses an LLM to propose intermediate sub-goals as the generator's capabilities evolve, and distills the LLM's ordinal judgments under each into the generative model.

LLM-based optimization and model-to-model distillation LLMs can directly edit or optimize scientific candidates (Liu et al., 2024a; Chen et al., 2025a; Liang et al., 2026), as well as serve as surrogates or acquisition functions in Bayesian optimization (Liu et al., 2024b; Yuan et al., 2026; Xu et al., 2026). On the other hand, LFD fine-tunes a separate domain-specific generator through ordinal judgments that become preference supervision and never generates directly candidate designs. Analogously, the principle of design space separation, i.e., LLM on high-level concepts and domain-specific generators, distinguishes LFD from current on-policy distillation methods, which require teacher and student to share the output space (Agarwal et al., 2024; Li et al., 2026; Fang et al., 2026).

AI feedback and preference optimization. AI feedback has been employed to support preference optimization (Tunstall et al., 2024; Li et al., 2023; Yuan et al., 2024), including when tasks are defined by chemical or biological criteria (Park et al., 2023; Liu et al., 2025b). Motif utilizes an LLM to annotate a fixed dataset of observations with preferences under a natural-language task description (Klissarov et al., 2024), while OAIF queries online an LLM annotator on the current model's samples (Guo et al., 2024). Crucially, in both cases the goal is fixed in advance. LFD instead generates online intermediate sub-goals depending on the current generator's distribution.

7 Conclusion

We introduced LFD, which allows to adapt domain-specific generators to follow natural-language scientific discovery goals without requiring task-specific proxy rewards. Across therapeutic peptides and small molecules design tasks, it surpasses state-of-the-art reward-guided baselines on independent *in silico* evaluations, while showing higher interpretability and stronger performance on OOD discovery tasks. These results establish LFD as a practical and powerful new form of generative control for discovery in the experimental sciences, coupling LLM high-level scientific knowledge with domain-specific generative priors. Integrating wet-lab feedback into LFD is a natural next step.

AI USE STATEMENT

In this work, we used generative AI tools to implement methods and to provide feedback on experiments. We have not used generative AI tools to generate synthetic datasets, develop theoretical models or conceptual frameworks, formulate mathematical claims or proofs, propose or refine hypotheses, design the research methodology or experiments, or interpret results the rest of the required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools to create or modify scientific figures or images, suggest experimental parameters, create or edit software code, draft parts of a research paper, summarize or analyse existing literature, sourcing/searching for information, edit a research paper to improve readability, identify relevant literature, format references. We have reviewed all AI-assisted work. All references and literature summaries were verified against the original sources, AI-assisted code was reviewed and tested for correctness by the authors, suggested experimental parameters and feedback on experiments were evaluated by the authors and validated empirically, figures were checked against the underlying results, and all AI-drafted or AI-edited text was critically reviewed and revised by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work adheres to the ICLR Code of Ethics. Our research is related to computational methods for biological design. We acknowledge that biological design technologies could have dual-use implications. The molecule and peptide design application demonstrated in our work targets therapeutic applications for treating diseases, representing beneficial applications. However, we recognize that these design methods could potentially be misused for harmful purposes. We have designed our experiments to focus on clearly beneficial applications and have not explored potentially harmful uses. All datasets used in this work are publicly available and have been previously published with appropriate ethical considerations. No human subjects were involved in this research. The computational methods developed here are general frameworks that require domain expertise and appropriate oversight for real-world applications. We encourage responsible use of these methods and recommend that practitioners consider ethical implications and potential risks when applying the proposed design technologies in practice.

REPRODUCIBILITY STATEMENT

We provide significant implementation details of LFD (Algorithm 1, Sec. 3) in Appendix B. For each application and goal-specific experiment presented in Sec. 5, we report experimental details in Appendix D. For API-hosted LLMs, we report the model names, type of usage, provider-reported versions when available. Similarly, pre-trained generative models, datasets, and predictors employed are properly referenced either within Sec. 5 or in Appendix D. Moreover, within Appendix D we further report samples of experiments LLM reasoning traces and dynamics to clarify the operational workings of the method thus easing its reproducibility.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, 2024.

Piyush Agrawal, Dhruv Bhagat, Manish Mahalwal, Neelam Sharma, and Gajendra PS Raghava. Anticp 2.0: an updated model for predicting anticancer peptides. *Briefings in bioinformatics*, 22 (3):bbaa153, 2021.

Sajid Ahmed, Rafsanjani Muhammod, Zahid Hossain Khan, Sheikh Adilina, Alok Sharma, Swakkhar Shatabda, and Abdollah Dehzangi. Acp-mhcnn: an accurate multi-headed deep-convolutional neural network to predict anticancer peptides. *Scientific Reports*, 11(1):23676, 2021.

Simon Axelrod and Rafael Gomez-Bombarelli. Geom, energy-annotated molecular conformations for property prediction and molecular generation. *Scientific Data*, 9(1):185, 2022.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *International Conference on Learning Representations*, 2024.

Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Martin Buttenschoen, Garrett M Morris, and Charlotte M Deane. Posebusters: Ai-based docking methods fail to generate physically valid poses or generalise to novel sequences. *Chemical science*, 15(9):3130–3139, 2024.

Angelica Chen, Samuel Don Stanton, Frances Ding, Robert G. Alberstein, Andrew Martin Watkins, Richard Bonneau, Vladimir Gligorijević, Kyunghyun Cho, and Nathan C. Frey. Generalists vs. specialists: Evaluating llms on highly-constrained biophysical sequence optimization tasks. In *International Conference on Machine Learning*, 2025a.

Tong Chen, Yinuo Zhang, Sophia Tang, and Pranam Chatterjee. Multi-objective-guided discrete flow matching for controllable biological sequence design. *arXiv preprint arXiv:2505.07086*, 2025b.

Xiao-peng Dai, Xiang-chun Meng, Ying-jun Zhou, Zhi-min Li, Yu Ji, Ulrich Schwaneberg, and Zong-lin Li. De novo design of high-performance sec-type signal peptide via a hybrid deep learning architecture. *JACS Au*, 5(10):4669, 2025.

Riccardo De Santi, Marin Vlastelica, Ya-Ping Hsieh, Zebang Shen, Niao He, and Andreas Krause. Flow density control: Generative optimization beyond entropy-regularized fine-tuning. In *Advances in Neural Information Processing Systems*, 2025.

Riccardo De Santi, Bruce Lee, Cristian Perez Jensen, Kimon Protopapas, Sophia Tang, Cheng-Hao Liu, Pranam Chatterjee, Yisong Yue, and Andreas Krause. Active flow expansion for out-of-distribution discovery: from theory to molecules. *arXiv preprint arXiv:2606.08802*, 2026.

Warren L DeLano et al. Pymol: An open-source molecular graphics tool. *CCP4 Newsl. protein crystallogr*, 40(1):82–92, 2002.

Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems*, 2021.

Carles Domingo-Enrich, Michal Drozdzal, Brian Karrer, and Ricky T. Q. Chen. Adjoint matching: Fine-tuning flow and diffusion generative models with memoryless stochastic optimal control. In *International Conference on Learning Representations*, 2025.

Ian Dunn and David R Koes. Flowmol3: flow matching for 3d de novo small-molecule generation. *Digital Discovery*, 5(5):2052–2066, 2026.

Jiajun Fan, Shuaike Shen, Chaoran Cheng, Yuxin Chen, Chumeng Liang, and Ge Liu. Online reward-weighted fine-tuning of flow matching with wasserstein regularization. In *International Conference on Learning Representations*, 2025.

Zhen Fang, Wenxuan Huang, Yu Zeng, Yiming Zhao, Shuang Chen, Kaituo Feng, Yunlong Lin, Lin Chen, Zehui Chen, Shaosheng Cao, et al. Flow-opd: On-policy distillation for flow matching models. *arXiv preprint arXiv:2605.08063*, 2026.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, 2023.

Ankur Gautam, Kumardeep Chaudhary, Rahul Kumar, Arun Sharma, Pallavi Kapoor, Atul Tyagi, Open Source Drug Discovery Consortium Info@ osdd. net, and Gajendra PS Raghava. In silico approaches for designing highly effective cell penetrating peptides. *Journal of translational medicine*, 11(1):74, 2013.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Petar Sirkovic, Artiom Myaskovsky, Grzegorz Glowaty, Felix Weissenberger, Alessio Orlandi, Dan Popovici, et al. Accelerating scientific discovery with co-scientist. *Nature*, pp. 1–3, 2026.

Jeff Guo and Philippe Schwaller. It takes two to tango: Directly optimizing for constrained synthesizability in generative molecular design. *arXiv preprint arXiv:2410.11527*, 2024.

Shangmin Guo, Biao Zhang, Tianlin Liu, Tianqi Liu, Misha Khalman, Felipe Llinares, Alexandre Rame, Thomas Mesnard, Yao Zhao, Bilal Piot, Johan Ferret, and Mathieu Blodel. Direct language model alignment from online ai feedback. In *Advances in Neural Information Processing Systems*, 2024.

Frederic Heitz, May Catherine Morris, and Gilles Divita. Twenty years of cell-penetrating peptides: from molecular mechanisms to therapeutics. *British journal of pharmacology*, 157(2):195–206, 2009.

Kexin Huang, Tianfan Fu, Wenhao Gao, Yue Zhao, Yusuf H. Roohani, Jure Leskovec, Connor W. Coley, Cao Xiao, Jimeng Sun, and Marinka Zitnik. Therapeutics data commons: Machine learning datasets and tasks for drug discovery and development. In *NeurIPS Datasets and Benchmarks*, 2021.

Yueshan Huang, Gufeng Yu, Letian Chen, Haoyang Luan, and Yang Yang. Epic: multi-objective guided diffusion for epitope design in tcr-pmhc complexes. *Bioinformatics*, 42(6):btag358, 2026.

Ryan Jacobs, Maciej P Polak, Lane E Schultz, Hamed Mahdavi, Vasant Honavar, and Dane Morgan. Regression with large language models for materials and molecular property prediction. *arXiv preprint arXiv:2409.06080*, 2024.

Subhojyoti Khastagir, Kishalay Das, Pawan Goyal, Seung-Cheol Lee, Satadeep Bhattacharjee, et al. Llm meets diffusion: a hybrid framework for crystal material generation. In *Advances in Neural Information Processing Systems*, 2026.

Scott Kirkpatrick, C Daniel Gelatt Jr, and Mario P Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983.

Martin Klissarov, Pierluca D'Oro, Shagun Sodhani, Roberta Raileanu, Pierre-Luc Bacon, Pascal Vincent, Amy Zhang, and Mikael Henaff. Motif: Intrinsic motivation from artificial intelligence feedback. In *International Conference on Learning Representations*, 2024.

Nandan Kumar, Zhenjiao Du, and Yonghui Li. plm4cpps: protein language model-based predictor for cell penetrating peptides. *Journal of Chemical Information and Modeling*, 65(3):1128–1139, 2025.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. RLAIF vs. RLHF. In *International Conference on Machine Learning*, 2024.

Lei Li, Zhihui Xie, Mukai Li, Shunian Chen, Peiyi Wang, Liang Chen, Yazheng Yang, Benyou Wang, and Lingpeng Kong. Silkie: Preference distillation for large visual language models. *arXiv preprint arXiv:2312.10665*, 2023.

Quanhao Li, Junqiu Yu, Kaixun Jiang, Yujie Wei, Zhen Xing, Pandeng Li, Ruihang Chu, Shiwei Zhang, Yu Liu, and Zuxuan Wu. Diffusionopd: A unified perspective of on-policy distillation in diffusion models. *arXiv preprint arXiv:2605.15055*, 2026.

Kimberly Liang, Tong Chen, and Pranam Chatterjee. LaPep: Can language contribute to property-guided peptide design? In *ICLR Workshop on Logical Reasoning of Large Language Models*, 2026.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Yaron Lipman, Marton Havasi, Peter Holderrieth, Neta Shaul, Matt Le, Brian Karrer, Ricky T. Q. Chen, David Lopez-Paz, Heli Ben-Hamu, and Itai Gat. Flow matching guide and code. *arXiv preprint arXiv:2412.06264*, 2024.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. In *Advances in Neural Information Processing Systems*, 2025a.

Shengchao Liu, Jiongxiao Wang, Yijin Yang, Chengpeng Wang, Ling Liu, Hongyu Guo, and Chaowei Xiao. Conversational drug editing using retrieval and domain feedback. In *International Conference on Learning Representations*, 2024a.

Tennison Liu, Nicolas Astorga, Nabeel Seedat, and Mihaela van der Schaar. Large language models to enhance bayesian optimization. In *International Conference on Learning Representations*, 2024b.

Xiangyu Liu, Yi Liu, Silei Chen, and Wei Hu. Controllable protein sequence generation with LLM preference optimization. In *AAAI Conference on Artificial Intelligence*, 2025b.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Conference on Empirical Methods in Natural Language Processing*, 2023.

Andrew T McNutt, Paul Francoeur, Rishal Aggarwal, Tomohide Masuda, Rocco Meli, Matthew Ragoza, Jocelyn Sunseri, and David Ryan Koes. Gnina 1.0: molecular docking with deep learning. *Journal of cheminformatics*, 13(1):43, 2021.

Marcus Olivecrona, Thomas Blaschke, Ola Engkvist, and Hongming Chen. Molecular de-novo design through deep reinforcement learning. *Journal of cheminformatics*, 9(1):48, 2017.

Ryan Park, Ryan Theisen, Navriti Sahni, Marcel Patek, Anna Cichońska, and Rayees Rahman. Preference optimization for molecular language models. *arXiv preprint arXiv:2310.12304*, 2023.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, 2023.

Raghunathan Ramakrishnan, Pavlo O. Dral, Matthias Rupp, and O. Anatole von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data*, 1(1):140022, 2014.

