DECO: DECOUPLED CONDITIONAL GENERATION FOR TRANSCRIPTOMICS-DRIVEN DRUG DESIGN

Anonymous authors
Paper under double-blind review

ABSTRACT

Transcriptomics-driven drug design seeks to generate molecules that induce desired cellular responses without requiring predefined molecular targets. A central challenge is that transcriptomic responses and molecular structures inhabit fundamentally different representation spaces, making it nontrivial to connect biological conditions to chemical generation. Existing approaches bridge these modalities through direct conditioning, cross-modal alignment, or coupled representation learning. Our analysis further reveals a weak and non-unique relationship between transcriptomic-response similarity and chemical similarity. Guided by this observation, we introduce DECO (**DE**coupled **CO**nditional generation), which learns chemical representations independently and models desired responses as conditional distributions over the resulting latent space. To make this decoupled space responsive to transcriptomic conditions, DECO combines global transition features for adaptive modulation with gene-level transition tokens for cross-attention. A generation-centered objective explicitly promotes condition dependence while retaining cross-modal alignment only as auxiliary semantic supervision. A prediction-error-aware decoder improves robustness to generator-induced latent deviations while preserving reconstruction fidelity. Across standard, cell-line OOD, and scaffold OOD splits, DECO achieves the best paired structural similarity and predicted transcriptomic consistency among evaluated methods, reducing Best@10 response MSE by 19.0–43.2% relative to the best-performing baseline in each split while maintaining 94.5–96.1% molecular validity.

1 INTRODUCTION

Transcriptomics-driven drug design asks whether molecules can be generated to induce a desired cellular response without assuming a predefined molecular target. Large-scale resources such as the Connectivity Map and its L1000-based expansion provide transcriptomic readouts of cellular responses to chemical perturbations Lamb et al. (2006); Subramanian et al. (2017). These resources make it possible to formulate the task conditionally: given a desired control-to-perturbed gene-expression transition in a specified cellular context, infer a distribution over candidate molecules rather than predict the response to a given molecule Xu et al. (2026).

Unlike structure-based molecular generation, where protein pockets and ligands share a three-dimensional coordinate frame Guan et al. (2023), transcriptomics-driven design must connect biological responses to molecular structures across fundamentally different representation spaces. Existing approaches address this challenge through direct transcriptome conditioning Liu et al. (2024); Wang et al. (2024), explicit cross-modal matching or chemically informed representation learning Cheng et al. (2024); Xu et al. (2026), and coupled or shared-latent modeling Guo & Liu (2026). These strategies provide different mechanisms for connecting biological and chemical representations. But is transcriptomic-response similarity a reliable signal for organizing chemical-space geometry? We therefore examine the empirical relationship between transcriptomic-response similarity and chemical similarity.

Within matched experimental contexts, we represent each transcriptional response by the perturbation-induced expression change, $\Delta\mathbf{x} = \mathbf{x}^{\mathrm{pert}} - \mathbf{x}^{\mathrm{ctrl}}$, and compute response similarity as cosine similarity between these change vectors. Response similarity shows little association with chemical similarity: median Morgan similarity remains near 0.10 across response-similarity bins,

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.

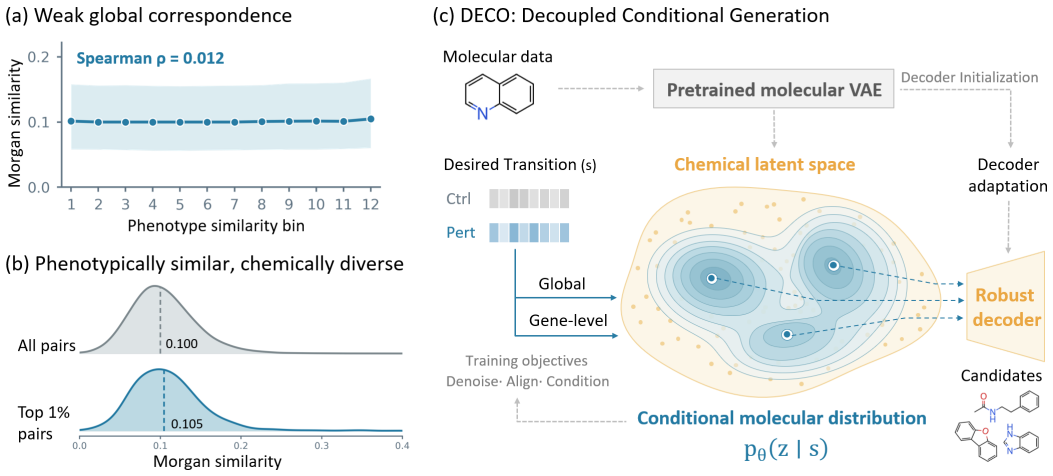


Figure 1: Motivation and overview of DECO. **(a)** Across matched-context compound pairs, the median Morgan similarity remains nearly constant across transcriptomic response-similarity bins (blue line), with the shaded region denoting the 10th–90th percentile interval; the overall association is weak (Spearman $\rho = 0.012$). **(b)** Even the pairs in the top 1% of response similarity within each experimental context exhibit low structural similarity: their median Morgan similarity is 0.105, compared with 0.100 across all pairs. **(c)** Motivated by the weak and non-unique response–structure relationship, DECO decouples chemical representation learning from transcriptomic conditioning and learns a conditional distribution over the independently learned chemical latent space.

with an overall pair-level Spearman correlation of $\rho = 0.012$ (Figure 1a). Even among pairs in the top 1% of response similarity within each experimental context, the median Morgan similarity is only 0.105, compared with 0.100 across all sampled pairs (Figure 1b). Qualitative examples of structurally distinct compounds with similar measured responses are provided in Appendix A.2. These observations suggest that transcriptomic similarity is not a reliable pointwise signal for organizing chemical-space geometry, while remaining informative for conditioning generation. This motivates separating two modeling roles that are often intertwined: learning a chemically meaningful representation space and conditioning generation on a desired biological response. Accordingly, we learn the chemical space independently and let the desired response specify a conditional distribution over candidate molecules within that space.

We implement this formulation with DECO. Given an independently learned chemical latent space, DECO models $p_\theta(\mathbf{z} \mid s)$, where s denotes the desired control-to-perturbed state transition and $\mathbf{z}$ a molecular latent.

This decoupling shifts the modeling burden from cross-modal representation construction to conditional generation. First, the transcriptomic condition must provide sufficiently informative control over the chemical space; DECO therefore combines global transition features with gene-level response signals through dual-scale conditioning. Second, the primary denoising loss is augmented with auxiliary condition-usage and lightweight alignment objectives. Finally, because diffusion-generated latents may deviate from encoder-produced latents seen during autoencoder training, we introduce prediction-error-aware decoding to improve robustness at the generator–decoder interface.

Our contributions are threefold:

- **Analysis and formulation.** We characterize the weak correlation between transcriptomic-response similarity and chemical similarity, and motivate conditional generation over an independently learned chemical space.
- **Method.** We develop DECO with dual-scale conditioning, auxiliary condition-usage and alignment objectives, and prediction-error-aware decoder adaptation.
- **Evidence.** Across standard, cell-line OOD, and scaffold OOD settings, DECO reduces Best@10 PRNet-predicted response MSE by 19.0–43.2% relative to the strongest baseline in each split while improving paired structural recovery.

2 Related Work

2.1 Transcriptomics-driven molecular generation

Early methods explored gene-expression-conditioned adversarial generation (Méndez-Lucio et al., 2020; Shayakhmetov et al., 2020). TRIOMPHE and Gex2SGen use expression-informed variational models (Kaitoh & Yamanishi, 2021; Das et al., 2023), while PaccMannRL optimizes molecules for transcriptomic cellular contexts using drug-sensitivity rewards (Born et al., 2021). TransGEM employs Transformer-based generation (Liu et al., 2024), and GLDM combines a molecular autoencoder with transcriptome-conditioned latent diffusion (Wang et al., 2024).

Explicit correspondence learning has also been studied for molecule–transcriptome retrieval (Finlayson et al., 2021) and generation. GexMolGen uses hierarchical matching and contrastive alignment in a "first-align-then-generate" pipeline (Cheng et al., 2024); MolGene-E learns cross-modal associations through supervised contrastive learning (Ohlan et al., 2025); and CURE aligns perturbation representations with complementary chemical views (Xu et al., 2026). XMolRL uses coupled VAEs for phenotype-guided generation, followed by multi-objective reinforcement learning for target-specific optimization (Guo & Liu, 2026). Building on these approaches, DECO uses response–structure analysis to motivate a decoupled conditional modeling strategy, with cross-modal alignment serving as auxiliary supervision.

