GXEVO: EVOLUTIONARY SEARCH FOR TRANSCRIPTOME-CONDITIONED MOLECULAR DESIGN

Anonymous authors
Paper under double-blind review

ABSTRACT

Transcriptome-based drug design (TBDD) seeks molecules that induce a desired transcriptional response. This is a one-to-many inverse problem: a single desired response can correspond to multiple molecular scaffolds, motivating the generation of structurally diverse, response-aligned candidates. Existing TBDD generators, however, struggle to jointly achieve high structural similarity to known actives and scaffold novelty, and their promising outputs do not directly guide subsequent proposals. We introduce GxEvo, a population-based evolutionary framework that uses candidate evaluations to guide further generation under a fixed transcriptional objective. GxEvo combines a condition-specific seed generator, a forward response model trained with response prediction and contrastive learning for candidate discrimination, and a learned mutation operator conditioned on both the parent molecule and the target condition. Across ten gene-knockdown targets and three held-out drug–cell splits, GxEvo achieves the highest mean structural similarity to reference compounds among the evaluated methods. On knockdown targets, it generates 23.8 structurally hit-like unseen scaffolds absent from the task corpus per 100 valid molecules, compared with 3.9 for the strongest prior method under its standard sampling protocol. The advantage persists under matched candidate budgets, while controlled substitutions of existing response predictors and variation operators demonstrate the benefits of the proposed forward model and mutation operator. These results establish evolutionary search as an effective computational approach to generating novel, hit-like molecular hypotheses for subsequent experimental validation. Our code is available at https://anonymous.4open.science/r/GxEvo.

1 INTRODUCTION

A central task in early drug discovery is to identify hit or lead compounds for subsequent optimization (Hughes et al., 2011; Macarron et al., 2011). Conventional screening addresses this task by searching predefined libraries, but these libraries cover only a minute fraction of drug-like chemical space (Reymond, 2015). To overcome this limitation, de novo molecular design seeks to construct new molecules tailored to a specified objective (Schneider & Fechner, 2005). In this context, modern generative models learn chemical priors from observed compounds and use them to concentrate molecular proposals in regions of chemical space that are more likely to satisfy the desired objective (Sanchez-Lengeling & Aspuru-Guzik, 2018).

Most goal-directed generators define the objective around a known protein target, typically through predicted binding or activity. While this formulation is effective when a disease-driving target is known and its modulation is expected to produce the desired phenotype, it is restrictive for phenotypes involving multiple or incompletely characterized pathways (Moffat et al., 2017; Swinney & Anthony, 2011). Transcriptome-based drug design (TBDD) provides an alternative by seeking molecules that induce a specified transcriptional response (Méndez-Lucio et al., 2020; Subramanian et al., 2017; Xu et al., 2026). Here, a transcriptional response refers to the change in gene expression induced by an intervention in a given cellular context.

TBDD is inherently a one-to-many inverse problem: a single desired transcriptional response can correspond to multiple molecular structures with different scaffolds. This non-uniqueness creates an opportunity to generate multiple structurally distinct candidates for the same response objective, allowing downstream selection to account for safety, physicochemical, and developability constraints (Waring et al., 2015). Generating such response-aligned scaffold hypotheses broadens the

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

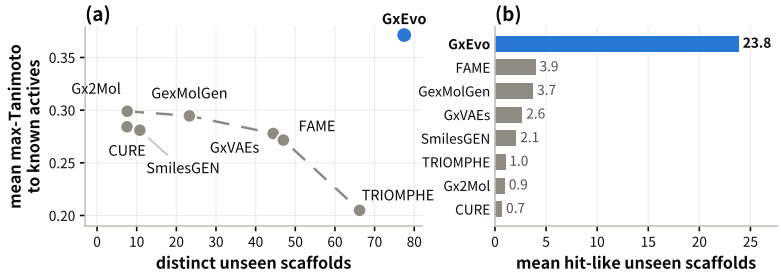

Figure 1: **Joint novelty and hit likeness on the knockdown benchmark.** Counts are per 100 valid molecules. **(a)** Mean maximum Tanimoto similarity on Morgan fingerprints to the target's known actives against distinct Bemis–Murcko scaffolds absent from the task corpus: the published generators lie on a trade-off frontier (dashed), while GxEvo improves both. **(b)** At the hit threshold $\tau = 0.4$, GxEvo produces 6.1× more hit-like unseen scaffolds than the best prior method.

set of starting points for subsequent validation, consistent with the practical motivation for scaffold hopping (Schneider et al., 1999; Brown & Jacoby, 2006; Sun et al., 2012; Callis et al., 2022).

An effective TBDD method should therefore produce candidates that are both relevant to the desired response and structurally diverse. Existing TBDD generators, however, struggle to achieve these goals simultaneously (Méndez-Lucio et al., 2020; Kaitoh & Yamanishi, 2021; Pham et al., 2022; Li & Yamanishi, 2024; 2025; Cheng et al., 2024; Liu et al., 2025; Xu et al., 2026). Figure 1 illustrates this limitation using structural similarity to known actives as a proxy for hit-likeness and the number of distinct scaffolds absent from the task corpus as a measure of scaffold novelty. We observed that methods that generate more unseen scaffolds tend to exhibit lower similarity to known actives, whereas methods with higher similarity yield fewer structurally novel candidates. Consequently, none of the evaluated prior methods provides a candidate set that is strong on both dimensions, limiting the breadth of hit-like structural hypotheses returned for a single transcriptional objective.

Overcoming this trade-off requires improving hit-likeness while preserving broad scaffold coverage. At inference time, however, the TBDD generators considered here do not maintain and update an explicit candidate population, so promising outputs do not directly guide subsequent proposals for the same target response. Using evaluated candidates as starting points for further generation could enable progressive refinement of promising structures while supporting exploration from multiple structural starting points. This motivates an iterative search in which candidate evaluation informs subsequent generation. We therefore formulate TBDD as population-based evolutionary search: selection favors candidates predicted to better match the target response, while variation generates new structural hypotheses from selected parents.

Existing molecular evolutionary algorithms (Jensen, 2019; Chandraghatgi et al., 2024; Wang et al., 2025), however, cannot be directly applied to this formulation without adapting how candidates are initialized, evaluated, and varied. First, initialization should incorporate the target transcriptional condition so that the search begins with candidates proposed specifically for the desired response. Second, the transcriptional response of an unmeasured candidate must be predicted to estimate its fitness. Existing perturbation-response predictors can provide these estimates, but their utility for evolutionary selection depends on how reliably they distinguish candidates by alignment with the target response. Low average error in predicting transcriptional responses alone does not guarantee this discrimination, motivating a forward model trained for both response prediction and discrimination between molecule-specific responses. Third, variation should account for both the selected parent's structure and the target transcriptional condition, whereas generic variation operators do not explicitly exploit the latter.

To meet these requirements, we introduce **GxEvo**, an evolutionary TBDD framework comprising three components. The **seed generator** initializes a candidate population conditioned on the target response and cellular context. The **forward response model** combines response prediction with contrastive learning to preserve molecule-specific response distinctions, providing a discriminative fitness signal for evolutionary selection. Fitness is computed as the cosine similarity between a candidate's predicted response and the target response. The **learned mutation operator** uses a selected parent together with the target condition to propose new molecular structures. Under a fixed target condition, GxEvo repeatedly selects parents, generates and scores offspring, and retains the highest-fitness candidates from the current population and newly generated offspring.

The main contributions of this work are summarized as follows:

- We introduce **GxEvo**, to our knowledge the first framework to formulate TBDD as population-based evolutionary optimization, enabling evaluated candidates to guide subsequent generation under a fixed transcriptional objective.
- We develop a candidate-discriminative forward response model for fitness evaluation and a learned mutation operator conditioned on both the parent molecule and the target condition. Controlled substitutions show that these components improve search performance over existing perturbation-response predictors and generic variation operators, respectively.
- Across gene-knockdown and held-out drug–cell benchmarks, GxEvo achieves the highest mean structural similarity to target-associated reference compounds among the evaluated methods and substantially increases novel hit scaffold yield on knockdown targets. On the knockdown benchmark, these gains persist under matched candidate budgets, demonstrating the benefit of evolutionary refinement beyond additional sampling.

2 Related work

Transcriptome-conditioned molecular generation. Existing TBDD methods primarily learn a direct mapping from a transcriptional condition to molecular structures. The baselines considered here span latent-variable models (Kaitoh & Yamanishi, 2021; Li & Yamanishi, 2024; Liu et al., 2025), fragment-based generation (Pham et al., 2022), cross-modal or sequence generation (Cheng et al., 2024; Li & Yamanishi, 2025), and perturbation-conditioned diffusion (Xu et al., 2026). Despite their architectural differences, these methods directly generate candidates from the condition, whereas GxEvo keeps its trained components fixed and explicitly evolves a candidate population for each target condition through selection, parent-conditioned variation, and survival.

Forward models for perturbation response. Forward response models estimate an unmeasured candidate's transcriptional perturbation from its structure and cellular context. Representative approaches include TranSiGen's dual-VAE latent mapping, PRnet's perturbation-conditioned profile generator, and XPert's context- and dose-aware transformer (Tong et al., 2024; Qi et al., 2024; Guo et al., 2026). Although these models primarily target predictive fidelity, evolutionary fitness additionally requires discriminative ranking of candidate molecules. We therefore train our forward model for both response prediction and drug discrimination, and evaluate its in-loop utility through controlled reward replacement.

Molecular evolutionary search. Evolutionary algorithms provide derivative-free optimization in discrete chemical spaces through graph-level operators (Jensen, 2019), fragment substitution under docking-based fitness (Chandraghatgi et al., 2024), and chemistry-aware language models driven by textual property objectives (Wang et al., 2025). These establish effective search for chemical-property or protein-level objectives, but their variation operators do not condition on a cell-specific transcriptional response. GxEvo instead couples evolutionary selection under a predicted-response reward with learned parent-aware variation within a transcriptome-conditioned search.

3 Method

Problem formulation and overview. Let $b, p \in \mathbb{R}^{n_g}$ denote the basal and perturbed expression profiles of a cell, respectively, where n_g is the number of genes considered. Let $\Delta = p - b$ denote the desired transcriptional response. Given a target condition $c = (b, \Delta)$, we seek a population of candidate molecules whose predicted responses align with Δ, with population-based search allowing multiple structural hypotheses to be explored.

GxEvo combines three components (Figure 2): a seed generator initializes condition-specific candidates, a forward response model assigns fitness to unmeasured molecules, and a learned mutation operator proposes children from selected parents. Evolution then repeatedly selects, mutates, scores, and retains candidates under the same target condition.

3.1 Population-level evolutionary search

Let $q_\theta(x \mid c)$ denote the conditional distribution defined by the seed generator, $M_\psi(x' \mid x, c)$ the conditional distribution defined by the mutation operator over children x' given parent x and target condition c, and $r_\phi(x; c)$ the fitness function computed using the forward response model. Let N

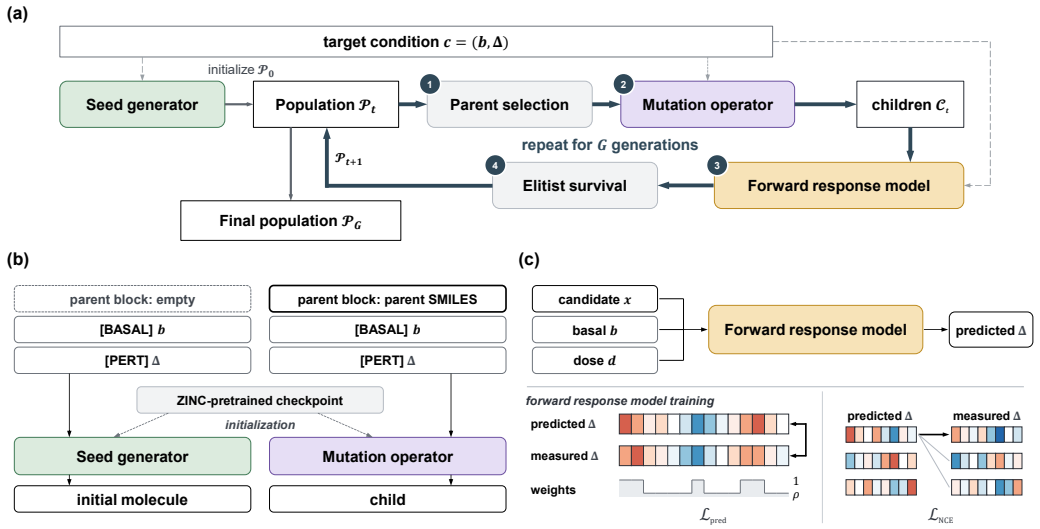

Figure 2: **Overview of GxEvo.** (a) A condition-specific population is iteratively updated through fitness-weighted parent selection, transcriptome-conditioned mutation, predicted-response scoring, and elitist survival. (b) The seed generator and mutation operator share one conditional Transformer, fine-tuned independently from a common ZINC-pretrained checkpoint. (c) The forward model's training objective and the search fitness, detailed in Section 3.3.

denote the population size, B the number of children proposed per generation, and G the number of generations. We denote by $\mathcal{P}_t$ the candidate population at the start of generation t. Given a target condition c, we initialize the search as follows:

Initialize. Sample N molecules from $q_\theta(\cdot \mid c)$, canonicalize them, and remove invalid outputs to form the initial population $\mathcal{P}_0$. Evaluate each candidate $x \in \mathcal{P}_0$ using the fitness function $r_\phi(x; c)$.

For each generation $t = 0, \ldots, G-1$, we then perform:

1. **Select.** Sample B parents from $\mathcal{P}_t$ with probability proportional to shifted nonnegative fitness.
2. **Mutate.** Generate one child for each selected parent from $M_\psi(\cdot \mid x, c)$.
3. **Score.** Canonicalize the generated children and remove invalid ones and any whose canonical SMILES was already scored in an earlier generation. Denote the remaining set of children by $\mathcal{C}_t$ and evaluate each candidate $x \in \mathcal{C}_t$ using the fitness function $r_\phi(x; c)$.
4. **Survive.** Pool the current population $\mathcal{P}_t$ and the newly generated children $\mathcal{C}_t$, and retain the N highest-fitness entries: $\mathcal{P}_{t+1} = \mathrm{Top}_{N, r_\phi}\big(\mathrm{pool}(\mathcal{P}_t, \mathcal{C}_t)\big)$.

The total proposal budget is therefore $N + GB$, and the final population $\mathcal{P}_G$ is returned. Unlike independent sampling, this search uses candidates retained in the population to guide subsequent proposals, allowing gains in fitness to accumulate across generations.

3.2 Conditional Molecular Proposals

The seed generator $q_\theta(x \mid c)$ and mutation operator $M_\psi(x' \mid x, c)$ serve complementary roles in the evolutionary loop: the seed generator initializes a condition-specific population, whereas the mutation operator proposes children from selected parents during evolution. As illustrated in Figure 2(b), both models use the same conditional Transformer architecture (Vaswani et al., 2017) but are fine-tuned independently with separate parameters. They share the following schematic token layout:

$$[\mathrm{BOS_E}]\ \ \mathrm{tok}(x_{\mathrm{parent}})\ [\mathrm{SEP}]\ [\mathrm{BASAL}]\ [\mathrm{PERT}]\ [\mathrm{SEP_{E2D}}] \parallel [\mathrm{BOS_D}]\ \ \mathrm{tok}(x_{\mathrm{out}})\ [\mathrm{EOS_D}], \quad (1)$$

where $\mathrm{tok}(\cdot)$ denotes SMILES tokenization. The subscripts E and D denote the conditioning prefix and the autoregressive output segment, respectively, and $\parallel$ denotes their concatenation. For the seed generator, the parent slot $\mathrm{tok}(x_{\mathrm{parent}})$ is empty and x_{out} denotes an initial candidate x. For the mutation operator, the parent slot contains the selected parent's SMILES tokens and x_{out} denotes a child x'. Learned projections of b and Δ provide the $[\mathrm{BASAL}]$ and $[\mathrm{PERT}]$ embeddings, respectively.

