PHAME: PHENOTYPE-AWARE MOLECULAR EDITING VIA LATENT DIFFUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Small-molecule drug discovery requires simultaneous optimization of numerous properties of candidate molecules. These properties can be investigated through the analysis of high-dimensional biological signatures, such as cell morphology and transcriptomic perturbations, which provide a rich perspective on the underlying biological mechanisms. However, existing generative methods, which use those signatures for optimization, fail to meet two key requirements: providing precise guidance toward desired phenotypic signatures while maintaining structural proximity to a known hit. We introduce PhAME (Phenotype-Aware Molecular Editing), a latent diffusion framework that overcomes this challenge by recasting molecular optimization as editing in the latent space of a pretrained graph-based VAE. Our central contribution is a compositional classifier-free guidance scheme with two independent scales, one for the target signature (phenotype or molecular property) and one for similarity to the seed structure, allowing practitioners to control the tradeoff between these two objectives. Empirical evaluations across diverse benchmarks, including docking score optimization and multimodal phenotypic generation, demonstrate that PhAME achieves state-of-the-art results while maintaining high chemical validity and novelty.

1 INTRODUCTION

Drug discovery is a search over a chemical space estimated at 10^{60} drug-like molecules (Polishchuk et al., 2013), but in practice it rarely starts from scratch. The dominant bottleneck is the iterative refinement of an already-active compound through *Hit-to-Lead* (H2L) and *Lead Optimization* (LO), reshaping a screening hit into a potent, drug-like, synthesizable candidate (Niazi, 2023). Compounding this, the field is shifting from single-target binding assays toward *phenotypic* readouts, such as gene-expression changes and cell-morphology profiles, that capture how a molecule perturbs a biological system as a whole (Chandrasekaran et al., 2021; Swinney, 2013). A generative model that can both edit a known hit and steer it toward such system-level signals would be a powerful tool for ML-driven discovery, yet no method today delivers both.

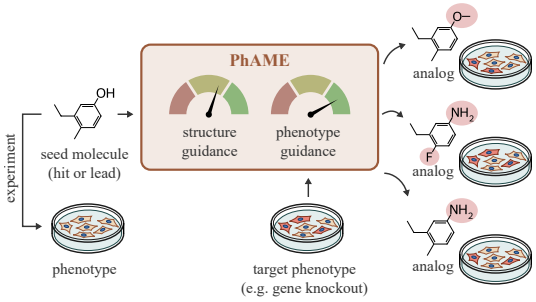


Figure 1: **Molecular editing with PhAME.** Given a seed molecule and target phenotype, the model generates close structural analogs that induce a target phenotype by balancing two independent guidance scales: structure and phenotype.

Ligand-centric generators (Gómez-Bombarelli et al., 2018; Jin et al., 2018; Jing et al., 2022) produce valid molecules but ignore cell-level response and offer no mechanism to anchor on a seed compound. Phenotype-conditioned generators close the first gap but not the second: GxVAEs (Li & Yamanishi, 2024b) and SmilesGEN (Liu et al., 2025c) both generate *de novo* from a phenotypic target and cannot be steered around a known hit. A concurrent work (Song & Liu, 2025) adds molecule editing via discrete mask-and-reconstruct on SELFIES substructures, but only under L1000 transcriptomics. Even classifier-free guidance (CFG) (Ho & Salimans, 2022), the standard steering tool for diffusion models, does not solve the problem. Its single guidance scale forces a trade-off between matching the target phenotype and staying close to the seed molecule.

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.

We propose **PhAME** (**Ph**enotype-**A**ware **M**olecular **E**diting) to close both gaps: conditioning on a desired phenotype (or, more generally, a target signature) and anchoring on a seed compound. PhAME reformulates design as an editing task in the latent space of a frozen graph-based VAE (Nguyen & Karolak, 2025). A seed molecule is partially noised and denoised under guidance, so the noise level controls the structural exploration radius. Structural proximity is measured in a learned embedding space of InfoAlign (Liu et al., 2025a) that reflects cellular response rather than raw chemical similarity. Crucially, we replace standard CFG with a *compositional* variant that splits the guidance signal into two terms with *independent* scales, one for the target signature (phenotype or molecular property) and one for structural anchoring, directly removing the trade-off baked into single-scale CFG.

Across docking-score optimization, transcriptomic-conditioned generation, and phenotype-driven mode of action (MoA) design, PhAME delivers the strongest results to date: the best novel hit ratio on four out of five MOOD (Lee et al., 2023) protein targets, and consistent superiority on phenotype-driven MoA retrieval, all while maintaining near-perfect validity and novelty. Our contributions can be summarized as follows:

- **Latent-diffusion molecular editing:** We introduce PhAME, which utilizes latent diffusion on graph-based embeddings to enable Hit-to-Lead guided by complex biological signatures.
- **Compositional guidance mechanism:** We generalize classifier-free guidance to use two independent scales, one for phenotype matching and one for seed proximity, replacing the forced trade-off with explicit control over each objective.
- **SOTA across modalities:** PhAME sets new state-of-the-art on docking-score optimization, transcriptomic-conditioned generation, and Cell-Painting-driven MoA generation, with near-perfect validity and novelty.

2 Related Work

Generative models for molecules. Molecular generation has progressed from SMILES-based VAEs (Gómez-Bombarelli et al., 2018) to graph-based VAEs (Jin et al., 2018), and more recently to latent diffusion models Rombach et al. (2022) that compress denoising into a learned manifold. Mol-CycleGAN Maziarka et al. (2020) already showed that the latent space of a graph VAE supports direct molecular optimization. Equivariant and graph variants of latent diffusion in chemistry include EDM Hoogeboom et al. (2022) and GCLDM Zhang et al. (2025). For property-driven optimization, score-based MOOD Lee et al. (2023), the flow-matching frameworks PropMolFlow Zeng et al. (2026) and MoltenFlow Lobo et al. (2026), and fragment-level discrete diffusion with GenMol (Lee et al., 2025) offer stable alternatives to reinforcement learning Loeffler et al. (2024). Conditioning is typically realized via classifier-free guidance (CFG) (Ho & Salimans, 2022; Weiss et al., 2023). Recent factorized variants in scPPDM Liang et al. (2025) and the multi-attribute CFGen Palma et al. (2024) expose multiple interpretable controls, but were developed for single-cell rather than molecular targets. PhAME instantiates a latent-diffusion-on-graph-VAE recipe on top of the DGVAE (Nguyen & Karolak, 2025) backbone and replaces single-scale CFG with a compositional scheme that scales the target signature and structural anchor independently.

Phenotype-conditioned molecular generation. A complementary line conditions generation on system-level biological readouts. Most methods consume L1000 transcriptomic signatures: the GAN of Méndez-Lucio et al. Méndez-Lucio et al. (2020), the VAEs Triomphe Kaitoh & Yamanishi (2021), GxVAE (Li & Yamanishi, 2024b) and SmilesGEN (Liu et al., 2025c), the graph latent diffusion GLDM Wang et al. (2024), and the chemistry LLM GEMGen Jiang et al. (2026). A smaller group conditions on Cell Painting morphology: CPMolGAN (Zapata et al., 2023), MGMG (Tang et al., 2025), and the GFlowNet of Lu et al. Lu et al. (2024). However, most of these methods generate molecules *de novo*. PhAME differs by editing in a graph-VAE latent, anchoring proximity through InfoAlign Liu et al. (2025a) embeddings aligned with both Cell Painting and L1000 via an information bottleneck, and exposing independent target signature and structural CFG scales.

3 Background

Latent diffusion models. Diffusion models (Ho et al., 2020; Song et al., 2021b) learn to generate data by reversing a gradual noising process. Operating this process directly on high-dimensional or

discrete data, such as SMILES strings or molecular graphs, is computationally expensive and requires careful handling of discrete structures. Latent diffusion models (LDMs) (Rombach et al., 2022) resolve this by first training an autoencoder to map data into a compact continuous latent space, and then applying the diffusion process entirely within that space. In our setting, a pretrained SMILES VAE with encoder $\mathcal{E}$ and decoder $\mathcal{D}$ provides the latent space: $z_0 = \mathcal{E}(m)$ for a molecule $m \in \mathcal{M}$.

The forward process progressively corrupts z_0 into Gaussian noise via $q(z_t \mid z_0) = \mathcal{N}(z_t;\, \sqrt{\bar{\alpha}_t}\, z_0,\, (1-\bar{\alpha}_t)\, I)$, where $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$ follows a predefined variance schedule and $z_T \approx \mathcal{N}(0, I)$ for sufficiently large T. A neural network $\epsilon_\theta(z_t, t)$ is trained to reverse this process by predicting the added noise, minimizing

$$\mathcal{L} = \mathbb{E}_{z_0,\epsilon,t}\left[\|\epsilon - \epsilon_\theta(z_t, t)\|^2\right], \qquad \epsilon \sim \mathcal{N}(0, I),\ t \sim \mathcal{U}\{1, \ldots, T\}. \tag{1}$$

At inference, new molecules are generated by sampling $\hat{z}_T \sim \mathcal{N}(0, I)$ and iteratively denoising:

$$\hat{z}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\hat{z}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\, \epsilon_\theta(\hat{z}_t, t)\right) + \sigma_t\, \eta, \qquad \eta \sim \mathcal{N}(0, I), \tag{2}$$

where σ_t controls sampling stochasticity ($\sigma_t = 0$ recovers DDIM deterministic sampling (Song et al., 2021a)). The final molecule is obtained by decoding: $m' = \mathcal{D}(\hat{z}_0)$.

Classifier-free guidance. To enable conditional generation without a separate classifier, classifier-free guidance (CFG) (Ho & Salimans, 2022) trains a single network $\epsilon_\theta(\hat{z}_t, t, c)$ on both conditional and unconditional objectives. During training, the conditioning signal c is randomly replaced with a null token $\varnothing$ with probability p_{uncond}, so that the network implicitly learns both the conditional and unconditional score functions. At inference, the noise prediction is interpolated as

$$\tilde{\epsilon}_\theta(\hat{z}_t, t, c) = \epsilon_\theta(\hat{z}_t, t, \varnothing) + w\left[\epsilon_\theta(\hat{z}_t, t, c) - \epsilon_\theta(\hat{z}_t, t, \varnothing)\right], \tag{3}$$

where $w \geq 0$ is the guidance scale. Setting $w = 0$ recovers unconditional sampling; $w = 1$ corresponds to standard conditional sampling; and $w > 1$ amplifies the influence of the conditioning signal at the expense of sample diversity. This mechanism is agnostic to the dimensionality and semantics of c, making it naturally suited for conditioning on signals ranging from scalar properties to high-dimensional phenotypic profiles.

4 METHODS

4.1 PROBLEM DEFINITION

Let $\mathcal{M}$ denote the space of valid molecules represented as SMILES strings. We write $m \in \mathcal{M}$ for a seed (lead) molecule and define a *target signature* $c \in \mathbb{R}^K$, where $K \geq 1$. Depending on the task, c may represent a scalar property target, a multi-property vector, a transcriptomic profile, or a cell morphology signature. Let $\phi : \mathcal{M} \to \mathbb{R}^K$ denote a (generally unknown) oracle mapping each molecule to its ground-truth signature, and let $\mathrm{sim} : \mathcal{M} \times \mathcal{M} \to [0, 1]$ be a structural similarity measure (e.g. Tanimoto similarity over Morgan fingerprints).

Objective. Given a seed molecule m and a target signature c, we seek an optimized molecule m^* that minimizes divergence from c while remaining structurally related to the seed:

$$m^* = \arg\min_{m' \in \mathcal{M}}\ d(\phi(m'),\, c) + \lambda(1 - \mathrm{sim}(m, m')), \tag{4}$$

where $m' \in \mathcal{M}$ denotes a candidate molecule, d is a divergence in the signature space, and $\lambda \geq 0$ governs the trade-off between property improvement and structural conservation. Setting $\lambda = 0$ recovers unconstrained de novo generation.

In practice, we do not optimize this objective directly. The trade-off is controlled implicitly through the *noise level* t in a partial diffusion process, which governs the structural exploration radius around the seed, and the *guidance scale* w in classifier-free guidance (Section 3), which governs the strength of conditioning toward c.

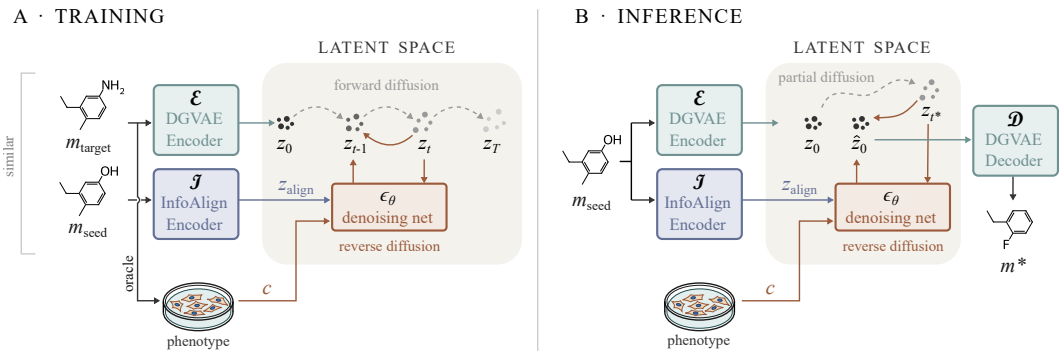


Figure 2: **Overview of PhAME.** *(Training, left)* A frozen DGVAE encoder $\mathcal{E}$ maps a target molecule to a latent z_0, which is noised along a forward diffusion, and a structural anchor (InfoAlign encoder) z_{align} of a seed molecule along with a target phenotypic signature c is used to train a denoising network. *(Inference, right)* A molecule is encoded into the latent space and partially noised; then, the denoising network conditioned on structural anchor z_{align} of the same molecule and phenotypic signature c is used to generate a molecular analog, which is decoded by the DGVAE decoder $\mathcal{D}$.

4.2 PhAME

PhAME (overview in Figure 2) operates in the latent space of DGVAE (Nguyen & Karolak, 2025), a pretrained VAE that encodes molecular graphs and decodes SMILES strings. The graph encoder yields latent codes $z_0 = \mathcal{E}(m)$ that are invariant to SMILES string ordering, while the SMILES decoder $\mathcal{D}$ maps back to valid molecules. The encoder and decoder are frozen; we only train the diffusion model on top of this latent space, following the formulation in Section 3.

The denoising network ϵ_θ is conditioned on two signals besides timestep t: a target signature $c \in \mathbb{R}^K$ and a molecular embedding z_{align} that anchors generation to the seed molecule's neighborhood. We obtain z_{align} from InfoAlign (Liu et al., 2025a), which contains an information bottleneck that aligns molecular structure with Cell Painting morphological profiles (Bray et al., 2016) and L1000 gene expression profiles (Subramanian et al., 2017). Since InfoAlign retains phenotype-predictive structural information, z_{align} encodes phenotypically relevant molecular similarity rather than raw fingerprint distance. The embedding is computed once per seed and held fixed during denoising.

