HITTING THE RIGHT TARGETS: SIMPLE PSEUDOBULK PREDICTORS FOR CHEMICAL PERTURBATION PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Predicting transcriptional responses to chemical perturbations can accelerate drug discovery. Single-cell perturbation models are increasingly complex, yet they are typically evaluated on aggregated pseudobulk profiles, where simple baselines remain competitive. We revisit chemical perturbation prediction at pseudobulk resolution with a lightweight MLP on frozen compound and cell-line embeddings, and benchmark twelve compound representations spanning molecular structure, morphology-aligned phenotype, and drug targets. On sci-Plex3 and Tahoe-100M, the predictor outperforms learning-free baselines and four single-cell models (ChemCPA, BioLord, Doloris, and PerturbDiff) on most metrics while using up to 55$\times$ fewer parameters. To understand why, we examine the two ingredients of the predictor: its prediction target and its compound representation. Aggregation yields a less noisy target, and retraining a representative single-cell model (ChemCPA) on pseudobulk profiles narrows, but does not close its gap to our model. On the representation side, more sophisticated molecular embeddings bring little consistent gain; instead, performance tracks how closely the geometry of a representation aligns with that of the transcriptional responses. These results suggest that predictive performance depends not only on model complexity but also on the data, and position pseudobulk predictors as strong baselines and potential priors for single-cell models.

1 INTRODUCTION

Chemical perturbation profiling is a cornerstone of modern drug discovery (Subramanian et al., 2017; Yofe et al., 2020), but generating perturbation-response data at scale is costly, especially across the combinatorial space of compounds, doses and cellular contexts (Srivatsan et al., 2020). Accurate *in silico* perturbation prediction is therefore critical to accelerate target discovery, prioritize experiments and reduce screening cost (Lotfollahi et al., 2019; Roohani et al., 2023).

Recent approaches increasingly formulate perturbation prediction at single-cell resolution, aiming to capture both the mean response to a perturbation (pseudobulk) and the heterogeneity of responses across cells. Yet model performance is often evaluated after aggregating predicted cells into condition-level pseudobulk profiles. At this resolution, simple learning-free predictors, such as the control mean or mean response across training perturbations, are natural baselines by capturing the basal state and shared transcriptional responses. Recent benchmarks show that such baselines can match or outperform more complex single-cell models on pseudobulk-level metrics (Ahlmann-Eltze et al., 2025; Kernfeld et al., 2025; Csendes et al., 2025; Wong et al., 2025; Bendidi et al., 2024; Liu et al., 2024), while model rankings vary considerably across evaluation criteria (Wu et al., 2024; Wei et al., 2026; Vinas Torne et al., 2025; Mejia et al., 2026). These observations raise a more fundamental question: *what determines successful chemical perturbation prediction: model complexity, the resolution of the prediction target, or the information encoded by the perturbation representation?*

We study this question by deliberately separating pseudobulk-level mean response prediction from the additional challenge of modeling cell-to-cell heterogeneity. We develop a lightweight embedding-based MLP that predicts pseudobulk transcriptional responses from compound and cell-line representations, and evaluate it on both sci-Plex3 and Tahoe-100M. Despite its simplicity, the model

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.

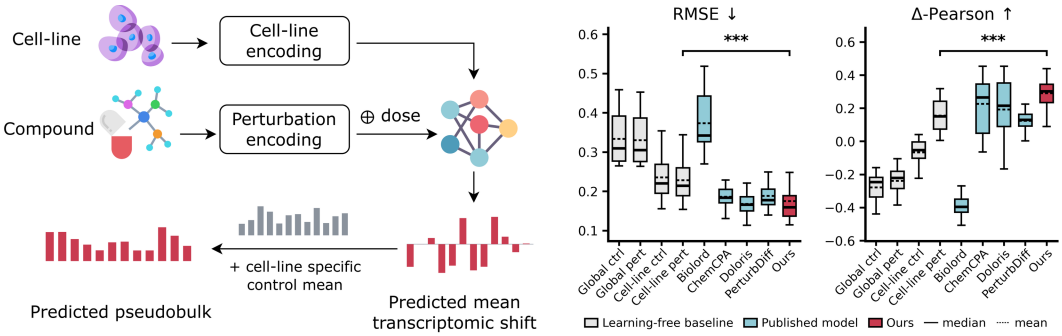


Figure 1: **Left:** A lightweight MLP-based pseudobulk perturbation predictor that takes frozen compound and cell-line embeddings as input. **Right:**On sci-Plex3 DEG50, the proposed predictor outperforms learning-free baselines and matches or exceeds representative single-cell models while requiring 2–55× fewer parameters.

outperforms learning-free baselines and four representative single-cell models on most pseudobulk metrics while using up to 55x fewer parameters. We then investigate why pseudobulk prediction is effective and what is lost through aggregation. Increasing the number of aggregated cells produces progressively more reproducible prediction targets, while training single-cell models such as ChemCPA directly at pseudobulk resolution improves its performance but does not close the gap to direct condition-level prediction. Although the pseudobulk predictor does not model cellular heterogeneity, shifting the observed control cells by the predicted mean response outperforms single-cell models on distribution metrics EMD and Wasserstein distance; notably, so does the unshifted control. Together, these results suggest that model complexity alone does not determine predictive performance.

Beyond model resolution, prediction for unseen drugs depends critically on how they are related to those observed during training. We therefore systematically compare molecular structure descriptors, molecular foundation-model embeddings, morphology-aligned representations, and target-aware representations under the same lightweight architecture, and ask not only *which* representations perform well, but *why*. More sophisticated representations do not consistently improve prediction; instead, on sci-Plex3, predictive performance is associated with how well representation-space similarity aligns with transcriptional-response similarity, while on Tahoe-100M differences between representations are small. Target-based representations benefit most when test drugs share targets with training compounds, and the predictor exploits them beyond simple averaging of target-sharing drugs. Together, these results suggest that representation complexity alone does not determine performance on unseen drugs; what matters is whether the representation captures perturbation relationships that are predictive of transcriptional response. We summarize our contributions below:

- We establish a strong baseline for chemical perturbation prediction: a lightweight pseudobulk MLP outperforms learning-free baselines and representative single-cell models on most metrics across sci-Plex3 and Tahoe-100M, with up to 55x fewer parameters.
- We show that pseudobulk aggregation yields a substantially less noisy prediction target, and that training the same single-cell model at pseudobulk resolution improves it on most metrics but still falls short of our lightweight predictor.
- We characterize what makes perturbation representations predictive, showing that downstream performance tracks the alignment between perturbation-representation and transcriptional-response geometry.

2 Related Work

Single-cell perturbation prediction has been approached through two broad modeling strategies (Lotfollahi et al., 2019; 2021; Wu et al., 2024; Wang et al., 2026). Latent generative approaches learn representations of cellular state and perturbation effects from observed expression profiles to generate responses under new conditions (Lotfollahi et al., 2021; Hetzel et al., 2022; Wang et al., 2024; Chi et al., 2026; Piran et al., 2024), whereas direct transition approaches learn mappings from control populations and perturbation information to perturbed states (Lotfollahi et al., 2019; Adduri et al.,

2026; Wenkel et al., 2026; Wang et al., 2026). Across both strategies, increasingly expressive architectures, including variational autoencoders, diffusion models, GNNs, and foundation models, have been explored to improve generalization across perturbations and cellular contexts.

Recent benchmarking studies, however, have questioned whether this increasing model complexity consistently improves perturbation prediction. Simple mean predictors and linear models (Ahlmann-Eltze et al., 2025; Vinas Torne et al., 2025), and lightweight learned predictors (Liu et al., 2024), can be competitive with substantially more complex models on commonly used perturbation-prediction benchmarks (Ahlmann-Eltze et al., 2025; Kernfeld et al., 2025; Csendes et al., 2025; Wong et al., 2025; Bendidi et al., 2024; Liu et al., 2024). These results suggest that strong pseudobulk-level predictive performance may be achievable with relatively simple models. However, much of this evidence has focused on genetic perturbations, leaving less explored how far simple learned models can be pushed for pseudobulk-level chemical perturbation prediction.

Chemical perturbations introduce an additional modeling choice: how to represent the compound itself. Existing methods have explored diverse molecular representations (Hetzel et al., 2022), but largely as alternative encodings of molecular structure. Complementary views capturing different aspects of a compound's chemical properties, phenotypic activity, and biological targets remain comparatively underexplored for chemical perturbation prediction. We therefore study pseudobulk-level chemical perturbation prediction with a deliberately simple learned model while systematically examining how predictive performance depends on the representation of the chemical perturbation.

3 METHOD

For a group of cells sharing cell line $c \in \mathcal{C}$, perturbation $p \in \mathrm{ctrl} \cup \mathcal{D}$ ($\mathcal{D}$ the drug set), and dosage $s \in \mathbb{R}_{\geq 0}$, the pseudobulk profile $\mathbf{x}^{(c,p,s)} \in \mathbb{R}^n$ is the average log-normalized expression over those cells (Wang et al., 2026; Roohani et al., 2023; Cui et al., 2024); control observations have $p = \mathrm{ctrl}$ and $s = 0$, written $\mathbf{x}^{(c,\mathrm{ctrl})}$. We predict the mean transcriptomic shift $\delta^{(c,p,s)} = \mathbf{x}^{(c,p,s)} - \mathbf{x}^{(c,\mathrm{ctrl})}$ and reconstruct the perturbed profile by adding it back to the observed control (Fig. 1, left):

$$\hat{\delta}^{(c,p,s)} \;=\; f(\psi(c),\, \phi(p),\, s)\,, \qquad \hat{\mathbf{x}}^{(c,p,s)} \;=\; \mathbf{x}^{(c,\mathrm{ctrl})} + \hat{\delta}^{(c,p,s)}, \tag{1}$$

where ϕ and ψ denote pretrained drug and cell-line embedding models, respectively.

3.1 ATTRIBUTE EMBEDDING

Perturbation encoding ϕ. Chemical perturbations act through diverse molecular mechanisms, making perturbation representation a first-order design decision. We therefore benchmark complementary views of each compound, spanning chemical structure, induced phenotype, and mechanism of action.

(1) Molecular structure representations encode compounds from their chemical structure including: RDKit2D descriptors (Hetzel et al., 2022), Morgan fingerprints (Rogers & Hahn, 2010), SMILES embeddings (token mean-pooling) from chemical language models ChemBERTa (Singh et al., 2026) and MolFormer (Ross et al., 2022), and 3D-geometric embeddings from the conformation-aware Uni-Mol (Zhou et al., 2023). Together, these let us test whether richer learned or geometric representations add information for pseudobulk prediction.

(2) Phenotypic (cell-morphology) representations encode a compound by the morphological cell state it induces, providing a readout of biological effect rather than compound structure (Sanchez-Fernandez et al., 2023; Lu et al., 2025). Given the limited coverage of raw morphological profiles, we use the text branch of CellCLIP (Lu et al., 2025), which is trained contrastively to align closely with the morphological features.

(3) Target-aware (mechanism-of-action) representations. Since a compound's transcriptomic effect is mediated by the proteins it engages, representing a drug by its target features provides a more direct view of mechanism than chemical structure alone (Li et al., 2026). Therefore, we approach this from two angles: **(a)** *Target-aware chemical embedding:* we use the chemical branch of UniPert (Li et al., 2026), whose representation is trained to lie close to its corresponding targets; mechanism is thus encoded implicitly through the training objective. **(b)** *Knowledge-derived target embedding:* we resolve each drug to its target gene(s) and encode those targets with gene-level features, spanning increasing semantic richness: Target ID, a fixed vector per gene that carries identity but no gene-gene

relationship and the text-derived gene embeddings GenePT-ada, GenePT-protein (Chen & Zou, 2024) together with Qwen3-Embeddings-8B (Zhang et al., 2025b) taken from (Hedley et al., 2026). Targets are retrieved from the literature with GPT5.4 and pooled over a drug's target set uniformly or with respect to binding probability; See Appendix E for target retrieval and quality assessment.

Cell-line encoding ψ. Rather than encoding cell-line with one-hot vector, we take the LLM-derived cell-line embeddings of BAITSAO (Liu et al., 2025b;a) so that the geometry between cell-lines reflects documented biological knowledge. We reduce the cell-line embedding $\psi(c)$ with two MLPs to obtain the FiLM parameters γ_c and β_c, which modulate the drug hidden representation as $(1+\gamma_c)\odot h+\beta_c$.

3.2 Training objective

The model's input carries information about the perturbation but not about the cell line's basal transcriptional state, so we predict the drug-induced effect $\delta_i^{c,p,s}$ directly and reconstruct the full expression profile by adding it to the observed control profiles, $\mathbf{x}^{c,\text{ctrl}}+\delta_i^{c,p,s}$. We supervise this with a bulk-level MSE between the predicted pseudobulk and the ground-truth pseudobulk.