Attention is bidirectional within the conditioning prefix, while each output position attends to the entire prefix and preceding output tokens.

Seed generator. The seed generator defines $q_\theta(x \mid c)$ and constructs the initial population $\mathcal{P}_0$ from the target condition alone. We initialize it from a Transformer pretrained on unpaired ZINC molecules and fine-tune it on LINCS molecule–response pairs using teacher-forced next-token cross-entropy.

Mutation operator. The mutation operator $M_\psi(x' \mid x, c)$ is fine-tuned independently from the same ZINC-pretrained checkpoint on LINCS molecule pairs with Morgan Tanimoto similarity in $[0.4, 0.8]$, providing structurally related but non-identical parent–child pairs. Each pair provides a parent molecule x and a child molecule x', with the child's measured response and cellular context defining the target condition c; both pair directions are used. Both the seed generator and mutation operator resample randomized SMILES during training (Arús-Pous et al., 2019).

3.3 Forward response model

Existing molecular evolutionary algorithms typically optimize objectives that can be evaluated directly from each candidate structure using predefined oracles. In contrast, a transcriptional target has no readily available oracle and therefore requires a surrogate fitness function to be learned from data.

The forward response model predicts the transcriptional response $\hat{\Delta}_\phi(x \mid b, d) \in \mathbb{R}^{n_g}$ of molecule x under basal state b and dose d. A pretrained MolDeBERTa encoder (de Oliveira & Saeed, 2026) maps x to a molecular representation h, which is conditioned on b and d through FiLM (Perez et al., 2018):

$$\tilde{h} = (1 + \gamma(b, \tilde{d})) \odot h + \beta(b, \tilde{d}), \qquad \tilde{d} = \log_{10}(1 + d), \tag{2}$$

where γ and β are learned functions that produce feature-wise scaling offsets and shifts, respectively, and $\odot$ denotes elementwise multiplication. A prediction head maps the conditioned representation $\tilde{h}$ to the predicted response $\hat{\Delta}_\phi(x \mid b, d)$.

Because evolutionary selection depends on relative fitness, a model that captures the average transcriptional response while collapsing molecule-specific variation can achieve low error while giving poor rankings and weak selection pressure. To emphasize reliable differential responses, we assign unit weight to 100 genes selected from compound-level replicate statistics and a lower weight ρ to the remaining genes; Appendix A gives the exact definition. We then combine a soft-DEG-aware response prediction loss with an InfoNCE batch-discrimination loss (van den Oord et al., 2018):

$$\mathcal{L}_{\text{resp}} = \mathcal{L}_{\text{pred}} + \lambda_{\text{NCE}} \mathcal{L}_{\text{NCE}}. \tag{3}$$

The prediction loss combines weighted Huber regression and cosine agreement; full loss definitions and hyperparameters are provided in Appendix A.

At search time, the proposal models are conditioned on $c = (b, \Delta)$, while the forward model scores every candidate at a fixed dose d_{eval} (Appendix A) using unweighted full-profile agreement:

$$r_\phi(x; c) = \cos\Big(\hat{\Delta}_\phi(x \mid b, d_{\text{eval}}), \Delta\Big). \tag{4}$$

This fitness determines both parent-selection probabilities and elitist survival.

4 Experiments

We ask whether GxEvo improves target alignment and scaffold novelty on gene-knockdown targets (Section 4.2), whether the gain survives a matched candidate budget (Section 4.3), whether it extends to held-out drug–cell conditions (Section 4.4), and how the reward and variation operator each contribute (Section 4.5).

Data and preprocessing. We construct conditions from the Level-3 (per-well) LINCS L1000 data redistributed by PRnet (Subramanian et al., 2017), restricting to `trt_cp` wells with a parsable structure and following its preprocessing over the 978 landmark genes. A condition is a (drug, cell-line) pair whose perturbed profile is the mean over its positive-dose wells and whose basal profile b is the mean over the cell line's control wells; the response is Δ = perturbed $- b$, and per-gene standardized b and Δ are used for modeling. The knockdown-axis corpus holds 85,310 conditions, 12,313 compounds, and 82 cell lines. To prevent leakage we exclude evaluation compounds and connectivity-equivalent structures from all retrainable corpora; Appendix B gives the filters in full.

Evaluation settings. We evaluate in two complementary settings. *Knockdown* evaluates ten MCF7 gene-knockdown targets: AKT1, AKT2, AURKB, CTSK, EGFR, HDAC1, MTOR, PIK3CA, SMAD3, and TP53. GxEvo pairs each knockdown signature with the MCF7 basal profile, while baselines follow their published conditioning conventions (Appendix B). Each molecule is compared with the ExCAPE actives for the corresponding gene (1,364–23,979 ligands per gene). *Held-out conditions* reconstruct the split design described by CURE (Xu et al., 2026), rather than reuse its released split code. A condition is a (compound, cell line) pair, and the random, scaffold, and cell splits hold out unseen pairs, unseen Bemis–Murcko scaffold clusters, and an entire unseen tissue, respectively. Each retrainable model is trained on its split's training partition and evaluated on 200 held-out conditions per split, each carrying a single ground-truth compound. Partition sizes, the held-out tissue, and the condition selection are given in Appendix G.

Baselines. Generation baselines are TRIOMPHE (Kaitoh & Yamanishi, 2021), FAME (Pham et al., 2022), GxVAEs (Li & Yamanishi, 2024), GexMolGen (Cheng et al., 2024), Gx2Mol (Li & Yamanishi, 2025), SmilesGEN (Liu et al., 2025), and CURE (Xu et al., 2026). FAME is evaluated only in the knockdown setting, for the reason given in Appendix D; every other baseline is evaluated in both. Variation-operator controls include BRICS fragment replacement (Degen et al., 2008), Graph GA (Jensen, 2019), the FDSL-DD fragment GA (Chandraghatgi et al., 2024), MOLLEO (Wang et al., 2025), and independent resampling. The reward-model analysis compares with TranSiGen (Tong et al., 2024), PRnet (Qi et al., 2024), and XPert (Guo et al., 2026).

Implementation details. GxEvo initializes and retains 100 molecules and generates 70 children for 20 generations, a proposal budget of 1,500 candidates per target. Reported knockdown values are means over three generation draws per baseline and four search seeds per search arm; Appendix E gives the replicate count, the two arms that admit none, and the dispersion of each. Appendices A and C give architecture and training details and the wall-clock cost of that budget.

4.1 Evaluation metrics

Structural reference alignment. For each molecule generated from a knockdown signature, we compute its maximum similarity to the corresponding gene's ExCAPE active set. We report the **population mean** and, for comparison with prior TBDD work, the **population max**, the largest such per-molecule similarity in the returned population. For held-out drug–cell conditions, we report **mean Morgan similarity** and **mean Fraggle similarity** to the held-out compound, along with **PRnet-MSE** (Qi et al., 2024) for comparison with CURE. PRnet-MSE is the mean squared error between the PRnet-predicted response of a generated molecule and the measured target response, evaluated at each condition's median positive dose rather than at the search reward's fixed d_{eval} (Appendix B).

Novel hit scaffolds. We call a generated molecule *hit-like* if its maximum Morgan similarity to the target reference set is at least $\tau = 0.4$; Appendix I evaluates the same populations over $\tau \in [0.2, 0.6]$ and GxEvo's lead is unchanged throughout. We call its Bemis–Murcko scaffold (Bemis & Murcko, 1996) *novel* if that scaffold is absent from the method's response-paired task corpus. **Novel hit scaffolds (NHS)** is their intersection: the number of distinct novel scaffolds represented by hit-like molecules, normalized per 100 valid outputs. The reference corpus differs between methods; Appendix M recomputes the metric against one corpus shared by all of them. We report this metric for knockdown targets only, because a held-out condition supplies a single ground-truth compound rather than a family of actives, so the same threshold asks a stricter and qualitatively different question there; Appendix L gives those values.

Generation quality and task-corpus overlap. Validity is the fraction of sampled candidates that parse, counted over every candidate a method draws. **Within-condition uniqueness** is the fraction of valid molecules with distinct canonical SMILES, and **novelty** is the fraction of these distinct molecules absent from the task corpus. To test whether generation reproduces condition-associated chemistry, we compare each valid output with the method's own task-specific LINCS corpus. **Median nearest-neighbor Morgan similarity (NN-T)** measures overall proximity to that corpus. **Connectivity overlap** is the fraction of valid outputs whose standard-InChI connectivity layer appears in that corpus, determined by matching the first InChIKey block (Heller et al., 2015); Appendix O compares it with the alternatives. We also report **QED** (Bickerton et al., 2012) and **synthetic accessibility (SA)** (Ertl & Schuffenhauer, 2009) for held-out conditions, following CURE's evaluation.

Table 1: **Knockdown-guided generation.** Every method is asked for a population of 100 molecules and is scored on what it returns, which is 66–100 for the baselines and 100 for GxEvo. Appendix B gives the draw budget behind each row.

Method	General generation			Reference similarity			Task-corpus overlap	
	Valid ↑	Unique ↑	Novel ↑	Mean ↑	Max ↑	NHS/100 ↑	NN-T ↓	Conn. overlap ↓
TRIOMPHE	0.993	0.934	0.983	.205	.384	0.99	**.253**	**.006**
FAME	.768	.990	.904	.272	.526	3.93	.519	.117
GxVAEs	.758	**.993**	.733	.278	.494	2.64	.797	.301
GexMolGen	**1.000**	.947	**1.000**	.295	.575	3.70	.792	.161
Gx2Mol	.826	.917	.246	.299	**.694**	0.93	1.000	.853
SmilesGEN	.926	.952	.248	.281	.510	2.05	1.000	.796
CURE	.656	.424	**1.000**	.284	.376	0.66	1.000	.644
GxEvo	0.965	.991	.975	**.371**	.617	**23.83**	.543	.033

4.2 KNOCKDOWN-GUIDED GENERATION

We first evaluate whether GxEvo can generate populations that align with a knockdown target while extending beyond the chemistry seen in the task corpus. The baselines in Table 1 reveal a gap between finding one close molecule and generating a strong population. Gx2Mol records the highest max similarity, but its lower population mean and novel hit scaffold yield are accompanied by 85.3% connectivity overlap with the task corpus. TRIOMPHE shows the opposite pattern, with little task-corpus overlap but weak reference alignment. GxEvo avoids both outcomes. It achieves the highest population mean and just over six times the novel hit scaffold yield of the strongest baseline while limiting connectivity overlap to 3.3%. Per-target results in Appendix H show that the aggregate gains are broadly distributed across targets. The advantage is preserved under a single shared novelty reference corpus (Appendix M). Appendix N bounds the same metrics from below with molecule sources that no method generated: random ZINC-250k, random task corpus, random ChEMBL, and known actives of the other nine targets. GxEvo is the only row that stands above all four in every column.

4.3 CANDIDATE-BUDGET CONTROL

To test whether the gains in Table 1 arise simply from evaluating more candidates, we match every method to the search's candidate and output budgets, using the same forward-model score for selection. We additionally compare GxEvo with budget-matched independent sampling from its own seed generator, isolating parent-conditioned evolutionary refinement from additional condition-only proposals.

Table 2: **Candidate-budget control on knockdown targets.**

Method	Candidate budget	Mean sim. ↑	Max sim. ↑	NHS/100 ↑
TRIOMPHE	1,500	.212	.378	1.50
CURE	1,500	.262	.384	2.35
FAME	≈ 1,500	.274	.544	5.10
GxVAEs	1,500	.271	.482	3.60
SmilesGEN	1,500	.294	.578	3.37
GexMolGen	1,500	.306	.575	6.10
Gx2Mol	1,500	.314	**.748**	1.53
Seed generator	100	.332	.596	12.64
Independent resampling	1,500	.344	.563	16.48
GxEvo	1,500	**.371**	.617	**23.83**

Under the matched budget, GxEvo remains best on both population mean and novel hit scaffold yield (Table 2). It also outperforms independent sampling from the same seed generator, demonstrating that evolutionary refinement contributes additional gains beyond increased sampling. Gx2Mol retains the highest max similarity score, reinforcing the distinction between recovering one close candidate and improving population-level quality. These controls support the premise of our formulation, that candidate evaluations are most useful when they influence where subsequent proposals are generated. Full per-method budget effects and diversity diagnostics are provided in Appendix F; search dynamics are reported in Appendix I.

4.4 GENERATION FOR HELD-OUT CONDITIONS

Having established that the knockdown improvement persists under matched candidate budgets, we next ask whether GxEvo generalizes beyond gene-knockdown targets to held-out drug–cell conditions. Here each method follows its own standard sampling protocol; Section 4.3 reports that the baselines gain at most $+.015$ mean similarity from a $15\times$ budget increase on the knockdown axis. Table 3 shows the same pattern across all three splits. GxEvo leads both structural similarity measures and

Table 3: **Generation for held-out conditions.** CURE is evaluated on 199 CELL conditions because one condition is absent from its test index.

		General generation					Target-conditioned performance		
Split	Method	Valid ↑	Unique ↑	Novel ↑	QED ↑	SA ↓	Morgan ↑	Fraggle ↑	PRnet-MSE ↓
RANDOM	TRIOMPHE	.987	.911	.968	.482	4.30	.148	.364	.01120
	GxVAEs	.637	**1.000**	.984	**.626**	3.19	.151	.387	.00998
	GexMolGen	**1.000**	.952	**1.000**	.460	3.69	.132	.373	.01262
	Gx2Mol	.828	.946	.479	.580	3.24	.184	.417	.00929
	SmilesGEN	.876	.976	.294	.581	3.19	.150	.389	.00979
	CURE	.443	.517	.879	.563	3.79	.167	.403	.01051
	GxEvo	0.981	.935	.908	.522	**3.14**	**.223**	**.456**	**.00888**
SCAFFOLD	TRIOMPHE	.971	.984	.992	.399	4.82	.114	.300	.01299
	GxVAEs	.652	**1.000**	.980	**.603**	3.39	.171	.412	.01067
	GexMolGen	**1.000**	.951	**1.000**	.461	3.70	.139	.378	.01298
	Gx2Mol	.848	.939	.518	.586	3.27	.209	.445	.00981
	SmilesGEN	.882	.961	.275	.580	3.21	.176	.418	.01034
	CURE	.430	.619	.932	.578	3.99	.159	.404	.01109
	GxEvo	0.985	.937	.908	.560	**3.16**	**.234**	**.469**	**.00957**
CELL	TRIOMPHE	.987	.916	.972	.492	4.28	.117	.326	.01980
	GxVAEs	.619	**1.000**	.985	**.599**	3.28	.140	.377	.01846
	GexMolGen	**1.000**	.955	**1.000**	.462	3.69	.129	.361	.02071
	Gx2Mol	.823	.952	.513	.574	3.16	.147	.380	.01775
	SmilesGEN	.863	.976	.305	.580	3.12	.131	.364	.01837
	CURE	.533	.492	.841	.584	3.71	.134	.365	.01871
	GxEvo	0.979	.963	.951	.546	**2.92**	**.185**	**.424**	**.01754**

PRnet-MSE, so its advantage is shared by global fingerprint similarity, fragment-level similarity, and a response-based proxy rather than tied to one evaluator. Paired comparisons over the shared conditions in each split consistently favor GxEvo against every baseline on all three target-conditioned metrics (Appendix E).

