MECHMOE: DISCOVERING TASK-ALIGNED PERTURBATIONAL MECHANISMS FOR OUT-OF-DISTRIBUTION MOLECULAR PERTURBATION PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Predicting transcriptional responses to molecular perturbations is a key challenge in computational drug discovery, particularly when generalizing to unseen compounds and cellular contexts. Although biological priors such as chemical similarity, annotated targets, and mechanism-of-action labels can provide useful structure, small-molecule responses often arise through indirect, context-dependent, and off-target effects that are only partially captured by predefined annotations. We introduce MechMoE, a mechanism-aware mixture model for out-of-distribution molecular perturbation prediction. MechMoE encodes molecular perturbations, basal cellular state, and expression-conditioned gene tokens, then uses learnable mechanism queries to extract candidate perturbational programs from contextualized gene representations. A sparse router selects a small subset of mechanism slots for each drug-cell pair, producing a perturbation-specific mechanism mixture used to predict the perturbation-induced expression change. We evaluate MechMoE on molecular perturbation data under strict held-out molecule and held-out cell-line settings. MechMoE improves transcriptome-wide response prediction across correlation, cosine similarity, reconstruction error, direction accuracy, and perturbation discrimination metrics, while remaining competitive on differential-expression recovery. Ablations show that sparse routing and the global perturbation decoder are both important for performance. These results suggest that small-molecule perturbation responses are effectively modeled as sparse combinations of learned, task-aligned latent mechanisms rather than fixed predefined biological categories.

1 INTRODUCTION

Predicting cellular responses to molecular perturbations is a central challenge in computational biology and drug discovery. Accurate models of perturbation-induced gene expression changes have the potential to accelerate therapeutic development, prioritize candidate compounds, identify mechanisms of action, and enable *in silico* exploration of cellular responses across diverse biological contexts. Recent advances in large-scale perturbation atlases, including single-cell and high-throughput transcriptomic profiling experiments, have made it possible to train increasingly expressive machine learning models that map molecular structure and cellular state to downstream transcriptional responses.

Despite this progress, perturbation prediction remains fundamentally difficult. The primary challenge is one of out-of-distribution generalization. In practical drug discovery settings, models are expected to predict responses for compounds that have never been experimentally observed and in cellular contexts where perturbation measurements are unavailable. Existing methods have made significant progress on perturbation modeling Adduri et al. (2025); Dong et al. (2026); Maleki et al. (2025), but many evaluations remain restricted to partially observed settings in which perturbations from a held-out context are still available during training. As illustrated in Figure 1, realistic deployment requires robustness along both axes: to previously unseen molecules and to cellular contexts not observed during training.

A common strategy for improving generalization and interpretability is to incorporate biological prior knowledge into the learning process. Examples include pathway annotations, drug-target

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.

relationships, mechanism-of-action labels, gene regulatory networks, and curated biological concepts. The underlying assumption is that these abstractions provide a biologically meaningful basis for organizing perturbational responses. However, whether such priors are actually aligned with the prediction objective remains unclear. In principle, concepts that are useful for human interpretation are not necessarily the concepts that best explain transcriptomic response variation.

To investigate this question, we first analyze the relationship between several widely used biological priors and transcriptomic perturbation similarity. Surprisingly, we find that chemical similarity, shared protein targets, and shared mechanism-of-action annotations explain only a small fraction of the variance in perturbational response similarity (Figure 1C). These results suggest that predefined biological concepts are only weakly aligned with the structure most relevant for perturbation prediction. While such annotations remain valuable for downstream interpretation, they may not provide an optimal inductive bias for learning predictive models of cellular response.

Motivated by these observations, we ask whether perturbation models should instead learn the latent biological programs most useful for prediction directly from data. We introduce **MechMoE**, a mechanism-aware mixture-of-experts architecture for transcriptomic perturbation prediction. MechMoE combines gene-level transformer representations with sparse mechanism routing to discover reusable perturbational programs without relying on predefined pathway supervision. Drug representations, baseline cellular states, and gene-level features are jointly contextualized through self-attention, after which sparse expert routing identifies a small subset of latent mechanisms that explain the predicted response. Unlike curated pathways or manually defined biological concepts, these mechanisms emerge directly from the perturbation prediction objective and are optimized end-to-end.

We further formulate perturbation prediction as a residual modeling problem. Rather than reconstructing treated expression profiles from scratch, MechMoE predicts perturbation-induced changes relative to the observed baseline state. This counterfactual-style formulation explicitly models the perturbational effect and encourages the model to separate basal cellular identity from treatment-induced transcriptional responses.

Experimentally, we evaluate MechMoE on challenging out-of-distribution perturbation prediction tasks involving both unseen molecules and fully unseen cell lines. In contrast to prior evaluations Adduri et al. (2025) that expose a subset of perturbations from held-out contexts during training, our cell-line benchmark excludes held-out cellular contexts entirely. Across both generalization settings, MechMoE achieves strong performance on transcriptomic reconstruction, perturbation retrieval, and differential expression recovery. Moreover, post hoc analysis reveals that the learned experts correspond to biologically coherent transcriptional programs, including stress-response, apoptosis-associated, and lineage-specific regulatory states, despite the absence of explicit pathway supervision. Our contributions are summarized as follows: (1) We introduce MechMoE, a mechanism-aware mixture-of-experts architecture that discovers sparse latent perturbational programs directly from data. (2) We evaluate perturbation prediction under challenging unseen-molecule and fully unseen-cell-line settings and demonstrate strong performance across transcriptomic and biological metrics. (3) We show that the learned latent mechanisms are interpretable, reusable across perturbations, and organize biologically meaningful transcriptional response programs.