2.2 Latent molecular generative modeling

Molecular VAEs established continuous representations for molecular generation and optimization (Gómez-Bombarelli et al., 2018). JT-VAE and HierVAE incorporate graph and substructure organization into molecular encoding and decoding (Jin et al., 2018; 2020), while GeoLDM extends latent diffusion to 3D molecular generation (Xu et al., 2023). These approaches provide the representational foundation for learning generative distributions in chemical latent space.

2.3 Robust latent decoding

Accurate reconstruction of encoded molecules does not guarantee reliable decoding of generated latents. Notin et al. use decoder uncertainty to guide latent-space optimization away from unreliable regions (Notin et al., 2021). In molecular latent diffusion, TopVAE identifies sampling-accessible "dark areas" that decode into invalid or disconnected structures, and improves robustness through topology-aware decoding and chemical constraint learning (Wang et al., 2026). This issue is relevant to transcriptomics-driven generation, where molecular latents are predicted from biological conditions rather than obtained by encoding known molecules. Accordingly, DECO trains the decoder to tolerate generator-like latent deviations while retaining reconstruction fidelity on clean encoder-produced latents, supporting reliable decoding of the learned response-conditioned distribution.

3 Method

3.1 Problem Setup and Decoupled Formulation

Let $\mathbf{x}^{\mathrm{ctrl}}, \mathbf{x}^{\mathrm{pert}} \in \mathbb{R}^G$ denote the control and desired perturbed transcriptomic states, respectively, where G is the number of measured genes. Given this state pair, our goal is to generate candidate molecules compatible with the desired transcriptional transition. A transcriptomic encoder represents this transition as

$$s = \mathcal{S}\left(\mathbf{x}^{\mathrm{ctrl}}, \mathbf{x}^{\mathrm{pert}}\right). \tag{1}$$

The chemical representation defines the space in which generation takes place, while the transcriptomic condition guides the distribution of candidate molecular latents within that space. This formulation accommodates multiple chemical candidates for a desired response.

Figure 2 summarizes the three stages of DECO: independent chemical representation learning, dual-scale conditional latent generation, and prediction-error-aware decoder adaptation.

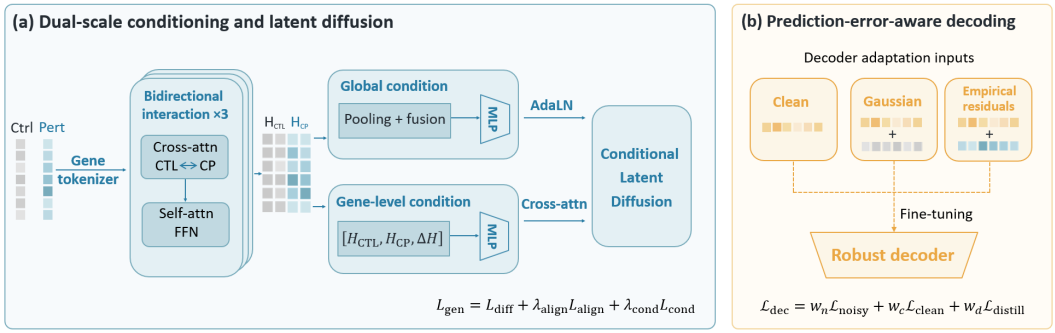


Figure 2: **Architecture of DECO. (a)** Control and desired perturbed transcriptomic states are processed by bidirectional interaction blocks to obtain $\mathbf{H}_{\mathrm{CTL}}$ and $\mathbf{H}_{\mathrm{CP}}$. Global pooled features condition latent diffusion through AdaLN, while gene-level features $[\mathbf{H}_{\mathrm{CTL}}, \mathbf{H}_{\mathrm{CP}}, \Delta\mathbf{H}]$, with $\Delta\mathbf{H} = \mathbf{H}_{\mathrm{CP}} - \mathbf{H}_{\mathrm{CTL}}$, are injected through cross-attention. Training combines the primary diffusion objective with auxiliary semantic-alignment and condition-usage losses. **(b)** The decoder is adapted using clean latents, Gaussian perturbations, and empirical generator residuals, with reconstruction and distillation objectives improving robustness to generator-induced latent deviations.

3.2 Independent Chemical Representation Learning

We adopt the SMILES-based molecular VAE architecture introduced by Kim et al. (2025) and retrain it from scratch using molecular reconstruction and KL regularization. For a molecule m, the encoder parameterizes $q_\phi(\mathbf{z} \mid m)$. We use its posterior mean as the deterministic representation for conditional generation:

$$\mathbf{z}_0 = \boldsymbol{\mu}_\phi(m) \in \mathbb{R}^{d_z}. \tag{2}$$

The molecular encoder remains fixed during subsequent conditional learning, so transcriptomic supervision does not update the chemical representation. The pretrained decoder D_ψ initializes the decoder adaptation described in Section 3.5. Architecture details are provided in Appendix B.

3.3 Dual-Scale Transcriptomic Conditioning

We encode the desired transition at two complementary scales: a global representation of the state pair and gene-level representations of their changes. Each gene expression value is embedded together with a learnable gene-identity embedding and a state embedding indicating control or perturbed origin. The resulting control and perturbed token sequences are processed by bidirectional interaction blocks. Each block applies cross-attention in both directions, followed by within-state self-attention and feed-forward transformations. We denote the refined representations by $\mathbf{H}_{\mathrm{CTL}}, \mathbf{H}_{\mathrm{CP}} \in \mathbb{R}^{G \times d_h}$, omitting the batch dimension for clarity.

Global transition condition. Attention pooling summarizes the two refined sequences, and an adaptive fusion module followed by an MLP produces the global condition:

$$\mathbf{C}_{\mathrm{global}} = f_{\mathrm{global}}(\mathrm{Pool}(\mathbf{H}_{\mathrm{CTL}}), \mathrm{Pool}(\mathbf{H}_{\mathrm{CP}}))\,. \tag{3}$$

This condition modulates the denoiser through adaptive normalization, providing a global summary of the desired transition.

Gene-level transition condition. To retain gene-resolved information, we concatenate the refined control and perturbed representations with their difference, and project the result into gene-level condition tokens:

$$\mathbf{C}_{\mathrm{gene}} = f_{\mathrm{gene}}([\mathbf{H}_{\mathrm{CTL}}, \mathbf{H}_{\mathrm{CP}}, \mathbf{H}_{\mathrm{CP}} - \mathbf{H}_{\mathrm{CTL}}])\,. \tag{4}$$

These tokens condition the denoiser through cross-attention, allowing it to access gene-specific transition information throughout generation. Together, the two pathways define $s = (\mathbf{C}_{\mathrm{global}}, \mathbf{C}_{\mathrm{gene}})$. Network configurations are given in Appendix B.

3.4 CONDITIONAL LATENT DIFFUSION

For each training state pair and its associated molecule, we perturb the molecular latent $\mathbf{z}_0$ at diffusion timestep t:

$$\mathbf{z}_t = \sqrt{\bar{\alpha}_t}\mathbf{z}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{5}$$

where $\bar{\alpha}_t$ is the cumulative signal-retention coefficient of the noise schedule. A latent Transformer predicts the clean latent as $\hat{\mathbf{z}}_0 = f_\theta(\mathbf{z}_t, t, s)$, using global adaptive modulation and gene-level cross-attention. The primary learning objective is molecular latent denoising:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{(\mathbf{z}_0,s),\,t,\,\boldsymbol{\epsilon}}\left[\left\|f_\theta(\mathbf{z}_t, t, s) - \mathbf{z}_0\right\|_2^2\right]. \tag{6}$$

Two auxiliary objectives support response-conditioned generation. For semantic alignment, let $\mathbf{h}_i^p$ and $\mathbf{h}_i^m$ denote the ℓ_2-normalized projections of the global transcriptomic representation and molecular latent for example i, respectively, and define

$$s_{ij} = \frac{(\mathbf{h}_i^p)^\top \mathbf{h}_j^m}{\tau}. \tag{7}$$

Let ℓ_i denote the drug identity of example i. Using drug identity to define the positive set $\mathcal{P}(i) = \{j : \ell_j = \ell_i\}$, the transcriptome-to-molecule supervised contrastive loss is

$$\mathcal{L}_{p\to m} = -\frac{1}{B}\sum_{i=1}^{B}\frac{1}{|\mathcal{P}(i)|}\sum_{j\in\mathcal{P}(i)}\log\frac{\exp(s_{ij})}{\sum_{k=1}^{B}\exp(s_{ik})}. \tag{8}$$

The reverse direction $\mathcal{L}_{m\to p}$ is defined analogously, giving

$$\mathcal{L}_{\text{align}} = \frac{1}{2}\left(\mathcal{L}_{p\to m} + \mathcal{L}_{m\to p}\right). \tag{9}$$

This alignment is applied only in the auxiliary semantic space; the molecular encoder remains fixed.