While the MSE loss penalizes the magnitude and sign errors, it does not explicitly enforce recovery of the relative transcriptional response pattern induced by a perturbation. So we add a Pearson loss between the predicted $\hat{\delta}_i^{c,p,s}$ and the true $\delta_i^{c,p,s}$ to align the shape of the responses. Since the δ-term can be dominated by the cross-perturbation signature that most drugs share within a cell line, we propose to center the delta by subtracting the frozen, cell line-specific train-only delta mean $\bar{\delta}^c$ to highlight the perturbation-unique information. Given batch size B and n genes, our loss function is as below:

$$\mathcal{L}=\underbrace{\frac{1}{Bn}\sum_{i=1}^{B}\|\hat{\mathbf{x}}_i^{c,p,s}-\mathbf{x}_i^{c,p,s}\|_2^2}_{\text{bulk-level MSE}}+\underbrace{1-\frac{1}{B}\sum_{i=1}^{B}\rho\Big(\hat{\delta}_i^{c,p,s}-\bar{\delta}^c,\delta_i^{c,p,s}-\bar{\delta}^c\Big)}_{\text{Pearson loss}} \tag{2}$$

4 Experimental setup

We evaluate on sci-Plex3 (Srivatsan et al., 2020) and Tahoe-100M (Zhang et al., 2025a). sci-Plex3 profiles three human cancer cell lines treated with 188 compounds; Tahoe-100M profiles 50 cancer cell lines exposed to 379 drugs across 14 plates. Because plate effects in Tahoe-100M are strongly confounded with perturbation effects, we analyze one plate at a time: plate 6 and the biological replicate plate 14 (Zhang et al., 2025a) (94 drugs) for the main results, and plates 3, 9, and 12 (92/93/94 drugs; fewer than three shared drugs) for robustness. sci-Plex3 uses an 80/10/10 Bemis–Murcko scaffold split, so test drugs have unseen scaffolds, repeated 7 times; Tahoe-100M uses a random drug split (80/10/10), repeated 10 times, because a single plate contains fewer drugs. For all methods, we fix the dose at its highest value (10 μM in sci-Plex3, 5 μM in Tahoe-100M) to isolate compound effects (all doses: Appendix B). We report mean $\pm$ std across repetitions and cell lines.

We report Root Mean Square Error (RMSE), coefficient of determination (R^2), and Pearson correlation following Hetzel et al. (2022); Roohani et al. (2023); Cui et al. (2024) To capture perturbation-induced change, we report Δ-Pearson and Δ-Direction (agreement in up- versus down-regulation calls) (Adduri et al., 2026). We compare against four learning-free baselines: cell-line specific and global control and perturbation mean (all calculated on the training data) following (Vinas Torne et al., 2025; Adduri et al., 2026; Ahlmann-Eltze et al., 2025) and four single-cell models ChemCPA (Hetzel et al., 2022), Biolord (Piran et al., 2024), diffusion-based Doloris (Chi et al., 2026) and state-of-the-art PerturbDiff (Yuan et al., 2026) on full gene panel and condition-specific top-50 differentially expressed genes (DEG50).

5 Result

5.1 Simple predictors are strong baselines

A lightweight pseudobulk predictor outperforms the learning-free baselines and single-cell models significantly with 17-55$\times$ fewer parameters. On sci-Plex3 (Table 1), our pseudobulk MLP attains the best mean on every metric except full-panel Δ-Direction with using 17–55$\times$

fewer parameter than PerturbDiff. On Tahoe-100M (Table 2), all learned representations improve substantially over the learning-free baselines on the full panel and on DEG50 ΔR^2 and slightly on DEG50 Δ-Pearson, whereas the cell-line perturbation mean remains best on DEG50 Δ-Direction. Parameter counts on Tahoe-100M are similar to those in Table 1 and are omitted for space.

More sophisticated molecular representations do not consistently improve prediction. Neither 3D molecular representations nor chemical language-model embeddings consistently outperform simple RDKit-based features, suggesting that greater representational complexity alone provides limited benefit. On sci-Plex3, CellCLIP's morphology-aligned embedding performs poorly on most metrics, whereas target-identity representations perform particularly well, especially on DEG50. Interestingly, GenePT-Ada, GenePT-protein, and Qwen3 yield nearly identical performance across both gene panels, despite varying either the underlying biological text or the embedding model, suggesting that prediction performance is relatively insensitive to these choices among text-derived gene representations. On Tahoe-100M, in contrast, differences between representations are small relative to their variability, and CellCLIP performs on par with the best representations.

Table 1: **sci-Plex3, full gene panel and DEG50.** Absolute performance (R^2) together with residual-based (Δ) metrics. Mean $\pm$ std across cell lines and seeds, **bold** marks the best mean per column. **M** is the number of trainable parameters (millions). Stars on a bold cell give its significance against the best learning-free baseline of that column ($*\ p < 0.05$, $**\ p < 0.01$, $***\ p < 0.001$; paired t-test)

		Full gene panel			DEG50			
Model / representation		$R^2\uparrow$	Δ-Pearson$\uparrow$	Δ-Direction$\uparrow$	$R^2\uparrow$	Δ-Pearson$\uparrow$	Δ-Direction$\uparrow$	M$\downarrow$
Learning-free baselines								
Cell-line control mean		0.824 $\pm$0.063	0.024 $\pm$0.028	0.000 $\pm$0.000	0.701 $\pm$0.099	−0.066 $\pm$0.078	0.000 $\pm$0.000	0
Cell-line perturbation mean		0.840 $\pm$0.058	0.336 $\pm$0.049	0.561 $\pm$0.019	0.714 $\pm$0.092	0.156 $\pm$0.104	0.596 $\pm$0.069	0
Global control mean		0.469 $\pm$0.027	0.133 $\pm$0.066	0.477 $\pm$0.026	0.200 $\pm$0.165	−0.278 $\pm$0.078	0.290 $\pm$0.086	0
Global perturbation mean		0.480 $\pm$0.025	0.154 $\pm$0.065	0.483 $\pm$0.029	0.204 $\pm$0.167	−0.238 $\pm$0.075	0.323 $\pm$0.070	0
Single-cell models (trained on single cells)								
BioLord (Piran et al., 2024)		0.447 $\pm$0.011	0.181 $\pm$0.042	0.510 $\pm$0.010	−0.045 $\pm$0.328	−0.395 $\pm$0.059	0.216 $\pm$0.057	42.4
ChemCPA (Hetzel et al., 2022)		0.811 $\pm$0.174	0.251 $\pm$0.260	**0.593** $\pm$0.059	0.625 $\pm$0.066	0.227 $\pm$0.029	0.396 $\pm$0.022	11.3
Doloris (Chi et al., 2026)		0.828 $\pm$0.051	0.174 $\pm$0.046	0.572 $\pm$0.012	0.708 $\pm$0.081	0.192 $\pm$0.066	0.478 $\pm$0.080	58.2
PerturbDiff (Yuan et al., 2026)		0.715 $\pm$0.074	0.162 $\pm$0.038	0.519 $\pm$0.014	0.673 $\pm$0.071	0.124 $\pm$0.038	0.613 $\pm$0.022	104.5
Ours: pseudobulk MLP on a frozen perturbation embedding								
Molecular structure	RDKit-2D	0.880 $\pm$0.054	**0.360** $\pm$0.059*	0.576 $\pm$0.024	0.787 $\pm$0.084	0.206 $\pm$0.140	0.608 $\pm$0.075	1.9
	Morgan FP	0.849 $\pm$0.050	0.282 $\pm$0.115	0.560 $\pm$0.022	0.759 $\pm$0.090	0.247 $\pm$0.089	0.627 $\pm$0.047	3.3
	Uni-Mol	0.873 $\pm$0.039	0.352 $\pm$0.067	0.565 $\pm$0.035	0.773 $\pm$0.071	0.197 $\pm$0.124	0.615 $\pm$0.066	2.2
Chemical LLM	ChemBERTa	0.858 $\pm$0.054	0.333 $\pm$0.055	0.565 $\pm$0.036	0.743 $\pm$0.087	0.203 $\pm$0.104	0.614 $\pm$0.040	2.5
	MoLFormer	0.844 $\pm$0.049	0.332 $\pm$0.069	0.559 $\pm$0.031	0.717 $\pm$0.081	0.181 $\pm$0.098	0.608 $\pm$0.045	2.5
Phenotype	CellCLIP	0.820 $\pm$0.066	0.189 $\pm$0.136	0.553 $\pm$0.023	0.725 $\pm$0.086	0.164 $\pm$0.100	0.583 $\pm$0.066	2.2
Target-aware	UniPert	0.876 $\pm$0.052	0.342 $\pm$0.060	0.581 $\pm$0.031	0.771 $\pm$0.079	0.186 $\pm$0.107	0.599 $\pm$0.061	2.0
	Target ID (binding prob.)	**0.886** $\pm$0.050**	0.356 $\pm$0.058	0.584 $\pm$0.024	0.805 $\pm$0.074	**0.291** $\pm$0.095***	**0.657** $\pm$0.046***	2.2
	Target ID (avg.)	0.884 $\pm$0.054	0.351 $\pm$0.062	0.583 $\pm$0.025	**0.809** $\pm$0.073***	0.277 $\pm$0.087	0.653 $\pm$0.047	2.2
	GenePT-ada (avg.)	0.824 $\pm$0.064	0.126 $\pm$0.045	0.566 $\pm$0.014	0.701 $\pm$0.100	0.266 $\pm$0.067	0.522 $\pm$0.062	3.3
	GenePT-protein (avg.)	0.824 $\pm$0.064	0.124 $\pm$0.045	0.565 $\pm$0.013	0.701 $\pm$0.101	0.265 $\pm$0.068	0.528 $\pm$0.065	4.8
	Qwen3 (avg.)	0.824 $\pm$0.064	0.125 $\pm$0.045	0.565 $\pm$0.013	0.701 $\pm$0.101	0.267 $\pm$0.068	0.529 $\pm$0.065	5.9

5.2 Understanding the Performance and Trade-offs of Pseudobulk Prediction

Where does the pseudobulk advantage come from, and what does aggregation cost? On sci-Plex3, which contains more drugs than a single Tahoe-100M plate, we ask whether pseudobulking (i) yields a less noisy prediction target, (ii) alone explains the gap to single-cell models, and (iii) combined with observed control cells, can reconstruct single-cell response distributions.

Pseudobulk aggregation substantially improves the reproducibility of the prediction target

A prediction target is only useful to the extent that it is reproducible: if two independent sets of cells from the same condition yield different responses, part of what a model is asked to fit is sampling noise and/or technical variation. Single-cell models are trained at the noisiest end of this spectrum, whereas our predictor is trained on profiles aggregated over many cells. We therefore measure how reliably N cells capture the condition-level response. For each condition, we sample $2N$ cells, split them into two non-overlapping groups of N, aggregate each group into a pseudobulk, and measure the similarity between the two as a function of N.Fig. 2 (left) shows means over 100 repetitions.

As the number of aggregated cells increases, the MSE between the two pseudobulks decreases consistently while the Pearson correlation of the gene expression profiles rapidly approaches to one (Fig. 2, left). Pearson correlation between residual response is lower than for absolute expression, likely due to removal of the shared basal transcriptional program, but increases with higher N. With few cells, the estimated perturbation response can vary substantially even between samples from the

Table 2: **Tahoe residual performance on the Full and DEG50 panel.** Results are reported as mean ± std across cell lines and seeds. **Bold** marks the best mean per metric within each panel.

