PHENTTP: PHENOTYPE-DRIVEN MOLECULAR GENERATION WITH TRANSFERABLE TARGET PRIORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Drug discovery ultimately seeks molecules that elicit desired biological responses. Phenotype-driven generation pursues this functional objective directly from cellular responses without requiring a predefined target. However, the inverse mapping from downstream cellular response to molecular structure is underdetermined. To reduce this ambiguity, we propose PhenTTP, a phenotype-driven molecular generation framework with transferable target priors. PhenTTP uses a state-transition phenotype encoder and proceeds in three stages: it aligns phenotype-derived representations with target-conditioned priors, learns phenotype-conditioned generation from phenotype-molecule pairs, and transfers target-conditioned behavior through dual-route distillation, supporting both phenotype-only and phenotype-plus-target generation. Across chemical, scaffold, cell-line, and target out-of-distribution settings, the phenotype-only route lowers Vina scores by 0.13-0.54 kcal/mol over the best phenotype-only baseline despite receiving no targets at inference; when a target is available, the target-informed route further increases Boltz-2 binder probability by 6.6-18.3 percentage points over the same baselines while maintaining phenotype correlation similar to that of the phenotype-only route and achieving high molecular validity, uniqueness, and novelty. Together, these results show that PhenTTP improves target compatibility by transferring target-conditioned molecular priors while preserving phenotype fidelity, establishing a framework that links cellular functional objectives to molecular-level constraints in generative drug discovery.

1 INTRODUCTION

Generative AI has enabled large-scale molecular design, but drug discovery ultimately seeks molecules that elicit desired biological responses. A desired cellular response specifies what a molecule should achieve, whereas a molecular target provides a hypothesis about how that response may be achieved. However, the relevant target may remain unknown even when the desired response is specified. This limits the applicability of target-conditioned generators, which require explicit target information. Phenotype-conditioned molecular generation addresses this setting by generating molecules directly from gene-expression or cellular-perturbation profiles Mendez-Lucio et al. (2020); Kaitoh & Yamanishi (2021); Das et al. (2023); Li & Yamanishi (2024); XU et al. (2026), without requiring a predefined target.

However, phenotypic agreement alone does not establish the underlying mechanism of action. Similar transcriptomic responses may arise from distinct molecular interactions, including unintended target engagement or nonspecific cellular effects. Conditioning solely on the phenotype leaves these interactions without explicit constraints, limiting control over how the desired response is achieved. This motivates incorporating target information to constrain the molecular solution space while preserving the desired cellular response.

We therefore view target information as a molecular prior that constrains phenotype-compatible solutions, while retaining the phenotype as the primary design objective. Existing phenotype-conditioned generators generally lack an explicit target-conditioned prior, whereas target-conditioned generators require a specified target and typically omit the desired cellular response. XMolRL Guo & Liu (2026) connects the two settings by learning a phenotype-conditioned prior and subsequently refining it for a specified protein target through docking-based reinforcement learning.

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

This demonstrates the value of target information for refining phenotype-conditioned generation, but introduces target knowledge through target-specific optimization rather than transferring it as a reusable prior into phenotype-driven generation itself. It therefore remains unclear how target-conditioned knowledge can constrain phenotype-only generation while also supporting stronger target control when explicit target information is available.

Learning such a transferable prior is challenging because phenotype and target conditions differ in representation and biological meaning, while aligned phenotype-target-molecule triples are substantially less abundant than phenotype-molecule pairs. Here, we introduce PhenTTP, a three-stage target-prior transfer framework for phenotype-driven molecular generation. PhenTTP aligns phenotype-derived representations with target-conditioned priors, learns phenotype-conditioned generation from phenotype-molecule pairs, and transfers target-conditioned behavior through dual-route distillation. Crucially, the resulting generator supports two complementary inference routes: a phenotype-only route that generates directly from a desired phenotype without requiring target information, and a phenotype-plus-target route that additionally incorporates an explicit target when available. In this way, target-conditioned knowledge serves as a transferable prior rather than a mandatory input, constraining the phenotype-compatible molecular space while preserving the cellular objective and enabling optional target-specific control.

Our main contributions are summarized as follows:

- We formulate phenotype-driven molecular generation by treating target-conditioned molecular models as transferable priors, allowing target knowledge to guide generation without requiring target information at inference.
- We develop PhenTTP, a three-stage framework that aligns phenotype-derived representations with target-conditioned priors, learns phenotype-conditioned generation from phenotype-molecule pairs, and transfers target-conditioned behavior through dual-route distillation, supporting both phenotype-only and phenotype-target generation.
- We show that transferred target priors improve target compatibility without target input, while explicit targets provide further affinity gains alongside phenotype consistency across chemical, scaffold, cell-line, and target OOD settings.

2 Related Work

Target-conditioned Molecular Generation. Target-conditioned generation designs ligands for a predefined protein using either three-dimensional structure or sequence-derived representations. Structure-based methods such as Pocket2Mol Peng et al. (2022) and TargetDiff Guan et al. (2023) generate molecules from binding pockets, while DrugGPT Li et al. (2023), Prot2Drug Creanza et al. (2025), and ProtoBind-Diff Mistryukova et al. (2025) condition on protein sequences or embeddings. Despite different target representations, these methods share a common assumption: the target is known at inference. They therefore provide target-specific molecular control, but target-level conditioning alone does not ensure that the generated molecules induce the desired downstream cellular response.

Phenotype-conditioned Molecular Generation. Phenotype-conditioned methods instead generate molecules directly from cellular-response measurements, particularly transcriptomic perturbation profiles. Early work established gene-expression-to-molecule generation Mendez-Lucio et al. (2020), followed by methods using variational modeling, cross-modal alignment, gene interactions, and language models Li & Yamanishi (2025); Cheng et al. (2024); Koo et al. (2025); Jiang et al. (2026). Recent approaches further model cellular-state transitions or maintain phenotype guidance throughout generation XU et al. (2026); Xu et al. (2026). These methods make the desired cellular response the primary conditioning signal and do not require a predefined target, but phenotype conditioning alone leaves the molecular solution space underdetermined at the target level.

Connecting Phenotypes and Molecular Targets. Several studies connect transcriptional phenotypes with molecular targets through shared representations or target-specific refinement. TRIOMPHE Kaitoh & Yamanishi (2021) and MolGene-E Ohlan et al. (2025) relate transcriptomic, molecular, and target-associated information, whereas XMolRL Guo & Liu (2026) and HiTGen Yuan

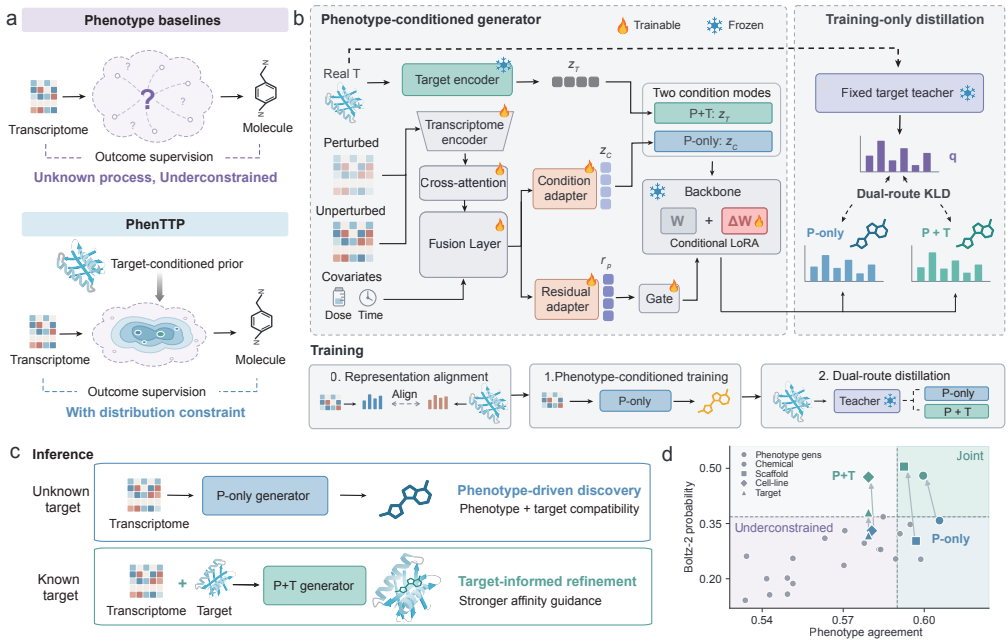


Figure 1: **Overview of PhenTTP. a.** Phenotype-to-molecule mapping is underdetermined because similar cellular responses can arise through different targets. **b.** PhenTTP transfers target-conditioned priors through representation alignment and dual-route distillation. **c.** It supports phenotype-only generation without target input and phenotype-plus-target generation when a target is available. **d.** Phenotype agreement (PRNet r_Δ) against Boltz-2 P_{bind} on the four OOD splits (Table 1). Gray circles are phenotype-conditioned generators. For P-only and P+T, marker shape indicates the split.

et al. (2026) refine phenotype-guided generation toward specified targets. These approaches either connect modalities or introduce target information through target-specific optimization. PhenTTP instead transfers a reusable target-conditioned molecular prior into phenotype-driven generation, allowing target knowledge to benefit phenotype-only generation and provide additional control when a target is available.

3 Method

3.1 Problem formulation

Phenotype-driven molecular generation aims to generate molecules that induce a desired cellular state. Let P denote a perturbational phenotype comprising cellular response measurements and experimental context, and let M denote a molecule. The objective is the conditional distribution $p_\theta(M \mid P)$, whose samples should reproduce the desired outcome. This inverse mapping is underconstrained: a phenotype reflects the consequence of perturbation rather than its molecular cause, and chemically distinct compounds acting through different targets can induce similar transcriptional responses (Figure 1a). Phenotype conditioning can therefore admit multiple molecular explanations of the same outcome, leaving the phenotype-compatible chemical space weakly constrained at the molecular level.

We use an optional protein target T to instantiate a molecular prior over this space, rather than as a substitute condition for P. The phenotype remains the functional objective, while target information narrows the set of molecular explanations consistent with that objective. When T is available, the generator realizes $p_\theta(M \mid P, T)$, concentrating the phenotype-compatible distribution toward candidates consistent with the specified target. When T is unavailable, the same generator realizes $p_\theta(M \mid P)$, with target-conditioned knowledge transferred during training to constrain the otherwise underdetermined distribution.

3.2 FRAMEWORK OVERVIEW

PhenTTP is organized as in Figure 1b. A phenotype encoder maps the cellular response to a latent z_P. A split projection then produces a target-space prompt z_C and a residual condition r_P. A frozen target-conditioned generator supplies $p_0(M \mid T)$. A specified target occupies the native prompt z_T; otherwise z_C occupies that slot. The residual r_P remains active in both routes. Training proceeds in three stages: alignment of z_C with frozen encodings z_T, phenotype-conditioned generation from pairs, and dual-route distillation on triples.

3.3 LEARNING PHENOTYPE RESPONSE REPRESENTATIONS

For target priors to constrain generation meaningfully, the phenotype representation must capture the perturbation-induced response rather than a raw cellular snapshot. A post-treatment profile alone conflates the perturbation effect with the pre-existing cellular state, while a cell-line identifier can become a shortcut for memorizing training contexts. We therefore represent each phenotype as a state transition measured under its experimental conditions.

