GENERALIZATION TO UNSEEN PERTURBATIONS VIA MULTIPLE BIOLOGICAL PRIORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Predicting cellular responses to genetic and chemical perturbations can reduce the cost of experimental screening. A key challenge is generalization to unseen perturbations, since perturbation identities alone do not specify how their responses relate to those observed during training. However, biological knowledge from existing research provides multiple priors that can connect observed and unseen perturbations through shared molecular and functional properties. We thus formulate unseen perturbation prediction as a generalization problem over multiple biological priors, and introduce a corresponding hypothesis space that accommodates different priors and predictor architectures. Using this methodology, we design a two-stage learning algorithm, connected to late fusion, stacked generalization, and boosting, that first trains a response predictor for each prior and then integrates their predictions in response space through a shared integration function. Across two chemical and two genetic perturbation datasets spanning three cell lines, our approach outperforms single-prior predictors and representation-level fusion methods across various metrics, providing an effective way to combine multiple priors.

1 INTRODUCTION

Characterizing cellular responses to genetic or chemical perturbations can reveal gene function, quantify drug effects, and identify potential therapeutic targets (Replogle et al., 2022; Srivatsan et al., 2020; Behan et al., 2019). Modern high-throughput assays have enabled these responses to be measured at large scale (Dixit et al., 2016; Srivatsan et al., 2020). However, experimentally testing every possible perturbation across relevant cellular contexts remains infeasible. This motivates computational methods that learn from *observed* perturbations, whose responses are available during training, to predict the responses of *unseen* perturbations.

Predicting the response to an unseen perturbation requires relating it to observed perturbations. Perturbation identity alone does not provide such relations because a gene or compound identifier is a categorical label with no inherent notion of similarity to previously observed perturbations. However, these genes and compounds have been characterized in other contexts, providing biological priors that relate unseen and observed perturbations through shared molecular or functional features. Throughout this paper, we use *priors* to refer to external biological information that provides structured relationships among perturbations. Existing methods for unseen perturbation prediction share a common principle of leveraging priors. They construct priors from molecular structures (Hetzel et al., 2022), gene regulatory networks (Kamimoto et al., 2023), curated knowledge graphs (Roohani et al., 2024), external perturbation databases (He et al., 2025; Littman et al., 2025), and foundation models pretrained on large-scale biological data (Cole et al., 2026).

A single perturbation can be represented by multiple such priors, each capturing a different aspect of its biology and providing complementary information about its response. Previous work suggests that the prior most predictive of the response can vary across perturbations (He et al., 2025; Cole et al., 2026), motivating methods that integrate multiple sources of information. Existing approaches typically combine prior-derived representations before mapping them to the response, a strategy we refer to as *representation-level fusion* (Littman et al., 2025; Cole et al., 2026; Guo et al., 2026). However, it remains unclear whether this is the most effective way to combine multiple priors. We instead propose *response-level fusion*, in which each prior is first mapped to a predicted response,

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.

and the resulting predictions are then integrated. This formulation allows the fusion model to use patterns across prior-specific predictions to learn how their errors should be corrected. Our contributions are as follows:

- We formulate unseen perturbation prediction as a generalization problem over multiple biological priors, connecting it to the framework of deep learning generalization (Kawaguchi et al., 2022), and introduce a corresponding hypothesis space that accommodates heterogeneous priors and predictor architectures.
- We propose a two-stage learning algorithm that first trains a response predictor for each prior and then combines their predictions using an integration rule shared across response genes and perturbations. We consider integration rules ranging from parameter-free statistics to a learned nonlinear function. We relate this approach to late fusion, stacked generalization, and the error-correction principle of boosting.
- Across two chemical and two genetic perturbation datasets, we show that this learning strategy significantly improves unseen-perturbation response prediction over single-prior predictors and representation-level fusion on multiple evaluation metrics, providing an effective approach for integrating biological priors.

2 Related Work

Generalization for Unseen Perturbations. Generalization in deep learning studies how predictors selected from a *hypothesis class* using observed training data perform on a target distribution that may contain unseen test examples (Kawaguchi et al., 2022). The difference between training and target risk, defined as the expected prediction loss over the corresponding data distribution, is commonly referred to as the generalization gap. A small generalization gap is often desired. Under suitable assumptions, Kawaguchi et al. (2018) relate this gap to how well the training data cover the target distribution in a chosen *representation* space and how much the prediction loss varies over that space. This perspective highlights how the choice of representation can affect generalization.

Unseen perturbation prediction poses a particular challenge under this perspective. Perturbation identities (genes or compounds) are categorical identifiers and, by themselves, lie in a representation space with no geometry for relating an unseen perturbation to those observed during training. For example, under a one-hot identity representation, an unseen perturbation is equally distant from every observed perturbation. Biological priors instead embed perturbations in informative feature spaces shaped by molecular or functional properties, allowing observed perturbations to provide information about unseen ones. Different priors may provide complementary information, motivating us to introduce a hypothesis class with multi-prior representations in Section 3.

Unseen Perturbation Prediction. Previous works on unseen perturbation prediction leverage different biological priors in various ways. For chemical perturbations, molecular structure can be encoded using fingerprints or pretrained molecular representations and used to condition the response predictor, e.g., in a variational autoencoder (Hetzel et al., 2022) or an invertible neural network (Yu et al., 2025). For genetic perturbations, regulatory relationships describing how perturbation effects propagate or functional relationships describing which genes share similar pathways can be encoded in a graph structure used by the response predictor, which can take the form of a linear graphical model (Kamimoto et al., 2023) or a graph neural network (Roohani et al., 2024). Other approaches use features derived from external perturbation resources in a modular neural network (He et al., 2025) or embeddings learned by pretrained biological foundation models in a simple nearest neighbor or multi-layer perceptron (MLP) response predictor (Cole et al., 2026).

Fusion Methods for Multiple Priors. When multiple priors are available, they induce different representations of the same perturbation that can provide complementary information. This connects to multimodal learning, where different modalities provide complementary information that can be fused for prediction of a downstream target (Baltrušaitis et al., 2018; Zhao et al., 2024). Multimodal fusion algorithms are typically classified into three classes based on when fusion occurs. Specifically, early fusion combines inputs before feature extraction, intermediate fusion combines learned representations, and late fusion combines modality-specific predictions of the same downstream tar-

get (Zhao et al., 2024). We collectively refer to early and intermediate fusion as *representation-level fusion*, since both combine information before mapping it to the prediction target.

Similar fusion approaches can be extended to combine multiple priors, where each prior can be viewed as a modality and the prediction target is the perturbation response. Recent multi-prior approaches primarily follow representation-level fusion, combining prior-derived representations through concatenation, attention, or other learned aggregation mechanisms before response prediction (Littman et al., 2025; Cole et al., 2026; Guo et al., 2026).

3 Problem Formulation

3.1 Problem setup

We now formally introduce the perturbation prediction problem. We consider perturbation data from single-cell expression assays. As the data usually contains multiple cells who received the same perturbation, we model the response to a perturbation as a distribution induced by cells receiving it. Each perturbation p corresponds to a distribution P^p over single-cell expression vectors $\boldsymbol{x}^p \in \mathbb{R}^d$, where d is the number of genes with measured expressions. The control distribution is denoted by $P^\emptyset$ over control expression vectors $\boldsymbol{x}^\emptyset$, and when clear from context, we write $\boldsymbol{x} := \boldsymbol{x}^\emptyset$.

Let $\mathcal{I}$ denote a set of perturbations and $\mathcal{I}_{\text{train}} \subset \mathcal{I}$ the perturbations whose responses are observed during training. We consider a training dataset $\mathcal{D} = \{(\boldsymbol{x}_i^{p_i}, p_i)\}_{i=1}^N$, where N is the number of observed cells, $p_i \in \mathcal{I}_{\text{train}} \cup \{\emptyset\}$ is the perturbation received by cell i, and $\boldsymbol{x}_i^{p_i}$ is the expression vector of cell i from P^{p_i}. We will use unbolded letter $x_{i,j}^{p_i}$ (or x_j^p) to denote the j-th coordinate of $\boldsymbol{x}_i^{p_i}$ (or $\boldsymbol{x}^p$). Given $\mathcal{D}$, our goal is to predict the response P^p for an unseen perturbation $p \in \mathcal{I} \setminus \mathcal{I}_{\text{train}}$.