		Full gene panel			DEG50		
Model / representation		$\Delta R^2 \uparrow$	Δ-Pearson $\uparrow$	Δ-Direction $\uparrow$	$\Delta R^2 \uparrow$	Δ-Pearson $\uparrow$	Δ-Direction $\uparrow$
Learning-free baselines							
Cell-line control mean		−0.093 ±0.056	−0.002 ±0.033	0.000 ±0.000	−0.256 ±0.176	−0.011 ±0.100	0.000 ±0.000
Cell-line perturbation mean		0.223 ±0.098	0.568 ±0.059	0.681 ±0.025	0.414 ±0.154	0.791 ±0.057	**0.893** ±0.041
Global control mean		−8.999 ±3.361	0.196 ±0.118	0.558 ±0.034	−7.112 ±4.028	0.387 ±0.131	0.623 ±0.101
Global perturbation mean		−8.864 ±3.364	0.242 ±0.109	0.580 ±0.031	−7.035 ±4.076	0.447 ±0.120	0.666 ±0.093
Single-cell models (trained on single cells)							
BioLord (Piran et al., 2024)		-14.401 ±1.414	0.399 ±0.041	0.582 ±0.033	-9.633 ±6.487	0.476 ±0.103	0.774 ±0.074
ChemCPA (Hetzel et al., 2022)		−0.087 ±0.052	0.519 ±0.018	0.653 ±0.011	0.099 ±0.059	0.667 ±0.019	0.852 ±0.013
Doloris (Chi et al., 2026)		−5.865 ±5.928	0.269 ±0.084	0.525 ±0.045	−1.159 ±1.549	0.528 ±0.080	0.676 ±0.075
PerturbDiff (Yuan et al., 2026)		−0.010 ±0.178	0.591 ±0.036	0.661 ±0.030	0.432 ±0.077	0.782 ±0.021	0.860 ±0.016
Ours: pseudobulk MLP on a frozen perturbation embedding							
Molecular structure	RDKit-2D	0.435 ±0.124	0.734 ±0.072	0.727 ±0.036	0.505 ±0.143	0.809 ±0.063	0.868 ±0.050
	Morgan FP	0.457 ±0.102	0.739 ±0.069	0.732 ±0.032	0.505 ±0.136	0.811 ±0.061	0.870 ±0.047
	Uni-Mol	0.463 ±0.101	0.740 ±0.068	0.733 ±0.032	0.507 ±0.135	0.811 ±0.061	0.871 ±0.046
Chemical LLM	ChemBERTa	0.459 ±0.110	0.740 ±0.069	0.733 ±0.034	0.506 ±0.138	0.812 ±0.061	0.870 ±0.048
	MoLFormer	0.459 ±0.105	0.739 ±0.069	0.732 ±0.032	0.509 ±0.135	0.812 ±0.061	0.871 ±0.047
Phenotype	CellCLIP	**0.468** ±0.102***	**0.742** ±0.069***	0.733 ±0.033	**0.513** ±0.136***	**0.813** ±0.061***	0.873 ±0.047
Target-aware	UniPert	**0.468** ±0.104***	**0.742** ±0.070***	**0.735** ±0.032***	0.510 ±0.137	0.812 ±0.062	0.872 ±0.047
	Target ID (avg.)	0.457 ±0.107	0.739 ±0.069	0.731 ±0.032	0.509 ±0.140	**0.813** ±0.061***	0.871 ±0.047
	Target ID (binding prob.)	0.401 ±0.199	0.726 ±0.078	0.726 ±0.041	0.493 ±0.141	0.804 ±0.065	0.865 ±0.049
	GenePT-ada (avg.)	0.455 ±0.110	0.737 ±0.069	0.731 ±0.034	0.509 ±0.136	0.811 ±0.062	0.871 ±0.048
	GenePT-protein (avg.)	0.424 ±0.148	0.728 ±0.073	0.726 ±0.038	0.506 ±0.143	0.807 ±0.065	0.867 ±0.050
	Qwen3 (avg.)	0.429 ±0.129	0.731 ±0.070	0.724 ±0.037	0.500 ±0.137	0.808 ±0.063	0.869 ±0.048

same condition, providing the model with a noisy training target. As N increases, this variability is progressively averaged out, yielding a more stable estimate of the underlying perturbation response. Thus, aggregation yields a less noisy learning target.

Test pseudobulks are themselves averages over finitely many cells and therefore deviate from the true condition-level response by sampling noise. A perfect predictor, one that outputs the true response exactly, would thus still not reach, for example, a Pearson correlation of one with the measured test profile. We define the *noise ceiling* as the score such a predictor would be expected to obtain, and estimate it from the split-half analysis above: for each test condition, we split all of its cells into two random halves, aggregate each half, and score one against the other, averaging over 100 splits, all test conditions and seeds. Because each half contains only half the cells, this split-half agreement understates the reliability of the full test pseudobulk. For Pearson and Δ-Pearson, we correct for this with the Spearman–Brown formula, $r_{\text{full}} = 2r_{\text{half}}/(1 + r_{\text{half}})$, which estimates the correlation between two independent full-size pseudobulks, and take $\sqrt{r_{\text{full}}}$, the expected correlation between the true response and a single measured pseudobulk. R^2 and Δ-Direction admit no such correction, so we report their split-half values, which are conservative: the true ceiling is higher. Ceilings are shown in Fig. 2 (middle, red; full panel). Our predictor approaches the ceiling on absolute metrics but remains well below it on residual metrics, indicating room for improvement in predicting perturbation effects.

Matching the training resolution improves ChemCPA, but does not eliminate performance gap

Having established that pseudobulking yields a less noisy target, we ask whether target resolution explains the gap to single-cell models, which are trained on individual cells but evaluated after aggregation. Unlike Doloris and PerturbDiff that model distributions over cells, ChemCPA maps each input profile to a predicted profile, so pseudobulks can replace single cells without architectural changes; We retrain it on the same pseudobulk profiles with all other settings fixed. Pseudobulk training improves ChemCPA on most metrics (Fig. 2, middle), consistent with reduced sampling noise, but its mean remains below our predictor on most metrics Thus, matching the training resolution narrows but does not close the gap.

From pseudobulk responses to single-cell distributions

We next ask whether pseudobulk predictor can support reconstruction of single-cell responses despite not explicitly modeling perturbation-induced heterogeneity. Following (Chi et al., 2026), we evaluate distribution-level agreement using Earth Mover's Distance (EMD) and Wasserstein distance. For a comparable evaluation, both approaches start from observed control cells: single-cell models learn perturbation-induced cell-level changes, whereas our model predicts only the condition-level mean response. We reconstruct a single-cell population by shifting each control cell by the predicted mean

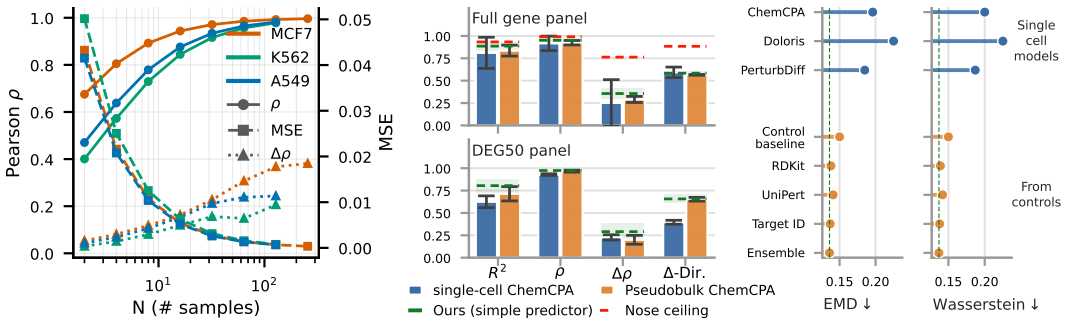


Figure 2: **Decomposing the pseudobulk advantage on sci-Plex3**. **Left**: agreement between pseudobulks from two disjoint sets of N cells rises with N; residual responses remain less reproducible than absolute expression. **Middle**: ChemCPA trained on single cells (blue) or pseudobulks (orange) versus ours (green, Target ID); red dashes mark the noise ceiling (full panel). Pseudobulk training improves ChemCPA, but it stays below our predictor on most metrics, so target denoising alone does not explain the gap. **Right**: EMD and Wasserstein distance (lower is better; DEG50) for control cells shifted by the predicted mean response versus single-cell models. Shifted controls are competitive, but so are unshifted controls, suggesting no clear benefit of modeling heterogeneity for unseen drugs.

response: $\tilde{\mathbf{x}}_{i,c} := \mathbf{x}^{\text{ctrl}}_{i,c} - \mu^{\text{ctrl},c} + \hat{\mathbf{x}}^{c,p,s}$ where $\mathbf{x}^{\text{ctrl}}_{i,c}$ is a randomly drawn control single cells from cell line c without replacement, the $\mu^{\text{ctrl},c}$ represents the population-level control mean from cell line c, and $\hat{\mathbf{x}}^{c,p,s}$ means the predicted perturbed mean-response. This construction assumes that the perturbation changes the population mean but preserves the cell-to-cell structure of the control population. It therefore cannot represent perturbation-induced changes in dispersion, covariance, subpopulation composition, or state-dependent responses (Adduri et al., 2026).

To isolate the contribution of the prediction, we compare against an unshifted control baseline $\tilde{\mathbf{x}}_{i,c} := \mathbf{x}^{\text{ctrl}}_{i,c}$. We shift control cells using mean responses predicted from RDKit, UniPert, and Target ID, and from their gene-wise average (Ensemble), draw 500 control cells, and average over three repeated runs. Despite its simplicity, the shifted-control reconstruction is competitive with single-cell models on DEG50 distributional metrics. This does not imply that perturbations preserve cell-level heterogeneity; rather, under the OOD setting, explicitly modeling perturbation-induced heterogeneity provides no clear advantage over combining the control distribution with a predicted mean shift.

5.3 What makes a perturbation representation predictive?

Predictive performance tracks alignment between perturbation-representation and transcriptional-response geometry

On sci-Plex3, target-aware representations, particularly Target ID, outperform chemical and morphology-derived ones (Table 1). To explain this ranking, we compare the drug-similarity geometry of each representation with that of the residual responses δ_d, asking whether drugs close in feature space also respond similarly (Pearson correlation).

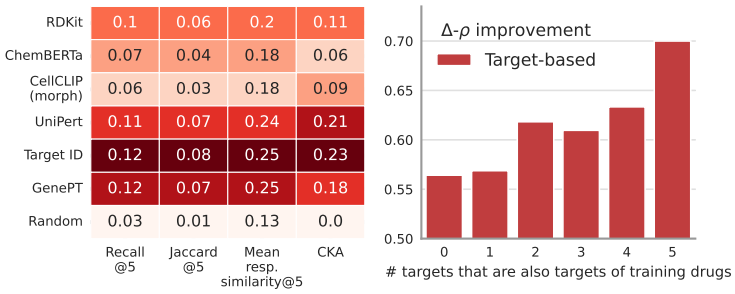
We measure drug–drug similarity by Pearson correlation ρ, computed between representations in feature space and between residual responses δ in response space. For each anchor drug d, let $\mathcal{N}^{\phi}_5(d)$ and $\mathcal{N}^{\delta}_5(d)$ denote its five most similar drugs in each space. Local correspondence is measured by the overlap between the two neighbor sets (Recall@5 and Jaccard@5) and by Mean Response Similarity@5, $1/5 \sum_{j \in \mathcal{N}^{\phi}_5(d)} \rho(\delta_d, \delta_j)$, the average response similarity of a drug's feature-space neighbors, which credits similarly responding neighbors even outside $\mathcal{N}^{\delta}_5(d)$. Global correspondence is measured by linear centered kernel alignment (Kornblith et al., 2019) between the drug-by-feature matrix and the drug-by-gene residual matrix. All metrics are computed per cell line and averaged, with local metrics first averaged over anchor drugs. As a sanity baseline, each drug is assigned a randomly sampled feature vector (Fig. 3, left).

On sci-Plex3, Target ID ranks first on all four metrics, matching GenePT locally while exceeding it in CKA; ChemBERTa and CellCLIP align worst, though above the random baseline. This ordering mirrors DEG50 performance: Target ID and GenePT give the highest DEG50 Δ-Pearson and

CellCLIP the lowest (Table 1), although GenePT is weaker on the full panel. On Tahoe-100M, where performance differs little across representations, local alignment is similarly compressed: all representations lie close to the random baseline, and even random neighbors respond similarly (Mean Response Similarity@5 of 0.47 vs. 0.13 on sci-Plex3), suggesting a strong response component shared across drugs within a plate (Appendix C).

Target ID benefits specifically when test-drug targets have been represented during training

Why does Target ID work so well on sci-Plex3? Our hypothesis is that a new drug is easier to predict when the training set already contains drugs acting on the same proteins. To check this, we group test drugs by how many of their targets are also hit by at least one training drug (Fig. 3, middle). The more targets a test drug shares with the training set, the larger Target ID's Δ-Pearson gain, although the trend is not strictly monotonic. This raises a simpler explanation: the model might just be copying the responses of training drugs that hit the same target. We test this with a "lookup" baseline that involves no learning. For each test drug, it averages the expression profiles of all training drugs that share at least one target with it. If the drug has no annotated targets, or none of its targets appear in training, it falls back to the cell-line perturbation mean. On the full gene panel (Fig. 3, right; gains over the cell-line perturbation mean in percentage points), this lookup barely helps (-0.20 Δ-Pearson, $+1.10$ R^2). The Target ID predictor improves both ($+1.99$ and $+4.56$), giving the largest R^2 gain in the panel. Grouping drugs by the hand-curated pathway labels of the original sci-Plex3 study (Srivatsan et al., 2020) gives a larger Δ-Pearson gain, but such labels are rarely available for new compounds. We therefore treat it as a reference point, not a practical alternative. **Shared targets make transfer possible, but the predictor does more than copy the responses of drugs with the same targets.**



	R^2	$\Delta\rho$	Δ-Dir
Overlapping targets	+1.10	−0.20	+0.23
Pathway (17)	+2.19	**+4.23**	**+2.62**
Finer pathways (55)	+3.29	+2.10	+2.41
ours	**+4.56**	+1.99	+2.25

Figure 3: Perturbation representations differ in how they connect drugs to observed transcriptional responses. **Left**: local (Recall@5, Jaccard@5, Mean Response Similarity@5) and global (CKA) alignment between feature-space and residual-response geometry on sci-Plex3 and Tahoe-100M; numbers are metric values and within-column ranks. **Middle**: Target-based performance improves as more test-drug targets are represented among training compounds, supporting transfer through recurring target mechanisms rather than direct target–output overlap. **Right**: improvement over the cell-line perturbation mean (sci-Plex3, full panel) when predicting each test drug as the average transcriptional responses (residual) of training drugs grouped by overlapping targets or by the sci-Plex3 pathway annotations (Srivatsan et al., 2020) (17 coarse, 55 fine groups).