2 Related Work

Perturbation response prediction. A growing body of work has developed generative and representation-learning models for predicting cellular responses to perturbations. Variational and disentangled latent-variable models, such as CPA (Hetzel et al., 2022), BioLORD (Piran et al., 2024), and SAMS-VAE (Bereket & Karaletsos, 2024), aim to separate basal cellular state from perturbation-induced effects and learn representations that support counterfactual prediction. SAMS-VAE is particularly related to our work in that it models perturbations through sparse additive latent mechanism shifts. However, these approaches typically operate at the level of sample- or perturbation-level latent variables, whereas MechMoE learns sparse perturbational mechanisms through gene-level contextualization and mechanism routing. This allows the model to capture both global perturbation structure and gene-specific transcriptional responses.

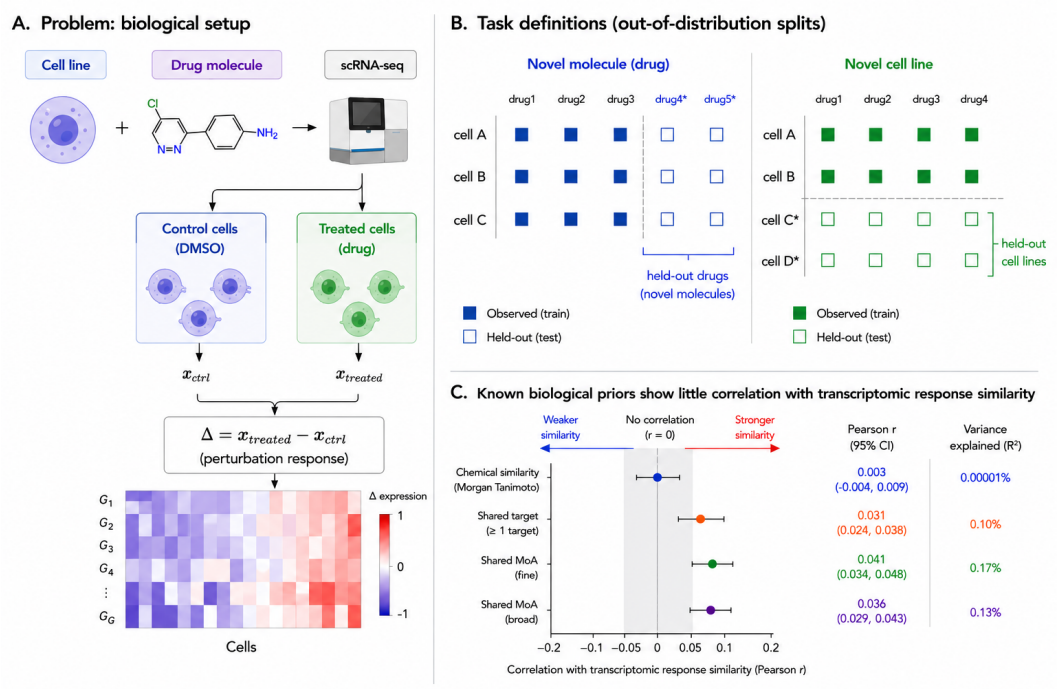

Figure 1: MechMoE addresses perturbation prediction under realistic out-of-distribution settings. **(A)** Biological setup: Given a cell line and a drug molecule, single-cell RNA sequencing is performed on control (DMSO) and drug-treated cells. The perturbation response is defined as the gene-level expression change $\Delta = x_{\text{treated}} - x_{\text{ctrl}}$. **(B)** Out-of-distribution generalization tasks: We evaluate perturbation prediction under two challenging settings: *novel molecule* and *novel cell line*. **(C)** Existing biological priors explain little of transcriptomic response similarity.

Single-cell foundation models. Recent single-cell foundation models have leveraged large-scale transcriptomic corpora to learn general-purpose cell and gene representations. Models such as scGPT (Cui et al., 2024), scFoundation (Hao et al., 2024), UCE (Rosen et al., 2023), Transcriptformer (Rosen et al., 2023), and STATE (Adduri et al., 2025) demonstrate that large-scale pretraining can improve transfer across datasets and biological contexts. These models provide powerful general representations, but many are not designed specifically for generalizing to previously unseen small molecules, and evaluations often focus on settings where some perturbation information from the target context is observed. In contrast, MechMoE focuses on out-of-distribution molecular perturbation prediction, including unseen molecules and fully held-out cell lines.

Knowledge-guided perturbation models. Another line of work incorporates prior biological knowledge into perturbation models. GEARS (Roohani et al., 2024), for example, uses a knowledge graph of gene–gene relationships to predict transcriptional outcomes of genetic perturbations, including novel multigene perturbations. Such priors are well aligned with genetic perturbation settings because the perturbed entities are genes, which are themselves nodes in known regulatory or interaction networks.

Molecular perturbation prediction presents a different challenge. Small molecules affect cells through binding, off-target effects, pathway modulation, metabolism, toxicity, and context-dependent cellular state changes. As a result, commonly available priors such as chemical similarity, annotated targets, and mechanism-of-action labels may only partially describe the downstream transcriptomic response. Our analysis in Figure 1C supports this distinction: these predefined features explain limited transcriptomic response similarity in the small-molecule setting. MechMoE therefore takes a complementary approach to knowledge-guided genetic perturbation models by learning task-aligned latent mechanisms directly from molecular perturbation data rather than imposing predefined biological concepts.

