ONE DRUG, MORE VIEWS: RESPONSE-DIRECTED MULTI-MODAL CONDITIONING FOR UNSEEN-DRUG PERTURBATION

Anonymous Authors

ABSTRACT

Predicting transcriptional responses to unseen drugs is central to prioritising compounds for experimental profiling, yet existing predictors generalise poorly to held-out drugs. We attribute this to two coupled limitations: drugs are represented by chemical structure alone, which only partly determines their mechanism, and models are trained to reconstruct whole expression profiles, an objective dominated by background variation rather than the perturbation effect. We introduce MoRe, a model-agnostic plug-in that addresses both. A lightweight module fuses chemical and target-informed representations into a single drug latent, and an auxiliary objective supervises the predicted mean response of each condition. Both components are trained jointly with the backbone. On an autoencoder (ChemCPA) and a diffusion model (PerturbDiff), evaluated on held-out drugs in sci-Plex3 and Tahoe-100M, MoRe improves the correlation between predicted and observed responses on the top 50 differentially expressed genes by up to 113%. Analysis shows that the drug views are complementary: they lead on different gene programmes, and fusion places each drug near training drugs with similar responses, whereas chemistry alone does not. By describing drugs by what they do as well as what they are, MoRe points to a promising direction for virtual-cell models of *in silico* drug perturbation.

1 INTRODUCTION

Predicting how cells respond transcriptionally to unseen chemical perturbations is central to drug discovery and a core capability of virtual-cell models (Bunne et al., 2024). Single-cell perturbation screens measure drug-induced expression changes at scale (Subramanian et al., 2017; Zhang et al., 2025), yet even the largest profile only thousands of drug and cell-line combinations, a small fraction of the combinatorial space of compounds, doses and cellular contexts. Models that generalise to untested compounds would allow this space to be explored computationally, prioritising experiments and extending perturbation atlases beyond what can be measured directly.

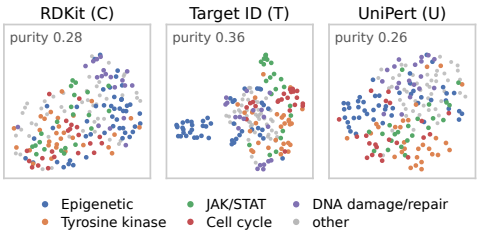

Figure 1: **Different drug embeddings induce distinct geometries over the same compounds.** UMAP of frozen drug embeddings for sci-Plex3 compounds, coloured by annotated pathway. *Purity*: fraction of a compound's ten nearest neighbours sharing its pathway label (chance: 0.13).

Generalising to unseen drugs requires mechanistic information that chemical structure provides only implicitly. A compound's transcriptional effect is mediated by its targets: it binds one or more protein targets, often with substantial polypharmacology (Zhang et al., 2016); and target engagement perturbs downstream signalling and regulatory programmes, which ultimately shift the cell's transcriptional state and phenotype (Replogle et al., 2022). However, existing predictors typically represent compounds by chemical structure alone (Hetzel et al., 2022; Yuan et al., 2026), leaving this mechanism to be inferred from a few hundred training compounds. Such inference is unreliable, because structurally dissimilar compounds can share targets (Wang et al., 2016) and close analogues can differ sharply in activity (Martin et al., 2002). Target annotations and target-informed embeddings such as UniPert, pretrained across molecular perturbagens (Li et al., 2026b), encode mechanistic information directly. These representations induce notably different geometries over the same set of compounds, each capturing a distinct view of how they relate (Figure 1). This

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.

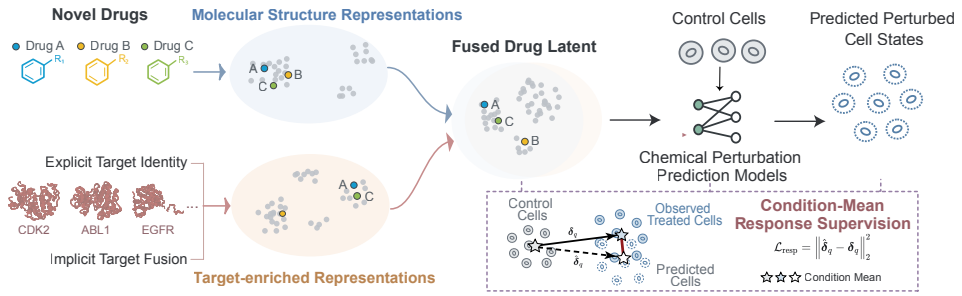

Figure 2: **Overview of MoRe.** Chemical structure and target-informed representations are fused into a single drug latent that conditions an existing perturbation predictor. During training, condition-mean response supervision aligns the predicted ($\hat{\boldsymbol{\delta}}_q$) and observed ($\boldsymbol{\delta}_q$) average responses relative to the control cells given condition q. Coupled with the backbone's own loss, the fusion module and backbone is trained end-to-end.

suggests that they complement chemical structure and that combining them offers a richer basis for generalising to unseen compounds.

Richer drug representations help only if the training signal rewards using them, which points to a second limitation. Expression profiles are dominated by variation shared across conditions, such as cell-state structure and cell-to-cell heterogeneity, whereas perturbation-induced changes are concentrated in a comparatively small set of genes. An objective defined over the whole profile therefore underweights these changes, an effect termed *signal dilution* (Mejia et al., 2026). Consistent with this, simple baselines that reproduce the average perturbed profile often match or outperform deep models (Ahlmann-Eltze et al., 2025; Vinas Torne et al., 2025), indicate that standard objectives and metrics largely reward this systematic background, which a model can reproduce while making little use of the drug. **The two limitations are therefore coupled: representation bounds what the objective can extract, and the objective bounds what the representation is used for.** This motivates the question we study: *how can complementary drug representations be exploited to predict responses to unseen drugs?*

In this work, we introduce MoRe, **a model-agnostic plug-in that lifts both limitations at once**: multimodal drug fusion widens what the objective can extract, and condition-mean response guidance directs how the representation is used (Figure 2). On the representation side, lightweight adapters project chemical structural descriptors and target-enriched representations into a shared space and fuse them into a single drug latent, letting the model relate an unseen compound to training drugs through shared mechanism rather than chemical similarity alone. On the objective side, we add an auxiliary loss that averages the predictor's outputs over the cells of each condition and matches this mean to the observed mean response, weighting genes so that the few differentially expressed genes (DEGs) are not outweighed by the many that barely respond. This mirrors how mechanism of action (MoA) is read from condition-mean response signatures rather than from individual cells (Lamb et al., 2006; Subramanian et al., 2017), and provides a more stable training signal concentrated on the perturbation effect itself. Both components are trained jointly with the backbone, and MoRe therefore applies to existing predictors regardless of their architectural family.

We demonstrate the effectiveness of MoRe on two backbones from different families, autoencoder-based ChemCPA (Hetzel et al., 2022) and diffusion-based PerturbDiff (Yuan et al., 2026), and evaluate on held-out drugs in sci-Plex3 (Srivatsan et al., 2020) and Tahoe-100M (Zhang et al., 2025). **MoRe improves every residual metric on every backbone and dataset**, raising the Pearson correlation between predicted and observed drug-induced effect (residual) on the top 50 differentially expressed genes (DEGs) by **up to 113**% (Section 5.1). The two components are sub-additive. On PerturbDiff, fusion alone improves this metric by 27% and condition-mean response supervision alone by 14%, while together they improve it by 29% (Section 5.2). This suggests that both components strengthen the drug-specific signal, fusion in the input and response supervision in the loss.

Fusion enriches drug views with complementary biology. No single view leads on every gene programme: target-informed views lead on programmes driven by specific proteins, such as cell-cycle and oestrogen signalling, while chemistry leads on others, and fusion outperforms the average single view in 20 of 22 programmes. and its margin grows with response strength (Section 5.3). The fused representation places held-out drugs near training drugs with similar responses (ρ=0.44,

versus -0.08 for chemistry alone), and these better-matched neighbours support more accurate predictions (Section 5.4). More broadly, our results suggest that describing drugs by what they do, not only by what they are, is a practical route to predicting responses to compounds that have never been profiled. We summarize our contribution as below:

- **A model-agnostic plug-in.** Motivated by two limitations of existing predictors, single-view drug representations and objectives diluted by variation unrelated to the drug, we introduce MoRe, which fuses chemical and target-informed views and complements the backbone's original loss with an auxiliary loss on average responses, without changing its architecture.
- **Consistent gains across model families.** On an autoencoder (ChemCPA) and a diffusion model (PerturbDiff), evaluated on held-out drugs in sci-Plex3 and Tahoe-100M, MoRe improves every residual metric, by up to 113% in DEG correlation.
- **Insight into generalization ability of MoRe.** We show that chemical and target-informed views capture different parts of the transcriptional perturbation response, and fusing them organises drugs by response, relating an unseen compound to seen drugs that *act* like it, not only to those that look like it.

2 Related work

Perturbation response models. Models of cellular perturbations differ in how they represent and generate changes in cell state. scGen transfers effects through a learned latent space (Lotfollahi et al., 2019); CPA and ChemCPA compose perturbation and covariate embeddings in an autoencoder (Lotfollahi et al., 2023; Hetzel et al., 2022); biolord learns disentangled representations (Piran et al., 2024); and PerturbDiff models responses with conditional diffusion (Yuan et al., 2026). Transport methods such as CellOT and CondOT instead learn maps between unpaired control and treated populations (Bunne et al., 2023; 2022). Single-cell foundation models such as Geneformer, scGPT and scFoundation are pretrained on large expression corpora and fine-tuned for perturbation prediction (Theodoris et al., 2023; Cui et al., 2024; Hao et al., 2024), and State scales set-level prediction across contexts using large drug screens (Adduri et al., 2026). Despite their different generative mechanisms, these models share two design choices: a drug is described by a single view, either a learned embedding or chemical structure, and training is driven by a per-cell or per-distribution loss on the whole expression profile. MoRe changes only these two components and leaves the generative mechanism untouched, so it applies across these families.

Drug representations for response prediction. Molecular fingerprints, descriptors and pretrained encoders capture different aspects of chemical structure (Rogers & Hahn, 2010; Wu et al., 2018; Hu et al., 2020; Wang et al., 2022). Drug–target models add biological information through protein interactions (Singh et al., 2023; Huang et al., 2020), and UniPert learns a shared representation of genetic and chemical perturbagens (Li et al., 2026b). Multimodal learning combines such heterogeneous descriptions of an entity (Baltrušaitis et al., 2019), and has been applied to drug response for viability and sensitivity prediction (Yu et al., 2025; Zhang et al., 2026). For transcriptomic response prediction, however, existing predictors condition on one view at a time. MoRe fuses different views into a single drug latent that is trained end to end through a given predictor.

Training objectives and evaluation. Perturbation effects are concentrated in a small subset of genes, so objectives over the whole profile underweight them, an effect termed signal dilution (Mejia et al., 2026). As a result, baselines that predict the average perturbed profile often match or outperform complicated models (Ahlmann-Eltze et al., 2025; Vinas Torne et al., 2025), and benchmarks now separate prediction fit from perturbation specificity (Wu et al., 2024; Vinas Torne et al., 2025). Existing approaches reconstruct full gene expression of individual cells using MSE loss (Hetzel et al., 2022), increase the weight of responsive genes Mejia et al. (2026); Chi et al. (2026), or match cell distributions using distributional objectives using maximum mean discrepancy, energy or optimal transport (Yuan et al., 2026; Adduri et al., 2026; Bunne et al., 2023). The former remain sensitive to single-cell level noise, artifact and shared background expression, while the latter can place greater emphasis on within-condition heterogeneity than on the systematic response, particularly in chemical screens where many compounds induce only modest shifts. MoRe instead matches the control-subtracted condition mean on the full gene panel weighted by DEGs that are calculated from the training data, targeting the systematic perturbation effect directly, as an auxiliary term alongside the backbone's native loss. Following these benchmarks, we report control-referenced

DEG metrics, a cell-line-specific perturbation-mean baseline and condition-retrieval rank Wu et al. (2024).

3 MoRe: MULTIMODAL FUSION WITH CONDITION-MEAN SUPERVISION

For an unseen drug, a predictor must infer the response from the training drugs, so it generalises best when drugs with similar representations respond similarly. Chemical structure alone does not ensure this: structurally similar drugs can respond differently, and dissimilar drugs can respond alike (Martin et al., 2002; Subramanian et al., 2017). In addition, existing predictors are trained on whole single-cell profiles, which are dominated by variation unrelated to the drug, so the drug effect is a small part of what the objective rewards. MoRe addresses both while keeping the backbone's architecture: a conditioning module fuses several views of each drug into its input (Section 3.2), and an auxiliary loss supervises the predicted cellular response in condition-mean level (Section 3.3).

3.1 PROBLEM FORMULATION

Let $q = (c, p, s)$ denote a condition comprising a cellular context c (e.g., cell line), a drug p, and a dose s. We study unseen-drug prediction in seen cell lines and doses. Given a control cell $\mathbf{x}_i^c$ from context c, a drug representation $\mathbf{z}_p$, and a dose s, the predictor F_θ outputs a predicted treated cell $F_\theta(\mathbf{x}_i^c, \mathbf{z}_p, s)$. Since a cell is measured either before or after treatment but never both, individual predictions have no ground truth and are compared with data through the population they form. We summarise this population by the predicted *condition mean* over n control cells:

$$\widehat{\boldsymbol{\mu}}_q = \frac{1}{n}\sum_{i=1}^{n} F_\theta(\mathbf{x}_i^c, \mathbf{z}_p, s), \qquad \mathbf{x}_i^c \sim \mathcal{D}_c^{\mathrm{ctrl}}. \tag{1}$$

Let $\boldsymbol{\mu}_q$ denote the condition mean of the observed treated cells of condition q, and $\boldsymbol{\mu}_c^{\mathrm{ctrl}}$ the mean of the training control cells of context c.[1] The observed and predicted responses are

$$\boldsymbol{\delta}_q = \boldsymbol{\mu}_q - \boldsymbol{\mu}_c^{\mathrm{ctrl}}, \qquad \widehat{\boldsymbol{\delta}}_q = \widehat{\boldsymbol{\mu}}_q - \boldsymbol{\mu}_c^{\mathrm{ctrl}}. \tag{2}$$

Intuitively, $\boldsymbol{\delta}_q$ is the average expression shift that drug p at dose s induces relative to unperturbed cells: subtracting the control mean removes the baseline expression shared by control cells of the according cell line, isolating the real drug effect. For a held-out drug, every input except $\mathbf{z}_p$ is observed during training so everything the model knows about the drug must come from $\mathbf{z}_p$.