Together, these results suggest that the most predictive representations are those that link an unseen drug to training drugs with similar responses, and that target annotations provide such links on sci-Plex3. This reflects generalization within a partially shared mechanism space rather than extrapolation to unseen mechanisms.

5.4 Ablation and Sensitivity Study

We ablate two components of our predictor: the Pearson-correlation loss and the LLM-derived BAITSAO cell-line representation used for FiLM conditioning (Table 3). For the latter, we replace the BAITSAO embedding with a one-hot cell-line encoding and learn the corresponding FiLM parameters from a zero-initialized lookup table. Removing the Pearson loss consistently degrades performance. The BAITSAO-based cell-line representation generally performs better on the full gene panel but worse on DEG50 compared with one-hot encoding. These trends remain consistent across perturbation representations, suggesting that the effects of both ablations are largely independent of the choice of perturbation embedding.

To assess whether the Tahoe-100M results generalize across plates, we additionally evaluate plates 3, 9, and 12, which share up to 3 compounds with each other (Appendix Table 4). Target ID is comparable to or better than the cell-line perturbation mean on most DEG50 metrics, with the largest gains in absolute metrics; the perturbation mean remains best on Δ-Pearson and Δ-Direction for plate 3. The two Target ID variants perform similarly, with neither consistently dominating.

Table 3: Residual-level test performance on the Full and DEG50 panels of the sci-Plex3 dataset (mean ± std across cell lines and seeds). Higher is better for all metrics. Each block starts from the configuration used in the main text (bulk-level MSE + Pearson, LLM cell-line embedding); ~~struck-out~~ terms are removed and → marks a replacement. **Bold** is the best mean per feature block/per column.

	Full panel			DEG50 panel	
Ablation term (bulk-level MSE + ...)	Δ-Pearson ↑	Δ-DEoverlap ↑	Δ-Direction ↑	Δ-Pearson ↑	Δ-Direction ↑
Molecular structure: RDKit-2D					
Pearson + LLM	**0.360** ±0.059	**0.465** ±0.025	0.576 ±0.024	0.206 ±0.140	0.608 ±0.075
~~Pearson~~ + LLM	0.344 ±0.047	0.447 ±0.030	0.567 ±0.020	0.160 ±0.125	0.592 ±0.081
Pearson + ~~LLM~~ → One-hot	0.292 ±0.062	0.448 ±0.021	0.576 ±0.014	**0.234** ±0.106	0.603 ±0.048
Pearson + ~~LLM~~ → FiLM	0.351 ±0.056	0.441 ±0.027	**0.581** ±0.015	0.196 ±0.111	**0.620** ±0.054
Target-aware: UniPert					
Pearson + LLM	**0.342** ±0.060	**0.458** ±0.026	**0.581** ±0.031	0.186 ±0.107	0.599 ±0.061
~~Pearson~~ + LLM	0.335 ±0.058	0.448 ±0.032	0.564 ±0.020	0.123 ±0.146	0.570 ±0.085
Pearson + ~~LLM~~ → One-hot	0.270 ±0.067	0.408 ±0.059	0.572 ±0.018	**0.232** ±0.080	**0.625** ±0.048
Pearson + ~~LLM~~ → FiLM	**0.342** ±0.052	0.446 ±0.023	0.580 ±0.017	0.217 ±0.101	0.618 ±0.060
Target-aware: Target ID (binding prob.)					
Pearson + LLM	**0.356** ±0.058	**0.457** ±0.028	**0.584** ±0.024	0.291 ±0.095	0.657 ±0.046
~~Pearson~~ + LLM	0.339 ±0.063	0.444 ±0.047	0.562 ±0.022	0.107 ±0.132	0.568 ±0.077
Pearson + ~~LLM~~ → One-hot	0.323 ±0.070	0.434 ±0.054	0.574 ±0.017	**0.306** ±0.082	**0.668** ±0.044
Pearson + ~~LLM~~ → FiLM	0.327 ±0.102	0.399 ±0.087	0.576 ±0.024	0.247 ±0.105	0.635 ±0.058

Table 4: **Cross-plate, DEG50 panel** Test performance on the DEG50 gene panel for plates 3, 9 and 12 of the Tahoe-100M dataset (mean ± std across cell lines and seeds). **Bold** marks the best mean per column within each plate. We here show the cell-line specific perturbation mean and pseudobulk MLP trained on Target ID performance. See Appendix C for other features.

Model / representation	R^2 ↑	Pearson ↑	RMSE ↓	ΔR^2 ↑	Δ-Pearson ↑	Δ-Direction ↑
Plate 3						
Cell-line perturbation mean	0.454 ±0.385	0.814 ±0.130	0.511 ±0.119	0.297 ±0.244	**0.760** ±0.041	**0.879** ±0.025
Target ID (avg.)	0.590 ±0.213	0.817 ±0.120	0.493 ±0.065	0.384 ±0.065	0.726 ±0.036	0.857 ±0.023
Target ID (binding prob.)	**0.601** ±0.207	**0.823** ±0.118	**0.489** ±0.062	**0.385** ±0.072	0.746 ±0.038	0.870 ±0.023
Plate 9						
Cell-line perturbation mean	0.730 ±0.308	0.896 ±0.133	0.321 ±0.124	0.548 ±0.125	0.817 ±0.067	0.904 ±0.038
Target ID (avg.)	**0.782** ±0.275	**0.912** ±0.126	**0.295** ±0.110	**0.581** ±0.136	**0.822** ±0.084	**0.908** ±0.045
Target ID (binding prob.)	**0.782** ±0.274	**0.912** ±0.126	**0.295** ±0.111	0.580 ±0.139	0.821 ±0.083	**0.908** ±0.045
Plate 12						
Cell-line perturbation mean	0.724 ±0.202	0.887 ±0.112	0.370 ±0.096	0.144 ±0.280	0.706 ±0.076	0.847 ±0.053
Target ID (avg.)	**0.738** ±0.197	**0.891** ±0.110	**0.365** ±0.093	**0.166** ±0.261	**0.713** ±0.074	**0.853** ±0.048
Target ID (binding prob.)	0.731 ±0.193	0.888 ±0.109	0.369 ±0.088	0.145 ±0.274	0.683 ±0.098	0.829 ±0.055

6 Conclusion

We revisit chemical perturbation prediction at pseudobulk resolution, where a lightweight MLP outperforms learning-free baselines and representative single-cell models on most metrics. This has three implications. First, single-cell models evaluated on aggregated profiles should be benchmarked against strong learned pseudobulk predictors, not only learning-free means, and showing a benefit from modeling heterogeneity requires metrics that the unperturbed control distribution cannot already satisfy. Second, progress on unseen drugs may come less from more complex molecular encoders than from representations that link new compounds to related profiled perturbations, such as through shared targets. Third, performance depends on the data, namely the noise in measured responses and whether test drugs have mechanistically related drugs in training. These results highlight the potential of simple pseudobulk predictors and motivate future work on using pseudobulk predictions as priors for modeling single-cell response heterogeneity.

AI USE STATEMENT

In this work, we used generative AI tools for grammar and writing polishing. All equations, analysis, and research contributions are our own. We also leveraged LLM for enriched targets retrieval. We verified the precision of the retrieved targets by comparing them with the ground truth annotations.

REPRODUCIBILITY STATEMENT

All precomputed compound and cell-line feature embeddings used in this work will be released publicly upon acceptance.

REFERENCES

Abhinav K Adduri, Dhruv Gautam, Beatrice Bevilacqua, Mohsen Naghipourfar, Alishba Imran, Rohan Shah, Noam Teyssier, Rishi Verma, Christopher Carpenter, Basak Eraslan, et al. Predicting cellular responses to perturbation across diverse contexts with State. *Cell*, 2026.

Constantin Ahlmann-Eltze, Wolfgang Huber, and Simon Anders. Deep-learning-based gene perturbation effect prediction does not yet outperform simple linear baselines. *Nature Methods*, 22(8):1657–1661, Aug 2025. ISSN 1548-7105. doi: 10.1038/s41592-025-02772-6. URL https://doi.org/10.1038/s41592-025-02772-6.

Ihab Bendidi, Shawn Whitfield, Kian Kenyon-Dean, Hanene Ben Yedder, Yassir El Mesbahi, Emmanuel Noutahi, and Alisandra K. Denton. Benchmarking transcriptomics foundation models for perturbation analysis : one PCA still rules them all, 2024. URL https://arxiv.org/abs/2410.13956.

Yiqun Chen and James Zou. GenePT: A simple but effective foundation model for genes and cells built from ChatGPT. *bioRxiv*, 2024. doi: 10.1101/2023.10.16.562533. URL https://www.biorxiv.org/content/early/2024/03/05/2023.10.16.562533.

Changxi Chi, Jun Xia, Yufei Huang, Zhuoli Ouyang, Cheng Tan, Yunfan Liu, Jingbo Zhou, Chang Yu, Liangyu Yuan, Siyuan Li, et al. Doloris: Dual conditional diffusion implicit bridges with sparsity masking strategy for unpaired single-cell perturbation estimation. In *International Conference on Learning Representations*, volume 2026, pp. 52808–52823, 2026.

Gerold Csendes, Gema Sanz, Kristóf Z Szalay, and Bence Szalai. Benchmarking foundation cell models for post-perturbation RNA-seq prediction. *BMC Genomics*, 26(1):393, April 2025.

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nature Methods*, 21(8):1470–1480, Aug 2024. ISSN 1548-7105. doi: 10.1038/s41592-024-02201-0. URL https://doi.org/10.1038/s41592-024-02201-0.

Jonathan G. Hedley, Philip H. S. Torr, and Kaspar Märtens. GenePT revisited: Do better text embeddings make better gene embeddings? *bioRxiv*, 2026. doi: 10.64898/2026.04.16.718976. URL https://www.biorxiv.org/content/early/2026/04/20/2026.04.16.718976.

Leon Hetzel, Simon Böhm, Niki Kilbertus, Stephan Günnemann, Mohammad Lotfollahi, and Fabian J Theis. Predicting cellular responses to novel drug perturbations at a single-cell resolution. In *NeurIPS 2022*, 2022.

Eric Kernfeld, Yunxiao Yang, Joshua S. Weinstock, Alexis Battle, and Patrick Cahan. A comparison of computational methods for expression forecasting. *Genome Biology*, 26(1):388, Nov 2025. ISSN 1474-760X. doi: 10.1186/s13059-025-03840-y. URL https://doi.org/10.1186/s13059-025-03840-y.

Huan Yee Koh, Anh TN Nguyen, Shirui Pan, Lauren T May, and Geoffrey I Webb. Physicochemical graph neural network for learning protein–ligand interaction fingerprints from sequence data. *Nature Machine Intelligence*, 6(6):673–687, 2024.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey E. Hinton. Similarity of neural network representations revisited. *CoRR*, abs/1905.00414, 2019. URL http://arxiv.org/abs/1905.00414.

Yiming Li, Min Zeng, Jun Zhu, Linjing Liu, Fang Wang, Longkai Huang, Fan Yang, Min Li, and Jianhua Yao. Unipert-G2CP bridges genetic and chemical screens from molecular representation to phenotype modeling. *Cell*, 2026.

Tianyu Liu, Kexing Zhou, Yuge Wang, and Hongyu Zhao. GenePert: Leveraging genept embeddings for gene perturbation prediction. *bioRxiv*, 2024. doi: 10.1101/2024.10.27.620513. URL https://www.biorxiv.org/content/10.1101/2024.10.27.620513.

Tianyu Liu, Tianqi Chen, Wangjie Zheng, Xiao Luo, Yiqun Chen, and Hongyu Zhao. scELMo: Embeddings from language models are good learners for single-cell data analysis. *bioRxiv*, 2025a. doi: 10.1101/2023.12.07.569910. URL https://www.biorxiv.org/content/early/2025/08/23/2023.12.07.569910.

Tianyu Liu, Tinyi Chu, Xiao Luo, and Hongyu Zhao. Building a unified model for drug synergy analysis powered by large language models. *Nature Communications*, 16(1):4537, May 2025b. ISSN 2041-1723. doi: 10.1038/s41467-025-59822-y. URL https://doi.org/10.1038/s41467-025-59822-y.