The lead remains on the scaffold split, where the target structural families are absent from training, and on the cell split, where the tissue context is unseen. Appendix J provides an additional condition-specificity check using an external response model and measured held-out responses, without relying on structural similarity or the search fitness. On this score, the evolved populations approach the ground-truth-drug reference and remain well separated from mismatched drugs and random molecules, so the advantage is not confined to the knockdown benchmark.

4.5 DESIGN CHOICES FOR EVOLUTIONARY TBDD

We next evaluate two design choices for evolutionary TBDD: a candidate-discriminative reward and learned parent-aware variation.

Candidate-Discriminative Reward. Low average response-prediction error does not guarantee a useful fitness function. Evolutionary search depends on relative rankings, so a model that maps different molecules to a similar average response can achieve low error while providing little selection pressure. We therefore measure discrimination with two complementary offline tests.

Table 4: **Forward-model discrimination and search utility.**

	Offline discrimination				In-loop search	
Reward model	Top-1 % (chance .25)	MRR ↑	Med. rank /400 ↓	Ligand AUC ↑	Mean ↑	NHS/100 ↑
TranSiGen	2.2	.049	163	.603	.354	5.95
PRnet	2.5	.071	76	.506	.325	9.30
XPert	1.5	.034	191	.613	.344	9.53
GxEvo	**4.0**	**.094**	**59**	**.635**	**.371**	**23.83**
w/o InfoNCE	3.5	.074	93	.623	.336	14.08

Response retrieval asks whether a held-out compound's predicted response identifies its own measured response, while *ligand ranking* asks whether a target's known actives rank above the actives of other targets. Both protocols are defined in Appendix B. Top-1, MRR and median rank evaluate response retrieval against 400 candidates, and ligand AUC evaluates ligand ranking. We also replace only the reward model inside the same search loop to test whether offline discrimination improves the final populations.

Our forward response model performs best on all four offline discrimination measures and yields the strongest final population on both in-loop metrics (Table 4), which is consistent with candidate discrimination mattering for evolutionary fitness. Among the baselines, however, the offline ranking does not match the in-loop one. Removing the InfoNCE term, which trains the model to tell

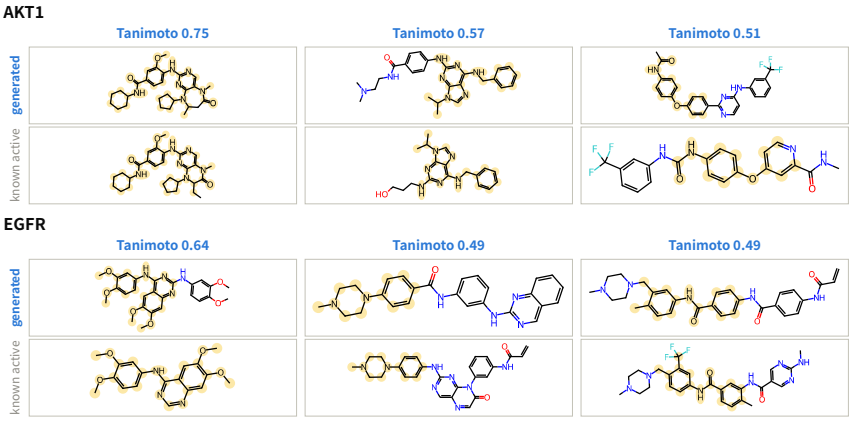

Figure 3: **Generated molecules and their nearest known actives, for AKT1 and EGFR.** Each generated molecule is shown above its nearest ExCAPE active, with the shared substructure shaded; Appendix K covers all ten targets.

candidate molecules apart rather than to predict any one response more accurately, lowers every offline measure and cuts novel hit scaffold yield from 23.83 to 14.08, so much of the gap comes from the discriminative objective rather than the backbone.

Learned Parent-aware Variation. We replace the learned mutation operator with BRICS replacement, Graph GA, FDSL-DD, or MOLLEO. The learned mutation operator outperforms all tested generic variation operators on both mean similarity and novel hit scaffold yield (Table 5). The initial population, reward, selection rule, and candidate budget are fixed across all arms, and every arm varies a selected parent, so the comparison isolates what the learned operator adds, namely conditioning on the target response and training on observed parent–child pairs.

Table 5: **Controlled comparison of variation operators.**

Variation operator	Mean sim. ↑	Novel hit scaffolds/100 ↑
BRICS	.283	3.83
Graph GA	.267	2.05
FDSL-DD	.267	1.25
MOLLEO	.313	2.50
GxEvo	**.371**	**23.83**

4.6 Qualitative case study

Figure 3 shows three generated molecules each for AKT1 and EGFR beside their nearest known active; Appendix K gives the selection rule and applies it to all ten targets.

Each pair shares a substructure while the surrounding context differs. The AKT1 examples retain fused heterocyclic cores and aryl-linked amide motifs while changing ring systems, linkers, and peripheral groups; the EGFR examples carry motifs characteristic of 4-anilinoquinazoline and acrylamide-bearing inhibitors on different surrounding scaffolds (Smaill et al., 2001). The six molecules have distinct scaffolds, match different known actives, and range from .49 to .75 in similarity, so GxEvo populates multiple regions of target-associated chemical space rather than variations of a single chemotype.

5 Conclusion

In this paper, we introduced GxEvo, a population-level evolutionary framework for TBDD in which a condition-specific seed population is iteratively refined using a candidate-discriminative forward response model and a learned parent-aware mutation operator. On knockdown targets it improves population mean similarity from .299 to .371 and novel hit scaffold yield from 3.93 to 23.83 relative to the best baseline on each metric, and it achieves the highest structural reference similarity on all three held-out drug–cell splits. Budget-matched and component-substitution controls show that these gains are not explained by additional sampling alone and that both the reward and variation mechanisms contribute. These results establish population-based evolutionary search as an effective strategy for generating response-aligned molecular hypotheses across diverse scaffolds. Because the reward and evaluation metrics are computational proxies, the generated molecules remain hypotheses requiring experimental validation.

AI USE STATEMENT

In this work, we used generative AI tools to implement methods: parts of the code for GxEvo and for the reproduced baselines were written with the assistance of an LLM coding assistant.

We have not used generative AI tools to generate synthetic data sets, to help develop theoretical models or conceptual frameworks, to formulate mathematical claims, to provide critical ingredients for proving mathematical claims, to assist in the writing of proofs, to propose or refine the hypotheses of this work, to design or provide feedback on our research methodology or experiments, to clean or reformat the datasets, to support qualitative or thematic data analysis, or to interpret the reported results. Assistance with translation is not applicable to this work.

Additionally, we used generative AI tools to create and edit software code, to create and modify figures, and to edit the manuscript for readability and layout.

We have reviewed all AI-assisted work. All AI-assisted code was read and tested by the authors, and every number reported in this paper was produced by executing that code on the stated data rather than generated by a language model. AI-assisted figures were checked against the underlying result files. All text in this paper was written by the authors; generative AI tools were used only to revise wording, formatting, and layout of text the authors had already written. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

We note separately that GPT-5 is used as a component of the MOLLEO variation-operator control (Appendix B); this is part of the evaluated method rather than of the authoring process.

REPRODUCIBILITY STATEMENT

Code for the seed generator, mutation operator, forward response model, and the evolutionary search loop, together with the search and evaluation scripts, is available at `https://anonymous.4open.science/r/GxEvo`. Baseline results are reproduced from each method's own released code under the filters described below, which we do not redistribute. All data are public: transcriptional profiles come from LINCS L1000 as redistributed by PRnet, reference ligands from ExCAPE-DB, and the pretraining and reference collections are ZINC-250k and ChEMBL. Section 3 specifies the search procedure, and Appendix A gives the architectures, training objectives, optimizer settings, gene-weighting rule, and search hyperparameters. Appendix B documents the evaluation protocol, the knockdown target construction, the draw budget behind each reported row, the leakage filters applied to every retrainable corpus, and the reproduction decisions taken for each baseline; Appendix D records the four baselines whose releases did not settle a training or conditioning choice on their own. Appendix G gives the three held-out splits, their partition sizes, and the fixed lists of 200 evaluation conditions that every method and arm is scored on, so that all comparisons are paired. Appendix E reports the replicate count and dispersion behind every knockdown number, and Appendix C the wall-clock cost of one search run and the hardware it was measured on.

ETHICS STATEMENT

This work uses only publicly available data. The LINCS L1000 profiles are measurements on immortalized human cell lines and carry no individual-level human-subject information; ExCAPE-DB, ChEMBL, and ZINC are public chemical databases. No human subjects, personally identifiable data, or animal experiments are involved, and we are not aware of conflicts of interest or sponsorship concerns affecting this work.

The molecules GxEvo produces are computational hypotheses. Both the search reward and the evaluation metrics are model-based or structural proxies, and we report no docking, retrosynthesis, selectivity profiling, or assay evidence for any generated structure (Appendix K). Structural similarity to known actives indicates that a molecule occupies a relevant chemical neighborhood; it is not a prediction of activity, safety, or synthesizability, and none of these molecules should be treated as a drug candidate without experimental validation.

Generative molecular design carries a dual-use consideration common to the field, in that methods that propose bioactive-looking structures for a desired cellular response could in principle be directed at harmful objectives. We judge the incremental risk here to be low: the method is trained and evaluated on public perturbation data, its targets are standard oncology-related genes, and its outputs

are unvalidated hypotheses that still require the same synthesis and testing as any other design proposal.

REFERENCES

Josep Arús-Pous, Simon Johansson, Oleksii Prykhodko, Esben Bjerrum, Christian Tyrchan, Jean-Louis Reymond, Hongming Chen, and Ola Engkvist. Randomized smiles strings improve the quality of molecular generative models. *Journal of Cheminformatics*, 11, 11 2019. doi: 10.1186/s13321-019-0393-0.

Guy W Bemis and Mark A Murcko. The properties of known drugs. 1. molecular frameworks. *Journal of medicinal chemistry*, 39(15):2887–2893, 1996.

G Richard Bickerton, Gaia V Paolini, Jérémy Besnard, Sorel Muresan, and Andrew L Hopkins. Quantifying the chemical beauty of drugs. *Nature chemistry*, 4(2):90–98, 2012.

Nathan Brown and Edgar Jacoby. On scaffolds and hopping in medicinal chemistry. *Mini reviews in medicinal chemistry*, 6(11):1217–1229, 2006.

Timothy B Callis, Taylor R Garrett, Andrew P Montgomery, Jonathan J Danon, and Michael Kassiou. Recent scaffold hopping applications in central nervous system drug discovery. *Journal of Medicinal Chemistry*, 65(20):13483–13504, 2022.

Rohan Chandraghatgi, Hai-Feng Ji, Gail L. Rosen, and Bahrad A. Sokhansanj. Streamlining computational fragment-based drug discovery through evolutionary optimization informed by ligand-based virtual prescreening. *Journal of Chemical Information and Modeling*, 64(9):3826–3840, 05 2024. ISSN 1549-9596. doi: 10.1021/acs.jcim.4c00234. URL https://doi.org/10.1021/acs.jcim.4c00234.

Jiabei Cheng, Xiaoyong Pan, Yi Fang, Kaiyuan Yang, Yiming Xue, Qingran Yan, and Ye Yuan. Gexmolgen: cross-modal generation of hit-like molecules via large language model encoding of gene expression signatures. *Briefings in bioinformatics*, 25(6):bbae525, 2024.

Gabriel Bianchin de Oliveira and Fahad Saeed. Moldeberta: Foundational model for physicochemical and substructure-informed molecular representation learning. *bioRxiv*, 2026. doi: 10.64898/2026.02.15.706011. URL https://www.biorxiv.org/content/early/2026/06/10/2026.02.15.706011.

Jorg Degen, Christof Wegscheid-Gerlach, Andrea Zaliani, and Matthias Rarey. On the art of compiling and using'drug-like'chemical fragment spaces. *ChemMedChem*, 3(10):1503, 2008.

Peter Ertl and Ansgar Schuffenhauer. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *Journal of cheminformatics*, 1(1):8, 2009.

Yue Guo, Hao Zhang, Haitao Hu, Jialu Wu, Ji Cao, Chang-Yu Hsieh, and Bo Yang. Modelling drug-induced cellular perturbation responses with a biologically informed dual-branch transformer. *Nature Machine Intelligence*, 8(1):96–112, 2026.

Stephen R Heller, Alan McNaught, Igor Pletnev, Stephen Stein, and Dmitrii Tchekhovskoi. Inchi, the iupac international chemical identifier. *Journal of cheminformatics*, 7(1):23, 2015.

James P Hughes, Stephen Rees, S Barrett Kalindjian, and Karen L Philpott. Principles of early drug discovery. *British journal of pharmacology*, 162(6):1239–1249, 2011.

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

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.

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, 2024.

Chen Li and Yoshihiro Yamanishi. Gx2mol: De novo generation of hit-like molecules from gene expression profiles. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 333–349. Springer, 2025.

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, 2025. ISBN 978-1-956792-06-5. doi: 10.24963/ijcai.2025/1008. URL https://doi.org/10.24963/ijcai.2025/1008.

Ricardo Macarron, Martyn N Banks, Dejan Bojanic, David J Burns, Dragan A Cirovic, Tina Garyantes, Darren VS Green, Robert P Hertzberg, William P Janzen, Jeff W Paslay, et al. Impact of high-throughput screening in biomedical research. *Nature reviews Drug discovery*, 10(3):188–195, 2011.

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.

John G Moffat, Fabien Vincent, Jonathan A Lee, Jörg Eder, and Marco Prunotto. Opportunities and challenges in phenotypic drug discovery: an industry perspective. *Nature reviews Drug discovery*, 16(8):531–543, 2017.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. Film: visual reasoning with a general conditioning layer. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press, 2018. ISBN 978-1-57735-800-8.

Thai-Hoang Pham, Lei Xie, and Ping Zhang. Fame: fragment-based conditional molecular generation for phenotypic drug discovery. In *Proceedings of the 2022 SIAM International Conference on Data Mining (SDM)*, pp. 720–728. SIAM, 2022.

Xiaoning Qi, Lianhe Zhao, Chenyu Tian, Yueyue Li, Zhen-Lin Chen, Peipei Huo, Runsheng Chen, Xiaodong Liu, Baoping Wan, Shengyong Yang, et al. Predicting transcriptional responses to novel chemical perturbations using deep generative model for drug discovery. *Nature Communications*, 15(1):9256, 2024.

Jean-Louis Reymond. The chemical space project. *Accounts of chemical research*, 48(3):722–730, 2015.

David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5):742–754, 2010.

Benjamin Sanchez-Lengeling and Alán Aspuru-Guzik. Inverse molecular design using machine learning: Generative models for matter engineering. *Science*, 361(6400):360–365, 2018.

Gisbert Schneider and Uli Fechner. Computer-based de novo design of drug-like molecules. *Nature reviews Drug discovery*, 4(8):649–663, 2005.

Gisbert Schneider, Werner Neidhart, Thomas Giller, and Gerard Schmid. "scaffold-hopping" by topological pharmacophore search: a contribution to virtual screening. *Angewandte Chemie International Edition*, 38(19):2894–2896, 1999.

Jeff B Smaill, HD Hollis Showalter, Hairong Zhou, Alexander J Bridges, Dennis J McNamara, David W Fry, James M Nelson, Veronika Sherwood, Patrick W Vincent, Bill J Roberts, et al. Tyrosine kinase inhibitors. 18. 6-substituted 4-anilinoquinazolines and 4-anilinopyrido [3, 4-d] pyrimidines as soluble, irreversible inhibitors of the epidermal growth factor receptor. *Journal of medicinal chemistry*, 44(3):429–440, 2001.

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, et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*, 171(6):1437–1452, 2017.

Hongmao Sun, Gregory Tawa, and Anders Wallqvist. Classification of scaffold-hopping approaches. *Drug discovery today*, 17(7-8):310–324, 2012.