For each experiment, we use the unperturbed transcriptome X_{pre}, the perturbed transcriptome X_{post}, their difference $X_\Delta = X_{\mathrm{post}} - X_{\mathrm{pre}}$, and covariates c describing dose and time. A single phenotype encoder E_P, shared by both generation routes, maps these inputs to a latent z_P. The two transcriptomic states are first mapped by view-specific projections and then coupled by bidirectional cross-attention,

$$\begin{aligned} H_{\mathrm{pre}} &= \mathrm{Proj}_{\mathrm{pre}}(X_{\mathrm{pre}}), & H_{\mathrm{post}} &= \mathrm{Proj}_{\mathrm{post}}(X_{\mathrm{post}}), \\ \tilde{H}_{\mathrm{pre}} &= \mathrm{XAttn}(H_{\mathrm{pre}}, H_{\mathrm{post}}), & \tilde{H}_{\mathrm{post}} &= \mathrm{XAttn}(H_{\mathrm{post}}, H_{\mathrm{pre}}). \end{aligned} \tag{1}$$

Pooled features of the updated states are then fused with an encoding of the expression change and an embedding of the experimental context,

$$z_P = F\Big(\mathrm{Pool}(\tilde{H}_{\mathrm{pre}}), \mathrm{Pool}(\tilde{H}_{\mathrm{post}}), E_\Delta(X_\Delta), e(c)\Big). \tag{2}$$

We omit a cell-line identity token, so cellular context is conveyed only through the measured states. Architectural details are given in Appendix A.

3.4 ADAPTING A TARGET-CONDITIONED MOLECULAR PRIOR

A pretrained target-conditioned generator supplies a molecular prior $p_0(M \mid T)$ and a target encoding $z_T = E_T(T)$. End-to-end fine-tuning would risk overwriting this prior, so we freeze both E_T and the generation backbone θ_0 and introduce phenotype information through two interfaces produced by a split projection,

$$(z_C, r_P) = A(z_P), \tag{3}$$

where $z_C = A_C(z_P)$ is a prompt in the native target-embedding space and r_P is a residual condition. The residual Δh is a phenotype-dependent increment added to each frozen decoder state h, so phenotype information enters without rewriting the backbone:

$$\Delta h(h, r_P) = \sigma(g(h, \bar{r}_P)) \odot W_r \bar{r}_P, \qquad h' = h + \Delta h(h, r_P), \tag{4}$$

where σ is the sigmoid, g is a learned gating network, and $\bar{r}_P$ is a pooled summary of r_P. Zero-initializing W_r and the auxiliary low-rank updates make the residual pathway inactive at initialization, recovering the frozen generator. We additionally use gated low-rank updates in the last two decoder layers; masking and layer details are given in Appendix A.2.

The residual condition r_P is derived solely from z_P and is shared by both inference routes. When a protein is specified, z_T already enters through the native target prompt; because r_P does not take z_T as input, the pretrained target condition remains distinct from phenotype-induced adaptation. The trainable phenotype modules are collected in ϕ, the encoder E_P, the split projection A, and the residual adapters, while the backbone parameters θ_0 remain frozen. Writing $\theta = (\theta_0, \phi)$ and $G_{\theta_0,\phi}$ for the resulting generator, the two routes are

$$\begin{aligned} p_\theta(M \mid P) &= G_{\theta_0,\phi}(M \mid z_C,\, r_P), \quad z_C = A_C(z_P), \\ p_\theta(M \mid P, T) &= G_{\theta_0,\phi}(M \mid z_T,\, r_P), \quad z_T = E_T(T). \end{aligned} \tag{5}$$

They share the residual pathway and differ only in whether the native target slot is filled by a phenotype-derived prompt or by an explicit target.

3.5 THREE-STAGE TRAINING

The two routes are trained in three stages on phenotype–molecule pairs $\mathcal{D}_P$ and phenotype–target–molecule triples $\mathcal{D}_{P,T}$.

We first align z_C with the frozen encodings z_T on target-annotated phenotypes,

$$\mathcal{L}_{\text{align}} = \mathcal{L}_{\cos} + \mathcal{L}_{\text{NCE}} + \tfrac{1}{2}\mathcal{L}_{\text{margin}}, \tag{6}$$

which pulls matched phenotype-derived and target encodings together and separates mismatched targets. Residual adapters are not updated in this stage, so z_C can occupy the native target slot before generation is trained.

After alignment, we train the phenotype encoder, the split projection, and the residual adapters on $\mathcal{D}_P$, keeping θ_0 frozen:

$$\mathcal{L}_{\text{gen}} = -\log p_\theta(M \mid P) + \lambda_{\text{ctr}}\mathcal{L}_{\text{ctr}} + \lambda_{\text{mom}}\mathcal{L}_{\text{mom}}. \tag{7}$$

The first term is teacher-forced molecular likelihood on the P-only route. The contrastive term prefers the paired phenotype over a context-matched negative, and the moment term keeps z_C in the distributional range of native target encodings. This stage establishes phenotype-conditioned generation from the more abundant pair supervision.

Pair supervision does not explicitly transfer target-conditioned behavior into the phenotype-only route. For each triple, disabling r_P gives the frozen teacher $q_{T,t}(\cdot) = p_0(\cdot \mid m_{<t}, T)$ under teacher forcing. The teacher is not updated, so the transfer is unidirectional. Let $d_{r,i}$ denote the mean teacher-to-student token-level KL for sample i on route $r \in \{P, (P,T)\}$. For a batch of B triples, let $\bar{d}_P$ average the P-only KL over samples with comparable molecule tokens. We optimize

$$\mathcal{L}_{\text{proj}} = \tfrac{1}{2}\big[\bar{d}_P - \varepsilon\big]_+^2 + \frac{1}{2B}\sum_{i=1}^{B}\big[d_{P,T,i} - \varepsilon\big]_+^2. \tag{8}$$

The first term transfers target-conditioned behavior into $p_\theta(M \mid P)$ even though T is withheld from the student. The second regularizes $p_\theta(M \mid P, T)$ toward the same prior while retaining phenotype conditioning through the shared residual pathway. The P-only term applies the hinge after batch averaging, whereas the P+T term applies it before averaging. The slack ε removes penalties below the corresponding KL threshold; token masking is specified in Appendix A.3.

The three stages optimize

$$\mathcal{L}^{(1)} = \mathcal{L}_{\text{align}}, \qquad \mathcal{L}^{(2)} = \mathcal{L}_{\text{gen}}, \qquad \mathcal{L}^{(3)} = (1-\rho)\,\mathcal{L}_{\text{triple}} + \rho\,\mathcal{L}_{\text{replay}}. \tag{9}$$

$\mathcal{L}_{\text{triple}}$ combines P+T molecule likelihood, phenotype and target-switch contrasts, alignment terms, and $\mathcal{L}_{\text{proj}}$. $\mathcal{L}_{\text{replay}}$ continues the P-only pair objective and is excluded from distillation. Inputs, labels, and the expansions of every term are given in Appendix A.3.

3.6 DUAL-ROUTE INFERENCE

At inference, a single checkpoint realizes both routes in equation 5 (Figure 1c). When no target is specified, the native slot is filled by z_C and the generator samples from $p_\theta(M \mid P)$. When a target is available, z_T replaces z_C while the same residual condition r_P remains active, realizing $p_\theta(M \mid P, T)$. Switching routes therefore changes only the source of the target-space prompt; it does not update parameters or require target-specific optimization. Target knowledge acts as a transferred prior in phenotype-only generation and provides additional target-specific control when explicitly available. Figure 1d places both routes against phenotype-conditioned generators on phenotype agreement and Boltz-2 P_{bind}, with one marker for each of the four distribution shifts. From P-only to P+T, phenotype agreement stays similar, and Boltz-2 P_{bind} rises on every split.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Data. We train and evaluate on LINCS L1000 landmark profiles (978 genes) Subramanian et al. (2017). Each condition provides X_{pre}, X_{post}, X_Δ, dose, time, and the molecule; a subset also has

Table 1: L1000 generation on four OOD splits. Phenotype metrics and VUN use $N = 2{,}000$ requests; affinity uses a fixed 100-request panel. Result style: **best**, second-best. Gray cells mark a metric shortfall.

Method	PRNet		CRISP		MAP		Molecular quality				Affinity	
	MSE↓	r_Δ ↑	MSE↓	r_Δ ↑	MSE↓	r_Δ ↑	VUN↑	QED↑	SA↓	FCD↓	Vina↓	P_{bind} ↑
Chemical OOD ($N = 2{,}000$, $n_{\text{chem}} = 485$)												
GexMolGen	10.36	0.549	17.87	0.391	11.32	0.431	5.7%	0.511	3.81	79.83	-7.28	0.202
Gx2Mol	13.62	0.247	15.29	0.225	13.81	0.222	31.7%	0.578	3.07	21.04	-7.72	0.280
CURE	11.54	0.493	11.43	0.529	10.79	0.474	90.8%	0.188	4.87	43.15	-4.95	0.237
GEMGen	9.78	0.568	12.60	0.532	10.24	0.516	74.0%	0.525	2.84	17.43	-7.67	0.348
XMolRL	10.34	0.517	13.49	0.473	10.67	0.470	47.4%	0.580	2.99	17.54	-7.58	0.282
PhenTTP (P-only)	**9.59**	**0.578**	**10.98**	**0.573**	**9.82**	**0.548**	**92.8%**	**0.636**	**2.65**	17.98	**-8.26**	0.358
PhenTTP (P+T)	9.97	0.557	11.28	0.552	10.07	0.531	90.5%	0.604	2.81	**16.78**	-8.12	**0.480**
Scaffold OOD ($N = 2{,}000$, $n_{\text{chem}} = 232$)												
GexMolGen	10.82	0.543	18.65	0.385	11.80	0.424	6.5%	0.513	3.79	82.67	-6.95	0.156
Gx2Mol	15.04	0.125	15.66	0.115	15.14	0.115	20.1%	0.580	2.97	25.43	-7.03	0.253
CURE	11.95	0.496	12.09	0.529	11.15	0.472	90.2%	0.179	4.99	52.43	-4.78	0.180
GEMGen	10.29	0.566	13.72	0.512	10.82	0.509	75.0%	0.542	2.83	**21.60**	-7.68	0.322
XMolRL	10.64	0.526	13.99	0.478	11.16	0.468	45.9%	0.577	3.03	24.49	-7.64	0.254
PhenTTP (P-only)	**10.09**	**0.570**	**11.73**	**0.558**	**10.29**	**0.539**	**92.8%**	**0.625**	**2.68**	22.16	-8.00	0.303
PhenTTP (P+T)	10.26	0.555	12.07	0.538	10.39	0.522	90.8%	0.614	2.75	21.87	**-8.10**	**0.505**
Cell-line OOD ($N = 2{,}000$, $n_{\text{chem}} = 884$)												
GexMolGen	10.50	0.549	17.22	0.397	11.09	0.447	6.3%	0.511	3.80	78.42	-7.04	0.158
Gx2Mol	10.70	0.513	13.45	0.462	10.79	0.462	14.7%	0.549	3.02	7.16	-7.37	0.368
CURE	12.01	0.471	12.71	0.475	11.39	0.432	90.5%	0.168	5.03	41.26	-5.12	0.232
GEMGen	10.34	0.547	13.91	0.473	10.76	0.479	71.4%	0.535	2.89	7.50	-7.83	0.331
XMolRL	11.40	0.453	14.19	0.398	11.60	0.408	71.2%	0.594	3.02	7.00	-7.33	0.233
PhenTTP (P-only)	**10.21**	**0.552**	**12.18**	**0.520**	**10.35**	**0.514**	**92.8%**	**0.608**	**2.67**	7.93	-7.96	0.331
PhenTTP (P+T)	10.29	0.539	12.44	0.499	10.42	0.499	91.0%	0.589	2.84	**6.96**	**-8.19**	**0.476**
Target OOD ($N = 2{,}000$, $n_{\text{chem}} = 467$)												
GexMolGen	11.37	0.534	18.99	0.374	11.89	0.429	5.7%	0.505	3.81	82.92	-6.82	0.141
Gx2Mol	14.21	0.229	15.53	0.205	14.35	0.204	12.2%	0.561	2.72	23.45	-7.17	0.310
CURE	11.72	0.502	12.49	0.511	11.15	0.480	91.8%	0.280	4.66	43.44	-5.33	0.172
GEMGen	10.63	**0.558**	13.72	0.506	10.84	0.512	72.7%	0.535	2.91	20.46	-7.92	0.318
XMolRL	10.88	0.532	14.13	0.478	11.16	0.476	30.7%	0.568	2.71	20.69	-7.82	0.297
PhenTTP (P-only)	**10.59**	0.549	**12.34**	**0.532**	**10.58**	**0.524**	**91.9%**	**0.630**	**2.66**	21.03	**-8.32**	0.317
PhenTTP (P+T)	10.77	0.543	12.43	0.525	**10.58**	0.521	91.3%	0.594	2.81	**18.98**	**-8.32**	**0.384**