Mohammad Lotfollahi, F. Alexander Wolf, and Fabian J. Theis. scGen predicts single-cell perturbation responses. *Nature Methods*, 16(8):715–721, Aug 2019. ISSN 1548-7105. doi: 10.1038/s41592-019-0494-8. URL https://doi.org/10.1038/s41592-019-0494-8.

Mohammad Lotfollahi, Anna Klimovskaia Susmelj, Carlo De Donno, Yuge Ji, Ignacio L. Ibarra, F. Alexander Wolf, Nafissa Yakubova, Fabian J. Theis, and David Lopez-Paz. Learning interpretable cellular responses to complex perturbations in high-throughput screens. *bioRxiv*, 2021. doi: 10.1101/2021.04.14.439903. URL https://www.biorxiv.org/content/early/2021/05/18/2021.04.14.439903.

Mingyu Lu, Ethan Weinberger, Chanwoo Kim, and Su-In Lee. CellCLIP – learning perturbation effects in cell painting via text-guided contrastive learning, 2025. URL https://arxiv.org/abs/2506.06290.

Gabriel Mateo Mejia, Henry E Miller, Francis J.A. Leblanc, BO WANG, Brendan Swain, and Lucas Paulo de Lima Camillo. Needles in the haystack: Addressing signal dilution improves scRNA-seq perturbation response modeling and evaluation. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=XsrXLPxBJw.

Nhung T H Nguyen, Dmytro S Lituiev, Zhimin Liu, Aditya Kashyap, Garrett Jenkinson, Kevin Kuhl, Christopher Corrado, Naisargi Manishkumar Patel, Kirti Snigdha, Sirwe Saeedi, David Smith, Nicholas Baro, and Timothy Schultz. Med.ai ASK: an agentic system for biomedical question answering. *Journal of the American Medical Informatics Association*, 33(6):1134–1145, 06 2026. ISSN 1527-974X. doi: 10.1093/jamia/ocag038. URL https://doi.org/10.1093/jamia/ocag038.

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

Zoe Piran, Niv Cohen, Yedid Hoshen, and Mor Nitzan. Disentanglement of single-cell data with Biolord. *Nature Biotechnology*, 42(11):1678–1683, 2024.

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

Yusuf Roohani, Kexin Huang, and Jure Leskovec. Predicting transcriptional outcomes of novel multigene perturbations with GEARS. *Nature Biotechnology*, 2023.

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, Dec 2022. ISSN 2522-5839. doi: 10.1038/s42256-022-00580-7. URL https://doi.org/10.1038/s42256-022-00580-7.

Ana Sanchez-Fernandez, Elisabeth Rumetshofer, Sepp Hochreiter, and Günter Klambauer. CLOOME: contrastive learning unlocks bioimaging databases for queries with chemical structures. *Nat. Commun.*, 14(1):7339, November 2023.

Nikhil Shenoy, David Errington, Emmanuel Bengio, Kacper Kapuśniak, Kerstin Klaeser, Yui Tik Pang, Vladimir Radenkovic, Prudencio Tossou, Therence Bois, Andrew Wedlake, and Francesco Di Giovanni. Nesso-1: Accelerating open-source binding affinity predictions. *bioRxiv*, 2026. doi: 10.64898/2026.08.01.742196. URL https://www.biorxiv.org/content/10.64898/2026.08.01.742196v1.

Riya Singh, Aryan Amit Barsainyan, Rida Irfan, Connor Joseph Amorin, Stewart He, Tony Davis, Arun Thiagarajan, Shiva Sankaran, Seyone Chithrananda, Walid Ahmad, Derek Jones, Kevin McLoughlin, Hyojin Kim, Anoushka Bhutani, Shreyas Vinaya Sathyanarayana, Venkat Viswanathan, Jonathan E. Allen, and Bharath Ramsundar. ChemBERTa-3: an open source training framework for chemical foundation models. *Digital Discovery*, 5(2):662–685, 02 2026. ISSN 2635-098X. doi: 10.1039/d5dd00348b. URL https://doi.org/10.1039/d5dd00348b.

Sanjay R. Srivatsan, José L. McFaline-Figueroa, Vijay Ramani, Lauren Saunders, Junyue Cao, Jonathan Packer, Hannah A. Pliner, Dana L. Jackson, Riza M. Daza, Lena Christiansen, Fan Zhang, Frank Steemers, Jay Shendure, and Cole Trapnell. Massively multiplex chemical transcriptomics at single-cell resolution. *Science*, 367(6473):45–51, 2020. doi: 10.1126/science.aax6234. URL https://www.science.org/doi/abs/10.1126/science.aax6234.

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

Ramon Vinas Torne, Maciej Wiatrak, Zoe Piran, Shuyang Fan, Liangze Jiang, Sarah A. Teichmann, Mor Nitzan, and Maria Brbic. Systema: A framework for evaluating genetic perturbation response prediction beyond systematic variation. *Nature Biotechnology*, 2025.

Chloe Wang, Mehran Karimzadeh, Neal G. Ravindra, Lexi R. Bounds, Nader Alerasool, Ann C. Huang, Shihao Ma, Daniel R. Gulbranson, Haotian Cui, Yongju Lee, Anusuya Arjavalingam, Elliot J. MacKrell, Matthew S. Wilken, Jieming Chen, Benjamin W. Herken, Jesse A. Weber, Massimo M. Onesto, Barbara Gonzalez-Teran, Nicole F. Leung, Sally Yu Shi, Byron J. Smith, Sharon K. Lam, Adam Barner, Philip Wright, Elizabeth M. Rumsey, Soohong Kim, Rene V. Sit, Adam J. Litterman, Ci Chu, and Bo Wang. X-cell: Scaling causal perturbation prediction across diverse cellular contexts via diffusion language models. *bioRxiv*, 2026. doi: 10.64898/2026.03.18.712807. URL https://www.biorxiv.org/content/early/2026/03/20/2026.03.18.712807.

Gefei Wang, Tianyu Liu, Jia Zhao, Youshu Cheng, and Hongyu Zhao. Modeling and predicting single-cell multi-gene perturbation responses with scLAMBDA. *bioRxiv*, 2024. doi: 10.1101/2024.12.04.626878. URL https://www.biorxiv.org/content/early/2024/12/08/2024.12.04.626878.

Zhiting Wei, Yiheng Wang, Yicheng Gao, Shuguang Wang, Ping Li, Duanmiao Si, Yuli Gao, Siqi Wu, Danlu Li, Kejing Dong, Xingbo Yang, Chen Tang, Shaliu Fu, Xiaohan Chen, Wannian Li, Yuzhou You, Chen Zhang, Aibin Liang, Guohui Chuai, and Qi Liu. Benchmarking algorithms

for generalizable single-cell perturbation response prediction. *Nature Methods*, 23(2):451–464, Feb 2026. ISSN 1548-7105. doi: 10.1038/s41592-025-02980-0. URL https://doi.org/10.1038/s41592-025-02980-0.

Frederik Wenkel, Wilson Tu, Cassandra Masschelein, Hamed Shirzad, Liam Hodgson, Ihab Bendidi, Cian Eastwood, Shawn T. Whitfield, Craig Russell, Yassir El Mesbahi, Jiarui Ding, Marta M. Fay, Berton Earnshaw, Emmanuel Noutahi, and Alisandra K. Denton. TxPert: using multiple knowledge graphs for prediction of transcriptomic perturbation effects. *Nature Biotechnology*, May 2026. ISSN 1546-1696. doi: 10.1038/s41587-026-03113-4. URL https://doi.org/10.1038/s41587-026-03113-4.

Daniel R Wong, Abby S Hill, and Rob Moccia. Simple controls exceed best deep learning algorithms and reveal foundation model effectiveness for predicting genetic perturbations. *Bioinformatics*, 41(6):btaf317, 06 2025. ISSN 1367-4811. doi: 10.1093/bioinformatics/btaf317. URL https://doi.org/10.1093/bioinformatics/btaf317.

Yan Wu, Esther Wershof, Sebastian M Schmon, Marcel Nassar, Błażej Osiński, Ridvan Eksi, Kun Zhang, and Thore Graepel. PerturBench: Benchmarking machine learning models for cellular perturbation analysis. In *NeurIPS 2024 Workshop on AI for New Drug Modalities*, 2024. URL https://openreview.net/forum?id=1EJXtK3AWr.

Ido Yofe, Rony Dahan, and Ido Amit. Single-cell genomic approaches for developing the next generation of immunotherapies. *Nature Medicine*, 26:171 – 177, 2020. URL https://api.semanticscholar.org/CorpusID:211017696.

Xinyu Yuan, Xixian Liu, Ya Shi Zhang, Zuobai Zhang, Hongyu Guo, and Jian Tang. PerturbDiff: Functional diffusion for single-cell perturbation modeling. *arXiv preprint arXiv:2602.19685*, 2026.

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*, 2025a. doi: 10.1101/2025.02.20.639398. URL https://www.biorxiv.org/content/early/2025/02/24/2025.02.20.639398.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 embedding: Advancing text embedding and reranking through foundation models, 2025b. URL https://arxiv.org/abs/2506.05176.

Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. Uni-Mol: A universal 3D molecular representation learning framework. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=6K2RM6wVqKu.

A Ablation

A.1 Cell-line embedding

In this section we compare three ways of encoding the cell line: FiLM modulation driven by the LLM-derived BAITSAO embedding (used in the main text), a one-hot cell-line indicator concatenated to the drug embedding, and FiLM parameters read from a zero-initialised learnable lookup table. Tables 5 and 6 compare the three on the top-3 perturbation encodings.

The ranking of cell-line encodings depends on the metric, the perturbation encoding and the panel. On the full panel, FiLM modulation driven by the LLM-derived BAITSAO embedding ranks first or ties on most of the Δ-metric of Table 5, losing only Δ-Direction under RDKit-2D and by 0.005, whereas the one-hot indicator is the stronger choice on the absolute metrics and on DEG50 Δ-Pearson. The margins lie within one standard deviation. The DEG50 panel is defined by the ground-truth top-50 DE genes of the held-out (drug, cell line) pair, which are unavailable for a compound whose mechanism is unknown, leaving full-panel Δ metrics as the setting a genuinely novel drug would be scored under. Therefore, we report the results of using the LLM-based FiLM-based cell line embedding method in the main manuscript.

Table 5: **Cell-line encoding ablation, full gene panel.** Test performance as mean $\pm$ std across cell lines and seeds. **Bold** marks the best mean per column within each perturbation encoding. Stars mark models that are significantly better than the best learning-free baseline of that column ($*$ $p < 0.05$, $**$ $p < 0.01$, $***$ $p < 0.001$; paired t-test).One-hot: cell-line indicator concatenated to the perturbation embedding; LLM: FiLM driven by the LLM-derived BAITSAO embedding (main-text setting); FiLM: FiLM parameters from a zero-initialised learnable lookup table.

Feature	Cell line	R^2 ↑	Pearson ↑	MSE ↓	RMSE ↓	Δ-Pearson ↑	Δ-DEoverlap ↑	Δ-Direction ↑
Learning-free baselines								
Cell-line control mean		0.824 ±0.063	0.919 ±0.026	0.004 ±0.002	0.065 ±0.011	0.024 ±0.028	0.437 ±0.018	0.000 ±0.000
Cell-line perturbation mean		0.840 ±0.058	0.925 ±0.024	0.004 ±0.001	0.062 ±0.011	0.336 ±0.049	0.428 ±0.025	0.561 ±0.019
Global control mean		0.469 ±0.027	0.736 ±0.013	0.013 ±0.001	0.113 ±0.003	0.133 ±0.066	0.329 ±0.096	0.477 ±0.026
Global perturbation mean		0.480 ±0.025	0.742 ±0.012	0.013 ±0.001	0.112 ±0.003	0.154 ±0.065	0.341 ±0.096	0.483 ±0.029
Ours: pseudobulk MLP on a frozen perturbation embedding								
RDKit-2D	One-hot	0.875 ±0.056*	0.945 ±0.022**	0.003 ±0.001**	0.053 ±0.012**	0.292 ±0.062	0.448 ±0.021*	0.576 ±0.014***
	LLM	**0.880** ±0.054**	**0.947** ±0.021**	0.003 ±0.001**	**0.051** ±0.012**	**0.360** ±0.059*	**0.465** ±0.025***	0.576 ±0.024**
	FiLM	0.871 ±0.052**	0.942 ±0.020**	0.003 ±0.001**	0.054 ±0.011**	0.351 ±0.056	0.441 ±0.027	**0.581** ±0.015***
UniPert	One-hot	0.876 ±0.057***	0.948 ±0.022***	0.003 ±0.001**	0.052 ±0.012***	0.270 ±0.067	0.408 ±0.059	0.572 ±0.018*
	LLM	0.876 ±0.052***	0.944 ±0.020***	0.003 ±0.001**	0.053 ±0.010***	0.342 ±0.060	**0.458** ±0.026***	**0.581** ±0.031**
	FiLM	**0.879** ±0.055**	0.948 ±0.020**	0.003 ±0.001**	0.052 ±0.010**	0.342 ±0.052	0.446 ±0.023	0.580 ±0.017***
Target ID (binding prob.)	One-hot	**0.898** ±0.047**	**0.957** ±0.016**	**0.002** ±0.001**	**0.046** ±0.009**	0.323 ±0.070	0.434 ±0.054	0.574 ±0.017**
	LLM	0.886 ±0.050**	0.951 ±0.018**	0.003 ±0.001**	0.050 ±0.010***	**0.356** ±0.058	**0.457** ±0.028***	**0.584** ±0.024**
	FiLM	0.879 ±0.045**	0.947 ±0.017**	0.003 ±0.001**	0.052 ±0.009***	0.327 ±0.102	0.399 ±0.087	0.576 ±0.024**

Table 6: **Cell-line encoding ablation, DEG50 panel.** Test performance as mean $\pm$ std across cell lines and seeds. **Bold** marks the best mean per column within each perturbation encoding. Stars mark models that are significantly better than the best learning-free baseline of that column ($*$ $p < 0.05$, $**$ $p < 0.01$, $***$ $p < 0.001$; paired t-test). Cell-line encodings as in Table 5.