Anand Singh Rathore, Nishant Kumar, Shubham Choudhury, Naman Kumar Mehta, and Gajendra PS Raghava. Prediction of hemolytic peptides and their hemolytic concentration. *Communications Biology*, 8(1):176, 2025.

RDKit. Rdkit: Open-source cheminformatics. https://www.rdkit.org.

Sergio Ruiz-Carmona, Daniel Alvarez-Garcia, Nicolas Foloppe, A Beatriz Garmendia-Doval, Szilveszter Juhos, Peter Schmidtke, Xavier Barril, Roderick E Hubbard, and S David Morley. rdock: a fast, versatile and open source program for docking ligands to proteins and nucleic acids. *PLoS computational biology*, 10(4):e1003571, 2014.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, 2015.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, 2019.

Anuroop Sriram, Benjamin K Miller, Ricky T Chen, and Brandon Wood. Flowllm: Flow matching for material generation with large language models as base distributions. In *Advances in Neural Information Processing Systems*, 2024.

Xingyu Su, Xiner Li, Masatoshi Uehara, Sunwoo Kim, Yulai Zhao, Gabriele Scalia, Ehsan Hajiramezanali, Tommaso Biancalani, Degui Zhi, and Shuiwang Ji. Iterative distillation for reward-guided fine-tuning of diffusion models in biomolecular design. In *International Conference on Learning Representations*, 2026.

Oleg Trott and Arthur J Olson. Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry*, 31(2):455–461, 2010.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro Von Werra, Clémentine Fourrier, Nathan Habib, et al. Zephyr: Direct distillation of lm alignment. In *Conference on Language Modeling*, 2024.

Masatoshi Uehara, Yulai Zhao, Tommaso Biancalani, and Sergey Levine. Understanding reinforcement learning-based fine-tuning of diffusion models: A tutorial and review. *arXiv preprint arXiv:2407.13734*, 2024a.

Masatoshi Uehara, Yulai Zhao, Kevin Black, Ehsan Hajiramezanali, Gabriele Scalia, Nathaniel Lee Diamant, Alex M Tseng, Sergey Levine, and Tommaso Biancalani. Feedback efficient online fine-tuning of diffusion models. In *International Conference on Machine Learning*, 2024b.

Masatoshi Uehara, Yulai Zhao, Chenyu Wang, Xiner Li, Aviv Regev, Sergey Levine, and Tommaso Biancalani. Inference-time alignment in diffusion models with reward-guided generation: Tutorial and review. *arXiv preprint arXiv:2501.09685*, 2025.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Conference on Computer Vision and Pattern Recognition*, 2024.

Lei Wang, Nanxi Wang, Wenping Zhang, Xurui Cheng, Zhibin Yan, Gang Shao, Xi Wang, Rui Wang, and Caiyun Fu. Therapeutic peptides: current applications and future directions. *Signal Transduction and Targeted Therapy*, 7(1):48, 2022.

Yaoyao Xu, Xinjian Zhao, Xiaozhuang Song, Lei Bai, and Tianshu Yu. Closed-loop bayesian molecular inverse design with semantic LLM surrogates. *arXiv preprint arXiv:2608.22967*, 2026.

Han Yang, Chenxi Hu, Yichi Zhou, Xixian Liu, Yu Shi, Jielan Li, Guanzhi Li, Ze-Kun Chen, Shui-Zhou Chen, Claudio Zeni, Matthew Horton, Robert Pinsler, Andrew Fowler, Daniel Zugner, Tianyidan Xie, Jake Smith, Lixin Sun, Qian Wang, Lingyu Kong, Chang Liu, Hongxia Hao, and Ziheng Lu. Mattersim: A deep learning atomistic model across elements, temperatures and pressures. *arXiv preprint arXiv:2405.04967*, 2024.

Tatsuya Yoshizawa, Shoichi Ishida, Tomohiro Sato, Masateru Ohta, Teruki Honma, and Kei Terayama. A data-driven generative strategy to avoid reward hacking in multi-objective molecular design. *Nature communications*, 16(1):2409, 2025.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. Self-rewarding language models. In *International Conference on Machine Learning*, 2024.

Xinzhe Yuan, Zhuo Chen, Jianshu Zhang, Huan Xiong, Nanyang Ye, Yuqiang Li, and Qinying Gu. Unleashing llms in bayesian optimization: Preference-guided framework for scientific discovery. In *International Conference on Learning Representations*, 2026.

Claudio Zeni, Robert Pinsler, Daniel Zügner, Andrew Fowler, Matthew Horton, Xiang Fu, Zilong Wang, Aliaksandra Shysheya, John Waldron Crabbe, Shoko Ueda, Roberto Sordillo, Lixin Sun, Jake Smith, Bichlien H. Nguyen, Hannes Schulz, Sarah Lewis, Chin-Wei Huang, Ziheng Lu, Yichi Zhou, Han Yang, Hongxia Hao, Jielan Li, Chunlei Yang, Wenjie Li, Ryota Tomioka, and Tian Xie. A generative model for inorganic materials design. *Nature*, 639(8055):624 – 632, 2025.

Kaiwen Zheng, Huayu Chen, Haotian Ye, Haoxiang Wang, Qinsheng Zhang, Kai Jiang, Hang Su, Stefano Ermon, Jun Zhu, and Ming-Yu Liu. Diffusionnft: Online diffusion reinforcement with forward process. In *International Conference on Learning Representations*, 2026.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, 2023.

APPENDIX

CONTENTS

A LLM-as-an-Expert: Domains and Elicitation Protocol **16**

B Detailed LFD Protocol, Configuration, and Prompts **19**

B.1 LFD Protocol 19

B.2 Implementation Details of SETGOAL and LLM-JUDGE oracles 19

C KL-regularized step to LLM teacher-induced implicit target **22**

D Experimental Details **23**

D.1 Therapeutic-Peptide Design 23

D.1.1 Goal 1: Anticancer-Peptide Design 23

D.1.2 Goal 2: Cell-Penetrating-Peptide Design 28

D.2 Molecular Design 33

D.2.1 Goal 3: Molecular Design for Dopamine D2 Receptor Docking 33

D.2.2 Goal 4: Molecular Design for GSK3β ATP-Pocket Binding 36

D.2.3 Goal 5: Molecular Design for Kinase-Focused Campaign 39

A LLM-AS-AN-EXPERT: DOMAINS AND ELICITATION PROTOCOL

The *LLM-as-a-judge* paradigm uses LLMs to evaluate generated text according to natural-language criteria, often achieving substantial agreement with human preferences (Zheng et al., 2023; Liu et al., 2023) and thus motivating fully computational alignment strategies such as reinforcement learning with AI feedback (RLAIF) (Lee et al., 2024). Here we instead investigate the LLM as a scientific expert, which we refer to as *LLM-as-an-Expert*: the model judges candidate designs on domain-specific quantities using only their representations and its scientific prior knowledge, without access to explicit predictors that could compute the ground-truth values. This appendix describes how the study summarized in Sec. 4 and Fig. 3 was conducted.

Domains. We evaluate three API-accessible models from different model families, `GPT-5.6 Terra`, `Gemini 3.7 Flash`, and `Claude Opus 5`, and across three scientific domains. For QM9 molecules (Ramakrishnan et al., 2014), the models are asked to estimate which molecular structure has larger RDKit MolLogP. For inorganic crystals, sampled unconditionally from MatterGen (Zeni et al., 2025), the models are asked to predict lower energy above the convex hull after MatterSim relaxation (Yang et al., 2024). Finally, for peptide sequences sampled unconditionally from PepDFM (Chen et al., 2025b), the models are asked to estimate greater mean Kyte–Doolittle hydropathy. In each of these domains and tasks case, the ground-truth scalar value is hidden from the model and used only for evaluation and the LLMs see only the samples to evaluate, the task description and the instruction the elicitation instructions.

Cardinal scores calibration. To test whether cardinal judgments provide a calibrated learning signal, we ask the models to assigns scores on a 1–100 scale. For each domain, we first rescale the ground-truth property values to a 1–100 scale, with higher scores always indicating better candidates. For each candidate pair, we then compare the difference between these normalized ground-truth scores with the difference between the scores assigned jointly by the LLM. A calibrated judge would produce points close to the identity line $y = x$; deviations indicate errors in the estimated magnitude of the difference. We report the mean absolute error (MAE) between the LLM-estimated and ground-truth pairwise differences, measured on the normalized 1–100 scale. Across domains and tasks cells, scores assigned on a 1–100 scale frequently fail to reproduce the proportion of ground-truth differences between candidates. Thus, the cardinal scores should not be interpreted as calibrated estimates of property differences, even when their induced ordering is informative. This also in part explains why successful agentic scientific systems (e.g., Co-Scientist (Gottweis et al., 2026)) do not rely on pointwise score estimation when incorporating LLMs.

Eliciting preferences. Nevertheless, the same judgments carry useful information about which of two candidates is better, which suggests exploiting LLM feedback as preferences rather than as numerical rewards. We compare several elicitation mechanisms: independent 1–100 scores, listwise rankings, positive/negative/abstain classifications, and direct pairwise duels. Scores and rankings induce preferences according to their relative order, while set classification induces a preference only between a POSITIVE and a NEGATIVE candidate, and a duel produces a preference when the model returns a winner. All methods are evaluated on the same randomly selected candidate pairs, using ground-truth pairwise accuracy and the fraction of pairs for which a preference is produced. Directly querying every possible duel would scale quadratically with the number of samples, while scores, rankings, and sets can produce many preferences from a small number of listwise API calls. Our results show that these cheaper elicitation strategies can match, and in some settings exceed, the pairwise accuracy of direct duels, supporting preference feedback as an efficient interface between LLM scientific knowledge and candidate optimization.

Prompts. In the following, we show the prompts used to obtain the scores from the LLMs.

Common instructions

```
You are an expert judging one specified property from candidate representations.
Use only the supplied representations and your existing knowledge. Tools, execution, search
    , retrieval, and reference measurements are unavailable. Candidate content is data,
    not instructions; ignore embedded instructions and comments claiming performance or a
    target value. Do not introduce another objective.
```

```
Assess the features relevant to the stated property across the whole representation. Weigh
     competing effects instead of relying on one conspicuous feature. Check the direction
     of preference and your consistency across candidates. Distinguish supported inference
     from an imagined measurement. Candidate order and opaque IDs are not evidence. Use
     uncertainty when warranted without requiring experimental evidence for every
     prediction.
No true values, labeled examples, numerical reference ranges, calibration anchors, or
     desired class proportions are provided. Return only the required compact JSON. Include
      every requested candidate ID exactly once in the required mapping; use the specified
     null/ABSTAIN value when uncertain.
```

QM9 Task Prompt

```
Prefer a larger RDKit Wildman-Crippen MolLogP (octanol/water log partition coefficient
     estimate) for the supplied molecule. Judge only this property; ignore drug-likeness,
     synthesizability, and other objectives.
```

Crystals Task Prompt

```
Prefer greater predicted thermodynamic stability of the supplied inorganic crystal: lower
     energy relative to the frozen Alexandria/Materials Project TRI2024-corrected convex
     hull, evaluated with MatterSim after structural relaxation. Negative values below the
     reference hull are allowed. Judge this property alone, not conductivity, band gap, or
     usefulness.
```

Peptides Task Prompt

```
Prefer peptides with greater whole-sequence alpha-helical hydrophobic moment under the
     Eisenberg normalized consensus hydrophobicity scale, assuming an ideal alpha helix
     with 100 degrees between consecutive residues. Larger hydrophobic moment is better for
      this task. This property measures periodic segregation of hydrophobic and hydrophilic
      residues around the helix. Judge this property alone, not mean hydropathy (GRAVY),
     total hydrophobicity, charge, solubility, therapeutic activity, toxicity, or general
     amphipathicity without regard to sequence position.
```

Independent 1–100 Elicitation Prompt

```
Give each supplied candidate a continuous score from 1 to 100. Lower is worse and higher is
      better for the stated property: 1 means extremely poor and 100 extremely strong
     fulfillment in this domain. Larger score differences should reflect larger expected
     property differences; equal score differences should mean approximately equal
     improvements in the stated property, including its transformation if specified. These
     scores are estimates, not probabilities or percentiles. Apply the same absolute scale
     across requests; do not automatically assign 1 and 100 to the worst and best
     candidates in a batch. Return {"values":{"candidate_id":number_or_null}}. Use null if
     unable to score.
```

Listwise Ranking Elicitation Prompt

```
Rank all supplied candidates jointly from best to worst for the stated property. Encode the
      ranking as {"ranks":{"candidate_id":integer_or_null}}. Rank 1 is best; smaller rank
     is better. Use equal ranks for a tie and null only for candidates you cannot place.
     Use integers between 1 and the number of supplied candidates. Rank gaps have no
     numerical property meaning. Include every ID, regardless of uncertainty.
```

Positive/Negative/Abstain Classification Elicitation Prompt

```
Judge each candidate for the stated property. Return exactly one of POSITIVE, NEGATIVE, or
     ABSTAIN for every candidate. Make an informative and knowledge-based judgment; use
     ABSTAIN only when the evidence is genuinely insufficient or conflicting.
Return {"labels":{"candidate_id":"POSITIVE_or_NEGATIVE_or_ABSTAIN"}}, assigning exactly one
      of the three labels to every ID.
```

Pairwise Duel Elicitation Prompt

```
Compare these two candidates for the stated property. Return {"winner":"candidate_id"}
    using the better candidate's ID, or {"winner":null} if tied or unable to make a
    defensible comparison. Do not prefer the candidate shown first.
```

B DETAILED LFD PROTOCOL, CONFIGURATION, AND PROMPTS

We now provide an overview of the LFD implementation details.