The condition-usage loss explicitly encourages dependence on the transcriptomic condition. For each retained training example, we evaluate the same noisy latent $\mathbf{z}_t$, timestep t, and target $\mathbf{z}_0$ under the matched, null, and shuffled conditions, yielding reconstruction errors e_i^+, $e_i^\varnothing$, and e_i^{shuf}. With retained-example set $\mathcal{K}$ and adaptive margin

$$m = \gamma\frac{1}{|\mathcal{K}|}\sum_{i\in\mathcal{K}} e_i^+, \tag{10}$$

where γ is the margin ratio, we define

$$\mathcal{L}_{\text{cond}} = \frac{1}{|\mathcal{K}|}\sum_{i\in\mathcal{K}}\frac{1}{2}\left[\left(e_i^+ - e_i^{\text{shuf}} + m\right)_+ + \left(e_i^+ - e_i^\varnothing + m\right)_+\right], \tag{11}$$

where $(x)_+ = \max(0, x)$. Thus, matched conditions are encouraged to yield lower latent reconstruction error than shuffled or null conditions.

The complete objective is

$$\mathcal{L}_{\text{gen}} = \mathcal{L}_{\text{diff}} + \lambda_{\text{align}}\mathcal{L}_{\text{align}} + \lambda_{\text{cond}}\mathcal{L}_{\text{cond}}, \tag{12}$$

with both auxiliary weights linearly warmed up to 0.1. Further details are provided in Appendix C.

At inference, DDIM sampling starts from $\mathbf{z}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and iteratively denoises it under the desired transcriptomic condition. The resulting latent is converted into a molecule by the adapted decoder.

3.5 PREDICTION-ERROR-AWARE DECODING

The molecular decoder is pretrained on encoder-posterior latents, whereas conditional generation can produce latents that deviate from this training distribution. We therefore adapt the decoder using clean molecular representations, synthetic Gaussian perturbations, and empirical generator residuals.

For a paired training example, we define the generator residual as

$$\mathbf{r}_i = \mathbf{z}_{\text{pred}}^{(i)} - \mathbf{z}_{\text{gt}}^{(i)}, \tag{13}$$

where $\mathbf{z}_{\text{pred}}^{(i)}$ is produced by the conditional generator and $\mathbf{z}_{\text{gt}}^{(i)}$ is the encoder-produced latent of the corresponding reference molecule.

To expose the decoder to generator-like deviations, we perturb clean molecular latents using sampled empirical residuals:

$$\tilde{\mathbf{z}}_{\text{emp}} = \mathbf{z}_0 + \lambda_{\text{emp}}\mathbf{r}, \qquad \mathbf{r} \sim \mathcal{R}, \tag{14}$$

where $\mathcal{R}$ denotes the residual bank.

The adapted decoder learns to reconstruct the original molecular sequence from both clean and perturbed latents. To retain its behavior on clean inputs, we additionally distill the token distributions of the frozen pretrained decoder into the adapted decoder, yielding

$$\mathcal{L}_{\text{dec}} = w_n\mathcal{L}_{\text{noisy}} + w_c\mathcal{L}_{\text{clean}} + w_d\mathcal{L}_{\text{distill}}. \tag{15}$$

The first two terms are token-level cross-entropy losses. The third is teacher-to-student KL divergence evaluated on clean latents. Only decoder-side parameters are updated during adaptation; the molecular encoder and shared input embedding remain frozen. This procedure is designed to improve tolerance to generator-induced latent deviations while retaining clean-latent reconstruction fidelity. The perturbation mixture and loss definitions are detailed in Appendix D.

4 Experiments

4.1 Experimental Setup

Data and splits. We use the LINCS L1000 dataset (Subramanian et al., 2017) distributed with chemCPA (Hetzel et al., 2022), combining Phase I (GSE92742) and Phase II (GSE70138). Retaining small-molecule treatments and vehicle controls yields 1,034,271 expression profiles across 978 landmark genes. We retain the supplied expression values without additional log transformation or standardization. We first reserve all A375 profiles for cell-line OOD evaluation. Among the remaining profiles, we hold out 135 of 1,354 unique Bemis–Murcko scaffolds, selected with seed 42, for scaffold OOD evaluation. The remaining profiles are randomly divided into training, validation, and test sets at an 8:1:1 ratio. Data preparation and split statistics are detailed in Appendix E.1.

Baselines. We compare DECO against three representative transcriptomics-conditioned molecular generators: TransGEM (Liu et al., 2024), a Transformer-based transcriptome-to-molecule generator; XMolRL (Guo & Liu, 2026), which combines phenotype-guided generation with target-aware reinforcement learning; and CURE (Xu et al., 2026), a perturbation-conditioned diffusion model with cross-modal chemical alignment. Together, these baselines represent three distinct paradigms for connecting transcriptomic conditions to molecular generation: direct conditioning, explicit cross-modal alignment, and coupled representation learning, providing one representative of each rather than multiple variants of a single approach. All evaluated generators are trained or retrained on our dataset splits to ensure a consistent evaluation setting. For XMolRL, we evaluate its phenotype-guided molecular generator without target-aware reinforcement-learning fine-tuning, since our task does not assume a predefined molecular target.

Evaluation protocol. For each transcriptomic condition, we sample $K = 10$ candidates and report validity, paired Fraggle and Morgan similarities, and PRNet-based response MSE and correlation error, with paired metrics reported as Best@10. Because structurally distinct molecules can induce similar transcriptomic responses, paired structural similarity is interpreted as recovery of the observed reference molecule rather than as a complete measure of conditional-generation quality. PRNet metrics assess response consistency under a frozen surrogate predictor. Full metric definitions and invalid-output handling are provided in Appendix E.2.

4.2 Main Results

Table 1 compares DECO with the evaluated baselines under the standard, cell-line OOD, and scaffold OOD settings. DECO achieves the highest paired Fraggle and Morgan similarities across all

Table 1: Main results across standard, cell-line OOD, and scaffold OOD splits. Structural and response metrics are reported as Best@10. Fraggle and Morgan similarities compare generated molecules with the paired reference molecule. Response MSE and correlation error compare PRNet-predicted transcriptomic responses of generated and reference molecules; both errors are multiplied by 10^3. Bold denotes the best value for each metric within each split, determined before rounding. The DECO model is shaded.

Split	Model	Valid. ↑	Structure		Response	
			Fraggle ↑	Morgan ↑	MSE ↓	Corr. Err. ↓
Test	TransGEM	**1.000**	.140	.080	.968	9.402
	XMolRL	.507	.457	.160	.577	5.200
	CURE	.830	.262	.078	.970	8.443
	DECO	.958	**.567**	**.278**	**.328**	**2.863**
Cell OOD	TransGEM	**1.000**	.159	.080	1.032	9.455
	XMolRL	.517	.448	.152	1.067	9.704
	CURE	.838	.274	.078	1.183	9.962
	DECO	.945	**.495**	**.175**	**.617**	**5.410**
Scaffold OOD	TransGEM	**1.000**	.141	.079	.374	3.412
	XMolRL	.461	.408	.130	.453	4.115
	CURE	.873	.271	.075	.747	6.341
	DECO	.961	**.477**	**.160**	**.303**	**2.702**

three splits, indicating improved recovery of the observed reference molecules. Given the non-unique between transcriptomic response and molecular structure illustrated in Figure 1, however, these metrics should be interpreted as measures of reference recovery rather than as a complete assessment of conditional-generation quality.

We therefore additionally evaluate predicted transcriptomic consistency using the frozen PRNet surrogate. DECO achieves the lowest response MSE and correlation error across all three evaluation settings. Relative to the strongest baseline in each split, Best@10 response MSE is reduced by 43.2% on the standard test split, 40.2% on cell-line OOD, and 19.0% on scaffold OOD. These improvements persist under both cell-line and scaffold holdouts, indicating that the gains are not confined to the standard profile-level split.

DECO also maintains 94.5–96.1% validity across the three evaluation settings. Additional set-level analyses show high internal diversity (0.888–0.894), while mean QED and SA remain close to the LINCS reference distribution (Appendix F.1). Qualitative examples of structurally distinct candidates generated under the same transcriptomic condition are provided in Appendix F.2.

4.3 Component Ablations

We examine the contributions of dual-scale transcriptomic conditioning and prediction-error-aware decoder adaptation in Table 2.