Feature	Cell line	R^2 ↑	Pearson ↑	MSE ↓	RMSE ↓	Δ-Pearson ↑	Δ-Direction ↑
Learning-free baselines							
Cell-line control mean		0.701 ±0.099	0.967 ±0.013	0.059 ±0.031	0.236 ±0.060	−0.066 ±0.078	0.000 ±0.000
Cell-line perturbation mean		0.714 ±0.092	0.968 ±0.012	0.055 ±0.029	0.228 ±0.058	0.156 ±0.104	0.596 ±0.069
Global control mean		0.200 ±0.165	0.780 ±0.086	0.116 ±0.047	0.334 ±0.066	−0.278 ±0.078	0.290 ±0.086
Global perturbation mean		0.204 ±0.167	0.776 ±0.085	0.113 ±0.046	0.330 ±0.065	−0.238 ±0.075	0.323 ±0.070
Ours: pseudobulk MLP on a frozen perturbation embedding							
RDKit-2D	One-hot	**0.789** ±0.088***	0.968 ±0.011	**0.035** ±0.023***	**0.180** ±0.056***	**0.234** ±0.106***	0.603 ±0.048
	LLM	0.787 ±0.084***	**0.970** ±0.009	0.036 ±0.022***	0.181 ±0.053***	0.206 ±0.140*	0.608 ±0.075
	FiLM	0.770 ±0.081***	0.969 ±0.010	0.040 ±0.023***	0.194 ±0.053***	0.196 ±0.111**	**0.620** ±0.054***
UniPert	One-hot	**0.801** ±0.078***	0.969 ±0.011	**0.033** ±0.020***	**0.173** ±0.052***	**0.232** ±0.080**	**0.625** ±0.048
	LLM	0.771 ±0.079***	**0.971** ±0.009*	0.040 ±0.021***	0.194 ±0.049***	0.186 ±0.107**	0.599 ±0.061
	FiLM	0.787 ±0.071***	0.969 ±0.009	0.036 ±0.020***	0.185 ±0.049***	0.217 ±0.101***	0.618 ±0.060*
Target ID (binding prob.)	One-hot	**0.839** ±0.065***	**0.975** ±0.009**	**0.025** ±0.014***	**0.154** ±0.039***	**0.306** ±0.082***	**0.668** ±0.044***
	LLM	0.805 ±0.074***	0.974 ±0.009*	0.033 ±0.019***	0.175 ±0.050***	0.291 ±0.095***	0.657 ±0.046***
	FiLM	0.793 ±0.066***	0.971 ±0.009*	0.035 ±0.019***	0.183 ±0.046***	0.247 ±0.105***	0.635 ±0.058***

A.2 Loss function

In this section, we demonstrate the importance of the pearson loss term for the model performance. We compare the results with using only MSE loss and the results with using both loss. We see that adding the pearson loss can improve the performance nearly across all the evaluation metrics, input features, and gene panels.

Table 7: **Loss-function ablation, full gene panel.** Test performance as mean ± std across cell lines and seeds. **Bold** marks the best mean per column within each perturbation encoding. Stars mark models that are significantly better than the best learning-free baseline of that column ($* \ p < 0.05$, $** \ p < 0.01$, $*** \ p < 0.001$; paired t-test). "With Pearson loss" is the main-text setting.

Feature	Loss	R^2 ↑	Pearson ↑	MSE ↓	RMSE ↓	Δ-Pearson ↑	Δ-DEoverlap ↑	Δ-Direction ↑
Learning-free baselines								
Cell-line control mean		0.824 ±0.063	0.919 ±0.026	0.004 ±0.002	0.065 ±0.011	0.024 ±0.028	0.437 ±0.018	0.000 ±0.000
Cell-line perturbation mean		0.840 ±0.058	0.925 ±0.024	0.004 ±0.001	0.062 ±0.011	0.336 ±0.049	0.428 ±0.025	0.561 ±0.019
Global control mean		0.469 ±0.027	0.736 ±0.013	0.013 ±0.001	0.113 ±0.003	0.133 ±0.066	0.329 ±0.096	0.477 ±0.026
Global perturbation mean		0.480 ±0.025	0.742 ±0.012	0.013 ±0.001	0.112 ±0.003	0.154 ±0.065	0.341 ±0.096	0.483 ±0.029
Ours: pseudobulk MLP on a frozen perturbation embedding								
RDKit-2D	Without Pearson loss	0.878 ±0.048*	0.946 ±0.021*	0.003 ±0.001*	0.052 ±0.011*	0.344 ±0.047	0.447 ±0.030	0.567 ±0.020*
	With Pearson loss	**0.880** ±0.054**	**0.947** ±0.021**	0.003 ±0.001**	**0.051** ±0.012**	**0.360** ±0.059*	**0.465** ±0.025***	**0.576** ±0.024**
UniPert	Without Pearson loss	0.872 ±0.050*	0.944 ±0.019*	0.003 ±0.001*	0.054 ±0.009**	0.335 ±0.058	0.448 ±0.032*	0.564 ±0.020
	With Pearson loss	**0.876** ±0.052***	0.944 ±0.020***	0.003 ±0.001**	**0.053** ±0.010***	**0.342** ±0.060	**0.458** ±0.026***	**0.581** ±0.031**
Target ID (binding prob.)	Without Pearson loss	0.884 ±0.045**	0.949 ±0.017**	0.003 ±0.001*	0.050 ±0.009**	0.339 ±0.063	0.444 ±0.047	0.562 ±0.022
	With Pearson loss	**0.886** ±0.050**	**0.951** ±0.018**	0.003 ±0.001**	0.050 ±0.010***	**0.356** ±0.058	**0.457** ±0.028***	**0.584** ±0.024**

Table 8: **Loss-function ablation, DEG50 panel.** Test performance as mean ± std across cell lines and seeds. **Bold** marks the best mean per column within each perturbation encoding. Stars mark models that are significantly better than the best learning-free baseline of that column ($* \ p < 0.05$, $** \ p < 0.01$, $*** \ p < 0.001$; paired t-test). "With Pearson loss" is the main-text setting.

Feature	Loss	R^2 ↑	Pearson ↑	MSE ↓	RMSE ↓	Δ-Pearson ↑	Δ-Direction ↑
Learning-free baselines							
Cell-line control mean		0.701 ±0.099	0.967 ±0.013	0.059 ±0.031	0.236 ±0.060	−0.066 ±0.078	0.000 ±0.000
Cell-line perturbation mean		0.714 ±0.092	0.968 ±0.012	0.055 ±0.029	0.228 ±0.058	0.156 ±0.104	0.596 ±0.069
Global control mean		0.200 ±0.165	0.780 ±0.086	0.116 ±0.047	0.334 ±0.066	−0.278 ±0.078	0.290 ±0.086
Global perturbation mean		0.204 ±0.167	0.776 ±0.085	0.113 ±0.046	0.330 ±0.065	−0.238 ±0.075	0.323 ±0.070
Ours: pseudobulk MLP on a frozen perturbation embedding							
RDKit-2D	Without Pearson loss	0.781 ±0.097***	0.969 ±0.010	0.036 ±0.024***	0.182 ±0.055***	0.160 ±0.125	0.592 ±0.081
	With Pearson loss	**0.787** ±0.084***	**0.970** ±0.009	0.036 ±0.022***	**0.181** ±0.053***	**0.206** ±0.140*	**0.608** ±0.075
UniPert	Without Pearson loss	0.770 ±0.070***	0.968 ±0.008	**0.038** ±0.018***	**0.190** ±0.043***	0.123 ±0.146	0.570 ±0.085
	With Pearson loss	**0.771** ±0.079***	**0.971** ±0.009*	0.040 ±0.021***	0.194 ±0.049***	**0.186** ±0.107**	**0.599** ±0.061
Target ID (binding prob.)	Without Pearson loss	0.786 ±0.075***	0.967 ±0.012	**0.032** ±0.016***	0.175 ±0.039***	0.107 ±0.132	0.568 ±0.077
	With Pearson loss	**0.805** ±0.074***	**0.974** ±0.009*	0.033 ±0.019***	0.175 ±0.050***	**0.291** ±0.095***	**0.657** ±0.046***

A.3 Tahoe-100M dataset

Table 9 repeats the objective and cell-line ablations on Tahoe-100M. Unlike on sci-Plex3, where the Pearson term improves every Δ metric for all three features, here it brings no gain. Without it the model gives the best mean on all five full-panel metrics for every feature, and on all four DEG50 metrics for RDKit-2D. For UniPert and Target ID (avg.) the two objectives trade places on DEG50. The differences are small, at most 0.024 in full-panel ΔR^2 (RDKit-2D, 0.459 vs. 0.435), and well within one standard deviation. A plausible reason is that MSE alone already captures the Tahoe-100M responses well (Δ-Pearson ≥ 0.73 on the full panel), which leaves the correlation term little to add. Replacing the LLM cell-line embedding with a learnable FiLM lookup lowers every metric for UniPert and Target ID (avg.) and leaves RDKit-2D essentially unchanged. These margins are also well below the spread across cell lines and seeds.

B Experiments with using all the doses

We here show the performance of using all the doses. We encode the dose as an one-hot vector and append it to the perturbation feature representation. Pooling all doses leaves the picture of the 10 μM setting largely unchanged. Every perturbation encoding matches or exceeds the learning-free baselines on the absolute metrics and on Δ-Pearson, and target-aware encodings again rank at the top.

Table 9: **Ablation, Tahoe-100M** Residual-level test performance on the Full and DEG50 panels of the Tahoe-100M dataset (mean ± std across cell lines and seeds). Higher is better for all metrics. Each block starts from the configuration used in the main text (bulk-level MSE + Pearson, LLM cell-line embedding), ~~struck-out~~ terms are removed, → marks a replacement, and **bold** marks the best mean per column within each feature block.

	Full panel					DEG50 panel			
Ablation term (bulk-level MSE + …)	ΔR^2 ↑	Δ-Pearson ↑	Δ-Spearman ↑	Δ-DEoverlap ↑	Δ-Direction ↑	ΔR^2 ↑	Δ-Pearson ↑	Δ-Spearman ↑	Δ-Direction ↑
Molecular structure: RDKit-2D									
Pearson + LLM	0.435 ±0.124	0.734 ±0.072	0.668 ±0.071	0.494 ±0.071	0.727 ±0.036	0.505 ±0.143	0.809 ±0.063	0.792 ±0.063	0.868 ±0.050
~~Pearson~~ + LLM	**0.459** ±0.110	**0.741** ±0.069	**0.677** ±0.067	**0.503** ±0.069	**0.738** ±0.032	**0.507** ±0.136	**0.812** ±0.061	**0.795** ±0.062	**0.871** ±0.048
Pearson + ~~LLM~~ → FiLM	0.443 ±0.122	0.734 ±0.071	0.666 ±0.070	0.494 ±0.073	0.725 ±0.035	0.503 ±0.136	0.809 ±0.061	0.793 ±0.061	0.870 ±0.047
Target-aware: UniPert									
Pearson + LLM	0.468 ±0.104	0.742 ±0.070	0.678 ±0.067	0.504 ±0.069	0.735 ±0.032	**0.510** ±0.137	0.812 ±0.062	0.795 ±0.062	**0.872** ±0.047
~~Pearson~~ + LLM	**0.469** ±0.101	**0.743** ±0.069	**0.681** ±0.066	**0.505** ±0.068	**0.742** ±0.030	0.508 ±0.136	**0.813** ±0.061	**0.796** ±0.061	0.871 ±0.048
Pearson + ~~LLM~~ → FiLM	0.444 ±0.121	0.733 ±0.072	0.666 ±0.069	0.495 ±0.070	0.727 ±0.031	0.498 ±0.142	0.807 ±0.068	0.790 ±0.067	0.867 ±0.050
Target-aware: Target ID (avg.)									
Pearson + LLM	0.457 ±0.107	0.739 ±0.069	0.674 ±0.067	0.501 ±0.069	0.731 ±0.032	**0.509** ±0.140	**0.813** ±0.061	**0.796** ±0.061	**0.871** ±0.047
~~Pearson~~ + LLM	**0.462** ±0.104	**0.741** ±0.069	**0.678** ±0.066	**0.502** ±0.068	**0.737** ±0.031	0.505 ±0.139	0.812 ±0.061	0.795 ±0.062	0.870 ±0.048
Pearson + ~~LLM~~ → FiLM	0.441 ±0.103	0.732 ±0.067	0.665 ±0.065	0.494 ±0.067	0.725 ±0.029	0.502 ±0.138	0.807 ±0.062	0.791 ±0.062	0.869 ±0.046

The gaps between encodings shrink, however. Lower doses induce weaker responses that lie closer to the control, which compresses the differences between models. One difference from the 10 μM setting stands out: the gene-text embeddings (GenePT, Qwen3) no longer collapse to the control mean and instead give the best scores on most metrics. A possible cause is that the additional dose conditions supply enough training signal for these embeddings to become informative. ChemCPA stays below our models on most metrics; it keeps an edge in full-panel Δ-Direction but falls well behind on the DEG50 panel.