We consider predictors that use the control distribution $P^\emptyset$ as anchor. In particular, let

$$f : \mathbb{R}^d \times \mathcal{I} \to \mathbb{R}^d \tag{1}$$

be a predictor that maps a control cell $\boldsymbol{x}$ and a perturbation p to a predicted response $f(\boldsymbol{x}, p)$. Applying f to all control cells from $P^\emptyset$ induces the predicted response distribution $\widehat{P}_f^p$. We evaluate the prediction for perturbation p using

$$\ell(f, p) = \mathsf{D}\left(\widehat{P}_f^p, P^p\right), \tag{2}$$

where we consider both *distribution-level* evaluations with D being the maximum mean discrepancy (MMD) (Gretton et al., 2012), and *mean-level* evaluations with D being a function of $\mathbb{E}_{x\sim\widehat{P}_f^p}[x]$ and $\mathbb{E}_{x\sim P^p}[x]$ (see Appendix B.2)

3.2 Multi-prior prediction

Section 2 motivates representing perturbations using multiple biological priors. Suppose that each perturbation p has M such priors,

$$h_m(p) \in \mathbb{R}^{q_m}, \qquad m = 1, \ldots, M,$$

where $h_m : \mathcal{I} \to \mathbb{R}^{q_m}$ denotes the vector representation derived from the m-th prior. Note that we assume below that this vector representation h_m of the m-th prior exists and is sufficient for downstream response prediction. This assumption covers most of the priors discussed in Section 2. For graph-neural-network-based approaches (e.g., Roohani et al. (2024)), the assumption is valid by interpreting h_m as the intermediate graph-based representation learned by the model. For simplicity, we consider the case where this vector representation is given prior to training, but the method easily extends to priors outside of this case.

These priors can be fused at two stages. Representation-level fusion combines the representations $h_1(p), \ldots, h_M(p)$ into a single input before predicting the response, and is the approach taken by most previous works (Littman et al., 2025; Cole et al., 2026). We instead propose response-level fusion, in which each prior first produces its own prediction of the response and these predictions are then combined. Response-level fusion circumvents the need to align priors of different dimensionality or semantics into a common representation space, since the alignment happens in the response space, which is shared by construction.

Formally, let $\mathcal{F}_m$ be a class of prior-specific predictors $f_m : \mathbb{R}^d \times \mathbb{R}^{q_m} \to \mathbb{R}^d$, and let $\mathcal{G}$ be a class of integration functions $g : \mathbb{R}^{Md} \to \mathbb{R}^d$. The corresponding *hypothesis class* of the response-level fusion predictor f in Eq. (1) we consider is thus

$$\mathcal{F} = \{(\boldsymbol{x}, p) \mapsto g(f_1(\boldsymbol{x}, h_1(p)), \ldots, f_M(\boldsymbol{x}, h_M(p))) : f_m \in \mathcal{F}_m,\ g \in \mathcal{G}\}\,. \tag{3}$$

This decomposition reveals two separate learning problems, where each f_m learns how one perturbation representation maps to the response, while g learns how the resulting predictions should be combined.

4 Learning with Response-level Fusion

We now describe our two-stage algorithm for learning a response-level fusion predictor $f \in \mathcal{F}$.

Single-prior Predictor. For each prior m, we fit a predictor $\hat{f}_m \in \mathcal{F}_m$ on the observed perturbations, with p as either a gene target or a compound at a specific dosage and $h_m(p)$ as the perturbation representation. This predictor and fitting process can take any form compatible with h_m. For example, among the approaches described in the second paragraph of Section 2, MORPH (He et al., 2025) and the MLP regressor (Cole et al., 2026) can be used to incorporate any such prior. All M predictors share the same training data $\mathcal{D}$ (Section 3) but have independent parameters. Given a control cell $\boldsymbol{x}$ and perturbation p, the prediction from prior m is

$$\hat{\boldsymbol{x}}^p_m = \hat{f}_m\big(\boldsymbol{x}, h_m(p)\big) \in \mathbb{R}^d, \qquad m = 1, \ldots, M. \tag{4}$$

Integration Function. The integration function g combines the M single-prior predictions into a final response. We apply the same rule to every gene and perturbation: for each gene j, a shared map $\gamma : \mathbb{R}^M \to \mathbb{R}$ receives the M predicted values for that gene and returns one,

$$\hat{x}^p_j = \gamma\big(\hat{x}^p_{1,j}, \ldots, \hat{x}^p_{M,j}\big), \qquad j = 1, \ldots, d, \tag{5}$$

which defines the fused prediction $\hat{\boldsymbol{x}}^p = g(\hat{\boldsymbol{x}}^p_1, ..., \hat{\boldsymbol{x}}^p_M)$. We consider three classes of γ of increasing complexity (see Appendix B.5 for details):

- *Learning-free aggregation.* The mean, median, or trimmed mean of the M predictions.
- *Learned weighted mean.* $\hat{x}^p_j = \sum_{m=1}^M w_m\, \hat{x}^p_{m,j}$ with $\boldsymbol{w}$ on the simplex, which learns one global reliability weight per prior.
- *Shared MLP.* A small MLP γ_η with learnable weights η, which results in a nonlinear combination rule.

Training and Inference. We first train the single-prior predictors $\hat{f}_m$ independently. We then fit γ on the predictions of the $\hat{f}_m$ for the observed perturbations during training. See Appendix B.5 for details. At inference, an unseen perturbation is represented by each prior h_m, each $\hat{f}_m$ produces a response, and γ combines them into the final profile.

Finite-sample Motivation. Biological priors can be high-dimensional relative to the number of measured perturbations. Representation-level fusion requires learning to combine multiple high-dimensional priors (with total dimension $q = \sum_{m=1}^M q_m$) and predict directly from the resulting representation. On the other hand, response-level fusion divides the learning part into two steps, where it combines M prior-specific response predictions in a separate optimization step. In a stylized linear model in Appendix A, we show that response-level fusion has lower expected test error when the variance reduction from learning with M rather than q inputs outweighs the signal discarded by the prior-specific predictors. This illustrates a bias–variance tradeoff that may favor response-level fusion over representation-level fusion in finite-data regimes.

Connection to Previous Works. Our two-stage framework is related to the late-fusion principle of combining predictions from separately trained models, using either fixed aggregation rules or a learned fusion model (Baltrušaitis et al., 2018). The learned second-stage MLP is also related to

stacked generalization (Wolpert, 1992) and, more specifically, multi-view stacking, which combines predictions from models trained on different feature representations (van Loon et al., 2020). It also mirrors the error-correction principle of boosting (Friedman, 2001), where the integrator can learn to correct errors of each single predictor.

5 Experiment Setup

We compare single-prior prediction, representation-level fusion, and response-level fusion for unseen-perturbation prediction across chemical and genetic perturbation datasets. Implementation details are provided in Appendix B.

5.1 Datasets and biological priors

Datasets. Our experiments use two chemical and two genetic single-cell perturbation screens. For chemical perturbations, we use two `SciPlex3` datasets on `K562` and `MCF7` cells (173,652 and 344,862 cells after preprocessing), comprising 188 compounds measured at four doses (Srivatsan et al., 2020). For genetic perturbations, we use the two `Replogle` datasets on `K562` and `RPE1` cells (300,998 and 239,802 cells), comprising 1,969 and 2,196 gene knockdowns, respectively, after preprocessing (Replogle et al., 2022). Each cell is represented as a 5,000-dimensional vector (observed variable $\boldsymbol{x}$) measuring the expressions of 5,000 highly variable genes. Full preprocessing details and dataset statistics are provided in Appendix B.1.

Biological Priors. We represent each perturbation using multiple biological priors. For the chemical screens, we use six representations: extended-connectivity fingerprints (`ECFP`) (Rogers & Hahn, 2010), contrastively trained molecule-assay embeddings (`CLAMP`) (Seidl et al., 2023), two SMILES-based molecular language-model embeddings (`ChemBERTa` and `MoLFormer`) (Ahmad et al., 2022)(Ross et al., 2022), molecule-text representations (`MolT5`) (Edwards et al., 2022), and 3D molecular representations (`Uni-Mol`) (Zhou et al., 2023). For the genetic screens, we use five representations: protein-sequence embeddings (`ESM-2`) (Lin et al., 2023), gene-dependency profiles (`DepMap`) (Tsherniak et al., 2017), text embeddings of gene descriptions from databases (`STRING`) (Szklarczyk et al., 2023), and two GenePT-derived text representations (Chen & Zou, 2024) with one based on NCBI gene summaries (`GenePT gene`) (Sayers et al., 2024) and the other on NCBI gene summaries combined with UniProt protein annotations (`GenePT protein`) (The UniProt Consortium et al., 2023). We also include a one-hot representation as a baseline. The complete set of priors and their dimensions is given in Appendix Table 3.

5.2 Evaluation strategy

We split the data into training and test by perturbation identity. For the chemical screens, we use four-fold cross-validation over compounds, such that all doses of a held-out compound are unseen during training. For the genetic screens, we use five-fold cross-validation over perturbations. Each comparison is restricted to the intersection of perturbations covered by all corresponding priors.

We evaluate each held-out perturbation on its top 50 differentially expressed genes (DEGs) relative to control, selected per perturbation. We report complementary metrics at the distribution and mean-response levels. Depending on which the *prediction setting* (described in the next section) is considered: at the distribution level, we report the MMD between predicted and observed cell populations; at the mean-response level, we report sign agreement (the fraction of genes whose direction of change is correctly predicted), RMSE between the predicted and observed responses, and $\mathrm{Pearson}_{\Delta}^{\mathrm{pert}}$ (Viñas Torné et al., 2026), which correlates predicted and observed response profiles after centering by the mean perturbed centroid rather than the control, reducing sensitivity to shared systematic effects. Full definitions are in Appendix B.2.