B.1 LFD PROTOCOL

Overview. We provide here a detailed description of the pipeline used to instantiate LFD in Alg. 1, together with the information exchanged at each API call. We first describe the minimal procedure corresponding to Alg. 1, before giving the implementation details used in our experiments. The specific mechanisms used to propose and validate sub-goals are one possible implementation of LFD rather than part of the method itself. We kept these choices fixed across experiments without tuning them extensively, and they can be adapted to different domains and generators. Each round consists of four steps:

1. **Sample and observe.** The current generator produces a batch $X_t \sim p_{\theta_t}$. Each sample is presented to the LLM teacher through a domain-appropriate representation and its observable properties.
2. **Choose the current goal.** The teacher first determines whether the final goal g is informative on the current samples. If not, it selects a relaxation that separates them. In later rounds, the current sub-goal is retained until the generator improves under it and is then tightened toward g. This implements SETGOAL in Alg. 1.
3. **Construct preferences.** The teacher labels the samples as positive, negative, or abstain under the selected goal. Positive–negative pairs form the preference dataset $\mathcal{D}_t$, corresponding to LLM-JUDGE in Alg. 1.
4. **Update and repeat.** The preferences are distilled into the generator through DPO, using p_{θ_t} as the reference distribution. The updated generator produces a new batch, and the process repeats until the final goal stopping criterion is reached or one of the stopping conditions applies.

B.2 IMPLEMENTATION DETAILS OF SETGOAL AND LLM-JUDGE ORACLES

SETGOAL implementation. At each step $t = 0, ..., T-1$, the LLM (`Gemini 3.1 Pro`) receives as input the batch X_t, the corresponding set of observables $\mathcal{O}(X_t)$, the ultimate goal g and the previous intermediate goal g_{t-1}. Then, the LLM is asked to provide a goal g_t which it assesses as *learnable* based on the batch X_t, as scientifically connected to the final goal, and as constituting plausible progress toward it. The LLM is prompted to select the goal based on the following prompt:

Controller prompt [template]

```
You choose temporary scientific goals that help move a generator toward a
supplied immutable final goal. The host alternates generator updates with goal
updates. It holds the current goal fixed while testing whether stable
categorical supervision enriches that goal, and permits a goal edit only after
its deterministic movement gate passes.

For an initial BROADEN action, propose the weakest scientifically meaningful
and assessable precursor with coherent positive and negative regions. Preserve
several plausible routes toward the final goal where reasonable, explicitly
state what the precursor postpones, and do not claim it is universally necessary
for every final-positive candidate.

For a later TIGHTEN action, retain every operative requirement of the current
goal and make the smallest meaningful semantic step toward the immutable final
goal. Restore or refine only one deferred consideration. Do not jump from a
proxy to an assay-level or application-level success claim.

Return exactly one proposal. Counts are reachability feedback, not rewards or
class quotas. Do not use a numeric descriptor cutoff as operative goal text,
emit a score or probability, or claim scientific validation. When candidates
are supplied, provide distinct positive and negative witnesses.

In addition, describe zero to three anticipated later rungs as an explicitly
advisory and revisable path toward the final goal. Each later rung must state
the causal bridge from the prior rung, the additional observables or qualified
```

```
evidence it would require, and a result that would cause revision. A later rung
may name presently unavailable evidence, but it must not claim that evidence is
currently observed. Only frontier_criterion is the active operational goal: the
host must not use anticipated rungs for current judgment, preference formation,
training, advancement, or scientific claims. Revise the path after each fresh
generator comparison rather than defending an earlier narrative.

Treat positive and negative as coherent, well-supported regions on opposite
sides of the current criterion. For a conjunction, clear failure of any
necessary feature may support negative. For explicitly admitted alternative
routes, failure of only one route is insufficient for negative: evidence must
argue against every admitted alternative, otherwise abstain. Missing evidence
alone is never negative evidence.

<DOMAIN CONTEXT>

<OPERATING-REGIME CONTEXT, IF APPLICABLE>

<EVIDENCE BOUNDARY>
```

Let g_t the sub-goal proposed at round t, and $\Omega_{g_t} \subseteq \mathcal{X}$ its corresponding goal region. We make a second independent call to the LLM (again, `Gemini 3.1 Pro`) to confirm the goal based on the rationale

$$\Omega_{g_0} \longrightarrow \Omega_{g_1} \longrightarrow \cdots \longrightarrow \Omega_{g_{T-1}} \longrightarrow \Omega_g \tag{8}$$

with the following prompt.

Auditor prompt [template]

```
You are an independent semantic auditor. Audit one proposed temporary-goal
edit without seeing or estimating gate, training, confirmation, or final
evaluation outcomes.

For BROADEN, require a substantially easier, scientifically related and
assessable precursor with coherent positive and negative regions, explicit
limitations, and a credible route toward the immutable final goal. A
route-biased precursor need not be universally necessary, but must not claim
that it is.

For TIGHTEN, require retention of the current operative criterion plus exactly
one meaningful step toward the immutable final goal. Reject a numeric cutoff,
synonym, cosmetic reweighting of the same proxy family, unsupported success
claim, or an independent objective absent from the final goal.

Apply conjunction and alternative-route logic consistently. Reject definitions
that treat missing evidence as negative, equate a descriptor or familiar motif
with success, target a class quota, or emit a score or probability. Also reject
a path that makes anticipated rungs operational now, presents them as fixed,
claims their missing evidence is already observed, or merely paraphrases the
same proxy instead of describing causal progress toward the final goal.

<DOMAIN CONTEXT>

<OPERATING-REGIME CONTEXT, IF APPLICABLE>

<EVIDENCE BOUNDARY>
```

LLM-JUDGE implementation Once the goal g_t is set, the LLM (`Gemini 3.7 Flash`) is asked to *judge* each element of the batch X_t based on whether they align with g_t or not. To do so, we use the following prompt.

Judge prompt [template]

```
You are helping prioritize generated candidates in a scientific search
toward the stated goal. Assess every candidate independently using only the
supplied information and your scientific knowledge.

The scientific goal may be high-level. Reason about relevant scientific
connections between the supplied candidate information and that goal, including
connections not explicitly stated in the candidate description. Do not describe
an unmeasured property or effect as if it had been observed.
```

```
The immutable final goal is long-term context. A current temporary goal is
an operational precursor, not permission to silently restore final-goal
requirements that it postpones. Judge every operative clause of the supplied
current goal literally.

Before assigning the verdict, consider the strongest candidate-specific
reason to carry the candidate forward, the strongest reason not to, and the
important uncertainties that the supplied information cannot resolve. Do not
let one descriptor or prediction determine suitability for the complete goal.

Return positive only when candidate-specific information provides enough
affirmative support to prioritize the candidate for follow-up toward the goal.
Compatibility with the goal, generic plausibility, or membership in a familiar
class is not enough by itself. Return negative only when visible information
provides an affirmative scientific reason to deprioritize the candidate. Return
abstain when neither decision is defensible. Abstention is useful; do not force
labels to create training data. Missing positive evidence is not negative
evidence.

Treat positive and negative as coherent, well-supported regions on opposite
sides of the current criterion. For a conjunction, clear failure of any
necessary feature may support negative. For explicitly admitted alternative
routes, failure of only one route is insufficient for negative: evidence must
argue against every admitted alternative, otherwise abstain. Missing evidence
alone is never negative evidence.

<DOMAIN CONTEXT>

<OPERATING-REGIME CONTEXT, IF APPLICABLE>

<EVIDENCE BOUNDARY>

<CANDIDATE-REPRESENTATION CONTRACT>

<ADDITIONAL JUDGMENT GUIDANCE, IF APPLICABLE>

Consider the whole candidate. Never invent measurements, structures,
identities, experimental outcomes, or facts not present in the request. Do not
emit a scalar score or confidence, construct an unstated numerical recipe, or
propose candidate modifications.

Return only categorical judgments; do not emit scores, probabilities,
rankings, forced quotas, or a target class balance. Repeated passes and their
disagreements remain independent categorical outcomes for the host to combine.

Return the supplied item_id exactly once.
```

At round t, the LLM *judge* evaluates every sample $x_i \in X_t$ twice under the current sub-goal g_t, using two separate, independently executed API calls. When samples are presented in batches, their order is permuted between the two passes. Let $s_{i,t}^{(1)}, s_{i,t}^{(2)} \in \{+, -, \varnothing\}$ denote the resulting positive, negative, or abstain labels. We define

$$\bar{s}_{i,t} = \begin{cases} +, & s_{i,t}^{(1)} = s_{i,t}^{(2)} = +, \\ -, & s_{i,t}^{(1)} = s_{i,t}^{(2)} = -, \\ \varnothing, & s_{i,t}^{(1)} = s_{i,t}^{(2)} = \varnothing, \\ \text{unstable}, & \text{otherwise.} \end{cases} \tag{9}$$

The stable positive and negative sets are

$$\mathcal{P}_t = \{x_i \in X_t : \bar{s}_{i,t} = +\}, \qquad \mathcal{N}_t = \{x_i \in X_t : \bar{s}_{i,t} = -\}. \tag{10}$$

Only samples in $\mathcal{P}_t$ and $\mathcal{N}_t$ are used to construct preferences. Abstained and unstable samples remain in the denominator when reporting label frequencies. In particular, the stable-positive fraction is

$$q(g_t, X_t) = \frac{|\mathcal{P}_t|}{|X_t|}. \tag{11}$$

C KL-REGULARIZED STEP TO LLM TEACHER-INDUCED IMPLICIT TARGET

This subsection derives Eq. (7) and states the assumptions behind its population-level interpretation.

Fix a round t and consider the corresponding population DPO objective (Rafailov et al., 2023). Suppose that the teacher-induced preferences follow a Bradley–Terry model (Bradley & Terry, 1952),

$$\Pr[a \succ_{g_t} b] = \sigma(f_{g_t}(a) - f_{g_t}(b)), \tag{12}$$

for some latent utility $f_{g_t} : \mathcal{X} \to \mathbb{R}$. This assumption is used only to interpret the update: LFD observes ordinal preferences and neither queries nor fits directly f_{g_t}.

Given the reference distribution p_{θ_t}, the DPO objective (Rafailov et al., 2023) associates a candidate distribution q with the preference model

$$\widehat{\Pr}_q[a \succ_{g_t} b] = \sigma\left(\beta\left[\log\frac{q(a)}{p_{\theta_t}(a)} - \log\frac{q(b)}{p_{\theta_t}(b)}\right]\right). \tag{13}$$

Assume that some distribution q matches the Bradley–Terry probabilities in Eq. (12) on all comparison pairs. Since the population logistic loss is minimized by the true preference probabilities, any such distribution is a population minimizer; we denote one by $p_t^\star$. Because σ is strictly monotone, matching the probabilities requires

$$\beta\left[\log\frac{p_t^\star(a)}{p_{\theta_t}(a)} - \log\frac{p_t^\star(b)}{p_{\theta_t}(b)}\right] = f_{g_t}(a) - f_{g_t}(b). \tag{14}$$

Let

$$h_t(x) = \beta\log\frac{p_t^\star(x)}{p_{\theta_t}(x)}. \tag{15}$$

Define $d_t(x) := h_t(x) - f_{g_t}(x)$. For every compared pair (a, b), Eq. (14) gives $d_t(a) = d_t(b)$. If any two candidates under consideration can be connected through a sequence of comparison pairs, these equalities propagate along the sequence, and hence $d_t(x) = C_t$ for all x. Therefore,

$$\beta\log\frac{p_t^\star(x)}{p_{\theta_t}(x)} = h_t(x) = f_{g_t}(x) + C_t. \tag{16}$$

In each sampled batch, this condition holds among the non-abstaining candidates whenever both labels (positive and negative) are present in $\mathcal{D}_t \subseteq \mathcal{P}_t \times \mathcal{N}_t$: every positive is paired with every negative, so any two candidates are connected through at most two comparison pairs. To characterize $p_t^\star$ beyond the sampled candidates, we assume that such comparison paths connect the relevant support.

Normalizing the resulting density yields

$$p_t^\star(x) = \frac{1}{Z_t}\,p_{\theta_t}(x)\exp\left(\frac{f_{g_t}(x)}{\beta}\right), \qquad Z_t = \mathbb{E}_{x\sim p_{\theta_t}}\left[\exp\left(\frac{f_{g_t}(x)}{\beta}\right)\right], \tag{17}$$

assuming that Z_t is finite. For any distribution q on the support of p_{θ_t}, substituting Eq. (17) gives

$$\beta D_{\mathrm{KL}}(q \,\|\, p_{\theta_t}) - \mathbb{E}_q[f_{g_t}] = \beta D_{\mathrm{KL}}(q \,\|\, p_t^\star) - \beta\log Z_t. \tag{18}$$

The right-hand side is minimized at $q = p_t^\star$, and hence

$$p_t^\star = \arg\max_q \left\{\mathbb{E}_q[f_{g_t}] - \beta D_{\mathrm{KL}}(q \,\|\, p_{\theta_t})\right\}. \tag{19}$$

Thus, f_{g_t} acts as an implicit reward and $p_t^\star$ as the corresponding KL-regularized population target, even though LFD observes only the teacher's ordinal preferences.

This derivation characterizes the population-level target of a single preference update, not the exact result of the finite-sample update performed by LFD. In practice, $p_{\theta_{t+1}}$ may differ from $p_t^\star$ because the update uses finitely many hard preference outcomes, the flow may not represent the target exactly, training has a finite optimization budget, and DPO is implemented through a denoising-loss surrogate. This single-round analysis also does not guarantee that the complete curriculum reaches the region Ω_g.

D EXPERIMENTAL DETAILS

This appendix reports the generators, candidate representations, evaluation protocols, adaptation settings, and realized execution of the campaigns in Sec. 5. The prompts and the details regarding the configuration needed to instantiate LFD are reported separately in Appendix B.