3.2 MULTIMODAL DRUG-CONDITIONING MODULE

Complementary views. Two drugs can introduce similar responses for different reasons: they engage the same protein, or they share physicochemical properties that govern uptake, off-target binding and scaffold-specific effects whatever the target (Sirci et al., 2017). A single view captures only one of these, so it cannot connect a held-out drug to training drugs that respond alike for the other reason. We therefore describe each drug with three views ($M = 3$), one per account: chemical descriptors (shared structure), explicit target identity (shared targets), and a target-informed chemical embedding from UniPert (Li et al., 2026b) (contrastively aligned with the binding targets).

For each view m, a trainable adapter maps the drug embedding $\phi_m(p)$ into a shared latent space:

$$\mathbf{h}_m(p) = \mathrm{normalize}(f_m(\mathrm{LN}_m(\phi_m(p)))), \tag{3}$$

where LN_m is layer normalization, f_m is an MLP, and $\mathrm{normalize}(\mathbf{v}) = \mathbf{v}/\|\mathbf{v}\|_2$ rescales a vector to unit length. Since the views differ in dimension and scale, layer normalization standardizes the input scale, and the MLP maps them to a common dimension and unit length prevents any view from dominating the fusion merely through its magnitude.

Integration before prediction. We concatenate the adapted views $\mathbf{h}_1(p), \ldots, \mathbf{h}_M(p)$ and pass them through an MLP f_{cat} to obtain the drug latent:

$$\mathbf{z}_p = \mathrm{normalize}(f_{\mathrm{cat}}(\mathrm{LN}_{\mathrm{cat}}([\mathbf{h}_1(p); \ldots; \mathbf{h}_M(p)]))), \tag{4}$$

and $\mathbf{z}_p$ is used to represent a drug. We choose concatenation for its simplicity: it imposes no assumption on how the views should combine and leaves the MLP to learn their integration. Nevertheless, we benchmarked more advanced fusion techniques where they show no clear gains (Sec. 5.2).

[1] On Tahoe-100M, controls are additionally matched by plate during training (Appendix C).

3.3 CONDITION-MEAN RESPONSE SUPERVISION

A richer $\mathbf{z}_p$ helps only if training makes the model use it. Existing backbones are optimised on individual cells or on distributions of cells, and both kinds of loss are dominated by variation among cells of a condition, much of it unrelated to the drug. The systematic drug effect of a condition therefore contribute little to the final loss and gives $\mathbf{z}_p$ a weak training signal. To address this, we propose an auxiliary objective on condition-mean responses, in which cell-to-cell variation averages out and the average drug effect remains; heterogeneity in how cells respond is left to the backbone's native loss. Below, we omit the superscript c; controls remain matched to the context of condition q.

Per-cell reconstruction dilutes the systematic drug effect. We consider the simplest and commonly used objective: the squared error between a predicted cell and an independently sampled observed treated cell of the same condition. Denote one observed treated cell as $\mathbf{y}_j = \boldsymbol{\mu}_q + \boldsymbol{\eta}_j$, where $\boldsymbol{\eta}_j = \mathbf{y}_j - \boldsymbol{\mu}_q$ is its *deviation* from the condition mean, whether from cell-cycle phase, sequencing depth, technical noise or heterogeneous response to the drug. Similarly, we write each predicted cell $\widehat{\mathbf{y}}_i = F_\theta(\mathbf{x}_i, \mathbf{z}_p, s)$, where $\bar{\boldsymbol{\mu}}_q = \mathbb{E}[\widehat{\mathbf{y}}_i]$ is the model's *expected condition mean* over control cells $\mathbf{x}_i \sim \mathcal{D}_c^{\text{ctrl}}$ and any sampling noise of the model. Since control and treated cells are unpaired, $\mathbf{x}_i$ and $\mathbf{y}_j$ are drawn independently, the expected loss decomposes as

$$\mathbb{E}\left\|\widehat{\mathbf{y}}_i - \mathbf{y}_j\right\|^2 = \left\|\bar{\boldsymbol{\mu}}_q - \boldsymbol{\mu}_q\right\|^2 + \mathbb{E}\|\boldsymbol{\eta}_j\|^2 + \mathbb{E}\|\widehat{\mathbf{y}}_i - \bar{\boldsymbol{\mu}}_q\|^2, \tag{5}$$

Eq. (5) is a bias–variance decomposition (Appendix E). **The first term is the bias**, the systematic error in the predicted drug effect, which $\mathbf{z}_p$ should inform. **The second represents the *observed* within-condition variance**; it depends only on the data, but each sampled cell introduces stochastic gradient proportional to $\boldsymbol{\eta}_j$. However, this variance typically far exceeds the drug effect (on sci-Plex3, the median $\|\boldsymbol{\delta}_q\|^2$ is only 0.15% of $\mathbb{E}\|\boldsymbol{\eta}_j\|^2$; Appendix A). Per-cell gradients are therefore dominated by this noise, which dilutes the signal from the first term. **The third is the variance of the *predicted* cells**, which arises because different control cells, and for stochastic backbones different noise draws, are mapped to different predictions. A single-cell model should capture this spread, yet the per-cell loss penalises it in full and pushes predictions to collapse onto their mean.

Condition-mean supervision. We therefore supervise the mean-error term of Eq. (5) directly while reducing cell-sampling variability. For a training condition q, we independently draw n matched control cells, $\mathbf{x}_1, \ldots, \mathbf{x}_n \sim \mathcal{D}_c^{\text{ctrl}}$ and penalize the systematic error in the predicted response:

$$\ell_q = \left\|\widehat{\boldsymbol{\delta}}_q - \boldsymbol{\delta}_q\right\|^2 = \left\|\left(\widehat{\boldsymbol{\mu}}_q - \boldsymbol{\mu}_c^{\text{ctrl}}\right) - \left(\boldsymbol{\mu}_q - \boldsymbol{\mu}_c^{\text{ctrl}}\right)\right\|^2 = \left\|\frac{1}{n}\sum_{i=1}^{n} F_\theta(\mathbf{x}_i, \mathbf{z}_p, s) - \boldsymbol{\mu}_q\right\|^2, \tag{6}$$

with $\boldsymbol{\delta}_q$ used only as a training target. Since the control mean cancels, ℓ_q compares an average of n predictions with the mean of all treated cells, rather than one prediction with one cell. Its expectation over the n draws is shown in Eq. (7); see Appendix E for the derivation.

$$\mathbb{E}\left[\ell_q\right] = \left\|\bar{\boldsymbol{\mu}}_q - \boldsymbol{\mu}_q\right\|^2 + \frac{1}{n}\,\mathbb{E}\left\|\widehat{\mathbf{y}} - \bar{\boldsymbol{\mu}}_q\right\|^2. \tag{7}$$

Intuitively, averaging within a condition reduces cell-sampling variability, providing a more precise estimate of the mean drug-induced response for supervision.

Gene weighting and joint optimization. The same dilution arises across genes: a drug changes only a few genes substantially, so an unweighted error over the full panel is dominated by the many genes that barely respond. Following Mejia et al. (2026), we therefore weight the error of Eq. (6) gene by gene, ℓ_q^W (See Appendix D for a two-set variant of their gene reweighting). Eq. (7) holds unchanged under this weighting (Appendix E). Averaging over a batch $\mathcal{B}$ of training conditions gives $\mathcal{L}_{\text{aux}} = |\mathcal{B}|^{-1}\sum_{q\in\mathcal{B}} \ell_q^W$, which is added to the backbone's native objective,

$$\mathcal{L}_{\text{train}} = \mathcal{L}_{\text{native}} + \lambda_{\text{aux}}\mathcal{L}_{\text{aux}}, \tag{8}$$

Its gradients update both the conditioning module and the backbone. Targets, DEG sets for reweighting and scales are computed from training conditions only, and inference is unchanged: the backbone takes its standard inputs, with no response target or additional decoder.

4 EXPERIMENTAL SET-UP

Datasets. We test generalization to unseen drugs on two datasets. On sci-Plex3 (Srivatsan et al., 2020), a fixed drug-level split holds out 30 of 187 drugs, which gives 360 held-out conditions (drug

Table 1: **MoRe improves every backbone on unseen-drug prediction in sci-Plex3 and Tahoe-100M.** Each backbone is shown in its original form (RDKit descriptors, native objective) and with MoRe added (shaded); *Rel. gain* is the relative improvement of MoRe over the same backbone (for rank, the relative reduction). **Bold** represents best mean per column. Means over three seeds with sample standard deviation as a subscript. Reference methods are shown for context. The perturbation mean assigns one response to every held-out drug, so its rank is 1 by construction.

	sci-Plex3					*Tahoe-100M*				
	DEG50		DEG20		Rank	DEG50		DEG20		Rank
Method	Pearson ↑	Dir. ↑	Pearson ↑	Dir. ↑	↓	Pearson ↑	Dir. ↑	Pearson ↑	Dir. ↑	↓
Reference methods										
Perturbation mean	0.26	**0.63**	0.27	0.63	1.00	0.54	0.73	0.57	0.75	1.00
CondOT	$0.12_{\pm 0.081}$	$0.61_{\pm 0.008}$	$0.12_{\pm 0.094}$	$0.61_{\pm 0.009}$	$0.45_{\pm 0.006}$	$0.61_{\pm 0.001}$	$0.73_{\pm 0.005}$	$0.63_{\pm 0.002}$	$0.75_{\pm 0.004}$	$0.18_{\pm 0.009}$
Biolord	$0.13_{\pm 0.003}$	$0.52_{\pm 0.011}$	$0.15_{\pm 0.004}$	$0.52_{\pm 0.013}$	$0.49_{\pm 0.005}$	$0.26_{\pm 0.001}$	$0.62_{\pm 0.000}$	$0.28_{\pm 0.002}$	$0.63_{\pm 0.000}$	$0.61_{\pm 0.007}$
Autoencoder backbone										
ChemCPA	$0.12_{\pm 0.024}$	$0.48_{\pm 0.016}$	$0.14_{\pm 0.019}$	$0.49_{\pm 0.014}$	$0.50_{\pm 0.017}$	$0.42_{\pm 0.026}$	$0.69_{\pm 0.008}$	$0.45_{\pm 0.028}$	$0.70_{\pm 0.009}$	$0.40_{\pm 0.009}$
+ MoRe (ours)	$0.25_{\pm 0.029}$	$0.56_{\pm 0.046}$	$0.27_{\pm 0.024}$	$0.58_{\pm 0.046}$	$0.49_{\pm 0.005}$	$0.51_{\pm 0.040}$	$0.71_{\pm 0.004}$	$0.54_{\pm 0.039}$	$0.74_{\pm 0.011}$	$0.32_{\pm 0.035}$
Rel. gain	↑113%	↑16%	↑89%	↑20%	↑2%	↑23%	↑3%	↑21%	↑5%	↑21%
Diffusion backbone										
PerturbDiff	$0.22_{\pm 0.015}$	$0.60_{\pm 0.008}$	$0.25_{\pm 0.010}$	$0.62_{\pm 0.008}$	$0.42_{\pm 0.002}$	$0.68_{\pm 0.003}$	$0.73_{\pm 0.002}$	$0.68_{\pm 0.005}$	$0.76_{\pm 0.002}$	$0.16_{\pm 0.002}$
+ MoRe (ours)	$\mathbf{0.29}_{\pm 0.015}$	$\mathbf{0.63}_{\pm 0.006}$	$\mathbf{0.33}_{\pm 0.017}$	$\mathbf{0.66}_{\pm 0.002}$	$\mathbf{0.37}_{\pm 0.022}$	$\mathbf{0.75}_{\pm 0.004}$	$\mathbf{0.76}_{\pm 0.002}$	$\mathbf{0.76}_{\pm 0.005}$	$\mathbf{0.80}_{\pm 0.004}$	$\mathbf{0.14}_{\pm 0.004}$
Rel. gain	↑29%	↑5%	↑31%	↑7%	↑11%	↑10%	↑4%	↑11%	↑5%	↑12%

× cell line × dose) across three cell lines and four doses. On Tahoe-100M (Zhang et al., 2025), we use three cell lines (A549, RKO and Hs766T) and an 80/20 drug split that holds out 76 of 379 drugs.

Drug representations. We represent each drug with three views. (C) *Chemical structure*: RDKit2D descriptors, following Hetzel et al. (2022). (T) *Target identity*: we represent each drug by its associated protein targets (Li et al., 2026a) where the targets span direct targets, complex dependencies and reported off-targets (Appendix B). Each target is assigned a fixed random 512-dimensional vector, and the drug is represented by the mean of its target vectors. This view encodes which targets drugs share, but carries no learned information about the proteins themselves. (U) *UniPert* (Li et al., 2026b): a 256-dimensional, target-informed chemical embedding.

Models and baselines. We compare against ChemCPA (Hetzel et al., 2022), PerturbDiff (Yuan et al., 2026), CondOT (Bunne et al., 2022) and Biolord (Piran et al., 2024). We apply MoRe to ChemCPA and PerturbDiff, which are representative autoencoder-based and diffusion-based approaches. We benchmark against each backbone in its original form (RDKit descriptors and its native objective) against MoRe + backbone, which uses three-view conditioning and condition-mean response supervision. For each backbone, we choose the training budget on the validation set and report the mean and standard deviation over 3 seeds on the same split. Appendix C gives the preprocessing, training budgets, selection protocol and operator ablations.

Metrics. To isolate each perturbation's effect from baseline expression, we compute all metrics on control-referenced residuals (perturbed minus control mean expression). For each condition, we report the Pearson correlation and sign agreement between predicted and observed residuals over its top 20 and top 50 DEGs (Appendix A), averaging over conditions with equal weight. To test condition specificity, we also report the rank metric of PerturBench (Wu et al., 2024) (from 0 to 1, lower is better), which measures how closely a prediction matches its own observed response compared with the other held-out conditions in the same cell line. As a reference, a perturbation-mean baseline predicts the mean expression of all perturbed training cells from the same cell line.