5.3 Prediction settings and compared methods

Prediction Settings. We evaluate all methods under three prediction settings that vary the response predictor backbone f_m and fitting process (see definitions in Section 4). In each setting, the same backbone is used to construct the single-prior predictors and the corresponding fusion methods. We

Table 1: Multi-prior fusion methods. We compare representation-level fusion, which combines prior representations before response prediction, and our proposed response-level fusion, which combines prior-specific predictions. Full implementation details are provided in Appendices B.4 and B.5.

Fusion Stage	Integration Method	Description of Integration Method
Representation	Concatenation	Concatenate all priors and pass into a single predictor
Representation	Soft-argmax selector	Project each prior to a common space and learn a near-discrete weighting over priors
Representation	Perceiver-style	Use cell-conditioned cross-attention to combine projected prior representations
Representation	PRESAGE-style	Combine global source weighting with perturbation-specific graph-attention pooling
Response	Mean	Mean of the prior-specific predictions
Response	Median	Median of the prior-specific predictions
Response	Trimmed mean	Remove the largest and smallest prediction before averaging
Response	Learned weighted mean	Learn a weighted avg over prior-specific predictions
Response	Shared MLP	Learn a shared function of the prior-specific predictions

consider: (i) minimizing distributional loss with MORPH as backbone; (ii) minimizing distributional loss with an MLP-based variational autoencoder (VAE) as backbone; and (iii) minimizing mean-response (pseudo-bulk) loss with MORPH. Settings (i)–(ii) vary the model architecture, and (i) and (iii) vary the prediction target in the fitting process. Within each setting, we fix the data splits across single- and multi-prior predictors. We report setting (i) in Table 2 and settings (ii) and (iii) in Appendix Tables 5 and 6, respectively. For (iii), as the model is not learned with a distributional-level loss, we report RMSE instead of MMD.

Learning-free Baselines. To calibrate the metrics, we include two learning-free baselines. *Random control* predicts randomly sampled control cells (no perturbation effect), and *random perturbed* predicts randomly sampled non-held out perturbed cells (a generic response that ignores perturbation identity). For the mean-response prediction setting (iii) above, we analogously use the control mean and the mean response over training perturbations.

Single-prior Predictors. For each prior, we train an independent response predictor using the backbone associated with the corresponding prediction setting. Since MORPH (He et al., 2025) was originally developed for genetic perturbations, we adapt it to the chemical setting by encoding chemical perturbations and scaling the latent perturbation representation by the log-transformed dose. We also use an MLP-based VAE as an alternative model architecture which still retains a probabilistic generative formulation for distribution-level prediction. For mean-response prediction, we train the model directly on pseudo-bulk perturbation responses using a corresponding loss function. See Appendix B.3 for details.

Multi-prior Predictors. We compare two classes of fusion methods: representation-level fusion, which combines priors before response prediction, and our proposed response-level fusion, which combines prior-specific predictions. Since the existing representation-level fusion method, PRESAGE (Littman et al., 2025), predicts mean responses rather than full single-cell distributions, we benchmark the original method in setting (iii) and implement a PRESAGE-style fusion variant for the distributional settings (i)–(ii). Table 1 summarizes the methods considered, and implementation details are provided in Appendices B.4 and B.5.

6 Experiment Results

6.1 The best prior varies across perturbations

Single-prior predictor performance varies across priors and datasets (Table 2). On the genetic screens, `DepMap` is the strongest prior on all reported metrics, whereas on the chemical screens no prior dominates across cell lines and metrics. To assess whether different priors are useful for different perturbations, we construct a per-perturbation oracle that selects the best single-prior pre-

dictor for each test perturbation post-hoc. On all four datasets, this oracle improves over the best single prior on every metric, reducing MMD by 9.1–18.0%, increasing sign accuracy by 2.4–4.9 percentage points, and improving $\mathrm{Pearson}_{\Delta}^{\mathrm{pert}}$ by 0.137–0.254. This gap indicates that the most informative prior can vary across perturbations, motivating methods that combine multiple priors.

6.2 Response-level fusion effectively combines multiple priors

Existing fusion methods typically combine priors at the representation level, before response prediction. We therefore first evaluate representation-level fusion and then compare it with our proposed response-level fusion.

Response-level fusion provides more consistent gains than representation-level fusion. Representation-level fusion yields mixed results relative to the strongest single prior, with limited benefits on chemical screens, a lack of consistent gains from more elaborate fusion architectures, and only sporadic improvements from concatenation on genetic screens. In contrast, when comparing the best-performing method within each class of models, response-level fusion improves on the strongest single-prior method in all 12 dataset–metric comparisons. It also outperforms the best representation-level fusion method in 10 of 12 comparisons and ties in one at the reported precision (Table 2).

The benefit of response-level fusion is apparent even with nonparametric aggregation (mean, median, and trimmed mean). The median, for example, outperforms *both* the best single-prior model and the best representation-level fusion method in 7 of 12 dataset–metric comparisons, while improving MMD relative to the best single prior by 0.6%–9.7% across all four datasets. By comparison, the best representation-level fusion improves MMD over the best single prior only on the two genetic screens (2.4% and 5.5%), and is 0.7%–2.1% *worse* on the two chemical screens. Learning a nonlinear combination rule further strengthens these gains, as the shared MLP matches or exceeds the strongest single-prior and representation-level alternatives in 11 of 12 comparisons, improving MMD over the best single prior by 2.4%–37.4% across all datasets.

Response-level fusion remains effective across prediction settings. We next test the robustness of response-level fusion to changes in the underlying prediction setup. We consider two complementary variants, replacing the response-prediction architecture and changing the prediction target, respectively. Response-level fusion remains consistently effective in both settings. Considering the best-performing method within each class, response-level fusion outperforms both the strongest single-prior model and the strongest representation-level fusion method in 11 of 12 dataset–metric comparisons with an MLP-based VAE response predictor (Appendix Table 5). The same pattern holds when changing the prediction target from the full response distribution to the mean response: response-level fusion is better in 11 of 12 comparisons and matches the best representation-level method in the remaining one (Appendix Table 6).

The mean-response setting enables direct comparison with PRESAGE on the genetic screens, an existing representation-level fusion method for predicting mean cellular response to unseen genetic perturbations Littman et al. (2025). Despite using only five priors, our proposed response-level fusion outperforms PRESAGE with 40 knowledge sources. For example, simple nonparametric median fusion already increases sign agreement by 9–14% across both genetic screens when compared to PRESAGE. Moreover, learning a nonlinear integration rule further improves performance, as the shared MLP outperforms PRESAGE on all six dataset–metric comparisons, reducing RMSE by 6–8%, increasing sign agreement by 15–22%, and $\mathrm{Pearson}_{\Delta}^{\mathrm{pert}}$ by 6–10% across `Replogle K562` and `RPE1` (Appendix Table 6).

7 Discussion

In this paper, we formulate unseen perturbation prediction as a generalization problem over multiple biological priors, and introduce a corresponding hypothesis space that accommodates heterogeneous priors and predictor architectures, identifying that the stage of prior integration is a critical modeling choice for downstream performance. While existing perturbation prediction methods primarily combine priors at the representation level (Littman et al., 2025; Cole et al., 2026), we instead propose response-level fusion, showing that mapping each prior separately into response space before

Table 2: Comparison of single-prior models, representation-level fusion, and response-level fusion. Single-prior models use one prior at a time, whereas fusion approaches combine multiple priors. Response-level fusion rows are shaded gray. Each entry is the mean across cross-validation folds, with the standard deviation across folds as a subscript. *Per-perturbation oracle* is a reference that selects the best single-prior predictor with the lowest MMD for each test perturbation using its observed response. Bold and underlining indicate the best and second-best mean in each column, respectively, excluding per-perturbation oracle.