a protein-target annotation. Training uses native L1000 phenotype–molecule pairs and phenotype–target–molecule triples formed by attaching an activity-derived protein target to a subset of those pairs (Appendix C). All comparisons use four precomputed splits that hold out unseen compounds, scaffolds, cell lines, or protein targets. Each split uses a frozen cohort of 2,000 test conditions, one molecule per condition, and no metric-based filtering or reranking. A separate frozen common-10 panel is reserved for Chemical, Scaffold, and Cell-line OOD, with ten conditions per target. The targets are EGFR, DRD2, HTR1B, PTGS2, HRH1, KDR, PPARG, PIK3CA, GSK3B, and HDAC1. Target OOD uses its own frozen panel of ten unseen targets. Within each split, all methods receive the same condition–target evaluation requests.

Metrics. Phenotype consistency is scored by frozen PRNet Qi et al. (2024), CRISP Wang et al. (2026), and MAP Feng et al. (2026) as MSE and Pearson r_Δ of predicted versus observed transcriptional changes; displayed MSE is $\times 10^3$. Molecular quality is VUN, QED Bickerton et al. (2012), SA Ertl & Schuffenhauer (2009), and FCD Preuer et al. (2018). VUN (valid, unique, and novel) counts distinct valid canonical molecules absent from the split-specific training set, divided by all generation requests. Target affinity is AutoDock Vina Trott & Olson (2010) and Boltz-2 P_{bind} Passaro et al. (2025). Formal definitions appear in Appendix D.1. Phenotype consistency and VUN use the full 2,000-request budget; invalid or unscorable outputs receive the corresponding training split's mean no-change MSE and $r_\Delta = 0$, listed for L1000 in Appendix Table 5. QED, SA, and FCD use a shared chemically scorable subset. Affinity means use each method's finite Vina–Boltz-2

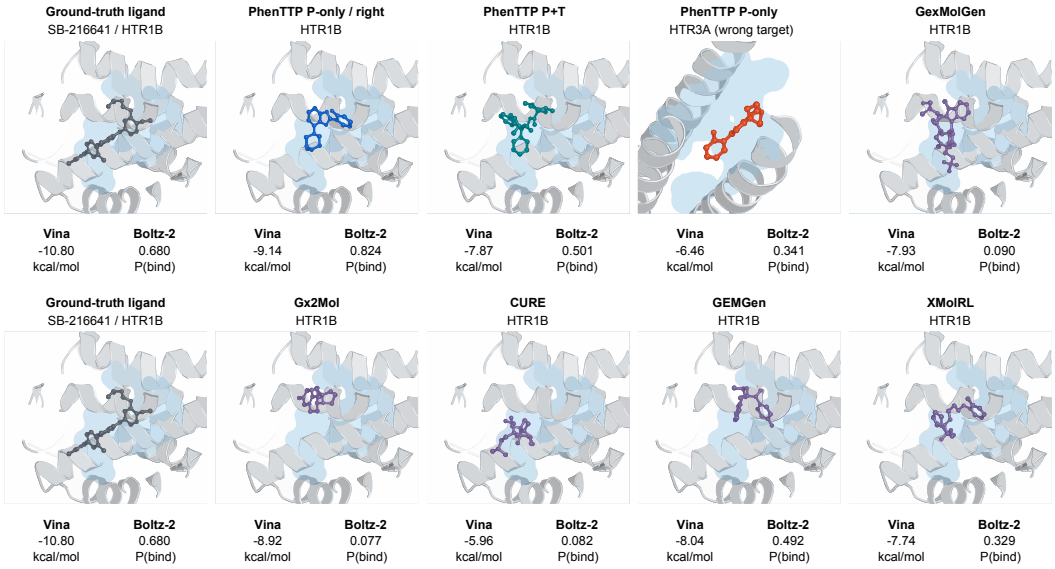


Figure 2: Target preference without target input. The P-only molecule is docked to the annotated target HTR1B and to the negative control HTR3A; P+T, the baselines, and ground-truth SB-216641 are evaluated on HTR1B. Lower Vina and higher P_{bind} are better.

pairs without imputation. Evaluator training is summarized in Appendix D.2. Equal-n scoring is in Appendix D.

Baselines. The phenotype-conditioned baselines are GexMolGen Cheng et al. (2024), Gx2Mol Li & Yamanishi (2025), CURE XU et al. (2026), GEMGen Jiang et al. (2026), and XMolRL Guo & Liu (2026) without its docking-based target RL stage. PhenTTP in these tables is the DrugGPT Li et al. (2023) instantiation of section 3.5, run from one checkpoint as P-only and P+T. Target-only DrugGPT, ProtoBind-Diff Mistryukova et al. (2025), and the matching PhenTTP variants on both backbones are in Appendix Table 8. Implementation details are in Appendix B.

4.2 Phenotype-Driven Generation with Transferable Target Priors

Phenotype conditioning states the desired cellular outcome but leaves the molecular route to it unspecified, and the P-only setting is where that gap is widest, since no protein reaches the model at inference. PhenTTP nonetheless leads on the phenotype objective itself, obtaining the lowest MSE in all 12 predictor–split comparisons and the highest r_Δ in 11 of 12 (Table 1). VUN stays near 92%, so these margins do not come from narrowing onto a small set of reliable molecules. Phenotype-conditioned baselines face a dilemma: methods that match the phenotype closely tend to sacrifice molecular quality or diversity, while those that maintain chemical validity often lack mechanistic grounding. PhenTTP resolves this by transferring a target-conditioned molecular prior during training, which regularizes the otherwise underconstrained phenotype-to-molecule mapping.

Target compatibility is where the phenotype-conditioned baselines fall behind. PhenTTP obtains the best Vina score on all four splits, improving from -7.72 to -8.26 kcal/mol on Chemical OOD, with no protein supplied at inference. Boltz-2 P_{bind} moves less and stays near the baselines, so the effect registers mainly through docking. Nothing at test time identifies a target, which leaves training as the only source of this behavior.

Averages alone cannot separate a genuine target preference from a molecule that docks well against anything. Figure 2 resolves this on an SB-216641-associated Chemical OOD condition, contrasting the annotated target HTR1B (Price et al., 1997) with HTR3A as a non-annotated comparator target. The same P-only molecule reaches -9.14 kcal/mol and $P_{\text{bind}} = 0.824$ on HTR1B against -6.46 and 0.341 on HTR3A, and on HTR1B it surpasses every phenotype-conditioned baseline (Gx2Mol

Table 2: Tahoe generation, macro-averaged over the four OOD splits. P+T uses a target prompt only when an annotation exists. Per-split scores are in Appendix Table 9. P-only ablations are in Appendix Table 10. Result style: **best**, second-best.

Method	PRNet		CRISP		MAP		Molecular quality				Affinity	
	MSE↓	r_Δ ↑	MSE↓	r_Δ ↑	MSE↓	r_Δ ↑	VUN↑	QED↑	SA↓	FCD↓	Vina↓	P_{bind} ↑
XMolRL	6.25	0.139	6.16	0.154	6.18	0.120	26.7%	0.531	3.47	45.93	-6.302	0.321
Gx2Mol	6.47	0.016	6.45	0.019	6.64	0.010	3.2%	0.525	3.61	**42.65**	-4.926	0.274
GexMolGen	6.23	0.059	6.29	0.062	6.81	0.038	11.1%	0.476	3.76	103.96	-5.809	0.238
GEMGen	5.84	0.388	5.50	0.435	6.04	0.319	75.1%	0.524	2.83	53.53	-6.779	0.366
CURE	5.77	0.454	5.39	0.508	10.06	0.278	88.7%	0.302	4.58	52.93	-5.015	0.254
PhenTTP (P-only)	**5.73**	**0.477**	**5.26**	**0.537**	**5.55**	**0.410**	**92.4%**	**0.593**	**2.80**	49.93	-6.432	0.345
PhenTTP (P+T)	5.78	0.452	5.31	0.508	5.60	0.389	87.7%	0.572	2.88	45.59	**-7.226**	**0.573**

-8.92; GEMGen 0.492). The transferred prior therefore acts on the annotated target within a single generation, and does not simply raise affinity across receptors.

4.3 Explicit Target Conditioning Refines the Transferred Prior

A target-conditioned prior should sharpen generation once the protein is supplied, without overriding what the phenotype has already fixed. The same checkpoint takes the annotation directly, with no retraining and no per-target optimization. P_{bind} rises on all four splits, from 0.358 to 0.480 on Chemical OOD, 0.303 to 0.505 on Scaffold OOD, 0.331 to 0.476 on Cell-line OOD, and 0.317 to 0.384 on Target OOD (Table 1). The smallest gain falls on Target OOD, whose proteins are never seen during training; without a transferred prior to refine, the explicit prompt cannot compensate for the absence of target-specific supervision.

Vina does not track P_{bind}. It agrees on Scaffold and Cell-line OOD, stays flat on Target OOD, and is marginally worse on Chemical OOD, so we report both throughout. FCD improves on every split, most clearly from 21.03 to 18.98 on Target OOD, indicating that the explicit target narrows the chemical distribution toward reference ligands rather than expanding exploration.

The phenotype pays little for this. PRNet r_Δ drops by at most 0.021. P_{bind} and FCD move together while phenotype metrics remain stable, which supports the interpretation that the target selects among phenotype-compatible solutions rather than redefining the cellular objective. The two shifts are one restriction seen twice. A higher binder probability and a lower FCD both say that fewer molecules remain admissible, and that those which remain sit closer to real target-conditioned ligands. The same ordering holds on a ProtoBind-Diff backbone (Appendix Table 8), so the behavior follows from the formulation and not from the DrugGPT instantiation. A single checkpoint therefore serves both regimes.

4.4 Generalization to Single-Cell Perturbations

Furthermore, we repeat the four splits on Tahoe single-cell perturbations, using P+T only where an annotation exists (Table 2; per-split scores in Appendix Table 9). This instantiation leaves the residual adapter off. P-only again leads the phenotype objective on all four splits, keeps VUN above 91%, and holds the highest QED among the phenotype-conditioned methods. Docking is the exception. The four-split average favors GEMGen (-6.779 vs. -6.432), because P-only falls behind on Chemical OOD at -3.460 against -6.926; P-only takes the best Vina score on Scaffold, Cell-line, and Target OOD. The Chemical OOD weakness suggests that implicit target priors transfer poorly to assays where protein annotations are sparse or misaligned; phenotype transfer remains intact, but docking gains require named targets.