Table 10: **Full gene panel, all doses** Test performance as mean ± std across cell lines and seeds. **Bold** marks the best mean per column; ties are all marked. Stars mark our models that are significantly better than the best learning-free baseline of that column ($* \ p < 0.05$, $** \ p < 0.01$, $*** \ p < 0.001$; paired t-test)

Model / representation		R^2 ↑	Pearson ↑	MSE ↓	RMSE ↓	Δ-Pearson ↑	Δ-Direction ↑
Learning-free baselines							
Cell-line control mean		0.910 ±0.035	0.958 ±0.016	0.002 ±0.001	0.046 ±0.009	0.023 ±0.016	0.000 ±0.000
Cell-line perturbation mean		0.917 ±0.031	0.961 ±0.014	0.002 ±0.001	0.044 ±0.009	0.255 ±0.049	0.544 ±0.014
Global control mean		0.505 ±0.014	0.756 ±0.008	0.012 ±0.000	0.110 ±0.002	0.068 ±0.037	0.466 ±0.021
Global perturbation mean		0.511 ±0.012	0.759 ±0.007	0.012 ±0.000	0.109 ±0.001	0.084 ±0.036	0.468 ±0.023
Single-cell models (trained on single cells)							
ChemCPA Hetzel et al. (2022)		0.823 ±0.035	0.920 ±0.016	0.004 ±0.001	0.058 ±0.005	0.233 ±0.043	**0.597** ±0.011
Ours: pseudobulk MLP on a frozen perturbation embedding							
Molecular structure	RDKit-2D	0.922 ±0.031**	0.964 ±0.014***	0.002 ±0.001	0.042 ±0.009**	0.280 ±0.038***	0.551 ±0.014*
	Morgan FP	0.921 ±0.030*	0.964 ±0.014*	0.002 ±0.001	0.043 ±0.009*	0.280 ±0.050***	0.549 ±0.021
	Uni-Mol	0.923 ±0.029**	0.964 ±0.013**	0.002 ±0.001	0.042 ±0.009**	0.274 ±0.053**	0.548 ±0.024
Chemical LLM	ChemBERTa	0.920 ±0.032*	0.962 ±0.015	0.002 ±0.001	0.043 ±0.009	0.266 ±0.039*	0.548 ±0.018
	MoLFormer	0.920 ±0.030*	0.962 ±0.014	0.002 ±0.001	0.043 ±0.008*	0.274 ±0.040**	0.551 ±0.022
Phenotype	CellCLIP	0.920 ±0.028	0.962 ±0.013	0.002 ±0.001	0.043 ±0.008	0.270 ±0.059***	0.549 ±0.022
Target-aware	UniPert	0.922 ±0.032***	0.964 ±0.014***	0.002 ±0.001	0.042 ±0.009**	0.268 ±0.038*	0.559 ±0.019**
	Target ID (binding prob.)	0.926 ±0.027*	0.965 ±0.012**	0.002 ±0.001	0.042 ±0.008**	0.289 ±0.051***	0.557 ±0.015***
	Target ID (avg.)	0.926 ±0.029**	0.966 ±0.013***	0.002 ±0.001	0.041 ±0.008***	0.290 ±0.051***	0.556 ±0.015***
	GenePT-ada (avg.)	0.930 ±0.028**	0.968 ±0.012**	0.002 ±0.001	**0.040** ±0.008***	0.296 ±0.042***	0.551 ±0.025
	GenePT-protein (avg.)	0.929 ±0.027**	0.968 ±0.012**	0.002 ±0.001	**0.040** ±0.008***	**0.302** ±0.049***	0.549 ±0.030
	Qwen3 (avg.)	**0.931** ±0.028**	**0.969** ±0.011**	0.002 ±0.001	**0.040** ±0.008**	0.299 ±0.048***	0.550 ±0.023

C TAHOE RESULTS

Here we also show the performance of other representative features on the plate 6, 9, 12 of the Tahoe-100M dataset.

D IMPLEMENTATION AND TRAINING DETAILS

Our predictor is a two-block MLP: the first block projects the frozen perturbation embedding $\phi(p)$ to a hidden representation h, a FiLM layer conditions h on the cell line, the second block processes

Table 11: **DEG50 panel, all doses** Test performance as mean ± std across cell lines and seeds. **Bold** marks the best mean per column; ties are all marked. Stars mark our models that are significantly better than the best learning-free baseline of that column ($* \ p < 0.05$, $** \ p < 0.01$, $*** \ p < 0.001$; paired t-test)

Model / representation		$R^2 \uparrow$	Pearson ↑	MSE ↓	RMSE ↓	Δ-Pearson ↑	Δ-Direction ↑
Learning-free baselines							
Cell-line control mean		0.867 ±0.051	0.980 ±0.007	0.025 ±0.014	0.153 ±0.042	−0.042 ±0.062	0.000 ±0.000
Cell-line perturbation mean		0.873 ±0.047	0.980 ±0.006	0.023 ±0.013	0.147 ±0.040	0.203 ±0.056	0.590 ±0.035
Global control mean		0.395 ±0.178	0.793 ±0.087	0.071 ±0.027	0.263 ±0.048	−0.112 ±0.037	0.404 ±0.044
Global perturbation mean		0.392 ±0.181	0.789 ±0.087	0.070 ±0.026	0.261 ±0.047	−0.076 ±0.030	0.427 ±0.035
Single-cell models (trained on single cells)							
ChemCPA Hetzel et al. (2022)		0.645 ±0.077	0.936 ±0.013	0.050 ±0.014	0.181 ±0.020	0.236 ±0.033	0.426 ±0.019
Ours: pseudobulk MLP on a frozen perturbation embedding							
Molecular structure	RDKit-2D	0.883 ±0.045***	0.980 ±0.006**	0.021 ±0.013***	0.141 ±0.040***	0.245 ±0.063***	0.609 ±0.031***
	Morgan FP	0.883 ±0.044***	0.980 ±0.006*	0.021 ±0.013***	0.140 ±0.040***	0.255 ±0.065***	0.615 ±0.033***
	Uni-Mol	0.883 ±0.044***	0.981 ±0.006*	0.021 ±0.012***	0.139 ±0.039***	0.234 ±0.065***	0.607 ±0.031***
Chemical LLM	ChemBERTa	0.877 ±0.048***	0.980 ±0.007	0.022 ±0.014***	0.144 ±0.042**	0.213 ±0.051	0.605 ±0.030***
	MoLFormer	0.877 ±0.045***	0.980 ±0.006	0.022 ±0.013***	0.144 ±0.039***	0.226 ±0.058***	0.606 ±0.029***
Phenotype	CellCLIP	0.880 ±0.043***	0.980 ±0.006	0.022 ±0.013***	0.144 ±0.039***	0.241 ±0.057***	0.611 ±0.029***
Target-aware	UniPert	0.883 ±0.046***	0.980 ±0.007**	0.021 ±0.013***	0.140 ±0.041***	0.236 ±0.068***	0.605 ±0.027**
	Target ID (binding prob.)	0.887 ±0.040***	0.981 ±0.006*	0.020 ±0.011***	0.137 ±0.037***	0.255 ±0.088***	0.613 ±0.033***
	Target ID (avg.)	0.889 ±0.041***	0.981 ±0.005	0.020 ±0.012***	0.135 ±0.039***	0.263 ±0.085***	0.616 ±0.031***
	GenePT-ada (avg.)	0.896 ±0.039***	**0.982** ±0.006***	**0.018** ±0.011***	**0.129** ±0.037***	0.266 ±0.073***	0.622 ±0.029***
	GenePT-protein (avg.)	0.896 ±0.040***	**0.982** ±0.006***	**0.018** ±0.011***	0.130 ±0.037***	**0.270** ±0.078***	0.622 ±0.031***
	Qwen3 (avg.)	**0.898** ±0.038***	**0.982** ±0.006***	**0.018** ±0.011***	**0.129** ±0.036***	**0.270** ±0.080***	**0.623** ±0.032***

Table 12: **Cross-plate, DEG50 panel** Test performance on the DEG50 gene panel for plates 3, 9 and 12 of the Tahoe-100M dataset (mean ± std across cell lines and seeds) where the number of overlapping drugs across plates is less than 3. **Bold** marks the best mean per column within each plate, over all rows of that plate. We here show the cell-line specific perturbation mean and pseudobulk MLP trained on Target ID performance.

Model / representation	$R^2 \uparrow$	Pearson ↑	RMSE ↓	$\Delta R^2 \uparrow$	Δ-Pearson ↑	Δ-Direction ↑
Plate 3						
Cell-line control mean	0.172 ±0.559	0.716 ±0.188	0.684 ±0.127	−0.296 ±0.256	−0.004 ±0.106	0.000 ±0.000
Cell-line perturbation mean	0.454 ±0.385	0.814 ±0.130	0.511 ±0.119	0.297 ±0.244	**0.760** ±0.041	**0.879** ±0.025
RDKit-2D	0.436 ±0.408	0.814 ±0.128	0.519 ±0.123	0.247 ±0.277	0.715 ±0.071	0.851 ±0.039
UniPert	0.462 ±0.376	0.815 ±0.129	0.507 ±0.118	0.305 ±0.236	0.743 ±0.057	0.869 ±0.034
Target ID (avg.)	0.590 ±0.213	0.817 ±0.120	0.493 ±0.065	0.384 ±0.065	0.726 ±0.036	0.857 ±0.023
Target ID (binding prob.)	**0.601** ±0.207	**0.823** ±0.118	**0.489** ±0.062	**0.385** ±0.072	0.746 ±0.038	0.870 ±0.023
Plate 9						
Cell-line control mean	0.522 ±0.533	0.832 ±0.162	0.504 ±0.129	−0.249 ±0.196	0.001 ±0.115	0.000 ±0.000
Cell-line perturbation mean	0.730 ±0.308	0.896 ±0.133	0.321 ±0.124	0.548 ±0.125	0.817 ±0.067	0.904 ±0.038
RDKit-2D	0.727 ±0.313	0.896 ±0.133	0.320 ±0.125	0.549 ±0.128	0.812 ±0.073	0.898 ±0.041
UniPert	0.728 ±0.307	0.895 ±0.132	0.324 ±0.123	0.529 ±0.147	0.800 ±0.080	0.890 ±0.046
Target ID (avg.)	**0.782** ±0.275	**0.912** ±0.126	**0.295** ±0.110	**0.581** ±0.136	**0.822** ±0.084	**0.908** ±0.045
Target ID (binding prob.)	**0.782** ±0.274	**0.912** ±0.126	**0.295** ±0.111	0.580 ±0.139	0.821 ±0.083	**0.908** ±0.045
Plate 12						
Cell-line control mean	0.574 ±0.449	0.830 ±0.215	0.474 ±0.127	−0.395 ±0.299	0.005 ±0.080	0.000 ±0.000
Cell-line perturbation mean	0.724 ±0.202	0.887 ±0.112	0.370 ±0.096	0.144 ±0.280	0.706 ±0.076	0.847 ±0.053
RDKit-2D	0.715 ±0.206	0.884 ±0.112	0.384 ±0.099	0.050 ±0.315	0.636 ±0.119	0.811 ±0.059
UniPert	0.736 ±0.192	0.890 ±0.109	0.370 ±0.094	0.135 ±0.263	0.670 ±0.098	0.829 ±0.052
Target ID (avg.)	**0.738** ±0.197	**0.891** ±0.110	**0.365** ±0.093	**0.166** ±0.261	**0.713** ±0.074	**0.853** ±0.048
Target ID (binding prob.)	0.731 ±0.193	0.888 ±0.109	0.369 ±0.088	0.145 ±0.274	0.683 ±0.098	0.829 ±0.055

the conditioned representation, and a linear head maps it to the pseudobulk delta. All perturbation encoders are frozen; only the MLP, the FiLM heads and the output layer are trained.