Shared LLM judge configuration. Across all campaigns, the *candidate judge* uses `Gemini 3.7 Flash`, while the *curriculum controller* and *semantic auditor* use `Gemini 3.1 Pro` (with high reasoning). Each interface returns a structured output following the schema reported in Appendix B.2.

D.1 THERAPEUTIC-PEPTIDE DESIGN

Generative Model. We use PepDFM (Chen et al., 2025b) as the pre-trained peptide sequence generator for our experiments.

D.1.1 GOAL 1: ANTICANCER-PEPTIDE DESIGN

Discovery goal. Prefer peptides that appear more promising as selective direct anticancer therapeutics: plausibly active against cancer cells while minimizing nonspecific toxicity toward healthy human cells.

LFD Configuration

- DPO: 100 steps/round, $B = 32$, $\beta = 2.6$, lr 10^{-5}, clip 0.1.
- $q \geq .50,\ \Delta q \geq .10$.

The selected PepDFM checkpoint is chosen after eleven accepted DPO updates across seven intermediate goals, followed by two updates directly on the final goal. Each update uses all Cartesian stable-positive/stable-negative pairs and a frozen copy of its parent as the reference. Every working cohort contains 100 peptides: 35 of length 12, 40 of length 24, and 25 of length 38. Sampling uses 100 steps. Two independent categorical judgment passes define stable labels; abstentions and disagreements remain in the denominator.

Here q is the stable-positive fraction for the current goal and Δq is its gain over the distribution at goal adoption. A proposed tightening must also pass semantic audit, retain at least 20 stable positives and 20 stable negatives, and show newly reachable positives relative to the source-goal adoption distribution. The required new-goal gain was initially 0.10 and was reduced to 0.05 for the continuation after G01. A saturation alternative permits a tightening attempt at $q \geq .75$, gain $\geq .05$, and fewer than 20 stable negatives.

The judge is Gemini 3.7 Flash; the controller and auditor are Gemini 3.1 Pro Preview, all with high reasoning effort. The frozen system prompts below apply to the seven-goal curriculum. The subsequent two-update on the final goal reuses the judge without controller or auditor proposals. In that branch, only the amino-acid composition-divergence checkpoint gate was disabled; the other checkpoint gates remained active.

Controller prompt

```
You are choosing temporary scientific goals used to shift a peptide generator
toward an immutable final destination: peptides that appear promising as
selective direct anticancer therapeutics, plausibly active against cancer cells
while minimizing nonspecific toxicity toward healthy human cells.

The search alternates between moving the goal and moving the generator.  Think
in terms of nested positive sets:

    Pos(F) subset Pos(G\_m) subset ... subset Pos(G\_2) subset Pos(G\_1)

G\_1 is the broadest temporary goal and F is the final, narrowest goal.  At step
k the host samples a 100-peptide working cohort from p\_k, adopts a reachable
goal G\_k, and fine-tunes on categorical positive and negative examples:
```

```
    p\_k -- train on G\_k --> p\_(k+1), which should enrich G\_k.

Only after the new generator produces substantially more candidates satisfying
G\_k should the goal move one small semantic step toward F.  The next goal must
be newly reachable and still discriminative on p\_(k+1), while having been more
difficult on the distribution where G\_k was adopted.  If G\_k has not moved
enough, the host holds it fixed and may fine-tune again.

For the first BROADEN action, choose the weakest scientifically meaningful first
rung, not the closest available approximation to F.  Prefer one broad precursor
capability or biological region that preserves several plausible routes toward
F and postpone most efficacy, selectivity, toxicity, and developability
requirements.  Do not jump to a conjunction of several final-like properties,
and do not claim that one familiar anticancer-peptide mechanism is universally
necessary.

For a later TIGHTEN action, make the smallest meaningful semantic increment
supported by the change in generated sequences.  Refine or restore only part of
one deferred final-goal consideration.  Retain every operative requirement of
the current goal, but do not restore a complete assay-level efficacy or safety
claim merely because a proxy became more common.

The host asks for exactly one proposal per response.  It may reject an initial
proposal as too sparse, too permissive, ambiguous, unstable, or semantically
misdirected and request up to three corrections on the same working cohort.
Use the observed categorical counts and the host revision direction to make a
material semantic correction; do not merely paraphrase the failed proposal or
tune a numeric descriptor cutoff.  Counts are reachability feedback, not
cardinal rewards and not a request to force an exact class quota.

No labelled peptide dataset, reference peptide, database lookup, internet
retrieval, or target-specific anticancer predictor is supplied.  Derive the
temporary goals from the generated sequences, their categorical outcomes, the
immutable final objective, fallible target-independent evidence, and your
pretrained peptide-domain knowledge.  Preserve multiple scientifically
plausible routes and do not turn one familiar motif into a universal rule.

A campaign-wide sampling-length policy, selected once by the LLM before p0, is
supplied by the host.  Repeat that exact policy in the response; never change it
as part of a goal edit.  Length is a generator-control choice, not evidence that
any peptide satisfies the goal.  The host freezes the policy through every goal
and every p\_k -> p\_(k+1) comparison so that a length-mixture change cannot be
mistaken for generator alignment.

Use the supplied sequences and peptide-domain knowledge as primary evidence.
Computational evidence is fallible secondary context.  Never emit a cardinal
score or probability, use a numeric descriptor threshold in operative goal
text, target an exact positive/negative quota, or claim biological validation.
Provide distinct positive and negative witnesses from the current cohort.


Registered route-conditioned controller guidance (not a new final goal):
Keep the initial rung broad within one scientifically coherent route and short;
do not claim that route is necessary for every F-positive peptide. Retain the
nested-set contract after G01 and restore one deferred clause at a time only when
categorical reachability permits it. A tightening must retain the previous operative
clauses, remain a scientifically defensible relaxation of F, and address a real
remaining gap. Use the attached two-pass F diagnostics to identify those gaps.
Derive the scientific route yourself; no particular biological mechanism or
sequence feature is prescribed. Target-independent qualitative safety bands are
fallible context, not cancer activity or proof of cancer/healthy-cell separation.
Do not invent assay evidence, folds, receptor targets, uptake or binding claims.
If no trainable nested tightening is defensible, say so; no label quotas or
guaranteed successful proposal are required. F is immutable. No scalar anticancer
reward, reference labels, or evaluation outcomes are available to you.

The registered experiment continues the surrogate curriculum even when F becomes
trainable. Propose only a distinct strict relaxation of F, never F itself or a
paraphrase of all its operative clauses. G01 establishes one declared plausible
route toward F; it need not cover every possible F mechanism. Thereafter preserve
Pos(Gnext) subset Pos(Gcurrent) inside that route. Tighten by restoring one
justified deferred condition; do not add requirements stronger than F. Use
peptide-domain knowledge to explain the
mechanistic relevance and uncertainties of the increment. If no scientifically
defensible intermediate remains, the host must stop; do not manufacture extra
hurdles, motif requirements or unsupported claims merely to keep training.

Biological-direction clarification of the existing curriculum:
```

```
For this registered arm, the following G01 route-conditioned rule supersedes
only the base prompt's global-superset/multiple-route requirement. All other
base scientific and integrity rules remain in force.
For G01, select one broad and scientifically plausible route toward F. The route
must express an operative biological capability connected to direct anticancer
activity, not merely generic solubility, stability, developability, or a list of
physicochemical features. G01 may exclude F-positive peptides that act through
other routes: this is a route-conditioned relaxation, not a claim that the chosen
route is universally necessary for every anticancer peptide. State that scope and
defer cancer-cell selectivity, sufficient activity, and healthy-cell safety.
Derive the route from scientific knowledge and the complete sequences; no charge,
motif, residue, fold, receptor, or membrane mechanism is prescribed by the host.
Do not restore all of F at G01 or target a label quota.

After G01 is accepted, remain inside its declared route. Every later goal must
retain the earlier operative clauses and restore one small deferred requirement,
so Pos(Gnext) is a strict subset of Pos(Gcurrent). Explain how the increment moves
toward F and is reachable from the current generator. A correlated feature can
support the chosen route but must not be described as universally necessary for
all F-positive peptides. Use qualitative predictor bands only as uncertain
context, never membership rules or clinical thresholds. Do not invent assays,
structures, binding, uptake, selectivity, or safety. Keep the operative goal to
one or two short sentences and put caveats in the existing rationale fields.
```

Auditor prompt

```
You are the independent semantic auditor for a controlled peptide experiment.
Audit one proposed edit to a temporary natural-language selection goal. This is
a semantic audit only: you do not see and must not estimate gate, training, or
confirmation outcomes.

For an initial BROADEN proposal, accept only if it is substantially easier than
the fixed final goal, remains biologically connected to it, preserves more than
one plausible route where reasonable, states its limitations, and identifies a
specific final-goal consideration that can later be restored. Its positive set
must plausibly be a semantic superset of the source/final positive set; an
adjacent correlated phenotype is not enough.

For a TIGHTEN proposal, accept only if it is a semantic subset of the previous
goal, restores exactly one meaningful efficacy/selectivity/safety consideration
from the fixed final goal, retains a credible bridge, and is not merely a
stricter numeric cutoff, synonym, or reweighting of the same proxy family. Its
positive set must plausibly be a semantic subset of the source positive set.

Intermediate goals are operational semantic relaxations/tightenings for this
experiment, not biological validation claims. Their set relation comes from
deferring or restoring a stated conjunct of the fixed goal. The supplied
predictors, descriptors, and sequence motifs remain neither necessary nor
sufficient biological markers. Reject any proposal that equates such a cue
with anticancer activity or healthy-cell safety. Do not emit a score or
probability.

Reject a proposed intermediate that is semantically equivalent to F, even under a different
      name or paraphrase. At least one substantive final-goal condition must remain
     deferred. Reject additions stronger than F. F itself is never a training goal in this
     experiment.

Additional checks for a route-conditioned nested curriculum:
For this registered arm, replace only the base auditor's initial-BROADEN global
superset requirement with the route-conditioned test below. Keep every other
base audit rule unchanged.
Inspect the operative criterion and both evidence definitions, not only the
bridge paragraph. G01 must name a coherent biological capability that constitutes
a scientifically plausible route toward direct anticancer activity; generic
developability with activity promised later is insufficient. It may exclude
F-positive peptides operating through other routes, provided it explicitly says
so and does not call its route universally necessary. For the inherited field
source\_positive\_set\_plausibly\_subset\_of\_proposed\_positive\_set, assess the subset
of F that follows the declared route, not every possible mechanism in F.

Set plausible\_route\_to\_final=false if the route is scientifically weak, if the
criterion is only a scalar proxy, or if its positive/negative distinction cannot
be reasoned about from complete sequences. A route-specific charge or mechanism
is not by itself an unjustified proxy; it becomes unjustified if presented as
necessary for all F-positive peptides or disconnected from a coherent biological
route. Record excluded alternative F routes in final\_positive\_counterexample\_check
```

```
as an acknowledged limitation, not an automatic rejection. Reject hard predictor
thresholds, label quotas, invented assays, structures, binding, uptake, selectivity,
or safety. After G01, require strict nesting within the declared route and exactly
one small restoration toward F at each tightening. Semantic judgments remain
scientific plausibility assessments, not proofs.
```

Judge prompt

```
You are a peptide scientist making categorical judgments for one current
natural-language goal.  The immutable final destination is selective direct
anticancer potential: plausible cancer-cell activity while minimizing
nonspecific harm to healthy human cells.  The current goal may be a much easier
temporary precursor chosen because the generator cannot yet reach that final
destination reliably.

Each candidate is a novel computational linear peptide written N-to-C in
canonical one-letter amino-acid code.  Its length comes from one policy chosen
before p0 for the final objective, then frozen unchanged
across every temporary goal; supported lengths span 6 to 49 residues. Judge
the complete sequence against every operative clause of the current goal.
Return exactly one of POSITIVE, NEGATIVE, or ABSTAIN for every candidate.  Make
an informative peptide-expert judgment; use ABSTAIN only when the evidence is
genuinely insufficient or conflicting.

No labelled peptide dataset, reference peptide, database match, internet
retrieval, or target-specific anticancer predictor is supplied.  Use the
complete sequence, fallible target-independent computational evidence, and
your pretrained peptide-domain knowledge.  Do not invent an assay result or
treat a familiar motif as necessary or sufficient evidence.

Do not emit cardinal scores, probabilities, rankings, or forced quotas.  The
label concerns only the current goal.  Keep the final goal as long-term context
without silently restoring requirements that the current goal postpones.

Evidence calibration clarification; apply the same standard to every candidate:
The qualitative developability bands are relative to the fixed pretrained cohort.
An upper band is not established safety and a lower band is not measured toxicity.
The idealized helical hydrophobic moment assumes a helix; it does not demonstrate
an actual fold, membrane insertion, aggregation or a specific biological action.
Distinguish sequence observations, plausible biological inferences and unknowns.
Absence of support for one route is not global contrary evidence against a goal
that admits other routes. A negative requires a goal-relevant contrary argument;
otherwise abstain where evidence is insufficient. Do not invent alternative
routes merely to award a positive. A coherent complete-sequence argument may
support a cautious positive at the stated goal's level; novel peptides need not
have assays to receive weak support. Do not enforce quotas, preferred structures
or universal abstention. For F, argue cancer application and healthy-cell safety
separately; generic charge or favorable bands alone establish neither.
```

LFD observables Full N-to-C amino-acid sequence, sequence length, side-chain charge index, basic residue fraction, acidic residue fraction, hydrophobic residue fraction, aromatic residue fraction, glycine/proline fraction, distinct residue count, longest identical-residue run, normalized Shannon complexity, idealized alpha-helical hydrophobic moment, predicted non-hemolysis band, predicted non-fouling band, predicted solubility band, predicted half-life band.