	SciPlex3 K562			SciPlex3 MCF7		
Method	MMD ↓	Sign ↑	$\text{Pearson}_{\Delta}^{\text{pert}}$ ↑	MMD ↓	Sign ↑	$\text{Pearson}_{\Delta}^{\text{pert}}$ ↑
Learning-free baselines						
Random control cell	$0.291_{\pm.036}$	$0.556_{\pm.001}$	$-0.024_{\pm.019}$	$0.328_{\pm.048}$	$0.553_{\pm.004}$	$-0.147_{\pm.010}$
Random perturbed cell	$\underline{0.284}_{\pm.035}$	$0.553_{\pm.002}$	$0.009_{\pm.028}$	$0.307_{\pm.046}$	$0.538_{\pm.009}$	$-0.032_{\pm.033}$
Single prior						
One-hot	$0.360_{\pm.017}$	$0.488_{\pm.018}$	$0.053_{\pm.017}$	$0.347_{\pm.043}$	$0.551_{\pm.010}$	$0.163_{\pm.021}$
ECFP	$0.333_{\pm.024}$	$0.536_{\pm.010}$	$0.121_{\pm.052}$	$0.290_{\pm.047}$	$0.584_{\pm.020}$	$0.231_{\pm.035}$
CLAMP	$0.347_{\pm.034}$	$0.535_{\pm.007}$	$0.094_{\pm.016}$	$0.321_{\pm.035}$	$0.569_{\pm.017}$	$0.203_{\pm.047}$
ChemBERTa	$0.329_{\pm.035}$	$0.541_{\pm.009}$	$0.113_{\pm.054}$	$0.299_{\pm.043}$	$0.572_{\pm.022}$	$0.194_{\pm.054}$
MoLFormer	$0.336_{\pm.026}$	$0.549_{\pm.012}$	$0.107_{\pm.032}$	$0.299_{\pm.043}$	$0.577_{\pm.020}$	$0.175_{\pm.055}$
MolT5	$0.342_{\pm.035}$	$0.535_{\pm.009}$	$0.079_{\pm.054}$	$0.307_{\pm.050}$	$0.559_{\pm.033}$	$0.180_{\pm.050}$
Uni-Mol	$0.341_{\pm.030}$	$0.547_{\pm.006}$	$0.102_{\pm.026}$	$0.295_{\pm.040}$	$0.575_{\pm.022}$	$0.210_{\pm.053}$
Multi-prior: representation-level fusion						
Concatenation	$0.341_{\pm.022}$	$0.555_{\pm.017}$	$\underline{0.146}_{\pm.061}$	$0.294_{\pm.038}$	$0.564_{\pm.016}$	$0.189_{\pm.046}$
Soft-argmax	$0.340_{\pm.035}$	$0.537_{\pm.011}$	$0.106_{\pm.061}$	$0.298_{\pm.039}$	$0.566_{\pm.025}$	$0.190_{\pm.045}$
Perceiver-style	$0.340_{\pm.032}$	$0.554_{\pm.017}$	$0.138_{\pm.033}$	$0.292_{\pm.041}$	$0.581_{\pm.038}$	$0.231_{\pm.058}$
PRESAGE-style	$0.336_{\pm.028}$	$0.542_{\pm.005}$	$0.110_{\pm.027}$	$0.292_{\pm.036}$	$0.578_{\pm.018}$	$0.247_{\pm.039}$
Multi-prior: response-level fusion						
Mean	$0.300_{\pm.032}$	$0.542_{\pm.005}$	$0.127_{\pm.044}$	$0.282_{\pm.048}$	$0.581_{\pm.006}$	$0.255_{\pm.049}$
Median	$0.297_{\pm.031}$	$\underline{0.705}_{\pm.074}$	$0.129_{\pm.043}$	$\underline{0.279}_{\pm.048}$	$\mathbf{0.928}_{\pm.016}$	$\underline{0.284}_{\pm.051}$
Trimmed mean	$0.297_{\pm.032}$	$0.662_{\pm.056}$	$0.129_{\pm.044}$	$\underline{0.279}_{\pm.048}$	$\underline{0.901}_{\pm.019}$	$0.279_{\pm.051}$
Learned weighted mean	$0.299_{\pm.032}$	$0.542_{\pm.005}$	$0.126_{\pm.044}$	$0.282_{\pm.048}$	$0.580_{\pm.006}$	$0.256_{\pm.049}$
Shared MLP	$\mathbf{0.206}_{\pm.040}$	$\mathbf{0.971}_{\pm.017}$	$\mathbf{0.534}_{\pm.158}$	$\mathbf{0.273}_{\pm.050}$	$0.859_{\pm.044}$	$\mathbf{0.342}_{\pm.058}$
Per-perturbation oracle	$0.299_{\pm.029}$	$0.573_{\pm.009}$	$0.258_{\pm.037}$	$0.244_{\pm.040}$	$0.610_{\pm.006}$	$0.485_{\pm.032}$

	Replogle K562			Replogle RPE1		
Method	MMD ↓	Sign ↑	$\text{Pearson}_{\Delta}^{\text{pert}}$ ↑	MMD ↓	Sign ↑	$\text{Pearson}_{\Delta}^{\text{pert}}$ ↑
Learning-free baselines						
Random control cell	$0.611_{\pm.019}$	$0.506_{\pm.003}$	$-0.141_{\pm.022}$	$1.062_{\pm.030}$	$0.512_{\pm.003}$	$-0.231_{\pm.013}$
Random perturbed cell	$0.556_{\pm.016}$	$0.578_{\pm.004}$	$0.023_{\pm.007}$	$0.795_{\pm.027}$	$0.692_{\pm.002}$	$-0.138_{\pm.012}$
Single prior						
One-hot	$0.597_{\pm.015}$	$0.542_{\pm.010}$	$-0.088_{\pm.015}$	$0.855_{\pm.031}$	$0.687_{\pm.007}$	$-0.126_{\pm.016}$
DepMap	$0.492_{\pm.009}$	$0.661_{\pm.006}$	$0.303_{\pm.017}$	$0.672_{\pm.035}$	$0.722_{\pm.003}$	$0.303_{\pm.036}$
GenePT gene	$0.503_{\pm.009}$	$0.641_{\pm.009}$	$0.241_{\pm.014}$	$0.712_{\pm.026}$	$0.709_{\pm.004}$	$0.216_{\pm.014}$
GenePT protein	$0.499_{\pm.010}$	$0.649_{\pm.007}$	$0.260_{\pm.019}$	$0.706_{\pm.029}$	$0.712_{\pm.003}$	$0.225_{\pm.025}$
STRING	$0.499_{\pm.009}$	$0.650_{\pm.005}$	$0.250_{\pm.010}$	$0.714_{\pm.023}$	$0.706_{\pm.005}$	$0.225_{\pm.024}$
ESM-2	$0.529_{\pm.015}$	$0.620_{\pm.011}$	$0.184_{\pm.041}$	$0.755_{\pm.019}$	$0.696_{\pm.013}$	$0.156_{\pm.018}$
Multi-prior: representation-level fusion						
Concatenation	$\mathbf{0.480}_{\pm.010}$	$0.673_{\pm.004}$	$\mathbf{0.327}_{\pm.021}$	$\underline{0.635}_{\pm.022}$	$0.732_{\pm.006}$	$\underline{0.371}_{\pm.014}$
Soft-argmax	$0.494_{\pm.009}$	$0.651_{\pm.010}$	$0.269_{\pm.023}$	$0.708_{\pm.036}$	$0.709_{\pm.011}$	$0.223_{\pm.031}$
Perceiver-style	$0.497_{\pm.016}$	$0.651_{\pm.007}$	$0.271_{\pm.024}$	$0.676_{\pm.023}$	$0.719_{\pm.006}$	$0.289_{\pm.024}$
PRESAGE-style	$\underline{0.486}_{\pm.008}$	$0.658_{\pm.007}$	$0.293_{\pm.021}$	$0.663_{\pm.022}$	$0.722_{\pm.010}$	$0.313_{\pm.012}$
Multi-prior: response-level fusion						
Mean	$0.504_{\pm.014}$	$0.674_{\pm.005}$	$0.319_{\pm.017}$	$0.669_{\pm.027}$	$0.735_{\pm.005}$	$0.289_{\pm.016}$
Median	$0.489_{\pm.014}$	$\underline{0.741}_{\pm.004}$	$0.308_{\pm.020}$	$0.649_{\pm.026}$	$\underline{0.791}_{\pm.004}$	$0.286_{\pm.016}$
Trimmed mean	$0.495_{\pm.014}$	$0.728_{\pm.004}$	$0.312_{\pm.018}$	$0.654_{\pm.027}$	$0.778_{\pm.004}$	$0.287_{\pm.016}$
Learned weighted mean	$0.502_{\pm.013}$	$0.675_{\pm.004}$	$\underline{0.320}_{\pm.017}$	$0.665_{\pm.027}$	$0.736_{\pm.004}$	$0.299_{\pm.019}$
Shared MLP	$\mathbf{0.480}_{\pm.021}$	$\mathbf{0.764}_{\pm.022}$	$0.302_{\pm.029}$	$\mathbf{0.559}_{\pm.027}$	$\mathbf{0.879}_{\pm.004}$	$\mathbf{0.380}_{\pm.026}$
Per-perturbation oracle	$0.427_{\pm.008}$	$0.710_{\pm.007}$	$0.447_{\pm.010}$	$0.551_{\pm.020}$	$0.766_{\pm.005}$	$0.495_{\pm.012}$

integration can provide a more effective way to combine heterogeneous information. This pattern persists across predictor architectures, perturbation modalities, and prediction targets.

The advantage of response-level fusion is that it does not rely solely on a powerful learned combiner, but instead leverages the integration of multiple learners. Even simple nonparametric integration improves over individual priors in many settings, while a shared nonlinear integration rule yields further gains, suggesting that there is straightforward, useful structure across the different single-prior predictions to exploit. We hypothesize that one reason for this advantage is that biological priors may be easier to combine after translation into the response space. Perturbation datasets typically contain relatively few training perturbations, while biological priors are often high-dimensional and heterogeneous. Representation-level fusion asks a single predictor to reconcile these representations while simultaneously learning the response distribution with limited perturbation-level supervision. Response-level fusion instead lets each prior-specific predictor first map its representation into a common response space, leaving the integration function the simpler problem of combining predictions in the same space.