Training. The model is trained on molecular pairs constructed as described in Section 4.3: for each pair, $z_0 = \mathcal{E}(m_{\text{target}})$ is the DGVAE latent of the target molecule with signature c, and z_{align} is the InfoAlign embedding of the corresponding seed molecule. We use a cosine noise schedule (Nichol & Dhariwal, 2021) and train with the denoising objective from Section 3, extended to the dual-conditioned network: $\mathcal{L}_{\text{diff}} = \mathbb{E}_{z_0,\epsilon,t}\left[\|\epsilon - \epsilon_\theta(z_t, t, c, z_{\text{align}})\|^2\right]$. Since $\mathcal{L}_{\text{diff}}$ alone does not constrain where in latent space the denoised sample $\hat{z}_0$ lands, we add an alignment loss

$$\mathcal{L}_{\text{align}} = \max(0,\ \tau - \cos(\hat{z}_0,\ \psi(z_{\text{align}}))) \,, \tag{5}$$

with margin τ=0.8, penalizing predictions that drift too far from the seed. Function ψ is a two-layer MLP that projects the compound representation into the VAE latent space. The full objective is $\mathcal{L} = \mathcal{L}_{\text{diff}} + \gamma\,\mathcal{L}_{\text{align}}$. During training, c and z_{align} are each independently replaced with a null token $\varnothing$ with probability $p_{\text{uncond}} = 0.1$, so the network sees all four conditioning combinations.

Compositional guidance. Standard CFG (Section 3) uses a single scale w for all conditioning, conflating signature guidance with structural anchoring. We decompose guidance into two terms with separate scales w_c and w_a. Let $\Delta_c = \epsilon_\theta(\hat{z}_t, t, c, \varnothing) - \epsilon_\theta(\hat{z}_t, t, \varnothing, \varnothing)$ and $\Delta_a = \epsilon_\theta(\hat{z}_t, t, \varnothing, z_{\text{align}}) - \epsilon_\theta(\hat{z}_t, t, \varnothing, \varnothing)$ denote the signature and alignment directions. The guided prediction is

$$\tilde{\epsilon}_\theta(\hat{z}_t, t, c, z_{\text{align}}) = \epsilon_\theta(\hat{z}_t, t, \varnothing, \varnothing) + w_c\,\Delta_c + w_a\,\Delta_a. \tag{6}$$

Here w_c controls how strongly generation is pushed toward the target signature, while w_a controls how tightly it stays near the seed structure.

Inference. PhAME can run in two modes. In *molecular optimization*, we encode the seed molecule $z_0 = \mathcal{E}(m)$, run $t^* < T$ forward noising steps, and denoise with Eq. 6; the noise level t^* controls

the exploration radius around the seed, while w_c and w_a independently steer signature and structural guidance. In *de novo generation*, we sample $\hat{z}_T \sim \mathcal{N}(0, I)$ and denoise from pure noise with $\gamma = 0$ and $w_a = 0$, reducing the model to a signature-conditioned generator with no structural anchor.

4.3 Dataset

For each task, we partition the dataset according to a predefined threshold (e.g., solubility or docking score), yielding two disjoint subsets corresponding to molecules with property values above and below the threshold. To construct the training data, molecules from the two subsets are paired by selecting the most structurally similar counterparts across the threshold. For more details see Appendix B.

5 Results

We evaluate the proposed approach across a range of benchmark and real-world molecular design tasks. In Section 5.1, we examine the contribution of individual model components within a controlled logP optimization setting, and benchmark our method against state-of-the-art approaches for molecular editing. In Section 5.2, we evaluate the model's ability to generate novel compounds with high predicted binding affinity while preserving drug-likeness and synthetic accessibility. Finally, in Sections 5.3 and 5.4, we demonstrate the model's capability for multimodal conditioning, enabling molecular generation guided by gene expression profiles and phenotypic signals.

5.1 Controlled Property Editing on logP

We first evaluate PhAME on a controlled logP optimization benchmark against state-of-the-art molecular editing methods. Moreover, we show that independently tuning signature guidance (w_c) and structural guidance (w_a) provides explicit control over the trade-off between property optimization and structural preservation.

Setup. We evaluate on the MOSES benchmark Polykovskiy et al. (2020) using the standard train/test split. Following Section 4.3, we partition molecules according to the median logP of the training set to define the optimization direction. The task is bidirectional: high-logP molecules are optimized downward and low-logP molecules upward, providing a controlled setting for assessing property-guided generation.

We compare PhAME against the oracle-free molecular editing and generation methods Mol-CycleGAN Maziarka et al. (2020), DiGress Vignac et al. (2023), and PURE Gupta et al. (2025), as well as three diagnostic baselines: random target-class retrieval, nearest target-class retrieval, and VAE reconstruction. Definitions and implementation details are provided in Appendix C.

Table 1: **LogP optimization with similarity preservation to the seed molecule.** While competing methods typically optimize either target success (*Tgt*) or structural similarity (*Sim*), PhAME achieves the strongest overall performance on the metric *NTS*. We report mean$_{\pm\text{std}}$ over 5 runs.

Method	Nov ↑	Tgt ↑	Sim ↑	NTS ↑
Random target class	0.00 ±0.00	1.00 ±0.00	0.14 ±0.00	0.00 ±0.00
Nearest target class	0.00 ±0.00	1.00 ±0.00	0.60 ±0.00	0.00 ±0.00
VAE reconstruction	0.99 ±0.00	0.19 ±0.00	0.59 ±0.00	0.11 ±0.00
Mol-CycleGAN	0.99 ±0.00	0.51 ±0.04	0.15 ±0.00	0.08 ±0.01
DiGress	1.00 ±0.00	1.00 ±0.00	0.13 ±0.00	0.13 ±0.00
PURE	1.00 ±0.00	0.45 ±0.00	0.58 ±0.00	0.26 ±0.00
PhAME w/o $\mathcal{L}_{\text{align}}$	0.99 ±0.00	0.81 ±0.01	0.35 ±0.00	0.28 ±0.00
PhAME (ours)	0.99 ±0.00	0.78 ±0.03	0.39 ±0.02	**0.30 ±0.01**

We evaluate generated molecules using four metrics. *Nov* (novelty) is the fraction of molecules not present in the training set. *Tgt* (target success) measures the fraction of molecules that reach the desired property region. *Sim* denotes the Tanimoto similarity to the seed molecule. *NTS* is defined as the product of *Nov*, *Tgt*, and *Sim*, capturing the ability to generate novel molecules that satisfy the target objective while remaining close to the seed structure. Additional quality metrics of generated molecules are reported in Appendix C.

Results. Table 1 highlights a clear trade-off between structural preservation and property optimization among the baselines. Mol-CycleGAN struggles to achieve both high target success and structural similarity. DiGress attains perfect target success but sacrifices structural alignment, whereas PURE exhibits the opposite behavior, preserving similarity while underperforming on the optimization

objective. In contrast, PhAME ($w_c = 6, w_a = 3$) performs strongly across all metrics (*Nov*, *Tgt*, and *Sim*) and achieves the best overall *NTS*, demonstrating a superior balance between novelty, target satisfaction, and structural preservation. Incorporating the alignment loss $\mathcal{L}_{\text{align}}$ further improves structural preservation relative to the variant without $\mathcal{L}_{\text{align}}$, increasing similarity while maintaining competitive target success and resulting in a higher *NTS* score.

Notably, PhAME is highly steerable, as illustrated in Figure 3 and detailed in Table 7 in Appendix C. This flexibility allows the model to match or approach the individual strengths of specialized methods, such as *Tgt* of DiGress or *Sim* of PURE. The compositional-guidance mechanism enables independent control over the optimization trade-off. Increasing property guidance (w_c) consistently improves target success, with diminishing returns at larger values and minimal effect on similarity. In contrast, increasing alignment guidance (w_a) improves similarity but weakens property optimization. These gains also saturate at larger w_a, while validity decreases.

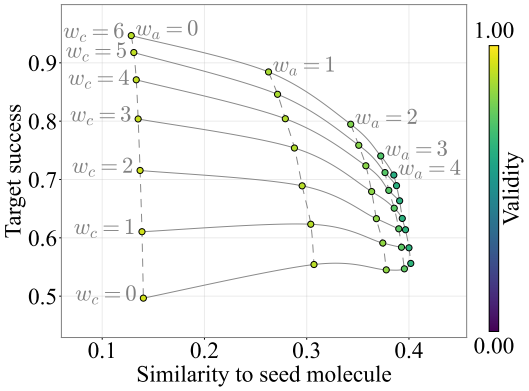


Figure 3: **Influence of guidance scales**. Trade-off between target success and similarity under varying guidance strengths. Increasing property guidance (w_c) improves target success, while increasing alignment guidance (w_a) enhances similarity to the seed molecule.

Overall, the results show that jointly leveraging both guidance signals is essential, as neither property nor structural guidance alone is sufficient. Their combination enables effective control over the trade-off between target optimization and similarity to the seed molecule.

5.2 Docking Score Optimization

We evaluate PhAME's ability to generate *novel*, high-affinity binders for established protein targets in both de novo generation and molecular editing settings. In the de novo setting, we sample latent representations from the VAE prior. In the editing setting, generation is initialized from promising molecules, allowing to assess whether PhAME can improve their predicted binding affinity. To push the model toward increasingly potent regions of chemical space, training follows a three-stage curriculum: each stage fine-tunes the denoising network on ZINC250k pairs split at a stricter binder/non-binder docking-score threshold than the previous, progressively sharpening the property gradient the model must learn to follow.

Setup. Following Lee et al. (Lee et al., 2023), we train on ZINC250k (Irwin & Shoichet, 2005) and generate 3,000 candidates each for the five protein targets PARP1, FA7, 5HT1B, BRAF, and JAK2, scoring them with QuickVina2 (Alhossary et al., 2015) alongside QED and SA. We report two metrics from Yang et al. (Yang et al., 2021) with the novelty constraint of Lee et al. (Lee et al., 2023): *novel top 5% docking score* (*Dock.*, kcal/mol), the average docking score over the top 5% of generations satisfying QED>0.5, SA<5, and max Tanimoto-to-train <0.4; and *novel hit ratio* (*Hit*), the fraction of generations passing those filters and achieving a target-specific docking threshold. We compare against the seven baselines reported by Lee et al. (Lee et al., 2023), a *Random VAE* sanity check that decodes Gaussian samples from the DGVAE prior without guidance, and an upper-reference computed on the top 5% of ZINC250k itself. Curriculum schedule, per-target thresholds, hyperparameters, and the Random VAE and PhAME editing definitions are in Appendix D.

Results. Table 2 shows PhAME setting a new state of the art across the benchmark. PhAME achieves the best *Dock.* score on *all five targets* and is the only method whose top-5% generations surpass the top-5% of ZINC250k uniformly: PhAME's 3,000 outputs contain a head that is more potent than the strongest 12,500 molecules of its own training set. On novel hit ratio, PhAME leads on four of five targets, with relative gains of $+379\%$ over the next-best method on FA7, $+29\%$ on JAK2, and $+61\%$ on PARP1. The single regression is 5HT1B, where PhAME (13.6%) trails MOOD (18.7%) and FREED-QS (16.8%). On this target, however, ZINC250k itself reaches a 27.4% novel

Table 2: **Novel top 5% docking score (*Dock.*) and novel hit ratio (*Hit*).** Docking scores are reported in kcal/mol and hit ratios in %. PhAME achieves the best docking scores for all targets and the highest hit ratio for 4 out of 5 proteins. Cells report mean$_{\pm\text{std}}$ over 5 runs; ZINC250k is a top-5% reference and has no std. Best benchmark results are in bold and second-best underlined. PhAME editing† starts from 100 promising molecules and is reported for analysis only; it is excluded from the benchmark ranking and rank-based highlighting.

	PARP1		FA7		5HT1B		BRAF		JAK2	
Method	Dock.↓	Hit↑	Dock.↓	Hit↑	Dock.↓	Hit↑	Dock.↓	Hit↑	Dock.↓	Hit↑
ZINC250k	-10.59	5.66	-8.48	1.61	-10.51	27.36	-10.28	1.62	-9.75	7.77
REINVENT	-8.70 ±0.52	0.48 ±0.34	-7.21 ±0.26	0.21 ±0.08	-8.77 ±0.32	2.45 ±0.56	-8.39 ±0.40	0.13 ±0.09	-8.17 ±0.28	0.61 ±0.17
MORLD	-7.53 ±0.26	0.05 ±0.05	-6.26 ±0.17	0.01 ±0.01	-7.87 ±0.65	0.88 ±0.74	-8.04 ±0.34	0.05 ±0.04	-7.82 ±0.13	0.23 ±0.12
HierVAE	-9.49 ±0.28	0.55 ±0.21	-6.81 ±0.27	0.01 ±0.01	-8.08 ±0.25	0.51 ±0.28	-8.98 ±0.53	0.21 ±0.22	-8.29 ±0.37	0.23 ±0.13
FREED	-10.43 ±0.18	3.63 ±0.96	-8.30 ±0.09	1.11 ±0.21	-10.43 ±0.33	10.19 ±3.31	-10.33 ±0.16	2.07 ±0.63	-9.62 ±0.10	4.52 ±0.67
FREED-QS	-10.58 ±0.10	4.63 ±0.73	-8.38 ±0.04	1.33 ±0.11	-10.71 ±0.18	16.77 ±0.90	-10.56 ±0.08	2.94 ±0.36	-9.74 ±0.02	5.80 ±0.30
GDSS	-9.97 ±0.03	1.93 ±0.21	-7.78 ±0.04	0.37 ±0.10	-9.46 ±0.10	4.67 ±0.31	-9.22 ±0.07	0.17 ±0.13	-8.93 ±0.09	1.17 ±0.28
MOOD	-10.87 ±0.11	7.02 ±0.43	-8.16 ±0.07	0.73 ±0.14	-11.15 ±0.04	**18.67 ±0.42**	-11.06 ±0.03	5.24 ±0.29	-10.15 ±0.06	9.20 ±0.52
Random VAE	-8.51 ±0.04	0.10 ±0.04	-7.17 ±0.03	0.04 ±0.03	-8.48 ±0.07	1.43 ±0.14	-8.58 ±0.04	0.03 ±0.03	-8.28 ±0.05	0.33 ±0.06
PhAME (ours)	**-11.23 ±0.08**	**11.27 ±0.84**	**-9.14 ±0.06**	**6.37 ±0.65**	**-11.27 ±0.14**	13.58 ±1.96	**-11.09 ±0.14**	**6.10 ±1.16**	**-10.43 ±0.11**	**11.89 ±1.12**
PhAME editing†	-11.27 ±0.05	12.51 ±0.51	-9.31 ±0.11	8.61 ±1.66	-11.24 ±0.22	13.40 ±1.84	-11.18 ±0.19	7.09 ±1.33	-10.60 ±0.09	15.78 ±1.41

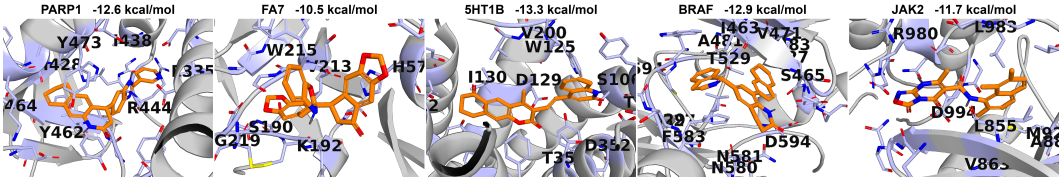


Figure 4: **Top-scoring docked poses of PhAME-generated molecules across five protein targets.** For each target, we visualize the highest-affinity molecule generated under docking-score guidance. The protein backbone is shown as cartoon; binding-pocket residues (any heavy-atom contact within 3.5 Å of the ligand) are drawn as light-blue sticks; the docked ligand is in orange.

hit rate, an order of magnitude above the other four, meaning the benchmark is dominated by the Tanimoto$<$0.4 novelty filter rather than by generative capacity: high-affinity 5HT1B chemotypes are densely populated in ZINC and any sufficiently active candidate is likely to share a scaffold with the training set. The gap between PhAME and MOOD here is small relative to the gaps PhAME opens on the other four targets.