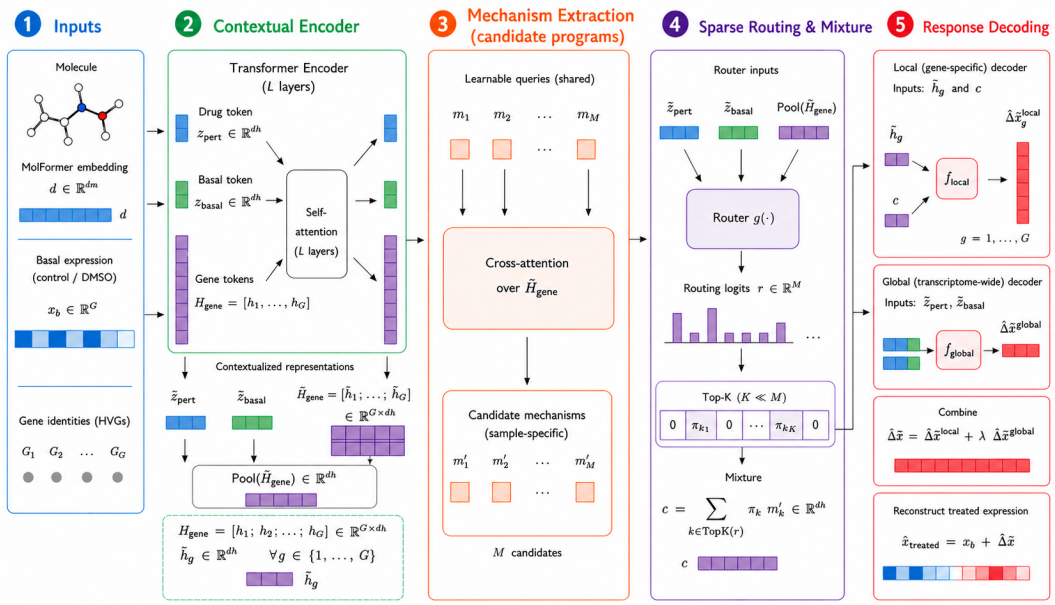

Figure 2: Architecture overview of MechMoE. Molecular perturbations, basal expression profiles, and expression-conditioned gene tokens are jointly encoded by a Transformer. Learnable mechanism queries attend over contextualized gene tokens to extract candidate mechanism representations. A sparse router selects a perturbation-specific subset of mechanism slots from the drug-cell context, producing sparse weights π. The selected candidate mechanisms are combined into a mechanism mixture c, which is used by local and global perturbation decoders to predict the perturbation-induced expression change $\widehat{\Delta x}$.

3 Method

3.1 Problem Formulation and Overview

Given a baseline cellular expression profile $x_b \in \mathbb{R}^G$, where G denotes the number of genes, together with a molecular perturbation represented by a drug embedding d, our goal is to predict the transcriptomic response induced by the perturbation.

Rather than reconstructing the treated expression profile directly, we formulate perturbation prediction as estimation of the perturbation effect $\Delta x = x_t - x_b$ where x_t denotes the treated expression profile. The model therefore learns $f(x_b, d) \rightarrow \widehat{\Delta x}$ from which the treated expression is recovered as $\hat{x}_t = x_b + \widehat{\Delta x}$.

The central hypothesis of MechMoE is that perturbational responses are better explained by a sparse combination of reusable latent mechanisms than by predefined biological concepts such as pathways, protein targets, or mechanism-of-action annotations.

Figure 2 provides an overview of the proposed framework. MechMoE first constructs contextualized representations of genes conditioned on both the molecular perturbation and the cellular state. These representations are then summarized into a collection of learnable latent mechanisms through cross-attention. Finally, a sparse routing network selects the subset of mechanisms most relevant to the current perturbation and combines them to predict the perturbational response.

3.2 Learning Latent Perturbational Mechanisms

Perturbation responses arise through context-dependent interactions between molecular perturbations, baseline cellular state, and gene-level transcriptional programs. MechMoE first constructs contextualized gene representations and then uses them to extract candidate latent mechanisms for each perturbation.

We represent each gene as an individual token. For gene g, the token combines a learned gene identity embedding with its baseline expression value,

$$h_g = E_g + E_x(x_{b,g}), \tag{1}$$

where E_g denotes a learned embedding for gene g, and E_x projects the baseline expression value into the latent space.

In addition to gene tokens, we introduce two summary tokens encoding the molecular perturbation and the global cellular state, $z_{\text{pert}} = f_d(d)$, $z_{\text{basal}} = f_b(x_b)$, where f_d and f_b denote neural encoders for the molecular representation and baseline expression profile, respectively.

Although each gene token contains its own baseline expression value, the basal token serves a different role: it summarizes transcriptome-wide cellular context, including lineage, cell state, and coordinated expression programs distributed across many genes. This global cell-state representation is used by both the mechanism router and the global perturbation decoder, allowing perturbation responses to be conditioned on the overall cellular state rather than on isolated gene-level measurements alone.

The complete input sequence is

$$H = z_{\text{pert}}, z_{\text{basal}}, h_1, \ldots, h_G. \tag{2}$$

A Transformer encoder processes these tokens jointly, $\tilde{H} = \text{Transformer}(H)$, allowing information to propagate between genes while conditioning on both perturbation identity and global cellular context.