Although we focus on genetic and chemical perturbations, our formulation and distinction between representation- and response-level fusion may extend to other perturbation-based prediction spaces with heterogeneous data representations. Open questions for future work include whether hybrid architectures that combine representation- and response-level fusion can yield further gains, and whether parameter sharing across prior-specific predictors can reduce the computational cost of response-level fusion without sacrificing its flexibility.

Taken together, these results highlight that deciding how to combine multiple biological priors is an important design choice for unseen-perturbation prediction. Specifically, we find that the most effective way to leverage multiple biological priors to increase downstream prediction performance involves treating prior-specific response prediction and cross-prior integration as separate learning problems, providing a flexible, modular framework for incorporating heterogeneous sources of biological information.

Reproducibility Statement

The problem formulation and the two-stage learning procedure are provided in Sections 3-4 and the experimental setup and evaluation protocol are described in Section 5. Appendix B.1 documents the datasets, preprocessing steps, train-test splits, and biological prior representations, while Appendix B.2 specifies the evaluation metrics. Model architectures, fusion methods, and training configurations are described in Appendices B.3–B.5. Source code to reproduce results is provided as a ZIP archive in the supplementary materials.

AI Use Statement

In this work, generative AI was used to "assist in the writing of proofs", "implement methods", and "clean and reformat datasets". We have not used generative AI tools to "help develop theoretical models or conceptual frameworks", "formulate mathematical claims", "provide critical ingredients for proving mathematical claims", "propose or refine hypotheses", "design or provide feedback on research methodology or experiments", "support qualitative and thematic data analysis", or "interpret results". "Generate synthetic data sets" and "assist with translation" are not applicable to this work. Additionally, we used generative AI tools to "suggest experimental parameters" and "create or edit software code". We have reviewed all AI-assisted works, where all AI-generated outputs are verified and revised under strict human supervision. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

References

Walid Ahmad, Elana Simon, Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. Chemberta-2: Towards chemical foundation models, 2022. URL https://arxiv.org/abs/2209.01712.

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, 2018.

Fiona M Behan, Francesco Iorio, Gabriele Picco, Emanuel Gonçalves, Charlotte M Beaver, Giorgia Migliardi, Rita Santos, Yanhua Rao, Francesco Sassi, Marika Pinnelli, et al. Prioritization of cancer therapeutic targets using crispr–cas9 screens. *Nature*, 568(7753):511–516, 2019.

Yiqun Chen and James Zou. Simple and effective embedding model for single-cell biology built from ChatGPT. *Nature Biomedical Engineering*, 9(4):483–493, December 2024. ISSN 2157-846X. doi: 10.1038/s41551-024-01284-6. URL https://www.nature.com/articles/s41551-024-01284-6.

Elijah Cole, Geert-Jan Huizing, Sohan Addagudi, Nicholas Ho, Euxhen Hasanaj, Merel Kuijs, Toby Johnstone, Maria Carilli, Alec Davi, Caleb Ellington, et al. Foundation models improve perturbation response prediction. *BioRxiv*, pp. 2026–02, 2026.

Atray Dixit, Oren Parnas, Biyu Li, Jenny Chen, Charles P. Fulco, Livnat Jerby-Arnon, Nemanja D. Marjanovic, Danielle Dionne, Tyler Burks, Raktima Raychowdhury, Britt Adamson, Thomas M. Norman, Eric S. Lander, Jonathan S. Weissman, Nir Friedman, and Aviv Regev. Perturb-Seq: Dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell*, 167(7):1853–1866.e17, 2016. doi: 10.1016/j.cell.2016.11.038.

Carl Edwards, Tuan Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. Translation between molecules and natural language, 2022. URL https://arxiv.org/abs/2204.11817.

Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001. doi: 10.1214/aos/1013203451.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012. URL https://www.jmlr.org/papers/v13/gretton12a.html.

Yue Guo, Hao Zhang, Haitao Hu, Jialu Wu, Ji Cao, Chang-Yu Hsieh, and Bo Yang. Modelling drug-induced cellular perturbation responses with a biologically informed dual-branch transformer. *Nature Machine Intelligence*, 8(1):96–112, 2026.

Chujun He, Jiaqi Zhang, Munther Dahleh, and Caroline Uhler. Morph predicts the single-cell outcome of genetic perturbations across conditions and data modalities. *bioRxiv*, 2025.

Leon Hetzel, Simon Boehm, Niki Kilbertus, Stephan Günnemann, Fabian Theis, et al. Predicting cellular responses to novel drug perturbations at a single-cell resolution. *Advances in Neural Information Processing Systems*, 35:26711–26722, 2022.

Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, Olivier J. Hénaff, Matthew M. Botvinick, Andrew Zisserman, Oriol Vinyals, and João Carreira. Perceiver IO: A general architecture for structured inputs & outputs. *CoRR*, abs/2107.14795, 2021. URL https://arxiv.org/abs/2107.14795.

Kenji Kamimoto, Blerta Stringa, Christy M Hoffmann, Kunal Jindal, Lilianna Solnica-Krezel, and Samantha A Morris. Dissecting cell identity via network inference and in silico gene perturbation. *Nature*, 614(7949):742–751, 2023.

Kenji Kawaguchi, Yoshua Bengio, Vikas Verma, and Leslie Pack Kaelbling. Generalization in machine learning via analytical learning theory. *arXiv preprint arXiv:1802.07426*, 2018.

Kenji Kawaguchi, Leslie Pack Kaelbling, and Yoshua Bengio. Generalization in deep learning. In *Mathematical Aspects of Deep Learning*. Cambridge University Press, 2022. doi: 10.1017/9781009025096.003. URL https://arxiv.org/abs/1710.05468.

Greg Landrum. RDKit: Open-source cheminformatics. https://www.rdkit.org.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. doi: 10.1126/science.ade2574. URL https://www.science.org/doi/abs/10.1126/science.ade2574.

Russell Littman, Jacob Levine, Sepideh Maleki, Yongju Lee, Vladimir Ermakov, Lin Qiu, Alexander Wu, Kexin Huang, Romain Lopez, Gabriele Scalia, Tommaso Biancalani, David Richmond, Aviv Regev, and Jan-Christian Hütter. Gene-embedding-based prediction and functional evaluation of perturbation expression responses with PRESAGE. *bioRxiv*, 2025. doi: 10.1101/2025.06.03.657653. URL https://doi.org/10.1101/2025.06.03.657653.

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, Nika Iremadze, Florian Oberstrass, Doron Lipson, Jessica L. Bonnar, Marco Jost, Thomas M. Norman, and Jonathan S. Weissman. Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq. *Cell*, 185(14):2559–2575.e28, 2022. doi: 10.1016/j.cell.2022.05.013.

David Rogers and Mathew Hahn. Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, May 2010. ISSN 1549-9596. doi: 10.1021/ci100050t. URL https://doi.org/10.1021/ci100050t.

Yusuf Roohani, Kexin Huang, and Jure Leskovec. Predicting transcriptional outcomes of novel multigene perturbations with gears. *Nature biotechnology*, 42(6):927–935, 2024.

Jerret Ross, Brian Belgodere, Vijil Chenthamarakshan, Inkit Padhi, Youssef Mroueh, and Payel Das. Large-scale chemical language representations capture molecular structure and properties, 2022. URL https://arxiv.org/abs/2106.09553.

Eric W Sayers, Jeff Beck, Evan E Bolton, J Rodney Brister, Jessica Chan, Donald C Comeau, Ryan Connor, Michael DiCuccio, Catherine M Farrell, Michael Feldgarden, Anna M Fine, Kathryn Funk, Eneida Hatcher, Marilu Hoeppner, Megan Kane, Sivakumar Kannan, Kenneth S Katz, Christopher Kelly, William Klimke, Sunghwan Kim, Avi Kimchi, Melissa Landrum, Stacy Lathrop, Zhiyong Lu, Adriana Malheiro, Aron Marchler-Bauer, Terence D Murphy, Lon Phan, Arjun B Prasad, Shashikant Pujar, Amanda Sawyer, Erin Schmieder, Valerie A Schneider, Conrad L Schoch, Shobha Sharma, Françoise Thibaud-Nissen, Barton W Trawick, Thilakam Venkatapathi, Jiyao Wang, Kim D Pruitt, and Stephen T Sherry. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 52(D1):D33–D43, January 2024. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkad1044. URL https://academic.oup.com/nar/article/52/D1/D33/7442534.

Philipp Seidl, Andreu Vall, Sepp Hochreiter, and Günter Klambauer. Enhancing activity prediction models in drug discovery with the ability to understand human language. *Proceedings of the 40th International Conference on Machine Learning (ICML)*, July 2023.

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.

Damian Szklarczyk, Rebecca Kirsch, Mikaela Koutrouli, Katerina Nastou, Farrokh Mehryary, Radja Hachilif, Annika L Gable, Tao Fang, Nadezhda T Doncheva, Sampo Pyysalo, Peer Bork, Lars J Jensen, and Christian von Mering. The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Research*, 51(D1):D638–D646, January 2023. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkac1000. URL https://academic.oup.com/nar/article/51/D1/D638/6825349.