We also evaluate PhAME in an editing setting, starting from a set of 100 promising molecules whose docking scores narrowly miss the hit threshold. Compared with de novo generation, editing achieves more favorable top-5% docking scores and higher novel hit ratios on four of the five targets. The strongest gains are observed for JAK2 and FA7, where the novel hit ratio increases from 11.89% to 15.78% and from 6.37% to 8.61%, respectively. Averaged across all targets, editing lowers the top-5% docking score by 0.09 kcal/mol and increases the novel hit ratio by 1.64 percentage points.

5.3 Transcriptomic Guided Generation

Complementary to the structure-based docking task, we evaluate PhAME on generating compounds conditioned on target-specific transcriptomic perturbations. This setting assesses whether the model can translate gene expression signatures into chemically plausible and target-relevant molecules, without relying on explicit structural or docking information.

Table 3: **Average molecular quality metrics across targets.** PhAME achieves the strongest molecular quality among gene-conditioned generation baselines. Cells report mean$_{\pm\text{std}}$ over targets.

Method	QED ↑	SA ↓	Uniqueness ↑	Novelty ↑
TRIOMPHE	0.39 ±0.03	6.02 ±0.29	**1.00 ±0.00**	**1.00 ±0.00**
GxVAEs	0.59 ±0.04	3.42 ±0.32	0.86 ±0.09	0.45 ±0.20
SmilesGEN	0.60 ±0.02	3.48 ±0.18	0.94 ±0.03	0.16 ±0.14
PhAME (ours)	**0.71 ±0.03**	**2.94 ±0.17**	**1.00 ±0.00**	**1.00 ±0.00**

Setup. We follow the evaluation protocol of Li & Yamanishi (Li & Yamanishi, 2024a; Liu et al., 2025b), where models are assessed on their ability to generate compounds conditioned on target-specific transcriptomic perturbations. Models are trained on paired transcriptomic–molecular data and evaluated against known ligands. For each target, we generate 100 molecules and report the maximum Tanimoto similarity of generated compounds

to known ligands. Knockdown profiles are used to generate inhibitors for *AKT1*, *AKT2*, *AURKB*, *CTSK*, *EGFR*, *HDAC1*, *MTOR*, and *PIK3CA*, while overexpression profiles are used for activators of *SMAD3* and *TP53*. Transcriptomic signatures are obtained from LINCS (Duan et al., 2014) using the *MCF7* cell line. While largely following Li & Yamanishi (Li & Yamanishi, 2024a), we identify data leakage due to overlap between training and evaluation molecules and remove the overlapping compounds, reducing the training set from 13,755 to 13,432 ($< 3\%$) (details in Appendix E).

Results. Table 3 shows that PhAME outperforms prior gene-conditioned models, achieving the highest drug-likeness (QED) and lowest synthetic accessibility (SA), while maintaining perfect uniqueness and novelty. In contrast, SmilesGEN exhibits low novelty, suggesting memorization of training molecules. Figure 5 shows that PhAME generates compounds structurally similar to known binders, achieving the highest Tanimoto similarity for 8 of 10 targets.

Figure 6 demonstrates PhAME's editing capability. Generation is initialized from the nearest training molecule in gene-expression space (cosine similarity). The left panel shows the starting compounds, and the right panel edited counterparts conditioned on target-specific expression profiles. In both cases, PhAME preserves the core molecular structure while increasing similarity to known binders, using only target knockdown profiles. For AKT1, PhAME improves similarity to known binders by 39%, while for HDAC1 the improvement reaches 90%.

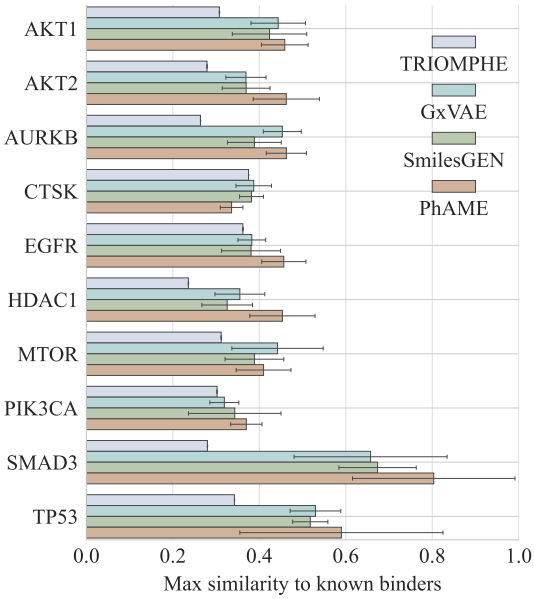


Figure 5: **Recovery of known binder motifs**. Maximum Tanimoto similarity to known binders for each target. PhAME generates compounds most similar to known binders for 8 out of 10 targets. Bars show mean ± std over 5 runs.

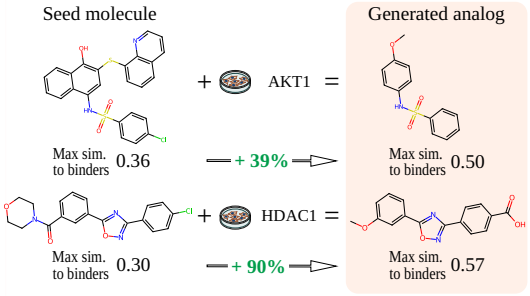


Figure 6: **Gene-driven recovery of known binders.** PhAME generates analogs with substantially increased similarity to known binders using only target-specific knockdown profiles. For AKT1, the maximum Tanimoto similarity to known binders improves by 39%, while for HDAC1 the improvement reaches 90%.

5.4 Cell Phenotype Guided Generation

We close with the most demanding setting in this paper: generating *de novo* molecules conditioned on a Cell Painting morphology centroid corresponding to a target Mode of Action (MoA). There is no synthesizable oracle that can rate a generated compound on its true cellular response, so we evaluate via a structural proxy: do the generations sit closer to the reference compounds of target MoA than to those of any other MoA, across multiple representation spaces?

Setup. We use the JUMP Cell Painting dataset (Chandrasekaran et al., 2023), for evaluation we intersected it with ChEMBL2K (Gaulton et al., 2012) and cross-referenced with the Broad Drug Repurposing Hub (Corsello et al., 2017) to obtain MoA labels, restricted to seven balanced MoA classes (8 reference compounds each). PhAME is run in its *de novo* mode (Section 4.2): we condition on the CellProfiler (Carpenter et al.) morphological centroid of each target MoA and sample 400 candidates per class (5 seeds $\times$ 80 valid samples). Dataset construction, feature provenance, contrastive-pairing thresholds, and inference hyperparameters are in Appendix F; baselines (DGVAE, CPMolGAN (Zapata et al., 2023), GFlowNet (Lu et al., 2024)) follow their original protocols.

Table 4: **MoA classification of *de novo* generated molecules evaluated in the ECFP structural embedding space.** Values are percentages (mean$_{\pm\text{std}}$ over 5 seeds). PhAME fundamentally outperforms competing methods, uniquely demonstrating superior results on all evaluated metrics.

	Top-1	MoA Retrieval Rate @k ↑			kNN Accuracy @k ↑		
Model	**Cluster Acc ↑**	**@1**	**@2**	**@3**	**@1**	**@3**	**@5**
DGVAE	14.0 ±0.9	14.2 ±2.0	28.5 ±1.1	42.7 ±2.1	14.5 ±0.6	14.6 ±1.0	13.7 ±1.6
CPMolGAN	14.3 ±1.0	13.4 ±1.3	28.8 ±2.4	42.0 ±0.9	12.9 ±1.2	13.5 ±1.4	14.1 ±1.4
GFlowNet	17.0 ±1.8	15.1 ±1.3	29.0 ±1.6	46.1 ±2.1	12.7 ±1.1	14.4 ±1.3	15.3 ±1.6
PhAME (ours)	**27.2 ±1.8**	**23.0 ±1.6**	**40.4 ±2.3**	**56.2 ±2.9**	**20.6 ±1.2**	**21.2 ±1.6**	**23.0 ±2.2**

Metrics. We score each generation against the seven MoA classes in three representation spaces with complementary inductive biases: ECFP (raw structural similarity), CLOOME embeddings (Sanchez-Fernandez et al., 2023) (jointly trained on molecule–cell pairs), and InfoAlign embeddings (information bottleneck over morphology and L1000). We report: *Top-1 Cluster Accuracy* (does the highest-scoring MoA class match the target?), *MoA Retrieval Rate* @k (does the target rank in the top k classes by average similarity to its n=3 nearest reference compounds?), and *kNN Accuracy* @k (majority-vote MoA among the k nearest references). The chance baseline for Top-1 Cluster Accuracy is $1/7 \approx 14.3\%$. The formal definitions of these metrics are given in Appendix F.

Results. Table 4 reveals a clean separation for evaluation on ECFP embedding space. DGVAE and CPMolGAN sit at the chance baseline (∼14.3% Top-1 Cluster Accuracy), recovering essentially no MoA information. GFlowNet shows only a marginal improvement over these baselines. In contrast, PhAME achieves a $+60\%$ relative gain over GFlowNet on Top-1 Cluster Accuracy (0.272 vs. 0.170), with the same pattern repeating across MoA Retrieval Rate @k and kNN Accuracy @k.

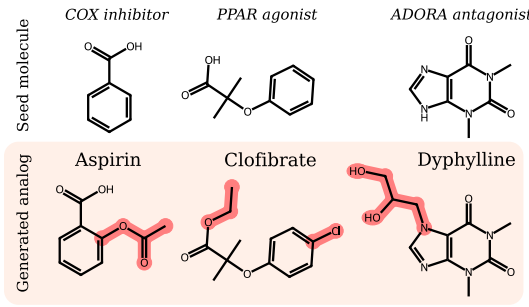


Figure 7: **Phenotype-driven recovery of known drugs.** PhAME, run in editing mode and seeded with a structurally similar hit, recovers analogs: aspirin (*cyclooxygenase inhibitor*, left), clofibrate (*PPAR receptor agonist*, middle) and dyphylline (*adenosine receptor antagonist*, right) when conditioned on the corresponding MoA centroid.

Crucially, extended results in Appendix F (Table 11) confirm that PhAME is the *only* method to clear the chance baseline across all three spaces (ECFP, CLOOME, InfoAlign). It significantly outperforms GFlowNet in InfoAlign (where the biological signal is richest) and recovers from a minor deficit in CLOOME Top-1 accuracy by dominating deeper retrieval metrics (MoA @1/2/3, kNN @1/3/5). Because these advantages hold consistently across structurally diverse representations, the gains reflect genuine MoA-aware chemistry rather than embedding-specific overfitting.

From de novo to drug recovery. The same model can be run in editing mode (Section 4.2) to recover known therapeutics: seeding with a structurally similar hit and conditioning on the target MoA, PhAME reconstructs aspirin, clofibrate and dyphylline from MoA-only guidance (Figure 7). Additional details are in Appendix G.

6 Conclusions

In this work, we introduced PhAME which combines discrete chemical graphs with high-dimensional biological signatures to formulate molecular design as an editing task within the latent space of a pretrained graph VAE. By introducing a novel compositional guidance mechanism that independently scales target signature and structural anchoring, PhAME resolves the problem of conflicting guidance signals. Consequently, it achieves SOTA performance across docking-score optimization, transcriptomic-conditioned generation, and Cell-Painting-driven Mode of Action generation while maintaining chemical validity and novelty. In future work, we plan to generalize this approach toward library design and hit generation tasks through multi-stage curriculum learning and to develop more robust biological oracles for evaluating de novo generated molecules.

AI USE STATEMENT

In this work, we used generative AI tools for polishing the text to improve readability and assisting with the implementation of our method. We have not used generative AI tools for proposing hypotheses, designing research methodology or experiments, cleaning and reformatting datasets, or supporting qualitative and thematic data analysis. The remaining tasks requiring disclosure were not applicable to this work. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims, code, and other artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The implementation of our model and the code for reproducing the experiments can be found in the supplementary material. The code will be publicly available under an MIT license upon publication of the paper. All experiments were conducted on a single NVIDIA GH200 GPU with 96 GB of memory and a single NVIDIA Grace 72-core CPU @ 3.1 GHz, using up to 32.0 GB of system RAM. Our experiments were carried out using Python 3.11, with PyTorch 2.10.0, PyTorch Lightning 2.5.1, TorchGeometric 2.7.0, RDKit 2023.09.6, InfoAlign 0.1.1. The full Python environment, including all additional requirements, is available in the code repository.

REFERENCES

Amr Alhossary, Stephanus Daniel Handoko, Yuguang Mu, and Chee-Keong Kwoh. Fast, accurate, and reliable molecular docking with quickvina 2. *Bioinformatics*, 31(13):2214–2216, 2015.

Mark-Anthony Bray, Shantanu Singh, Han Han, Chadwick T Davis, Blake Borgeson, Cathy Hartland, Maria Kost-Alimova, Sigrun M Gustafsdottir, Christopher C Gibson, and Anne E Carpenter. Cell painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. *Nature protocols*, 11(9):1757–1774, 2016.

A. E. Carpenter, T. R. Jones, M. R. Lamprecht, C. Clarke, I. H. Kang, O. Friman, D. A. Guertin, J. H. Chang, R. A. Lindquist, J. Moffat, P. Golland, and D. M. Sabatini. Cellprofiler: image analysis software for identifying and quantifying cell phenotypes.

Srinivas Niranj Chandrasekaran, Hugo Ceulemans, Justin D Boyd, and Anne E Carpenter. Image-based profiling for drug discovery: due for a machine-learning upgrade? *Nature reviews drug discovery*, 20(2):145–159, 2021.

Srinivas Niranj Chandrasekaran, Jeanelle Ackerman, Eric Alix, D. Michael Ando, John Arevalo, Melissa Bennion, Nicolas Boisseau, Adriana Borowa, Justin D. Boyd, and et al. Brino. Jump cell painting dataset: morphological impact of 136,000 chemical and genetic perturbations. *bioRxiv*, 2023. doi: 10.1101/2023.03.23.534023. URL https://www.biorxiv.org/content/early/2023/03/24/2023.03.23.534023.