5 Experiments

5.1 MoRe Consistently Improves Diverse Backbones

The perturbation mean, which assigns every held-out drug the average response of its cell line, is a strong reference: on sci-Plex3 no baseline exceeds it on any residual endpoint, and on Tahoe-100M only PerturbDiff consistently matches or exceeds it. **MoRe improves every endpoint for every backbone on both datasets**, raising DEG50 correlation by 113% (ChemCPA) and 29% (PerturbDiff) on sci-Plex3, and by 23% and 10% on Tahoe-100M. MoRe + PerturbDiff is the best method on

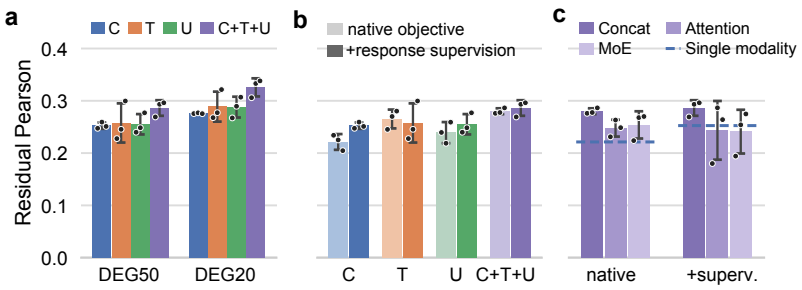

Figure 3: **Ablations of MoRe** on PerturbDiff and sci-Plex3. Bars show means over three seeds, error bars the sample SD. **(a)** Fusion improves on every single view. **(b)** Response supervision improves most single views but adds little to fusion **(c)** Concatenation matches or exceeds attention and gating, with lower variance across seeds; Dashed lines: RDKit baseline

every metric and the only one to exceed the perturbation mean on both datasets; the single exception is DEG50 direction on sci-Plex3, where the two tie.

Gains shrink as the backbone becomes stronger: they are smaller for PerturbDiff than for ChemCPA, and smaller on Tahoe-100M than on sci-Plex3, indicating that MoRe can supply drug information that weaker backbones cannot extract and stronger ones already partly capture. For ChemCPA, MoRe narrows but does not close the gap to the perturbation mean on either dataset, pointing to a limit of the backbone itself. For PerturbDiff, it lifts sci-Plex3 DEG50 correlation from below the perturbation mean to above it ($0.22 \rightarrow 0.29$, vs. 0.26) and widens its lead on Tahoe-100M (0.75 vs. 0.54). Although MoRe supervises the response at the condition-mean level, it preserves distributional fit as measured by energy distance and gene-wise Wasserstein distance (Appendix H).

5.2 Which components account for the improvement

To understand *what drives the gains* of MoRe and how its two components interact, we ablate one design choice at a time on PerturbDiff and sci-Plex3, holding everything else fixed (Figure 3).

Fusion outperforms every single view. On their own, the three views are interchangeable: all reach a DEG50 correlation of about 0.25, within seed dispersion of one another. Fusing all three improves on the best single view by 11% on DEG50 and 13% on DEG20 (Figure 3a). Correlations are consistently higher on DEG20 than on DEG50, as expected because the top-ranked genes carry the strongest response signal. ChemCPA shows the same pattern (Appendix H).

Condition-mean response supervision helps, but overlaps with fusion. Figure 3b toggles the response supervision for each representation. Response supervision improves most single views but adds little once the views are fused. Relative to the chemical-only baseline under the native objective, fusion alone gains 27%, supervision alone 14%, and the two together 29%. The two gains are strongly sub-additive, which suggests that they act on the same bottleneck, a drug latent that tracks the response poorly. On Tahoe-100M the roles reverse: fusion alone raises DEG50 correlation by only 2%, supervision alone by 9%, and both together by 10% (Appendix I). Which component dominates therefore depends on the dataset, but their combination is best on both. We examine this latent directly in Section 5.4.

Simple fusion works best. More elaborate fusion operators do not help. Under the native objective every operator improves on the chemical view alone, but once response supervision is added, only concatenation remains stable across seeds (Figure 3c; Appendix I). We attribute this to the limited number and diversity of training drugs: higher-capacity operators are easily overparameterized in this regime, and one can expect their advantage to emerge when larger drug collections are available.

5.3 MoRe Combines Views that Capture Different Responses

To understand *where each view contributes*, we break the response down by biological context. A drug shapes the transcriptional response both through its physicochemical properties and through the targets it engages. If different gene programmes follow different routes, each drug view should best predict the programmes of the route it captures. The three views indeed induce different geometries over the same set of compounds (Figure 1), and fusing them outperforms each alone (Section 5.2),

but aggregate accuracy cannot show whether each view contributes to a different part of the response. We therefore stratify DEG50 genes by MSigDB Hallmark programme (Liberzon et al., 2015), and conditions by the drug's mechanism-of-action (MoA) class or by response strength, to compare how well each view predicts the response within each group. We use direction accuracy, defined for every condition–gene pair and can therefore be pooled over any subset of genes.

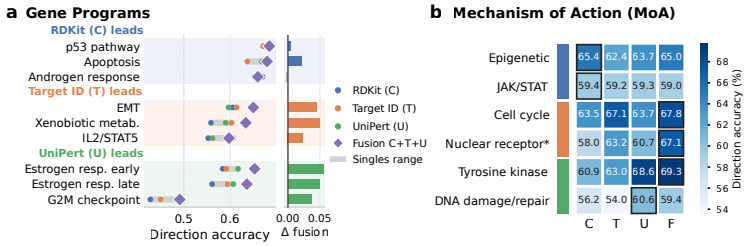

Figure 4: **No single drug view leads everywhere.** Direction accuracy on the top 50 DEGs (PerturbDiff, sci-Plex3). C, chemistry (RDKit); T, targetID; U, UniPert; F, fusion. **(a)** Hallmark gene programmes grouped by leading single view. **(b)** MoA classes with best value per row boxed.

No single view leads everywhere. Each single view leads in some Hallmark programmes (Liberzon et al., 2015) and MoA classes (Srivatsan et al., 2020) but not in others, whereas fusion exceeds the mean of the three views in 20 of 22 programmes and 9 of 10 MoA classes (Figure 4; all groups in Appendix K). Selecting the single view with the best overall score would therefore sacrifice accuracy on the groups where another view leads, whereas fusion performs well across them.

Chemistry and targets lead different programmes. Two drugs can produce the same response for two different reasons: their molecules look alike, or they act on the same proteins. The gene programmes in Figure 4a and Appendix K split along these lines. Chemistry-based representation do best on generic stress-response and cell-death programs such as the p53 signaling and apoptosis (Kastenhuber & Lowe, 2017), triggered by general chemical properties (Martin et al., 2002). Target-informed views do best on programs driven by specific proteins such as EMT and the G2M checkpoint, which are driven by specific signaling receptors and cell-cycle kinases (Lamouille et al., 2014). Grouping drugs by mechanism of action gives the same picture (Figure 4b): for example, target-informed views lead classes such as tyrosine kinase inhibitors, where drugs with different structures hit the same proteins and act alike (Srivatsan et al., 2020; Subramanian et al., 2017).

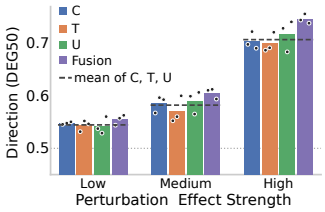

Figure 5: **Fusion helps most where the response is strong.** Direction agreement on the DEG50, binned by perturbation effect strength and averaged over three cell lines (PerturbDiff, sci-Plex3). Bars: seed means; dots: seeds; dashed line: mean of single views.

Gains concentrate where the drug effect is strong. If fusion's gain came from fitting variation shared across conditions rather than the drug effect, it would be largest where the drug effect is weak. Stratifying conditions by response strength shows the opposite (Figure 5). Fusion beats the best single view in the strong-response bin, and its margin grows with effect size; in the weakest bin, every arm sits near chance (0.5) and the advantage disappears. The gain therefore comes from conditions with a higher signal-to-noise ratio, where there is more real drug effect to predict. Together, these results confirm the complementary nature of drug representations at the level of individual programmes and drug MoA. We next ask how, by examining the learned drug latent space.

5.4 MoRe ORGANISES THE DRUG LATENT BY RESPONSE

For a held-out drug, every input except its latent $\mathbf{z}_p$ was seen during training. Therefore, prediction depends on where the unseen drug lies relative to the training drugs. **A useful drug space should therefore place drugs with similar responses close together.** Furthermore, this structure should extend to new drugs. We hypothesise that the fused latent is better structured than any single-view latent, and that this contributes to the performance gain.

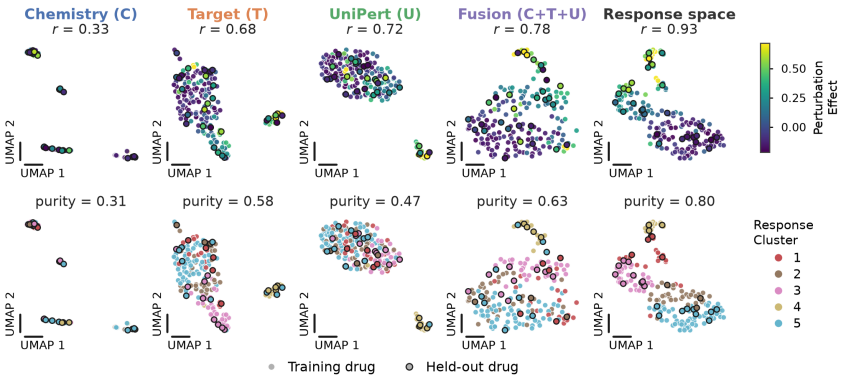

Figure 7: **Drug latent space of PerturbDiff.** UMAP of each drug latent and of the observed residuals; outlined dots are held-out drugs. **Top:** coloured by perturbation effect strength; r, correlation of each drug's effect with the mean of its 10 nearest neighbours. **Bottom:** coloured by residual k-means cluster (k=5); purity, fraction of 10 nearest neighbours in the same cluster.

Fusion better groups drugs by response. To test this, we measure *neighbour response similarity*: for each anchor drug, we rank all other drugs by cosine similarity in the latent, take the top k, and average the cosine similarity between the anchor's mean-centred residual and theirs, reporting this as a function of k (Figure 6). Chemical neighbourhoods are close to random, as structurally similar compounds can respond differently (Martin et al., 2002); target-informed views are more informative, and fusion is the most informative at every k. UMAP shows the same structure: the fused latents are ordered by perturbation effect and response cluster, whereas the chemical latent splits into clusters that mix strong and weak responders (Figure 7). The fused latent best reflects the latent geometry of observed perturbation response (r=0.78), with held-out drugs landing among training drugs of matching phenotype.

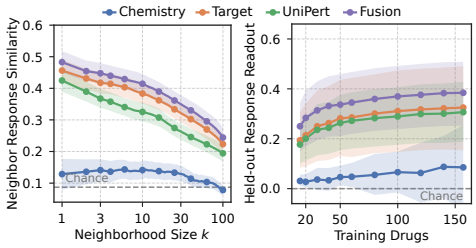

Figure 6: **The fused latent is organised by response and generalises to unseen drugs. Left:** response similarity of each drug's k nearest latent neighbours. **Right:** held-out Pearson r with 95% CI of a linear probe fitted on training drugs.

This structure generalises to unseen drugs. To test whether the latent space supports unseen-drug prediction, we fit a linear probe (Alain & Bengio, 2018) from latent to response on the training drugs and apply it to the 30 held-out drugs (Figure 6, right). Implementation details in Appendix L.1. We use a probe rather than neighbourhood similarity because distance weighs all latent directions equally, while the decoder, like a probe, can ignore directions irrelevant to the response. The fused latent predicts held-out responses most accurately at every number of training drugs and its advantage is largest when few training drugs are available.

6 Discussion

Predicting responses to unseen drugs is limited from two sides: chemical structure provides mechanistic information only implicitly, and objectives over the whole expression profile reward background shared across conditions rather than the drug effect, so that a predictor can score well while largely ignoring the drug. MoRe addresses both with a model-agnostic plug-in: it fuses chemical and target-informed views into one drug representation and adds an auxiliary loss on average responses. On an autoencoder and a diffusion model across two datasets, it improves every residual metric. The drug views are complementary: they lead on different gene programmes, and fusion places held-out drugs nearer training drugs that respond like them than any single view.

More broadly, MoRe reflects a different route to progress. **Rather than proposing yet another architecture, MoRe improves existing predictors in place, so their strengths are kept rather than rebuilt from scratch.** As perturbation models continue to improve, such plug-ins can carry these gains to each new backbone, moving towards virtual-cell models that can be queried with compounds outside the measured space.

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.

REPRODUCIBILITY NOTE

Experiment specifications record dataset preprocessing, modalities, splits, training budgets, and seeds. Response-loss weights use validation drugs; reported architecture selections are retrospective. Appendix C specifies feature preprocessing, condition-mean sampling, and the implemented response objective. All reported predictions use the final checkpoint, and test metrics are never used for selection.

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, 2025.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes, 2018. URL https://arxiv.org/abs/1610.01644.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2): 423–443, 2019. doi: 10.1109/TPAMI.2018.2798607.

Charlotte Bunne, Andreas Krause, and Marco Cuturi. Supervised training of conditional monge maps. *Advances in Neural Information Processing Systems*, 35:6859–6872, 2022.

Charlotte Bunne, Stefan G. Stark, Gabriele Gut, Jacobo Sarabia del Castillo, Mitch Levesque, Kjong-Van Lehmann, Lucas Pelkmans, Andreas Krause, and Gunnar Rätsch. Learning single-cell perturbation responses using neural optimal transport. *Nature Methods*, 20:1759–1768, 2023. doi: 10.1038/s41592-023-01969-x.

Charlotte Bunne, Yusuf Roohani, Yanay Rosen, Ankit Gupta, Xikun Zhang, et al. How to build the virtual cell with artificial intelligence: Priorities and opportunities. *Cell*, 187(25):7045–7063, 2024. doi: 10.1016/j.cell.2024.11.015.

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.