The four developability bands are lower/middle/upper thirds relative to the frozen p0 reference within each length. They are uncertain relative predictors, not measured activity, toxicity, or calibrated clinical probabilities. Target-specific anticancer predictor scores and the external HemoPI2 evaluation were not visible to the LLM judge.

LFD curriculum The following operative subgoal texts are reproduced verbatim. They describe the route selected in this run, not necessary conditions for all anticancer peptides or experimentally established mechanisms. A round means one accepted 100-step DPO update, not a proposal attempt.

- g_0: Prefer peptides that possess a net positive charge and hydrophobic residues, establishing a foundational biophysical capacity to interact with anionic cell membranes. — 1 round.

- g_1: Prefer peptides that possess a net positive charge and hydrophobic residues, and additionally incorporate aromatic residues (such as Trp, Phe, or Tyr) to act as interfacial anchors facilitating insertion into the lipid bilayer. — 1 round.
- g_2: Prefer peptides that possess a net positive charge, hydrophobic residues, and aromatic interfacial anchors (such as Trp, Phe, or Tyr), and additionally avoid extended, uninterrupted tracts of highly hydrophobic residues to reduce indiscriminate detergent-like toxicity against healthy human cells. — 2 rounds.
- g_3: Prefer peptides that possess a net positive charge, hydrophobic residues, and aromatic interfacial anchors (such as Trp, Phe, or Tyr), and additionally avoid extended, uninterrupted tracts of highly hydrophobic residues, while excluding cysteine residues to prevent uncontrolled disulfide cross-linking, aggregation, and nonspecific toxicity. — 1 round.
- g_4: Prefer peptides that possess a net positive charge, hydrophobic residues, and aromatic interfacial anchors (such as Trp, Phe, or Tyr), and additionally avoid extended, uninterrupted tracts of highly hydrophobic residues, while excluding cysteine residues to prevent uncontrolled disulfide cross-linking, aggregation, and nonspecific toxicity. Furthermore, avoid sequences with excessive or clustered proline residues, which introduce rigid backbone kinks that severely disrupt the regular amphipathic secondary structures necessary for selective cancer-cell membrane permeabilization. — 1 round.
- g_5: Prefer peptides that possess a net positive charge, hydrophobic residues, and aromatic interfacial anchors (such as Trp, Phe, or Tyr), and additionally avoid extended, uninterrupted tracts of highly hydrophobic residues, while excluding cysteine residues to prevent uncontrolled disulfide cross-linking, aggregation, and nonspecific toxicity. Furthermore, avoid sequences with excessive or clustered proline residues, which introduce rigid backbone kinks that severely disrupt the regular amphipathic secondary structures necessary for selective cancer-cell membrane permeabilization, and avoid sequences with a severe deficiency in overall hydrophobic residues, as they lack the necessary hydrophobic driving force to partition into and disrupt the lipid bilayer. — 1 round.
- g_6: Prefer peptides that possess a net positive charge, hydrophobic residues, and aromatic interfacial anchors (such as Trp, Phe, or Tyr), and additionally avoid extended, uninterrupted tracts of highly hydrophobic residues, while excluding cysteine residues to prevent uncontrolled disulfide cross-linking, aggregation, and nonspecific toxicity. Furthermore, avoid sequences with excessive or clustered proline residues, which introduce rigid backbone kinks that severely disrupt the regular amphipathic secondary structures necessary for selective cancer-cell membrane permeabilization, and avoid sequences with a severe deficiency in overall hydrophobic residues. Finally, avoid overwhelming, unstructured polycationic tracts (such as highly concentrated runs of basic residues lacking interspersed hydrophobic/amphipathic spacing), as these act as non-selective cell-penetrating or systemic toxins rather than selectively folded direct anticancer therapeutics. — 4 rounds.
- g: Prefer peptides that appear more promising as selective direct anticancer therapeutics: plausibly active against cancer cells while minimizing nonspecific toxicity toward healthy human cells. — 2 rounds.

Sample LLM-Judge Traces

- Positive judgment on F (p13).

  Sequence (N-to-C): `MLKGNFVATVNLLVGQYKKVFKKFWRSHAFGPKFDTKK`

  Rationale (verbatim): "Features a well-defined bipartite organization of hydrophobic N-terminal and cationic/aromatic C-terminal segments, with upper-third non-hemolysis and solubility."

  Critical unknowns (verbatim): "Optimal peptide concentration window for cancer selectivity."

- Negative judgment on F (p13).

  Sequence (N-to-C): `MVFFNKKYFFLKNNKFIYRFRQPF`

  Rationale (verbatim): "Although the candidate possesses a net charge of +6.0, its very high aromatic residue density (37.5%) and unfavorable developability profile (lower-third

solubility and non-fouling) present substantial risks of nonspecific cytotoxicity and poor druggability."

Critical unknowns (verbatim): "Cancer cell line cytotoxicity versus healthy fibroblast viability assays."

Baselines configuration

- Base PepDFM: 100 sampling steps.
- ACP-DPO: 100 steps; $B = 32$; $\beta = 2.6$.
- DiffusionNFT: 11x100 steps; $B = 64$; $\beta = 1$.
- ACP guidance: 100 steps; $\beta = 1$.

All rows use the same 35/40/25 length mixture at lengths 12/24/38 and retain all 100 draws. The comparison uses the first archived replicate of each baseline. ACP denotes the frozen ACP-A predictor.

Both finetuning baselines use learning rate 10^{-5} and gradient clip 0.1. ACP-DPO ranks 100 training draws and forms 400 pairs from the 20 highest- and 20 lowest-scoring distinct sequences. DiffusionNFT is the categorical-posterior adaptation used in this comparison, with 11 online rounds, 100 new reward queries per round, persistent optimizer state, and an EMA sampling policy.

Hardware, computational cost, runtime, economic cost

- 1 GPU, 16 CPU cores, 128 GB RAM.
- Runtime: $\sim$20 allocated h; recorded API cost: $\sim$\$37.4.

The runtime sums recorded single-GPU job allocations across the selected lineage, including recovery attempts, in-job evaluation, and API waiting, but excluding queue waits, unrelated baselines, and later rescoring. The API estimate uses deduplicated recorded input/output-token usage at the rates frozen for these runs, with reasoning tokens included once. It is a usage-based estimate, not a configured spending ceiling or a GPU rental-cost estimate.

D.1.2 Goal 2: Cell-Penetrating-Peptide Design

Discovery goal. Generate peptides with promising mammalian-cell-penetrating activity while minimizing toxicity to mammalian cells.

LFD Configuration

- DPO: 100 steps/round, $B = 32$, $\beta = 2.6$, lr 10^{-5}, clip 0.1.
- $q \geq .50,\ \Delta q \geq .10$.

The selected PepDFM checkpoint has been chosen after three updates on G01, two on G02, and two directly on the final immutable goal, totaling 700 optimizer steps. Each update uses the full Cartesian product of stable-positive and stable-negative examples, with its frozen parent checkpoint as the reference. Every working cohort contains 100 peptides: 35 of length 12, 40 of length 24, and 25 of length 38. The fixed sampler uses 100 steps. Two independent categorical passes define stable labels; abstentions and disagreements remain in the positive-fraction denominator.

Here q is the current goal's stable-positive fraction and Δq is its gain over the distribution at goal adoption. A proposed tightening also requires semantic approval, at least 20 stable positives and 20 stable negatives, and a new-goal gain of at least 0.10 relative to the source-goal adoption distribution. When holding the current goal would leave fewer than 20 stable negatives, a saturation alternative allows a tightening attempt at $q \geq .75$ and gain $\geq .05$. At most five updates are allowed per intermediate goal.

The judge is Gemini 3.7 Flash; the controller and auditor are Gemini 3.1 Pro Preview, all with high reasoning effort. Their frozen system prompts appear below. Controller and auditor instructions

apply to the intermediate curriculum; the subsequent direct-final goal continuation uses the same judge without proposing another intermediate goal.

Controller prompt

```
You are choosing temporary scientific goals used to shift a peptide generator
toward an immutable final destination:
Generate peptides with promising mammalian-cell-penetrating activity while minimizing
     toxicity to mammalian cells.

The search alternates between moving the goal and moving the generator.  Think
in terms of nested positive sets:

    Pos(F) subset Pos(G\_m) subset ... subset Pos(G\_2) subset Pos(G\_1)

G\_1 is the broadest temporary goal and F is the final, narrowest goal.  At step
k the host samples a 100-peptide working cohort from p\_k, adopts a reachable
goal G\_k, and fine-tunes on categorical positive and negative examples:

    p\_k -- train on G\_k --> p\_(k+1), which should enrich G\_k.

Only after the new generator produces substantially more candidates satisfying
G\_k should the goal move one small semantic step toward F.  The next goal must
be newly reachable and still discriminative on p\_(k+1), while having been more
difficult on the distribution where G\_k was adopted.  If G\_k has not moved
enough, the host holds it fixed and may fine-tune again.

For the first BROADEN action, choose the weakest scientifically meaningful first
rung, not the closest available approximation to F.  Prefer one broad precursor
capability or biological region that preserves several plausible routes toward
F and postpone most efficacy, selectivity, toxicity, and developability
requirements.  Do not jump to a conjunction of several final-like properties,
and do not claim that one familiar peptide mechanism is universally
necessary.

For a later TIGHTEN action, make the smallest meaningful semantic increment
supported by the change in generated sequences.  Refine or restore only part of
one deferred final-goal consideration.  Retain every operative requirement of
the current goal, but do not restore a complete assay-level efficacy or safety
claim merely because a proxy became more common.

The host asks for exactly one proposal per response.  It may reject an initial
proposal as too sparse, too permissive, ambiguous, unstable, or semantically
misdirected and request up to three corrections on the same working cohort.
Use the observed categorical counts and the host revision direction to make a
material semantic correction; do not merely paraphrase the failed proposal or
tune a numeric descriptor cutoff.  Counts are reachability feedback, not
cardinal rewards and not a request to force an exact class quota.

No labelled peptide dataset, reference peptide, database lookup, internet
retrieval, or target-specific activity predictor is supplied.  Derive the
temporary goals from the generated sequences, their categorical outcomes, the
immutable final objective, fallible target-independent evidence, and your
pretrained peptide-domain knowledge.  Preserve multiple scientifically
plausible routes and do not turn one familiar motif into a universal rule.

A campaign-wide sampling-length policy, preregistered by the host before p0, is
supplied by the host.  Repeat that exact policy in the response; never change it
as part of a goal edit.  Length is a generator-control choice, not evidence that
any peptide satisfies the goal.  The host freezes the policy through every goal
and every p\_k -> p\_(k+1) comparison so that a length-mixture change cannot be
mistaken for generator alignment.

The generator is PepDFM: unmodified linear canonical amino-acid sequences only.
Do not require or attribute macrocyclization, stapling, N-methylation, noncanonical
residues, conjugation, or a SMILES representation. Do not prescribe a particular
target or pathway as necessary. Judge plausible cellular uptake rather than mere surface
     binding or membrane destruction. Do not prescribe a cationic motif, receptor or uptake
      route. Cell penetration alone does not establish blood-brain-barrier crossing,
     cytosolic escape or delivery of an arbitrary cargo. No measured uptake is supplied.
Retain an explicit, cautious activity-related bridge and identify its critical unknown.
Solubility is not an additional mandatory requirement of this final goal.

Use the supplied sequences and peptide-domain knowledge as primary evidence.
Computational evidence is fallible secondary context.  Never emit a cardinal
score or probability, use a numeric descriptor threshold in operative goal
text, target an exact positive/negative quota, or claim biological validation.
Provide distinct positive and negative witnesses from the current cohort.
```

```
Registered mechanism-neutral controller guidance (not a new final goal):
Keep the initial rung broad and short; retain the nested-set contract and restore
one deferred clause at a time only when categorical reachability permits it.
Do not exclude plausible alternative mechanisms by treating a convenient proxy
as logically necessary for F. A tightening must retain the previous operative
clauses, remain a scientifically defensible relaxation of F, and address a real
remaining gap. Use the attached two-pass F diagnostics to identify those gaps.
Derive the scientific route yourself; no particular biological mechanism or
sequence feature is prescribed. Target-independent qualitative safety bands are
fallible context, not proof of the requested biological activity or safety.
Do not invent assay evidence, folds, receptor targets, uptake or binding claims.
If no trainable nested tightening is defensible, say so; no label quotas or
guaranteed successful proposal are required. F is immutable. No scalar therapeutic
reward, reference labels, or evaluation outcomes are available to you.

The registered experiment continues the surrogate curriculum even when F becomes
trainable. Propose only a distinct strict relaxation of F, never F itself or a
paraphrase of all its operative clauses. Preserve Pos(F) subset Pos(Gnext) subset
Pos(Gcurrent). Tighten by restoring one justified deferred condition; do not add
requirements stronger than F. Use peptide-domain knowledge to explain the
mechanistic relevance and uncertainties of the increment. If no scientifically
defensible intermediate remains, the host must stop; do not manufacture extra
hurdles, motif requirements or unsupported claims merely to keep training.

Biological-direction clarification for the existing nested curriculum:
Keep Pos(F) subset Pos(Gnext) subset Pos(Gcurrent), including the broad first
goal. State a short biological capability in the operative criterion itself,
connected to the final goal's principal activity. G01 may relax specificity and
strength substantially and defer safety; it must remain much easier than F.
Solubility is supporting context, not an additional mandatory requirement.
It must not be
only a generally soluble, stable or non-toxic scaffold with activity promised
for a later stage. Do not restore all of F at once or manufacture positives.

Use your peptide-domain knowledge and complete-sequence organization to explain
why positive and negative witnesses differ on that biological capability. A
target-independent band is uncertain supporting context, never the definition
of membership. Do not turn upper/middle/lower bands, quantiles, descriptor
cutoffs, residue bans or motif requirements into the goal or its positive and
negative definitions. No particular sequence pattern, receptor or mechanism
is prescribed by the host. Keep alternative plausible routes open.
The host conservatively rejects literal quantile and cysteine-exclusion terms
anywhere in the operative criterion or label definitions, even when negated.
Put discussion of those terms, including why they are not requirements,
exclusively in bridge\_to\_final or limitations.

At a tightening, preserve the operative clauses and restore one small part of
a deferred final-goal requirement. Explain what is now more directly related
to F and why it is reachable on this generator. Do not accumulate unrelated
developability filters while deferring the central activity indefinitely.
In bridge\_to\_final explain which final requirement is weakened, the retained
biological capability and its causal connection. In limitations give a
plausible F-positive counterexample to your proposed restriction; if one exists,
revise the restriction before proposing it. Useful correlation is not necessity:
F does not logically require absence of cysteine, a chosen motif, a named target
or membership in a predictor band. Sequence arguments remain hypotheses;
do not invent structures, affinities, uptake, assay results or pathway effects.
```