Steven M Corsello, Joshua A Bittker, Zihan Liu, Joshua Gould, Patrick McCarren, Jodi E Hirschman, Stephen E Johnston, Anita Vrcic, Bang Wong, Mariya Khan, Jacob Asiedu, Rajiv Narayan, Christopher C Mader, Aravind Subramanian, and Todd R Golub. The drug repurposing hub: a next-generation drug library and information resource. *Nature medicine*, 2017.

Qiaonan Duan, Corey Flynn, Mario Niepel, Marc Hafner, Jeremy L Muhlich, Nicolas F Fernandez, Andrew D Rouillard, Christopher M Tan, Edward Y Chen, Todd R Golub, et al. Lincs canvas browser: interactive web app to query, browse and interrogate lincs l1000 gene expression signatures. *Nucleic acids research*, 42(W1):W449–W460, 2014.

Anna Gaulton, Louisa J. Bellis, A. Patricia Bento, Jon Chambers, Mark Davies, Anne Hersey, Yvonne Light, Shaun McGlinchey, David Michalovich, Bissan Al-Lazikani, and John P. Overington. Chembl: a large-scale bioactivity database for drug discovery. *Nucleic Acids Research*, 2012.

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

Abhor Gupta, Barathi Lenin, Sean Current, Rohit Batra, Balaraman Ravindran, Karthik Raman, and Srinivasan Parthasarathy. Pure: policy-guided unbiased representations for structure-constrained molecular generation. *Journal of Cheminformatics*, 17(1):1–18, 2025.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022. URL https://arxiv.org/abs/2207.12598.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf.

Emiel Hoogeboom, Vıctor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3d. In *International conference on machine learning*, pp. 8867–8887. PMLR, 2022.

Jameed Hussain and Ceara Rea. Computationally efficient algorithm to identify matched molecular pairs (mmps) in large data sets. *Journal of chemical information and modeling*, 50(3):339–348, 2010.

John J Irwin and Brian K Shoichet. Zinc- a free database of commercially available compounds for virtual screening. *Journal of chemical information and modeling*, 45(1):177–182, 2005.

Qun Jiang, Xing Ye, Zekun Guo, Yingce Xia, Zequn Liu, Jiaying Xu, Peiran Jin, Fusong Ju, Huanhuan Xia, Shangya Feng, et al. Phenotype-guided in silico molecular generation using large language models. *bioRxiv*, pp. 2026–01, 2026.

Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Junction tree variational autoencoder for molecular graph generation. In *International conference on machine learning*, pp. 2323–2332. PMLR, 2018.

Bowen Jing, Gabriele Corso, Jeffrey Chang, Regina Barzilay, and Tommi Jaakkola. Torsional diffusion for molecular conformer generation. *Advances in neural information processing systems*, 35:24240–24253, 2022.

Kazuma Kaitoh and Yoshihiro Yamanishi. Triomphe: Transcriptome-based inference and generation of molecules with desired phenotypes by machine learning. *Journal of Chemical Information and Modeling*, 61(9):4303–4320, 2021.

Seul Lee, Jaehyeong Jo, and Sung Ju Hwang. Exploring chemical space with score-based out-of-distribution generation. In *International Conference on Machine Learning*, pp. 18872–18892. PMLR, 2023.

Seul Lee, Karsten Kreis, Srimukh Prasad Veccham, Meng Liu, Danny Reidenbach, Yuxing Peng, Saee Gopal Paliwal, Weili Nie, and Arash Vahdat. Genmol: A drug discovery generalist with discrete diffusion. In *International Conference on Machine Learning*, pp. 33205–33226. PMLR, 2025.

Chen Li and Yoshihiro Yamanishi. Gxvaes: Two joint vaes generate hit molecules from gene expression profiles. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(12):13455–13463, Mar. 2024a. doi: 10.1609/aaai.v38i12.29248. URL https://ojs.aaai.org/index.php/AAAI/article/view/29248.

Chen Li and Yoshihiro Yamanishi. Gxvaes: Two joint vaes generate hit molecules from gene expression profiles. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 13455–13463, 2024b.

Zhaokang Liang, Shuyang Zhuang, Xiaoran Jiao, Weian Mao, Hao Chen, and Chunhua Shen. scppdm: A diffusion model for single-cell drug-response prediction. *arXiv preprint arXiv:2510.11726*, 2025.

Gang Liu, Srijit Seal, John Arevalo, Zhenwen Liang, Anne E Carpenter, Meng Jiang, and Shantanu Singh. Learning molecular representation in a cell. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=BbZy8nI1si.

Hui Liu, Shiye Tian, and Xuejun Liu. Phenotypic profile-informed generation of drug-like molecules via dual-channel variational autoencoders. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, IJCAI '25, 2025b. ISBN 978-1-956792-06-5. doi: 10.24963/ijcai.2025/1008. URL https://doi.org/10.24963/ijcai.2025/1008.

Hui Liu, Shiye Tian, and Xuejun Liu. Phenotypic profile-informed generation of drug-like molecules via dual-channel variational autoencoders. *arXiv preprint arXiv:2506.02051*, 2025c.

Alexander Arjun Lobo, Urvi Awasthi, and Leonid Zhukov. Property-guided molecular generation and optimization via latent flows. *arXiv preprint arXiv:2603.26889*, 2026.

Hannes H Loeffler, Jiazhen He, Alessandro Tibo, Jon Paul Janet, Alexey Voronov, Lewis H Mervin, and Ola Engkvist. Reinvent 4: Modern ai–driven generative molecule design. *Journal of Cheminformatics*, 16(1):20, 2024.

Stephen Zhewen Lu, Ziqing Lu, Ehsan Hajiramezanali, Tommaso Biancalani, Yoshua Bengio, Gabriele Scalia, and Michał Koziarski. Cell morphology-guided small molecule generation with gflownets. In *ICML 2024 Workshop on Structured Probabilistic Inference {\&} Generative Modeling*, 2024.

Łukasz Maziarka, Agnieszka Pocha, Jan Kaczmarczyk, Krzysztof Rataj, Tomasz Danel, and Michał Warchoł. Mol-cyclegan: a generative model for molecular optimization. *Journal of Cheminformatics*, 12(1):2, 2020.

Oscar Méndez-Lucio, Benoit Baillif, Djork-Arné Clevert, David Rouquié, and Joerg Wichard. De novo generation of hit-like molecules from gene expression signatures using artificial intelligence. *Nature communications*, 11(1):10, 2020.

Flavio M. Morelli, Vladislav Kim, Franziska Hecker, Sven Geibel, and Paula A. Marín Zapata. unidino: Assay-independent feature extraction for fluorescence microscopy images. *Computational and Structural Biotechnology Journal*, 27:928–936, 2025. ISSN 2001-0370. doi: https://doi.org/10.1016/j.csbj.2025.02.020. URL https://www.sciencedirect.com/science/article/pii/S2001037025000479.

Nikita Moshkov, Michael Bornholdt, Santiago Benoit, Matthew Smith, Claire McQuin, Allen Goodman, Rebecca A. Senft, Yu Han, Mehrtash Babadi, Peter Horvath, Beth A. Cimini, Anne E. Carpenter, Shantanu Singh, and Juan C. Caicedo. Learning representations for image-based profiling of perturbations. *bioRxiv*, 2022. doi: 10.1101/2022.08.12.503783. URL https://www.biorxiv.org/content/early/2022/10/03/2022.08.12.503783.

Trieu Nguyen and Aleksandra Karolak. Expanding molecular design with graph variational autoencoders: A comparative study of pair-encoding and character tokenization. *ChemRxiv*, 2025(0523), 2025. doi: 10.26434/chemrxiv-2025-j1p93.

Sarfaraz K Niazi. The coming of age of ai/ml in drug discovery, development, clinical testing, and manufacturing: the fda perspectives. *Drug Design, Development and Therapy*, pp. 2691–2725, 2023.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pp. 8162–8171. PMLR, 2021.

Alessandro Palma, Till Richter, Hanyi Zhang, Manuel Lubetzki, Alexander Tong, Andrea Dittadi, and Fabian Theis. Multi-modal and multi-attribute generation of single cells with cfgen. *arXiv preprint arXiv:2407.11734*, 2024.

Pavel G Polishchuk, Timur I Madzhidov, and Alexandre Varnek. Estimation of the size of drug-like chemical space based on gdb-17 data. *Journal of computer-aided molecular design*, 27(8):675–679, 2013.

Daniil Polykovskiy, Alexander Zhebrak, Benjamin Sanchez-Lengeling, Sergey Golovanov, Oktai Tatanov, Stanislav Belyaev, Rauf Kurbanov, Aleksey Artamonov, Vladimir Aladinskiy, Mark Veselov, et al. Molecular sets (moses): a benchmarking platform for molecular generation models. *Frontiers in pharmacology*, 11:565644, 2020.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Ana Sanchez-Fernandez, Elisabeth Rumetshofer, Sepp Hochreiter, and Günter Klambauer. Cloome: contrastive learning unlocks bioimaging databases for queries with chemical structures. 2023. doi: 10.1101/2022.11.17.516915.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021a.

Ran Song and Hui Liu. Phenomoler: Phenotype-guided molecular optimization via chemistry large language model. *arXiv preprint arXiv:2509.21424*, 2025.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021b.

Aravind Subramanian, Rajiv Narayan, Steven M. Corsello, David D. Peck, Ted E. Natoli, Xiaodong Lu, Joshua Gould, John F. Davis, Andrew A. Tubelli, Jacob K. Asiedu, David L. Lahr, Jodi E. Hirschman, Zihan Liu, Melanie Donahue, Bina Julian, Mariya Khan, David Wadden, Ian C. Smith, Daniel Lam, Arthur Liberzon, Courtney Toder, Mukta Bagul, Marek Orzechowski, Oana M. Enache, Federica Piccioni, Sarah A. Johnson, Nicholas J. Lyons, Alice H. Berger, Alykhan F. Shamji, Angela N. Brooks, Anita Vrcic, Corey Flynn, Jacqueline Rosains, David Y. Takeda, Roger Hu, Desiree Davison, Justin Lamb, Kristin Ardlie, Larson Hogstrom, Peyton Greenside, Nathanael S. Gray, Paul A. Clemons, Serena Silver, Xiaoyun Wu, Wen-Ning Zhao, Willis Read-Button, Xiaohua Wu, Stephen J. Haggarty, Lucienne V. Ronco, Jesse S. Boehm, Stuart L. Schreiber, John G. Doench, Joshua A. Bittker, David E. Root, Bang Wong, and Todd R. Golub. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*, 171(6):1437–1452.e17, 2017. ISSN 0092-8674. doi: https://doi.org/10.1016/j.cell.2017.10.049. URL `https://www.sciencedirect.com/science/article/pii/S0092867417313090`.

David C Swinney. Phenotypic vs. target-based drug discovery for first-in-class medicines. *Clinical Pharmacology & Therapeutics*, 93(4):299–301, 2013.

Qiaosi Tang, Daoyun Ding, Xiaoyong Yuan, Gustavo Seabra, Peter A Ramdhan, Chi-Yuan Liu, My T. Thai, Chenglong Li, Hendrik Luesch, and Yanjun Li. Mgmg: Cell morphology-guided molecule generation for drug discovery. *bioRxiv*, 2025. doi: 10.1101/2025.07.11.664424.

Clément Vignac, Igor Krawczuk, Antoine Siraudin, Bohan Wang, Volkan Cevher, and Pascal Frossard. Digress: Discrete denoising diffusion for graph generation. In *Proceedings of the 11th International Conference on Learning Representations*, 2023.

Conghao Wang, Hiok Hian Ong, Shunsuke Chiba, and Jagath C Rajapakse. Gldm: hit molecule generation with constrained graph latent diffusion model. *Briefings in bioinformatics*, 25(3): bbae142, 2024.

Tomer Weiss, Eduardo Mayo Yanes, Sabyasachi Chakraborty, Luca Cosmo, Alex M Bronstein, and Renana Gershoni-Poranne. Guided diffusion for inverse molecular design. *Nature Computational Science*, 3(10):873–882, 2023.

Soojung Yang, Doyeong Hwang, Seul Lee, Seongok Ryu, and Sung Ju Hwang. Hit and lead discovery with explorative rl and fragment-based molecule generation. *Advances in Neural Information Processing Systems*, 34:7924–7936, 2021.

Paula A Marin Zapata, Oscar Méndez-Lucio, Tuan Le, Carsten Jörn Beese, Jörg Wichard, David Rouquié, and Djork-Arné Clevert. Cell morphology-guided de novo hit design by conditioning gans on phenotypic image features. *Digital discovery*, 2(1):91–102, 2023.

Cheng Zeng, Jirui Jin, Connor Ambrose, George Karypis, Mark Transtrum, Ellad B Tadmor, Richard G Hennig, Adrian Roitberg, Stefano Martiniani, and Mingjie Liu. Propmolflow: property-guided molecule generation with geometry-complete flow matching. *Nature Computational Science*, pp. 1–10, 2026.

Qunhao Zhang, Jun Xiao, Dongjiang Niu, Zhixin Zhang, Shanyang Ding, and Zhen Li. Geometry-complete latent diffusion model for 3d molecule generation. *Bioinformatics*, 41(8):btaf426, 2025.

A Training Details

Python environment. All experiments were carried out using Python 3.11.5. The key libraries are listed in Table 5. Other requirements are included in the code repository.

Table 5: **Key libraries used in the experiments.**

Library	Version
PyTorch	2.10.0
PyTorch Lightning	2.5.1
TorchGeometric	2.7.0
RDKit	2023.09.6
InfoAlign	0.1.1
Scikit-learn	1.8.0
Pandas	3.0.2
Numpy	1.26.4

Computational resources. Experiments were conducted on a single NVIDIA Grace 72-core CPU @ 3.1 GHz and a single NVIDIA GH200 GPU with 96 GB of memory, using up to 32.0 GB of system RAM.

B Dataset

Specifically, for each molecule in one subset, we identify the nearest neighbor in the opposite subset based on molecular similarity and form a training pair. Structural similarity is computed using the Tanimoto similarity between Morgan fingerprints. This pairing strategy provides supervision signals that encourage the model to learn transformations that move molecules across the threshold while preserving structural similarity to the source structure. We denote this strategy as nearest-neighbor and use by default in all experiments.

C Details of logP Optimization

Dataset. To construct the training and testing datasets of pairs, we compute the logP values using the RDKit implementation of the Crippen model (i.e., `Crippen.MolLogP`). We then calculate the median logP value on the training set and use it as a threshold to split both sets. The training set contains an equal number of samples below and above this threshold, while the test set is split using the same value and undersampled to match the minority class. The distribution of logP values for both splits is shown in Figure 8.

Model. For this task, we used the c-CFG (Compositional Classifier-Free Guidance) variant of PhAME. Unless otherwise stated, all reported results use $w_c = 6$ and $w_a = 3$. The model was trained for 1000 epochs with a learning rate of 1×10^{-4}, a batch size of 512, and 1000 diffusion steps. The alignment loss weight was set to $\gamma = 1$.