The UniProt Consortium, Alex Bateman, Maria-Jesus Martin, Sandra Orchard, Michele Magrane, Shadab Ahmad, Emanuele Alpi, Emily H Bowler-Barnett, Ramona Britto, Hema Bye-A-Jee, Austra Cukura, Paul Denny, Tunca Dogan, ThankGod Ebenezer, Jun Fan, Penelope Garmiri, Leonardo Jose Da Costa Gonzales, Emma Hatton-Ellis, Abdulrahman Hussein, Alexandr Ignatchenko, Giuseppe Insana, Rizwan Ishtiaq, Vishal Joshi, Dushyanth Jyothi, Swaathi Kandasaamy, Antonia Lock, Aurelien Luciani, Marija Lugaric, Jie Luo, Yvonne Lussi, Alistair MacDougall, Fabio Madeira, Mahdi Mahmoudy, Alok Mishra, Katie Moulang, Andrew Nightingale, Sangya Pundir, Guoying Qi, Shriya Raj, Pedro Raposo, Daniel L Rice, Rabie Saidi, Rafael Santos, Elena Speretta, James Stephenson, Prabhat Totoo, Edward Turner, Nidhi Tyagi, Preethi Vasudev, Kate Warner, Xavier Watkins, Rossana Zaru, Hermann Zellner, Alan J Bridge, Lucila Aimo, Ghislaine Argoud-Puy, Andrea H Auchincloss, Kristian B Axelsen, Parit Bansal, Delphine Baratin, Teresa M Batista Neto, Marie-Claude Blatter, Jerven T Bolleman, Emmanuel Boutet, Lionel Breuza, Blanca Cabrera Gil, Cristina Casals-Casas, Kamal Chikh Echioukh, Elisabeth Coudert, Beatrice Cuche, Edouard De Castro, Anne Estreicher, Maria L Famiglietti, Marc Feuermann, Elisabeth Gasteiger, Pascale Gaudet, Sebastien Gehant, Vivienne Gerritsen, Arnaud Gos, Nadine Gruaz, Chantal Hulo, Nevila Hyka-Nouspikel, Florence Jungo, Arnaud Kerhornou, Philippe Le Mercier, Damien Lieberherr, Patrick Masson, Anne Morgat, Venkatesh Muthukrishnan, Salvo Paesano, Ivo Pedruzzi, Sandrine Pilbout, Lucille Pourcel, Sylvain Poux, Monica Pozzato, Manuela Pruess, Nicole Redaschi, Catherine Rivoire, Christian J A Sigrist, Karin Sonesson, Shyamala Sundaram, Cathy H Wu, Cecilia N Arighi, Leslie Arminski, Chuming Chen, Yongxing Chen, Hongzhan Huang, Kati Laiho, Peter McGarvey, Darren A Natale, Karen Ross, C R Vinayaka, Qinghua Wang, Yuqi Wang, and Jian Zhang. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research*, 51(D1):D523–D531, January 2023. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkac1052. URL https://academic.oup.com/nar/article/51/D1/D523/6835362.

Aviad Tsherniak, Francisca Vazquez, Phil G. Montgomery, Barbara A. Weir, Gregory Kryukov, Glenn S. Cowley, Stanley Gill, William F. Harrington, Sasha Pantel, John M. Krill-Burger, Robin M. Meyers, Levi Ali, Amy Goodale, Yenarae Lee, Guozhi Jiang, Jessica Hsiao, William F.J. Gerath, Sara Howell, Erin Merkel, Mahmoud Ghandi, Levi A. Garraway, David E. Root, Todd R. Golub, Jesse S. Boehm, and William C. Hahn. Defining a Cancer Dependency Map. *Cell*, 170(3):564–576.e16, July 2017. ISSN 00928674. doi: 10.1016/j.cell.2017.06.010. URL https://linkinghub.elsevier.com/retrieve/pii/S0092867417306517.

Wouter van Loon, Marjolein Fokkema, Botond Szabó, and Mark de Rooij. Stacked penalized logistic regression for selecting views in multi-view learning. *Information Fusion*, 61:113–123, 2020. doi: 10.1016/j.inffus.2020.03.007.

Ramon Viñas Torné, Maciej Wiatrak, Zoe Piran, Shuyang Fan, Liangze Jiang, Sarah A. Teichmann, Mor Nitzan, and Maria Brbić. Systema: a framework for evaluating genetic perturbation response prediction beyond systematic variation. *Nature Biotechnology*, 44(6):1050–1059, June 2026. ISSN 1546-1696. doi: 10.1038/s41587-025-02777-8. URL https://doi.org/10.1038/s41587-025-02777-8.

David H. Wolpert. Stacked generalization. *Neural Networks*, 5(2):241–259, 1992. doi: 10.1016/S0893-6080(05)80023-1.

Hengshi Yu, Weizhou Qian, Yuxuan Song, and Joshua D Welch. Perturbnet predicts single-cell responses to unseen chemical and genetic perturbations. *Molecular systems biology*, 21(8):960–982, 2025.

Fei Zhao, Chengcui Zhang, and Baocheng Geng. Deep multimodal data fusion. *ACM computing surveys*, 56(9):1–36, 2024.

Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. Uni-mol: A universal 3d molecular representation learning framework. *ChemRxiv*, 2023(0307), 2023. doi: 10.26434/chemrxiv-2022-jjm0j-v4. URL https://chemrxiv.org/doi/abs/10.26434/chemrxiv-2022-jjm0j-v4.

A FINITE-SAMPLE COMPARISON OF REPRESENTATION- AND RESPONSE-LEVEL FUSION

We compare representation-level and response-level fusion for mean-level prediction. For concreteness, we study concatenation as the instantiation of representation-level fusion: fitting a predictor to the concatenated prior representations. Response-level fusion instead fits a predictor to the prior-specific response predictions. We assume a single response gene, writing y for its scalar response.

Let the concatenated prior representation be

$$\mathbf{z}(p) = \left[h_1(p)^\top, \ldots, h_M(p)^\top\right]^\top \in \mathbb{R}^q, \qquad h_m(p) \in \mathbb{R}^{q_m}, \qquad q = \sum_{m=1}^{M} q_m.$$

For a randomly sampled perturbation, assume

$$\mathbf{z} \sim \mathcal{N}(\mathbf{0}, I_q), \qquad y = \boldsymbol{\beta}^\top \mathbf{z} + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2),$$

where ε is independent of $\mathbf{z}$. Assume $1 \le M < q$, and let $B = [\boldsymbol{b}_1, \ldots, \boldsymbol{b}_M] \in \mathbb{R}^{q \times M}$ be fixed and independent of the training sample. We define column $\boldsymbol{b}_m$, the linear prediction direction associated with the mth prior, to have unit norm and be supported only on the coordinates of prior h_m, so the columns are orthonormal. The response-space summaries are

$$\mathbf{s} = B^\top \mathbf{z} \in \mathbb{R}^M.$$

Thus, $s_m = \boldsymbol{b}_m^\top \mathbf{z}$ represents the fixed linear prediction obtained from prior h_m. Define the variance of the signal omitted by these predictions as

$$\delta = \left\|(I_q - BB^\top)\boldsymbol{\beta}\right\|_2^2.$$

Given $n > q + 1$ independent training perturbations, let $\widehat{\boldsymbol{\beta}}$ be the ordinary least-squares (OLS) estimate of $\boldsymbol{\beta}$. Let $\widehat{\boldsymbol{\alpha}}$ denote the coefficient vector obtained by regressing y on the prior-specific predictions $\mathbf{s}$ using OLS. For an independent test pair $(\mathbf{z}_0, y_0)$, define the prediction risks $R_{\text{representation}}$ and R_{response} as the expected squared test errors of representation-level and response-level fusion, respectively:

$$R_{\text{representation}} = \mathbb{E}\left[(y_0 - \widehat{\boldsymbol{\beta}}^\top \mathbf{z}_0)^2\right], \qquad R_{\text{response}} = \mathbb{E}\left[(y_0 - \widehat{\boldsymbol{\alpha}}^\top B^\top \mathbf{z}_0)^2\right],$$

where the expectation is over both the training sample and the test pair. Then

$$R_{\text{representation}} = \sigma^2 \left(1 + \frac{q}{n - q - 1}\right), \tag{6}$$

$$R_{\text{response}} = (\sigma^2 + \delta) \left(1 + \frac{M}{n - M - 1}\right). \tag{7}$$

Consequently,

$$R_{\text{response}} - R_{\text{representation}} = \frac{n-1}{n - M - 1} \left(\delta - \sigma^2 \frac{q - M}{n - q - 1}\right). \tag{8}$$

Thus, response-level fusion has lower expected test error than representation-level fusion if and only if

$$\delta < \sigma^2 \frac{q - M}{n - q - 1}. \tag{9}$$

Proof. Let $Z \in \mathbb{R}^{n \times q}$ be the representation-space design matrix, whose rows are the training representations, and let $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\top$ be the training-noise vector. Conditional on Z, ordinary least squares satisfies

$$\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} = (Z^\top Z)^{-1} Z^\top \boldsymbol{\varepsilon}.$$

Because the test representation has covariance I_q, its conditional excess prediction risk is

$$\mathbb{E}\left[\left(\mathbf{z}_0^\top (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\right)^2 \,\middle|\, Z\right] = \sigma^2 \operatorname{tr}\left((Z^\top Z)^{-1}\right).$$

Since $Z^\top Z$ follows a Wishart distribution with n degrees of freedom and scale matrix I_q,

$$\mathbb{E}\big[(Z^\top Z)^{-1}\big] = \frac{I_q}{n-q-1}.$$

Adding the irreducible noise σ^2 gives Equation 6.