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, 2024.

Minsheng Hao, Jing Gong, Xin Zeng, Chiming Liu, Yucheng Guo, Xingyi Cheng, Taifeng Wang, Jianzhu Ma, et al. Large-scale foundation model on single-cell transcriptomics. *Nature Methods*, 21(8):1481–1491, 2024. doi: 10.1038/s41592-024-02305-7.

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.

Weihua Hu, Bowen Liu, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay Pande, and Jure Leskovec. Strategies for pre-training graph neural networks. In *International Conference on Learning Representations (ICLR)*, 2020.

Kexin Huang, Tianfan Fu, Lucas M Glass, Marinka Zitnik, Cao Xiao, and Jimeng Sun. DeepPurpose: a deep learning library for drug–target interaction prediction. *Bioinformatics*, 36(22-23): 5545–5547, 2020. doi: 10.1093/bioinformatics/btaa1005.

Edward R Kastenhuber and Scott W Lowe. Putting p53 in context. *Cell*, 170(6):1062–1078, 2017.

Justin Lamb, Emily D Crawford, David Peck, Joshua W Modell, Irene C Blat, Matthew J Wrobel, Jim Lerner, Jean-Philippe Brunet, Aravind Subramanian, Kenneth N Ross, et al. The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *science*, 313(5795):1929–1935, 2006.

Samy Lamouille, Jian Xu, and Rik Derynck. Molecular mechanisms of epithelial–mesenchymal transition. *Nature reviews Molecular cell biology*, 15(3):178–196, 2014.

Bo Li, Tingyang Yu, Scott Oloff, Rui Liao, Tianyu Cui, and Zhoubing Xu. Simple predictors, strong baselines: Revisiting chemical perturbation prediction. 2026a.

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*, 2026b.

Arthur Liberzon, Chet Birger, Helga Thorvaldsdóttir, Mahmoud Ghandi, Jill P. Mesirov, and Pablo Tamayo. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Systems*, 1(6):417–425, 2015. doi: 10.1016/j.cels.2015.12.004.

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

Mohammad Lotfollahi, Anna Klimovskaia Susmelj, Carlo De Donno, Leon Hetzel, Yuge Ji, Ignacio L Ibarra, Sanjay R Srivatsan, Mohsen Naghipourfar, Riza M Daza, Beth Martin, Jay Shendure, Jose L McFaline-Figueroa, Pierre Boyeau, F Alexander Wolf, Nafissa Yakubova, Stephan Günnemann, Cole Trapnell, David Lopez-Paz, and Fabian J Theis. Predicting cellular responses to complex perturbations in high-throughput screens. *Mol. Syst. Biol.*, 19(6):e11517, June 2023.

Yvonne C Martin, James L Kofron, and Linda M Traphagen. Do structurally similar molecules have similar biological activity? *Journal of medicinal chemistry*, 45(19):4350–4358, 2002.

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.

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

Joseph M Replogle, Reuben A Saunders, Angela N Pogson, Jeffrey A Hussmann, Alexander Lenail, Alina Guna, Lauren Mascibroda, Eric J Wagner, Karen Adelman, Gila Lithwick-Yanai, et al. Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq. *Cell*, 185(14):2559–2575, 2022.

David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 2010. doi: 10.1021/ci100050t.

Rohit Singh, Samuel Sledzieski, Bryan Bryson, Lenore Cowen, and Bonnie Berger. Contrastive learning in protein language space predicts interactions between drugs and protein targets. *Proceedings of the National Academy of Sciences*, 120(24), 2023. doi: 10.1073/pnas.2220778120.

Francesco Sirci, Francesco Napolitano, Sandra Pisonero-Vaquero, Diego Carrella, Diego L. Medina, and Diego di Bernardo. Comparing structural and transcriptional drug networks reveals signatures of drug activity and toxicity in transcriptional responses. *npj Systems Biology and Applications*, 3(1):23, Aug 2017. ISSN 2056-7189. doi: 10.1038/s41540-017-0022-3. URL https://doi.org/10.1038/s41540-017-0022-3.

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

Christina V. Theodoris, Ling Xiao, Anant Chopra, Mark D. Chaffin, Zeina R. Al Sayed, Matthew C. Hill, Helene Mantineo, Elizabeth M. Brydon, et al. Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624, 2023. doi: 10.1038/s41586-023-06139-9.

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.

Yuyang Wang, Jianren Wang, Zhonglin Cao, and Amir Barati Farimani. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence*, 4(3):279–287, 2022. doi: 10.1038/s42256-022-00447-x.

Zhonghua Wang, Lu Liang, Zheng Yin, and Jianping Lin. Improving chemical similarity ensemble approach in target prediction. *Journal of cheminformatics*, 8(1):20, 2016.

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.

Zhenqin Wu, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay Pande. MoleculeNet: a benchmark for molecular machine learning. *Chemical Science*, 9(2):513–530, 2018. doi: 10.1039/C7SC02664A.

Jie Yu, Cheng Shi, Yiran Zhou, Ningfeng Liu, Xiaolin Zong, Zhenming Liu, and Liangren Zhang. DTLCDR: A target-based multimodal fusion deep learning framework for cancer drug response prediction. *Journal of Pharmaceutical Analysis*, 15(8):101315, 2025. doi: 10.1016/j.jpha.2025.101315.

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

Jinxiong Zhang, Hedong Xu, Chunyan Tang, Dandan Ni, and Xuyang Shen. MDCDR: predicting cancer drug response via multimodal feature fusion and feature disentanglement. *Briefings in Bioinformatics*, 27(3), 2026. doi: 10.1093/bib/bbag299.

Wenjuan Zhang, Yaofei Bai, Yonghua Wang, and Wei Xiao. Polypharmacology in drug discovery: a review from systems pharmacology perspective. *Current pharmaceutical design*, 22(21):3171–3181, 2016.

APPENDIX OVERVIEW

Table 2 summarises each appendix section and the part of the main text it supports.

Table 2: **Structure of the appendix.**

	Section	Contents
A	Datasets and evaluation protocol	Drug splits and held-out conditions, validation drugs, gene panel, DEG construction, control references, condition variance analysis
B	Drug representations	The chemistry, target-identity and UniPert views
C	Implementation and provenance	Feature preprocessing, dose handling and response-loss implementation, DEG reweighting, fusion alternatives, selection and reporting
D	Gene weighting	Balancing full-panel and DEG errors in the response objective
E	Loss decomposition	Assumptions and derivation of the per-cell and population losses and their gradients (supports Eqs. 5 and 7)
F	Training configuration and baselines	Hyperparameters, response-loss weight selection and baseline settings
G	Metrics	Residual Pearson and direction, retrieval rank, distribution metrics
H	Complete results	Every representation and fusion operator under both objectives, for both backbones and datasets (extends Table 1); expression-space and distribution-level metrics
I	Ablations in detail	Direction agreement and Tahoe-100M ablation
J	Output-space ensembles	Learned fusion versus averaging separately trained single-view models
K	Biological analysis	All gene programmes and MoA classes; perturbation effect strength per cell line (extends Figures 4 and 5)
L	Latent-space analysis	Other seeds, latent statistics, robustness to the similarity definition, neighbourhood–prediction coupling; linear probe on held-out drugs (extends Section 5.4)

A DATASETS AND EVALUATION PROTOCOL

sci-Plex3. We use the three cell lines (A549, K562, MCF7) and four doses of sci-Plex3 (Srivatsan et al., 2020). A fixed drug-level split randomly holds out 30 of 187 drugs, giving 360 held-out conditions (drug $\times$ cell line $\times$ dose). For selecting the response-loss weight, 16 training drugs from the train set (10%) are held out as a drug-level validation set. For each model, we select once using RDKit embedding and then fix it for all its experiments.

Tahoe-100M. We use Tahoe-100M Zhang et al. (2025) with the fixed 80/20 train/test split shared by all methods. The evaluation covers 76 unseen drugs across A549, RKO, and Hs766T cell lines. These three cell lines are randomly selected across three different organs. We use the same $2{,}000$-gene panel from PerturbDiff (Yuan et al., 2026). For control-based references, DMSO cells are pooled across plates within each cell line.

Differentially expressed genes. On sci-Plex3, we compute DEGs using Wilcoxon rank-sum test (scanpy `rank_genes_groups`) ranked per drug; the same panel is used for every method. On Tahoe-100M they come from the official pseudobulk DESeq2 Wald test against plate-matched DMSO and genes ranked by $|\log_2 \mathrm{FC}|$ then adjusted p-value.

Control references. On sci-Plex3 every residual subtracts the training-control mean of the condition's cell line. On Tahoe-100M, we subtract the training-control mean of the cell line with controls pooled across plates.

Drug effect and within-condition variation in sci-Plex3. Table 3 compares, for each training condition, the squared drug effect $\|\boldsymbol{\delta}_q\|^2$ with the within-condition variance $\mathbb{E}\|\boldsymbol{\eta}_j\|^2$. For the median condition the drug effect is 0.15% of the within-condition variance, and even at the 90th percentile it is below 4%.

Table 3: **Size of the drug effect relative to within-condition variation of sci-Plex3** For each condition, $\|\boldsymbol{\delta}_q\|^2/\mathbb{E}\|\boldsymbol{\eta}_j\|^2$ is the squared shift of the condition mean from the control mean of its cell line, divided by the total variance of its treated cells. A lower value means the drug effect is smaller relative to the variation among individual cells, so a single cell carries less information about it.

Quantile over conditions	10%	25%	50%	75%	90%
$\|\boldsymbol{\delta}_q\|^2/\mathbb{E}\|\boldsymbol{\eta}_j\|^2$	0.013%	0.056%	0.15%	0.70%	3.5%

B Drug representations

Views. Chemistry (C) is the 178-dimensional RDKit2D descriptor vector of Hetzel et al. (2022). Target identity (T) assigns each annotated protein target a fixed random 512-dimensional vector and represents a drug by the mean over its targets, so it encodes which targets drugs share but nothing about the proteins themselves. UniPert (U) is the 256-dimensional target-informed chemical embedding of Li et al. (2026b). Feature preprocessing is described in Appendix C.

C Implementation and provenance

Feature preprocessing. RDKit, target-identity, and UniPert inputs have 178, 512, and 256 features, respectively. The recorded feature preprocessing replaces nonfinite entries with zero, standardizes each feature using the frozen all-drug table, substitutes a denominator of one for standard deviations below 10^{-8}, and clips standardized values to $[-10^4, 10^4]$. This is transductive *feature-only* preprocessing: held-out-drug embeddings. The modality adapters then apply layer normalization, an MLP, and ℓ_2 normalization.

Dose and response-loss implementation. ChemCPA uses its native amortized scalar doser. PerturbDiff uses the categorical drug–dose bridge and its pinned DDIM prediction recurrence. Direct response supervision samples four training conditions every four native update cycles and averages predictions from 32 matched training controls per condition. Its loss averages full-panel and DEG-only squared errors with separate training-response scales and mixing fraction $\rho = 0.5$, as defined in Section 3.3. Tahoe-100M training signatures use cell-line–plate matching and sci-Plex3 uses cell-line matching; plate enters the predictor as a training covariate for PerturbDiff following the original paper (Yuan et al., 2026). At evaluation, every residual on both datasets, including those of the baselines, subtracts the training-control mean of the condition's cell line (Appendix A).

DEG Reweighting (Mejia et al., 2026) For optimization-training conditions $\mathcal{Q}_{\text{train}}$, the implementation sets

$$\sigma^2_{\text{all}} = \max\left(10^{-8}, \frac{1}{|\mathcal{Q}_{\text{train}}|}\sum_{q\in\mathcal{Q}_{\text{train}}}\frac{1}{n}\sum_{g}\delta^2_{qg}\right), \tag{9}$$

$$\sigma^2_{\text{DEG}} = \max\left(10^{-8}, \frac{1}{|\mathcal{Q}_{\text{train}}|}\sum_{q\in\mathcal{Q}_{\text{train}}}\frac{1}{|\mathcal{G}(q)|}\sum_{g\in\mathcal{G}(q)}\delta^2_{qg}\right). \tag{10}$$

The loss can equivalently be written as $|\mathcal{B}|^{-1}\sum_q\sum_g w_{qg}(\widehat{\delta}_{qg}-\delta_{qg})^2$ with

$$w_{qg} = \frac{1-\rho}{n\sigma^2_{\text{all}}} + \frac{\rho\,\mathbf{1}[g\in\mathcal{G}(q)]}{|\mathcal{G}(q)|\sigma^2_{\text{DEG}}}. \tag{11}$$

These weights encode the implemented full-panel/DEG balance; they are not a fixed, unnormalized DEG multiplier. A further normalization by $\sum_g w_{qg}$ would rescale the loss.

Fusion alternatives. The comparison includes individual views, chemical-plus-target concatenation, and three-view concatenation, MoE dense gating, and self-attention. Self-attention treats modality latents as tokens, adds learned modality embeddings, and prepends a learned fusion token:

$$\mathbf{z}_p = \text{normalize}(\text{Transformer}([\mathbf{e}_{\text{fuse}}, \mathbf{h}_1(p) + \mathbf{e}_1, \ldots, \mathbf{h}_M(p) + \mathbf{e}_M])_{\text{fuse}}) \,. \tag{12}$$

Dense-gating comparisons use declared balance weights 10^{-4}, 10^{-3}, and 10^{-2}. The aggregate balancing term discourages collapse of mean routing weights across training drugs; it does not force each drug to assign identical weights to every view. Concatenation is the operator reported in every main-text table, for both backbones and both datasets; the alternatives are ablations only. We fix concatenation throughout rather than selecting an operator per setting, which also removes operator choice as a source of retrospective selection.

Selection and reporting. Representation comparisons within each backbone share the split, training budget, dose mechanism and context conditioning. Response-loss weights are selected on a deterministic drug-OOD validation split; the fusion operator is fixed to concatenation throughout and is therefore not selected on test data. The perturbation-mean baseline is a cell-weighted mean rather than an equally weighted average of condition means, and has no training-seed uncertainty; its construction is shared by backbones evaluated under the same dataset protocol. For retrieval rank, zero-norm residuals make the cosine rank undefined, and ranks are averaged across conditions and then across seeds.