During inference, we initialize the process from a partially noised seed molecule at timestep $T = 900$, and perform the reverse diffusion process over the remaining steps.

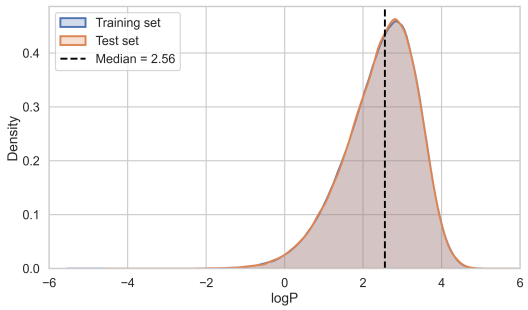


Figure 8: **Distribution of logP in the MOSES Dataset.** Distribution of logP values for the training and test sets derived from the MOSES dataset. The dashed vertical line denotes the global median logP of the train set.

Baselines implementation. For our baselines, we adopted Mol-CycleGAN, which has a process most similar to our approach, it utilizes paired training. For DiGress, we used classifier guidance to steer the model toward the appropriate logP. However, it is worth noting that this model lacks an explicit mechanism for direct editing. PURE was trained following the setup described in its original paper. Crucially, our experimental setup does not allow for oracle-based optimization or filtering, as we rely solely on the provided labeled dataset. Therefore, to ensure a fair and equitable comparison, all post-processing steps were strictly disabled across all evaluated models.

Extended results. Table 7 summarizes the contribution of individual components. Standard Classifier-Free Guidance (denoted as CFG) achieves a reasonable trade-off, with moderate target success and similarity, resulting in limited performance in terms of *NTS*. Employing Compositional Classifier-Free Guidance (denoted as c-CFG) with only the property signal active (c-CFG(w_c=6, w_a=0)) maximizes target success but severely degrades similarity, indicating that optimizing logP alone leads to substantial structural drift. Conversely, using only alignment guidance (c-CFG(w_c=0, w_a=3)) preserves high similarity but substantially reduces target success. Combining both signals (c-CFG(w_c=6, w_a=3)) yields a better balance between objectives. Removing the alignment loss decreases similarity and overall *NTS*, indicating that it helps the model learn stronger structural alignment with the seed molecule. Finally, we demonstrate that the model can be aggressively steered to reach specific performance peaks. (c-CFG(w_c=25, w_a=0)) forces the model to achieve perfect target success. Applying a strong joint conditioning strategy (c-CFG(w_c=12, w_a=3)) yields the highest absolute *NTS*, though the increased guidance signal comes at the cost of molecular validity.

As shown in Tables 6, 7, quality metrics such as *Val*, *Uniq*, *QED*, and *SA* remain broadly comparable across methods, with no approach consistently dominating all criteria. One notable exception is standard CFG, which significantly reduces validity, indicating that stronger guidance can lead to chemically invalid structures. In contrast, methods incorporating alignment, including PhAME, maintain more stable validity while preserving competitive performance across the remaining quality metrics.

Table 6: **Comparison of methods for logP optimization with similarity preservation to the seed molecule.** Notably, PhAME achieves the highest score on the aggregated *NTS* metric while maintaining strong, balanced performance across individual ones. Cells show mean ± std over 5 runs.

Method	Val ↑	Uniq ↑	QED ↑	SA ↓	Nov ↑	Tgt ↑	Sim ↑	NTS ↑
Random target class	1.00 ±0.00	0.67 ±0.00	0.81 ±0.00	2.45 ±0.00	0.00 ±0.00	1.00 ±0.00	0.14 ±0.00	0.00 ±0.00
Nearest target class	1.00 ±0.00	0.34 ±0.00	0.82 ±0.00	2.28 ±0.00	0.00 ±0.00	1.00 ±0.00	0.60 ±0.00	0.00 ±0.00
VAE reconstruction	0.91 ±0.00	0.53 ±0.00	0.80 ±0.00	2.70 ±0.00	0.99 ±0.00	0.19 ±0.00	0.59 ±0.00	0.11 ±0.00
Mol-CycleGAN	0.72 ±0.00	0.90 ±0.01	**0.82 ±0.01**	**2.57 ±0.07**	0.99 ±0.00	0.51 ±0.04	0.15 ±0.00	0.08 ±0.01
DiGrees	0.83 ±0.03	**0.99 ±0.00**	0.77 ±0.01	2.87 ±0.07	**1.00 ±0.00**	**1.00 ±0.00**	0.13 ±0.00	0.13 ±0.00
PURE	**0.86 ±0.00**	0.96 ±0.00	0.54 ±0.01	2.83 ±0.02	**1.00 ±0.00**	0.45 ±0.00	**0.58 ±0.00**	0.26 ±0.00
PhAME	0.71 ±0.03	0.88 ±0.01	0.74 ±0.01	2.96 ±0.05	0.99 ±0.00	0.78 ±0.03	0.39 ±0.02	**0.30 ±0.01**

Table 7: **Ablation of PhAME for logP optimization with similarity preservation to the seed molecule.** CFG refers to standard Classifier-Free Guidance, while c-CFG denotes compositional CFG. Cells show mean ± std over 5 runs. The results illustrate a clear trade-off across all metrics depending on the guidance scales.

Method	Val ↑	Uniq ↑	QED ↑	SA ↓	Nov ↑	Tgt ↑	Sim ↑	NTS ↑
CFG(3)	0.42 ±0.02	0.89 ±0.01	0.70 ±0.01	**2.80 ±0.06**	0.99 ±0.00	0.68 ±0.01	0.40 ±0.00	0.27 ±0.01
c-CFG(6, 0)	**0.87 ±0.02**	**0.97 ±0.01**	0.74 ±0.01	2.95 ±0.07	**1.00 ±0.00**	0.96 ±0.01	0.13 ±0.00	0.12 ±0.00
c-CFG(0, 3)	0.73 ±0.03	0.88 ±0.00	**0.76 ±0.01**	2.93 ±0.03	0.99 ±0.00	0.59 ±0.01	**0.41 ±0.01**	0.24 ±0.00
c-CFG(6, 3) w/o $\mathcal{L}_{\text{align}}$	0.80 ±0.02	0.94 ±0.00	0.74 ±0.00	2.81 ±0.02	0.99 ±0.00	0.81 ±0.01	0.35 ±0.00	0.28 ±0.00
c-CFG(6, 3)	0.71 ±0.03	0.88 ±0.01	0.74 ±0.01	2.96 ±0.05	0.99 ±0.00	0.78 ±0.03	0.39 ±0.02	0.30 ±0.01
c-CFG (25, 0)	0.67 ±0.10	0.86 ±0.03	0.62 ±0.07	2.90 ±0.15	**1.00 ±0.00**	**1.00 ±0.01**	0.12 ±0.01	0.12 ±0.01
c-CFG (12, 3)	0.67 ±0.04	0.89 ±0.01	0.71 ±0.02	2.99 ±0.07	0.99 ±0.00	0.87 ±0.04	0.36 ±0.02	**0.31 ±0.01**

D Details of Docking Score Optimization

Three-stage curriculum. PhAME is trained in three stages on ZINC250k, following the pairing protocol of Section 4.3: at each stage, training pairs consist of molecules with QuickVina2 docking scores above and below a percentile-based threshold (non-binders vs. binders), with the most structurally similar pair selected across the threshold. Across stages, the threshold tightens from the 50th to the 90th and finally to the 99.8th percentile, so each stage fine-tunes the denoising network on a progressively higher-affinity property partition, a difficulty schedule rather than a self-distillation loop. Per-target percentile thresholds for ZINC250k are listed in Table 8; per-stage training hyperparameters and target docking value (between the 99.8th and 100th percentile) are summarized in Table 9.

Model. For this task, we used the c-CFG implementation of PhAME. All hyperparameters are listed in Table 9.

Hit thresholds. As in MOOD (Lee et al., 2023), the docking-score thresholds used in the *novel hit ratio* are 10.0 for PARP1, 8.5 for FA7, 8.7845 for 5HT1B, 9.1 for JAK2, and 10.3 for BRAF.

Random VAE baseline. *Random VAE* samples 3,000 latents from $\mathcal{N}(0, I)$ in the DGVAE latent space and decodes them without guidance, providing a no-conditioning lower-bound on docking performance.

PhAME editing. We use the same five final Stage 3 checkpoints as for PhAME generation. For each target, we select 100 molecules with docking scores closest to, but below, the hit threshold. Their DGVAE latents are noised to diffusion step $t^* = 998$ and subsequently denoised using the target-specific property guidance scales and target docking-score conditioning values reported in Table 9.The molecular guidance scales w_a are set to 1.0, 4.0, 2.5, 4.0, and 4.0 for PARP1, FA7, 5HT1B, BRAF, and JAK2, respectively.

Table 8: **Percentile thresholds and numbers of compounds exceeding the threshold for different targets in the ZINC250k dataset.**

Percentile	Count ($\geq$ Thr.)	PARP1 Thr.	FA7 Thr.	5HT1B Thr.	BRAF Thr.	JAK2 Thr.
0.10	227,000	6.4	5.5	5.9	6.6	6.4
0.20	200,000	7.1	6.0	6.5	7.2	7.0
0.50	126,000	8.2	6.7	7.9	8.2	7.9
0.80	55,000	9.2	7.4	9.1	9.0	8.6
0.90	27,000	9.7	7.8	9.6	9.5	9.0
0.95	14,000	10.1	8.1	10.0	9.8	9.4
0.99	3,000	10.9	8.7	10.8	10.5	10.0
0.998	600	11.5	9.2	11.4	11.1	10.5
1.0	1	14.9	16.0	16.2	14.7	13.4

Table 9: **Training hyperparameters for different targets.**

	PARP1	FA7	5HT1B	BRAF	JAK2
Stage 1 threshold	8.2	7.4	7.9	9.0	7.9
Stage 2 threshold	10.9	8.5	10.8	10.5	10.0
Stage 3 threshold	11.5	8.7	11.4	11.0	10.5
Stage 1 LR	1e$-$3	1e$-$3	1e$-$3	1e$-$3	1e$-$3
Stage 1 epochs	1000	1000	1000	1000	1000
Stage 2 LR	1e$-$3	1e$-$3	1e$-$3	1e$-$4	1e$-$3
Stage 2 epochs	500	500	1500	500	900
Stage 3 LR	1e$-$6	1e$-$4	1e$-$6	1e$-$5	1e$-$6
Stage 3 epochs	500	800	500	800	500
γ	0.01	0.01	0.01	0.01	0.01
Batch size	512	512	512	512	512
w_c	3.5	2.0	5.0	4.0	3.0
w_a	0	0	0	0	0
Target	12.0	13.0	13.0	14.7	12.5

Table 10: **Properties of the top 5 molecules ranked by docking score across all targets, satisfying the constraints QED > 0.5, SA < 5, and a maximum Tanimoto similarity to the training set of < 0.4.**

	ID	Similarity	Docking Score ↓	QED ↑	SA ↓
PARP1	12	0.39	-12.60	0.50	3.09
	105	0.30	-12.50	0.52	4.05
	43	0.38	-12.40	0.58	3.08
	107	0.38	-12.40	0.57	3.63
	30	0.36	-12.40	0.67	3.44
FA7	25	0.38	-10.60	0.67	3.75
	43	0.27	-10.40	0.51	3.33
	64	0.33	-10.20	0.51	4.02
	70	0.37	-10.20	0.51	3.30
	91	0.38	-10.10	0.68	3.23
5HT1B	14	0.35	-13.30	0.56	3.12
	29	0.38	-12.30	0.60	3.45
	39	0.33	-12.20	0.52	2.98
	142	0.37	-12.20	0.63	3.76
	83	0.36	-12.20	0.58	3.21
BRAF	9	0.29	-12.90	0.57	3.46
	17	0.36	-12.80	0.52	3.53
	107	0.34	-12.40	0.78	4.12
	128	0.39	-12.30	0.67	3.42
	99	0.32	-12.30	0.57	2.63
JAK2	13	0.36	-12.00	0.53	4.25
	46	0.31	-11.70	0.56	3.03
	65	0.39	-11.50	0.51	2.48
	93	0.38	-11.40	0.62	3.93
	116	0.36	-11.30	0.53	2.83

Three-stage curriculum ablation. To attribute the gain reported in Section 5.2 to the iterative threshold-refresh curriculum rather than to the DGVAE backbone or compositional guidance alone, we evaluate a single-stage variant of PhAME on FA7: training only with the stage-1 threshold (50th percentile) and skipping the bootstrapped stage-2 and stage-3 fine-tuning. As can be seen in Figure 9, the single-stage variant achieves a novel hit ratio of 3.7% (vs. 6.37% for three-stage PhAME), confirming that the curriculum is the primary driver of the improvement.

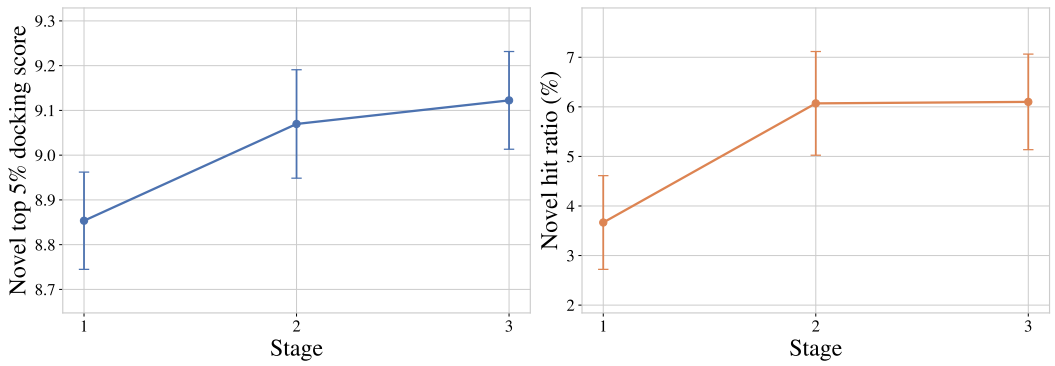


Figure 9: **Generative performance across training stages for FA7.** Novel top 5% docking score (left) and novel hit ratio (right) for molecules generated from Gaussian noise after each training stage. Points show means across five random seeds, and error bars indicate $\pm$ one standard deviation. Higher values indicate better performance for both metrics.

Quantitative top-5 hits. Table 10 shows the quality metrics for the top five molecules for each hit. We report the QuickVina2 docking score, QED, SA, and the maximum Tanimoto similarity to any molecule in the ZINC250k training set.

E Details of Transcriptomic Guided Generation

Model. For this task, we used the c-CFG implementation of PhAME with condition guidance $w_c = 1$ and molecular-alignment guidance $w_a = 1$. Training was performed on the `MCF7` dataset using five random seeds $\{0, 1, 2, 3, 4\}$. For each molecule, up to three matched training pairs were used, selected with a minimum expression cosine distance of 0.5/. For testing 10% of the data was held out.