Jiangming Sun, Nina Jeliazkova, Vladimir Chupakhin, Jose-Felipe Golib-Dzib, Ola Engkvist, Lars Carlsson, Jörg Wegner, Hugo Ceulemans, Ivan Georgiev, Vedrin Jeliazkov, et al. Excape-db: an integrated large scale dataset facilitating big data analysis in chemogenomics. *Journal of cheminformatics*, 9(1):17, 2017.

David C Swinney and Jason Anthony. How were new medicines discovered? *Nature reviews Drug discovery*, 10(7):507–519, 2011.

Xiaochu Tong, Ning Qu, Xiangtai Kong, Shengkun Ni, Jingyi Zhou, Kun Wang, Lehan Zhang, Yiming Wen, Jiangshan Shi, Sulin Zhang, et al. Deep representation learning of chemical-induced transcriptional profile for phenotype-based drug discovery. *Nature Communications*, 15(1):5378, 2024.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. 2018. URL https://arxiv.org/abs/1807.03748.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Haorui Wang, Marta Skreta, Cher-Tian Ser, Wenhao Gao, Lingkai Kong, Felix Strieth-Kalthoff, Chenru Duan, Yuchen Zhuang, Yue Yu, Yanqiao Zhu, Yuanqi Du, Alan Aspuru-Guzik, Kirill Neklyudov, and Chao Zhang. Efficient evolutionary search over chemical space with large language models. In *International Conference on Learning Representations*, volume 2025, pp. 51694–51727, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/804b5e300c9ed4e3ea3b073f186f4adc-Paper-Conference.pdf.

Michael J Waring, John Arrowsmith, Andrew R Leach, Paul D Leeson, Sam Mandrell, Robert M Owen, Garry Pairaudeau, William D Pennie, Stephen D Pickett, Jibo Wang, et al. An analysis of the attrition of drug candidates from four major pharmaceutical companies. *Nature reviews Drug discovery*, 14(7):475–486, 2015.

Ziyu Xu, Zijian Zhang, Liang Wang, Zhiyuan Liu, Qiang Liu, Shu Wu, and Liang Wang. Reading the cell, designing the cure: Perturbation-conditioned molecular diffusion for function-oriented drug design. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=7pinsbGVLt.

Supplementary Material

- GXEvo: Evolutionary Search for Transcriptome-Conditioned Molecular Design -

A Model and search implementation **15**

B Evaluation and baseline details **16**

C Inference cost **18**

D Reproduction fidelity of prior methods **19**

E Replication and dispersion **20**

F Budget-matched baselines **21**

G Held-out-condition splits **23**

H Per-gene knockdown results **24**

I Search dynamics and robustness **25**

J Functional condition-specificity **26**

K Case-study selection and interpretation **29**

L Novel hit scaffolds on held-out conditions **29**

M Novel hit scaffolds under a common reference corpus **29**

N Non-generative floors for the structural metrics **32**

O Identity audit of fingerprint ties **33**

A MODEL AND SEARCH IMPLEMENTATION

Conditional molecular models. The seed generator and mutation operator use the same Transformer backbone with hidden size 512, eight layers, eight attention heads, feed-forward size 2048, dropout 0.1, and maximum sequence length 256. Both are initialized from the same active-filtered ZINC checkpoint and fine-tuned separately. The seed generator is fine-tuned for ten epochs, while the mutation operator is fine-tuned for three epochs with parent dropout 0.2. We optimize both models with AdamW using batch size 128, learning rate 2×10^{-4}, weight decay 0.01, and $(\beta_1, \beta_2) = (0.9, 0.95)$. The learning rate follows a cosine schedule after a linear warm-up over the first 5% of updates. Training uses bfloat16 precision. At inference, both models sample with temperature 1.0 and nucleus probability 0.95. During search, generated children are canonicalized before comparison with previously scored molecules. Per-gene standardization statistics are computed from the training partition only and applied unchanged to validation and test conditions.

Forward response model. The forward response model fine-tunes MolDeBERTa end to end. It uses attention-mask-weighted mean pooling followed by FiLM conditioning and a two-layer prediction head. We train for three epochs with batch size 128 and AdamW. The MolDeBERTa parameters use learning rate 2×10^{-5}, while the conditioning and prediction layers use 2×10^{-4}. Weight decay is 0.01, and the learning rate follows a cosine schedule with 5% linear warm-up. We set $\lambda_{\mathrm{rec}} = 1$, $\lambda_{\cos} = 0.5$, and $\lambda_{\mathrm{NCE}} = 1$.

Gene weights are defined per compound from replicate statistics rather than from the individual well being fitted. For compound u with n_u training wells, let $\bar{\Delta}_g(u)$ denote its mean response at gene g, and let σ_g denote the within-replicate-group standard deviation of gene g, pooled over all (compound, cell, dose) groups containing at least two wells. We rank genes by

$$t_g(u) = \frac{\left|\bar{\Delta}_g(u)\right| \sqrt{n_u}}{\sigma_g}.$$

The weight vector $w^{(u)}$ assigns unit weight to the 100 genes with the largest $t_g(u)$ and weight $\rho = 0.1$ to the remaining coordinates. Compounds with fewer than three training wells carry insufficient replicate information for this ranking and therefore receive a uniform weight vector. Incorporating replicate variability avoids preferentially weighting genes whose large observed responses arise from high measurement noise.

Write $\hat{\Delta}_i, \Delta_i \in \mathbb{R}^{978}$ for the predicted and measured responses of example i in a shuffled minibatch of size m, and let $w_i = w^{(u_i)}$ denote the gene weights associated with its compound u_i. The soft-DEG-aware prediction loss combines a weight-normalized Huber term and a weighted cosine term:

$$\mathcal{L}_{\mathrm{pred}} = \lambda_{\mathrm{rec}} \mathcal{L}_{\mathrm{rec}} + \lambda_{\cos} \mathcal{L}_{\cos}.$$

The reconstruction term is

$$\mathcal{L}_{\mathrm{rec}} = \frac{1}{m} \sum_{i=1}^{m} \frac{\sum_g w_{ig}\, h_\delta\left(\hat{\Delta}_{ig} - \Delta_{ig}\right)}{\sum_g w_{ig}}, \qquad h_\delta(x) = \begin{cases} \frac{1}{2}x^2, & |x| \le \delta, \\ \delta\left(|x| - \frac{1}{2}\delta\right), & |x| > \delta, \end{cases}$$

and the cosine term is

$$\mathcal{L}_{\cos} = 1 - \frac{1}{m} \sum_{i=1}^{m} \cos\left(\sqrt{w_i} \odot \hat{\Delta}_i,\, \sqrt{w_i} \odot \Delta_i\right).$$

Here $\odot$ denotes elementwise multiplication and the square root is applied elementwise, so that the weights enter the inner product linearly. We set $\delta = 1$; the standardized residuals are an order of magnitude smaller, so the Huber term remains in its quadratic regime throughout training.

To preserve molecule-specific distinctions beyond prediction fidelity, we additionally use an InfoNCE batch-discrimination objective. Define

$$s_{ij} = \frac{\cos(\hat{\Delta}_i, \Delta_j)}{\kappa}, \qquad \kappa = 0.07,$$

where κ is the softmax temperature, distinct from the hit threshold τ in Section 4.1. The batch-discrimination loss is

$$\mathcal{L}_{\mathrm{NCE}} = -\frac{1}{m} \sum_{i=1}^{m} \log \frac{\exp(s_{ii})}{\exp(s_{ii}) + \sum_{j:u_j \neq u_i} \exp(s_{ij})}.$$

Measurements of the same compound are excluded from the denominator so that they are not treated as negatives; the remaining examples in the minibatch provide negative responses.

The full objective is therefore

$$\mathcal{L}_{\text{resp}} = \mathcal{L}_{\text{pred}} + \lambda_{\text{NCE}}\mathcal{L}_{\text{NCE}}.$$

The *w/o InfoNCE* arm of Table 4 sets $\lambda_{\text{NCE}} = 0$ and is otherwise trained identically, on the same corpus and for the same number of epochs, so the two differ only in the discrimination term.

Evolutionary search. We set the population size to $N = 100$, the number of children per generation to $B = 70$, and the number of generations to $G = 20$, giving a total proposal budget of $N + GB = 1{,}500$ decoded candidates per target. Candidate responses are scored at the fixed forward-model dose $d_{\text{eval}} = 10\,\mu\text{M}$ in both evaluation settings.

Candidate scoring and target construction treat dose differently. The forward model is trained on individual wells with $\log_{10}(d + 1)$ as an input and therefore predicts a dose-conditional response, whereas the target Δ for a condition is marginalized over the doses at which that compound–cell pair was measured. Scoring every candidate at $d_{\text{eval}} = 10\,\mu\text{M}$ therefore compares a dose-conditional prediction with a dose-marginal target. Among the retained LINCS wells, $10\,\mu\text{M}$ is the modal dose (49.9%), although 40.1% of the resulting conditions have no measurement at that dose.

Parents are sampled with probability proportional to shifted fitness: from a population carrying fitness values $\{f_k\}$, parent j is drawn with replacement with probability $(f_j - \min_k f_k + \varepsilon)/\sum_l (f_l - \min_k f_k + \varepsilon)$, where $\varepsilon = 10^{-6}$. The shift is required because the fitness is a cosine and can be negative; it is monotone, so it preserves the ordering the rule is built on. Survival uses elitist truncation. We use the same 20-generation endpoint for every comparison. The primary population mean has plateaued by generation 14, while novel hit scaffold yield changes by 0.1 between generations 18 and 20.

B Evaluation and baseline details

Evaluation protocol. Unless noted otherwise, metrics are computed over the population a method returns when it is asked for 100 molecules per target. That population holds 100 valid molecules only for FAME, GexMolGen and GxEvo; averaged over the ten genes and the replicate draws the others return 66 (CURE), 76 (GxVAEs), 83 (Gx2Mol), 91 (TRIOMPHE) and 93 (SmilesGEN). Rates such as novel hit scaffold yield are normalized per 100 valid molecules, so rows of different size stay comparable. Structural similarities use radius-2, 2048-bit Morgan fingerprints (Rogers & Hahn, 2010) without chirality. NN-T in the knockdown comparison is the median over all valid molecules a method emits, pooled over targets and replicate draws.

Knockdown target construction. The knockdown targets are MCF7 `trt_sh.cgs` signatures for the ten evaluation genes. Because the compared generators were developed using different LINCS representations, we follow each method's published conditioning convention. GxEvo constructs its target from Level-3 instances using the same normalization and aggregation pipeline as its generator training data, whereas methods whose published pipelines use Level-5 consensus signatures, including TRIOMPHE and GxVAEs, are conditioned on the corresponding published Level-5 CGS profiles. All targets refer to the same gene perturbations in MCF7 and are represented over the 978 landmark genes.

Draw budgets behind the validity column. Validity is a parse rate over the candidates a method draws, so it depends on a budget that is not the same for every row. TRIOMPHE, CURE, GxVAEs, Gx2Mol, SmilesGEN, and GexMolGen each draw the 100 candidates their reported population is built from. FAME is specified in valid molecules rather than draws and spends 768–$1{,}792$ per gene to reach 100 (Appendix F). GxEvo draws $1{,}500$ per gene or condition, of which 100 are seed-generator samples and 20×70 are mutation-operator children. Four of these methods would report 1.000 if the rate were read off the returned rows instead: the TRIOMPHE, CURE, and FAME reference implementations write only successfully decoded molecules, and our search discards an invalid child without redrawing, so it never enters the population. For GxEvo the two stages decode at different rates, and the reported figure is their pooled value: on knockdown targets the seed generator parses $.940$ of its draws and the operator $.967$ of its children, giving $.965$ over all $60{,}000$ candidates the four seeds draw; on held-out conditions the seed generator parses $.964$–$.966$ and the operator

.980–.987. The operator rate is flat across the run — between .957 and .979 at every one of the twenty generations — so it is a property of the decoder rather than of how far the search has moved from its seeds. TRIOMPHE additionally removes duplicates and regenerated source molecules before writing, so its uniqueness and novelty are read on the same valid draws for the same reason: the condition-axis reproduction already emits every draw, and on the knockdown axis the draws are reconstructed with a sampler that records them before the writer filters.

Leakage filtering and baseline retraining. For the knockdown benchmark, we remove the ten evaluation genes' ExCAPE-DB actives (Sun et al., 2017) and connectivity-equivalent structures from every retrainable molecular corpus. We also exclude the held-out ranking compounds from training the forward response model. After filtering, the ZINC-250k pretraining corpus contains 249,039 molecules, and the full-data mutation operator used for the knockdown experiment is trained on 1,078,984 bidirectional parent–child examples mined from LINCS. Held-out-condition operators use the corresponding split-specific training corpora. TRIOMPHE, FAME, GxVAEs, Gx2Mol, SmilesGEN, CURE, and GxEvo are retrained under these filters. For each reproduced baseline we follow the authors' released code, and where the release omits a component or leaves a setting unspecified we follow the corresponding paper; where neither specifies it, we reuse that method's own setting from the nearest comparable stage. GexMolGen is evaluated from its published checkpoint because its training code was not released; its training corpus therefore cannot be filtered or fully audited. This asymmetry favors GexMolGen.

Similarity to the held-out compound. Morgan and Fraggle similarities on held-out conditions follow CURE's evaluation, with one correction. CURE's released routine computes Fraggle, Morgan and MACCS inside a single exception handler and returns zero for all three if any of them raises; Fraggle is computed first, and RDKit's Fraggle implementation raises on a few percent of pairs because its own fragmentation step returns a null molecule. Every such pair is therefore recorded as Morgan similarity zero although its Morgan fingerprints are well defined. The rate depends on molecular structure and so differs by method — 1.8% of our pairs against 4.1–6.3% for the baselines — which biases the comparison rather than shifting it uniformly, and in our favor. We therefore compute Morgan independently of Fraggle, and treat a pair on which the Fraggle routine raises as missing for Fraggle rather than as a similarity of zero. Molecule selection, the per-condition averaging, and the fingerprint settings are unchanged. The correction leaves the ordering of Table 3 unchanged on both measures.

Forward-model comparison. The response-retrieval benchmark uses 400 MCF7 compounds measured at dose 10 μM. These compounds belong to the held-out split used by our forward response model and PRnet. Because the published TranSiGen and XPert training sets contain many of these compounds, we retrain both models after excluding the complete 400-compound retrieval pool. The retrained models otherwise retain their published architectures and training objectives. The *w/o InfoNCE* arm is evaluated on the same 400-compound pool and the same ligand-ranking protocol as every other row, and drives the search under the identical settings, so its in-loop columns are comparable to the rest of Table 4.

Response retrieval ranks the 400 measured responses against each prediction by cosine similarity, and we report top-1 accuracy, mean reciprocal rank, and the median rank of the correct response; scoring a prediction against every measured response, rather than against its own alone, is what makes the test sensitive to molecule-specific discrimination rather than to average accuracy. Because that test stays internal to the perturbation corpus, ligand ranking supplies an external check: held-out ExCAPE molecules are scored by the directional agreement between their predicted response and each target's knockdown profile, and AUROC measures whether that target's known actives rank above the actives of the other nine targets.

MOLLEO control. Our MOLLEO port samples two parents in proportion to shifted fitness, uses GPT-5 with low reasoning effort for one crossover proposal per child, and applies the prescribed Graph GA mutation with probability .067. We request the explanation and molecule through a JSON schema rather than parsing free text. To preserve the common budget of 1,500 decoded candidates, we generate exactly 70 children per generation and omit MOLLEO's additional over-generation and similarity-pruning stage. The vendored Graph GA fallback and mutation use their module-level random-number generator, which is not controlled by the search seed. Because one child requires one language-model request, we report one seed over all ten targets for this control.