The four single-cell models ChemCPA Hetzel et al. (2022), BioLord Piran et al. (2024), Doloris Chi et al. (2026) and PerturbDiff Yuan et al. (2026) are trained following their official GitHub repository with the default hyperparameter on the same data split as the pseudobulk models.

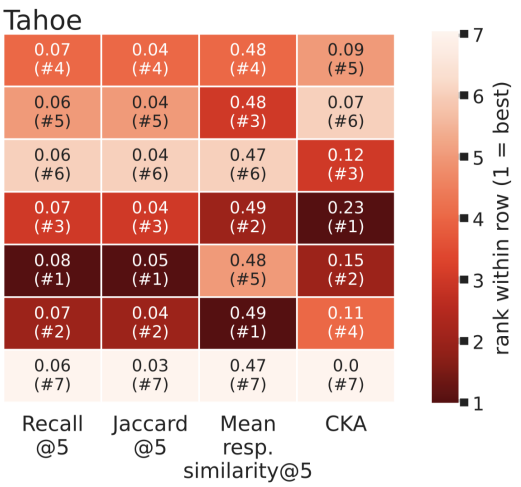


Figure 4: Perturbation representations differ in how they connect drugs to observed transcriptional responses. local (Recall@5, Jaccard@5, Mean Response Similarity@5) and global (CKA) alignment between feature-space and residual-response geometry on Tahoe-100M; numbers are metric values and within-column ranks. From top to bottom the corresponding features are: RDkit, ChemBERTa, CellCLIP(morph), UniPert, Target ID, GenePT, and Random retrieval. We observe the CellCLIP (morph) features rank the number 6th in terms of local alignment but number 3rd in terms of global alignment.However, the features are less distinguishable compared to the sci-Plex3 dataset.

E TARGET PREPARATION

To systematically identify drug-associated genes and categorize them according to the strength and nature of supporting evidence, we employ GPT5.4 Nguyen et al. (2026). The prompt instructed the LLM to analyze a single drug at a time using the provided drug name and SMILES representation, retrieve evidence from PubMed and CCCPapers, and classify genes into three evidence tiers: (i) high-confidence direct targets, (ii) complex dependency genes, and (iii) experimentally reported off-targets. The complete prompt is reproduced below.

We first show the distribution of the number of retrieved targets per drug in Fig. 5. We observe that the majority of drugs have approximately 10–20 associated targets. Additionally, we validate the accuracy of the retrieved targets by calculating the coverage between the retrieved targets and the target annotations in the Tahoe-100M dataset. Among the 379 drugs, 264 have at least one annotated target (51% of the drugs have one annotated target, and 36% have 2–3 annotated targets). We observe that, among these 264 drugs, 219 have at least one annotated target included in the retrieved target list, corresponding to a coverage of 83%.

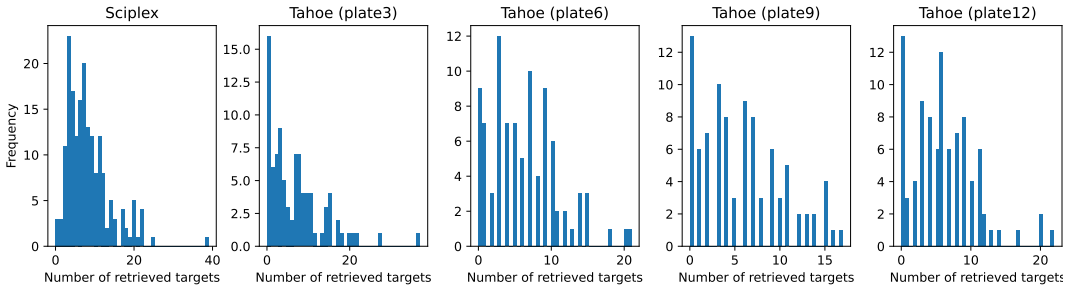


Figure 5: Distribution of the number of associated targets per drug across datasets and plates.

A natural explanation for the strength of target-based representations is that the predictor exploits a direct correspondence between a drug's annotated targets and the genes it is asked to predict. We test this by stratifying held-out sci-Plex3 drugs by the number of their annotated targets that fall

within the evaluated gene panel (Fig. 6). If performance relied on this correspondence, scores should increase with overlap. They do not: drugs whose targets are entirely absent from the panel are predicted comparably to those with several targets in it, and neither Δ-Pearson nor Δ-Direction shows a monotone trend. The modestly higher scores at one or two overlapping targets do not persist at higher counts. This is consistent with the mechanism of action being a protein-level event whose transcriptional consequences are propagated downstream, rather than a change in the target's own transcript. Together with the finding that target-based representations benefit most when test drugs share targets with training compounds, this suggests the representation helps by placing an unseen drug near mechanistically related training perturbations, not by pointing the predictor at specific output genes.

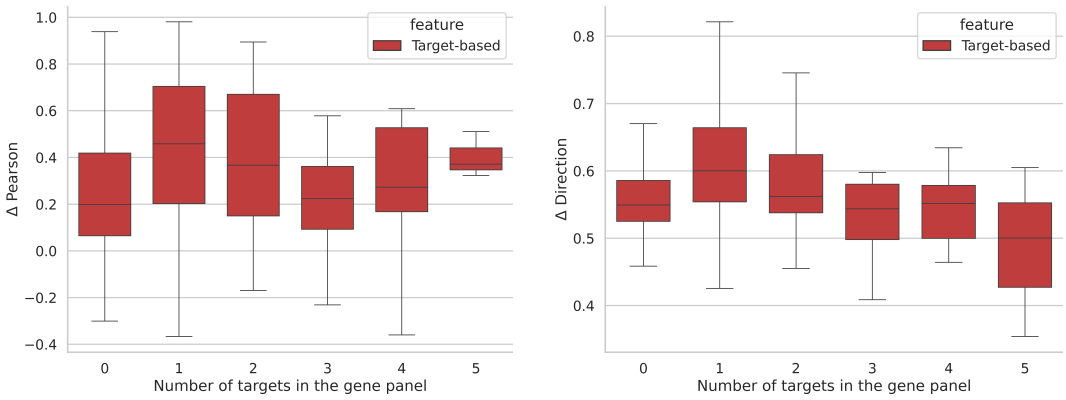


Figure 6: Performance of the target-based representation on held-out sci-Plex3 drugs, stratified by the number of annotated drug targets contained in the evaluated full panel. Left: Δ-Direction; right: Δ-Pearson. Numbers of unique drugs per bin: 46/21/15/8/5/1. Neither metric increases with target overlap.

Target set. For each drug d, we pool the target genes identified across all three evidence tiers into a single target set $\mathcal{T}_d = \{g_1, g_2, \ldots, g_{n_d}\}$, where $n_d = |\mathcal{T}_d|$ is the number of targets associated with drug d. Throughout, we refer to targets by their official gene symbols; each symbol denotes the corresponding protein product, which is the physical entity engaged by the drug. Embeddings are defined at the gene level (each gene $g \in \mathcal{T}_d$ is represented by a k-dimensional embedding $\mathbf{e}_g \in \mathbb{R}^k$), whereas binding scores are computed at the protein level. Note that the embedding dimension k depends on the choice of gene embedding. The drug representation is then obtained by aggregating the embeddings of its targets.

Target embeddings. We consider two choices of embedding $\mathbf{e}_g$:

- **Target-identity embeddings (*Target ID*).** Each gene is assigned a fixed vector drawn independently from a standard Gaussian, $\mathbf{e}_g \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_k)$, and left untrained (i.e., a randomly initialized `nn.Embedding` with no gradient updates). In high dimension such vectors are approximately orthogonal, so this encoding treats genes as distinct, mutually independent identities and carries no biological similarity structure, serving as a one-hot–style baseline; we call it *Target ID* in the tables to avoid suggesting a literal indicator vector.
- **Embeddings that carry biological information** Each gene is represented by its pretrained feature embedding, including GenePT-ada, GenePT-protein, and GenePT-ada-Qwen3, which encode functional and biological similarity between genes, so that related targets occupy nearby regions of the embedding space.

Binding-probability scoring. For every (d, g) pair we compute a predicted binding probability $w_{d,g} \in [0, 1]$ between drug d and the protein encoded by gene g, using the PSICHIC drug–target affinity predictor Koh et al. (2024), which scores interactions from the drug structure and the target protein sequence. These probabilities provide a per-target relevance weight for the drug. Note, more advanced binding models including Boltz2 Passaro et al. (2025) and NESSO-1 Shenoy et al. (2026) can potentially give more accurate binding predictions, but we leave that as future work.

Aggregation. We form the drug embedding $\mathbf{z}_d$ by pooling the target embeddings in one of two ways:

$$(\textit{avg.}) \qquad \mathbf{z}_d = \frac{1}{n_d} \sum_{g \in \mathcal{T}_d} \mathbf{e}_g, \tag{3}$$

$$(\textit{binding prob.}) \qquad \mathbf{z}_d = \frac{\sum_{g \in \mathcal{T}_d} w_{d,g}\, \mathbf{e}_g}{\sum_{g \in \mathcal{T}_d} w_{d,g}}. \tag{4}$$

The unweighted mean (*avg.*) treats all targets as equally informative, whereas the binding-probability–weighted mean (*binding prob.*) emphasizes targets that PSICHIC predicts to bind the drug more strongly, down-weighting weakly engaged targets.

```
You are a biomedical drug-target extraction assistant specialized in
    evidence-based analysis of a single drug at a time.

Your task is to analyze exactly one drug provided by the user, using:
- drug name {drug_name}
- SMILES string {smiles}

and return a structured evidence-based classification of target-related
    genes for that drug.

## Core objective

For the given drug, identify genes and separate them into exactly these
    four tiers:

### Tier 1 -- High-confidence direct targets
Include only genes supported by strong direct evidence such as:
- biochemical binding, inhibition, or affinity assays
- structural biology evidence such as co-crystal or direct binding
    structures
- validated resistance mutations in the target gene
- authoritative mechanism-of-action descriptions
- direct target-identification studies

### Tier 2 -- Complex dependency genes
Include genes that are not necessarily direct binding targets but are
    strongly implicated in drug sensitivity, resistance, or function
    because they are part of the same target complex or are required for
    drug response, such as:
- target-complex subunits
- proteasome-complex dependency genes
- genes supported by strong CRISPR, knockout, knockdown, rescue, deletion,
     mutation, or overexpression evidence
- genes whose perturbation consistently alters sensitivity or resistance

### Tier 3 -- Experimentally reported off-targets
Include genes with experimental evidence of non-primary target engagement,
     such as:
- direct off-target inhibition in biochemical assays
- experimentally demonstrated secondary target activity in cellular
    assays

Do not include speculative, docking-only, computational-only, or
    similarity-only off-targets unless clearly labeled as weak evidence.

## Mandatory methodology

For this drug:
- You must search both PubMed and CCCPapers.
- PubMed and CCCPapers are required sources.
- You may use additional sources only if useful, but all nontrivial
    claims must be grounded in retrieved evidence.
```

```
- Use only evidence retrieved from tools or retrieved source content.
- Do not rely on unsupported prior knowledge.
- Prefer official human gene symbols where possible.
- Use the SMILES string only as contextual information.
- Do not infer targets solely from chemical class assumptions.
- If the drug name and SMILES conflict, explicitly note the inconsistency.


## Extraction principles

- Separate evidence classes strictly.
- Do not elevate downstream biomarkers to direct targets.
- Do not call a gene a direct target without direct evidence.
- If evidence is weak, indirect, conflicting, inferred from class analogy,
    or computational-only, label it clearly.
- If no strong target evidence is found, explicitly say so.
- If the drug acts on a protein complex, include the directly bound or
   primary mechanistic subunit(s) in Tier 1 only if direct evidence
   exists.
- Place related complex members in Tier 2 only if supported by genetic or
    functional dependency evidence.
- Be concise but evidence-dense.
- Support every nontrivial assertion with citations from retrieved
   sources.

## Internal reasoning guide

Internally determine:
1. What is the strongest direct evidence for a target?
2. Are there genetically or functionally supported complex/dependency
   genes?
3. Are there experimentally measured off-targets?
4. Which genes are only downstream effectors and must remain separate?

## Output constraints

- If no evidence is found for a tier, return an empty list for that tier.
- Do not hallucinate gene names.
- Do not invent citations.
- Do not add external links unless they were retrieved from tools.
- Preserve reference formatting from retrieved tools as faithfully as
   possible.
- If PubMed IDs are available, include them in citation or reference
   strings as: PubMed: XXXXXXXX.
- Be conservative and explicit about uncertainty.
```