Adding the target lifts P_{bind} everywhere, from 0.345 to 0.573 on average across the four splits. On Chemical OOD, P_{bind} rises from 0.400 to 0.497 and Vina recovers from -3.460 to -7.317, confirming that explicit target conditioning acts as a refinement step rather than a fallback: where the unnamed prior is weak, naming the target restores compatibility. FCD tightens alongside affinity, from 49.93 to 45.59 on average, indicating that the target pulls the generated distribution toward reference ligands.

Table 3: L1000 component analysis under P-only inference. Evaluation metrics and data follow Table 1. A checkmark indicates an enabled training component; a dash indicates its removal. The shaded row is the full model. Per-split scores, including MSE, are in Appendix Table 6.

P–T alignment	Dual KD	PRNet $r_\Delta \uparrow$	CRISP $r_\Delta \uparrow$	MAP $r_\Delta \uparrow$	VUN↑	QED↑	SA↓	FCD↓	Vina↓	$P_{\text{bind}} \uparrow$
✓	✓	0.562	0.546	0.531	92.6%	0.625	2.67	17.28	-8.14	0.327
✓	–	0.544	0.523	0.512	83.3%	0.652	2.71	17.34	-7.32	0.284
–	✓	0.564	0.543	0.531	93.2%	0.631	2.66	17.37	-8.15	0.329
–	–	0.526	0.501	0.495	85.9%	0.644	2.65	17.47	-7.55	0.301

Target OOD is where the joint route strains. PRNet r_Δ falls from 0.484 to 0.395, and VUN from 93.2% to 76.0%. On unseen proteins, the explicit prompt has no transferred prior to refine, so generation is drawn toward the target at the expense of the phenotype. P-only remains stable in this regime, which confirms that prior transfer occurs during training and not at inference by naming a protein. The dual-route architecture therefore adapts to what is available: when the target is unknown or unseen, phenotype transfer alone suffices; when the target is named and has been supervised, the explicit route sharpens compatibility without requiring model retraining. P-only ablations are in Appendix Table 10.

4.5 Ablation of transferable framework

To transfer a target-conditioned molecular prior to phenotype-only generation, PhenTTP combines P–T alignment at the representation level with dual knowledge distillation at the generation level. Alignment maps the phenotype-derived prompt into the native target-embedding space, whereas dual KD transfers the frozen teacher's target-conditioned generation behavior to the student. Compared with disabling both components, the full model increases PRNet r_Δ from 0.526 to 0.562 and VUN from 85.9% to 92.6%, while improving Vina from -7.55 to -8.14. Per-split results are in Appendix Table 6. This ablation focuses on the effect of target prior transfer on phenotype generation. Table 3 follows the same evaluation metrics and experimental data as Table 1.

Without dual KD, enabling alignment improves phenotype agreement across all three evaluators. PRNet, CRISP, and MAP r_Δ increase from $0.526/0.501/0.495$ to $0.544/0.523/0.512$, respectively, relative to disabling both components. This comparison supports the value of representation-level target supervision for phenotype-conditioned generation. However, VUN decreases from 85.9% to 83.3%, and neither affinity metric improves. Alignment alone therefore improves phenotype agreement without ensuring that target compatibility carries over to the generated molecules.

Adding dual KD to the aligned model further improves all three phenotype correlations and raises VUN from 83.3% to 92.6%. Vina improves from -7.32 to -8.14, and P_{bind} increases from 0.284 to 0.327, supporting direct supervision of generation behavior beyond representation alignment. Dual KD alone also approaches the full model across the reported metrics, with several scores slightly better. This pattern is consistent with partially overlapping supervision rather than strictly additive benefits. Overall, alignment provides a useful representation-level learning signal, while dual KD extends prior transfer to improved target compatibility under phenotype-only inference.

5 Conclusion

We presented PhenTTP, which transfers the molecular prior of a pretrained target-conditioned generator into phenotype-driven molecular generation through representation alignment and dual-route distillation. Across four distribution shifts, two inference regimes, and a second perturbation dataset, it improves both phenotype agreement and predicted target compatibility without a protein at inference, concentrates that improvement on the annotated target, and gains further target control at a small cost in phenotype agreement once a protein is supplied. Therefore, PhenTTP shows that target-conditioned molecular priors can be inherited while the cellular response remains the objective, bridging cellular function and molecular constraint in generative drug discovery.

6 Ethics Statement

This work is computational. We train and evaluate generative models on publicly available transcriptomic perturbation datasets (LINCS L1000 and Tahoe) and score proposed molecules with frozen response predictors and in silico affinity estimators. No experiments involving pathogens, human subjects, or animals were performed, and no molecule proposed here has been synthesized or assayed.

7 Reproducibility Statement

Code and trained checkpoints will be released. Appendix A specifies the phenotype encoder, adapters, and three-stage training. Appendix B lists optimizer settings, decoding hyperparameters, and hardware. Appendix D documents metric definitions, equal-n scoring, invalid-output handling, and the frozen affinity panel. Appendix E reports per-split L1000 ablations and additional variants. Appendix F records the alternative ProtoBind-Diff backbone. Appendix G reports per-split Tahoe results and P-only ablations.

AI Use Statement

In this work, we used generative AI tools to polish the manuscript for readability. We have reviewed all AI-assisted work. We have not used generative AI tools to generate synthetic datasets, develop theoretical or conceptual frameworks, formulate mathematical claims, assist with proofs, propose or refine hypotheses, design or give feedback on methodology or experiments, implement methods, translate text, clean or reformat datasets, perform qualitative or thematic analysis, or interpret results; the remaining tasks that require disclosure are not applicable to this work. The authors checked the numerical results, scientific claims, method descriptions, and final interpretations against the experiments and the literature. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

References

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. doi:10.1038/nchem.1243.

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. doi:10.1093/bib/bbae525.

Teresa Maria Creanza, Domenico Alberga, Cosimo Patruno, Giuseppe Felice Mangiatordi, and Nicola Ancona. Transformer decoder learns from a pretrained protein language model to generate ligands with high affinity. *Journal of Chemical Information and Modeling*, 65(3):1258–1277, 2025. doi:10.1021/acs.jcim.4c02019.

Dibyajyoti Das, Broto Chakrabarty, Rajgopal Srinivasan, and Arijit Roy. Gex2sgen: designing drug-like molecules from desired gene expression signatures. *Journal of chemical information and modeling*, 63(7):1882–1893, 2023. doi:10.1021/acs.jcim.2c01301.

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. doi:10.1186/1758-2946-1-8.

European Molecular Biology Laboratory. ChEMBL database release 37. European Bioinformatics Institute, May 2026. URL `https://doi.org/10.6019/CHEMBL.database.37`.

Jinghao Feng, Ziheng Zhao, Xiaoman Zhang, Mingfei Liu, Jingyi Chen, Xingran Quan, Boyang Fu, Jian Zhang, Yanfeng Wang, Ya Zhang, and Weidi Xie. MAP: A knowledge-driven framework for predicting single-cell responses for unprofiled drugs. *bioRxiv*, 2026. doi:10.64898/2026.02.25.708091.

Michael K. Gilson, Tiqing Liu, Michael Baitaluk, George Nicola, Linda Hwang, and Jenny Chong. BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Research*, 44(D1):D1045–D1053, 2016. doi:10.1093/nar/gkv1072.

Jiaqi Guan, Wesley Wei Qian, Xingang Peng, Yufeng Su, Jian Peng, and Jianzhu Ma. 3d equivariant diffusion for target-aware molecule generation and affinity prediction. In *International Conference on Learning Representations*, 2023.

Haotian Guo and Hui Liu. Bridging the phenotype-target gap for molecular generation via multi-objective reinforcement learning. *Bioinformatics*, 42(1):btag242, 2026. doi:10.1093/bioinformatics/btag242.

Qun Jiang, Xing Ye, Zekun Guo, Yingce Xia, Zequn Liu, Jiaying Xu, Peiran Jin, Fusong Ju, Huanhuan Xia, Haiguang Liu, Tao Qin, Pan Deng, and Sida Shao. Phenotype-guided in silico molecular generation using large language models. *bioRxiv*, 2026. doi:10.64898/2026.01.03.697483.

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. doi:10.1021/acs.jcim.1c00967.

Bonil Koo, Bo Kyung Park, and Sun Kim. Transcriptome-conditioned molecule generation via gene interaction-aware fragment modeling with a gpt-based architecture. *Scientific Reports*, 15:38523, 2025. doi:10.1038/s41598-025-17439-7.

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. doi:10.1609/aaai.v38i12.29248.

Chen Li and Yoshihiro Yamanishi. Gx2mol: De novo generation of hit-like molecules from gene expression profiles. In *Machine Learning and Knowledge Discovery in Databases. Research Track*, volume 16015 of *Lecture Notes in Computer Science*, pp. 333–349. Springer, 2025. doi:10.1007/978-3-032-06066-2_20.

Yuesen Li, Chengyi Gao, Xin Song, Xiangyu Wang, Yungang Xu, and Suxia Han. Druggpt: A gpt-based strategy for designing potential ligands targeting specific proteins. *bioRxiv*, 2023. doi:10.1101/2023.06.29.543848.

Oscar Mendez-Lucio, Benoit Baillif, Djork-Arne Clevert, David Rouquie, and Joerg Wichard. De novo generation of hit-like molecules from gene expression signatures using artificial intelligence. *Nature Communications*, 11:10, 2020. doi:10.1038/s41467-019-13807-w.

Lukia Mistryukova, Vladimir Manuilov, Konstantin Avchaciov, and Peter O. Fedichev. Protobinddiff: A structure-free diffusion language model for protein sequence-conditioned ligand design. *bioRxiv*, 2025. doi:10.1101/2025.06.16.659955.

Rahul Ohlan, Raswanth Murugan, Li Xie, Mohammedsadeq Mottaqi, Shuo Zhang, and Lei Xie. Molgene-e: Inverse molecular design to modulate single cell transcriptomics. *bioRxiv*, 2025. doi:10.1101/2025.02.19.638723.

Saro Passaro, Gabriele Corso, Jeremy Wohlwend, Mateo Reveiz, Stephan Thaler, Vignesh Ram Somnath, Noah Getz, Tally Portnoi, Julien Roy, Hannes Stark, David Kwabi-Addo, Dominique Beaini, Tommi Jaakkola, and Regina Barzilay. Boltz-2: Towards accurate and efficient binding affinity prediction. *bioRxiv*, 2025. doi:10.1101/2025.06.14.659707.

Xingang Peng, Shitong Luo, Jiaqi Guan, Qi Xie, Jian Peng, and Jianzhu Ma. Pocket2mol: Efficient molecular sampling based on 3d protein pockets. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 17644–17655. PMLR, 2022.

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

G. W. Price, M. J. Burton, L. J. Collin, M. Duckworth, L. Gaster, M. Göthert, B. J. Jones, C. Roberts, J. M. Watson, and D. N. Middlemiss. SB-216641 and BRL-15572 – compounds to pharmacologically discriminate h5-HT1B and h5-HT1D receptors. *Naunyn-Schmiedeberg's Archives of Pharmacology*, 356(3):312–320, 1997. doi:10.1007/pl00005056.

Xiaoning Qi, Lianhe Zhao, Chenyu Tian, Yueyue Li, Zhen-Lin Chen, Peipei Huo, Runsheng Chen, Xiaodong Liu, Baoping Wan, Shengyong Yang, and Yi Zhao. Predicting transcriptional responses to novel chemical perturbations using deep generative model for drug discovery. *Nature Communications*, 15:9256, 2024. doi:10.1038/s41467-024-53457-1.