The model was optimized for 4000 epochs with a learning rate of 1×10^{-4}, batch size 512, 1000 diffusion steps, and $\mathcal{L}_{\text{align}} = 0.01$. Gene-expression conditions and molecular-alignment embeddings were projected into the latent conditioning space before being concatenated with the noisy latent representation and the time embedding. Specifically, the gene-expression condition was mapped from the full 978-dimensional input space to the latent dimension using a linear projection layer, while the molecular embedding was projected through a two-layer MLP with a SiLU activation.

As a seed molecule, we selected one of the ten most similar molecules based on cosine similarity between its gene expression and the target expression from the training set. During benchmarking, 100 molecules were generated for each target. Each of the ten seed molecules was used 10 times.

Data leakage handling. While largely following the original setup of Li & Yamanishi (Li & Yamanishi, 2024a; Liu et al., 2025b), we introduce a minor but important modification to prevent potential data leakage. Specifically, we identify a subset of molecules that appear both in the training set and among the reference ligands used for evaluation. These overlapping molecules are removed from the training data to ensure a strict separation between training and evaluation. This filtering reduces the training set size from 13,755 to 13,432 molecules (less than 3% of the data).

Dataset creation. We construct a paired compound–transcriptome dataset as described in Section 4.3. Unlike the property-based setting, transcriptomic data does not admit a natural threshold for partitioning, and we therefore adopt a similarity-based pairing strategy. Specifically, we compute cosine similarity between normalized gene expression profiles. For each molecule, we select the most structurally similar candidates based on Tanimoto similarity over ECFP fingerprints, while enforcing transcriptomic dissimilarity, $\cos(g_i, g_j) < 0.5$, where g denotes the gene expression profile. This results in training pairs that are structurally similar but induce distinct transcriptional responses.

F Details of Cell Phenotype Guided Generation

Dataset. PhAME was trained via a two-stage strategy that uses distinct threshold parameters to construct both the general and the fine-tuning datasets. For the general dataset, the threshold was set to $\tau_{\text{cos}} = -0.3$. For fine-tuning, we selected a much stricter threshold: $\tau_{\text{cos}} = -0.6$. In both stages, we first isolated all candidates whose phenotypic cosine similarity fell below τ_{cos} (computed on features extracted from CellProfiler). Among these, we selected the pair with the highest Tanimoto similarity. This protocol is crucial because mining a sufficiently large dataset of highly divergent biological responses and strongly similar chemical structures is difficult. The proposed regime provides the volume needed for stable pre-training and then strictly fine-tunes the model on the desired dataset. As outlined, the dataset was created from the CP-JUMP database, with features extracted from CellProfiler. The test set was generated using the CP-JUMP database and intersected with ChEMBL2K and the Broad Drug Repurposing Hub. Crucially, to ensure a perfectly balanced and unbiased evaluation (without intersection of MoA classes), we restricted our focus to seven distinct MoA classes comprising exactly 8 representative molecules per class: *acetylcholine receptor antagonist*, *adrenergic receptor agonist*, *adrenergic receptor antagonist*, *cyclooxygenase inhibitor*, *dopamine receptor antagonist*, *histamine receptor antagonist*, and *serotonin receptor antagonist*.

Feature provenance. At inference, PhAME is conditioned on the per-MoA centroid (mean over the 8 reference compounds in each class) of CellProfiler (Carpenter et al.) morphological features

computed on each compound's CP-JUMP (Chandrasekaran et al., 2023) profiles. We use the specific preprocessed feature flavor distributed by the InfoAlign authors (Liu et al., 2025a) on their public HuggingFace repository, which we denote *CellProfiler* in our ablations. A head-to-head comparison of CellProfiler against two alternative cellular feature extractors (*DeepProfiler* and *uniDino*) is reported in the Morphology Representation ablation below.

Training details and baselines. For the training of the other baselines CPMolGAN and GFlowNet, we strictly follow their originally proposed training procedures and hyperparameter configurations. We omit the recent method MGMG (Tang et al., 2025) from our evaluation framework, as its codebase is not publicly available. PhAME was optimized for 2000 epochs with a learning rate of 1×10^{-5} in fine-tune stage and 1×10^{-3} in general stage, batch size 512, 1000 diffusion steps, with early stopping for 50 epochs, warmup 200 epochs, dropout 0.1. However, because this specific evaluation focuses on *de novo* generation rather than localized structural optimization, we adjusted the training and inference parameters accordingly. During training, we incorporated EMA (decay factor of 0.999) optimization to stabilize training and reduced the structural-alignment weight γ to 0.00001. We set the structural guidance scale to $w_a = 0$ to allow unconstrained exploration, and the conditioning scale to $w_c = 34$ to enforce a strong cellular signal, a value determined empirically to yield the highest biological fidelity.

Specifically, for each of the 7 MoA classes, the generative process was conditioned directly on the pre-computed biological centroid of that specific class. To ensure statistical robustness in our evaluation, we conducted sampling across 5 independent random seeds, generating 10 valid compound samples per representative molecule, yielding 80 generated molecules per MoA class in total.

Metrics. Let $\hat{m} \in \mathcal{M}$ denote a generated molecule with target MoA $C^* \in \mathcal{Y} = \{C_1, \ldots, C_7\}$, and let $\mathcal{R}_C \subset \mathcal{M}$ collect the reference compounds annotated with MoA C ($|\mathcal{R}_C|$=8 in our balanced benchmark). All three metrics below share a representation-specific pairwise similarity $s(\cdot, \cdot) \in [0, 1]$ between two molecules: for the ECFP space, s is Tanimoto similarity over Morgan fingerprints; for the learned spaces (CLOOME, InfoAlign), s is cosine similarity over the corresponding embeddings of the two molecules.

- **Top-1 Cluster Accuracy** scores each class by measuring the mean similarity between the generated molecule and all individual members of reference class in evaluation representation,
$$\sigma(\hat{m}, C) \;=\; \frac{1}{|\mathcal{R}_C|} \sum_{r \in \mathcal{R}_C} s(\hat{m},\, r)\,,$$
and reports the fraction of generations for which the target attains the highest score, i.e. $C^* = \arg\max_{C \in \mathcal{Y}} \sigma(\hat{m}, C)$.
- **MoA Retrieval Rate @k** relaxes Top-1 to admit the target anywhere in the top k ranked classes, and replaces the centroid-based score above with a top-n mean of pairwise similarities that focuses on each class's most similar references:
$$\rho(\hat{m}, C) \;=\; \frac{1}{n} \sum_{r \in \mathrm{Top}_n(\hat{m},\, \mathcal{R}_C)} s(\hat{m}, r), \qquad n = 3.$$
Sorting classes by $\rho(\hat{m}, \cdot)$ in descending order yields $\mathrm{rank}(C^*)$; the metric is the fraction of generations with $\mathrm{rank}(C^*) \leq k$.
- **kNN Accuracy @k** discards the per-class structure and runs a k-nearest-neighbour classifier over the pooled reference set $\mathcal{R} = \bigcup_{C \in \mathcal{Y}} \mathcal{R}_C$. It takes the k references most similar to $\hat{m}$ under s and assigns the majority MoA among them as the predicted label; the metric is the fraction of generations whose prediction equals C^*.

Extended Results. Table 11 demonstrates a contrast in performance across the three representation spaces. Both DGVAE and CPMolGAN achieve results near a random-guess baseline (∼14.3% Top-1 Cluster Accuracy) regardless of the embedding used. Although GFlowNet performs well within CLOOME, its performance drops when evaluated in InfoAlign and ECFP spaces. PhAME is the *only* method that clears chance in all three spaces simultaneously, and the gap is largest exactly where the biological signal is richest: $+68\%$ relative over GFlowNet on InfoAlign Top-1 Cluster

Table 11: **MoA classification of de novo generated molecules across three representation spaces.** Values are percentages (mean$_{\pm\text{std}}$ over 5 seeds). Best in bold, second-best underlined. PhAME achieves the best performance in **20 out of the 21** evaluated combinations of metrics and representation spaces.

		Top-1	MoA Retrieval Rate @k ↑			kNN Accuracy @k ↑		
	Model	**Cluster Acc ↑**	**@1**	**@2**	**@3**	**@1**	**@3**	**@5**
CLOOME	DGVAE	14.6 ±1.2	13.7 ±1.8	29.1 ±2.0	42.9 ±2.5	13.6 ±2.0	14.0 ±2.7	14.4 ±1.5
	CPMolGAN	15.4 ±1.5	14.8 ±1.4	28.8 ±1.8	42.4 ±2.2	14.9 ±1.3	15.2 ±1.3	15.2 ±1.8
	GFlowNet	**21.2 ±1.2**	**19.4 ±2.5**	34.4 ±2.1	51.5 ±2.2	18.0 ±1.1	18.6 ±1.1	19.3 ±1.2
	PhAME	17.8 ±1.1	**19.4 ±0.5**	**38.2 ±0.9**	**53.6 ±2.1**	**20.5 ±1.2**	**20.4 ±1.0**	**20.4 ±0.7**
InfoAlign	DGVAE	13.4 ±0.5	14.4 ±1.0	28.6 ±1.9	43.2 ±3.2	14.0 ±1.7	14.3 ±1.8	14.2 ±1.3
	CPMolGAN	13.1 ±0.9	13.5 ±1.2	26.9 ±0.7	39.6 ±1.0	13.5 ±1.2	13.1 ±1.1	13.6 ±0.8
	GFlowNet	17.1 ±0.7	16.3 ±1.3	32.4 ±0.8	50.1 ±1.7	14.6 ±0.5	15.7 ±0.6	15.9 ±1.3
	PhAME	**28.7 ±1.7**	**28.7 ±2.7**	**45.2 ±1.8**	**58.0 ±1.0**	**25.4 ±2.2**	**27.0 ±1.8**	**28.3 ±1.4**
ECFP	DGVAE	14.0 ±0.9	14.2 ±2.0	28.5 ±1.1	42.7 ±2.1	14.5 ±0.6	14.6 ±1.0	13.7 ±1.6
	CPMolGAN	14.3 ±1.0	13.4 ±1.3	28.8 ±2.4	42.0 ±0.9	12.9 ±1.2	13.5 ±1.4	14.1 ±1.4
	GFlowNet	17.0 ±1.8	15.1 ±1.3	29.0 ±1.6	46.1 ±2.1	12.7 ±1.1	14.4 ±1.3	15.3 ±1.6
	PhAME	**27.2 ±1.8**	**23.0 ±1.6**	**40.4 ±2.3**	**56.2 ±2.9**	**20.6 ±1.2**	**21.2 ±1.6**	**23.0 ±2.2**

Table 12: **Evaluation of quality and diversity generated molecules for Cell Phenotype task.** Average metrics for cell-phenotype-guided generation, evaluated across five distinct seed molecules.

Method	QED ↑	SA ↓	Uniqueness ↑	Novelty ↑
DGVAE	**0.797 ±0.001**	**2.515 ±0.014**	**1.000 ±0.000**	**1.000 ±0.001**
CPMolGAN	0.607 ±0.010	3.042 ±0.037	**1.000 ±0.000**	**1.000 ±0.000**
GFlowNet	0.649 ±0.005	4.457 ±0.024	0.999 ±0.002	**1.000 ±0.000**
PhAME	0.674 ±0.003	3.233 ±0.027	0.960 ±0.007	**1.000 ±0.001**

Accuracy (0.287 vs. 0.171) and $+60\%$ on ECFP (0.272 vs. 0.170), with similar gains reflected in MoA Retrieval Rate @k and kNN Accuracy @k. On CLOOME, PhAME trails GFlowNet on Top-1 (0.178 vs. 0.212); PhAME's lead reappears once we move to deeper metrics (MoA Retrieval Rate @1/2/3, kNN Accuracy @1/3/5). Because the InfoAlign and ECFP advantages hold across two structurally unrelated representations, the improvements reflect a true understanding of structure-activity relationships rather than merely adapting to a single representational topology.

For completeness, Table 12 reports extended statistics on molecular quality and diversity metrics (QED, SA, Uniqueness and Novelty). These results should be contextualized with the primary Table 11. While baselines like DGVAE and CPMolGAN achieved slightly better QED or SA, they do not perform effectively on the actual MoA retrieval task. Importantly, PhAME still outperforms its primary competitor, GFlowNet, in both QED and SA.

Guidance scale sweep. We analyze the influence of the phenotype (w_c) and structural guidance (w_a) scales on Cell Phenotype Guided Generation benchmark. All configurations follow the evaluation protocol described previously. We denote a configuration by c-CFG(w_c, w_a). First, we fix the structural guidance at $w_a = 0$ and sweep the phenotype guidance scale w_c. As shown in Table 13, the configuration c-CFG$(34, 0)$ achieves the best results, with the metrics consistently improving as w_c increases from 0 to 34. At $w_c = 40$, we observe a degradation in the results. This improved retrieval performance is accompanied by reduction in validity from 79.5% to 70.3%, showing the expected trade-off between the strength of the phenotype signal and the fraction of valid generated molecules. Next, we fix the phenotype guidance at $w_c = 34$ and sweep structural guidance scale w_a. Table 14 shows that increasing structural guidance degrade results almost on all metrics. This behaviour is expected because the anchor in this *de novo* benchmark is randomly sampled and is not related to the target phenotype. Therefore, we use $w_a = 0$ and $w_c = 34$ for Cell Phenotype Guided Generation.

Table 13: **Sweep over the phenotype conditioning guidance scale w_c.** Evaluation on ECFP embedding space and validity metric with structural guidance scale fixed at $w_a = 0$. Values are percentages (mean$_{\pm\text{std}}$ over 5 seeds). Best in bold, second-best underlined.

	Top-1	MoA Retrieval Rate @k ↑			kNN Accuracy @k ↑			Validity ↑
Configuration	**Cluster Acc ↑**	**@1**	**@2**	**@3**	**@1**	**@3**	**@5**	
c-CFG(0,0)	14.5 ±2.0	14.8 ±1.6	29.9 ±1.6	44.1 ±1.6	14.9 ±1.2	14.4 ±1.5	13.9 ±1.8	**79.5 ±1.8**
c-CFG(4,0)	18.7 ±1.8	17.1 ±1.5	32.3 ±2.2	47.5 ±3.2	16.8 ±2.1	17.0 ±1.1	16.7 ±1.6	76.3 ±2.7
c-CFG(8,0)	19.6 ±1.9	17.4 ±1.8	33.6 ±1.4	49.5 ±1.9	17.1 ±1.6	17.9 ±1.3	18.1 ±2.0	74.3 ±1.9
c-CFG(12,0)	21.2 ±2.0	18.9 ±1.6	36.6 ±1.2	52.7 ±1.1	17.0 ±0.8	18.0 ±1.6	18.6 ±1.0	72.7 ±3.4
c-CFG(20,0)	24.8 ±2.5	21.5 ±2.2	38.2 ±1.7	54.0 ±1.7	19.4 ±1.3	20.0 ±1.8	20.9 ±2.0	71.4 ±2.3
c-CFG(28,0)	25.9 ±1.8	22.2 ±2.6	39.3 ±3.0	54.8 ±2.0	19.4 ±1.6	21.0 ±2.2	22.1 ±3.0	70.0 ±1.6
c-CFG(34,0)	**27.2 ±1.8**	**23.0 ±1.6**	**40.4 ±2.3**	**56.2 ±2.9**	**20.6 ±1.2**	**21.2 ±1.6**	**23.0 ±2.2**	70.3 ±2.1
c-CFG(40,0)	25.2 ±0.7	20.7 ±0.7	**40.4 ±1.7**	55.2 ±1.6	18.4 ±0.9	18.3 ±1.2	20.2 ±1.3	71.3 ±2.9

Table 14: **Sweep over the structural-guidance scale w_a.** Evaluation on ECFP embedding space and validity metric with phenotype conditioning scale fixed at $w_c = 34$. Values are percentages (mean$_{\pm\text{std}}$ over 5 seeds). Best in bold, second-best underlined.