Rather than associating perturbations with predefined pathways, MechMoE introduces a collection of learnable latent mechanism queries,

$$\mathcal{M} = m_1, \ldots, m_M. \tag{3}$$

These queries are used to probe the contextualized gene representations for recurring transcriptional response programs. For each mechanism query, cross-attention extracts a candidate mechanism representation,

$$m'_k = \text{CrossAttn}(m_k, \tilde{H}), \tag{4}$$

where m'_k denotes the candidate mechanism extracted from the current perturbation-cell context. Thus, the raw vectors m_k act as learnable queries, while the contextualized outputs m'_k represent sample-specific latent mechanisms.

Biological perturbations typically activate only a small subset of the transcriptional programs available within a cell. We therefore perform perturbation-specific mechanism selection using a sparse routing network. The router receives the perturbation representation, the basal cell-state representation, and pooled contextualized gene features,

$$r = f_{\text{route}}\left(z_{\text{pert}}, z_{\text{basal}}, \text{Pool}(\tilde{H})\right), \tag{5}$$

and produces routing logits over the candidate mechanisms. We apply sparse Top-K routing, $\pi = \text{TopKSoftmax}(r)$, with $\sum_{k=1}^{M} \pi_k = 1$.

The perturbation-specific mechanism mixture is then computed as

$$c = \sum_{k=1}^{M} \pi_k m'_k. \tag{6}$$

This formulation separates mechanism extraction from mechanism selection: cross-attention extracts candidate transcriptional programs from contextualized genes, while sparse routing selects the subset most relevant to the current drug-cell pair. As a result, different perturbations can recruit different combinations of latent mechanisms, promoting specialization while allowing mechanisms to be reused across compounds and cellular contexts.

3.3 PREDICTING PERTURBATIONAL RESPONSES

The learned mechanism representation is used to predict perturbation-induced transcriptional changes at both local and global scales.

For each gene, the contextualized representation is concatenated with the routed mechanism representation, $h_g^{\text{dec}} = \text{concat}(\tilde{h}_g, c)$, and passed through a local decoder to estimate gene-specific perturbation effects, $\widehat{\Delta x}_g^{\text{local}} = f_{\text{local}}(h_g^{\text{dec}})$.

In parallel, a second decoder predicts a global perturbation vector directly from the latent perturbation and cellular representations,

$$\widehat{\Delta x}^{\text{global}} = f_{\text{global}}(z_{\text{pert}}, z_{\text{basal}}). \tag{7}$$

The final perturbation prediction combines both components,

$$\widehat{\Delta x} = \widehat{\Delta x}^{\text{local}} + \lambda \widehat{\Delta x}^{\text{global}}, \tag{8}$$

where λ balances local gene-specific effects with global transcriptomic shifts.

This hybrid formulation captures both broad perturbational programs and fine-grained gene-level responses, improving generalization across unseen molecules and cellular contexts.

3.4 TRAINING OBJECTIVE

MechMoE is trained end-to-end using a combination of prediction objectives and regularization terms,

$$\mathcal{L} = \lambda_r \mathcal{L}_{\text{retr}} + \lambda_d \mathcal{L}_{\Delta} + \lambda_g \mathcal{L}_{\text{global}} + \lambda_v \mathcal{L}_{\text{div}}. \tag{9}$$

The primary objectives supervise perturbation prediction. An InfoNCE retrieval loss encourages perturbations with similar transcriptomic responses to occupy nearby locations in latent space. Regression losses supervise both the predicted perturbation effect and the global perturbation decoder.

The remaining terms regularize the mechanism space. A diversity loss encourages different mechanisms to attend to distinct subsets of genes and specialize to different transcriptional programs.

Together, these objectives enable MechMoE to discover sparse, reusable perturbational mechanisms that are optimized directly for transcriptomic response prediction rather than predefined biological annotations.

4 EXPERIMENTS

We evaluate MechMoE on the Tahoe-100M perturbation dataset Zhang et al. (2025), a large-scale collection of transcriptomic responses to chemical perturbations across diverse cancer cell lines. Following prior work Adduri et al. (2025), perturbation responses are represented using the top 2,000 highly variable genes (HVGs), and all models operate within the same transcriptomic feature space.

For each perturbation, we define the perturbational response as the difference between treated and control expression profiles, $\Delta = x_{\text{treated}} - x_{\text{control}}$, where x_{control} denotes the corresponding DMSO control profile. Consistent with the formulation introduced in Section 3, MechMoE predicts perturbation-induced expression changes rather than treated expression directly.

All molecular perturbations are represented using pretrained MolFormer embeddings Ross et al. (2022). During training, perturbation examples are constructed as pseudobulk profiles by averaging cells sharing the same perturbation and cellular context. This reduces single-cell measurement noise while preserving biologically meaningful perturbational signals Crowell et al. (2020); Squair et al. (2021).

Table 1: Novel cell line generalization results. Mean denotes prediction using the mean response of the training set.