Dual-scale transcriptomic conditioning. Removing gene-level conditioning increases response MSE and correlation error across all three splits, indicating that gene-resolved information contributes to response-conditioned generation beyond the global representation. Removing global conditioning also increases both response errors on the standard and scaffold OOD splits, whereas cell-line OOD performance remains nearly unchanged, with slightly lower errors than the full model. The full model additionally achieves higher validity than either single-scale variant across all splits. These results support the complementary roles of the two conditioning pathways, while showing that the benefit of global conditioning varies across evaluation settings.

Prediction-error-aware decoder adaptation. Removing decoder adaptation reduces validity from 95.8% to 92.6% on the standard split, 94.5% to 91.2% on cell-line OOD, and 96.1% to 93.7% on scaffold OOD. Both paired structural similarities also decrease across all splits. The effect extends to predicted response consistency: without adaptation, response MSE increases from 0.328 to 0.512, 0.617 to 0.652, and 0.303 to 0.337, respectively, in the reported 10^3-scaled units. Correla-

Table 2: Component ablations across standard, cell-line OOD, and scaffold OOD splits, using the same evaluation protocol as Table 1. Structural and response metrics are reported as Best@10. PRNet response MSE and correlation error are multiplied by 10^3. Bold denotes the best value for each metric within each split, determined before rounding. The full DECO model is shaded.

Split	Variant	Valid. ↑	Structure		Response	
			Fraggle ↑	Morgan ↑	MSE ↓	Corr. Err. ↓
Test	**DECO**	**.958**	**.567**	**.278**	**.328**	**2.863**
	w/o global condition	.934	.553	.237	.420	3.648
	w/o gene-level condition	.920	.553	.248	.477	4.148
	w/o decoder adaptation	.926	.543	.236	.512	4.450
Cell OOD	**DECO**	**.945**	**.495**	**.175**	.617	5.410
	w/o global condition	.933	.494	.174	**.614**	**5.405**
	w/o gene-level condition	.929	.461	.170	.723	6.425
	w/o decoder adaptation	.912	.481	.171	.652	5.768
Scaffold OOD	**DECO**	**.961**	**.477**	**.160**	**.303**	**2.702**
	w/o global condition	.938	.469	.159	.342	3.081
	w/o gene-level condition	.953	.467	.159	.354	3.134
	w/o decoder adaptation	.937	.467	.156	.337	2.964

tion error follows the same trend. These results support adapting the decoder to generator-produced latent deviations as a means of improving both molecular validity and the structural and predicted-response quality of decoded candidates.

Auxiliary training objectives. We further examine the auxiliary condition-usage and semantic-alignment objectives in Appendix F. The condition-usage objective improves response consistency across all three evaluation settings, whereas the contribution of semantic alignment is setting-dependent. Increasing λ_{align} does not yield monotonic improvements, indicating that stronger cross-modal alignment is not uniformly beneficial for conditional generation and supporting our use of alignment as lightweight auxiliary supervision.

4.4 Effect of Transcriptomic Conditioning

We probe condition usage by evaluating each validation example under matched, shuffled, and null conditions while holding the noisy latent, diffusion timestep, and target latent fixed. Matched conditioning reaches a final MSE of 0.083, compared with 0.115 and 0.097 for shuffled and null conditions, respectively (Figure 3a), and yields lower sample-level error in 79.1% and 71.1% of the corresponding comparisons (Figure 3b). These results indicate systematic use of transcriptomic information.

5 Discussion

Our results support decoupled conditional modeling for transcriptomics-driven molecular generation: DECO learns chemical representations independently and uses cross-modal alignment only as auxiliary supervision. Across all three evaluation settings, it improves paired structural recovery and predicted transcriptomic consistency; conditioning and component analyses support effective condition use, dual-scale conditioning, and prediction-error-aware decoding, while the benefit of alignment remains setting-dependent. Together, these results support response-conditioned generation over an independently learned chemical space for the weak and non-unique response–structure relationship.

Several limitations remain. First, the observed response–structure relationship is specific to the LINCS L1000 data, experimental contexts, and similarity measures considered here, and may differ in other perturbational datasets. Second, PRNet-based response metrics rely on a frozen surrogate predictor and therefore do not establish that generated molecules experimentally induce the desired transcriptomic state. Third, the accessible chemical space remains constrained by the pretrained molecular representation and decoder. Finally, the cell-line and scaffold OOD evaluations measure

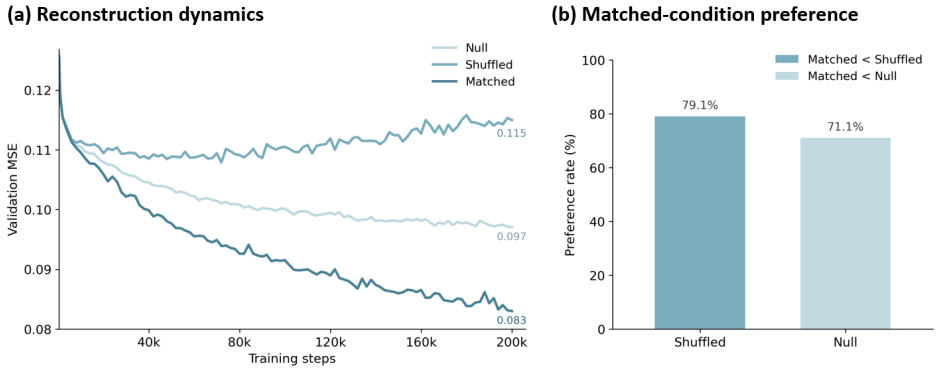


Figure 3: **Effect of transcriptomic conditioning on validation latent reconstruction.** **(a)** Validation latent reconstruction MSE during training under matched, shuffled, and null transcriptomic conditions, with the noisy latent, diffusion timestep, and target latent held fixed. **(b)** Fraction of validation samples at the final checkpoint for which matched conditioning yields lower reconstruction error than shuffled or null conditioning.

generalization under specific held-out settings but do not establish therapeutic efficacy, safety, or synthesizability. Experimental validation and evaluation on broader perturbational datasets remain important directions for future work.

6 Conclusion

We introduce DECO, a decoupled framework that models transcriptomic conditions as distributions over an independently learned chemical latent space. Across standard and OOD settings, dual-scale conditional diffusion and prediction-error-aware decoding improve paired structural recovery and surrogate-predicted response consistency while maintaining high molecular validity.

AI use statement

Generative AI tools were used to assist with language editing, literature retrieval and discovery, research discussion and feedback, and the drafting or revision of selected passages based on author-provided technical content. All AI-assisted content was reviewed, edited, and verified by the authors. The authors determined the final methodology, analyses, and scientific claims and take full responsibility for the final manuscript.

Ethics statement

This work uses previously collected public perturbational gene-expression data and involves no new human-subject recruitment. Generated molecules are computational hypotheses and should not be interpreted as evidence of safety or therapeutic efficacy without experimental validation.

Reproducibility Statement

We provide detailed descriptions of the model architecture, training objectives, data preparation and splits, implementation settings, and evaluation protocol in the main text and appendices. These include the construction of transcriptomic conditions, conditional diffusion and decoder training procedures, auxiliary objectives, sampling configuration, and metric definitions. The source code and configurations required to reproduce the experiments will be made publicly available upon acceptance.

REFERENCES

Guy W. Bemis and Mark A. Murcko. The properties of known drugs. 1. molecular frameworks. *Journal of Medicinal Chemistry*, 39(15):2887–2893, 1996. doi: 10.1021/jm9602928.

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.

Jannis Born, Matteo Manica, Ali Oskooei, Joris Cadow, Greta Markert, and María Rodríguez Martínez. PaccMannRL: De novo generation of hit-like anticancer molecules from transcriptomic data via reinforcement learning. *iScience*, 24(4):102269, 2021. doi: 10.1016/j.isci.2021.102269.

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.

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.

Samuel G. Finlayson, Matthew B. A. McDermott, Alex V. Pickering, Scott L. Lipnick, and Isaac S. Kohane. Cross-modal representation alignment of molecular structure and perturbation-induced transcriptional profiles. *Pacific Symposium on Biocomputing*, 26:273–284, 2021. doi: 10.1142/9789811232701_0026.

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

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 *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=kJqXEPXMsE0.

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

Leon Hetzel, Simon Boehm, Niki Kilbertus, Stephan Günnemann, Mohammad Lotfollahi, and Fabian J Theis. Predicting cellular responses to novel drug perturbations at a single-cell resolution. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=vRrFVHxFiXJ.

Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Junction tree variational autoencoder for molecular graph generation. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 2323–2332. PMLR, 2018.

Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Hierarchical generation of molecular graphs using structural motifs. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 4839–4848. PMLR, 2020.

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.