Table 6: **Inference cost for one knockdown target (AKT1).** Wall-clock seconds to serve one 100-molecule request. The candidate count for GxEvo denotes its explicit search budget; baseline timings follow their standard generation protocols.

Method	candidates	seconds
TRIOMPHE	100	3.8
FAME	100	8.0
GxVAEs	100	5.1
GexMolGen	100	304.9
Gx2Mol	100	3.1
SmilesGEN	100	3.3
CURE	100	69.9
GxEvo seed generator only	100	29.4
GxEvo (full search)	1,500	401.3

FDSL-DD control. FDSL-DD represents molecules as BRICS fragment lists and applies fragment substitution, single-point crossover, and elitist survival. Its second evolutionary stage requires protein pocket information and is not applicable to transcriptional targets. We therefore adapt its first stage using the common phenotype reward and a leakage-filtered LINCS fragment vocabulary.

C Inference cost

We measure the wall-clock cost of producing the reported deliverable for a single knockdown target: one 100-molecule request for one gene. All methods run sequentially on one RTX A6000 with four pinned CPU cores, in a single session on an otherwise idle machine: wall-clock on a shared box is dominated by scheduler contention, so rows measured at different times are not comparable. For GxEvo the search diagnostics are disabled, leaving generation, reward, selection, and survival unchanged. Each method is also run once requesting a single molecule, which separates the one-time cost of interpreter start-up, CUDA initialization, and checkpoint and corpus loading from the cost that scales with the number of molecules requested. We report that split in the text.

Table 6 shows that our cost is the candidate budget rather than the search machinery. Marginal cost is .264 seconds per candidate for the full search, .244 for the seed generator alone, and .238–.240 for independent resampling at the same budget, so the learned mutation operator adds roughly 10%. Drawing all 1,500 candidates in a single pass instead of over 20 generations takes 365.5 seconds against the 401.3 of the table, so the evolutionary loop itself accounts for about 9% of the total. The remainder is decoding and reward evaluation for 1,500 molecules, which is the price of search and not of any particular variation operator.

The one-time cost varies by two orders of magnitude across methods and is paid once per run rather than once per target. GexMolGen loads a single-cell language model and a hierarchical graph decoder, and its beam decode is below our measurement resolution: one molecule costs 305.7 seconds against 304.9 for a hundred. CURE spends 27 of its 70 seconds on loading; for the five remaining baselines, which return 100 molecules in 3–8 seconds, loading is the whole of it. Since all methods process genes sequentially, the ten-target cost is ten times the per-target generation cost plus one loading cost: approximately 66 minutes for GxEvo, 7.5 for CURE, and 5.1 for GexMolGen.

Two qualifications apply. First, methods differ in how many usable molecules a 100-molecule request returns, so cost per usable molecule is higher than the table implies for most of the baselines. Two of them write fewer than 100 rows: in the run timed here, on AKT1, CURE wrote 71 and TRIOMPHE 85. The reason differs between them. CURE's reference implementation emits only successfully decoded molecules, so its shortfall is decode failure. TRIOMPHE decodes almost everything it draws (validity .993) but additionally drops duplicates and regenerated source molecules before writing (Appendix B), so its shortfall is that filter rather than the decoder. The rest write 100 rows, but a fraction of those do not parse for GxVAEs, Gx2Mol and SmilesGEN (Table 1). Only FAME, GexMolGen and GxEvo return 100 usable molecules. Second, a repeated run of the same search took 400.0 seconds against the 401.3 reported here, a difference of 0.3%, so the reported ratios are stable.

D REPRODUCTION FIDELITY OF PRIOR METHODS

Four of the reproduced baselines required a decision that the release does not settle on its own: a training-target convention that changes what the method appears to do (GxVAEs), a release that ships neither data nor checkpoints (SmilesGEN), a conditioning pathway that is constant across examples as released (CURE), and a conditioning input the release cannot produce for a held-out condition (FAME).

GxVAEs training targets. GxVAEs is trained with randomized, non-canonical SMILES targets, following its published recipe. Retraining it instead with canonical targets on the same leakage-filtered corpus leaves mean target similarity unchanged (.277 against .278) but lowers standard novelty from .733 to .249 and novel hit scaffold yield from 2.64 to 0.95, while raising the fraction of outputs that are exact training molecules from .301 to .822. Table 1 therefore reports the stronger of the two arms for this baseline.

SmilesGEN data and checkpoints. SmilesGEN ships neither training data nor checkpoints: the files its loaders read are replaced by pointers to the primary databases, and the published preprocessing is not specific enough to rebuild them. Both of its stages are therefore trained from scratch, on the knockdown axis from the same leakage-filtered MCF7 corpus used for GxVAEs, which makes the two baselines training-matched and gives them a common novelty denominator, and on the condition axis from each split's own training partition like every other retrained baseline. Its released tokenizer encodes every molecule of that corpus, so no molecule is dropped. One deviation was unavoidable. Its training objective reconstructs an untreated profile alongside the perturbed one, and that target is read from the missing profile file; because our corpus stores differential expression, whose center is the untreated state, we supply the per-gene mean the released code computes. That target is easier to fit than in the absolute space the authors used, so this channel contributes less here than intended; an absolute basal profile instead would force the decoder to span two incompatible scales and would break the training match with GxVAEs.

GxVAEs and SmilesGEN on held-out conditions. Both methods consume a single 978-dimensional landmark profile and carry no cell-line input, so a held-out condition is the same object their released code already accepts; both are retrained from scratch on each split's training conditions, which also removes the leakage question a fixed corpus would raise. GxVAEs is driven by the differential channel, the quantity its level-5 training signatures already are. SmilesGEN is driven by the treated profile with the matching per-cell basal as the untreated target of its second reconstruction branch, so the substitute described above is not needed here: this axis stores the profile that objective was written for. Two properties of the released code required a driver of our own rather than a change to it. Their generation paths consume one condition per invocation, and SmilesGEN negates the conditioning vector for any target outside the ten knockdown genes its evaluation script names, which is the intended behavior for reversing a disease signature but the opposite of the requirement here, so that branch is not taken. Epoch counts are step-matched to the released settings, since these training partitions are about seven times the size of the released corpus. GxVAEs is trained here with the same randomized, non-canonical SMILES targets as on the knockdown axis, so one recipe — its published one — is used on both axes. On this axis that choice is not the one favorable to us: retraining it on canonical targets instead raises its Morgan similarity to .187/.211/.149 on the RANDOM/SCAFFOLD/CELL splits, its Fraggle similarity to .420/.450/.385, and lowers its PRnet-MSE to .00937/.00986/.01777. GxEvo still leads all three measures on all three splits against that arm, by a smaller margin than Table 3 shows.

FAME on held-out conditions FAME conditions on a 64-dimensional embedding rather than on a landmark profile, and the release does not include the contrastive denoiser that produces it. The knockdown embeddings ship with the code, but nothing in it maps a new condition into that space, so a held-out condition has no conditioning input FAME can consume; supplying one would mean reimplementing an unreleased model from its description. We therefore evaluate FAME on the knockdown axis only. This is a property of the release rather than of the method.

CURE conditioning pathway. In the released configuration, CURE's conditioning pathway is constant across examples. We restore the condition-dependent input specified by the model and evaluate both versions on the same original set of 500 conditions per split. Table 7 shows that the restored pathway improves Morgan-to-ground-truth similarity and PRnet-MSE on all three splits.

Table 7: **CURE conditioning-pathway audit.** Restored denotes CURE after restoring the condition-dependent input. Released denotes the constant-conditioning configuration. Each comparison uses the same 500 conditions within a split. Morgan and Fraggle follow the corrected definition of Appendix B; the Fraggle routine fails on 5.9–6.3% of the restored arm's pairs against 2.2–3.4% of the released arm's, so the uncorrected metric understated the restored model on both measures.

Split	Configuration	Valid ↑	Morgan ↑	Fraggle ↑	PRnet-MSE ↓
RANDOM	Restored	.455	.170	.407	.01084
	Released	.838	.134	.376	.01267
SCAFFOLD	Restored	.451	.163	.413	.01103
	Released	.805	.140	.389	.01280
CELL	Restored	.555	.137	.367	.01874
	Released	.850	.133	.376	.02011

The improvement is accompanied by substantially lower validity, and Fraggle similarity decreases slightly on the cell split.

Restoring condition dependence also reduces the three-split average connectivity overlap with CURE's task corpus from .757 to .581. Thus, the restored model is less dominated by training structures despite its lower validity. This table uses the original matched 500-condition audit. Table 3 instead reports the restored model regenerated on the shared evaluation lists for each split; CURE is scored on 199 rather than 200 cell-split conditions because one condition is absent from its test index.

E REPLICATION AND DISPERSION

Every knockdown value in the paper is a mean over replicates, and the unit of replication differs by comparison. A generation baseline is replicated by drawing its population again, so the three draws of Table 1 measure that method's own sampling variability. A search arm is replicated by rerunning the full search under a different random seed. Controlled arms share the same initial population within each replicate, while populations differ across replicates. The four seeds of Tables 2, 4 and 5 capture stochasticity in initialization, selection, and variation. Table 8 gives the spread of each. The means themselves are the values already printed in those tables and are not repeated here.

Two arms admit no dispersion. GexMolGen generates by deterministic beam search with greedy decoding, so redrawing returns the same population and the seed axis is a no-op; this is a property of the released method, not a shortfall of the reproduction. The MOLLEO control issues one language-model request per child, so a second seed would repeat the full generation cost of the arm; we therefore report a single seed over all ten targets, as Appendix B states.

Held-out conditions. A held-out-condition comparison replicates over the 200 evaluation conditions of its split rather than over seeds, and every method generates for the same conditions, so the comparisons are paired. Table 9 reports the paired difference and a Wilcoxon signed-rank test for each target-conditioned quantity of Table 3. Novel hit scaffolds are zero for both arms on most conditions — between 122 and 151 of the 200 comparisons are ties — so that test is decided by the minority of conditions on which either method places a novel scaffold above threshold.

GxEvo improves both structural similarities and PRnet-MSE against every baseline on every split; Table 9 reports the paired differences together with unadjusted Wilcoxon signed-rank tests. The novel hit scaffold column is reported here rather than in Table 3 for the reason given in Section 4.1, and it behaves differently: the advantage holds against four baselines on all three splits and against TRIOMPHE on two, but GxVAEs is statistically indistinguishable from GxEvo on all three. That arm reaches its yield by a trade the other columns make visible—its validity is .62–.65 versus about .98 for GxEvo, and every molecule it does return is further from the ground-truth compound.

The spread is small relative to the differences the paper rests on. On the budget comparison, the gap between independent resampling and the full search is .027 in mean similarity against seed standard deviations of .001 and .006, and 7.35 in novel hit scaffolds against 0.94 and 1.79. On the operator comparison every generic operator is at least .058 below GxEvo in mean similarity, against standard deviations of at most .010. The reward comparison is the one place where the margin is

Table 8: **Replicate count and dispersion of the knockdown comparisons.** n is the number of replicates behind the reported mean; s.d. is the sample standard deviation over those replicates, each replicate first averaged over the ten targets, so the spread is across replicates and not across (replicate, target) cells. Baseline rows replicate by redrawing the population; search rows replicate by reseeding the search. Floor rows redraw 100 molecules from a fixed 3,000-molecule pool, so their spread is that of the draw and not of the source database.

		s.d. over replicates	
Arm	n	mean similarity	NHS/100
Generation (Table 1); a replicate redraws the population			
TRIOMPHE	3	.001	0.12
FAME	3	.004	0.95
GxVAEs	3	.005	0.15
GexMolGen	1	deterministic beam search	
Gx2Mol	3	.004	0.31
SmilesGEN	3	.002	0.36
CURE	3	.002	0.11
Candidate budget (Table 2); a replicate reseeds the search			
Seed generator	4	.003	0.70
Independent resampling	4	.001	0.94
GxEvo	4	.006	1.79
Reward model (Table 4); a replicate reseeds the search			
PRnet	4	.014	2.42
TranSiGen	4	.018	1.53
XPert	4	.009	1.27
w/o InfoNCE	4	.013	4.01
Variation operator (Table 5); a replicate reseeds the search			
BRICS replacement	4	.010	2.21
Graph GA	4	.010	0.79
FDSL-DD	4	.006	0.48
MOLLEO	1	one language-model run	
Non-generative floors (Table 17); a replicate redraws 100 *molecules*			
ZINC-250k random	50	.005	0.90
LINCS task-corpus random	50	.005	0.00
ChEMBL random	50	.003	0.91
ExCAPE off-target active	50	.003	1.16

of the same order as the spread on one metric: TranSiGen reaches $.354 \pm .018$ in mean similarity against $.371 \pm .006$. Its novel hit scaffold yield, however, is 5.95 ± 1.53 against 23.83 ± 1.79, which is the separation the accompanying claim rests on, since the claim is that only our forward model improves both quantities at once.

F BUDGET-MATCHED BASELINES

This appendix expands the baseline side of Table 2. Each prior method is resampled to 1,500 candidates per knockdown target, the budget one search run spends, and each pool is reduced to 100 molecules by our reward. Three points fix the protocol. First, budgets count draws, including invalid ones, because the search also discards invalid children without redrawing. Second, selection keeps the best-scoring copy of each canonical SMILES before taking the top 100; without this step a generator that repeats molecules would return 100 copies of one structure and record a diversity collapse it did not commit, so the reported setting is the one favorable to the baseline. Third, the TRIOMPHE, CURE, and FAME reference implementations write only valid molecules, so their pools hold fewer than 1,500 rows while still consuming 1,500 draws.

FAME requires one further adjustment, because it is the only prior method whose sampling loop is specified in valid molecules rather than draws. It samples $n_{\text{sig}} \times L \times S$ candidates per round, where L and S are its per-signature latent and sample counts, and stops once 100 valid molecules have accumulated, so its 100-molecule row in Table 1 already consumes 768–1,792 draws depending on

Table 9: **Paired comparisons on held-out conditions.** Each entry is the mean paired difference in favor of GxEvo over the conditions of that split, with a Wilcoxon signed-rank test: $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$. PRnet-MSE is lower-is-better, so its column reports the baseline's error minus ours, scaled by 10^3. Fraggle is defined on the pairs its routine evaluates, so its n is 198–200. $n = 200$ conditions, except CURE on the CELL split (199; one condition is absent from its test index). Unadjusted Wilcoxon signed-rank p-values are shown for reference.