D GENE WEIGHTING IN THE RESPONSE OBJECTIVE

Mejia et al. (2026) show that objectives over the whole gene panel dilute the perturbation signal and propose reweighting genes by their differential-expression statistics, applied per cell. We use a two-set simplification of this idea on the population-level error of Section 3.3. With the mean squared error of condition q over a gene set S,

$$e_S(q) = \frac{1}{|S|} \sum_{g \in S} \left(\widehat{\delta}_{qg} - \delta_{qg} \right)^2 , \tag{13}$$

the objective for a batch $\mathcal{B}$ of training conditions is

$$\mathcal{L}_{\text{aux}} = \frac{1}{|\mathcal{B}|} \sum_{q \in \mathcal{B}} \left[(1-\rho) \, \frac{e_{\mathcal{V}}(q)}{\sigma^2_{\text{all}}} \; + \; \rho \, \frac{e_{\mathcal{G}(q)}(q)}{\sigma^2_{\text{DEG}}} \right] , \tag{14}$$

where $\mathcal{V}$ is the full panel, $\mathcal{G}(q)$ the DEG set of condition q, and σ^2_{all}, σ^2_{DEG} the mean squared response magnitudes of the two sets over training conditions. Averaging within each set and dividing by its typical response size prevents the panel from dominating by count and the DEGs by magnitude. We use $\rho = 0.5$ throughout.

E DECOMPOSITION OF THE PER-CELL AND CONDITION LOSSES

Given a condition $q = (c, p, s)$. A prediction is $\widehat{\mathbf{y}}_i = F_\theta(\mathbf{x}_i, \mathbf{z}_p, s; \boldsymbol{\epsilon}_i)$ with $\mathbf{x}_i \sim \mathcal{D}_c^{\text{ctrl}}$ and backbone noise $\boldsymbol{\epsilon}_i$ (absent for a deterministic backbone), and a treated cell is $\mathbf{y}_j \sim \mathcal{D}_q$. Let $\boldsymbol{\mu}_q = \mathbb{E}_{\mathbf{y} \sim \mathcal{D}_q}[\mathbf{y}]$ be the observed condition mean and $\bar{\boldsymbol{\mu}}_q = \mathbb{E}_{\mathbf{x}_i, \boldsymbol{\epsilon}_i}[\widehat{\mathbf{y}}_i]$ the model's mean prediction, and define

$$\mathbf{a}_i = \widehat{\mathbf{y}}_i - \bar{\boldsymbol{\mu}}_q, \qquad \boldsymbol{\eta}_j = \mathbf{y}_j - \boldsymbol{\mu}_q, \qquad \mathbf{b} = \bar{\boldsymbol{\mu}}_q - \boldsymbol{\mu}_q. \tag{15}$$

Note, these are purely definitions, not a model of the data: $\mathbf{y}_j = \boldsymbol{\mu}_q + \boldsymbol{\eta}_j$ does not assume additive drug effect, noise distribution, or independence between $\boldsymbol{\mu}_q$ and $\boldsymbol{\eta}_j$, and simply names each cell's deviation from its condition mean. Intuitively, $\mathbf{a}_i$ is how far one predicted cell deviates from the model's average prediction, $\boldsymbol{\eta}_j$ is how far one observed treated cell deviates from the true condition mean, and $\mathbf{b}$ is how far the model's average prediction is from the true condition mean, *i.e.* the error in the predicted drug effect.

By definition $\mathbb{E}[\mathbf{a}_i] = \mathbb{E}[\boldsymbol{\eta}_j] = \mathbf{0}$, and $\mathbf{b}$ is non-random: it depends on θ but not on the sampled cells or noise as $\bar{\boldsymbol{\mu}}_q$ represents the model's mean prediction over the whole control pool and noise, as opposed to the mini-batch pseudobulk $\widehat{\boldsymbol{\mu}}_q$. We assume that (i) $(\mathbf{x}_i, \boldsymbol{\epsilon}_i)$ is drawn independently of $\mathbf{y}_j$, as the data are unpaired; (ii) the n pairs $(\mathbf{x}_i, \boldsymbol{\epsilon}_i)$ in Eq. (6) are i.i.d.; and (iii) all second moments are finite.

Per-cell loss. Since $\widehat{\mathbf{y}}_i - \mathbf{y}_j = \mathbf{a}_i + \mathbf{b} - \boldsymbol{\eta}_j$,

$$\mathbb{E}\|\widehat{\mathbf{y}}_i - \mathbf{y}_j\|^2 = \|\mathbf{b}\|^2 + \mathbb{E}\|\boldsymbol{\eta}_j\|^2 + \mathbb{E}\|\mathbf{a}_i\|^2 + 2\,\mathbb{E}[\mathbf{a}_i]^\top \mathbf{b} - 2\,\mathbf{b}^\top \mathbb{E}[\boldsymbol{\eta}_j] - 2\,\mathbb{E}[\mathbf{a}_i^\top \boldsymbol{\eta}_j]. \tag{16}$$

The first two cross terms vanish because $\mathbb{E}[\mathbf{a}_i] = \mathbb{E}[\boldsymbol{\eta}_j] = \mathbf{0}$, and the third because, by (i), $\mathbb{E}[\mathbf{a}_i^\top \boldsymbol{\eta}_j] = \mathbb{E}[\mathbf{a}_i]^\top \mathbb{E}[\boldsymbol{\eta}_j] = 0$. This gives Eq. (5).

Condition-mean loss. The control mean cancels in $\widehat{\boldsymbol{\delta}}_q - \boldsymbol{\delta}_q$, so with $\widehat{\boldsymbol{\mu}}_q = \frac{1}{n}\sum_{i=1}^n \widehat{\mathbf{y}}_i$ (Eq. (2) with $m = n$) and the batch average $\widehat{\mathbf{a}} = \frac{1}{n}\sum_{i=1}^n \mathbf{a}_i$,

$$\ell_q = \|\widehat{\boldsymbol{\mu}}_q - \boldsymbol{\mu}_q\|^2 = \|\widehat{\mathbf{a}} + \mathbf{b}\|^2 = \|\widehat{\mathbf{a}}\|^2 + 2\,\widehat{\mathbf{a}}^\top \mathbf{b} + \|\mathbf{b}\|^2. \tag{17}$$

The cross term has zero expectation since $\mathbb{E}[\widehat{\mathbf{a}}] = \mathbf{0}$. By (ii), $\mathbb{E}[\mathbf{a}_i^\top \mathbf{a}_k] = \mathbb{E}[\mathbf{a}_i]^\top \mathbb{E}[\mathbf{a}_k] = 0$ for $i \neq k$, and $\mathbb{E}\|\mathbf{a}_i\|^2$ is the same for all i, so

$$\mathbb{E}\|\widehat{\mathbf{a}}\|^2 = \frac{1}{n^2}\sum_{i=1}^n \sum_{k=1}^n \mathbb{E}[\mathbf{a}_i^\top \mathbf{a}_k] = \frac{1}{n}\,\mathbb{E}\|\mathbf{a}_i\|^2. \tag{18}$$

Hence $\mathbb{E}[\ell_q] = \|\mathbf{b}\|^2 + \frac{1}{n}\,\mathbb{E}\|\mathbf{a}_i\|^2$, which is Eq. (7).

Gene weighting. Only the bilinearity of the inner product was used, so both results hold with $\|\cdot\|$ replaced by $\|\cdot\|_{W_q}$ for the fixed diagonal weights W_q of Appendix D.

F Training configuration and baselines

Hyperparameters. The response-loss weight λ_{aux} is selected per backbone and dataset on the drug-level validation set by DEG50 residual Pearson, from $\{0.01, 0.1, 1.0\}$. Every reported prediction comes from the final checkpoint; test-set metrics are logged during training but never used for selection. The response loss averages the predictions of $m = 32$ control cells per condition, drawn from a context-matched pool of 256 control cells.

ChemCPA. ChemCPA (Hetzel et al., 2022) is an autoencoder that encodes an expression profile into a basal-state latent and decodes it conditioned on drug, dose and cell-line embeddings. At inference, it encodes a control cell and decodes it under the embedding of the queried drug. We use the default configuration of the official repository and keep it fixed across all drug representations and objectives, training for $48{,}000$ steps on sci-Plex3 and $96{,}000$ on Tahoe-100M at $2{,}048$ cells per batch; with MoRe, its drug embedding is replaced by the fused latent.

PerturbDiff. PerturbDiff (Yuan et al., 2026) is a conditional diffusion model over cell populations. Its public implementation indexes drugs by identity and therefore cannot predict unseen drugs, so we condition it on a drug representation encoded by an MLP, combined with dose as described in Appendix C. The training budget is shared by all representations and objectives: $5{,}000$ steps on sci-Plex3 and $10{,}000$ on Tahoe-100M at $2{,}048$ cells per batch.

CondOT. CondOT (Bunne et al., 2022) learns a conditional optimal-transport map from control to treated cells. Because every condition differs in drug, dose and cell line, the map is conditioned on the RDKit drug representation and dose. We use a 128-dimensional latent and hidden width, batch size 256, learning rate 10^{-4} and 1.1×10^6 optimiser steps. At evaluation, it transports control cells of each queried condition.

Biolord. Biolord (Piran et al., 2024) learns a disentangled representation of cells with separate latent factors for known attributes such as drug, dose and cell line, and predicts a condition by recombining them. We condition it on the RDKit drug representation so that it can represent unseen drugs.

Common setup. All methods use the same drug split, preprocessing and gene panel, are trained with three seeds, and are scored with the same evaluator on the same held-out conditions (360 on sci-Plex3 and 683 on Tahoe-100M).

G METRICS

Residual metrics. For condition q with DEG set $\mathcal{G}(q)$, residual Pearson is the correlation between $\widehat{\boldsymbol{\delta}}_q$ and $\boldsymbol{\delta}_q$ restricted to $\mathcal{G}(q)$, and direction agreement is the fraction of $g \in \mathcal{G}(q)$ with $\operatorname{sign} \widehat{\delta}_{qg} = \operatorname{sign} \delta_{qg}$, using a tolerance of 10^{-5} below which a value counts as zero. Scores are averaged over conditions with equal weight within a run and then over seeds.

Retrieval rank. Following PerturBench (Wu et al., 2024), for each condition i in a pool of P held-out conditions of the same cell line we compute $d_{ji} = 1 - \cos(\widehat{\boldsymbol{\delta}}_j, \boldsymbol{\delta}_i)$ over the full gene panel and report

$$\text{rank}_i = \frac{|\{\, j \neq i : d_{ji} \leq d_{ii} \,\}|}{P - 1}. \tag{19}$$

Zero is best and 0.5 is the expectation for an uninformative predictor. Ties count against the model, so a predictor that outputs the same response for every condition of a cell line, such as the perturbation mean, scores 1. Pools are formed within cell line, so cell-line identity cannot improve the score.

Distribution metrics. We also report the multivariate energy distance (E-distance) and the mean gene-wise one-dimensional Wasserstein-1 distance (EMD) between predicted and observed single-cell populations in the appendix following Doloris (Chi et al., 2026).

H COMPLETE RESULTS

Tables 4–7 report every drug representation and fusion operator under both objectives, of which Table 1 shows the native chemistry row and the response-supervised three-view concatenation row. Table 8 adds expression-space metrics: every method, including the control mean that predicts no response, scores close to ceiling on expression, which is why the main text scores control-referenced residuals.

Distribution-level metrics. Table 9 scores the predicted single-cell distributions on sci-Plex3 with the energy distance and the gene-wise Wasserstein-1 distance (Appendix G). Response supervision targets the condition mean and leaves the distribution-level fit essentially unchanged: for PerturbDiff, fusion slightly improves both distances on DEG50 while the all-gene energy distance increases slightly, and for ChemCPA all three configurations lie within about 1% of each other.

I ABLATIONS IN DETAIL

Direction agreement. Figure 8 repeats Figure 3 with direction agreement.

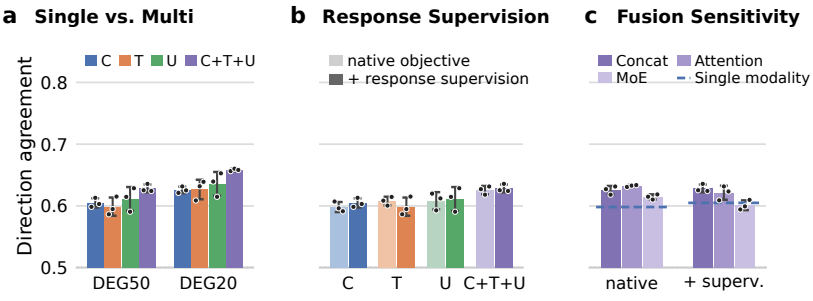

Figure 8: **Direction agreement for the ablations of Figure 3** (PerturbDiff, sci-Plex3). Panels, colours, error bars and dots as in Figure 3; the dotted line is chance.

Table 4: **sci-Plex3, PerturbDiff: every drug representation under both objectives.** 360 held-out conditions; mean over three seeds with sample SD as subscript. Rank is the within-cell-line retrieval rank over all genes (lower is better). Bold marks the best value within each objective.