Hyunho Kim, Bongsung Bae, Minsu Park, Yewon Shin, Trey Ideker, and Hojung Nam. A genotype-to-drug diffusion model for generation of tailored anti-cancer small molecules. *Nature Communications*, 16:5628, 2025. doi: 10.1038/s41467-025-60763-9.

Justin Lamb, Emily D. Crawford, David Peck, Joshua W. Modell, Irene C. Blat, Matthew J. Wrobel, Jim Lerner, Jean-Philippe Brunet, Aravind Subramanian, Kenneth N. Ross, Michael Reich, Haley Hieronymus, Guo Wei, Scott A. Armstrong, Stephen J. Haggarty, Paul A. Clemons, Ru Wei, Steven A. Carr, Eric S. Lander, and Todd R. Golub. The connectivity map: Using gene-expression signatures to connect small molecules, genes, and disease. *Science*, 313(5795):1929–1935, 2006. doi: 10.1126/science.1132939.

Yanguang Liu, Hailong Yu, Xinya Duan, Xiaomin Zhang, Ting Cheng, Feng Jiang, Hao Tang, Yao Ruan, Miao Zhang, Hongyu Zhang, and Qingye Zhang. TransGEM: a molecule generation model based on transformer with gene expression data. *Bioinformatics*, 40(5):btae189, 2024. doi: 10.1093/bioinformatics/btae189.

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. doi: 10.1038/s41467-019-13807-w.

Pascal Notin, José Miguel Hernández-Lobato, and Yarin Gal. Improving black-box optimization in VAE latent space using decoder uncertainty. In *Advances in Neural Information Processing Systems*, volume 34, pp. 802–814, 2021.

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.

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.

Rim Shayakhmetov, Maksim Kuznetsov, Alexander Zhebrak, Artur Kadurin, Sergey Nikolenko, Alexander Aliper, and Daniil Polykovskiy. Molecular generation for desired transcriptome changes with adversarial autoencoders. *Frontiers in Pharmacology*, 11:269, 2020. doi: 10.3389/fphar.2020.00269.

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.

Conghao Wang, Hiok Hian Ong, Shunsuke Chiba, and Jagath C. Rajapakse. GLDM: hit molecule generation with constrained graph latent diffusion model. *Briefings in Bioinformatics*, 25(3):bbae142, 2024. doi: 10.1093/bib/bbae142.

Xi Wang, Jiahan Li, Yuxuan Xia, Yingcheng Wu, Shaoyi Zheng, and Shengjie Wang. Smoothing dark areas in molecular latent diffusion, 2026.

Minkai Xu, Alexander S. Powers, Ron O. Dror, Stefano Ermon, and Jure Leskovec. Geometric latent diffusion models for 3D molecule generation. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 38592–38610. PMLR, 2023.

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.

A Transcriptomic Response–Structure Relationship in the Dataset

A.1 Aggregate Analysis Protocol

For each compound and experimental context, we construct a 978-dimensional perturbational response signature by subtracting a matched vehicle-control prototype from each treated profile and averaging across replicate profiles associated with the same canonical molecule and context. Experimental contexts are defined by exact matches in cell line, treatment time, LINCS phase, perturbation dose, and dose unit.

We retain the 30 largest contexts containing at least 200 molecules and sample 20,000 unique unordered, non-self molecule pairs within each context, yielding 600,000 pairs in total. Transcriptomic response similarity is computed as cosine similarity between ℓ_2-normalized response vectors. Chemical similarity is computed as Tanimoto similarity between 2,048-bit Morgan fingerprints with radius 2.

For the binned analysis, response similarities are assigned to 12 fixed bins, and we report the median and 10th–90th percentile interval of Morgan similarity within each bin. The overall association is summarized by pair-level Spearman correlation across all sampled pairs. Highly response-similar pairs are defined as the top 1% of response similarities within each experimental context and are then pooled across contexts for analysis.

A.2 Qualitative Examples of Phenotype–Structure Non-uniqueness

To complement the aggregate analysis in Figure 1, we further visualize representative examples from the dataset. For each reference compound, we identify compounds with highly similar transcriptomic responses under matched experimental contexts and compare their molecular structures. Figure 4 shows that high transcriptomic similarity can coexist with low chemical similarity across a range of examples, providing an intuitive view of the non-unique relationship between transcriptomic response and molecular structure.

B Architecture and Training Details

Molecular autoencoder. We adopt the SMILES-based molecular VAE architecture introduced by Kim et al. (2025) and retrain it from scratch using molecular reconstruction and KL regularization. The SMILES-based molecular VAE uses three-layer LSTM encoder and decoder networks with hidden dimension 256 and latent dimension $d_z = 128$. Although VAE pretraining uses posterior sampling, downstream molecular representations are deterministic posterior means, $\mathbf{z}_0 = \boldsymbol{\mu}_\phi(m)$. The same representation is used for diffusion targets and clean inputs during decoder adaptation.

Transcriptomic condition encoder. For a minibatch of size B, the control and perturbed transcriptomes have shape $B \times G$. The tokenizer combines expression-value, gene-identity, and state embeddings to produce two token sequences of shape $B \times G \times 64$. Three bidirectional interaction blocks refine these sequences through cross-state attention, within-state self-attention, and feed-forward transformations.

The global pathway independently pools the control and perturbed sequences, combines their pooled representations through adaptive fusion, and applies a two-layer MLP to produce $\mathbf{C}_{\text{global}} \in \mathbb{R}^{B\times 256}$. The gene-level pathway concatenates the refined control representation, perturbed representation, and their difference, producing features of shape $B \times G \times 192$. A projection network with the structure $\text{Linear} \to \text{GELU} \to \text{LayerNorm} \to \text{Linear}$ maps these features to $\mathbf{C}_{\text{gene}} \in \mathbb{R}^{B\times G\times 256}$.

Conditional denoiser and sampling. The latent Transformer predicts the clean molecular latent rather than the injected Gaussian noise. Its global conditioning pathway uses adaptive normalization, while cross-attention accesses the gene-level condition tokens. At inference, DDIM sampling begins from a standard Gaussian latent and uses the same transcriptomic condition throughout the reverse process.