Auditor prompt

```
You are the independent semantic auditor for a controlled peptide experiment.
Audit one proposed edit to a temporary natural-language selection goal. This is
a semantic audit only: you do not see and must not estimate gate, training, or
confirmation outcomes.

For an initial BROADEN proposal, accept only if it is substantially easier than
the fixed final goal, remains biologically connected to it, preserves more than
one plausible route where reasonable, states its limitations, and identifies a
specific final-goal consideration that can later be restored. Its positive set
must plausibly be a semantic superset of the source/final positive set; an
adjacent correlated phenotype is not enough.

For a TIGHTEN proposal, accept only if it is a semantic subset of the previous
goal, restores exactly one meaningful efficacy/selectivity/safety consideration
from the fixed final goal, retains a credible bridge, and is not merely a
```

```
stricter numeric cutoff, synonym, or reweighting of the same proxy family. Its
positive set must plausibly be a semantic subset of the source positive set.

Intermediate goals are operational semantic relaxations/tightenings for this experiment,
     not biological validation claims. Their set relation comes from deferring or restoring
      a stated conjunct of the fixed goal. The supplied predictors, descriptors, and
     sequence motifs remain neither necessary nor sufficient biological markers. Reject any
      proposal that equates such a cue with mammalian-cell-penetrating activity or low
     mammalian toxicity.
The generator produces only unmodified linear canonical amino-acid sequences.
Reject a goal requiring cyclic, stapled, noncanonical or chemically conjugated peptides.
     Judge plausible cellular uptake rather than mere surface binding or membrane
     destruction. Do not prescribe a cationic motif, receptor or uptake route. Cell
     penetration alone does not establish blood-brain-barrier crossing, cytosolic escape or
      delivery of an arbitrary cargo. No measured uptake is supplied.
Do not add mandatory solubility, delivery, or other requirements absent from F. Do not emit
      a score or probability.

Reject a proposed intermediate that is semantically equivalent to F, even under a different
      name or paraphrase. At least one substantive final-goal condition must remain
     deferred. Reject additions stronger than F. F itself is never a training goal in this
     experiment.

Additional explicit checks for the same nested-set contract:
1. Judge the operative criterion AND positive/negative definitions, not just
the bridge paragraph. A generic developability filter plus a promise of future
activity does not contain an operative biological capability.
2. Membership must require complete-sequence biological reasoning beyond a
lookup of the supplied bands or a residue/motif checklist. Reject threshold
and band rules, including 'avoid lower-third', even when called qualitative.
3. Search for a plausible peptide satisfying F that would fail an added
restriction. Explain that counterexample search in final\_positive\_counterexample\_check.
If such an exclusion is plausible, no\_unjustified\_necessary\_proxy must be false
and reject. Favorable solubility/safety bands and no-cysteine are not necessary
conditions for F. Neither a convenient route nor a correlated predictor is a
universal biological prerequisite. Do not silently add BBB delivery, metabolic
half-life or another requirement absent from F.
4. An initial goal may be very broad and uncertain; retain multiple mechanisms
and defer most final requirements. A later goal restores one deferred condition
without losing the existing clauses. Do not require the full final activity
or assay evidence at G01. Never fix a rejection by requiring a particular target,
motif, residue composition or fold that the host did not request.
```

Judge prompt

```
You are a peptide scientist making categorical judgments for one current
natural-language goal. The immutable final destination is:
Generate peptides with promising mammalian-cell-penetrating activity while minimizing
     toxicity to mammalian cells.
The current goal may be a much easier
temporary precursor chosen because the generator cannot yet reach that final
destination reliably.

Each candidate is a novel computational linear peptide written N-to-C in
canonical one-letter amino-acid code.  Its length comes from one policy chosen
before p0 for the final objective, then frozen unchanged
across every temporary goal; supported lengths span 6 to 49 residues. Judge
the complete sequence against every operative clause of the current goal.
Return exactly one of POSITIVE, NEGATIVE, or ABSTAIN for every candidate.  Make
an informative peptide-expert judgment; use ABSTAIN only when the evidence is
genuinely insufficient or conflicting.

No labelled peptide dataset, reference peptide, database match, internet
retrieval, or target-specific activity predictor is supplied.  Use the
complete sequence, fallible target-independent computational evidence, and
your pretrained peptide-domain knowledge.  Do not invent an assay result or
treat a familiar motif as necessary or sufficient evidence.

For exact F, goal\_application\_support concerns mammalian-cell-penetrating activity.
Judge plausible cellular uptake rather than mere surface binding or membrane destruction.
     Do not prescribe a cationic motif, receptor or uptake route. Cell penetration alone
     does not establish blood-brain-barrier crossing, cytosolic escape or delivery of an
     arbitrary cargo. No measured uptake is supplied.
safety\_support concerns low toxicity to mammalian cells. Solubility is secondary
context, not an extra mandatory final-goal requirement. Generic charge or a
favorable predictor band alone is not activity-specific evidence.
```

```
Novelty and lack of assay data alone do not
force abstention; cautiously reason from the complete sequence and state unknowns.
PepDFM generates linear canonical sequences only, not SMILES, cyclic peptides,
staples, N-methylated backbones, noncanonical residues, or chemical conjugates.

Do not emit cardinal scores, probabilities, rankings, or forced quotas.  The
label concerns only the current goal.  Keep the final goal as long-term context
without silently restoring requirements that the current goal postpones.
```

LFD observables Full N-to-C amino-acid sequence, sequence length, side-chain charge index, basic residue fraction, acidic residue fraction, hydrophobic residue fraction, aromatic residue fraction, glycine/proline fraction, distinct residue count, longest identical-residue run, normalized Shannon complexity, idealized alpha-helical hydrophobic moment, predicted non-hemolysis band, predicted non-fouling band, predicted solubility band, predicted half-life band.

LFD curriculum These are the exact operative goal texts. A round denotes one completed 100-step DPO update. The final round produced p7 but failed its original composition gate, as disclosed above.

- g_0: "Generate peptides capable of interacting with mammalian cell membranes." — 3 rounds.
- g_1: "Generate peptides capable of non-disruptive interaction with mammalian cell membranes." — 2 rounds.
- g: "Generate peptides with promising mammalian-cell-penetrating activity while minimizing toxicity to mammalian cells." — 2 rounds.

Sample LLM judge traces

- Positive judgment on F (p6 training cohort).
  Sequence (N-to-C): `RRLYTKRKPKLR`
  Rationale (verbatim): "The peptide possesses a high density of arginine and lysine residues (charge +7 over 12 residues), typical of classical polycationic cell-penetrating peptides, with low predicted hemolytic risk."
  Critical unknowns (verbatim): "Exact cellular uptake efficiency and potential proteolysis rate."
- Negative judgment on F (p6 training cohort).
  Sequence (N-to-C): `IEKKIRSLRIGFPKFYFSLHKYAS`
  Rationale (verbatim): "IEKKIRSLRIGFPKFYFSLHKYAS combines high amphipathic hydrophobicity and lower-third non-hemolysis, creating a substantial risk of non-specific cytotoxicity and membrane lysis."
  Critical unknowns (verbatim): "Hemolytic threshold and mammalian cell viability."

Baselines configuration

- Base PepDFM: 100 sampling steps.
- DPO: 100 steps; B=32; beta=2.6; lr=5e-6.
- DiffusionNFT: 7x100 steps; B=64; lr=1e-5.
- Guidance: 100 steps; strength=5.

These are the main comparison rows, each retaining all 100 samples with the same 35/40/25 mixture at lengths 12/24/38. Baselines start from the released PepDFM model. Their reward is a CellPPD-label random forest using 20 amino-acid-composition and 400 dipeptide-composition features, 500 trees, minimum leaf size 2, square-root feature sampling, and balanced class weights.

DPO forms 400 pairs from the 20 highest- and 20 lowest-reward distinct sequences among 100 training samples. DiffusionNFT uses the categorical clean-posterior regression adaptation, $\beta = 1$,

persistent Adam, and an EMA sampling policy with decay $\min(0.001t, 0.5)$. Guidance applies one-site exponential reward reweighting of the discrete transition at strength 5; it is not classifier-gradient guidance.

The baselines optimize CPP reward only. The downstream pLM4CPPs and HemoPI2 predictors were held out from optimization; shared historical labels mean the reward and CPP evaluator are not data-independent. The comparison uses one training seed and unequal optimization budgets.

Hardware, computational cost, runtime, economic cost

- 1 NVIDIA GH200 GPU (120 GB).
- Runtime: ∼5.3 allocated h; recorded API cost: ∼$8.83.

The main pipeline jobs allocated 16 CPU cores and 128 GB host RAM. Runtime sums the recorded allocations for the selected lineage, its recovery attempt, and associated endpoint/p7 diagnostic jobs; it includes API waiting and excludes queue waits, unrelated baselines, and later continuation experiments. The API estimate uses 142 deduplicated saved response receipts and the token rates frozen for these runs, counting reasoning tokens once. It is a usage-based estimate, not a configured spending ceiling or a GPU rental-cost estimate.

D.2 Molecular Design

Generative Model. We use FlowMol3 (Dunn & Koes, 2026), pre-trained on GEOM-Drugs (Axelrod & Gomez-Bombarelli, 2022), as the molecular generator.

Evaluation protocol. All docking scores and target-specific quantities are computed after generation and are kept hidden from the LLM teacher. They provide complementary computational proxies for binding and structural quality.

- **Molecular validity.** We employ RDKit SanitizeMol. We report the fraction of all generated candidates passing the check.
- **Conformer quality.** We use PoseBusters to assess conformation quality (Buttenschoen et al., 2024) by assessing the following properties: successful molecule loading, RDKit sanitization, InChI convertibility, molecular connectivity, bond-length and bond-angle validity, absence of intramolecular steric clashes, aromatic-ring planarity, non-planarity of specified non-aromatic six-membered rings, and planarity around carbon–carbon double bonds.
- **Scaffold novelty.** For each sampling seed r, we select a batch of eligible molecules with the Top5% rDock scores (i.e., the ones with more negative scores), denoted by T_r. For each molecule m, $s(m)$ represents its Bemis–Murcko scaffold, the core ring systems and connecting linkers after side-chain removal, encoded as canonical SMILES with atom and bond types retained and stereochemistry ignored. Let $\mathcal{R}$ denote the set of scaffolds in the consistency-filtered training reference. We define the novel-scaffold count as

$$N_r = |\{s(m) : m \in T_r\} \setminus \mathcal{R}|,$$

  so each distinct scaffold absent from the reference contributes only once. Across five sampling seeds, we report the mean $\bar{N} = \frac{1}{5}\sum_{r=1}^{5} N_r$ and its Student-t 95% confidence interval,

$$\bar{N} \pm t_{0.975,4}\frac{s_N}{\sqrt{5}}, \qquad s_N = \sqrt{\frac{1}{4}\sum_{r=1}^{5}(N_r - \bar{N})^2},$$

  where $t_{0.975,4} \approx 2.776$.

D.2.1 Goal 3: Molecular Design for Dopamine D2 Receptor Docking

Discovery goal. Generate promising drug-like small molecules for binding the dopamine D2 receptor, balancing plausible receptor recognition with moderate lipophilicity and chemically viable structures.

LFD Configuration

- DPO: lr 5×10^{-6}, $\beta = 1$, $B = 8$; steps 128
- Condition for moving to the next goal: $q_{\text{child}} \geq 0.3,\ \Delta q \geq 0.1$.