	DEG50 residual		DEG20 residual		
Drug representation	Pearson	Direction	Pearson	Direction	Rank ↓
Perturbation mean	0.256	0.630	0.272	0.628	1.000
Native objective					
Chemistry (C)	$0.221_{\pm 0.015}$	$0.598_{\pm 0.008}$	$0.249_{\pm 0.010}$	$0.616_{\pm 0.008}$	$0.417_{\pm 0.002}$
Target identity (T)	$0.265_{\pm 0.018}$	$0.608_{\pm 0.007}$	$0.281_{\pm 0.022}$	$0.625_{\pm 0.001}$	$\mathbf{0.354_{\pm 0.019}}$
UniPert (U)	$0.239_{\pm 0.020}$	$0.608_{\pm 0.014}$	$0.262_{\pm 0.021}$	$0.626_{\pm 0.015}$	$0.376_{\pm 0.018}$
C+T, concatenation	$0.264_{\pm 0.008}$	$0.616_{\pm 0.006}$	$0.292_{\pm 0.010}$	$0.638_{\pm 0.007}$	$0.380_{\pm 0.005}$
C+T+U, concatenation	$\mathbf{0.280_{\pm 0.006}}$	$0.626_{\pm 0.007}$	$\mathbf{0.310_{\pm 0.009}}$	$0.648_{\pm 0.010}$	$0.362_{\pm 0.023}$
C+T+U, self-attention	$0.248_{\pm 0.016}$	$\mathbf{0.632_{\pm 0.002}}$	$0.279_{\pm 0.015}$	$\mathbf{0.655_{\pm 0.003}}$	$0.385_{\pm 0.002}$
C+T+U, MoE λ=10^{-4}	$0.254_{\pm 0.026}$	$0.615_{\pm 0.004}$	$0.296_{\pm 0.018}$	$0.633_{\pm 0.007}$	$0.368_{\pm 0.006}$
C+T+U, MoE λ=10^{-3}	$0.233_{\pm 0.045}$	$0.616_{\pm 0.007}$	$0.279_{\pm 0.029}$	$0.636_{\pm 0.008}$	$0.371_{\pm 0.012}$
C+T+U, MoE λ=10^{-2}	$0.266_{\pm 0.014}$	$0.616_{\pm 0.004}$	$0.306_{\pm 0.010}$	$0.636_{\pm 0.002}$	$0.363_{\pm 0.010}$
With response supervision					
Chemistry (C)	$0.253_{\pm 0.006}$	$0.605_{\pm 0.007}$	$0.276_{\pm 0.001}$	$0.626_{\pm 0.005}$	$0.433_{\pm 0.004}$
Target identity (T)	$0.258_{\pm 0.037}$	$0.599_{\pm 0.015}$	$0.289_{\pm 0.029}$	$0.627_{\pm 0.016}$	$0.377_{\pm 0.009}$
UniPert (U)	$0.255_{\pm 0.019}$	$0.611_{\pm 0.019}$	$0.288_{\pm 0.020}$	$0.636_{\pm 0.019}$	$0.384_{\pm 0.012}$
C+T, concatenation	$0.261_{\pm 0.030}$	$0.605_{\pm 0.007}$	$0.301_{\pm 0.018}$	$0.633_{\pm 0.009}$	$0.386_{\pm 0.014}$
C+T+U, concatenation	$\mathbf{0.286_{\pm 0.015}}$	$\mathbf{0.628_{\pm 0.006}}$	$\mathbf{0.326_{\pm 0.017}}$	$\mathbf{0.658_{\pm 0.002}}$	$\mathbf{0.373_{\pm 0.022}}$
C+T+U, self-attention	$0.244_{\pm 0.056}$	$0.621_{\pm 0.011}$	$0.282_{\pm 0.044}$	$0.653_{\pm 0.009}$	$0.392_{\pm 0.006}$
C+T+U, MoE λ=10^{-4}	$0.241_{\pm 0.042}$	$0.601_{\pm 0.008}$	$0.290_{\pm 0.037}$	$0.630_{\pm 0.007}$	$0.377_{\pm 0.018}$
C+T+U, MoE λ=10^{-3}	$0.131_{\pm 0.162}$	$0.597_{\pm 0.005}$	$0.158_{\pm 0.181}$	$0.621_{\pm 0.004}$	$0.389_{\pm 0.022}$
C+T+U, MoE λ=10^{-2}	$0.231_{\pm 0.044}$	$0.597_{\pm 0.014}$	$0.277_{\pm 0.042}$	$0.626_{\pm 0.014}$	$0.378_{\pm 0.013}$

Tahoe-100M. Figure 9 repeats the ablation on Tahoe-100M. Response supervision raises every representation, and once it is on, the single views and the fused model lie within 0.01 of each other in DEG50 correlation, with fusion highest.

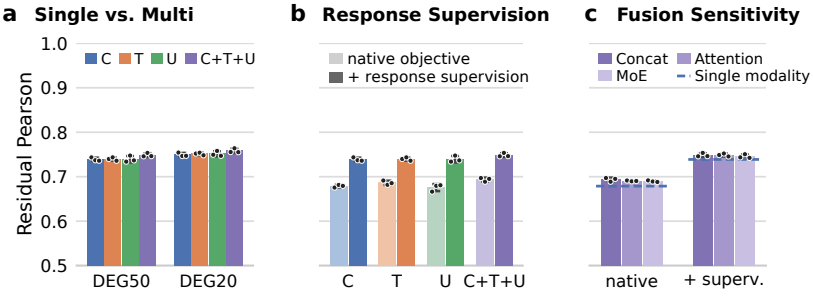

Figure 9: **Ablations on Tahoe-100M** (PerturbDiff), as Figure 3; MoE at $\lambda = 10^{-4}$. All residuals subtract the cell-line training-control mean.

J Learned fusion versus output-space ensembles

An alternative way to combine views is to train one model per view and average their predicted condition means (*output-space ensemble*); no weights are fitted. Tables 10 and 11 compare this ensemble with MoRe. The three-view ensemble also improves on the best single view for PerturbDiff on both datasets, independently confirming that the views are complementary. MoRe performs at the same level: on sci-Plex3 PerturbDiff the ensemble has the higher DEG50 correlation, while MoRe

Table 5: **sci-Plex3, ChemCPA: every drug representation under both objectives.** 360 held-out conditions; mean over three seeds with sample SD as subscript. Rank is the within-cell-line retrieval rank over all genes (lower is better). Bold marks the best value within each objective. Latent dimension 32 for both objectives.

	DEG50 residual		DEG20 residual		
Drug representation	Pearson	Direction	Pearson	Direction	Rank ↓
Perturbation mean	0.256	0.630	0.272	0.628	1.000
Native objective					
Chemistry (C)	$0.117_{\pm 0.024}$	$0.481_{\pm 0.016}$	$0.142_{\pm 0.019}$	$0.487_{\pm 0.014}$	$0.497_{\pm 0.017}$
Target identity (T)	$0.138_{\pm 0.061}$	$0.498_{\pm 0.019}$	$0.153_{\pm 0.061}$	$0.507_{\pm 0.025}$	$0.494_{\pm 0.002}$
UniPert (U)	$\mathbf{0.229_{\pm 0.015}}$	$\mathbf{0.550_{\pm 0.017}}$	$\mathbf{0.250_{\pm 0.028}}$	$\mathbf{0.572_{\pm 0.011}}$	$0.485_{\pm 0.005}$
C+T, concatenation	$0.157_{\pm 0.053}$	$0.502_{\pm 0.027}$	$0.183_{\pm 0.053}$	$0.516_{\pm 0.034}$	$0.489_{\pm 0.014}$
C+T+U, concatenation	$0.095_{\pm 0.065}$	$0.489_{\pm 0.031}$	$0.107_{\pm 0.062}$	$0.494_{\pm 0.037}$	$0.495_{\pm 0.010}$
C+T+U, self-attention	$0.109_{\pm 0.042}$	$0.484_{\pm 0.014}$	$0.128_{\pm 0.045}$	$0.489_{\pm 0.015}$	$0.497_{\pm 0.010}$
C+T+U, MoE λ=10^{-4}	$0.115_{\pm 0.068}$	$0.530_{\pm 0.013}$	$0.120_{\pm 0.068}$	$0.539_{\pm 0.003}$	$0.507_{\pm 0.013}$
C+T+U, MoE λ=10^{-3}	$0.169_{\pm 0.043}$	$0.534_{\pm 0.014}$	$0.193_{\pm 0.049}$	$0.547_{\pm 0.017}$	$0.487_{\pm 0.011}$
C+T+U, MoE λ=10^{-2}	$0.158_{\pm 0.005}$	$0.524_{\pm 0.013}$	$0.176_{\pm 0.004}$	$0.537_{\pm 0.021}$	$\mathbf{0.482_{\pm 0.012}}$
With response supervision					
Chemistry (C)	$0.221_{\pm 0.013}$	$0.555_{\pm 0.033}$	$0.235_{\pm 0.023}$	$0.574_{\pm 0.034}$	$0.491_{\pm 0.008}$
Target identity (T)	$0.245_{\pm 0.028}$	$0.544_{\pm 0.025}$	$0.266_{\pm 0.021}$	$0.564_{\pm 0.034}$	$0.486_{\pm 0.004}$
UniPert (U)	$0.247_{\pm 0.017}$	$0.549_{\pm 0.049}$	$0.266_{\pm 0.011}$	$0.571_{\pm 0.054}$	$\mathbf{0.471_{\pm 0.001}}$
C+T, concatenation	$0.216_{\pm 0.054}$	$0.521_{\pm 0.036}$	$0.240_{\pm 0.051}$	$0.536_{\pm 0.040}$	$0.486_{\pm 0.009}$
C+T+U, concatenation	$0.249_{\pm 0.029}$	$0.562_{\pm 0.046}$	$0.268_{\pm 0.024}$	$0.586_{\pm 0.046}$	$0.486_{\pm 0.005}$
C+T+U, self-attention	$0.253_{\pm 0.036}$	$0.559_{\pm 0.016}$	$0.269_{\pm 0.035}$	$0.585_{\pm 0.021}$	$0.498_{\pm 0.010}$
C+T+U, MoE λ=10^{-4}	$0.258_{\pm 0.024}$	$\mathbf{0.564_{\pm 0.031}}$	$0.279_{\pm 0.028}$	$\mathbf{0.588_{\pm 0.040}}$	$0.483_{\pm 0.011}$
C+T+U, MoE λ=10^{-3}	$\mathbf{0.274_{\pm 0.022}}$	$0.553_{\pm 0.043}$	$\mathbf{0.286_{\pm 0.026}}$	$0.583_{\pm 0.039}$	$0.480_{\pm 0.021}$
C+T+U, MoE λ=10^{-2}	$0.238_{\pm 0.033}$	$0.549_{\pm 0.018}$	$0.260_{\pm 0.034}$	$0.573_{\pm 0.020}$	$0.480_{\pm 0.003}$

is better on the other three endpoints with lower seed dispersion; on sci-Plex3 ChemCPA MoRe is at least as good as either average on every endpoint; on Tahoe-100M PerturbDiff the ensemble is slightly better on every endpoint (by at most 0.007); and on Tahoe-100M ChemCPA the chemistry view alone has the highest correlation. The ensemble, however, requires training and running one model per view, whereas MoRe uses a single model, and only MoRe yields a shared drug latent organised by response (Section 5.4).

K Full biological analysis

Figure 10 shows every Hallmark programme and MoA class that passes the inclusion criteria, of which Figure 4 shows a subset. Programmes are included when at least 200 condition–gene pairs remain after removing the lowest decile of absolute observed responses; MoA classes marked with an asterisk are represented by one compound. Figure 11 splits Figure 5 by cell line: fusion is the best representation in the high-effect tertile of every cell line.

L Latent-space analysis: full versions

Other seeds. Figure 12 shows the remaining PerturbDiff seeds, and Table 12 reports every latent statistic for both backbones. The PerturbDiff ordering (fusion above the target-informed views, chemistry near chance) holds in every seed. ChemCPA's latents are all close to one-dimensional (PC1 carries 82–99% of the variance) and no view clearly leads, so the geometric account of Section 5.4 is specific to PerturbDiff, although ChemCPA gains from fusion at the output level.

Table 6: **Tahoe-100M (A549, RKO, Hs766T), PerturbDiff: every drug representation under both objectives.** 683 held-out conditions from 76 drugs; mean over three seeds with sample SD as subscript. All rows, including the perturbation mean, subtract the training-control mean of the condition's cell line (controls pooled across plates). Rank is the within-cell-line retrieval rank over all genes (lower is better). Bold marks the best value within each objective.

	DEG50 residual		DEG20 residual		
Drug representation	Pearson	Direction	Pearson	Direction	Rank ↓
Perturbation mean	0.538	0.729	0.566	0.748	1.000
Native objective					
Chemistry (C)	$0.679_{\pm 0.003}$	$0.729_{\pm 0.002}$	$0.684_{\pm 0.005}$	$0.755_{\pm 0.002}$	$0.162_{\pm 0.002}$
Target identity (T)	$0.686_{\pm 0.005}$	$0.729_{\pm 0.004}$	$0.693_{\pm 0.006}$	$0.755_{\pm 0.004}$	$0.151_{\pm 0.006}$
UniPert (U)	$0.676_{\pm 0.008}$	$0.721_{\pm 0.004}$	$0.682_{\pm 0.008}$	$0.749_{\pm 0.003}$	$0.168_{\pm 0.001}$
C+T, concatenation	$\mathbf{0.694_{\pm 0.008}}$	$0.731_{\pm 0.005}$	$\mathbf{0.699_{\pm 0.008}}$	$\mathbf{0.759_{\pm 0.008}}$	$0.149_{\pm 0.010}$
C+T+U, concatenation	$0.694_{\pm 0.004}$	$\mathbf{0.731_{\pm 0.007}}$	$0.698_{\pm 0.006}$	$0.758_{\pm 0.008}$	$\mathbf{0.149_{\pm 0.006}}$
C+T+U, self-attention	$0.690_{\pm 0.001}$	$0.726_{\pm 0.003}$	$0.697_{\pm 0.003}$	$0.752_{\pm 0.004}$	$0.153_{\pm 0.002}$
C+T+U, MoE λ=10^{-4}	$0.690_{\pm 0.002}$	$0.728_{\pm 0.004}$	$0.697_{\pm 0.002}$	$0.752_{\pm 0.002}$	$0.151_{\pm 0.003}$
C+T+U, MoE λ=10^{-3}	$0.689_{\pm 0.000}$	$0.728_{\pm 0.005}$	$0.696_{\pm 0.000}$	$0.752_{\pm 0.003}$	$0.151_{\pm 0.003}$
C+T+U, MoE λ=10^{-2}	$0.688_{\pm 0.001}$	$0.728_{\pm 0.005}$	$0.696_{\pm 0.001}$	$0.753_{\pm 0.003}$	$0.151_{\pm 0.002}$
With response supervision					
Chemistry (C)	$0.739_{\pm 0.004}$	$\mathbf{0.761_{\pm 0.001}}$	$0.749_{\pm 0.004}$	$\mathbf{0.797_{\pm 0.002}}$	$0.151_{\pm 0.003}$
Target identity (T)	$0.740_{\pm 0.004}$	$0.760_{\pm 0.003}$	$0.751_{\pm 0.003}$	$0.796_{\pm 0.006}$	$0.149_{\pm 0.003}$
UniPert (U)	$0.739_{\pm 0.007}$	$0.759_{\pm 0.004}$	$0.751_{\pm 0.006}$	$0.794_{\pm 0.004}$	$0.152_{\pm 0.005}$
C+T, concatenation	$0.745_{\pm 0.002}$	$0.760_{\pm 0.002}$	$0.755_{\pm 0.002}$	$0.796_{\pm 0.003}$	$0.147_{\pm 0.006}$
C+T+U, concatenation	$0.749_{\pm 0.004}$	$0.761_{\pm 0.002}$	$0.758_{\pm 0.005}$	$0.796_{\pm 0.004}$	$\mathbf{0.142_{\pm 0.004}}$
C+T+U, self-attention	$\mathbf{0.749_{\pm 0.003}}$	$0.761_{\pm 0.002}$	$\mathbf{0.759_{\pm 0.003}}$	$0.796_{\pm 0.003}$	$0.148_{\pm 0.003}$
C+T+U, MoE λ=10^{-4}	$0.745_{\pm 0.004}$	$0.761_{\pm 0.004}$	$0.756_{\pm 0.004}$	$0.796_{\pm 0.006}$	$0.148_{\pm 0.002}$
C+T+U, MoE λ=10^{-3}	$0.745_{\pm 0.007}$	$0.761_{\pm 0.004}$	$0.756_{\pm 0.006}$	$0.796_{\pm 0.006}$	$0.148_{\pm 0.002}$
C+T+U, MoE λ=10^{-2}	$0.746_{\pm 0.006}$	$0.759_{\pm 0.003}$	$0.755_{\pm 0.005}$	$0.794_{\pm 0.004}$	$0.146_{\pm 0.002}$

Robustness to the definition of response similarity. Figure 13 repeats the neighbourhood analysis under six definitions of response similarity, from the leading response axis to the full residual profile with no dimensionality reduction.