Split	Baseline	Δ Morgan	Δ Fraggle	Δ NHS/100	Δ PRnet-MSE ($\times 10^3$)
RANDOM	TRIOMPHE	$+0.074^{***}$	$+0.092^{***}$	$+0.64\,n.s.$	$+2.33^{***}$
	GxVAEs	$+0.072^{***}$	$+0.069^{***}$	-0.16 n.s.	$+1.11^{***}$
	GexMolGen	$+0.091^{***}$	$+0.083^{***}$	$+2.11^{***}$	$+3.75^{***}$
	Gx2Mol	$+0.039^{***}$	$+0.039^{***}$	$+1.43^{***}$	$+0.42^{***}$
	SmilesGEN	$+0.072^{***}$	$+0.066^{***}$	$+2.48^{***}$	$+0.91^{***}$
	CURE	$+0.055^{***}$	$+0.052^{***}$	$+2.09^{***}$	$+1.63^{***}$
SCAFFOLD	TRIOMPHE	$+0.120^{***}$	$+0.170^{***}$	$+2.66^{***}$	$+3.42^{***}$
	GxVAEs	$+0.063^{***}$	$+0.057^{***}$	-1.54 n.s.	$+1.10^{***}$
	GexMolGen	$+0.095^{***}$	$+0.091^{***}$	$+2.58^{***}$	$+3.41^{***}$
	Gx2Mol	$+0.025^{***}$	$+0.024^{***}$	$+1.96^{***}$	$+0.24^{***}$
	SmilesGEN	$+0.058^{***}$	$+0.051^{***}$	$+3.12^{***}$	$+0.77^{***}$
	CURE	$+0.075^{***}$	$+0.065^{***}$	$+2.98^{***}$	$+1.51^{***}$
CELL	TRIOMPHE	$+0.068^{***}$	$+0.098^{***}$	$+1.24^{***}$	$+2.25^{***}$
	GxVAEs	$+0.045^{***}$	$+0.047^{***}$	$+0.12$ n.s.	$+0.92^{***}$
	GexMolGen	$+0.056^{***}$	$+0.064^{***}$	$+1.53^{***}$	$+3.17^{***}$
	Gx2Mol	$+0.038^{***}$	$+0.044^{***}$	$+1.69^{***}$	$+0.21^{**}$
	SmilesGEN	$+0.054^{***}$	$+0.060^{***}$	$+1.95^{***}$	$+0.83^{***}$
	CURE	$+0.051^{***}$	$+0.059^{***}$	$+1.91^{***}$	$+1.12^{***}$

how many CRISPR signatures the gene has. One round exhausts the gene's signatures ($L = S = 8$ over its 12–28 of them) and yields 590–1,380 valid molecules, so which 100 of those are reported is fixed by the order the accumulator visits them: it walks the signatures round-robin, taking each one's first valid molecule, then each one's second, and stops at 100. The reported population is therefore spread evenly over the gene's signatures — roughly the first four to eight valid draws of each — rather than taken from the front of one signature or sampled at random. That row is therefore not a 100-candidate row, and FAME is the one baseline for which the matched budget is a small increase rather than a $15\times$ one. To equalize draws we set $S = \lceil 1500/(n_{\text{sig}}L) \rceil$ per gene and keep every valid molecule from that single round, which realizes 1,536–1,568 draws per gene.

The budget is not the explanation. The largest gain any baseline obtains is $+.015$ mean similarity, against a remaining gap to GxEvo of .058 to .160, and novel hit scaffolds stay at or below 6.10 against 23.83. Four methods respond in ways that identify their own limitation rather than a budget shortfall.

CURE exhausts its output distribution. 1,500 draws yield only 71–87 distinct molecules per target, so the selection step has fewer than 100 candidates to choose from and the emitted population averages 79.6 molecules. Mean similarity falls on all ten targets. Additional sampling returns the same structures, which is the same behavior its low uniqueness in Table 1 reports at the smaller budget.

Widening GexMolGen's beam costs diversity. Its generation is a deterministic beam search, so the only way to spend a larger budget is a wider beam, and beam ranks beyond the first hundred are increasingly redundant: distinct scaffolds fall from 76.3 to 54.3 even as mean similarity rises slightly. For this method the budget buys depth in a ranked list, not independent candidates.

TRIOMPHE does not respond to the budget. Its metrics change only slightly. Its candidates are drawn around five fixed source molecules per target, so a larger sample explores the same neighborhoods more densely.

Table 10: **Effect of the matched candidate budget on each baseline.** Each entry is the change from the 100-candidate population of Table 1 to the 1,500-candidate, reward-selected population of Table 2. Each Δ is taken before rounding, so adding a baseline row's Table 1 entry to its Δ reproduces its Table 2 entry to within one unit in the last digit. Positive means the larger budget helped. The seed generator row is not such a difference: it is our own single-shot reference put through this table's protocol at both budgets, and its endpoints are mean similarity .333 $\rightarrow$ .342, NHS/100 11.91 $\rightarrow$ 15.73. It is a separate run from the generation-0 seed row of Table 2, and the 1,500-draw arm reported there is independent resampling, which resamples over 20 generations under survival rather than in the single pass used here. Every row uses the same arm of its method as Table 1; for GxVAEs that is the randomized-target reproduction of Appendix D. Under the canonical-target arm the same two cells read $+.011$ and $+1.05$.

Method	Δ mean similarity	Δ NHS/100	scaffolds 100 $\rightarrow$ 1,500
TRIOMPHE	$+.007$	$+0.51$	68.3 $\rightarrow$ 86.2
CURE	$-.022$	$+1.69$	22.2 $\rightarrow$ 56.3
FAME	$+.002$	$+1.17$	87.8 $\rightarrow$ 88.3
GxVAEs	$-.006$	$+0.96$	72.5 $\rightarrow$ 91.3
SmilesGEN	$+.013$	$+1.32$	83.0 $\rightarrow$ 87.7
GexMolGen	$+.012$	$+2.40$	76.3 $\rightarrow$ 54.3
Gx2Mol	$+.015$	$+0.60$	68.2 $\rightarrow$ 88.1
Seed generator	$+.010$	$+3.82$	83.5 $\rightarrow$ 91.6

Table 11: **Held-out-condition partitions.** Counts are conditions; fractions are of the 107,906-condition universe. The random and scaffold splits realize the target 85:10:5 ratio. The cell split cannot, because it partitions whole tissues and a tissue is indivisible.

Split	Held-out unit	train	valid	test
RANDOM	compound–cell pair	91,720 (.850)	10,790 (.100)	5,396 (.050)
SCAFFOLD	Bemis–Murcko scaffold cluster	91,716 (.850)	10,792 (.100)	5,398 (.050)
CELL	tissue	73,029 (.677)	19,590 (.182)	15,287 (.142)

FAME was already sampling at scale. Its mean similarity moves by $+.002$ and its scaffold count by $+0.5$, the smallest response in the table. This is expected given the preceding paragraph: FAME's reported row already draws roughly 950 candidates per gene on average, so the matched budget adds about 60% more draws rather than 15$\times$, and the metric it most plausibly gains on, novel hit scaffolds, rises only from 3.93 to 5.10.

Retrieval-like generation is favored by the order statistic. Gx2Mol gains the most on max similarity (.694 $\rightarrow$ .748), widening our paired deficit to $-.131$. Its population mean gains only $+.015$, however, and novel hit scaffolds remain at 1.53. A generator that reproduces training actives converts extra draws into a better single best candidate; the population-level metrics do not follow.

G HELD-OUT-CONDITION SPLITS

A condition is a (compound, cell line) pair whose replicate wells, including those at different doses and exposure times, are averaged. The three splits partition one universe of 107,906 conditions over 13,104 compounds and 82 cell lines, drawn from LINCS L1000 (Subramanian et al., 2017) under the same filters as the generator corpus (nonzero dose, a basal profile for the cell line, and minimum replicate and cell-line counts). The knockdown-axis corpus of 85,310 conditions in Section 4 is a differently filtered view of the same LINCS base, taken after the knockdown evaluation actives are removed.

The scaffold split assigns Bemis–Murcko scaffolds to 4,419 clusters and partitions clusters, so no scaffold spans training and test. Holding out whole clusters also shrinks the available chemistry: its training partition contains 11,196 distinct compounds against 13,087 for the random split.

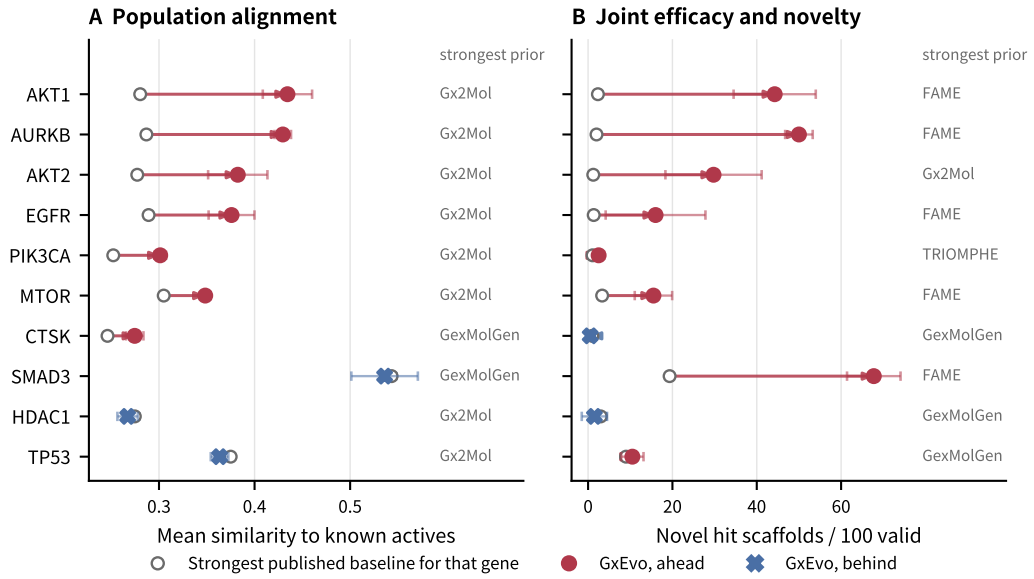

Figure 4: **Per-gene robustness of the knockdown result.** Each row compares GxEvo with the strongest published baseline for that gene and metric, named in the right-hand column because it differs from gene to gene. The left panel uses population mean similarity to known actives, the right uses novel hit scaffolds per 100 valid molecules. Filled circles mark the genes where GxEvo is ahead and crosses the genes where it is behind; error bars give the spread over the four search seeds. Genes are ordered by the population-mean margin, with the same order in both panels.

The cell split groups cell lines by the `primary_site` field of the LINCS `cell_info` table and holds out whole tissues. Training covers 16 named tissues plus two groups the table leaves unlabeled, validation is lung, and test is large intestine, comprising CL34, HCT116, HELA, HT115, HT29, LOVO, MDST8, NCIH508, RKO, SNU1040, SNUC5, SW480, SW620, and SW948. We follow the LINCS annotation as distributed, which assigns HELA to large intestine; HELA supplies 12 of the 200 evaluated conditions. This split is therefore a transfer to an unseen tissue rather than merely to unseen cell lines, and it is the hardest of the three: every method attains its worst PRnet-MSE there, and six of the seven attain their lowest Morgan similarity there (Table 3).

The evaluated conditions. Generating and scoring every test condition for every method is not affordable, so each split has a fixed list of 200 evaluation conditions, drawn once and reused by every method and every arm, so that all comparisons are paired. We verified that all 600 keys lie in their split's test partition, with no overlap with any training or validation partition. The random and cell lists cover 196 distinct compounds and the scaffold list 161; the random and scaffold lists span 43 cell lines and the cell list the 14 held-out ones.

These partitions reconstruct the protocol CURE describes rather than reusing its released split code, which selects test compounds by a random sample of SMILES within a single hard-coded cell line. We therefore retrain every method whose training code is available on each split's own training partition, as described in Appendix B.

H Per-gene knockdown results

Table 1 aggregates over the ten knockdown genes. This appendix gives the per-gene breakdown behind it: Figure 4 for the two headline metrics against the strongest baseline for each gene, and Table 12 for the max similarity statistic.

Table 12: **Per-gene max similarity** from Table 1, averaged over replicate draws.

Gene	GxEvo	Gx2Mol	GexMolGen	FAME	SmilesGEN	GxVAEs	TRIOMPHE	CURE
AKT1	**.738**	.715	.569	.428	.419	.398	.329	.324
AKT2	.743	**.791**	.448	.452	.404	.365	.254	.299
AURKB	.725	**.790**	.470	.538	.453	.420	.431	.313
CTSK	.432	.432	**.473**	.398	.365	.390	.267	.297
EGFR	.621	**.716**	.500	.518	.515	.577	.322	.344
HDAC1	.398	**.681**	.471	.503	.451	.399	.369	.312
MTOR	.522	**.744**	.436	.540	.560	.504	.400	.338
PIK3CA	.436	**.480**	.384	.412	.340	.393	.431	.325
SMAD3	.916	.808	**1.000**	.824	.842	.905	.486	.790
TP53	.642	.781	**1.000**	.644	.747	.593	.555	.417
mean	.617	**.694**	.575	.526	.510	.494	.384	.376

Gx2Mol wins the draw-averaged max similarity statistic against GxEvo on $7/10$ genes, but loses the *mean* of the same per-molecule scores on $8/10$. The aggregate comparison likewise reverses from .617 versus .694 in the extreme tail to .371 versus .299 for the population mean. Gx2Mol is also the only one of the seven baselines in Table 1 that exceeds GxEvo on this statistic; the other six fall below it by .042 to .241. We therefore report the maximum but do not rest the main claim on it. Gx2Mol's median NN-T is 1.000 and 85.3% of its molecules are training molecules, so its extreme tail frequently selects a retrieved molecule. GexMolGen attains exact active matches on SMAD3 and TP53, but its unavailable training code prevents the same corpus-level audit. We therefore do not interpret those matches as evidence of de novo recovery.

I Search dynamics and robustness

Search dynamics. Over the reported search, novel hit scaffolds increase from 12.64 to 23.83 per 100 valid molecules, while the number of scaffolds in the population remains nearly flat ($84.3 \rightarrow 85.2$), so selection reallocates mass toward scaffolds that cross the hit-likeness threshold rather than collapsing the portfolio to a few scaffolds. Decomposing the $1.89\times$ novel-hit increase gives a $1.62\times$ higher hit rate, $1.29\times$ greater scaffold novelty conditional on being a hit, and $0.90\times$ scaffolds per hit. After mean similarity has plateaued (generations 14–20), the remaining increase is driven mainly by conditional scaffold novelty ($.931 \rightarrow .947$).

Robustness to the hit threshold. Figure 6 evaluates the same populations over the threshold range $\tau \in [0.2, 0.6]$ rather than only at the pre-specified $\tau = 0.4$ of Table 1. GxEvo is first at every threshold in the range; the prior methods reorder among themselves.

Robustness to initialization. Table 13 transfers the evolutionary loop to external initial populations. Each run uses the baseline's own emitted knockdown population as generation 0 and is otherwise identical to the reported search.

The search improves every external initialization by .057–.089 in mean similarity, reaching final means of .352–.367, and multiplies novel hit scaffold yield from each external starting point by $4.7\times$ to $21.6\times$. Its benefit is therefore not specific to candidates from our seed generator. The two columns differ in how much of the outcome the starting population explains. Our seed generator gives the highest final mean, ahead of all three transplants. On novel hit scaffolds the starting populations span 0.93 to 12.64, a factor of 13.6, while the populations they reach span 17.40 to 23.83, a factor of 1.4; the gap between the arm started from our own seed generator and the arm started from GxVAEs is 0.11, against a run-to-run standard deviation of 2.91 for the latter. On this metric the search, rather than the population it is given, sets where the run ends up. All four initializations are themselves transcriptome-conditioned, so this experiment varies *which* conditioned generator supplies the population, not *whether* the population is conditioned.

That the outcome is set by the search invites the converse question, which Appendix J takes up: whether the run follows its target condition or a direction the reward would supply regardless.