Model	Δ Pearson	Δ Cosine	Δ MSE ↓	Δ MAE ↓	Direction Acc	PDS	AUPRC	DE@100
MEAN	0.192	0.191	**0.001**	**0.009**	0.950	0.001	0.309	0.350
scDCA	0.381	0.389	0.001	0.011	0.949	0.845	0.411	0.452
STATE	0.327	0.328	0.007	0.028	0.889	0.841	**0.468**	**0.506**
SAMS-VAE	0.275	0.276	0.004	0.021	0.522	0.790	0.376	0.452
BioLORD	0.280	0.281	0.005	0.026	0.573	0.838	0.358	0.437
MechMoE	**0.397**	**0.396**	**0.001**	**0.009**	**0.954**	**0.857**	0.438	0.476

Table 2: Novel molecule generalization results. Mean denotes prediction using the mean response of the training set.

Model	Δ Pearson	Δ Cosine	Δ MSE ↓	Δ MAE ↓	Direction Acc	PDS	AUPRC	DE@100
Mean	0.215	0.215	0.013	**0.001**	0.927	0.000	0.368	0.317
scDCA	0.545	0.541	0.001	0.012	0.936	0.835	**0.643**	0.558
SAMS-VAE	0.521	0.520	0.001	0.013	0.930	0.850	0.552	0.504
BioLORD	0.570	0.568	0.001	0.011	0.936	0.858	0.606	0.538
MechMoE	**0.580**	**0.576**	**0.000**	0.008	**0.953**	**0.901**	0.564	**0.559**

4.1 Out-of-Distribution Generalization Tasks

A primary goal of perturbation modeling is to generalize beyond experimentally observed perturbations and cellular contexts. We therefore evaluate MechMoE under two challenging out-of-distribution (OOD) settings illustrated in Figure 1B.

Fully unseen cell-line generalization. In this setting, models are evaluated on five held-out cell lines that are entirely excluded from training. Unlike prior evaluations that expose a subset of perturbations from held-out contexts during training, this benchmark evaluates fully out-of-context generalization.

Unseen molecule generalization. We additionally evaluate generalization to unseen perturbations by splitting molecules according to molecular scaffold. No non-control molecule appears in both training and testing partitions. This benchmark assesses whether models can extrapolate perturbational responses to previously unseen compounds based solely on molecular representation and learned perturbational structure (scaffold split).

Together, these tasks evaluate two complementary dimensions of generalization: transfer across cellular contexts and transfer across chemical space.

4.2 Baselines

We compare MechMoE against representative perturbation prediction models spanning foundation-model, latent-variable, and disentangled representation-learning approaches. scDCA fine-tunes single cell foundation models to adapt them for molecular perturbation tasks. STATE is a set-based single-cell foundation model for perturbation response prediction across cellular contexts. SAMS-VAE models perturbational effects through sparse additive latent mechanism shifts. BioLORD learns disentangled representations of biological attributes and perturbational variation. All baselines are evaluated in the same 2,000-dimensional HVG feature space and under the same train/validation/test splits as MechMoE.

4.3 Main Results

Generalization to unseen cell lines. As shown in Table 1, MechMoE achieves the strongest performance across most transcriptomic reconstruction metrics. In particular, MechMoE substantially improves Delta Pearson correlation and Delta cosine similarity while reducing reconstruction error. These gains indicate improved recovery of perturbation-induced transcriptional changes in cellular contexts that were never observed during training.

Table 3: Ablation study on the held-out test set (novel molecule split).

Model	PDS	Pearson	Dir. Acc.	MAE ↓	DE top-50
MechMoE (full)	**0.901**	**0.580**	**0.953**	**0.008**	**0.505**
w/o MoE	0.851	0.558	0.938	0.014	0.488
Dense routing	0.838	0.564	0.938	0.013	0.487
w/o global decoder	0.786	0.508	0.934	0.015	0.453
w shallow MLP	0.867	0.564	0.932	0.012	0.460

The largest improvements are observed on response-level and perturbation-discrimination metrics, including direction accuracy and PDS. MechMoE achieves substantially higher differential-expression direction accuracy and perturbation discrimination performance than competing approaches. These results suggest that the model learns perturbational effects rather than merely reconstructing baseline cellular identity.

Generalization to unseen molecules. Table 2 evaluates the more challenging setting of predicting responses for previously unseen compounds. MechMoE maintains strong performance across transcriptomic and biological metrics, achieving the highest Delta Pearson correlation, lowest reconstruction error, and strongest perturbation discrimination score. STATE cannot be evaluated in this setting.

These results demonstrate that the learned latent mechanisms transfer beyond molecules observed during training and capture perturbational structure that generalizes across chemical space.

We observe that no method dominates uniformly across all evaluation criteria. In particular, STATE achieves higher AUPRC and DE@100 in the held-out cell-line setting, and BioLORD achieves higher AUPRC in the held-out molecule setting. These metrics emphasize recovery of the highest-ranked differentially expressed genes, whereas MechMoE is optimized for accurate full-response prediction and mechanism-level generalization.

4.4 Ablation Study

We performed an ablation study to evaluate the contribution of the key architectural components of MechMoE. Specifically, we considered four variants: (1) **w/o MoE**: removes the latent mechanism module entirely and predicts perturbation responses directly from the contextualized transformer representations. (2) **Dense routing**: replaces sparse top-K routing with a dense softmax over all mechanisms, allowing every perturbation to utilize all experts simultaneously. (3) **w/o global decoder**: removes the global perturbation branch and relies exclusively on the local mechanism-conditioned decoder. (4) **shallow MLP**: replaces the Transformer layers with feed-forward MLPs. Table 3 reports results on the held-out novel-molecule benchmark.