L.1 Linear readout of held-out drug responses

Representations and targets. We test whether response information in the learned drug latent is accessible to a linear predictor on unseen compounds. We freeze representations from PerturbDiff models trained with the native objective and direct condition-mean response supervision, comparing chemistry (C), target identity (T), UniPert (U), and concatenation-based CTU fusion over three seeds. All four representations use response supervision; this comparison does not isolate its contribution. For each drug, observed control-relative responses are averaged across the four doses within each of A549, K562, and MCF7, then concatenated across cell lines. Each response vector $\mathbf{r}_p$ therefore contains $3 \times 1{,}621 = 4{,}863$ entries. We centre these vectors by the mean over the 157 training drugs, applying the same reference to the 30 held-out drugs:

$$\widetilde{\mathbf{r}}_p = \mathbf{r}_p - \frac{1}{|\mathcal{P}_{\text{train}}|} \sum_{p' \in \mathcal{P}_{\text{train}}} \mathbf{r}_{p'}. \tag{20}$$

Probe and evaluation. For each representation and seed, a ridge regression maps the frozen, unit-normalised drug latent to the centred response. On a training subset $\mathcal{S}$, we fit

$$(\widehat{\mathbf{W}}, \widehat{\mathbf{b}}) = \arg\min_{\mathbf{W},\mathbf{b}} \sum_{p \in \mathcal{S}} \|\mathbf{W}\mathbf{z}_p + \mathbf{b} - \widetilde{\mathbf{r}}_p\|_2^2 + \alpha \|\mathbf{W}\|_F^2. \tag{21}$$

The regularisation parameter is selected by `RidgeCV` using only the fitting subset, over 13 logarithmically spaced values from 10^{-3} to 10^{3}. A single fitted map predicts all 30 held-out drugs; the

Table 7: **Tahoe-100M (A549, RKO, Hs766T), ChemCPA: every drug representation under both objectives.** 683 held-out conditions from 76 drugs; mean over three seeds with sample SD as subscript. All rows, including the perturbation mean, subtract the training-control mean of the condition's cell line (controls pooled across plates). Rank is the within-cell-line retrieval rank over all genes (lower is better). Bold marks the best value within each objective.

	DEG50 residual		DEG20 residual		
Drug representation	Pearson	Direction	Pearson	Direction	Rank ↓
Perturbation mean	0.538	0.729	0.566	0.748	1.000
Native objective					
Chemistry (C)	$0.419_{\pm 0.026}$	$0.687_{\pm 0.008}$	$0.449_{\pm 0.028}$	$0.698_{\pm 0.009}$	$0.401_{\pm 0.009}$
Target identity (T)	$0.442_{\pm 0.016}$	$0.688_{\pm 0.001}$	$0.480_{\pm 0.021}$	$0.700_{\pm 0.002}$	$0.394_{\pm 0.005}$
UniPert (U)	$0.421_{\pm 0.028}$	$0.689_{\pm 0.005}$	$0.457_{\pm 0.033}$	$0.702_{\pm 0.008}$	$0.401_{\pm 0.005}$
C+T, concatenation	$0.424_{\pm 0.020}$	$0.688_{\pm 0.003}$	$0.464_{\pm 0.021}$	$0.698_{\pm 0.008}$	$0.403_{\pm 0.027}$
C+T+U, concatenation	$0.446_{\pm 0.030}$	$0.694_{\pm 0.007}$	$0.481_{\pm 0.027}$	$0.706_{\pm 0.008}$	$0.405_{\pm 0.020}$
C+T+U, self-attention	$0.454_{\pm 0.025}$	$0.689_{\pm 0.010}$	$0.494_{\pm 0.023}$	$0.699_{\pm 0.013}$	$\mathbf{0.382_{\pm 0.011}}$
C+T+U, MoE λ=10^{-4}	$\mathbf{0.463_{\pm 0.037}}$	$\mathbf{0.695_{\pm 0.009}}$	$\mathbf{0.499_{\pm 0.041}}$	$\mathbf{0.706_{\pm 0.010}}$	$0.402_{\pm 0.010}$
C+T+U, MoE λ=10^{-3}	$0.431_{\pm 0.034}$	$0.689_{\pm 0.006}$	$0.469_{\pm 0.038}$	$0.698_{\pm 0.005}$	$0.398_{\pm 0.013}$
C+T+U, MoE λ=10^{-2}	$0.455_{\pm 0.013}$	$0.693_{\pm 0.003}$	$0.492_{\pm 0.018}$	$0.702_{\pm 0.003}$	$0.407_{\pm 0.010}$
With response supervision					
Chemistry (C)	$\mathbf{0.522_{\pm 0.060}}$	$\mathbf{0.711_{\pm 0.016}}$	$\mathbf{0.546_{\pm 0.058}}$	$0.734_{\pm 0.023}$	$0.311_{\pm 0.046}$
Target identity (T)	$0.449_{\pm 0.118}$	$0.680_{\pm 0.034}$	$0.471_{\pm 0.125}$	$0.704_{\pm 0.043}$	$0.309_{\pm 0.017}$
UniPert (U)	$0.499_{\pm 0.033}$	$0.696_{\pm 0.024}$	$0.526_{\pm 0.040}$	$0.719_{\pm 0.026}$	$\mathbf{0.298_{\pm 0.018}}$
C+T, concatenation	$0.421_{\pm 0.132}$	$0.674_{\pm 0.047}$	$0.453_{\pm 0.138}$	$0.695_{\pm 0.056}$	$0.330_{\pm 0.059}$
C+T+U, concatenation	$0.514_{\pm 0.040}$	$0.711_{\pm 0.004}$	$0.543_{\pm 0.039}$	$\mathbf{0.736_{\pm 0.011}}$	$0.316_{\pm 0.035}$
C+T+U, self-attention	$0.494_{\pm 0.022}$	$0.696_{\pm 0.008}$	$0.523_{\pm 0.029}$	$0.720_{\pm 0.008}$	$0.313_{\pm 0.013}$
C+T+U, MoE λ=10^{-4}	$0.507_{\pm 0.042}$	$0.709_{\pm 0.006}$	$0.535_{\pm 0.037}$	$0.730_{\pm 0.012}$	$0.311_{\pm 0.027}$
C+T+U, MoE λ=10^{-3}	$0.445_{\pm 0.065}$	$0.680_{\pm 0.029}$	$0.476_{\pm 0.071}$	$0.701_{\pm 0.032}$	$0.324_{\pm 0.021}$
C+T+U, MoE λ=10^{-2}	$0.480_{\pm 0.059}$	$0.690_{\pm 0.022}$	$0.509_{\pm 0.067}$	$0.717_{\pm 0.021}$	$0.324_{\pm 0.013}$

Table 8: **sci-Plex3: expression-space metrics and full-panel residual metrics.** Expression metrics are near ceiling for every method, including the control mean, which predicts no response at all; this is why the main text scores control-referenced residuals on DEGs. DEG overlap is the fraction of the observed top-50 DEGs recovered among the predicted top-50. Mean over seeds with sample SD as subscript; residual metrics over all genes were not exported for the latent-32 ChemCPA native runs (—).

	Expression, all genes		Expression, DEG50		Residual, all genes	
Method	Pearson	R^2	Pearson	R^2	Pearson	DEG overlap
Control mean	0.955	0.908	0.976	0.847	—	0.022
Perturbation mean	0.959	0.915	0.977	0.856	0.324	0.338
CondOT	$0.957_{\pm 0.001}$	$0.911_{\pm 0.001}$	$0.974_{\pm 0.001}$	$0.850_{\pm 0.003}$	$0.170_{\pm 0.054}$	$0.307_{\pm 0.008}$
Biolord	$0.644_{\pm 0.010}$	$-1.661_{\pm 0.339}$	$0.705_{\pm 0.008}$	$-1.110_{\pm 0.245}$	$0.030_{\pm 0.003}$	$0.212_{\pm 0.007}$
ChemCPA	$0.956_{\pm 0.000}$	$0.910_{\pm 0.001}$	$0.974_{\pm 0.002}$	$0.837_{\pm 0.005}$	—	—
PerturbDiff	$0.957_{\pm 0.001}$	$0.912_{\pm 0.001}$	$0.974_{\pm 0.000}$	$0.852_{\pm 0.002}$	$0.204_{\pm 0.017}$	$0.258_{\pm 0.016}$
MoRe + ChemCPA	$0.956_{\pm 0.000}$	$0.911_{\pm 0.002}$	$0.973_{\pm 0.003}$	$0.851_{\pm 0.006}$	$0.184_{\pm 0.027}$	$0.287_{\pm 0.026}$
MoRe + PerturbDiff	$0.958_{\pm 0.001}$	$0.910_{\pm 0.003}$	$0.967_{\pm 0.001}$	$0.866_{\pm 0.004}$	$0.213_{\pm 0.011}$	$0.270_{\pm 0.010}$

fitting drugs are not selected as neighbours of each query. For each test drug, we compute Pearson correlation over the concatenated response vector and then average over drugs.

Results. Fusion achieves the highest mean readout correlation across the evaluated fitting-set sizes. With all 157 training drugs, correlations are 0.38 for fusion, 0.33 for target identity, 0.31 for UniPert, and 0.09 for chemistry. This supports the presence of linearly accessible response information that

Table 9: **Single-cell distribution metrics on sci-Plex3** (lower is better). E-distance, multivariate energy distance; W1, mean gene-wise one-dimensional Wasserstein-1 distance; computed over all genes and over each condition's DEG50. Mean over three seeds with sample SD as subscript; bold marks the best value per backbone.

		All genes		DEG50	
Backbone	Configuration	E-distance ↓	W1 ↓	E-distance ↓	W1 ↓
PerturbDiff	Native, RDKit	$0.537_{\pm 0.003}$	$0.0250_{\pm 0.0002}$	$0.434_{\pm 0.009}$	$0.1054_{\pm 0.0007}$
	Response supervision, RDKit	$0.565_{\pm 0.009}$	$0.0257_{\pm 0.0003}$	$0.435_{\pm 0.006}$	$0.1019_{\pm 0.0013}$
	Response supervision, C+T+U concatenation (MoRe)	$0.566_{\pm 0.005}$	$0.0257_{\pm 0.0003}$	$0.420_{\pm 0.003}$	$0.0979_{\pm 0.0022}$
ChemCPA	Native, RDKit	$1.966_{\pm 0.013}$	$0.0779_{\pm 0.0004}$	$1.206_{\pm 0.006}$	$0.2713_{\pm 0.0012}$
	Response supervision, RDKit	$1.988_{\pm 0.005}$	$0.0791_{\pm 0.0001}$	$1.212_{\pm 0.008}$	$0.2751_{\pm 0.0015}$
	Response supervision, C+T+U concatenation (MoRe)	$1.993_{\pm 0.023}$	$0.0792_{\pm 0.0008}$	$1.215_{\pm 0.010}$	$0.2752_{\pm 0.0025}$

Table 10: **Learned fusion versus averaging the outputs of single-view models** (sci-Plex3, response supervision). Output averages take the mean of the predicted condition means of independently trained single-view models within each seed; no weights are fitted. The best single view is chosen on DEG50 Pearson. Mean over three seeds with sample SD as subscript.

		DEG50 residual		DEG20 residual	
Backbone	Predictor	Pearson	Direction	Pearson	Direction
PerturbDiff	Output average, C+T	$0.281_{\pm 0.023}$	$0.599_{\pm 0.012}$	$0.303_{\pm 0.016}$	$0.624_{\pm 0.012}$
	Output average, C+T+U	$\mathbf{0.294_{\pm 0.020}}$	$0.603_{\pm 0.012}$	$0.317_{\pm 0.014}$	$0.631_{\pm 0.009}$
	Learned fusion, C+T+U (MoRe)	$0.286_{\pm 0.015}$	$\mathbf{0.628_{\pm 0.006}}$	$\mathbf{0.326_{\pm 0.017}}$	$\mathbf{0.658_{\pm 0.002}}$
ChemCPA	Output average, C+T	$0.242_{\pm 0.009}$	$0.552_{\pm 0.036}$	$0.260_{\pm 0.012}$	$0.575_{\pm 0.042}$
	Output average, C+T+U	$0.249_{\pm 0.010}$	$0.552_{\pm 0.044}$	$0.267_{\pm 0.006}$	$0.574_{\pm 0.052}$
	Learned fusion, C+T+U (MoRe)	$\mathbf{0.249_{\pm 0.029}}$	$\mathbf{0.560_{\pm 0.046}}$	$\mathbf{0.268_{\pm 0.024}}$	$\mathbf{0.584_{\pm 0.046}}$

generalises to unseen compounds. The highest mean does not establish statistically significant superiority over every single view, and the probe does not identify how the native predictor uses this information. In particular, it complements rather than establishes a nearest-neighbour explanation of the predictive gains.