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.e17, 2017. doi:10.1016/j.cell.2017.10.049.

Oleg Trott and Arthur J. Olson. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry*, 31(2):455–461, 2010. doi:10.1002/jcc.21334.

Yixuan Wang, Xinyuan Liu, Yimin Fan, Binghui Xie, James Cheng, Kam Chung Wong, Peter Cheung, Irwin King, and Yu Li. Predicting drug responses of unseen cell types through transfer learning with foundation models. *Nature Computational Science*, 6(1):39–52, 2026. doi:10.1038/s43588-025-00887-6.

Yaxin Xu, Taojie Kuang, Shuang Ge, Haomin Wu, Mingqing Wang, Huan Xu, Fengwei An, Zhengyu Ma, Qiang Cheng, and Zhixiang Ren. Phenotype-driven de novo molecular design from gene expression signatures. *bioRxiv*, 2026. doi:10.64898/2026.07.21.739736.

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.

Hang Yuan, Chen Li, Wenjun Ma, and Yuncheng Jiang. Human-in-the-loop targeted molecule design informed by transcriptomes, 2026. URL https://openreview.net/forum?id=jL3nlH3Rjg.

A Additional Method Details

A.1 Phenotype encoder

The phenotype encoder is randomly initialized and does not receive a cell-line identity embedding. Landmark L1000 profiles provide $X_{\mathrm{pre}}, X_{\mathrm{post}}, X_{\Delta} \in \mathbb{R}^{978}$. The view-specific projections $\mathrm{Proj}_{\mathrm{pre}}$ and $\mathrm{Proj}_{\mathrm{post}}$ in the main text map each of X_{pre} and X_{post} to four tokens of dimension 256, which are updated by two bidirectional cross-attention blocks. X_{Δ} is projected to a single token by E_{Δ} and fused with the pooled pre- and post-states. Writing $c = (d, \tau)$ for dose and time, these covariates enter as $\log(1+d)$ and $\log(1+\tau)$ through the embedding $e(c)$. The encoder output is $z_P \in \mathbb{R}^{256}$.

A split projection then maps z_P to the two interfaces used in section 3.4:

$$(z_C, r_P) = A(z_P), \tag{10}$$

where $z_C = A_C(z_P)$ is a short token sequence in the native target-embedding space and r_P is a residual sequence of the same length (four tokens in our implementation). Only z_C occupies the target-conditioning slot, and only when no protein prompt is supplied. The residual r_P is consumed by the gated decoder pathway on both routes, so that dose, time, and residual cellular state are not forced into the protein prompt.

A.2 Residual adaptation

The pretrained generator is an autoregressive target-conditioned language model Li et al. (2023). The native target representation $z_T = E_T(T)$ is the frozen token embedding of the protein-sequence prompt; we do not train a separate protein encoder. Gated residual modules are installed in every decoder layer and take r_P as their conditioning memory. For a decoder state h, the residual used in the main text is

$$\Delta h(h, r_P) = \sigma\big(g(h, \bar{r}_P)\big) \odot W_r \bar{r}_P, \tag{11}$$

where σ is the sigmoid, g is a learned gating network, $\bar{r}_P$ is a pooled summary of r_P, and W_r is zero-initialized. Conditional low-rank updates (rank 8) are enabled only in the last two decoder layers and are gated by the same summary. A query mask restricts both updates to hidden states that predict molecule tokens, so protein-prompt states remain on the frozen native path. The backbone, its token embeddings, and the target prompt pathway are never updated. Disabling r_P therefore recovers $p_0(M \mid T)$ at initialization.

A.3 Training procedure

Training inputs and supervision. We use phenotype–molecule pairs $\mathcal{D}_P = \{(P_i, M_i)\}$ and phenotype–target–molecule triples $\mathcal{D}_{P,T} = \{(P_i, T_i, M_i)\}$. Here $M_i = (m_{i,1}, \ldots, m_{i,L_i})$ is the molecular token sequence. For contrastive terms, P_i^- is a context-matched phenotype from a different molecule and T_i^- is a mismatched target. Stage 1 uses the matched target T_i (and its target identity) as supervision for the phenotype-derived prompt. Stage 2 uses M_i as the hard molecular label. Stage 3 uses both M_i and the frozen target-only next-token distribution

$$q_{i,t}(\cdot) = p_0(\cdot \mid m_{i,<t}, T_i) \tag{12}$$

as hard and soft supervision, respectively.

Representation alignment. Using only $\mathcal{D}_{P,T}$, we freeze the generator and the residual head and train E_P together with A_C. No molecule likelihood is used. Let

$$u_i = \frac{\mathrm{Pool}(z_{C,i})}{\|\mathrm{Pool}(z_{C,i})\|_2}, \qquad v_i = \frac{\mathrm{Pool}(z_{T,i})}{\|\mathrm{Pool}(z_{T,i})\|_2}, \tag{13}$$

and let $\mathcal{P}(i) = \{j : y_j = y_i\}$ contain all batch elements sharing the target identity of sample i. With batch size B,

$$\begin{aligned} \mathcal{L}_{\cos} &= \frac{1}{B}\sum_i \big(1 - u_i^\top v_i\big), \\ \mathcal{L}_{P\to T} &= -\frac{1}{B}\sum_i \log \frac{\sum_{j\in\mathcal{P}(i)} \exp(u_i^\top v_j/\tau)}{\sum_k \exp(u_i^\top v_k/\tau)}, \qquad \mathcal{L}_{\mathrm{NCE}} = \tfrac{1}{2}\big(\mathcal{L}_{P\to T} + \mathcal{L}_{T\to P}\big), \end{aligned} \tag{14}$$

where $\mathcal{L}_{T\to P}$ is the symmetric term with u and v swapped. Let $\mathcal{V}_T$ be the samples that have a valid mismatched target. Then

$$\mathcal{L}_{\text{margin}} = \frac{1}{|\mathcal{V}_T|} \sum_{i\in\mathcal{V}_T} \left[\gamma_T - u_i^\top v_i + u_i^\top v_i^-\right]_+, \tag{15}$$

and $\mathcal{L}_{\text{align}} = \mathcal{L}_{\cos} + \mathcal{L}_{\text{NCE}} + \frac{1}{2}\mathcal{L}_{\text{margin}}$.

Phenotype-conditioned generation. Define the teacher-forced token NLL under a condition C by

$$\ell_i(C) = -\frac{1}{L_i} \sum_{t=1}^{L_i} \log p_\theta(m_{i,t} \mid m_{i,<t}, C). \tag{16}$$

On $\mathcal{D}_P$, the hard label is M_i and the condition is P_i. Let $\mathcal{V}_P$ be the samples with a valid context-matched negative. Then

$$\mathcal{L}^P_{\text{NLL}} = \frac{1}{B}\sum_i \ell_i(P_i), \qquad \mathcal{L}^P_{\text{ctr}} = \frac{1}{|\mathcal{V}_P|} \sum_{i\in\mathcal{V}_P} \left[\gamma_P + \ell_i(P_i) - \ell_i(P_i^-)\right]_+. \tag{17}$$

Thus the same molecule M_i should be more likely under the paired phenotype P_i than under P_i^-. Let c_i be the pooled $z_{C,i}$, let (μ_C, σ_C) be the current batch statistics of $\{c_i\}$, and let (μ_T, σ_T) be frozen statistics of native target encodings, with encoding dimension d_z. The moment term is

$$\mathcal{L}_{\text{mom}} = \text{MSE}(\mu_C, \mu_T) + \text{MSE}(\sigma_C, \sigma_T) + \frac{0.1}{d_z}\left(\frac{1}{B}\sum_i \|c_i\|_2 - \sqrt{d_z}\right)^2. \tag{18}$$

The pair objective in the main text is $\mathcal{L}_{\text{gen}} = \mathcal{L}^P_{\text{NLL}} + \lambda_{\text{ctr}}\mathcal{L}^P_{\text{ctr}} + \lambda_{\text{mom}}\mathcal{L}_{\text{mom}}$.

Target-guided distillation. On $\mathcal{D}_{P,T}$, the hard label remains M_i, but the native slot is filled by T_i, so the molecule likelihood is the P+T NLL $\ell_i(P_i, T_i)$. The corresponding phenotype contrast holds the target fixed,

$$\mathcal{L}^{P,T}_{\text{NLL}} = \frac{1}{B}\sum_i \ell_i(P_i, T_i), \qquad \mathcal{L}^{P,T}_{\text{ctr}} = \frac{1}{|\mathcal{V}_P|} \sum_{i\in\mathcal{V}_P} \left[\gamma_P + \ell_i(P_i, T_i) - \ell_i(P_i^-, T_i)\right]_+. \tag{19}$$

A target-switch contrast holds the phenotype fixed and swaps the protein prompt,

$$\mathcal{L}_{\text{switch}} = \frac{1}{|\mathcal{V}_T|} \sum_{i\in\mathcal{V}_T} \left[\gamma_T + \ell_i(P_i, T_i) - \ell_i(P_i, T_i^-)\right]_+. \tag{20}$$

The frozen teacher is $q_{i,t}(\cdot) = p_0(\cdot \mid m_{i,<t}, T_i)$, obtained by a no-gradient forward with r_P disabled. The two student next-token distributions are $p^P_{i,t} = p_\theta(\cdot \mid m_{i,<t}, P_i)$ and $p^{P,T}_{i,t} = p_\theta(\cdot \mid m_{i,<t}, P_i, T_i)$. For route r, $d_{r,i}$ averages $D_{\text{KL}}(q_{i,t}\|p^r_{i,t})$ over comparable molecule-token positions, excluding protein-prompt tokens. P-only positions are aligned to teacher positions by their molecule-relative index. Let $\mathcal{I}_P$ contain samples with at least one comparable position and $\bar{d}_P = |\mathcal{I}_P|^{-1}\sum_{i\in\mathcal{I}_P} d_{P,i}$. Set $\bar{d}_P = 0$ if $\mathcal{I}_P$ is empty and $d_{P,T,i} = 0$ for a P+T sample without valid positions. For B target-annotated triples, the implemented reduction is

$$\mathcal{L}_{\text{proj}} = \tfrac{1}{2}\left[\bar{d}_P - \varepsilon\right]^2_+ + \frac{1}{2B}\sum_{i=1}^{B}\left[d_{P,T,i} - \varepsilon\right]^2_+. \tag{21}$$

The teacher is not updated. The triple-batch objective is

$$\begin{aligned}\mathcal{L}_{\text{triple}} = {} & \lambda^{(3)}_{\text{nll}}\mathcal{L}^{P,T}_{\text{NLL}} + \lambda^{(3)}_{\text{ctr}}\mathcal{L}^{P,T}_{\text{ctr}} + \lambda^{(3)}_{\text{mom}}\mathcal{L}_{\text{mom}} + \lambda^{(3)}_{\cos}\mathcal{L}_{\cos} + \lambda^{(3)}_{\text{nce}}\mathcal{L}_{\text{NCE}} + \lambda^{(3)}_{\text{margin}}\mathcal{L}_{\text{margin}} \\ & + \beta\mathcal{L}_{\text{proj}} + \lambda_{\text{switch}}\mathcal{L}_{\text{switch}}.\end{aligned} \tag{22}$$

A replay stream from $\mathcal{D}_P$ continues the P-only pair losses and is excluded from distillation,

$$\mathcal{L}_{\text{replay}} = \lambda^{(R)}_{\text{nll}}\mathcal{L}^P_{\text{NLL}} + \lambda^{(R)}_{\text{ctr}}\mathcal{L}^P_{\text{ctr}} + \lambda^{(R)}_{\text{mom}}\mathcal{L}_{\text{mom}}. \tag{23}$$

The stagewise objectives are $\mathcal{L}^{(1)} = \mathcal{L}_{\text{align}}$, $\mathcal{L}^{(2)} = \mathcal{L}_{\text{gen}}$, and $\mathcal{L}^{(3)} = (1-\rho)\mathcal{L}_{\text{triple}} + \rho\mathcal{L}_{\text{replay}}$.