Mechanism routing improves perturbation prediction. Removing the mechanism module leads to consistent performance degradation across all metrics. Perturbation discrimination score (PDS) decreases from 0.901 to 0.851, while differential-expression recovery also declines. These results suggest that explicitly modeling latent perturbational mechanisms provides useful inductive structure beyond standard transformer representations.

Sparse routing is more effective than dense routing. Replacing sparse routing with dense routing reduces performance across nearly all evaluation metrics. While dense routing allows every perturbation to access all mechanisms, it weakens specialization and encourages mechanisms to become redundant. In contrast, sparse routing encourages different perturbations to activate distinct subsets of latent programs, resulting in improved perturbation discrimination and biological specificity.

Global perturbation structure is important. The largest performance drop is observed when removing the global perturbation decoder. Without this component, PDS decreases from 0.901 to 0.786 and Pearson correlation falls from 0.580 to 0.508. This finding supports our hypothesis that perturbational responses contain both local gene-level effects and broader transcriptome-wide

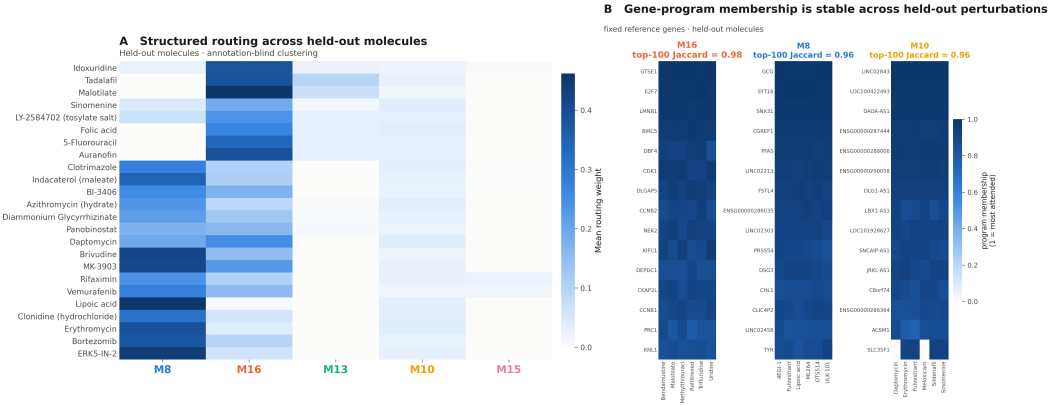

Figure 3: **MechMoE learns structured and stable perturbation programs. (A)** Routing profiles of held-out molecules across learned mechanisms. **(B)** Stability of mechanism-associated gene programs across held-out molecules, measured by top-100 gene-set Jaccard similarity.

shifts. Modeling only local effects is insufficient for accurately recovering the global structure of perturbational responses. Overall, the ablation study demonstrates that all components contribute to performance, with the global perturbation branch and sparse mechanism routing providing the largest gains.

4.5 Learned mechanisms exhibit structured routing and stable gene programs

We next examined whether the latent mechanisms learned by MechMoE capture reproducible structure beyond predictive performance. Rather than assigning mechanisms to predefined pathways, we analyzed the learned routing patterns and gene-to-mechanism attention directly on held-out molecules.

Figure 3A shows routing profiles for held-out molecules using the corrected model input vocabulary. For visualization, we display five high-variance, non-redundant mechanisms selected using a predefined routing-based criterion. The resulting heatmap reveals distinct routing patterns across held-out molecules. Some perturbations preferentially route through M8, whereas others assign greater weight to M16 or combinations involving M13, M10, and M15. Importantly, these patterns arise for molecules that were not observed during training, indicating that the router reuses a shared mechanism bank.

We then asked whether a learned mechanism retains a recognizable gene program when recruited by different held-out molecules. For each mechanism, we constructed a fixed reference gene-attention profile from the in-distribution data and evaluated its gene-program membership across held-out perturbations. The displayed genes in Figure 3B were selected once from the reference profile and were not reselected for individual held-out molecules. Thus, stability in the heatmaps reflects preservation of the same mechanism-associated gene set across molecular contexts rather than post hoc gene selection.

Across the three representative mechanisms shown, the top-attended gene sets were highly conserved across held-out molecules. The top-100 gene-set Jaccard similarities were 0.98 for M16, 0.96 for M8, and 0.96 for M10. For example, the M16 program consistently retained genes including *E2F7*, *CCNB2*, *NEK2*, *CCNB1*, and *PRC1* across distinct held-out perturbations. This stability suggests that, although routing weights vary between molecules, the gene program associated with a given latent mechanism can remain reproducible under out-of-distribution molecular perturbations.

Together, these analyses support a conservative interpretation of the learned mechanism layer: MechMoE discovers a shared set of latent perturbation programs that are recruited in different combinations across molecules, while the gene membership of individual programs remains stable across held-out perturbations.

AI USE STATEMENT

Generative AI tools were used to assist with experimental design, software implementation and debugging, interpretation and presentation of experimental results, and editing of the manuscript. They were also used to assist with figure design, literature search, and improving the clarity and organization of the paper. All experimental results were obtained from our computational experiments, and all AI-assisted code, analyses were reviewed and verified by the authors. The authors take full responsibility for the final content of this work.