Table 11: **Learned fusion versus averaging the outputs of single-view models on Tahoe-100M.** All models use response supervision and cell-line-specific controls. Output averages combine predicted condition means from independently trained single-view models of the same seed with equal weights. Values are means over three seeds with sample SD as subscript; bold marks the best mean per backbone and metric.

		DEG50 residual		DEG20 residual		
Backbone	Predictor	Pearson ↑	Direction ↑	Pearson ↑	Direction ↑	Rank ↓
PerturbDiff	Output average, C+T	$0.751_{\pm 0.002}$	$0.766_{\pm 0.001}$	$0.761_{\pm 0.002}$	$0.803_{\pm 0.002}$	$0.141_{\pm 0.001}$
	Output average, C+T+U	$\mathbf{0.754}_{\pm 0.001}$	$\mathbf{0.766}_{\pm 0.001}$	$\mathbf{0.764}_{\pm 0.002}$	$\mathbf{0.803}_{\pm 0.001}$	$\mathbf{0.140}_{\pm 0.002}$
	Learned fusion, C+T+U (MoRe)	$0.749_{\pm 0.004}$	$0.761_{\pm 0.002}$	$0.758_{\pm 0.005}$	$0.796_{\pm 0.004}$	$0.142_{\pm 0.004}$
ChemCPA	Output average, C+T	$0.495_{\pm 0.082}$	$0.702_{\pm 0.022}$	$0.519_{\pm 0.082}$	$0.725_{\pm 0.029}$	$0.300_{\pm 0.024}$
	Output average, C+T+U	$0.500_{\pm 0.059}$	$0.703_{\pm 0.016}$	$0.525_{\pm 0.061}$	$0.725_{\pm 0.024}$	$\mathbf{0.292}_{\pm 0.018}$
	Learned fusion, C+T+U (MoRe)	$\mathbf{0.514}_{\pm 0.040}$	$\mathbf{0.711}_{\pm 0.004}$	$\mathbf{0.543}_{\pm 0.039}$	$\mathbf{0.736}_{\pm 0.011}$	$0.316_{\pm 0.035}$

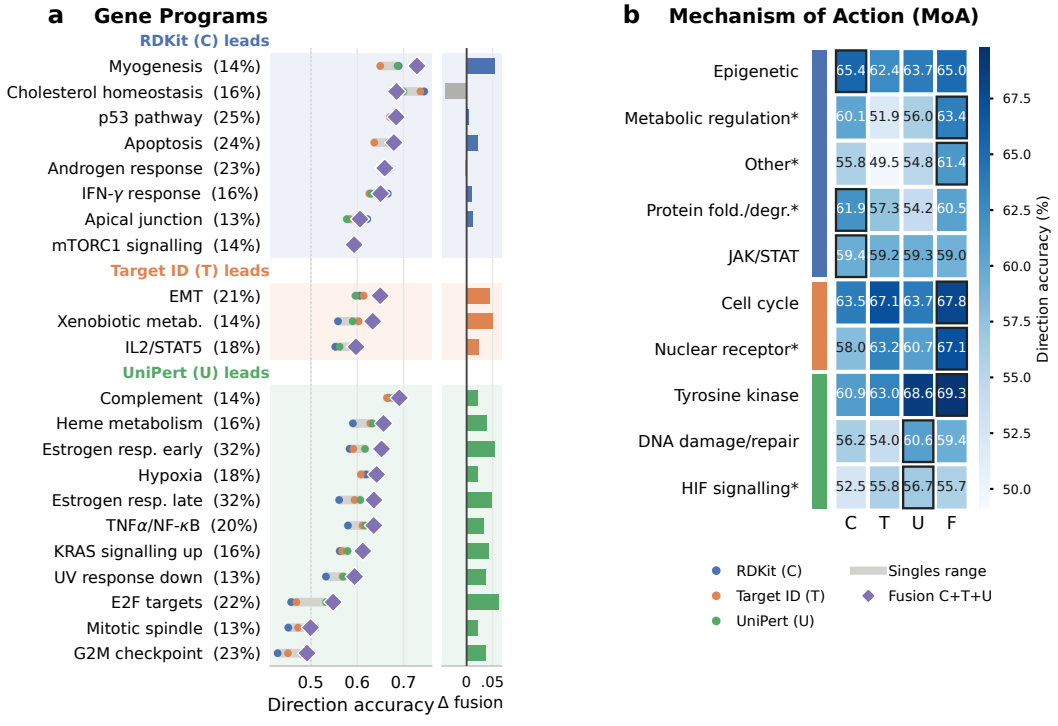

Figure 10: **All gene programmes and MoA classes, PerturbDiff** (full version of Figure 4). Direction accuracy on DEG50 with response supervision. **(a)** Every Hallmark programme, grouped by leading single view; percentages give the programme's coverage of the gene panel; right bars, fusion minus the mean of the single views. **(b)** Every MoA class; the best value per row is boxed.

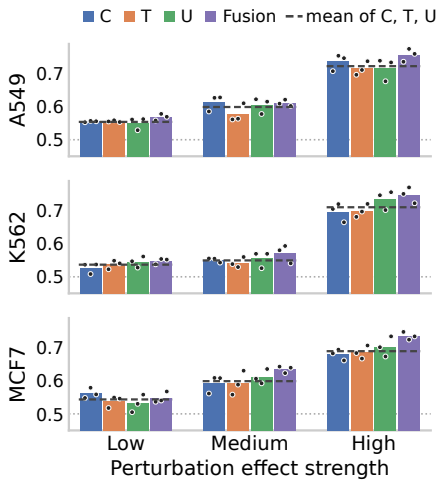

Figure 11: **Fusion gain by perturbation effect strength, per cell line** (PerturbDiff, sci-Plex3; per-cell-line version of Figure 5). Direction agreement on the top 50 DEGs by effect-strength tertile. Bars, seed means; dots, individual seeds; dashed line, mean of the three single views; dotted line, chance.

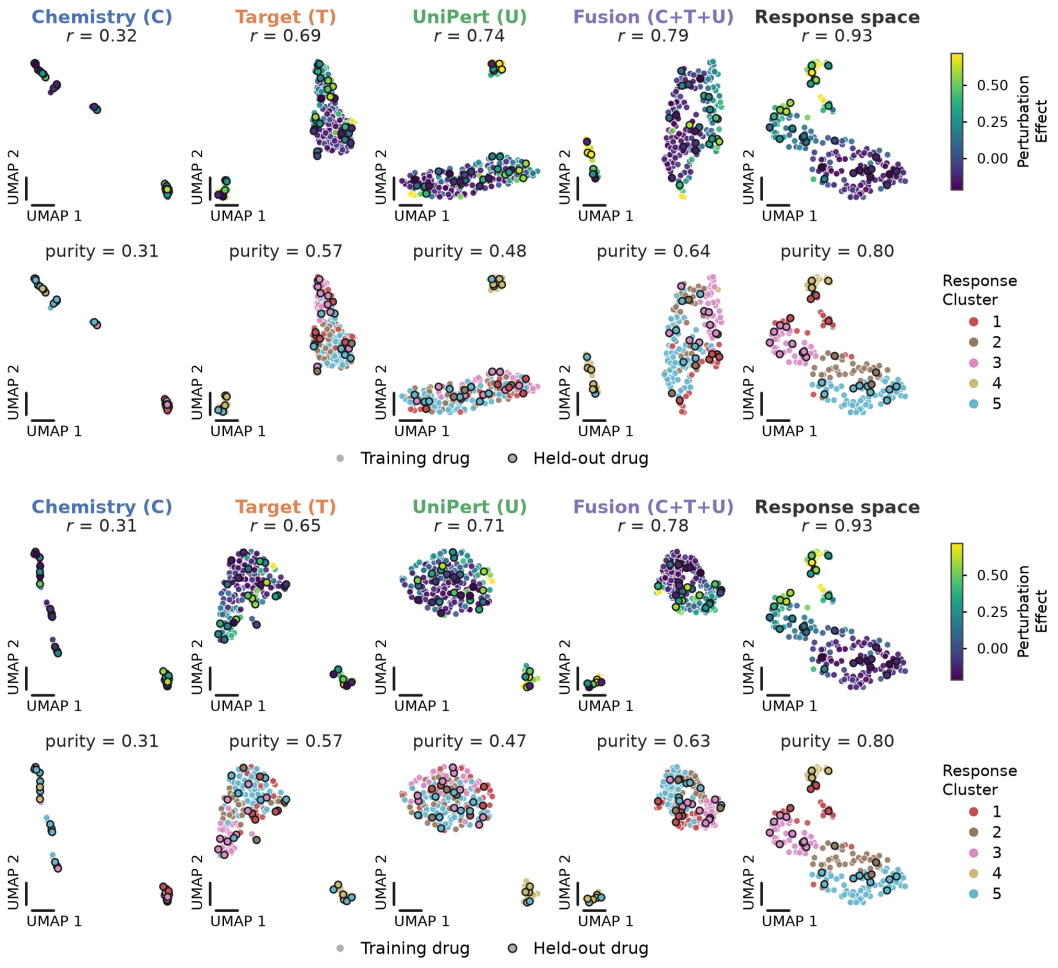

Figure 12: **Drug latents for PerturbDiff seeds 1 (top block) and 2 (bottom block)**, as Figure 7. Scores above each panel are for the seed shown; Figure 7 reports the mean over seeds.

Table 12: **Response organisation of learned drug latents.** All models use response supervision; subscripts report sample SD over three seeds. Neighbour response similarity averages response cosine similarity over each drug's k nearest latent neighbours among the other 186 drugs (chance: 0.087). Held-out ρ is the Spearman correlation between latent and response similarity across held-out–training pairs. Effect r correlates each drug's response magnitude ($\log_{10}$ residual L2 norm) with the mean magnitude of its ten nearest neighbours. Cluster and MoA purity measure the fraction of these neighbours sharing its response cluster (k-means, five clusters; chance: 0.25) or annotated MoA (chance: 0.13). PC1 reports the fraction of latent variance explained by the first principal component.

	Neighbour response similarity			Held-out	Effect	Cluster	MoA	
Latent	$k=1$	$k=10$	$k=30$	ρ	r	purity	purity	PC1
C	$0.128_{\pm 0.001}$	$0.141_{\pm 0.003}$	$0.130_{\pm 0.000}$	$-0.081_{\pm 0.000}$	$0.325_{\pm 0.023}$	$0.309_{\pm 0.003}$	$0.175_{\pm 0.002}$	89%
T	$0.456_{\pm 0.007}$	$0.383_{\pm 0.007}$	$0.322_{\pm 0.006}$	$0.342_{\pm 0.037}$	$0.683_{\pm 0.025}$	$0.576_{\pm 0.005}$	$0.369_{\pm 0.002}$	30%
U	$0.425_{\pm 0.026}$	$0.325_{\pm 0.009}$	$0.263_{\pm 0.011}$	$0.369_{\pm 0.036}$	$0.717_{\pm 0.024}$	$0.468_{\pm 0.012}$	$0.302_{\pm 0.014}$	24%
Fusion	$0.483_{\pm 0.012}$	$0.414_{\pm 0.002}$	$0.350_{\pm 0.005}$	$0.440_{\pm 0.024}$	$0.782_{\pm 0.011}$	$0.625_{\pm 0.017}$	$0.286_{\pm 0.002}$	54%

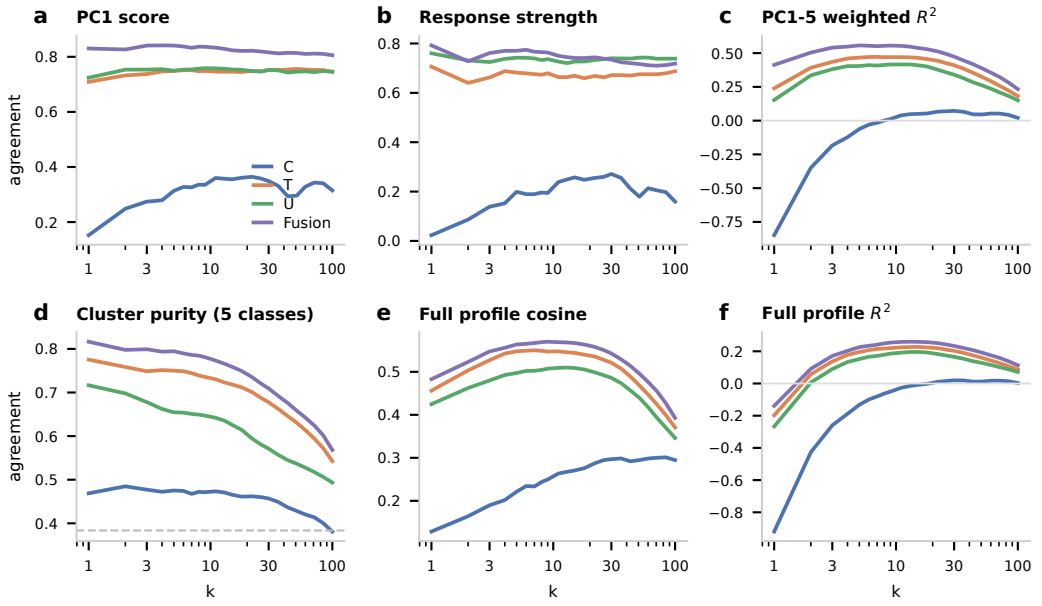

Figure 13: **Neighbourhood agreement under six definitions of response similarity** (PerturbDiff, all 187 drugs, three seeds). **(a)** Leading principal component of the response. **(b)** Response magnitude. **(c)** Variance-weighted R^2 on the first five components. **(d)** Fraction of neighbours in the same response cluster (hierarchical clustering into five clusters; dashed, chance). **(e)** Cosine between a drug's full response profile and the mean profile of its neighbours. **(f)** R^2 of that reconstruction.