Algorithm 1 Target-constrained phenotype generation

Require: Phenotype–molecule pairs $\mathcal{D}_P$, aligned triples $\mathcal{D}_{P,T}$, pretrained generator p_0

1: Freeze p_0 and the native target pathway; initialize E_P, the split projection A, and the residual adapters
2: On $\mathcal{D}_{P,T}$, align z_C with frozen encodings z_T using $\mathcal{L}_{\text{align}}$
3: On $\mathcal{D}_P$, train the phenotype-only route by $\mathcal{L}_{\text{gen}}$
4: For each triple in $\mathcal{D}_{P,T}$, form frozen teacher $q_{i,t} = p_0(\cdot \mid m_{i,<t}, T_i)$ and optimize $\mathcal{L}_{\text{triple}}$
5: Mix a replay stream from $\mathcal{D}_P$ with $\mathcal{L}_{\text{replay}}$, excluded from distillation
6: Update E_P, A, and the residual adapters with $\mathcal{L}^{(3)} = (1-\rho)\mathcal{L}_{\text{triple}} + \rho\mathcal{L}_{\text{replay}}$
7: **return** Unified generator for target-free and target-informed inference

A.4 Algorithm

B Implementation details

We instantiate the method on a frozen DrugGPT backbone Li et al. (2023). The phenotype encoder, split projection, and residual adapters are trained; the generator, its token embeddings, and the native target pathway are not updated. The split projection emits four tokens in the target-embedding space. Conditional LoRA of rank 8 (scale $\alpha_{\text{LoRA}} = 8$) is applied in the last two decoder layers. Both inference routes share these modules. The four out-of-distribution splits use the same optimizer settings and are trained independently.

Optimization uses AdamW with weight decay 10^{-4} and gradient clipping at 1.0. The batch size is 8. Learning rates are held constant within each stage (Table 4); we do not use warmup or cosine decay. The training seed is 20260828. All reported training runs use NVIDIA A100 GPUs (16 GPUs in total).

Table 4: Stagewise learning rates. Residual/LoRA rates apply only when that group is updated.

Stage	Steps	TFE	Adapter	Residual/LoRA
Representation alignment	5,000	1×10^{-4}	5×10^{-5}	—
Phenotype-conditioned generation	10,000	1×10^{-4}	4×10^{-5}	2×10^{-5}
Target-guided distillation	10,000	5×10^{-5}	2×10^{-5}	1×10^{-5}

The first stage optimizes $\mathcal{L}^{(1)} = \mathcal{L}_{\text{align}}$ with cosine and NCE weights 1 and margin weight 0.5. The second optimizes $\mathcal{L}^{(2)} = \mathcal{L}_{\text{gen}}$ with contrast weight 0.25 and moment weight 0.05. The third uses $\mathcal{L}^{(3)} = (1-\rho)\mathcal{L}_{\text{triple}} + \rho\mathcal{L}_{\text{replay}}$ with $\rho = 1/2$, P+T NLL weight 0.5, phenotype-contrast and alignment weights 0.25, moment weight 0.02, $\beta = 2$, target-switch weight 0.5, and slack $\varepsilon = 0.03$.

Generation uses one molecule per condition, seed 20260829, temperature 1.0, top-$k = 9$, top-$p = 0.9$, and at most 192 new tokens. These decoding settings are shared by both of our routes and are not used to filter or rerank candidates.

B.1 Computational resources

Experiments were run on two servers, each equipped with eight NVIDIA A100-SXM4 GPUs with 80 GB of memory per GPU. This shared pool supported PhenTTP training and inference, baseline reproduction, and the training and evaluation of the PRNet, CRISP, and MAP response predictors. Independent runs and evaluation shards were scheduled across available GPUs; the total of 16 GPUs describes the shared resources, not the allocation to each individual run.

Boltz-2 affinity predictions used GPU inference, with independent single-GPU workers processing separate evaluation shards. AutoDock Vina docking ran on the host CPUs, with independent docking jobs parallelized across processes. Ligand preparation and RDKit-based molecular validity, uniqueness, novelty, QED, and SA calculations also ran on CPUs. The same GPU pool and CPU-based docking pipeline were used for PhenTTP and the baseline evaluations.

C DATASET CONSTRUCTION

Phenotype–molecule pairs $\mathcal{D}_P$ are native LINCS L1000 conditions: each row supplies X_{pre}, X_{post}, X_Δ, dose, time, and the corresponding molecule Subramanian et al. (2017). L1000 does not annotate a protein target for every condition.

Phenotype–target–molecule triples $\mathcal{D}_{P,T}$ are constructed afterwards. We collect human protein–ligand activities from ChEMBL 37 European Molecular Biology Laboratory (2026) and BindingDB Gilson et al. (2016), retaining K_i, K_d, IC_{50}, and EC_{50} values reported in nM. ChEMBL records are restricted to single-protein targets with assay confidence at least 7. Activities are joined to L1000 molecules by RDKit-standardized InChIKey, then standardized SMILES, then exact SMILES, then the LINCS InChIKey list. A match is treated as active if $p_{\text{activity}} = 9 - \log_{10}(\text{value}_{\text{nM}}) \geq 6$. When a molecule matches several active targets, we keep the most potent one and attach its UniProt sequence, falling back to the ChEMBL component sequence if needed. The resulting molecule-level label is inherited by every L1000 condition of that molecule, and by a Tahoe condition when the same molecule appears.

This procedure annotates 2,130 of 17,766 L1000 molecules (2,127 with a sequence). Conditions without a label remain in $\mathcal{D}_P$ and are never assigned a synthetic target. Representation alignment and dual-route distillation therefore use only the annotated subset; phenotype-conditioned generation trains on the full pair pool.

D EVALUATION PROTOCOL

D.1 METRIC DEFINITIONS

The scores in section 4.1 are defined as follows. Let N be the number of generation requests and G the number of evaluated genes. For condition i, a frozen evaluator $f \in \{\text{PRNet}, \text{CRISP}, \text{MAP}\}$ predicts the transcriptional change $\widehat{X}_{\Delta,i} = f(M_i; X_{\text{pre},i}, c_i)$ of the generated molecule M_i. The observed change is $X_{\Delta,i} = X_{\text{post},i} - X_{\text{pre},i} \in \mathbb{R}^G$. Then

$$\text{MSE}_i = \frac{1}{G}\big\|\widehat{X}_{\Delta,i} - X_{\Delta,i}\big\|_2^2, \qquad r_{\Delta,i} = \text{Pearson}\big(\widehat{X}_{\Delta,i}, X_{\Delta,i}\big). \tag{24}$$

Reported MSE and r_Δ are means over all N requests. Tables display $10^3 \times \text{MSE}$. If M_i is invalid or unscorable, MSE_i is replaced by the no-change value specified below and $r_{\Delta,i} = 0$.

Let $\mathcal{M}$ be the set of distinct valid canonical SMILES among the N requests that do not appear in the split-specific training set. Molecular quality uses

$$\text{VUN} = \frac{|\mathcal{M}|}{N} = \text{Validity} \times \text{Uniqueness} \times \text{Novelty}, \tag{25}$$

where validity is N_{valid}/N, uniqueness is $N_{\text{valid,unique}}/N_{\text{valid}}$, and novelty is $N_{\text{valid,unique,novel}}/N_{\text{valid,unique}}$. QED Bickerton et al. (2012) and SA Ertl & Schuffenhauer (2009) are means on the shared chemically scorable subset. FCD Preuer et al. (2018) is the Fréchet ChemNet distance between generated and paired true molecules on that subset. Target affinity is the AutoDock Vina score (kcal/mol) Trott & Olson (2010) and the Boltz-2 binder probability P_{bind} Passaro et al. (2025), each averaged over finite Vina–Boltz-2 pairs on the frozen panel.

Table 5: No-change replacement values for invalid or unscorable L1000 generations. Each MSE is the corresponding training partition's mean error for predicting $X_{\text{post}} = X_{\text{pre}}$. The same values are used for PRNet, CRISP, and MAP across all generation methods. Values are rounded for display.

OOD split	MSE (raw)	MSE ($\times 10^3$)	Assigned r_Δ
Chemical	0.016492	16.492	0
Scaffold	0.016537	16.537	0
Cell-line	0.017010	17.010	0
Target	0.016667	16.667	0

Table 5 applies only to L1000. Tahoe instead uses the condition-specific no-change MSE, $G^{-1}\|X_{\mathrm{post},i} - X_{\mathrm{pre},i}\|_2^2$, where G is the number of evaluated genes; it does not use these split-level constants.

The molecule–phenotype benchmark uses the frozen $2{,}000$-condition cohort. Every requested generation enters phenotype means and VUN. Invalid or unscorable outputs receive the corresponding training split's mean no-change MSE and $r_\Delta = 0$, retaining all $2{,}000$ requests. The L1000 replacement values are those in Table 5; r_Δ is 0 because a zero predicted change has undefined Pearson correlation. $\mathrm{VUN} = N_{\mathrm{valid,unique,novel}}/N$ on canonical SMILES against split-specific training molecules, which factors as validity times uniqueness among valid outputs times novelty among unique valid molecules. QED, SA, and FCD use one shared set of chemically scorable conditions per split; FCD is computed against the paired true molecules on that set. That common set is the intersection of conditions for which every method in the table produces a chemically scorable molecule, so n_{chem} and the QED, SA, and FCD values can differ when the method set changes (Table 1 versus Table 8). Phenotype metrics, VUN, and affinity are unaffected.

Affinity is generated independently on a frozen ten-target panel (ten conditions per target; 100 requests). The unseen-target split has its own panel. Within a split, all methods are scored on the same condition–target pairs, with target-macro averages. Each row uses its finite Vina–Boltz-2 pairs, so n_{aff} may differ across methods.

The main table compares phenotype-only generators, including both of our routes, on the shared $2{,}000$-condition cohort. Target-only DrugGPT and ProtoBind-Diff, and the corresponding PhenTTP backbones, appear in Table 8; unmatched routes are not inserted into the main table. Optimizer and decoding settings are in section B.

L1000 ablations use the same cohorts and scoring as Table 1. Table 3 reports a four-split macro-average; per-split scores are in Table 6, and joint training in Table 7. Table 2 reports a four-split Tahoe macro-average; per-split scores are in Table 9, and P-only ablations in Table 10. Tahoe uses the same metric families on a collapse-aware fixed budget: a valid, split-novel molecule is weighted by reciprocal frequency so that duplicates contribute total weight one, and the unused budget receives the condition-matched no-change MSE and $r_\Delta = 0$. Affinity is a frozen panel of 100 requests on Chemical, Scaffold, and Cell-line OOD, and 60 requests on Target OOD (six held-out targets). P+T is used only when a target annotation is present; otherwise the P-only route is used. Target-OOD P+T is scored on the $1{,}199$ annotated requests. The Tahoe instantiation reported as PhenTTP leaves the residual adapter off.