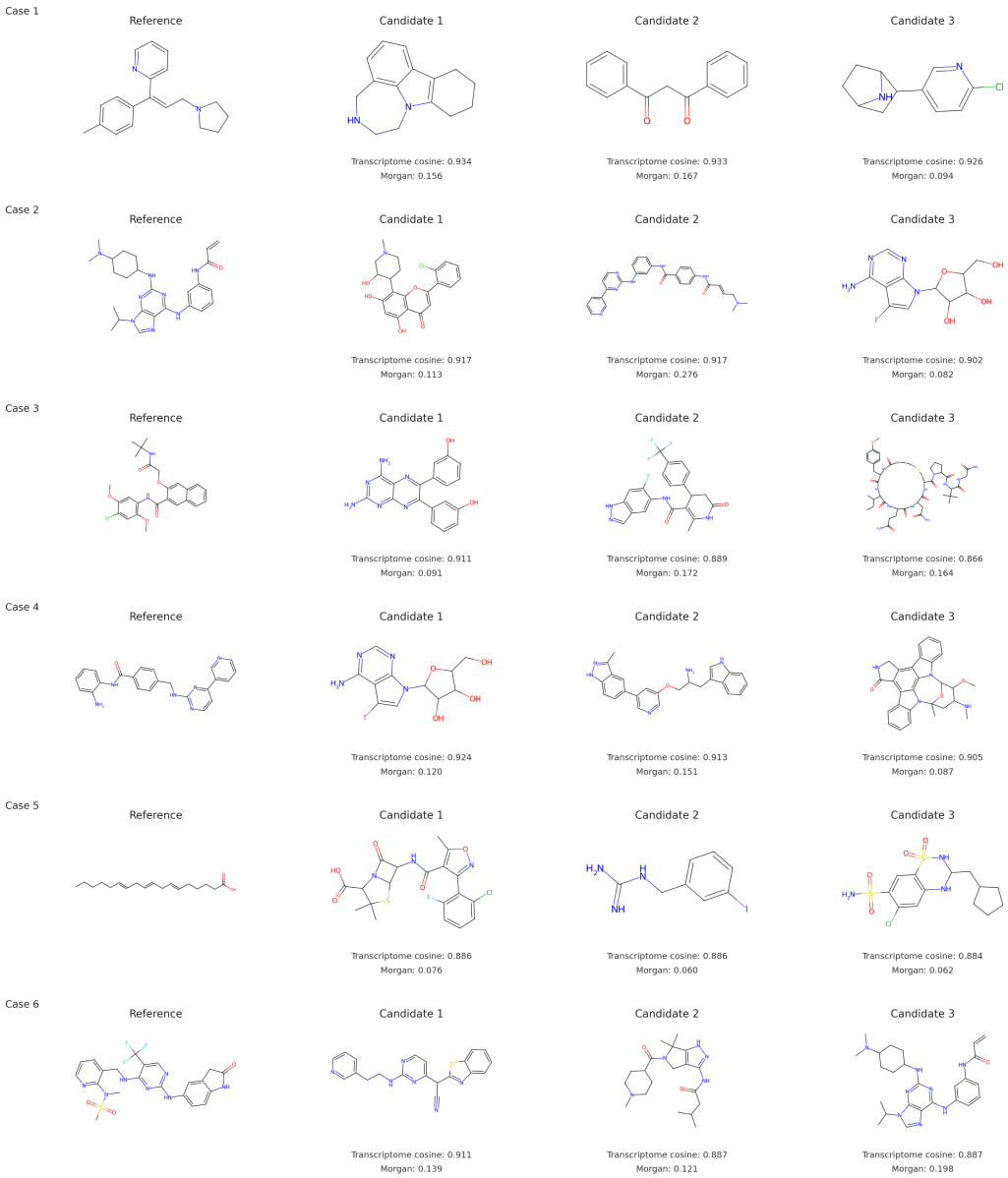


Figure 4: **Representative examples of response-similar but structurally diverse compounds in the dataset.** Each row shows a reference compound and three compounds with highly similar transcriptomic responses under matched experimental contexts. Response cosine denotes cosine similarity between the corresponding perturbation-induced expression-change vectors, while Morgan similarity measures molecular structural similarity. Across the six examples, compounds with response cosine similarities typically around 0.87–0.93 can exhibit substantially lower Morgan similarities, illustrating that similar transcriptional responses do not necessarily correspond to closely related molecular structures. All compounds shown here are retrieved from the dataset and are not generated by DECO.

Training stages. Chemical representation learning precedes conditional generator training. During conditional training, the molecular encoder is fixed. Decoder adaptation uses an empirical residual bank obtained from the conditional generator and updates only decoder-side parameters, while retaining a frozen copy of the pretrained decoder for distillation.

C Auxiliary Objectives for Conditional Generation

C.1 Semantic Alignment

The semantic alignment objective operates in an auxiliary semantic space through projections of the global phenotype representation and molecular latent. Let ℓ_i denote the drug identity of example i in a minibatch. Its cross-modal positive set is

$$\mathcal{P}(i) = \{j : \ell_j = \ell_i\}. \tag{16}$$

Bidirectional supervised contrastive learning uses drug identity to define positive sets in both phenotype-to-molecule and molecule-to-phenotype directions. Multiple transcriptomic responses associated with the same drug are therefore treated as positives, while samples with different drug identities are treated as negatives.

This identity-based objective provides auxiliary discriminative supervision but does not capture all potentially valid response-level equivalences between chemically distinct molecules. We therefore use it only as auxiliary supervision to the conditional generative objective. The molecular encoder remains fixed throughout alignment training.

Only examples with valid drug labels ($\ell_i \geq 0$) participate in the alignment objective. Both projection outputs are ℓ_2-normalized before computing cross-modal similarities, and the same temperature τ is used in both directions. If fewer than two valid examples are available in a minibatch, $\mathcal{L}_{\text{align}}$ is set to zero.

C.2 Condition Usage

For each training example, matched, shuffled, and null conditions are evaluated using the same noisy latent $\mathbf{z}_t$, timestep t, and clean target $\mathbf{z}_0$. Let s^+, s^{shuf}, and $s^{\varnothing}$ denote the matched, shuffled, and null conditions, respectively. Their per-example latent reconstruction errors are

$$e_i^a = \text{MSE}(f_\theta(\mathbf{z}_{t,i}, t_i, s_i^a), \mathbf{z}_{0,i}), \qquad a \in \{+, \text{shuf}, \varnothing\}. \tag{17}$$

The condition-usage objective is evaluated on the subset $\mathcal{K}$ of examples for which transcriptomic conditioning is retained during condition dropout. An adaptive margin is defined from the matched-condition reconstruction error as

$$m = \gamma \frac{1}{|\mathcal{K}|} \sum_{i \in \mathcal{K}} e_i^+, \tag{18}$$

where γ is the margin ratio. For each retained example, the shuffled- and null-condition ranking terms are

$$\ell_i^{\text{shuf}} = \max\big(0,\, e_i^+ - e_i^{\text{shuf}} + m\big), \tag{19}$$

$$\ell_i^{\varnothing} = \max\big(0,\, e_i^+ - e_i^{\varnothing} + m\big). \tag{20}$$

The resulting condition-usage loss is

$$\mathcal{L}_{\text{cond}} = \frac{1}{|\mathcal{K}|} \sum_{i \in \mathcal{K}} \frac{1}{2} \left(\ell_i^{\text{shuf}} + \ell_i^{\varnothing}\right). \tag{21}$$

Because the noisy latent, diffusion timestep, and reconstruction target are shared across the three branches, the ranking objective isolates the effect of changing the transcriptomic condition. It therefore encourages the matched condition to provide more useful information for molecular-latent reconstruction than shuffled or null conditioning.

Both auxiliary objective weights are linearly warmed up to their default values, $\lambda_{\text{align}} = \lambda_{\text{cond}} = 0.1$.

D Prediction-Error-Aware Decoder Adaptation

D.1 Empirical Residual Construction

For each paired example, let $\mathbf{z}_{\mathrm{gt}}^{(i)} = \boldsymbol{\mu}_\phi(m_i)$ be the posterior mean of the reference molecule, and let $\mathbf{z}_{\mathrm{pred}}^{(i)}$ be the latent produced by a fixed conditional-generator sampling trajectory under the corresponding transcriptomic condition. We construct the residual bank as

$$\mathcal{R} = \left\{ \mathbf{r}_i = \mathbf{z}_{\mathrm{pred}}^{(i)} - \mathbf{z}_{\mathrm{gt}}^{(i)} \right\}_{i=1}^{N_{\mathcal{R}}}, \tag{22}$$

where $N_{\mathcal{R}}$ is the number of stored residuals.

During decoder adaptation, residuals are randomly sampled from this bank and reused across molecular examples:

$$\tilde{\mathbf{z}}_{\mathrm{emp}} = \mathbf{z}_0 + \lambda_{\mathrm{emp}} \mathbf{r}, \qquad \mathbf{r} \sim \mathcal{R}. \tag{23}$$

Consequently, an injected residual need not originate from the molecule being reconstructed. This procedure transfers empirical perturbation directions and magnitudes to clean molecular latents; it does not model a separate residual distribution for each transcriptomic condition.

Because the inverse task admits multiple chemical solutions, these residuals measure deviation from paired reference latents rather than establishing that the generated molecules are biologically incorrect. Their role is to supply generator-derived perturbations for decoder adaptation.

The residual bank is constructed from the training split using the corresponding best conditional-generator checkpoint selected by validation loss, with 100 DDIM sampling steps, $\eta = 0$, no classifier-free guidance, and random seed 42, yielding $N_{\mathcal{R}} = 128$ residuals.

D.2 Latent Perturbation Mixture

For each clean latent $\mathbf{z}_0$, the augmented input is drawn from a mixture of clean, synthetic, and empirical perturbations:

$$\tilde{\mathbf{z}} = \begin{cases} \mathbf{z}_0, & \text{with probability } 0.55, \\ \mathbf{z}_0 + \sigma \boldsymbol{\epsilon}, & \text{with probability } 0.25, \\ \mathbf{z}_0 + 0.30\, \mathbf{r}, & \text{with probability } 0.20, \end{cases} \tag{24}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $\mathbf{r} \sim \mathcal{R}$, and σ is sampled from a predefined set of noise levels. Thus, the empirical residual scale is $\lambda_{\mathrm{emp}} = 0.30$.

A separate clean-input reconstruction branch is retained for every training example. Accordingly, the clean component of the augmentation mixture is additional to the clean reconstruction term in the training objective.

D.3 Reconstruction and Distillation Objectives

Let $y = (y_1, \ldots, y_L)$ be the target molecular token sequence. We denote the frozen pretrained decoder by p_T and the adapted student decoder by p_S. The reconstruction terms are token-level cross-entropy losses:

$$\mathcal{L}_{\mathrm{noisy}} = -\frac{1}{L} \sum_{k=1}^{L} \log p_S(y_k \mid y_{<k}, \tilde{\mathbf{z}}), \tag{25}$$

$$\mathcal{L}_{\mathrm{clean}} = -\frac{1}{L} \sum_{k=1}^{L} \log p_S(y_k \mid y_{<k}, \mathbf{z}_0). \tag{26}$$

Here, $\mathcal{L}_{\mathrm{noisy}}$ denotes the augmented-input branch, whose input can also be clean under the mixture above.