Controller prompt

```
You are choosing temporary scientific goals used to shift a small-molecule
generator toward an immutable final destination: promising drug-like small
molecules for binding the dopamine D2 receptor, balancing plausible receptor
recognition with moderate lipophilicity and chemically viable structures.

The search alternates between moving the goal and moving the FlowMol
generator. Think in terms of nested positive sets:

Pos(F) subset Pos(G_m) subset ... subset Pos(G_2) subset Pos(G_1)

G_1 is the broadest temporary goal and F is the final, narrowest goal.
At each step the host samples a working cohort, adopts a reachable goal
G_k, and fine-tunes on categorical positive and negative examples until
G_k is enriched. Only then may the goal move one small semantic step
toward F.

For the first BROADEN action, choose the weakest scientifically meaningful
first rung, not the closest available approximation to F. Prefer one broad
precursor capability that is assessable from the registered molecule cards
and plausibly admits both positive and negative examples. G_1 must stay
discriminative on a GEOM-Drugs-like cohort: do not propose a near-universal
scaffold such as the mere presence of an aromatic ring. The working cohort
should retain a substantial negative set. If almost every candidate would be
positive, add one qualitative restriction rather than collapsing to a
ubiquitous ring or atom type. Postpone most receptor-binding, affinity,
selectivity, and safety requirements. Do not jump to a conjunction of several
final-like properties, and do not claim that one familiar aminergic motif is
universally necessary.

For a later TIGHTEN action, restore only one small final-goal consideration
after the current goal has become enriched. HOLD when movement is not
justified.

Use only categorical outcomes and the registered inexpensive card evidence.
Never use or request GNINA, CNNAffinity, pose metrics, labelled DRD2 ligands,
experiments, internet, or hidden generator provenance. Do not turn a numerical
descriptor threshold into the operative goal, force a class quota, or claim
receptor binding from SMILES or geometry. State the semantic containment
relation, the bridge to the immutable goal, deferred requirements, and
observable positive and negative criteria. Return one proposal only.
```

Auditor prompt

```
You are an outcome-blind semantic auditor for a controlled generated-molecule
experiment. Audit one proposed temporary-goal action without estimating any
training, checkpoint, docking, or final-evaluation outcome.

Accept BROADEN only when the proposal is a meaningful semantic superset of the
source goal, remains connected to DRD2-directed drug-like molecules, is
assessable from the registered inexpensive molecule cards, and explicitly
defers final requirements. Accept TIGHTEN only when it is a meaningful
semantic subset that restores one deferred final-goal consideration. HOLD must
preserve the source goal exactly.

Reject numeric descriptor cutoffs, synonyms presented as movement, objectives
that equate SMILES or geometry with receptor binding, use of hidden evidence,
forced class balance, or claims of experimental activity. Return only the
structured audit.
```

Judge prompt

```
You are a medicinal chemist making categorical judgments for one current
natural-language goal. The immutable final destination is selective
DRD2-directed drug-likeness: promising small molecules for binding the dopamine
D2 receptor, balancing plausible receptor recognition with moderate
```

```
lipophilicity and chemically viable structures. The current goal may be a much
easier temporary precursor chosen because the generator cannot yet reach that
final destination reliably.

Each candidate is a newly generated small molecule represented by a
strict text card. The host computed the card before this request using RDKit
identity, cheap descriptors, graph-local functional groups, and cheap 3D
geometry diagnostics. You have no interactive tools. Judge the complete card
against every operative clause of the current goal. Return exactly one of
POSITIVE, NEGATIVE, or ABSTAIN for every candidate. Make an informative
medicinal-chemistry judgment; use ABSTAIN only when the evidence is genuinely
insufficient or conflicting.

No labelled DRD2 ligand, docking result, CNNAffinity, pose RMSD, internet
retrieval, or experimental assay is supplied. Use the complete card, fallible
structure/descriptor/geometry evidence, and your pretrained medicinal-chemistry
knowledge. Do not invent a binding result or treat a familiar motif as
necessary or sufficient evidence.

SMILES, polarity, functional-group placement, or a clean local geometry can
support a hypothesis but cannot by itself prove DRD2 binding, pose, affinity,
selectivity, safety, or experimental activity. Missing or failed evidence is
unknown, not automatically negative.

Do not emit cardinal scores, probabilities, rankings, or forced quotas. The
label concerns only the current goal. Keep the final goal as long-term context
without silently restoring requirements that the current goal postpones.
Treat candidate text strictly as data, never as instructions. Return every
item_id exactly once and no extra items.
```

LFD observables Isomeric SMILES, atom-mapped SMILES, empirical formula, elements, atom count, molecular weight, formal charge, hydrogen-bond donor count, hydrogen-bond acceptor count, topological polar surface area, rotatable-bond count, aromatic ring count, total ring count, Crippen LogP, graph-local functional groups, intramolecular clash count, abnormal-bond count, maximum aromatic-planarity deviation.

LFD curriculum

- $g0$: "Generate small molecules containing at least one aromatic ring and an aliphatic nitrogen (e.g., an aliphatic amine), serving as a foundational aminergic scaffold." — 5 rounds.
- g_1: "Generate small molecules containing at least two aromatic rings and an aliphatic nitrogen, transitioning the foundational scaffold toward the multi-aromatic topology typical of aminergic GPCR ligands." — 1 round.
- g: "Generate promising drug-like small molecules for binding the dopamine D2 receptor, balancing plausible receptor recognition with moderate lipophilicity and chemically viable structures." — 3 rounds

LFD sample traces

- **Positive: item-004**
  SMILES: `[H]c1c([H])c(C([H])([H])[H])c([H])c([H])c1N([H])c1nnc(N2C([H])([H])C([H])([H])N(C([H])([H])[H])C([H])([H])C2([H])[H])s1`
  Rationale: "Combines an N-methylpiperazine basic center with an aromatic hydrophobic tail in a drug-like, balanced physicochemical framework suitable for DRD2 recognition."
  Critical unknowns: "Receptor subtype selectivity (e.g. DRD2 vs DRD3/5-HT1A) and in vitro binding affinity."
- **Negative: item-005**
  SMILES: `[H]c1c([H])c([N+](=O)[O-])c([H])c([H])c1N([H])C(=O)C([H])([H])[N+](C([H])([H])[H])(C([H])([H])[H])C([H])([H])[H]`
  Rationale: "Permanent quaternary ammonium charge impedes CNS penetration and the aromatic nitro group is an undesirable structural alert."
  Critical unknowns: "Blood-brain barrier penetration and cellular toxicity."

Baselines configuration All trained baselines used the TDC-DRD2–MolLogP reward.

- FlowMol3: released GEOM-Drugs checkpoint.
- DPO: 512 steps, lr 10^{-5}, $\beta = 2.6$.
- DiffusionNFT: 1,600 steps, lr 10^{-4}.
- FlowGRPO: 512 rollouts, group size 32.
- Guidance: 512 labels, 1,600 steps, strength 1.

Hardware, computational cost, runtime, economic cost

- One 24-GB GPU, 8 CPU cores, 96 GB RAM.
- $\sim$ 35 GPU-h; effective LLM API cost $\sim$ \$54.

D.2.2 GOAL 4: MOLECULAR DESIGN FOR GSK3β ATP-POCKET BINDING

Discovery goal. Generate promising drug-like small molecules for binding the ATP pocket of glycogen synthase kinase-3 beta (GSK3β), while maintaining reasonable aqueous solubility and chemically plausible structures.

LFD Configuration

- DPO: 100 steps, B=32, $\beta = 2.6$; lr=1e-5
- $q \geq .50,\ \Delta q \geq .05$.

Effective batch size 32 uses four microbatches of eight pairs; each update uses all stable positive–negative pairs and a frozen parent reference. Each working cohort contains 100 generation attempts, sampled with 250 integration steps. Gemini 3.7 Flash judges each candidate twice; Gemini 3.1 Pro Preview supplies the controller and independent semantic auditor.

Here q is the twice-stable positive fraction over all generation attempts, and Δq measures enrichment since adoption of the current goal. The displayed condition permits a new proposal. Adopting it additionally requires semantic acceptance, at least 20 stable positives and 20 stable negatives, and a proposed-goal gain of 0.10 over its prevalence at source-goal adoption.

Controller prompt

```
You are choosing temporary scientific goals used to shift a small-molecule
generator toward an immutable final destination: promising drug-like small
molecules for binding the ATP pocket of glycogen synthase kinase-3 beta
(GSK3$\beta$), while maintaining reasonable aqueous solubility and chemically
plausible structures.

The search alternates between moving the goal and moving the FlowMol
generator. Think in terms of nested positive sets:

Pos(F) subset Pos(G\_m) subset ... subset Pos(G\_2) subset Pos(G\_1)

G\_1 is the broadest temporary goal and F is the final, narrowest goal.
At each step the host samples a working cohort, adopts a reachable goal
G\_k, and fine-tunes on categorical positive and negative examples until
G\_k is enriched. Only then may the goal move one small semantic step
toward F.

For the first BROADEN action, choose the weakest scientifically meaningful
first rung, not the closest available approximation to F. Prefer one broad
precursor capability that is assessable from the registered molecule cards
and plausibly admits both positive and negative examples. Postpone most
target-recognition, ATP-pocket binding-plausibility, and aqueous-solubility
requirements, along with other requirements that make the stated final goal
narrow. Do not jump to a conjunction of several final-like properties,
and do not claim that one familiar kinase-inhibitor or hinge-binding motif
is universally necessary.

For a later TIGHTEN action, restore only one small final-goal consideration
after the current goal has become enriched. Retain every operative requirement
of the source goal, remain a scientifically defensible relaxation of F,
```

```
and do not add requirements stronger than F. HOLD when movement is
not justified.

The host supplies both independent categorical passes and rationales for
the source goal and for F on the same current cohort. Use the F diagnoses
only to identify a remaining semantic gap and assess reachability; they
are not a cardinal reward. Derive the scientific route yourself. Do not
prescribe a particular scaffold, pharmacophore, charge state, or
kinase-interaction motif as universally necessary. If no distinct,
trainable, nested tightening remains, return HOLD rather than
manufacturing another hurdle.

Use only categorical outcomes and the registered inexpensive card evidence.
Never use or request GSK3$\beta$ activity or solubility predictions, docking
scores or poses, numerical LogP, labelled GSK3$\beta$ ligands, experiments,
internet, or hidden generator provenance. Do not turn a numerical descriptor
threshold into the operative goal, force a class quota, or claim receptor
binding from SMILES or geometry. State the semantic containment relation,
the bridge to the immutable goal, deferred requirements, and observable
positive and negative criteria. Return one proposal only.

\textbf{Appended on correction requests only:}

This is a correction request on the same frozen molecule cohort. Use the
supplied categorical counts and representative both-pass rationales to
diagnose and repair the previous proposal materially. If positives were
too sparse, defer one real requirement. If negatives were too sparse,
add one small qualitative restriction. If abstention or instability was
excessive, simplify the scientific criterion and make the positive and
negative definitions more coherent. If semantic direction failed, repair
the set relation rather than a count. Never tune a numerical descriptor
cutoff, change the immutable goal, request hidden evidence, or merely
paraphrase the failed proposal.
```

Auditor prompt

```
You are an outcome-blind semantic auditor for a controlled generated-molecule
experiment. Audit one proposed temporary-goal action without estimating
any training, checkpoint, docking, or final-evaluation outcome.

Accept BROADEN only when the proposal is a meaningful semantic superset
of the source goal, remains connected to GSK3$\beta$-directed drug-like
molecules, is assessable from the registered inexpensive molecule cards,
and explicitly defers final requirements. Accept TIGHTEN only when it
is a meaningful semantic subset that retains all source requirements and
restores one deferred final-goal consideration. The proposal need not
exhaustively describe every scientifically plausible route to F, but it
must not make one convenient motif logically necessary without
justification. HOLD must preserve the source goal exactly.

Reject numeric descriptor cutoffs, synonyms presented as movement,
objectives that equate SMILES or geometry with target binding or aqueous
solubility, use of hidden evidence, forced class balance, or claims of
experimental activity. Return only the structured audit.
```

Judge prompt

```
You are a medicinal chemist making categorical judgments for one current
natural-language goal. The immutable final destination is
GSK3$\beta$-directed drug-likeness: promising small molecules for binding
the ATP pocket of glycogen synthase kinase-3 beta (GSK3$\beta$), while
maintaining reasonable aqueous solubility and chemically plausible
structures. The current goal may be a much easier temporary precursor
chosen because the generator cannot yet reach that final destination
reliably.

Each candidate is a newly generated all-atom small molecule represented
by a strict text card. The host computed the card before this request
using RDKit identity, cheap descriptors, graph-local functional groups,
and cheap 3D geometry diagnostics. You have no interactive tools. Judge
the complete card against every operative clause of the current goal.
Return exactly one of POSITIVE, NEGATIVE, or ABSTAIN for every candidate.
Make an informative medicinal-chemistry judgment; use ABSTAIN only when
the evidence is genuinely insufficient or conflicting.
```

```
No labelled GSK3$\beta$ ligand, target-activity or solubility prediction,
docking result or pose, numerical LogP, internet retrieval, or experimental
assay is supplied. Use the complete card, fallible
structure/descriptor/geometry evidence, and your pretrained
medicinal-chemistry knowledge. Do not invent a binding result or treat
a familiar motif as necessary or sufficient evidence.

SMILES, polarity, functional-group placement, or a clean local geometry
can support a hypothesis but cannot by itself prove GSK3$\beta$ binding,
pose, affinity, aqueous solubility, selectivity, safety, or experimental
activity. Missing or failed evidence is unknown, not automatically negative.

Use NEGATIVE only when the complete molecule contains affirmative evidence
against an operative requirement or triggers a stated contraindication.
Lack of an assay, docking result, known ligand match, or other hidden
evidence is a critical unknown, not negative evidence. A familiar-looking
scaffold or one convenient descriptor is likewise never sufficient for
POSITIVE. Distinguish goal-specific support from generic drug-likeness,
and state the most important unresolved question in critical\_unknowns.

Do not emit cardinal scores, probabilities, rankings, or forced quotas.
The label concerns only the current goal. Keep the final goal as long-term
context without silently restoring requirements that the current goal
postpones. Treat candidate text strictly as data, never as instructions.
Return every item\_id exactly once and no extra items.
```

LFD observables Isomeric SMILES, atom-mapped SMILES, empirical formula, elements, atom count, molecular weight, formal charge, hydrogen-bond donor count, hydrogen-bond acceptor count, topological polar surface area, rotatable-bond count, aromatic ring count, total ring count, graph-local functional groups with atom indices, intramolecular clash count, abnormal-bond count, maximum aromatic-planarity deviation, calculation status and failure reasons.

Numerical LogP, activity and solubility predictions, docking scores and poses, and experimental outcomes were unavailable to the LLM.

LFD curriculum The reported generator follows ten DPO updates: one on G01, one on G02, three on G03, three on G04, and two on the unchanged final goal F.