For response-level regression, write

$$\boldsymbol{\alpha} = B^\top \boldsymbol{\beta}, \qquad \boldsymbol{\beta}_\perp = (I_q - BB^\top)\boldsymbol{\beta}.$$

Then

$$y = \boldsymbol{\alpha}^\top \mathbf{s} + \underbrace{\boldsymbol{\beta}_\perp^\top \mathbf{z} + \varepsilon}_{\widetilde{\varepsilon}}.$$

The Gaussian and orthogonality assumptions imply that $\widetilde{\varepsilon}$ is independent of $\mathbf{s}$, with variance

$$\mathrm{Var}(\widetilde{\varepsilon}) = \sigma^2 + \|\boldsymbol{\beta}_\perp\|_2^2 = \sigma^2 + \delta.$$

Moreover, $\mathbf{s} \sim \mathcal{N}(\mathbf{0}, I_M)$. Repeating the same ordinary-least-squares calculation with input dimension M and noise variance $\sigma^2 + \delta$ gives Equation 7. Subtracting Equation 6 from Equation 7 and simplifying yields Equation 8. Because $(n-1)/(n-M-1) > 0$, the response-level risk is smaller exactly when Equation 9 holds. □

B Additional methodological details

B.1 Data preprocessing, splits, and prior representations

Preprocessing. We remove cells with fewer than 100 total counts and genes with fewer than 5 total counts. Counts are normalized to 10^4 per cell and transformed with $\log(1+x)$. Chemical doses are also transformed with $\log(1+x)$. We then restrict each dataset to the top $d = 5{,}000$ highly variable genes. Perturbations with fewer than 32 cells are excluded from evaluation due to insufficient sample size.

Splits and prior coverage. For SciPlex3, we use four-fold cross-validation over compounds, so all four doses of a held-out compound remain unseen during training. For the Replogle screens, we use five-fold cross-validation over perturbations. Within each experiment, we restrict all models to the intersection of perturbations covered by the priors being compared. This prevents a change in representation from also changing the training or evaluation set. All six chemical priors cover all 188 compounds, retaining all 752 (compound, dose) perturbations. The shared genetic intersection retains 1,969 of 2,033 perturbed genes in K562 and 2,196 of 2,264 in RPE1. Every genetic prior individually covers at least 98% of the perturbed genes.

Datasets. After preprocessing and restriction to shared prior coverage, SciPlex3 contains 173,652 cells in K562 and 344,862 cells in MCF7. The Replogle K562 screen contains 1,969 perturbations and 300,998 cells, and the RPE1 screen contains 2,196 perturbations and 239,802 cells.

Prior representations. The chemical priors capture complementary notions of molecular similarity. `ECFP` encodes local chemical substructures using a fixed circular fingerprint (Rogers & Hahn, 2010), whereas `ChemBERTa`, `MoLFormer`, and `MolT5` provide learned representations of molecular structure from SMILES strings (Ahmad et al., 2022; Ross et al., 2022; Edwards et al., 2022). `Uni-Mol` additionally incorporates three-dimensional molecular geometry through a 3D conformer (Zhou et al., 2023), while `CLAMP` incorporates functional information by learning compound representations through contrastive training against bioassay descriptions (Seidl et al., 2023).

The genetic priors likewise capture complementary sources of biological information. `DepMap` represents each gene by its CRISPR knockout fitness effects across cancer cell lines, providing an experimentally derived functional profile (Tsherniak et al., 2017). `GenePT gene` encodes natural-language descriptions of gene function from NCBI (Sayers et al., 2024), while `GenePT protein` augments the text source with UniProt protein annotations (The UniProt Consortium et al., 2023; Chen & Zou, 2024). The `STRING` prior is a text embedding of gene annotations from the STRING

Table 3: Prior representations and their widths. Every biological prior is evaluated individually and is also included in the corresponding fused model.

Prior	Model	Pooling	q_m
Chemical, SciPlex3			
ECFP	RDKit (Landrum) Morgan fingerprint, radius 2	—	2048
CLAMP	CLAMP compound encoder	—	768
ChemBERTa	`DeepChem/ChemBERTa-77M-MTR`	`[CLS]`	384
MoLFormer	`ibm/MoLFormer-XL-both-10pct`	pooler output	768
MolT5	`laituan245/molt5-large` encoder	masked mean	1024
Uni-Mol	`UniMolRepr`, 3D conformer	`cls_repr`	512
		$\sum_m q_m$	5504
Genetic, Replogle			
DepMap	CRISPR knockout fitness profile	—	1100
GenePT gene	`text-embedding-ada-002` on NCBI gene summaries	—	1536
STRING	`text-embedding-ada-002` on STRING-derived gene text	—	1536
GenePT protein	`text-embedding-3-large` on NCBI+UniProt text	—	3072
ESM-2	`esm2_t33_650M_UR50D` on protein sequence	mean over residues	1280
		$\sum_m q_m$	8524
Control			
One-hot	One-hot vector per compound or perturbed gene	—	188 / $\lvert\mathcal{I}\rvert$

database (Szklarczyk et al., 2023), and `ESM-2` derives a representation directly from the amino-acid sequence of the encoded protein (Lin et al., 2023). Together, these priors span experimentally measured perturbational phenotypes, text-derived gene and protein annotations, and protein sequence. We additionally use a one-hot representation of perturbation identity as a non-transferable control: unlike the biological priors, it contains no information relating an unseen perturbation to perturbations observed during training and is not included in any fused model. Table 3 lists all representations and their widths.

B.2 Evaluation metrics

For each held-out perturbation p, we rank genes by the absolute differential-expression score relative to control and evaluate on the top 50. Let $\bar{\boldsymbol{x}}^p$ and $\hat{\bar{\boldsymbol{x}}}^p \in \mathbb{R}^{50}$ denote the observed and predicted pseudo-bulk means over these genes, and let $\bar{\boldsymbol{x}}^c$ denote the control mean. We define $\Delta^p = \bar{\boldsymbol{x}}^p - \bar{\boldsymbol{x}}^c$ and $\hat{\Delta}^p = \hat{\bar{\boldsymbol{x}}}^p - \bar{\boldsymbol{x}}^c$. We also define the mean perturbed centroid

$$\bar{\mathbf{o}} = \frac{1}{|\mathcal{I}_{\text{train}}|} \sum_{q \in \mathcal{I}_{\text{train}}} \bar{\boldsymbol{x}}^q,$$

and write $\rho(\cdot, \cdot)$ for Pearson correlation across the 50 genes.

MMD is computed at the single-cell level. Predicted and observed cells are subsampled to equal size, split into batches of 32 cells, and the squared MMD is averaged across batches. The remaining metrics compare pseudo-bulk profiles. Metrics are first averaged over held-out perturbations within a fold, and we report the mean and standard deviation across folds.

MMD compares the full predicted and observed cell populations. Sign agreement measures the change relative to the control population. These control-referenced metrics can reward response components that are shared across many perturbations. Following (Viñas Torné et al., 2026), $\text{Pearson}_{\Delta}^{\text{pert}}$ instead centers each response at the mean perturbed centroid and measures the component that distinguishes one perturbation from the average perturbed state. We use RMSE for evaluation on mean-response predictors. All evaluation metrics are summarized in Table 4.

B.3 Single-prior predictors

MORPH. We use MORPH (He et al., 2025) as the single-prior predictor for distributional prediction, following the published architecture, training procedure, and hyperparameters. MORPH combines latent representations of a control cell and a perturbation to generate the perturbed response. For prior m, the perturbation encoder ϕ maps $h_m(p)$ to the perturbation latent z_p. Since MORPH was originally developed for genetic perturbations, we extend it to chemical perturbations

Table 4: Evaluation metrics. All metrics are computed on the 50 DEGs of each held-out perturbation. For MMD, each batch contains $n = 32$ predicted and observed cell-expression vectors, and $k(a, b) = \sum_{r=0}^{9} \exp\big(-\|a - b\|_2^2/(2^{r-5}\sigma)\big)$ with $\sigma = 200$. RMSE is reported for the mean-response prediction setting.