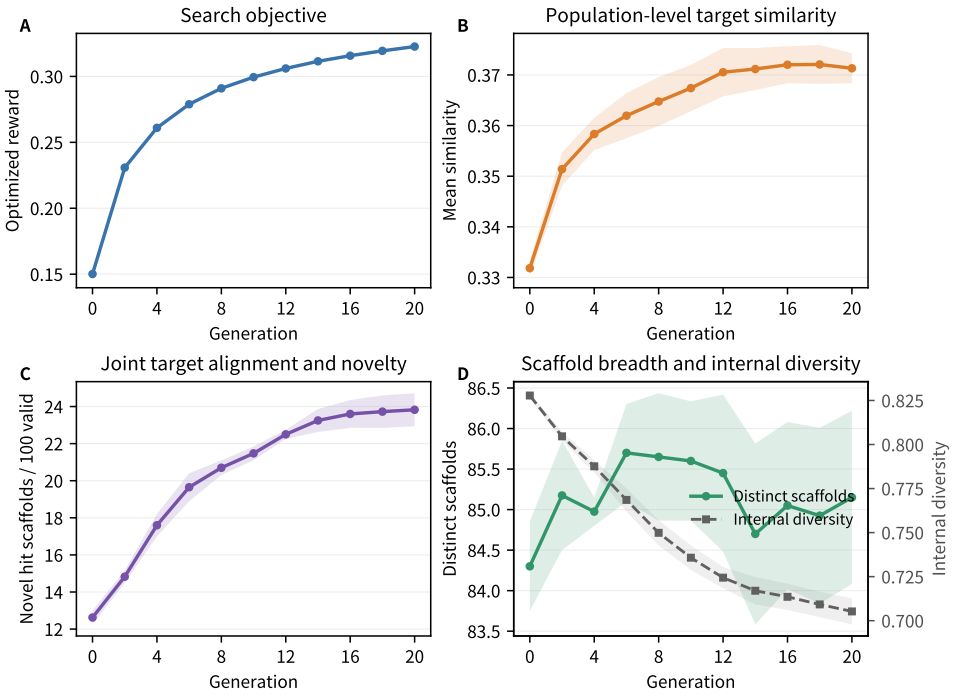

Figure 5: **Knockdown search dynamics under the fixed** 1,500**-candidate budget.** Evolution increases the optimized reward, population-level similarity, and novel hit scaffold yield. The number of distinct scaffolds in the population remains nearly constant, while internal diversity falls from .828 to .705: selection concentrates the population within a scaffold set it does not narrow.

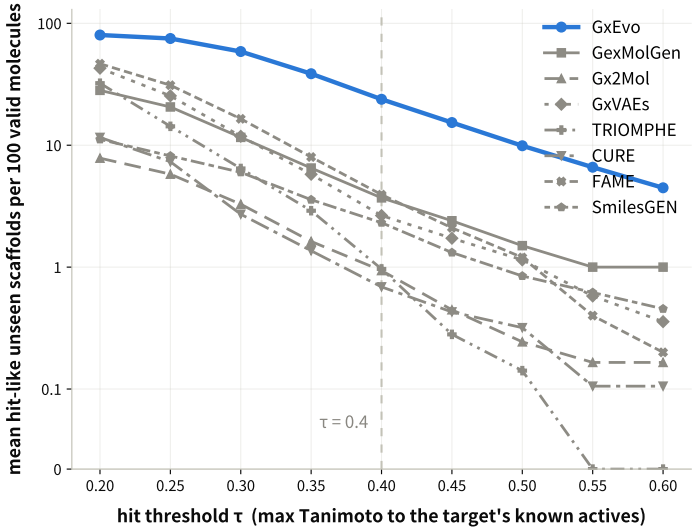

Figure 6: **Sensitivity of novel hit scaffold yield to the hit threshold.** Counts are per 100 valid molecules. The ranking is evaluated over thresholds $\tau \in [0.2, 0.6]$ rather than only the pre-specified value $\tau = 0.4$ used in the main text.

J FUNCTIONAL CONDITION-SPECIFICITY

Table 13 shows that novel hit scaffold yield converges from starting populations that differ by a factor of 13.6, so the search, rather than the population it is handed, sets where a run ends up. That

Table 13: **Evolution from external initial populations.** The same mutation, reward, and survival procedure runs for 20 generations from each baseline population. Novel hit scaffolds follow the definition of Section 4.1: because a transplanted population descends from two libraries – the baseline's corpus supplied generation 0, ours trained the mutation operator – a scaffold counts as unseen only when it is outside both, so each row is scored against the union of the corpora its population descends from. Gx2Mol was retrained on our leakage-filtered corpus, so that union is ours. The GxVAEs row uses the same randomized-target arm as Table 1; its generation-0 novel hit scaffold yield is 2.60 here rather than the 2.64 of Table 1 because the union corpus is the stricter novelty reference. Under the canonical-target arm the row reads .277 → .368 and 0.95 → 23.92. Recomputing every row against the single 631,637-scaffold reference set of Appendix M lowers all four generation-20 yields by 7.7–9.9% and leaves the pattern below unchanged.

	mean similarity			novel hit scaffolds/100		
Initial population	gen. 0	gen. 20	change	gen. 0	gen. 20	change
GxVAEs	.278	.367	+.089	2.60	23.72	+21.12
GexMolGen	.295	.352	+.057	3.70	17.40	+13.70
Gx2Mol	.299	.356	+.057	0.93	20.07	+19.14
GxEvo seed generator	**.332**	**.371**	+.039	**12.64**	**23.83**	+11.19

raises the complementary question of what the search is following. This appendix answers it on the held-out-condition axis, which is where the question can be posed: a condition carries a measured response and a single ground-truth compound, so a molecule's predicted response can be scored against the response it was asked to produce and against other responses measured in the same cell. A knockdown signature has no such per-condition ground-truth compound, so the null ladder below cannot be built on that axis.

Every target-conditioned column of Table 3 compares a generated molecule with the single ground-truth compound of its condition, so a method that proposes a structurally different molecule with the same transcriptional effect is scored as a failure. This appendix reports a test that does not use structure at all. For a molecule m proposed for condition i (cell line c_i, dose d_i) where d_i is the median positive dose among the wells aggregated into condition i, we compute

$$\text{diag} = \text{MSE}(\text{PRnet}(m; c_i, d_i),\ \Delta_i)\,, \qquad \text{off} = \frac{1}{K}\sum_{j\in\mathcal{O}_i} \text{MSE}(\text{PRnet}(m; c_i, d_i),\ \Delta_j)\,,$$

where Δ_i is the *measured* response of condition i and $\mathcal{O}_i$ is a fixed set of $K = 10$ other conditions in the same cell line. Restricting the off-targets to the same cell line cancels the basal state, so the contrast isolates the drug response. The reported quantity is the lift $\text{off} - \text{diag}$. A positive lift means the molecule's predicted response is closer to the response it was asked to produce than to other responses measured in the same cell. The test uses measured held-out responses and an external response model, and therefore shares no term with the fitness used during search, which scores a candidate against the *target* profile through our own forward response model.

A lift is only interpretable against a null ladder, since a molecule could score well merely by being drug-like. We therefore evaluate three reference populations under the identical protocol. They are the condition's own ground-truth drug as the ceiling, another condition's ground-truth drug from the same cell line (GT-PERM), and random ZINC250k molecules as the floor. GT-PERM is the decisive control because it is a real drug placed at the wrong condition. If it scored like the ceiling, the test would be measuring drug-likeness rather than condition-specificity. It does not: on the normalized scale it reaches -2.7%, 2.1% and 13.5% of the floor-to-ceiling range on the three splits (Table 14), so the separation this test reports is condition-specific.

On the raw scale the ground-truth drug attains lifts of $(3.31, 2.82, 2.82) \times 10^{-3}$ on the three splits, while random ZINC molecules attain $(0.59, -0.41, 0.56) \times 10^{-3}$. GxEvo attains $(3.20, 2.55, 2.54) \times 10^{-3}$, which is 96%, 92% and 87% of the way from the random floor to the ground-truth ceiling. GxEvo has the highest normalized lift on the random and cell splits and is within 1.4 points of the best value on the scaffold split. GexMolGen falls below the random floor on the random split and remains close to it on the scaffold and cell splits, even though Table 3 credits it with the highest standard novelty; CURE recovers about half of the floor-to-ceiling range once its condition-dependent

Table 14: **Functional condition-specificity on held-out conditions.** Lift normalized between the two reference populations, $100 \cdot (\text{lift} - \text{lift}_{\text{ZINC}})/(\text{lift}_{\text{GT}} - \text{lift}_{\text{ZINC}})$, over the same evaluation conditions as Table 3 per split and 20 molecules per condition. Because the random-molecule floor differs from zero on the raw lift scale, we report the raw lifts in the text and use the normalized values for cross-split comparison.

	normalized lift (%) ↑		
Method	RANDOM	SCAFFOLD	CELL
GT-PERM (a real drug at the wrong condition)	−2.7	2.1	13.5
TRIOMPHE	42.2	19.0	42.9
GxVAEs	53.0	54.9	41.7
GexMolGen	−5.0	7.1	11.2
Gx2Mol	71.6	76.3	66.6
SmilesGEN	55.0	65.0	53.0
CURE	48.5	49.2	54.1
GxEvo seed generator only	83.6	92.8	83.9
GxEvo independent resampling control	87.4	**93.1**	76.1
GxEvo	**95.9**	91.7	**87.4**

Molecules behind each row. Each condition contributes the first 20 outputs its method returned, of which the ones RDKit parses are scored; no prediction was discarded as non-finite in any row. A method whose validity is below one therefore scores on fewer than the 4,000 molecules the ceiling allows: Gx2Mol on 3,315/3,425/3,279, TRIOMPHE on 3,941/3,878/3,929, GxVAEs on 2,476/2,609/2,456, and SmilesGEN on 3,517/3,561/3,465. Each is near its Table 3 validity times 4,000; the residual difference is that this test reads the first 20 outputs of a condition rather than the whole population. GexMolGen and the three GxEvo rows contribute the full 4,000. CURE (3,801/3,681/3,961) does not follow that arithmetic, and its count is not .443/.430/.533 of 4,000. Its released writer leaves a row blank where a decode fails instead of writing an unparseable string, so the blanks are removed before the 20-per-condition cap applies and every molecule it does contribute parses. What its shortfall counts is the conditions that returned fewer than 20 molecules at all — 25, 49 and 3 of them — not a fraction of its outputs. Its CELL entry additionally runs over 199 conditions, one being absent from CURE's test index, so the ceiling there is 3,980; that same condition gives the CELL entry its own 199-condition ceiling and floor, 2.75 and 0.47×10^{-3}, rather than the 200-condition pair the other rows are normalized against.

input is restored (Appendix D), which places it above TRIOMPHE but below Gx2Mol. The ordering otherwise broadly follows the structural table. We use this test only as a secondary check.

Condition difficulty. Per-condition lift is strongly anti-correlated with per-condition Morgan similarity (Spearman $\rho = -.25, -.30, -.41$ on the three splits). This is a property of the conditions and not of the generated molecules. The ground-truth drug shows the same anti-correlation on its own condition ($-.23, -.24, -.41$) and so do random ZINC molecules ($-.39, -.51, -.48$), and our per-condition lift tracks the ground-truth drug's at $\rho = +.96$ or above on every split. Conditions whose measured responses are hard to tell apart give every population a low lift, so comparisons must be made within a condition. All comparisons reported here are paired by condition. The same effect also bounds what the test can resolve. The ground-truth drug attains a positive lift on only 151, 140 and 134 of the 200 conditions, so on roughly 30% of conditions the instrument cannot separate the correct response even for the correct drug. Restricting to the resolvable conditions leaves the method ordering unchanged.

Operator versus independent resampling. The test also separates the two components of the search. Against the matched independent resampling control, which draws the same number of candidates from the seed generator instead of mutating selected parents, the mutation operator gives a higher mean lift on the random (+0.23) and cell (+0.25) splits and a similar lift on the scaffold split. The same directional pattern appears across the evolutionary configurations we measured: the reported setting and tournament selection of sizes 1 and 4. Relative to the seed population, evolution with the operator improves the random-split lift by 0.33 and leaves it nearly unchanged elsewhere, whereas independent resampling gains 0.10 on random and degrades over generations on the cell split (lift 2.46 at generation 2 to 2.29 at generation 18). These results suggest that the operator helps keep the population on-condition while selection concentrates it.

K CASE-STUDY SELECTION AND INTERPRETATION

Table 1 reports distributions. It does not say what the molecules look like, and the diversity claim in particular is one a reader should be able to inspect: a population can reach a high mean similarity either by returning many variations of one chemotype or by returning several unrelated ones. Figure 3 shows the second case for two knockdown targets.

Each column is one generated molecule above the known active it most resembles, and the three columns of a row are three different scaffolds matched to three different actives.

The molecules are not chosen by eye. They come from the final population of the reported search (generation 20, four search seeds pooled, 400 molecules per target). Every molecule shown has a Bemis–Murcko scaffold and a molecular connectivity absent from our training corpus, so each one is counted as a novel hit scaffold by the metric in Table 1. Within that set we take the highest-similarity molecule subject to four constraints: a scaffold and a nearest active not already used anywhere in the figure, Tanimoto at most .45 to every other molecule shown, and a molecular weight at most 600 with QED at least .30. The constraints cost similarity rather than buy it. Dropping the three distinctness constraints and keeping everything else – that is, taking the three highest-similarity novel-scaffold molecules per target that pass the same property screen – raises the mean similarity from the .58 shown here to .70, but on both targets those three molecules are near-duplicates of one another (pairwise Tanimoto .72–.82 on AKT1, .61–.85 on EGFR) and two of the three match the same known active, which is the opposite of what the figure is drawn to show. The realized similarities are .49 to .75 and the largest pairwise similarity among the six molecules shown is .33. AKT1 and EGFR are used because their inhibitor chemotypes are the most widely documented of the ten, which lets a reader judge the pairs without consulting the active set.

Figures 7 and 8 apply the identical procedure to all ten knockdown targets. The two targets in Figure 3 are those rows of the same selection, so the molecules discussed in Section 4.6 are the molecules shown there. Across the thirty pairs no scaffold and no nearest active is repeated, every generated molecule keeps a scaffold and a connectivity absent from the training corpus, no two generated molecules exceed Tanimoto .44 to each other, and the similarities to the matched actives run from .35 to .75.

Nearest-neighbor similarity and motif preservation are evidence that the generated structures occupy relevant chemical neighborhoods; neither is a prediction of activity. No docking, retrosynthesis, selectivity profiling, or assay evidence is offered for these molecules.

L NOVEL HIT SCAFFOLDS ON HELD-OUT CONDITIONS

Section 4.1 reports novel hit scaffolds for knockdown targets only, because a held-out condition supplies a single ground-truth compound rather than a reference set and $\tau = 0.4$ against one structure is not the same question as $\tau = 0.4$ against a family of actives. The values are given here under both normalizations: per 100 valid outputs, which is the definition used for knockdown, and per 100 generated, which charges a method for the outputs it wastes. The two differ only for methods that return fewer than 100 valid molecules from the draws their population is built from. GxEvo returns a full population of 100 valid molecules, so the two normalizations coincide for it.

GxEvo leads three of the six columns. GxVAEs exceeds it on the SCAFFOLD split under both normalizations and on RANDOM under the first; the paired tests of Table 9 do not separate the two arms on any split. GxVAEs reaches that yield with validity .62–.65, the lowest of any row in Table 3, and is behind GxEvo on Morgan, Fraggle and PRnet-MSE on every split. Counting distinct novel scaffolds rewards a population that spreads, and a method that returns many unusable strings and few close ones can spread further than one that does not.

M NOVEL HIT SCAFFOLDS UNDER A COMMON REFERENCE CORPUS

Novel hit scaffolds are counted against each method's response-paired task corpus, with generic unpaired pretraining collections excluded for every method because equivalent corpora are not available for all baselines (Section 4.1). The reference set therefore differs between methods, and GexMolGen's unpaired pretraining corpus was never released, so its count cannot be placed on the same footing at all.