Table 3: Expression-profile counts in each split. These counts describe the complete dataset partitions, rather than the sampled generation-evaluation subsets.

Split	Profiles
Train	724,403
Validation	90,550
Test	90,551
Cell OOD	91,542
Scaffold OOD	37,225
Total	**1,034,271**

Distillation is evaluated exclusively on clean latents. Teacher and student token distributions are computed from their logits at temperature $\tau = 1.0$, and the student is regularized through teacher-to-student KL divergence:

$$\mathcal{L}_{\text{distill}} = \frac{1}{L} \sum_{k=1}^{L} D_{\text{KL}}(p_T(\cdot \mid y_{<k}, \mathbf{z}_0) \,\|\, p_S(\cdot \mid y_{<k}, \mathbf{z}_0)) \,. \tag{27}$$

The complete decoder objective is

$$\mathcal{L}_{\text{dec}} = w_n \mathcal{L}_{\text{noisy}} + w_c \mathcal{L}_{\text{clean}} + w_d \mathcal{L}_{\text{distill}}, \qquad (w_n, w_c, w_d) = (1.0, 1.0, 2.0). \tag{28}$$

The teacher decoder remains frozen. Only student decoder-side parameters are updated; the molecular encoder and shared input embedding are fixed. The augmented-input reconstruction term promotes tolerance to latent perturbations, while clean reconstruction and distillation retain supervision on the original molecular representations.

E Additional Experimental Details

E.1 Data Preparation and Splits

We use the `lincs_full.h5ad` dataset distributed with chemCPA (Hetzel et al., 2022), containing LINCS L1000 Phase I (GSE92742) and Phase II (GSE70138) profiles (Subramanian et al., 2017). We retain small-molecule treatments (`trt_cp`) and vehicle controls (`ctl_vehicle`). The resulting dataset contains 1,034,271 expression profiles over 978 landmark genes. Expression values are used on their supplied scale, without additional log transformation or standardization.

Splits are constructed sequentially. First, all profiles from the A375 cell line are assigned to the cell-line OOD split. Next, Bemis–Murcko scaffolds (Bemis & Murcko, 1996) are computed from canonical molecular SMILES. Among the non-A375 data, 135 of the 1,354 valid unique scaffolds are randomly selected with seed 42. All profiles associated with these scaffolds are assigned to the scaffold OOD split. The remaining profiles are randomly partitioned into training, validation, and test sets at an 8:1:1 ratio. Table 3 reports profile counts, including both treatment and control profiles.

E.2 Metric Definitions

Validity and canonicalization. Chemical processing uses RDKit. An output is valid if it can be parsed and sanitized and contains a single connected molecular component. Empty outputs, decoder failures, invalid SMILES, and multi-component structures are counted as invalid. Validity is

$$\text{Validity} = \frac{N_{\text{valid}}}{N_{\text{attempts}}}. \tag{29}$$

Valid outputs are converted to canonical isomeric SMILES for duplicate detection.

Uniqueness and diversity. Uniqueness is computed over the valid outputs of each model within each evaluation split:

$$\text{Uniqueness} = \frac{N_{\text{unique}}}{N_{\text{valid}}}. \tag{30}$$

For the set $\mathcal{U}$ of unique valid molecules, internal diversity is

$$\text{Diversity} = 1 - \frac{2}{|\mathcal{U}|(|\mathcal{U}|-1)} \sum_{\substack{a,b\in\mathcal{U} \\ a<b}} \text{Tan}\big(\text{FP}(a), \text{FP}(b)\big), \tag{31}$$

where FP denotes a Morgan fingerprint with radius 2 and 2,048 bits, and Tan is Tanimoto similarity. This statistic measures diversity across the generated set, rather than diversity within each individual condition.

Chemical properties. We report mean QED (Bickerton et al., 2012) and synthetic accessibility (SA) score (Ertl & Schuffenhauer, 2009) over unique valid molecules. QED uses RDKit's default weighted implementation. SA uses the raw Ertl–Schuffenhauer score, with lower values indicating greater estimated synthetic accessibility. Both are descriptive properties of the generated molecular distribution rather than optimization objectives.

Paired structural similarity. Let m_i^* denote the reference molecule for condition i, and let $m_{i1}, \ldots, m_{iK}$ be its generated candidates, where $K = 10$. For a structural similarity function S, we compute

$$S_i^{\text{Best@10}} = \max_{1\le k\le K} S(m_{ik}, m_i^*). \tag{32}$$

Invalid candidates receive similarity zero. Morgan similarity is the Tanimoto similarity between radius-2, 2,048-bit Morgan fingerprints. Fraggle similarity uses RDKit's `GetFraggleSimilarity` implementation with `tverskyThresh=0.8`, passing the generated molecule as the first argument and the reference as the second. Scores are averaged over evaluated conditions. These metrics assess recovery of the paired reference structure; they do not identify all chemically distinct molecules that could produce a compatible response.

Predicted response consistency. We use the LINCS-trained PRNet (Qi et al., 2024) checkpoint `lincs_best_epoch_all.pt` to predict transcriptional profiles for generated and reference molecules. PRNet parameters remain frozen throughout evaluation. Let $\hat{\mathbf{y}}_{ik}$ and $\hat{\mathbf{y}}_i^*$ denote these predictions, respectively. For a condition with a nonempty set $\mathcal{V}_i$ of valid, evaluable candidates, we compute

$$\text{MSE}_i^{\text{Best@10}} = \min_{k\in\mathcal{V}_i} \frac{1}{D} \|\hat{\mathbf{y}}_{ik} - \hat{\mathbf{y}}_i^*\|_2^2, \tag{33}$$

$$\text{CorrErr}_i^{\text{Best@10}} = 1 - \max_{k\in\mathcal{V}_i} \text{PCC}\left(\hat{\mathbf{y}}_{ik}, \hat{\mathbf{y}}_i^*\right). \tag{34}$$

Here, D is the number of genes in the PRNet output profile, and PCC is computed across these genes. Each metric selects its best candidate independently. Condition-level response scores are averaged and multiplied by 10^3 for reporting. Response metrics are defined only for conditions with at least one valid, evaluable candidate, i.e., $|\mathcal{V}_i| > 0$. All reported evaluations satisfy this requirement; otherwise, no aggregate response metric is produced for the corresponding split. This prediction-to-prediction evaluation measures consistency under PRNet, rather than experimentally verified induction of the desired transcriptomic state.

F Additional Experimental Results

F.1 Generation Quality and Chemical Properties

Table 4 complements the main results with set-level generation statistics and chemical properties. DECO maintains internal diversity of 0.888–0.894 across splits, approaching XMolRL and exceeding TransGEM and CURE. Its uniqueness is substantially higher than TransGEM's, although it falls to 0.816 on scaffold OOD. Mean QED and SA remain close to the LINCS reference values, particularly on the OOD splits. Together with the main results, these statistics show that improved response consistency is accompanied by diverse molecular outputs and chemical properties close to the dataset reference.

Table 4: Generation quality and chemical properties across evaluation splits. Uniqueness is computed over valid outputs; diversity, QED, and SA are computed over unique valid molecules. QED and SA are descriptive statistics rather than optimization objectives. Bold denotes the best uniqueness and diversity within each split, determined before rounding. The full DECO model is shaded. The LINCS reference has mean QED $= 0.566$ and SA $= 3.232$.

Split	Model	Generation		Chemical Properties	
		Unique. ↑	Diversity ↑	QED	SA
Test	TransGEM	.128	.714	.591	2.895
	XMolRL	.982	**.894**	.608	2.843
	CURE	**1.000**	.835	.192	5.120
	DECO	.948	.888	.556	3.346
Cell OOD	TransGEM	.156	.735	.592	2.937
	XMolRL	.980	**.898**	.601	2.961
	CURE	**1.000**	.834	.195	5.086
	DECO	.974	.893	.573	3.257
Scaffold OOD	TransGEM	.130	.708	.588	2.921
	XMolRL	.978	**.896**	.607	2.942
	CURE	**1.000**	.835	.168	5.204
	DECO	.816	.894	.567	3.223
LINCS reference		–	–	.566	3.232

F.2 Qualitative Examples of Generated Molecular Candidates

Figure 5 presents six selected cases, each containing a reference molecule and three candidates generated under the same transcriptomic condition. Cases are selected to include a candidate with relatively high reference similarity alongside candidates with different scaffolds, while retaining low PRNet-predicted response MSE. These examples complement the aggregate results by illustrating structural variation among candidates with similar predicted responses.