D.2 Training the phenotype evaluators

We train dataset-adapted PRNet, CRISP, and MAP response predictors separately on L1000 and Tahoe using their method-specific training objectives. The selected checkpoints are then frozen and shared across all generation methods. Training uses the fixed L1000 `drug_split_4` and Tahoe Chemical-OOD partitions, rather than separate evaluator training folds for each generator OOD split. These scores are used only for evaluation, not for training PhenTTP, selecting its checkpoints, or reranking generated molecules.

E Additional L1000 ablations

Table 6 is the per-split counterpart of Table 3. Table 7 reports joint training, which optimizes the three stage objectives simultaneously and remains competitive with the staged schedule.

F Alternative target-conditioned backbone

Table 8 repeats the four OOD settings with target-only DrugGPT and ProtoBind-Diff Mistryukova et al. (2025), each followed by the matching PhenTTP P-only and P+T variants. Scoring follows section D. Because this table includes additional generators, the shared chemically scorable cohort is larger than in Table 1, so QED, SA, and FCD are not directly comparable across the two tables.

Table 6: Per-split L1000 ablations corresponding to the four-OOD macro-average in Table 3. Result style: **best**, second-best.

Method	PRNet		CRISP		MAP		Molecular quality				Affinity	
	MSE↓	r_Δ ↑	MSE↓	r_Δ ↑	MSE↓	r_Δ ↑	VUN↑	QED↑	SA↓	FCD↓	Vina↓	P_{bind} ↑
Chemical OOD ($N = 2{,}000$, $n_{\text{chem}} = 485$)												
Full model (P-only)	**9.59**	**0.578**	**10.98**	**0.573**	**9.82**	**0.548**	92.8%	0.636	2.65	17.98	**-8.26**	0.358
w/o target supervision	10.03	0.538	11.49	0.531	10.28	0.508	86.9%	**0.657**	**2.52**	17.74	-7.81	0.312
w/o dual KD	9.77	0.560	11.10	0.559	9.90	0.534	82.4%	0.652	2.71	16.79	-7.62	0.299
w/o P–T alignment	9.65	0.576	11.18	0.567	9.84	0.546	**93.8%**	0.643	2.67	17.67	-8.13	0.318
Full model (P+T)	9.97	0.557	11.28	0.552	10.07	0.531	90.5%	0.604	2.81	**16.78**	-8.12	**0.480**
Scaffold OOD ($N = 2{,}000$, $n_{\text{chem}} = 232$)												
Full model (P-only)	10.09	0.570	**11.73**	**0.558**	**10.29**	**0.539**	92.8%	0.625	2.68	22.16	-8.00	0.303
w/o target supervision	10.44	0.533	12.18	0.516	10.65	0.504	85.7%	0.641	**2.64**	23.35	-7.19	0.306
w/o dual KD	10.09	0.566	11.91	0.548	10.31	0.535	81.0%	**0.661**	2.74	23.89	-7.38	0.294
w/o P–T alignment	**10.06**	**0.572**	11.93	0.551	10.36	0.534	**93.5%**	0.618	2.66	22.45	-8.03	0.335
Full model (P+T)	10.26	0.555	12.07	0.538	10.39	0.522	90.8%	0.614	2.75	**21.87**	**-8.10**	**0.505**
Cell-line OOD ($N = 2{,}000$, $n_{\text{chem}} = 884$)												
Full model (P-only)	10.21	**0.552**	12.18	0.520	10.35	**0.514**	**92.8%**	0.608	**2.67**	7.93	-7.96	0.331
w/o target supervision	10.61	0.515	13.03	0.467	10.72	0.478	85.2%	0.631	2.74	8.51	-7.67	0.312
w/o dual KD	10.25	0.541	12.42	0.501	10.48	0.496	84.9%	**0.650**	2.73	7.60	-6.69	0.284
w/o P–T alignment	**10.13**	**0.552**	**11.99**	**0.524**	**10.27**	**0.514**	92.6%	0.616	2.69	8.44	-8.16	0.314
Full model (P+T)	10.29	0.539	12.44	0.499	10.42	0.499	91.0%	0.589	2.84	**6.96**	**-8.19**	**0.476**
Target OOD ($N = 2{,}000$, $n_{\text{chem}} = 467$)												
Full model (P-only)	10.59	0.549	**12.34**	**0.532**	10.58	0.524	91.9%	0.630	2.66	21.03	**-8.32**	0.317
w/o target supervision	10.87	0.519	12.90	0.491	10.89	0.491	85.6%	**0.648**	2.68	20.26	-7.54	0.272
w/o dual KD	11.06	0.508	13.00	0.483	11.07	0.481	84.9%	0.643	2.67	21.09	-7.60	0.258
w/o P–T alignment	**10.52**	**0.557**	12.53	0.528	**10.48**	**0.531**	**92.9%**	0.646	**2.61**	20.93	-8.26	0.350
Full model (P+T)	10.77	0.543	12.43	0.525	10.58	0.521	91.3%	0.594	2.81	**18.98**	**-8.32**	**0.384**

Table 7: Joint training versus the staged Table 1 checkpoint, all under P-only inference. Scoring follows Table 6. Result style: **best**, second-best.

Method	PRNet		CRISP		MAP		Molecular quality				Affinity	
	MSE↓	r_Δ ↑	MSE↓	r_Δ ↑	MSE↓	r_Δ ↑	VUN↑	QED↑	SA↓	FCD↓	Vina↓	P_{bind} ↑
Chemical OOD ($N = 2{,}000$, $n_{\text{chem}} = 485$)												
Full model (P-only)	**9.59**	**0.578**	**10.98**	**0.573**	**9.82**	**0.548**	92.8%	0.636	**2.65**	17.98	**-8.26**	**0.358**
Joint training	9.60	0.575	11.31	0.562	9.83	0.544	**93.0%**	**0.649**	2.66	**17.21**	-8.14	0.322
Scaffold OOD ($N = 2{,}000$, $n_{\text{chem}} = 232$)												
Full model (P-only)	10.09	0.570	**11.73**	**0.558**	10.29	0.539	**92.8%**	0.625	2.68	**22.16**	-8.00	0.303
Joint training	**9.99**	**0.575**	11.91	0.553	**10.26**	**0.541**	**92.8%**	**0.634**	**2.59**	22.26	**-8.21**	**0.331**
Cell-line OOD ($N = 2{,}000$, $n_{\text{chem}} = 884$)												
Full model (P-only)	10.21	0.552	12.18	0.520	10.35	0.514	**92.8%**	0.608	**2.67**	**7.93**	-7.96	0.331
Joint training	**10.15**	**0.556**	**12.14**	**0.523**	**10.24**	**0.517**	92.5%	**0.635**	2.72	8.19	**-8.16**	**0.343**
Target OOD ($N = 2{,}000$, $n_{\text{chem}} = 467$)												
Full model (P-only)	**10.59**	0.549	12.34	**0.532**	10.58	0.524	91.9%	**0.630**	2.66	**21.03**	**-8.32**	0.317
Joint training	**10.59**	**0.552**	**12.33**	0.531	**10.48**	**0.526**	**92.3%**	0.622	**2.61**	21.94	-8.26	**0.357**

G TAHOE PER-SPLIT RESULTS

Table 9 is the per-split counterpart of the four-OOD macro-average in Table 2. Table 10 is the matching P-only ablation. Scoring is in section D.

Table 8: L1000 generation with target-only baselines and matching PhenTTP variants, grouped by backbone. Phenotype metrics and VUN use $N = 2{,}000$ requests; affinity uses a fixed 100-request panel. QED, SA, and FCD use the common chemically scorable cohort of this larger method set and are not matched to Table 1. Result style: **best**, second-best. Gray cells mark a metric shortfall.

Method	PRNet		CRISP		MAP		Molecular quality				Affinity	
	MSE↓	r_Δ ↑	MSE↓	r_Δ ↑	MSE↓	r_Δ ↑	VUN↑	QED↑	SA↓	FCD↓	Vina↓	P_{bind} ↑
Chemical OOD ($N = 2{,}000$, $n_{\text{chem}} = 592$)												
GexMolGen	10.36	0.549	17.87	0.391	11.32	0.431	5.7%	0.511	3.84	79.50	-7.28	0.202
Gx2Mol	13.62	0.247	15.29	0.225	13.81	0.222	31.7%	0.579	3.07	20.49	-7.72	0.280
CURE	11.54	0.493	11.43	0.529	10.79	0.474	90.8%	0.187	4.89	43.34	-4.95	0.237
GEMGen	9.78	0.568	12.60	0.532	10.24	0.516	74.0%	0.532	2.86	16.93	-7.67	0.348
XMolRL	10.34	0.517	13.49	0.473	10.67	0.470	47.4%	0.581	3.03	17.40	-7.58	0.282
DrugGPT (T-only)	9.66	0.587	10.82	0.590	9.61	0.563	90.0%	0.536	2.90	18.08	**-8.57**	**0.611**
PhenTTP–DrugGPT (P-only)	9.59	0.578	10.98	0.573	9.82	0.548	92.8%	0.632	2.68	17.46	-8.26	0.358
PhenTTP–DrugGPT (P+T)	9.97	0.557	11.28	0.552	10.07	0.531	90.5%	0.602	2.81	**16.51**	-8.12	0.480
ProtoBind-Diff (T-only)	9.59	0.591	10.93	0.593	9.55	0.568	**100.0%**	0.593	2.77	16.70	-8.29	0.414
PhenTTP–ProtoBind-Diff (P-only)	9.59	0.589	10.99	0.589	9.60	0.565	97.9%	**0.646**	2.96	18.09	-6.70	0.267
PhenTTP–ProtoBind-Diff (P+T)	**9.36**	**0.598**	**10.73**	**0.595**	**9.52**	**0.572**	98.2%	0.636	**2.61**	16.61	-7.15	0.358
Scaffold OOD ($N = 2{,}000$, $n_{\text{chem}} = 296$)												
GexMolGen	10.82	0.543	18.65	0.385	11.80	0.424	6.5%	0.512	3.80	79.84	-6.95	0.156
Gx2Mol	15.04	0.125	15.66	0.115	15.14	0.115	20.1%	0.590	2.98	24.55	-7.03	0.253
CURE	11.95	0.496	12.09	0.529	11.15	0.472	90.2%	0.183	4.93	52.59	-4.78	0.180
GEMGen	10.29	0.566	13.72	0.512	10.82	0.509	75.0%	0.544	2.86	**20.59**	-7.68	0.322
XMolRL	10.64	0.526	13.99	0.478	11.16	0.468	45.9%	0.573	3.04	23.02	-7.64	0.254
DrugGPT (T-only)	10.08	0.582	11.89	0.566	10.19	0.552	90.2%	0.544	2.90	23.33	**-8.56**	**0.613**
PhenTTP–DrugGPT (P-only)	10.09	0.570	11.73	0.558	10.29	0.539	92.8%	0.613	2.74	21.93	-8.00	0.303
PhenTTP–DrugGPT (P+T)	10.26	0.555	12.07	0.538	10.39	0.522	90.8%	0.613	2.82	21.96	-8.10	0.505
ProtoBind-Diff (T-only)	9.96	0.591	11.74	0.577	10.12	0.557	**99.8%**	0.596	2.76	22.02	-8.18	0.438
PhenTTP–ProtoBind-Diff (P-only)	9.78	0.597	11.61	0.582	10.00	0.562	99.0%	**0.675**	2.79	23.46	-7.11	0.241
PhenTTP–ProtoBind-Diff (P+T)	**9.69**	**0.602**	**11.57**	**0.583**	**9.96**	**0.568**	99.2%	0.667	**2.52**	22.84	-7.18	0.382
Cell-line OOD ($N = 2{,}000$, $n_{\text{chem}} = 1063$)												
GexMolGen	10.50	0.549	17.22	0.397	11.09	0.447	6.3%	0.510	3.82	77.90	-7.04	0.158
Gx2Mol	10.70	0.513	13.45	0.462	10.79	0.462	14.7%	0.554	3.02	6.46	-7.37	0.368
CURE	12.01	0.471	12.71	0.475	11.39	0.432	90.5%	0.166	5.03	40.82	-5.12	0.232
GEMGen	10.34	0.547	13.91	0.473	10.76	0.479	71.4%	0.541	2.89	6.67	-7.83	0.331
XMolRL	11.40	0.453	14.19	0.398	11.60	0.408	71.2%	0.594	3.03	6.55	-7.33	0.233
DrugGPT (T-only)	10.16	0.561	11.89	0.535	10.12	0.523	91.0%	0.527	2.93	7.23	**-8.59**	**0.598**
PhenTTP–DrugGPT (P-only)	10.21	0.552	12.18	0.520	10.35	0.514	92.8%	0.602	2.70	7.41	-7.96	0.331
PhenTTP–DrugGPT (P+T)	10.29	0.539	12.44	0.499	10.42	0.499	91.0%	0.592	2.86	**6.37**	-8.19	0.476
ProtoBind-Diff (T-only)	10.09	0.571	**11.78**	**0.548**	**10.01**	**0.534**	**99.9%**	0.567	2.79	7.47	-8.35	0.486
PhenTTP–ProtoBind-Diff (P-only)	**9.99**	**0.573**	12.54	0.524	10.10	0.530	98.9%	**0.658**	2.97	8.21	-7.18	0.261
PhenTTP–ProtoBind-Diff (P+T)	10.04	0.567	12.19	0.528	10.11	0.531	97.9%	0.624	**2.60**	7.63	-7.58	0.366
Target OOD ($N = 2{,}000$, $n_{\text{chem}} = 587$)												
GexMolGen	11.37	0.534	18.99	0.374	11.89	0.429	5.7%	0.504	3.80	83.33	-6.82	0.141
Gx2Mol	14.21	0.229	15.53	0.205	14.35	0.204	12.2%	0.565	2.69	22.95	-7.17	0.310
CURE	11.72	0.502	12.49	0.511	11.15	0.480	91.8%	0.278	4.66	43.61	-5.33	0.172
GEMGen	10.63	0.558	13.72	0.506	10.84	0.512	72.7%	0.527	2.90	20.24	-7.92	0.318
XMolRL	10.88	0.532	14.13	0.478	11.16	0.476	30.7%	0.568	2.73	20.02	-7.82	0.297
DrugGPT (T-only)	10.61	0.565	12.07	0.555	10.32	0.544	84.7%	0.523	3.06	20.23	-7.38	**0.523**
PhenTTP–DrugGPT (P-only)	10.59	0.549	12.34	0.532	10.58	0.524	91.9%	0.629	2.71	20.08	**-8.32**	0.317
PhenTTP–DrugGPT (P+T)	10.77	0.543	12.43	0.525	10.58	0.521	91.3%	0.594	2.83	**18.36**	**-8.32**	0.384
ProtoBind-Diff (T-only)	10.57	0.572	12.10	0.561	10.22	0.553	**100.0%**	0.579	2.88	19.57	-8.24	0.412
PhenTTP–ProtoBind-Diff (P-only)	10.40	0.578	12.51	0.550	10.19	0.554	99.2%	**0.647**	2.87	19.31	-7.41	0.253
PhenTTP–ProtoBind-Diff (P+T)	**10.27**	**0.584**	**12.06**	**0.564**	**10.14**	**0.560**	99.4%	0.613	**2.62**	18.64	-7.55	0.339