REFERENCES

Abhinav K. Adduri, Dhruv Gautam, Beatrice Bevilacqua, Alishba Imran, Rohan Shah, Mohsen Naghipourfar, Noam Teyssier, Rajesh Ilango, Sanjay Nagaraj, Mingze Dong, Chiara Ricci-Tam, Christopher Carpenter, Vishvak Subramanyam, Aidan Winters, Sravya Tirukkovular, Jeremy Sullivan, Brian S. Plosky, Basak Eraslan, Nicholas D. Youngblut, Jure Leskovec, Luke A. Gilbert, Silvana Konermann, Patrick D. Hsu, Alexander Dobin, Dave P. Burke, Hani Goodarzi, and Yusuf H. Roohani. Predicting cellular responses to perturbation across diverse contexts with state. *bioRxiv*, 2025. doi: 10.1101/2025.06.26.661135. URL https://www.biorxiv.org/content/early/2025/06/27/2025.06.26.661135.

Michael Bereket and Theofanis Karaletsos. Modelling cellular perturbations with the sparse additive mechanism shift variational autoencoder, 2024. URL https://arxiv.org/abs/2311.02794.

Helena L. Crowell, Charlotte Soneson, Pierre-Luc Germain, Daniela Calini, Ludovic Collin, Catarina Raposo, Dheeraj Malhotra, and Mark D. Robinson. On the discovery of population-specific state transitions from multi-sample multi-condition single-cell rna sequencing data. *Nature Communications*, 11(1):6077, 2020. doi: 10.1038/s41467-020-19894-4.

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Feng Luo, Bo Duan, Nan Wang, Bo Liu, Jianhua Guo, Xuegong Wang, Xiang Zhang, Peng Chen, Meng Liu, Xuan Wang, Sheng Li, Jun Wang, Kai Wang, Xiaoyang Liu, Hao Chen, Yue Jiang, Jian Ma, Aidong Zhang, Qian Liu, Xiaowei Zhang, Dana Pe'er, Eric P. Xing, Xiang Chen, Fabian J. Theis, and Jue Wang. scGPT: Toward building a foundation model for single-cell multi-omics using generative ai. *Nature Methods*, 21(8):1470–1480, August 2024. doi: 10.1038/s41592-024-02201-0. URL https://doi.org/10.1038/s41592-024-02201-0.

Mingze Dong, Abhinav Adduri, Dhruv Gautam, Christopher Carpenter, Rohan Shah, Chiara Ricci-Tam, Yuval Kluger, Dave P. Burke, and Yusuf H. Roohani. Stack: In-context learning of single-cell biology. *bioRxiv*, 2026. doi: 10.64898/2026.01.09.698608. URL https://www.biorxiv.org/content/early/2026/01/09/2026.01.09.698608.

Minsheng Hao, Jiawei Gong, Xuegong Zeng, et al. Large-scale foundation model on single-cell transcriptomics. *Nature Methods*, 21(8):1481–1491, August 2024. doi: 10.1038/s41592-024-02305-7. URL https://doi.org/10.1038/s41592-024-02305-7.

Leon Hetzel, Simon Böhm, Niki Kilbertus, Stephan Günnemann, Mohammad Lotfollahi, and Fabian Theis. Predicting cellular responses to novel drug perturbations at a single-cell resolution, 2022. URL https://arxiv.org/abs/2204.13545.

Sepideh Maleki, Jan-Christian Huetter, Kangway V. Chuang, David Richmond, Gabriele Scalia, and Tommaso Biancalani. Efficient fine-tuning of single-cell foundation models enables zero-shot molecular perturbation prediction, 2025. URL https://arxiv.org/abs/2412.13478.

Zoe Piran, Niv Cohen, Yedid Hoshen, et al. Disentanglement of single-cell data with biolord. *Nature Biotechnology*, 42:1678–1683, 2024. doi: 10.1038/s41587-023-02079-x. URL https://doi.org/10.1038/s41587-023-02079-x.

Yusuf Roohani, Kexin Huang, and Jure Leskovec. Predicting transcriptional outcomes of novel multigene perturbations with GEARS. *Nature Biotechnology*, 42(6):927–935, June 2024. doi: 10.1038/s41587-023-01905-6. URL https://doi.org/10.1038/s41587-023-01905-6.

Yanay Rosen, Yusuf Roohani, Ayush Agarwal, Leon Samotorčan, Tabula Sapiens Consortium, Stephen R. Quake, and Jure Leskovec. Universal cell embeddings: A foundation model for cell biology. *bioRxiv*, 2023. doi: 10.1101/2023.11.28.568918. URL https://www.biorxiv.org/content/early/2023/11/29/2023.11.28.568918.

Jerret Ross, Brian Belgodere, Vijil Chenthamarakshan, Inkit Padhi, Youssef Mroueh, and Payel Das. Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence*, 4 (12):1256–1264, 2022. doi: 10.1038/s42256-022-00580-7.

Jordan W. Squair, Matthieu Gautier, Claudia Kathe, Mark A. Anderson, Nicholas D. James, Thomas H. Hutson, Romain Hudelle, Taha Qaiser, Kevin J. E. Matson, Quentin Barraud, Ariel J. Levine, Gioele La Manno, and Grégoire Courtine. Confronting false discoveries in single-cell differential expression. *Nature Communications*, 12(1):5692, 2021. doi: 10.1038/s41467-021-25960-2.