F.3 Auxiliary Objectives and Alignment-Weight Sensitivity

Auxiliary training objectives. Table 5 examines the auxiliary alignment and condition-usage objectives. Removing $\mathcal{L}_{\text{cond}}$ increases response MSE by 3.5%, 8.7%, and 11.4% on the standard, cell-line OOD, and scaffold OOD splits, respectively, with higher correlation error in all three settings. This supports the contribution of explicit condition-usage supervision to predicted response consistency. Removing $\mathcal{L}_{\text{align}}$ increases response MSE by 25.0% on the standard split and 10.4% on scaffold OOD, while also reducing standard-split structural similarity. On cell-line OOD, however, removing alignment yields lower response errors, indicating that its benefit depends on the evaluation setting.

Alignment-weight sensitivity. We vary λ_{align} over $\{0, 0.1, 0.5, 1, 3\}$ while keeping $\lambda_{\text{cond}} = 0.1$ (Table 6). The default weight of 0.1 achieves the lowest response MSE and correlation error on the standard and scaffold OOD splits. It also yields the lowest unweighted mean response MSE across the three splits: 0.416, compared with 0.445, 0.441, 0.428, and 0.444 for weights 0, 0.5, 1, and 3, respectively, in the reported 10^3-scaled units. Cell-line OOD exhibits a different pattern, with the lowest MSE at zero alignment weight and the lowest correlation error at 1. Thus, increasing alignment strength does not consistently improve response consistency. These results support using alignment as a lightly weighted auxiliary objective in the evaluated configuration, rather than assuming that stronger correspondence constraints are preferable.

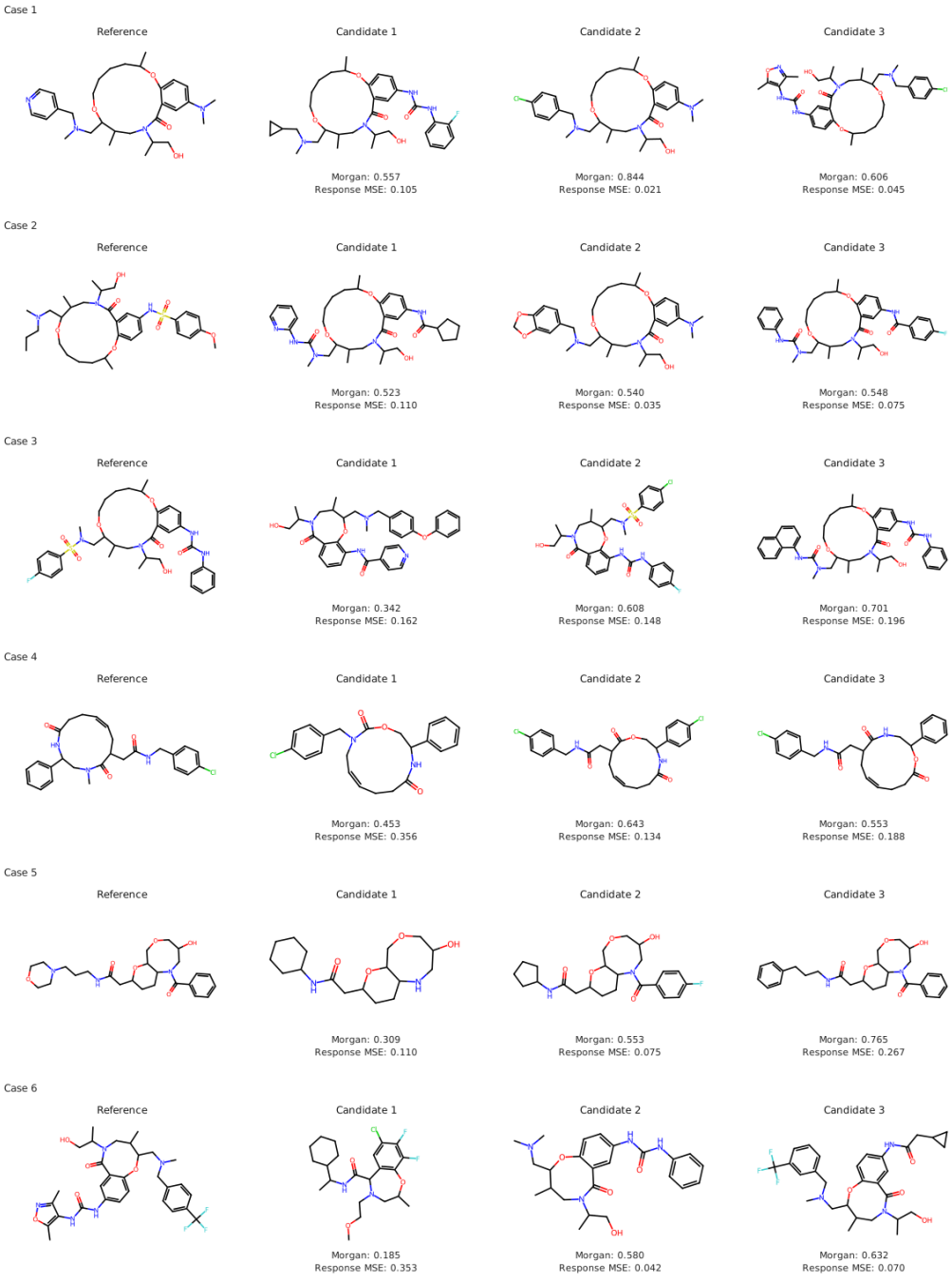


Figure 5: **Selected molecular candidates generated under shared transcriptomic conditions.** Each row shows a reference molecule and three candidates generated under the same condition. Candidates are selected to illustrate both reference-like structures and different scaffolds with low PRNet-predicted response MSE. Morgan similarity and response MSE are reported for each displayed candidate relative to the reference molecule and its PRNet-predicted response, respectively. Metric definitions are provided in Appendix E.2. Response MSE is multiplied by 10^3.

Table 5: Auxiliary-loss ablations across evaluation splits. The full model uses $\lambda_{\text{align}} = \lambda_{\text{cond}} = 0.1$; each ablation sets the corresponding weight to zero. Structural and response metrics are reported as Best@10. PRNet response MSE and correlation error are multiplied by 10^3. Bold denotes the best value for each metric within each split, determined before rounding. The full model is shaded.

Split	Variant	Valid. ↑	Structure		Response	
			Fraggle ↑	Morgan ↑	MSE ↓	Corr. Err. ↓
Test	**DECO**	**.958**	.567	**.278**	**.328**	**2.863**
	w/o $\mathcal{L}_{\text{align}}$	.952	.531	.214	.410	3.589
	w/o $\mathcal{L}_{\text{cond}}$	.950	**.569**	.270	.340	2.956
Cell OOD	**DECO**	.945	.495	.175	.617	5.410
	w/o $\mathcal{L}_{\text{align}}$	**.946**	**.497**	**.175**	**.591**	**5.293**
	w/o $\mathcal{L}_{\text{cond}}$	.936	.489	.170	.671	5.942
Scaffold OOD	**DECO**	.961	.477	.160	**.303**	**2.702**
	w/o $\mathcal{L}_{\text{align}}$	.955	**.479**	.160	.334	2.981
	w/o $\mathcal{L}_{\text{cond}}$	**.974**	.469	**.168**	.338	2.971

Table 6: Alignment-weight sensitivity across evaluation splits, with $\lambda_{\text{cond}} = 0.1$ fixed. Structural and response metrics are reported as Best@10. PRNet response MSE and correlation error are multiplied by 10^3. Bold denotes the best value for each metric within each split, determined before rounding. The default weight of 0.1 is shaded.

Split	λ_{align}	Valid. ↑	Structure		Response	
			Fraggle ↑	Morgan ↑	MSE ↓	Corr. Err. ↓
Test	0	.952	.531	.214	.410	3.589
	0.1	.958	.567	**.278**	**.328**	**2.863**
	0.5	.953	.560	.247	.406	3.532
	1	.948	**.575**	.258	.374	3.227
	3	**.959**	.561	.248	.381	3.333
Cell OOD	0	.946	.497	.175	**.591**	5.293
	0.1	.945	.495	.175	.617	5.410
	0.5	.948	.500	.186	.600	5.272
	1	.951	**.504**	**.188**	.593	**5.249**
	3	**.957**	.495	.172	.621	5.571
Scaffold OOD	0	.955	.479	.160	.334	2.981
	0.1	.961	.477	.160	**.303**	**2.702**
	0.5	.949	.468	.165	.317	2.812
	1	**.964**	.476	.162	.317	2.756
	3	**.964**	**.491**	**.165**	.328	2.925