Table 9: Per-split Tahoe generation corresponding to the four-OOD macro-average in Table 2. P+T uses a target prompt only when an annotation exists. Result style: **best**, second-best.

Method	PRNet		CRISP		MAP		Molecular quality				Affinity	
	MSE↓	r_Δ ↑	MSE↓	r_Δ ↑	MSE↓	r_Δ ↑	VUN↑	QED↑	SA↓	FCD↓	Vina↓	P_{bind} ↑
Chemical OOD												
XMolRL	6.61	0.169	6.53	0.186	6.55	0.144	32.6%	0.544	3.61	**50.87**	-5.973	0.308
Gx2Mol	6.87	0.009	6.86	0.010	6.97	0.006	1.8%	0.501	3.74	51.31	-4.698	0.148
GexMolGen	6.53	0.058	6.67	0.060	7.09	0.038	10.7%	0.477	3.75	99.48	-5.986	0.265
GEMGen	6.12	0.381	5.97	0.420	6.36	0.309	73.6%	0.532	2.82	53.30	-6.926	0.413
CURE	6.02	0.454	5.86	0.498	10.17	0.284	88.0%	0.308	4.55	52.70	-4.943	0.309
PhenTTP (P-only)	**5.95**	**0.480**	5.73	**0.529**	**5.88**	**0.406**	**92.5%**	0.580	2.81	53.83	-3.460	0.400
PhenTTP (P+T)	5.96	0.471	**5.72**	0.520	5.89	0.398	91.2%	**0.590**	**2.79**	52.69	**-7.317**	**0.497**
Scaffold OOD												
XMolRL	6.26	0.186	6.16	0.206	6.19	0.161	35.7%	0.537	3.58	60.79	-6.961	0.330
Gx2Mol	6.62	0.007	6.63	0.008	6.69	0.005	1.4%	0.534	3.91	**53.22**	-2.454	0.482
GexMolGen	6.40	0.058	6.50	0.061	6.96	0.038	10.9%	0.477	3.76	110.11	-7.061	0.217
GEMGen	5.96	0.391	5.82	0.432	6.17	0.320	75.3%	0.530	**2.80**	66.29	-6.894	0.389
CURE	5.83	0.461	5.64	0.510	10.02	0.284	89.0%	0.308	4.54	57.25	-5.522	0.281
PhenTTP (P-only)	**5.79**	**0.481**	**5.49**	**0.535**	**5.66**	**0.413**	**92.4%**	**0.600**	**2.80**	64.70	-8.179	0.350
PhenTTP (P+T)	5.81	0.473	5.50	0.528	5.67	0.408	91.1%	0.589	2.91	57.13	**-8.537**	**0.560**
Cell-line OOD												
XMolRL	5.96	0.012	5.94	0.013	5.96	0.010	2.3%	0.515	3.09	3.01	-6.954	0.338
Gx2Mol	5.94	0.037	5.87	0.042	6.33	0.021	7.3%	0.514	3.12	**2.68**	-6.471	0.301
GexMolGen	5.87	0.051	5.81	0.056	6.40	0.032	10.0%	0.477	3.78	78.32	-5.427	0.259
GEMGen	5.57	0.381	4.92	0.434	5.70	0.319	75.3%	0.510	2.87	11.21	-6.745	0.308
CURE	5.52	0.444	4.82	0.500	9.82	0.270	88.4%	0.281	4.63	33.87	-5.037	0.231
PhenTTP (P-only)	**5.50**	0.462	4.66	0.527	5.22	0.396	91.5%	**0.571**	**2.83**	11.70	**-7.205**	0.327
PhenTTP (P+T)	**5.50**	**0.467**	**4.65**	**0.532**	**5.21**	**0.401**	**92.4%**	0.560	2.85	10.86	-6.693	**0.583**
Target OOD												
XMolRL	6.16	0.187	5.99	0.212	6.02	0.164	36.1%	0.528	3.61	69.06	-5.320	0.309
Gx2Mol	6.45	0.011	6.42	0.014	6.57	0.006	2.3%	0.550	3.68	63.38	-6.079	0.167
GexMolGen	6.10	0.067	6.17	0.072	6.81	0.042	12.6%	0.474	3.74	127.93	-4.761	0.210
GEMGen	5.72	0.397	5.30	0.452	5.92	0.329	76.1%	0.524	2.82	83.30	-6.549	0.355
CURE	5.72	0.458	5.22	0.525	10.23	0.274	89.3%	0.310	4.60	67.91	-4.556	0.196
PhenTTP (P-only)	**5.69**	**0.484**	**5.14**	**0.555**	**5.43**	**0.424**	**93.2%**	**0.620**	**2.74**	69.47	**-6.883**	0.304
PhenTTP (P+T)	5.83	0.395	5.38	0.452	5.62	0.347	76.0%	0.550	2.95	**61.68**	-6.357	**0.651**

Table 10: P-only ablations of the Tahoe model in Table 2, under the per-split layout of Table 9. Later rows remove dual-route distillation or phenotype–target alignment, or train the stages jointly. Result style: **best**, second-best.

Method	PRNet		CRISP		MAP		Molecular quality				Affinity	
	MSE↓	r_Δ ↑	MSE↓	r_Δ ↑	MSE↓	r_Δ ↑	VUN↑	QED↑	SA↓	FCD↓	Vina↓	P_{bind} ↑
Chemical OOD												
PhenTTP	**5.95**	**0.480**	5.73	**0.529**	5.88	0.406	**92.5%**	0.580	2.81	53.82	-3.460	0.400
w/o dual KD	6.12	0.396	5.96	0.436	6.05	0.336	76.3%	**0.611**	**2.75**	**33.87**	-6.015	0.387
w/o P–T align	5.98	0.460	5.72	0.509	5.92	0.389	88.6%	0.566	2.82	33.98	-6.644	0.381
Joint	**5.95**	0.479	**5.71**	**0.529**	**5.86**	**0.407**	92.3%	0.597	2.78	35.01	**-6.796**	**0.407**
Scaffold OOD												
PhenTTP	**5.79**	**0.481**	**5.49**	**0.535**	**5.66**	**0.413**	**92.4%**	0.600	**2.80**	63.73	**-8.179**	0.350
w/o dual KD	5.88	0.424	5.63	0.474	5.77	0.366	81.7%	0.600	2.88	**35.39**	-7.973	0.328
w/o P–T align	5.85	0.443	5.63	0.493	5.72	0.381	85.2%	0.585	2.89	35.70	-8.095	0.341
Joint	5.81	0.471	5.60	0.523	5.68	0.405	90.6%	**0.611**	2.81	36.29	-7.995	**0.364**
Cell-line OOD												
PhenTTP	**5.50**	**0.462**	**4.66**	**0.527**	**5.22**	**0.396**	**91.5%**	0.571	2.83	11.66	-7.205	0.327
w/o dual KD	5.57	0.393	4.88	0.447	5.33	0.337	77.7%	**0.592**	**2.81**	11.41	-7.037	0.304
w/o P–T align	5.75	0.206	5.39	0.236	5.63	0.177	40.6%	0.562	2.94	**7.79**	-6.689	**0.341**
Joint	5.63	0.361	4.98	0.411	5.40	0.309	71.4%	0.551	2.93	9.44	**-7.354**	0.327
Target OOD												
PhenTTP	**5.69**	0.484	5.14	0.555	5.43	0.424	93.2%	0.620	2.74	69.45	-6.883	0.304
w/o dual KD	**5.69**	**0.494**	**5.10**	**0.567**	**5.40**	**0.434**	**95.6%**	**0.690**	**2.43**	69.43	**-7.400**	0.299
w/o P–T align	5.71	0.466	5.20	0.534	5.46	0.409	90.1%	0.596	2.81	**61.94**	-7.084	**0.343**
Joint	5.75	0.477	5.24	0.545	5.48	0.418	92.3%	0.615	2.75	63.45	-6.208	0.314