Jesse Zhang, Airol A. Ubas, Richard de Borja, Valentine Svensson, Nicole Thomas, Neha Thakar, Ian Lai, Aidan Winters, Umair Khan, Matthew G. Jones, Vuong Tran, Joseph Pangallo, Efthymia Papalexi, Ajay Sapre, Hoai Nguyen, Oliver Sanderson, Maria Nigos, Olivia Kaplan, Sarah Schroeder, Bryan Hariadi, Simone Marrujo, Crina Curca, Alec Salvino, Guillermo Gallareta Olivares, Ryan Koehler, Gary Geiss, Alexander Rosenberg, Charles Roco, Daniele Merico, Nima Alidoust, Hani Goodarzi, and Johnny Yu. Tahoe-100m: A giga-scale single-cell perturbation atlas for context-dependent gene function and cellular modeling. *bioRxiv*, 2025. doi: 10.1101/2025.02.20.639398. URL `https://doi.org/10.1101/2025.02.20.639398`. Preprint.

A Appendix

B Datasets and Evaluation Setting

We evaluated our model on the Tahoe-100M perturbation dataset using the same held-out cell line generalization paradigm introduced by the State framework. Following prior work, perturbation responses were represented using the top 2,000 highly variable genes (HVGs). We used the preprocessed `X_hvg` representation provided by the Tahoe preprocessing pipeline to ensure feature-space compatibility with the State baseline.

B.1 Evaluation Metrics

We evaluate perturbation prediction using transcriptomic reconstruction, perturbation retrieval, and differential expression metrics.

Transcriptomic reconstruction. We compare predicted and observed perturbation effects using Δ Pearson correlation, Δ cosine similarity, mean squared error, and mean absolute error.

Directionality. We categorize genes as upregulated, downregulated, or unchanged according to a fixed effect-size threshold and report direction accuracy.

Differential-expression recovery. We evaluate whether models recover high-effect responsive genes using DE overlap at top-k, AUPRC, and AUROC.

Perturbation discrimination. Following prior perturbation benchmarks, we report perturbation discrimination score (PDS), which measures whether each predicted response is closer to its matched observed perturbation than to alternative perturbations. Our experiments focused on predicting perturbation-induced transcriptomic changes in held-out cellular contexts.

We trained on 40 cell lines and evaluated on 5 unseen cell lines:

- `CVCL_0320`
- `CVCL_0069`
- `CVCL_0397`
- `CVCL_0459`
- `CVCL_0332`

Importantly, unlike the original State paper, which uses an "underrepresented context" split where 30% of perturbations from the held-out context are included during training Adduri et al. (2025), our evaluation used a stricter setting in which held-out cell lines were entirely excluded from training. Consequently, our benchmark evaluates fully out-of-context generalization.

B.2 Baselines

We compared our method against a reproduced implementation of State Adduri et al. (2025). Due to incompatibilities between the released implementation and current PyTorch versions, several engineering modifications were required for inference and training reproducibility. We trained and evaluated State using the publicly released ST-HVG-Tahoe architecture and checkpoints.

For fair comparison, both methods were evaluated in the same 2,000-dimensional HVG space. We evaluated STATE predictions using the predicted embeddings stored in `X_state_pred` or `X_hvg`, depending on the inference configuration.

One important caveat is that the released STATE perturbation mapping collapsed perturbation dosage information during inference, effectively treating perturbations at different concentrations as identical perturbation labels. In contrast, our model explicitly incorporates dosage information during training and prediction.

C REPRODUCIBILITY DETAILS

Software environment. All experiments were run using Python 3.10.20 with PyTorch 2.7.1+cu118. The main scientific Python dependencies were NumPy 2.2.6, pandas 2.3.3, SciPy 1.15.3, scikit-learn 1.7.2, AnnData 0.11.4, and Scanpy 1.11.5. Experiments were executed from a fixed Conda environment.

Hardware. All model comparisons were run on NVIDIA A100 GPUs with 80 GB of memory.

Dataset and preprocessing. Experiments used the Tahoe perturbation dataset stored as an AnnData object containing 1,884,935 observations and 10,000 genes in CSR sparse format. The variable index corresponds to Ensembl gene identifiers. Molecular perturbations were represented using pretrained MolFormer embeddings, with each SMILES string mapped to a 768-dimensional vector.

We used the existing dataset split stored in `obs['split']`, comprising 243 training molecules, 68 validation molecules, and 65 test molecules. Each split contains measurements across 50 cell lines. Gene expression values were normalized using a `log1p` transformation relative to a median library size of 1760; this preprocessing is not equivalent to CP10K normalization.

For perturbation prediction, cells were aggregated into one pseudobulk profile per `(cell_type, condition)` pair by averaging across all corresponding cells. Perturbation effects were defined as

$$\Delta x = x_{\text{treated}} - x_{\text{DMSO}},$$

where the matched DMSO profile serves as the baseline cellular state.

Novel-molecule split and feature space. For the novel-molecule benchmark, the existing molecule split was used together with the precomputed highly variable gene representation. Models operated on 2,000 highly variable genes, with HVGs defined using the Seurat v3 procedure. The model additionally used a PCA dimensionality of 128 and MolFormer molecular embeddings of dimension 768.

Training configuration. The reference MechMoE configuration contained 7,844,341 trainable parameters and was trained for 30 epochs using a batch size of 64, learning rate 2×10^{-4}, and random seed 0. The model used 32 latent mechanisms with top-k sparse routing, where $k = 4$. The global perturbation contribution was scaled by 0.50. Differential-expression evaluation used a threshold of 0.05. The global L_1 loss weight was 0.20 and the attention-diversity regularization weight was 0.01.