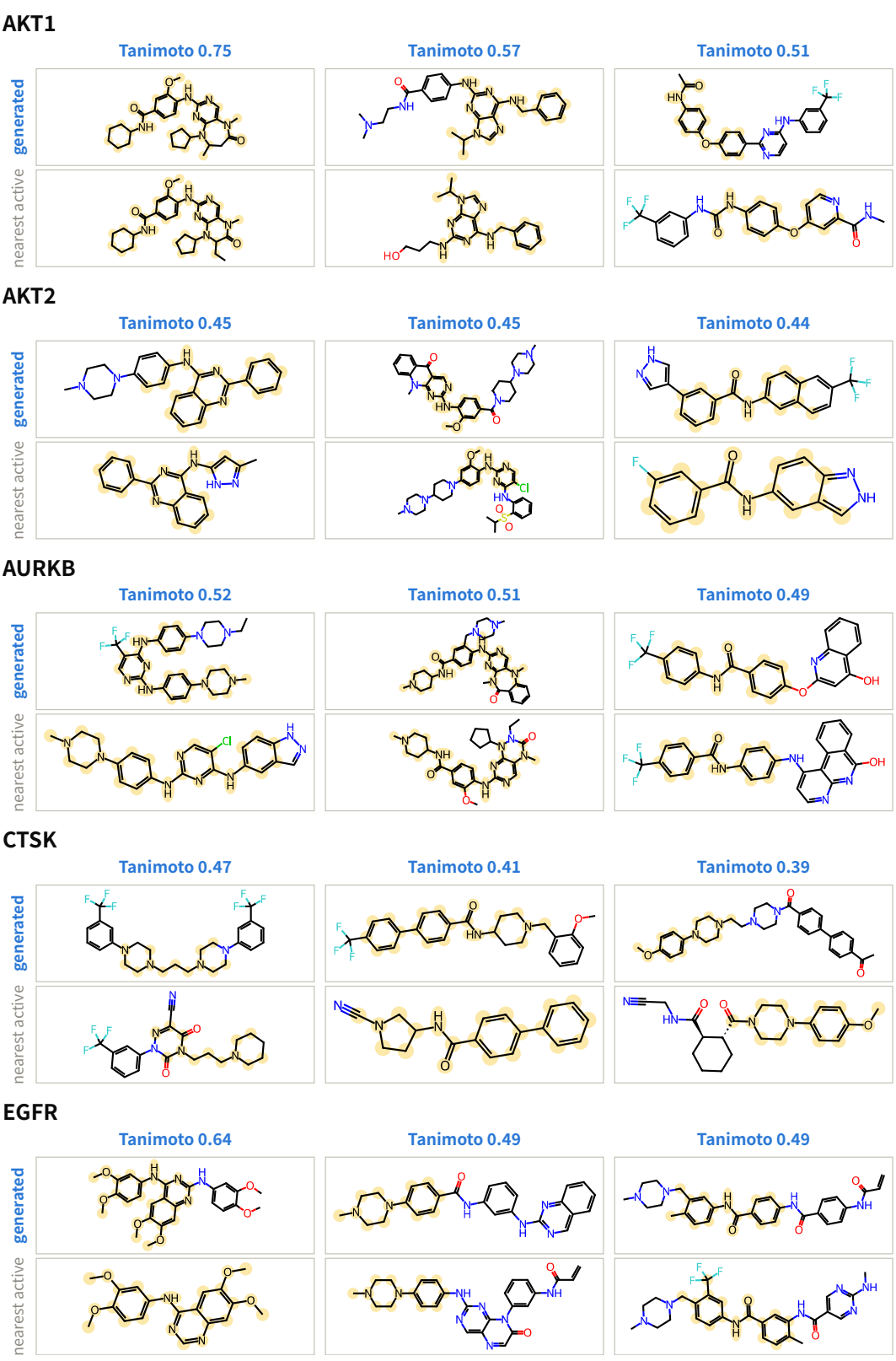

Figure 7: **Case study, targets AKT1–EGFR.** Same selection and same layout as Figure 3, drawn larger: generated molecules above their nearest known actives, with the Morgan Tanimoto similarity of each pair in the heading and the pair's maximum common substructure shaded.

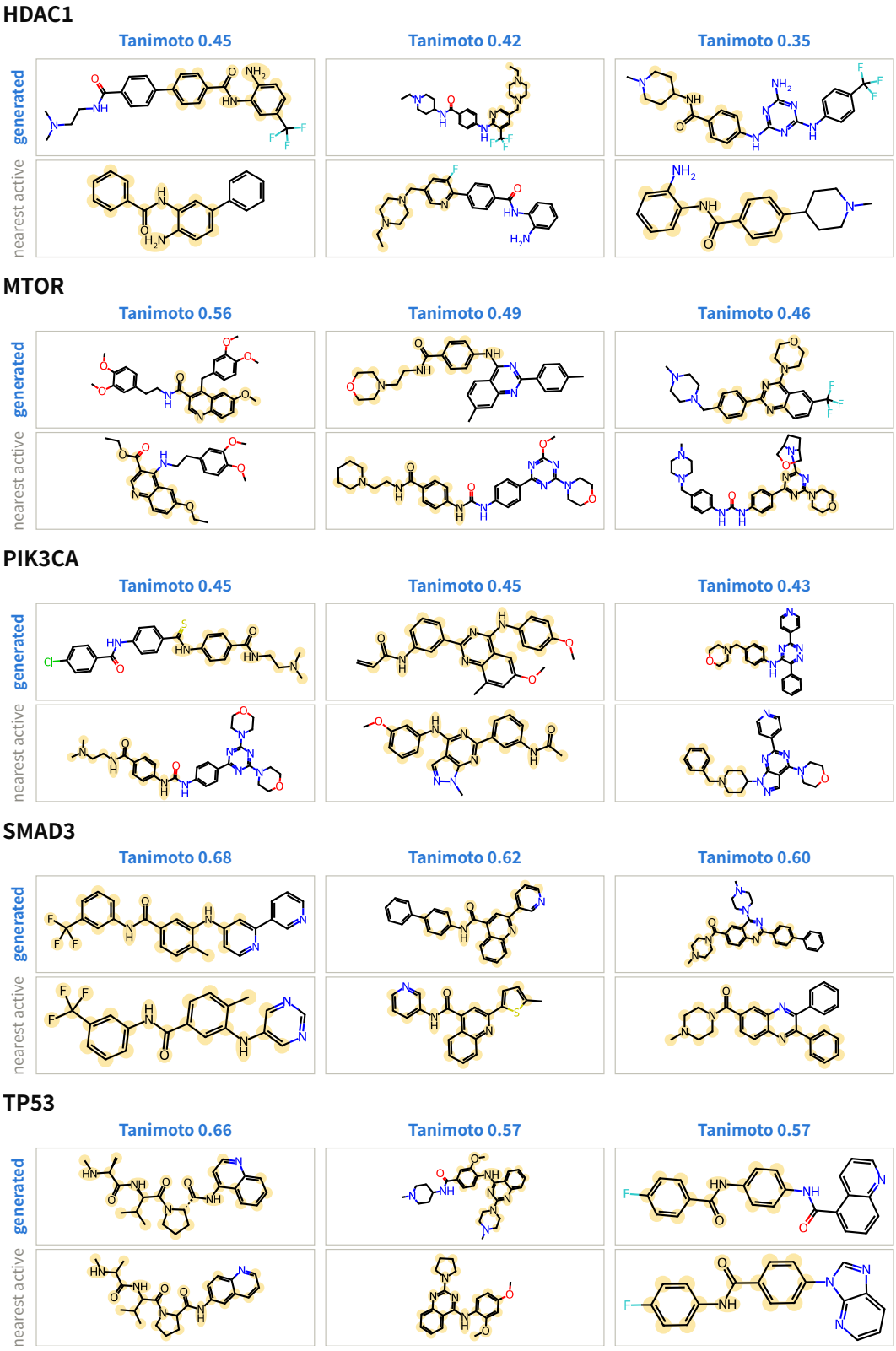

Figure 8: **Case study, targets HDAC1–TP53.** Continuation of Figure 7.

Table 15: **Novel hit scaffolds on held-out conditions**, at the prespecified $\tau = 0.4$, over the same 200 conditions per split. Hit-likeness here is maximum Morgan similarity to the condition's single ground-truth compound.

	per 100 valid			per 100 generated		
Method	RANDOM	SCAFFOLD	CELL	RANDOM	SCAFFOLD	CELL
TRIOMPHE	2.09	0.84	0.82	2.08	0.83	0.82
GxVAEs	**2.88**	**5.04**	1.94	2.15	**3.90**	1.48
GexMolGen	0.61	0.92	0.53	0.61	0.92	0.53
Gx2Mol	1.30	1.54	0.37	1.19	1.45	0.34
SmilesGEN	0.24	0.38	0.11	0.23	0.35	0.10
CURE	0.63	0.52	0.16	0.20	0.29	0.08
GxEvo	2.73	3.50	**2.06**	**2.73**	3.50	**2.06**

Table 16: **Knockdown novel hit scaffold yield under a reference corpus shared by all methods.** Counts are per 100 valid molecules. Only the set of scaffolds treated as already seen changes between the columns.

Method	own task corpus (reported)	shared reference set	change
TRIOMPHE	0.99	0.77	-22.2%
FAME	3.93	2.37	-39.8%
GxVAEs	2.64	1.94	-26.7%
GexMolGen	3.70	2.60	-29.7%
Gx2Mol	0.93	0.65	-30.4%
SmilesGEN	2.05	1.48	-28.1%
CURE	0.66	0.56	-15.4%
GxEvo	**23.83**	**21.60**	-9.3%

We therefore recompute the knockdown yields against one reference set shared by every method: the union of all six task corpora – SmilesGEN was trained on the GxVAEs corpus, so seven methods draw on six – the leakage-filtered ZINC-250k corpus that initializes our seed generator and mutation operator, and the 1,844,948-molecule ChEMBL collection released with FAME, which covers the chemistry GexMolGen's unreleased corpus is drawn from. The union holds 631,637 scaffolds against 3,928–6,069 for the individual task corpora, and contains each method's own corpus, so each recomputed count is that method's reported count restricted to scaffolds outside the union.

Table 16 shows that every method loses scaffolds and GxEvo loses the fewest, 9.3% against 15.4–39.8%, so its generated scaffolds show the least overlap with the shared reference chemistry; the ratio to the strongest prior method rises from 6.06 to 8.31 and no longer depends on a corpus that was not released. We keep the own-corpus definition in Table 1 because it is the pre-specified one and because it is the definition under which a scaffold is unseen *by the method that produced it*.

Widening to ZINC-250k alone isolates the part of our own count that this exclusion could inflate, since that corpus initializes both our models: it removes 0.25 of our 23.83, and the rest of the reduction comes from ChEMBL. The effect sits in the initial population rather than the final one — the seed generator loses 16% under the same widening — so the scaffolds evolution adds are less ZINC-derived than the ones it starts from.

N NON-GENERATIVE FLOORS FOR THE STRUCTURAL METRICS

Table 1 bounds its metrics from below by the weakest published *method*. Table 17 adds the bound from the weakest *molecule*: four sources that no generator produced and that carry no information about the target, scored through the same metric code, at the same $\tau = 0.4$, over the same ten targets. ZINC-250k random is the collection that initializes our seed generator and mutation operator; LINCS task-corpus random is the response-paired corpus; ChEMBL random is drug-like chemistry that is bioactive somewhere but was not selected for these targets; ExCAPE off-target is a known active

Table 17: **Structural metrics for molecule sources that no method generated.** Knockdown targets, $\tau = 0.4$, averaged over the ten targets. Hit rate is the fraction of a population that is hit-like; novel hit scaffolds are counted per 100 valid molecules, against each row's own response-paired task corpus and against the 631,637-scaffold set shared by all methods (Appendix M). The non-generative rows are scored for novelty against our task corpus, the strictest one available to them. Dispersion is in Table 8.

	mean similarity	hit rate	novel hit scaffolds/100	
Source	↑	↑	task corpus	shared set
published generators				
TRIOMPHE	.205	.012	0.99	0.77
FAME	.272	.081	3.93	2.37
GxVAEs	.278	.078	2.64	1.94
GexMolGen	.295	.109	3.70	2.60
Gx2Mol	.299	.137	0.93	0.65
SmilesGEN	.281	.078	2.05	1.48
CURE	.284	.083	0.66	0.56
non-generative sources				
ZINC-250k random	.286	.097	8.08	0.00
LINCS task-corpus random	.286	.085	0.00	0.00
ChEMBL random	.328	.191	16.87	1.32
ExCAPE off-target active	.334	.195	17.18	1.11
GxEvo seed generator	.332	.184	12.64	8.20
GxEvo	**.371**	**.298**	**23.83**	**21.60**

of one of the other nine targets, which makes it the strongest floor on every column but the shared reference set, where random ChEMBL is marginally higher.

GxEvo is the only row above all four floors in every column. Five of the seven published generators are at or below the ZINC-250k floor in mean similarity, and all seven fall below the ChEMBL and off-target floors in mean similarity, in hit rate, and in novel hit scaffolds counted against their own task corpora. The seed generator alone sits at those floors rather than above them — .332 mean similarity and .184 hit rate against .328/.191 for random ChEMBL and .334/.195 for off-target actives — so on these columns it is the search, not the conditioned initialization, that carries the population past chemistry selected for nothing in particular.

The two metric families respond differently. Mean similarity and hit rate carry no novelty condition, so their floors are properties of the chemistry: random bioactive chemistry is hit-like at $\tau = 0.4$ for .191 of its molecules against .097 for random ZINC, so at this threshold these columns are read as a comparison between rows rather than as an absolute level. Novel hit scaffolds carries one, so its floor is a property of the reference set: our task corpus contains neither ChEMBL nor ZINC, which credits a random ChEMBL molecule as novel for free and lifts that source to 16.87. Against the shared reference set of Appendix M, which contains all four sources, the floors fall to 0.00–1.32 while GxEvo retains 21.60. That the LINCS draw scores exactly 0.00 against our own corpus is the internal check that the novelty filter excludes what it is defined to exclude.

The off-target floor is generous to the floor rather than to us, since six of the ten targets are kinases and AKT1 and AKT2 share a chemotype. Per-target values for all four sources are released with the paper. Table 1 itself is unchanged, on its pre-specified threshold and novelty corpus.

O IDENTITY AUDIT OF FINGERPRINT TIES

Exact SMILES novelty and fingerprint distance disagree when methods handle stereochemistry differently. CURE assigns no stereocenters in our reproduction, although 79.8% of its training molecules carry a stereo annotation, so an output can differ in canonical SMILES while keeping Morgan similarity 1.0 to a training molecule. The two measures differ in two respects at once: the set they run over and the equivalence class they use. Table 18 changes one respect per rung, and the entire gap opens at the stereochemical layer.

Table 18: **Where the reported novelty gap comes from.** Each rung is computed on the same molecules and differs from the rung above it in exactly one respect. The top rung is the standard definition reported in Table 1; the bottom rung is one minus the connectivity overlap reported there.

Membership test	CURE	GexMolGen
canonical SMILES, distinct molecules *(standard novelty)*	1.000	1.000
canonical SMILES, every valid output	1.000	1.000
InChIKey, every valid output	1.000	1.000
InChIKey first block, every valid output	**.356**	**.839**

The same ladder spans less than .02 for TRIOMPHE and GxEvo, whose standard novelty can therefore be read at face value. Deduplication is not harmless in general, but it acts elsewhere: it is what lets CURE report high novelty while emitting about 2.4 copies of each distinct molecule, which Table 1 records as within-condition uniqueness .424.

We report the first InChIKey block rather than the complete key. The complete key carries the stereochemical layer and partitions molecules almost exactly as canonical SMILES does — the two differ by at most .0004 across every arm on both axes — so it restates standard novelty rather than tightening it, and under it a molecule could count as novel while the same table reported Morgan similarity 1.0 to a training compound. Only the first block places membership in the stereo-insensitive class our structural measurements already use. It is also the conservative choice for GxEvo: 44.8% of our own fingerprint ties are distinct stereoisomers of a training molecule, which the first block counts as overlap and canonical SMILES does not — it raises our reported overlap from .031 to .033. This audit concerns measurement resolution, not the biological quality of either method.