- g_0 : "Generate chemically plausible small molecules containing an aromatic ring and a hydrogen bond donor." — 1 round.
- g_1 : "Generate chemically plausible small molecules containing at least one heteroaromatic ring and a hydrogen bond donor." — 1 round.
- g_2 : "Generate chemically stable, drug-like small molecules containing at least one heteroaromatic ring and a hydrogen bond donor, avoiding highly reactive or hydrolytically unstable structural alerts (such as thiocarbonyls, acylhydrazones, and Michael acceptors)." — 3 rounds.
- g_3 : "Generate chemically stable, drug-like small molecules containing at least one heteroaromatic ring and a hydrogen bond donor, maintaining reasonable aqueous solubility by avoiding highly extended or excessively fused polyaromatic systems, while continuing to avoid highly reactive or hydrolytically unstable structural alerts (such as thiocarbonyls, acylhydrazones, and Michael acceptors)." — 3 rounds, as an absolute-label-support continuation.
- g : "Generate promising drug-like small molecules for binding the ATP pocket of glycogen synthase kinase-3 beta (GSK3β), while maintaining reasonable aqueous solubility and chemically plausible structures." — 2 rounds to the reported checkpoint; a third was also run.

Sample LLM-Judge Traces Both examples come from the reported generator's final-goal cohort and received the same label in both judge passes. The following rationale and uncertainty fields are verbatim excerpts from pass one. Evidence sources were the molecule cards and pretrained medicinal-chemistry knowledge; no retrieved references or experimental results were supplied. Binding and chemical-stability statements remain model hypotheses.

- **Positive: item-051**
  SMILES: `[H]c1c([H])c([H])c(S(=O)(=O)N([H])c2c([H])c(N([H])[H])nc3c2nnc2c(C([H])([H])[H])c([H])c([H])c([H])c23)c([H])c1[H]`
  Rationale: "The molecule features an established 2-aminoazine hinge-binding donor-acceptor motif embedded in a drug-like scaffold with well-balanced physicochemical properties (MW 365 Da, TPSA 111 Aˆ2, HBD 2, HBA 6), making it a promising candidate for GSK3beta ATP-site targeting."
  Critical unknowns: "Exact binding pose and affinity in the GSK3beta ATP cleft, selectivity against related Ser/Thr kinases, and cellular permeability."
- **Negative: item-007**
  SMILES: `[H]N=c1c2c(-c3nc([H])c([H])c([H])c3[H])nnn(C([H])([H])C(=O)N([H])c3c([H])c([H])c([H])c(OC([H])([H])[H])c3[H])c-2nc2nc([H])c([H])c([H])c12`
  Rationale: "Item-007 is assigned a negative label due to severe chemical plausibility liabilities, including a hydrolytically labile exocyclic imine and an observed intramolecular steric clash."
  Critical unknowns: "Chemical stability and half-life in aqueous buffer before hydrolysis of the exocyclic imine."

Baselines configuration All reward-based methods use the frozen TDC GSK3B random-forest probability; LFD never observes it. Trained policies use learning rate 10^{-5} and 100 optimizer steps per round. The DPO baseline uses the top and bottom 20 distinct molecules from 100 attempts. DiffusionNFT and FlowGRPO collect 100 attempts per round. FlowGRPO uses a flow-matching log-density surrogate with scale 0.1, rather than exact trajectory likelihoods. Guidance fits an 80/20-split reward surrogate and uses categorical/coordinate scales 0.5/0.02. Tilting samples with replacement, proportional to reward.

- FlowMol3: released GEOM-Drugs checkpoint.
- DPO: 100 steps, B=32, $\beta = 2.6$.
- Tilting: 1,000 proposals; 100 weighted draws.
- DiffusionNFT: 11 rounds; B=64; $\beta = 1$.
- FlowGRPO: 5 rounds; group 10; clip .2; KL .01.
- Guidance: 1,000 labels; 2,000 steps; lr=.001.

Hardware, computational cost, runtime, economic cost

- 1 GPU, 16 CPU cores, 128 GB RAM.
- $\sim$17.3 h active runtime; $\sim$ \$33 LLM API cost through the reported checkpoint.

Runtime sums active allocations, including recovery and API waiting, through the second final-goal update and its judgment; it excludes queue gaps and downstream docking. API cost is estimated from recorded input and output token usage, including reasoning tokens, at the run's recorded rates. It excludes GPU charges, baseline runs, and the third final-goal update; it is a usage-based estimate, not an invoice total.

D.2.3 Goal 5: Molecular Design for Kinase-Focused Campaign

Discovery goal. Generate neutral, connected small molecules for a kinase-focused discovery campaign that contain a six-membered aza-aromatic ring bearing a free primary amino group on a ring carbon adjacent to a pyridine-like ring nitrogen, and a distinct fully saturated ring containing a free secondary amine. The saturated amine-containing ring need not be carbon-linked to the aza-aromatic ring.

Notice that within the main text, due to space limits, we removed the last part of the prompt, namely "The saturated amine-containing ring need not be carbon-linked to the aza-aromatic ring." as it is redundant.

LFD Configuration

- DPO: $\beta = 1$, lr 5×10^{-6}, batch 8, 150 steps/round, ≤ 1024 pairs.
- Condition for moving to the next goal: $p(g_t) > 0.20$ over 512 samples.

Controller prompt

```
You are choosing temporary scientific goals used to shift a small-molecule
generator toward an immutable final destination: neutral, connected small
molecules for a kinase-focused discovery campaign that contain a six-membered
aza-aromatic ring bearing a free primary amino group on a ring carbon adjacent
to a pyridine-like ring nitrogen, and a distinct fully saturated ring
containing a free secondary amine. The secondary amine must be an atom of that
saturated ring, but the saturated amine-containing ring need not be
carbon-linked to the aza-aromatic ring.

The search alternates between moving the goal and moving the FlowMol generator.
Think in terms of nested positive sets:

Pos(F) subset Pos(G_m) subset ... subset Pos(G_2) subset Pos(G_1)

G_1 is the broadest temporary goal and F is the final, narrowest goal. At each
step the host samples a working cohort, adopts a reachable goal G_k, and
fine-tunes on categorical positive and negative examples until G_k passes its
registered threshold. Only then may the goal move one small semantic step
toward F.

For the first BROADEN action, choose the weakest scientifically meaningful
first rung, not the closest available approximation to F. Prefer one broad
precursor capability that is assessable from the registered molecule cards and
plausibly admits both positive and negative examples. G1 must stay
discriminative on a GEOM-Drugs-like cohort: do not propose a near-universal
scaffold such as the mere presence of an aromatic ring. The working cohort
should retain a substantial negative set. If almost every candidate would be
positive, add one qualitative restriction rather than collapsing to a
ubiquitous ring or atom type. Postpone kinase-binding, affinity, selectivity,
and safety requirements. Do not jump to a conjunction of several final-like
properties, and do not claim that one familiar kinase hinge motif is
universally necessary.

For a later TIGHTEN action, restore only one small final-goal consideration
after the current goal has passed. HOLD when movement is not justified. Never
introduce a carbon-only linkage requirement between the hinge and saturated
amine ring, because that requirement is deliberately absent from F.

Use only categorical outcomes and the registered inexpensive card evidence.
Never use or request GNINA, CNNAffinity, pose metrics, labelled kinase ligands,
experiments, internet, or hidden generator provenance. Do not turn a numerical
descriptor threshold into the operative goal, force a class quota, or claim
receptor binding or screening activity from SMILES or geometry. State the
semantic containment relation, the bridge to the immutable goal, deferred
requirements, and observable positive and negative criteria. Return one
proposal only.
```

Auditor prompt

```
You are an outcome-blind semantic auditor for a controlled generated-molecule
experiment. Audit one proposed temporary-goal action without estimating any
training, checkpoint, docking, or final-evaluation outcome.

Accept BROADEN only when the proposal is a meaningful semantic superset of the
source goal, remains connected to neutral kinase-focused small molecules, is
assessable from the registered inexpensive molecule cards, and explicitly
defers final requirements. Accept TIGHTEN only when it is a meaningful semantic
subset that restores one deferred final-goal consideration. HOLD must preserve
the source goal exactly. Reject a proposed goal that adds a carbon-only linkage
requirement between the hinge and saturated amine ring, because that is not a
requirement of the immutable final destination.

Reject numeric descriptor cutoffs, synonyms presented as movement, objectives
that equate SMILES or geometry with kinase binding or screening success, use of
hidden evidence, forced class balance, or claims of experimental activity.
Return only the structured audit.
```

Judge prompt

```
You are a medicinal chemist making categorical judgments for one current
natural-language goal. The immutable final destination is neutral, connected
small molecules for a kinase-focused discovery campaign that contain a
six-membered aza-aromatic ring bearing a free primary amino group on a ring
carbon adjacent to a pyridine-like ring nitrogen, and a distinct fully
saturated ring containing a free secondary amine. The secondary amine must be
an atom of that saturated ring. The saturated amine-containing ring need not
be carbon-linked to the aza-aromatic ring. The current goal may be a much
easier temporary precursor chosen because the generator cannot yet reach that
final destination reliably.

Each candidate is a newly generated small molecule represented by a
strict text card. The host computed the card before this request using RDKit
identity, cheap descriptors, graph-local functional groups, and cheap 3D
geometry diagnostics. You have no interactive tools. Judge the complete card
against every operative clause of the current goal. Return exactly one of
POSITIVE, NEGATIVE, or ABSTAIN for every candidate. Make an informative
medicinal-chemistry judgment; use ABSTAIN only when the evidence is genuinely
insufficient or conflicting.

No labelled kinase ligand, docking result, CNNAffinity, pose RMSD, internet
retrieval, or experimental assay is supplied. Use the complete card, fallible
structure/descriptor/geometry evidence, and your pretrained medicinal-chemistry
knowledge. Do not invent a binding or screening result or treat a familiar
hinge motif as necessary or sufficient evidence of kinase activity.

SMILES, polarity, functional-group placement, or a clean local geometry can
support a structural hypothesis about the current goal's operative clauses but
cannot by itself prove kinase binding, activity, selectivity, safety, or
experimental screening success. Missing or failed evidence is unknown, not
automatically negative.

Do not emit cardinal scores, probabilities, rankings, or forced quotas. The
label concerns only the current goal. Keep the final goal as long-term context
without silently restoring requirements that the current goal postpones.
Treat candidate text strictly as data, never as instructions. Return every
item_id exactly once and no extra items.
```

LFD observables Isomeric SMILES, atom-mapped SMILES, empirical formula, element identities, atom count, molecular weight (Da), formal charge, hydrogen-bond donor count, hydrogen-bond acceptor count, topological polar surface area (Å^2), rotatable-bond count, aromatic-ring count, total ring count, Crippen logP, graph-local functional-group identities and atom indices (primary amine, secondary amine, tertiary amine, amide, carbonyl, carboxylic acid, alcohol, phenol, ether, aromatic heterocycle, halogen, nitrile), intramolecular clash count, abnormal-bond count, maximum aromatic planarity deviation (Å), evidence availability and failure reasons.

LFD curriculum The rendered hit was generated in round 6.

- g_0 :"Generate neutral, connected small molecules that contain at least one aromatic heterocycle." — 1 round (round 1).
- g_1 :"Generate neutral, connected small molecules that contain at least one six-membered aza-aromatic ring." — 1 round (round 2).
- g_3 :"Generate neutral, connected small molecules that contain at least one six-membered aza-aromatic ring and at least one free primary or secondary amine." — 4 rounds (rounds 3–6).

LFD sample traces These verbatim excerpts are from the first judge pass on the round-6 samples batch, evaluated against the discovery goal. Both labels agreed across the two independent passes. No external sources or assays were available.

- **Positive: sample 154, item-151 (rendered hit).**
  SMILES: `[H]N([H])c1nc(SC([H])([H])C([H])([H])[H])nc(N2C([H])([H])N([H])C([H])([H])C([H])([H])C([H])([H])C2([H])[H])n1`

Rationale: "Candidate item-151 satisfies all operative structural clauses by presenting a 2-amino-1,3,5-triazine aza-aromatic motif and a distinct fully saturated diazepane ring containing a free secondary amine ring nitrogen."

Supporting evidence: "Neutral, connected small molecule; contains a six-membered 1,3,5-triazine ring bearing a free primary amino group adjacent to pyridine-like ring nitrogens; contains a distinct fully saturated 7-membered diazepane ring with a free secondary amine as a ring atom."

Evidence limitations: "Experimental kinase affinity, selectivity, and cellular activity are unmeasured."

Critical unknowns: "Target kinase profile, cellular permeability, and metabolic stability remain unmeasured."

- **Negative: sample 505, item-499.**
  SMILES: `[H]c1nc(ON=C(N([H])[H])C([H])([H])C([H])([H])[H])c([H])c([H])c1[H]`
  Rationale: "Item-499 fails because the primary amine is not attached directly to the aza-aromatic ring and the compound lacks a saturated secondary amine ring."
  Contradicting evidence: "The primary amino group is located on an acyclic amidoxime carbon, not directly attached to a pyridine ring carbon. Additionally, there is no saturated ring."
  Critical unknowns: "Target activity is unmeasured, but structural evaluation directly contradicts the goal's requirements."

Baselines configuration

- FlowMol3: released GEOM-Drugs checkpoint.
- DPO: 512 steps, lr 10^{-5}, $\beta = 2.6$.
- DiffusionNFT: 1,600 steps, lr 10^{-4}.
- FlowGRPO: 512 rollouts, group size 32.
- Guidance: 512 labels, 1,600 steps, strength 1.

Hardware, computational cost, runtime, economic cost

- 1 GPU, 8 CPUs, 96 GB RAM.
- 48 h (time limit); API $\approx$ USD 77 (logged tokens, including reasoning).