	Top-1	MoA Retrieval Rate @k ↑			kNN Accuracy @k ↑			Validity ↑
Configuration	**Cluster Acc ↑**	**@1**	**@2**	**@3**	**@1**	**@3**	**@5**	
c-CFG(34,0)	**27.2 ±1.8**	**23.0 ±1.6**	**40.4 ±2.3**	56.2 ±2.9	**20.6 ±1.2**	**21.2 ±1.6**	**23.0 ±2.2**	70.3 ±2.1
c-CFG(34,1)	26.8 ±2.9	22.0 ±2.0	39.8 ±2.9	56.2 ±2.6	20.4 ±2.0	20.3 ±1.4	21.0 ±1.3	70.4 ±2.1
c-CFG(34,2)	26.7 ±2.8	21.9 ±1.7	39.6 ±2.6	**56.3 ±2.5**	20.4 ±1.9	20.4 ±1.5	21.0 ±1.2	**70.5 ±2.1**

Qualitative MoA guided molecules. Figure 12 shows selected molecules conditioned by cell morphology described in Section 5.4. To analyze the structural relevance, we visualize the nearest generated molecules based on Tanimoto similarity to any of the molecules from each MoA.

G Details of Cell Phenotype Optimization

Setup and target selection. To evaluate the capacity of H2L optimization, we formulated a recovery task of drugs. We selected three different molecules with known MoAs. The selected target therapeutics and their respective MoAs are as follows:

- **Aspirin:** *Cyclooxygenase inhibitor*
- **Clofibrate:** *PPAR receptor agonist*
- **Dyphylline:** *Adenosine receptor antagonist*

For each target therapeutic, we selected a set of structurally similar "hit" molecules to serve as starting seeds in order to evaluate the model's performance across varying levels of optimization difficulty. Crucially, to strictly prevent data leakage, the target molecule was entirely removed from the reference dataset prior to inference. Specifically, when constructing the conditioning signal, the target MoA centroid was computed exclusively from the phenotypic representations of the remaining class members, ensuring the model had no direct access to the target drug's biological signature.

Evaluation. PhAME was evaluated using the configuration from Appendix F, with guidance scales specifically tuned for structural editing $w_a = 1$ and $w_c = 14$. Unlike *de novo* generation, which starts from pure Gaussian noise, we start the process from a noised seed molecule at the corresponding timesteps $T = 400$ and $T = 500$. We found that each seed has a different optimal initial noise level: the level of noise depends on the structural similarity between lead and hit. For instance, among our visualized examples, only one seed molecule (the aspirin hit presented in row 2, column 2) required a significantly higher initial noise level ($T = 700$) to successfully bridge the structural gap. For this evaluation, we generated a total of 100,000 candidate samples. On the same setup, we conducted

tests with the GFlowNet model. For the seed molecules presented in Figure 7 conditioned with the proper morphology, we could not recover any of the drugs.

To illustrate PhAME's structural editing capabilities, we present extended qualitative visualizations in Figures 13, 14, 15. We provide comprehensive trajectories for each of the three targeted therapeutics: aspirin, clofibrate, and dyphylline. For each target lead, we display successful recoveries starting from six distinct hit molecules. These examples have varying complexity in the required transformation, ranging from minimal, straightforward functional-group edits to more advanced ones.

H Ablations

H.1 Morphology Representation.

To evaluate the robustness of our model to different conditioning representations from different biological manifolds, we conducted an ablation study comparing various cellular representations. Specifically, we selected three different types of feature representations extracted from the CP-JUMP dataset. Our primary cell morphology signal, denoted *CellProfiler*, corresponds to the exact same CellProfiler features utilized by InfoAlign. We compared it against two alternative representations: *DeepProfiler* (Moshkov et al., 2022) features extracted from a CNN-based model, retrieved from the official CP-JUMP repository and *uniDino* (Morelli et al., 2025) representations extracted from the uniDino model vision transformer. Crucially, to ensure a fair comparison across these three modalities, we restricted this study to a single experimental source (*Source 2* from CP-JUMP), which reduces the dataset size by approximately 50% compared to our main experiments.

Figure 10 highlights the advantage of our chosen representation. It shows that *CellProfiler* achieves the highest overall performance on the Top-1 Cluster Accuracy and MoA Retrieval Rate@k metrics, averaged across all evaluation spaces (CLOOME, InfoAlign, and ECFP).

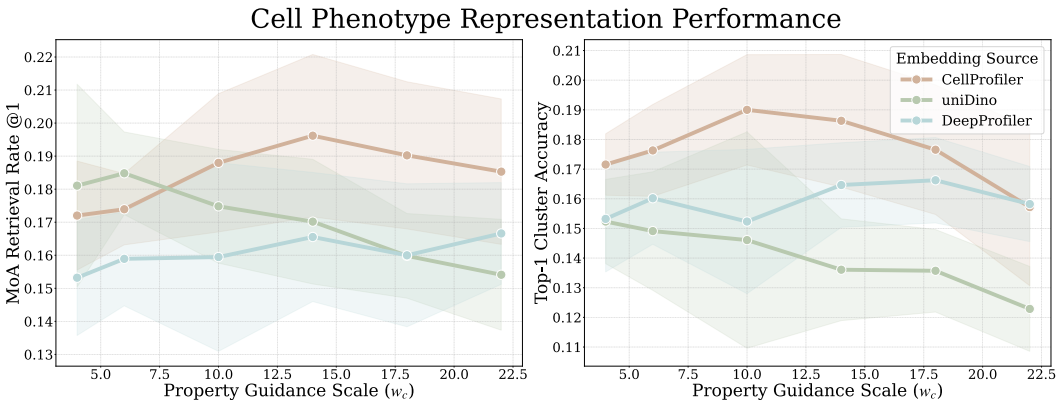


Figure 10: **Performance of Phenotypic Feature Extractors**. Comparison of cell phenotypic representations across different feature extractors: CellProfiler, UniDino and DeepProfiler. Performance is evaluated using the Top-1 Cluster Accuracy and MoA Retrieval Rate, metrics described in Appendix F. The CellProfiler representation achieves the best results across its head-to-head competitors as a signal for conditioning (w_c).

H.2 Structural Alignment Representation

To validate our choice of structural alignment representation z_{align} in Equations 5, 6, we evaluated PhAME on three distinct molecular representation embedding spaces: Morgan Fingerprint (ECFP), DGVAE latent space and InfoAlign embeddings in the setup of logP optimization described in Appendix C, with limited dataset to 20% of original size. As shown in Figure 11, InfoAlign slightly decreases similarity alignment to achieve superior property-target alignment. We accept a minor, negligible trade-off in structural similarity because InfoAlign offers a much richer representation.

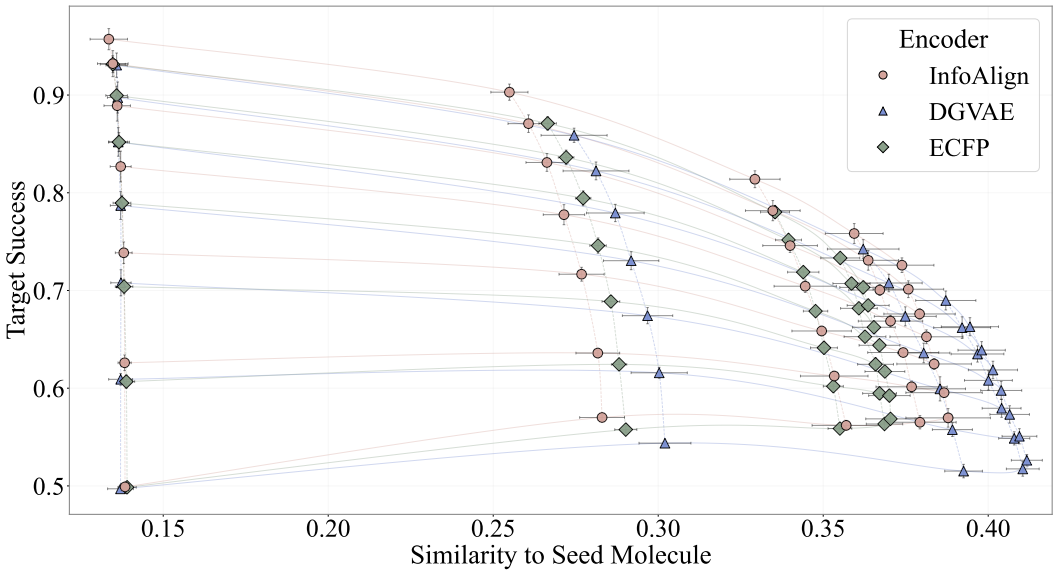


Figure 11: **Impact of Representations on Controllable Generation.** Trade-off between Target Success and Similarity to Seed Molecule under varying guidance strengths for different type encoded representation for conditioning w_a. Increasing property guidance (w_c) improves target success, while increasing seed guidance (w_a) enhances similarity to the input molecule. InfoAlign achieves the highest Target Success, showing only a marginal decrease in Similarity to Seed Molecule relative to DGVAE, which excels in seed similarity but struggles with Target Success. In contrast, ECFP shows weak performance across both criteria, further highlighting the advantage of InfoAlign. The layout and interpretation of this plot are identical to those of Figure3

Table 15: **Effect of pair construction on logP optimization.** Values are reported as mean ± std. Best values are shown in bold and second-best values are underlined.

Pair construction	Val ↑	Tgt ↑	Sim ↑	NTS ↑
MMP	0.65 ±0.06	0.76 ±0.01	**0.39 ±0.00**	<u>0.29 ±0.00</u>
Random	**0.74 ±0.06**	**0.96 ±0.01**	<u>0.12 ±0.00</u>	0.11 ±0.00
NN	<u>0.71 ±0.03</u>	<u>0.78 ±0.03</u>	**0.39 ±0.02**	**0.30 ±0.01**

H.3 Pair Construction

To assess the effect of training-pair construction, we compare three pairing strategies. Our approach, denoted as the nearest-neighbor (NN) strategy, pairs each molecule with its most Tanimoto-similar molecule satisfying the similarity threshold defined in Section 4.3. As an ablation, we consider random threshold pairing, where each molecule is instead paired with a randomly selected molecule satisfying the same target threshold, allowing us to isolate the effect of nearest-neighbor selection. Finally, we evaluate matched molecular pairs (MMPs), which capture chemically intuitive substitution patterns by constructing pairs using the MMP algorithm Hussain & Rea (2010), subject to the same threshold criterion.

logP optimization. Table 15 shows the results for logP optimization. MMP strategy does not improve upon NN pairing. Random pairing achieves the highest target success and validity, but substantially reduces similarity to the source molecule and the overall NTS score. This indicates that the property threshold drives target optimization, whereas selecting structurally related pairs is important for preserving the input molecule. Overall, NN pairing provides the best balance between target success and structural similarity.

Table 16: **Effect of pair construction on Cell Phenotype Guided Generation benchmark.** Values are percentages (mean ± std). Best values are shown in bold and second-best values are underlined.

Pair construction	Cluster Acc ↑	MoA Retrieval Rate @1 ↑	kNN Accuracy @1 ↑
MMP	16.5 ±1.1	14.5 ±1.2	15.3 ±0.5
Random	<u>18.1 ±1.3</u>	<u>16.9 ±1.8</u>	<u>16.0 ±2.1</u>
NN	**27.2 ±1.8**	**23.0 ±1.6**	**20.6 ±1.2**

Cell Phenotype guided generation. The effect of pair construction is more evident in the Cell Phenotype Guided Generation benchmark. As shown in Table 16, NN pairing outperforms both MMP and random pairing across all evaluated metrics. Although MMP produces structurally close pairs, only a limited number of these pairs satisfy the required morphological-divergence criterion. Constructing a sufficiently large MMP training set to satisfy budget used in NN strategy, therefore required relaxing the phenotype thresholds from $\tau_{\text{cos}} = (-0.6, -0.3)$ to $\tau_{\text{cos}} = (-0.3, 0.2)$. Consequently, the resulting pairs were more similar in phenotype, weakening the supervision signal.

Overall, Tanimoto-based NN pairing provides the best balance between structural consistency, supervision quality and dataset size. We therefore use NN pairing in all experiments.

H.4 SINGLE CFG

To evaluate the benefit of introduced c-CFG (Compositional Classifier-Free Guidance) over standard Classifier-Free Guidance, we compare the two techniques across a broad range of guidance scales. Standard CFG is described in Section 3 (paragraph *Classifier-free guidance*), and c-CFG in Section 4 (paragraph *Compositional guidance*). We denote a c-CFG configuration by $\text{c-CFG}(w_c, w_a)$, where w_c is the guidance scale of target signature and w_a is the structural-guidance scale. Standard CFG is denoted by $\text{CFG}(w)$, with a single guidance scale.

Table 17: **Comparison of standard CFG and compositional CFG for logP optimization.** Values are reported as mean ± std. Best values are shown in bold and second-best values are underlined.

Model	Val ↑	QED ↑	SA ↓	Tgt ↑	Sim ↑	NTS ↑
CFG(0)	<u>0.84 ± 0.01</u>	**0.83 ± 0.00**	3.44 ± 0.05	0.50 ± 0.00	0.13 ± 0.00	0.06 ± 0.00
CFG(1)	**0.89 ± 0.01**	<u>0.79 ± 0.00</u>	<u>2.94 ± 0.03</u>	0.65 ± 0.00	0.31 ± 0.01	0.20 ± 0.00
CFG(3)	0.42 ± 0.02	0.70 ± 0.01	**2.80 ± 0.06**	<u>0.68 ± 0.01</u>	**0.40 ± 0.00**	<u>0.27 ± 0.01</u>
CFG(5)	0.13 ± 0.03	0.17 ± 0.04	4.33 ± 0.16	0.65 ± 0.01	0.19 ± 0.02	0.13 ± 0.01
c-CFG(6,3)	0.71 ± 0.03	0.74 ± 0.01	2.96 ± 0.05	**0.78 ± 0.03**	<u>0.39 ± 0.02</u>	**0.30 ± 0.01**

logP optimization. Table 17 presents a comparison of standard CFG with c-CFG for logP optimization. Increasing the standard CFG scale improves target optimization only at the cost of generation quality. Validity decreases from 89% at CFG(1) to 42% at CFG(3) and 13% at CFG(5). In comparison, c-CFG(6,3) retains 71% validity while achieving the highest target success and NTS. None of the evaluated standard CFG scales provides the same flexibility and balance between target optimization and structural similarity.

Cell Phenotype guided generation. For the Cell Painting phenotype task, we observe a similar trend in Table 18. Standard CFG achieves its best results at CFG(8), while further increasing the scale degrades the evaluated metrics. The c-CFG(34,0) configuration was selected based on the experiments reported in the *Guidance scale sweep* paragraph of Appendix G, with the results presented in Tables 13 and 14. This configuration outperforms CFG(8) across all evaluated metrics, increasing Cluster Accuracy from 0.190 to 0.272, MoA Retrieval Rate @3 from 0.489 to 0.562, and kNN Accuracy @5 from 0.170 to 0.230. Standard CFG cannot represent this configuration because it applies one scale to the joint conditioning signal, whereas $\text{c-CFG}(34, 0)$ applies strong phenotype guidance while disabling seed guidance.

Table 18: **Comparison of standard CFG and compositional CFG for Cell Painting-guided generation.** Values are percentages (mean ± std). Best values are shown in bold and second-best values are underlined.

Model	Cluster Acc ↑	MoA @1 ↑	MoA @3 ↑	kNN @1 ↑	kNN @5 ↑
CFG(2)	16.8 ± 1.9	17.2 ± 2.0	44.8 ± 1.8	16.1 ± 2.4	16.2 ± 1.6
CFG(8)	<u>19.0 ± 0.9</u>	<u>17.7 ± 0.6</u>	<u>48.9 ± 0.9</u>	<u>16.6 ± 0.2</u>	<u>17.0 ± 0.8</u>
CFG(20)	15.8 ± 1.6	13.6 ± 1.3	47.3 ± 3.2	13.0 ± 0.7	14.7 ± 1.9
CFG(28)	13.6 ± 1.8	11.2 ± 2.0	46.6 ± 2.2	10.4 ± 2.2	11.7 ± 2.9
c-CFG(34,0)	**27.2 ± 1.8**	**23.0 ± 1.6**	**56.2 ± 2.9**	**20.6 ± 1.2**	**23.0 ± 2.2**

These results demonstrate that independently controlling the target and seed conditions provides a better trade-off between target optimization, structural similarity, and molecular quality than standard single-scale CFG.

H.5 Effect of a non-additive interaction term

To test whether interactions between the target and structural conditions provide additional guidance, we augment the additive c-CFG formulation with an explicit interaction term:

$$\widetilde{\epsilon}_\theta = \epsilon_\theta(\varnothing, \varnothing) + w_c\Delta_c + w_a\Delta_a + w_{ca}\Delta_{ca}, \tag{7}$$

where

$$\Delta_{ca} = \epsilon_\theta(c, z_{\text{align}}) - \left[\epsilon_\theta(c, \varnothing) + \epsilon_\theta(\varnothing, z_{\text{align}}) - \epsilon_\theta(\varnothing, \varnothing)\right]. \tag{8}$$

Here, Δ_{ca} is the residual joint prediction that cannot be explained by adding the separate target and structural guidance directions. The original additive formulation is recovered when $w_{ca} = 0$.

We sweep w_{ca} for logP optimization using the main configuration $(w_c, w_a) = (6, 3)$. As shown in Table 19, introducing the interaction term with $w_{ca} = 1$ produces no measurable improvement over the additive baseline. Results up to $w_{ca} = 3$ remain within run-to-run variation, with NTS unchanged at 0.30. Larger interaction weights slightly increase similarity, from 0.39 at $w_{ca} = 0$ to 0.41 at $w_{ca} = 10$, but reduce target success from 0.78 to 0.67 and consequently lower NTS from 0.30 to 0.27.

Table 19: **Effect of the non-additive interaction weight w_{ca} on logP optimization at $(w_c, w_a) = (6, 3)$.** Cells report mean ± std over five runs.

w_{ca}	Val ↑	QED ↑	SA ↓	Tgt ↑	Sim ↑	NTS ↑
0	0.71 ± 0.03	0.74 ± 0.01	2.96 ± 0.05	**0.78 ± 0.03**	0.39 ± 0.02	**0.30 ± 0.01**
1	0.71 ± 0.03	0.74 ± 0.01	2.96 ± 0.05	**0.78 ± 0.03**	0.39 ± 0.02	**0.30 ± 0.01**
3	<u>0.72 ± 0.03</u>	0.74 ± 0.01	2.95 ± 0.04	<u>0.76 ± 0.03</u>	<u>0.40 ± 0.01</u>	**0.30 ± 0.01**
6	<u>0.72 ± 0.03</u>	**0.75 ± 0.01**	<u>2.93 ± 0.04</u>	0.72 ± 0.02	<u>0.40 ± 0.01</u>	<u>0.29 ± 0.01</u>
10	**0.73 ± 0.04**	**0.75 ± 0.01**	**2.92 ± 0.03**	0.67 ± 0.01	**0.41 ± 0.01**	0.27 ± 0.01

The interaction term therefore changes the target–similarity trade-off but does not improve the combined objective. We consequently retain the simpler additive formulation, which avoids introducing an additional guidance hyperparameter.

I Limitations

While PhAME mitigates the trade-offs in standard classifier-free guidance, the framework remains sensitive to extreme guidance scales, where overly strong structural guidance can degrade the chemical validity of the generated molecules. Additionally, the model's performance relies on the

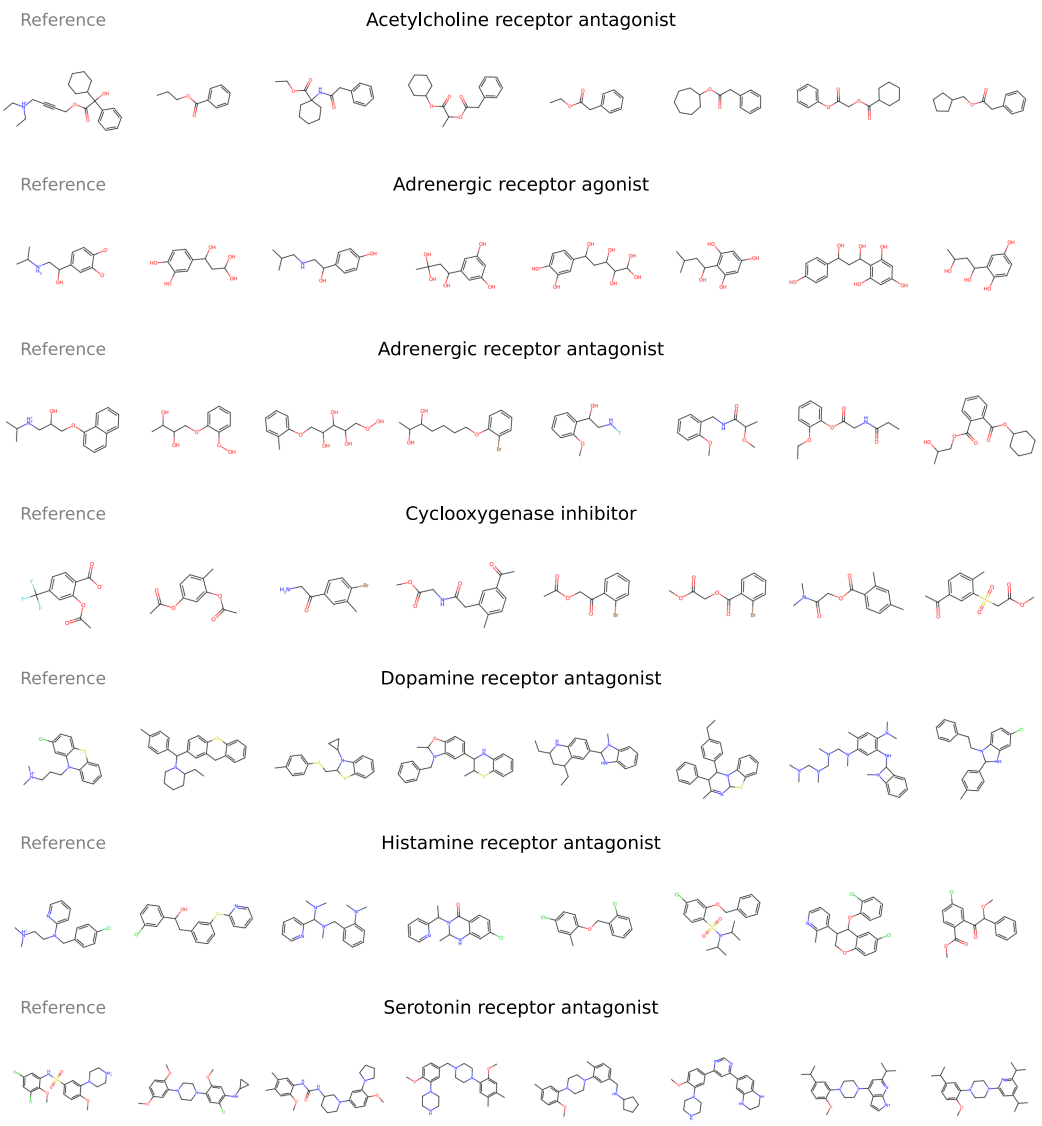


Figure 12: **Qualitative Assessment of PhAME-Generated molecules.** Qualitative examples of molecules generated by PhAME, conditioned by MoA embedding centroids. Each row presents a different Mode of Action. The first column captioned as Reference represents one of the selected molecules from our dataset with a known MoA, while the remaining columns visualize the nearest generated molecules based on Tanimoto similarity to any molecule in the MoA cluster. Each class contains specific, distinct molecular structures.

availability and quality of high-dimensional paired data, such as Cell Painting morphologies and L1000 transcriptomic profiles.

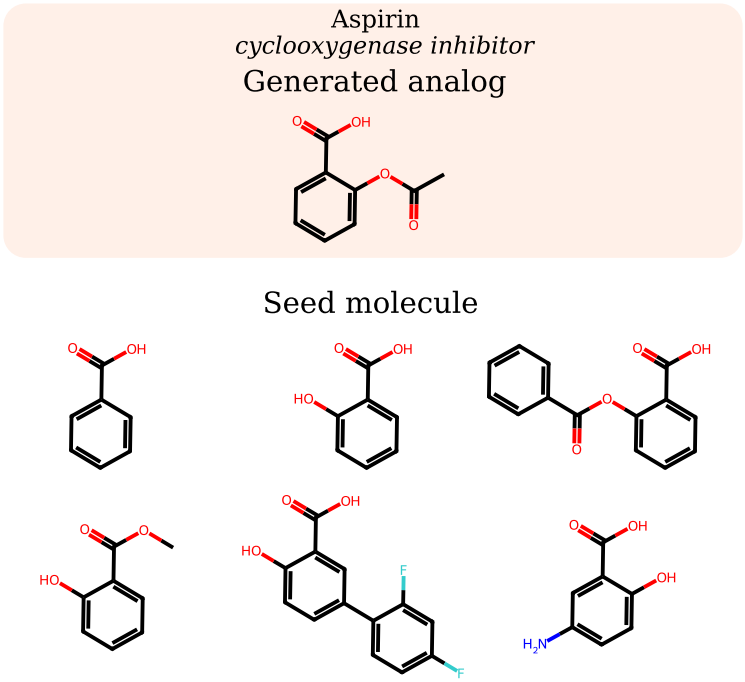


Figure 13: **Recovery of Aspirin.** Qualitative examples of PhAME recovering Aspirin Lead (Generated Analog). Six distinct Hits (Seed molecule) were evaluated using structural anchor and biological conditioning of cyclooxygenase inhibitor.

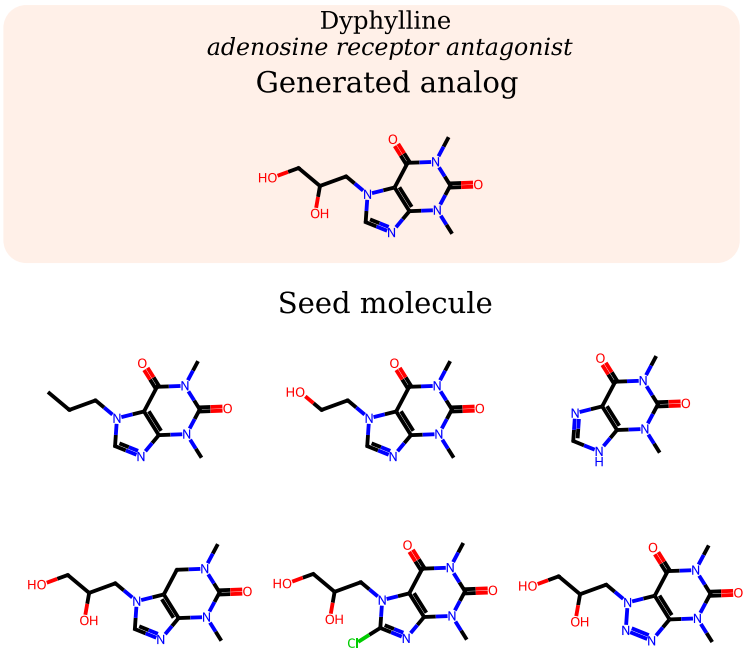


Figure 14: **Recovery of Dyphylline.** Qualitative examples of PhAME recovering Dyphylline Lead (Generated Analog). Six distinct Hits (Seed molecule) were evaluated using structural anchor and biological conditioning of adenosine receptor antagonist.

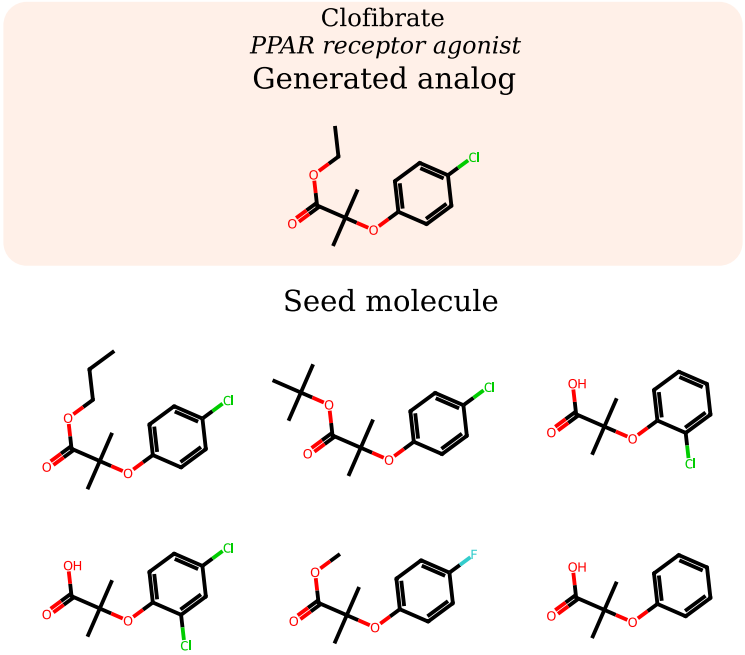


Figure 15: **Recovery of Clofibrate.** Qualitative examples of PhAME recovering Clofibrate Lead (Generated Analog). Six distinct Hits (Seed molecule) were evaluated using structural anchor and biological conditioning of PPAR receptor agonist.