Metric	Definition	Better	Reference
MMD	$\frac{1}{n^2}\sum_{i,j} k(\hat{\boldsymbol{x}}_i, \hat{\boldsymbol{x}}_j) + \frac{1}{n^2}\sum_{i,j} k(\boldsymbol{x}_i, \boldsymbol{x}_j) - \frac{2}{n^2}\sum_{i,j} k(\hat{\boldsymbol{x}}_i, \boldsymbol{x}_j)$	Lower	—
RMSE	$\sqrt{\frac{1}{50}\sum_{g=1}^{50}\big(\hat{\bar{x}}_g^p - \bar{x}_g^p\big)^2}$	Lower	—
Sign agreement	$\frac{1}{50}\sum_{g=1}^{50}\mathbf{1}\Big[\mathrm{sign}(\hat{\Delta}_g^p) = \mathrm{sign}(\Delta_g^p)\Big]$	Higher	Control mean $\bar{\boldsymbol{x}}^c$
$\mathrm{Pearson}_\Delta^{\mathrm{pert}}$	$\rho\big(\hat{\bar{\boldsymbol{x}}}^p - \bar{\mathbf{o}}, \bar{\boldsymbol{x}}^p - \bar{\mathbf{o}}\big)$	Higher	Mean perturbed centroid $\bar{\mathbf{o}}$ of the fold's training perturbations

by scaling the perturbation latent by the log-transformed dose:

$$z_p = \mathrm{dose}_p \cdot \phi(h_m(p)),$$

where dose_p is the log-transformed dose for chemical perturbations and 1 for genetic perturbations.

MLP-based VAE. We additionally implemented an MLP-based VAE model for distribution-level perturbation prediction. A cell encoder maps a control cell to a 50-dimensional latent z_{cell}, while a perturbation encoder maps the perturbation representation to a separate 50-dimensional latent z_p, which is similarly scaled by dose. The two latents are concatenated and decoded by

$$\mathbb{R}^{100} \to \mathbb{R}^{128} \to \mathbb{R}^d,$$

with a leaky-ReLU activation (slope 0.2) after each layer. The model is trained with MMD on perturbed cells, control reconstruction loss, and KL regularization.

Mean-response predictor. We adapt MORPH for mean-response prediction, i.e., predicting one mean response per perturbation. We retain the original architecture with the following changes. First, the model input changes to the mean control gene expression, together with the perturbation representation, and the output becomes the perturbation-induced change in mean response, Δ^p. Since this input is a single profile rather than a population of cells, we use the mean of the cell encoder's posterior as the cell latent instead of sampling from it. We also replace the distributional objective, including the control-reconstruction and KL terms, with mean squared error (MSE) against the observed mean change. We increase the maximum number of epochs from 100 to 2000 and the early-stopping patience from 20 to 100. In the mean-response setting, each epoch contains substantially fewer optimization steps because training is performed at the perturbation level rather than over individual cells. We therefore increase the epoch budget and patience to provide sufficient optimization steps. These values are fixed across all datasets and methods in this setting.

B.4 Representation-level fusion

Every representation-level fusion method combines the M prior representations into a single fused perturbation representation u_{fused}. This fused representation is then passed through the same response predictor used in the single-prior model, in place of the single-prior embedding. The fusion module and the response predictor are trained jointly under the same objective. We consider four variants of fusion that increase in flexibility, from direct concatenation to perturbation-specific and cell-conditioned attention.

Concatenation. Concatenation stacks the M prior representations directly, giving input dimensions of 5,504 for the chemical priors and 8,524 for the genetic priors.

Soft-argmax selector. Each prior $h_m(p) \in \mathbb{R}^{q_m}$ is mapped to width $e = 384$ using

$$h_m(p) \mapsto \mathrm{Linear}(q_m \to \mathbb{R}^{384}) \to \mathrm{LayerNorm} \to \mathrm{GELU} \to \mathrm{Dropout}(0.1).$$

Writing the projected representation as u_m, a learned direction $\theta \in \mathbb{R}^{384}$ assigns each prior a scalar score,

$$s_m = \left|u_m^\top \theta\right|, \qquad w = \mathrm{softmax}(s/\tau), \qquad u_{\mathrm{fused}} = \sum_{m=1}^{M} w_m u_m,$$

with fixed temperature $\tau = 0.1$. The low temperature makes the weighting nearly discrete, so the model behaves approximately as a learned prior selector for each perturbation, while still being differentiable.

Perceiver-style fusion (Jaegle et al., 2021). For each prior m, the representation $h_m(p)$ is projected using the same projector as the soft-argmax selector to $u_m \in \mathbb{R}^{384}$, and receives a learned identity embedding t_m,

$$\tilde{u}_m = u_m + t_m, \qquad U = \begin{bmatrix} \tilde{u}_1^\top \\ \vdots \\ \tilde{u}_M^\top \end{bmatrix} \in \mathbb{R}^{M \times 384}.$$

The model uses four learned latent queries $Q \in \mathbb{R}^{4 \times 384}$. We condition these queries on the control-cell latent $z_{\mathrm{cell}} \in \mathbb{R}^{50}$ through

$$Q' = Q + \mathbf{1}(W \cdot z_{\mathrm{cell}} + b)^\top.$$

One cross-attention layer then updates the queries,

$$H = Q' + \mathrm{MHA}(\mathrm{LayerNorm}_{\mathrm{Q}}(Q'), \mathrm{LayerNorm}_{\mathrm{KV}}(U), \mathrm{LayerNorm}_{\mathrm{KV}}(U)),$$

where multi-head attention (MHA) uses four heads and dropout 0.1. The fused representation is

$$u_{\mathrm{fused}} = \frac{1}{4} \sum_{\ell=1}^{4} H_\ell.$$

Because the queries depend on the control-cell latent, the effective contribution of each prior can vary across cells receiving the same perturbation.

PRESAGE-style fusion (Littman et al., 2025). Each prior is projected to width 512 by a source-specific two-layer MLP with hidden width 128 and LeakyReLU hidden/output state activations, giving $u_m \in \mathbb{R}^{512}$. A global branch learns one fixed weight per source, represented as a trainable parameter $g \in \mathbb{R}^M$, to calculate

$$\pi = \mathrm{softmax}(g/0.15), \qquad u_{\mathrm{glob}} = \sum_{m=1}^{M} \pi_m u_m.$$

A second branch learns perturbation-specific weights using graph attention. The M projected priors are connected to a learned virtual node $u_{\mathrm{virtual}} \in \mathbb{R}^{512}$, and a graph-attention layer with five averaged heads, self-loops, LeakyReLU activations, and a learned bias is applied twice with shared weights. The second-pass virtual-node representation, u_{att}, therefore summarizes the priors using perturbation-specific attention. The global and attention branches are then combined and projected,

$$u_{\mathrm{fused}} = W_o(0.85\, u_{\mathrm{att}} + 0.15\, u_{\mathrm{glob}}) + b_o.$$

Unlike the Perceiver-style model, this fusion depends on the perturbation but not on the control cell.

B.5 Response-level fusion

Nonparametric aggregation. The mean and median operate coordinate-wise across the M prior-specific predictions. The trimmed mean removes the largest and smallest prediction for each coordinate before averaging the remaining values.

Learned weighted mean. The learned weighted mean learns one scalar parameter a_m per prior, with $w_m = \mathrm{softmax}(a)_m$, and predicts for gene j:

$$\hat{x}_j^p = \sum_{m=1}^{M} \mathrm{softmax}(a)_m \hat{x}_{m,j}^p.$$

The same learned weights are used for every gene, cell, and perturbation.

Shared MLP. The shared MLP uses

$$\mathbb{R}^M \to \mathbb{R}^{64} \to \mathbb{R}^{64} \to \mathbb{R}$$

with ReLU activations.

Learned combiner training. We fit the learned combiners (the learned weighted mean and shared MLP) on the small set of validation perturbations used for early stopping of the prior-specific predictors. We reserve 10% of the training perturbations in each fold as a validation set for early stopping. 15% of these have all of their cells held out, while the rest contribute half of their cells. The learned combiners are fitted only on the fully held-out validation perturbations, about 1.5% of the training perturbations of a fold. This prevents the combiner from exploiting prior-specific training-set artifacts or overfitting patterns.

In the distribution-level prediction setting, each combiner minimizes the MMD between the combined and observed cell populations on the 50 DEGs of each perturbation, using kernel width $\sigma = 200$. Because the combiner learns a shared integration rule across perturbations and genes, we compute its loss on the 50 DEGs of each perturbation so that optimization is driven by perturbation-responsive genes rather than a large number of genes with little or no change. Both combiners are trained with Adam (learning rate 3×10^{-4}, no weight decay) for 100 epochs, with gradients clipped to a global norm of 1.0. We keep the parameters from the epoch with the lowest training loss. In the mean-response setting, both combiners instead minimize the mean squared error between the combined and observed mean-level changes, and all other training settings remain unchanged.

B.6 PRESAGE BASELINE AND COMPARISON

On the genetic screens, we additionally compare against the published PRESAGE model (Littman et al., 2025). PRESAGE provides a useful comparison because it also combines multiple biological knowledge sources to generalize to unseen gene perturbations, but does so entirely in representation space before predicting the response. It therefore serves as an external representation-level fusion baseline for testing whether the advantage of response-level fusion persists against a method designed specifically for multi-source biological integration.

We use the same perturbation folds and shared prior coverage as the main experiments. We consider two PRESAGE configurations: one using the same five gene priors used throughout our experiments and one using the 40 knowledge sources from the published PRESAGE model. During training, we use the dataset-specific learning rates for `Replogle` reported in the PRESAGE paper. Because PRESAGE represents perturbations as genes and does not model dose, we evaluate it only on the genetic screens.

Table 5: The same comparison as Table 2, using an MLP-based VAE as the prior-specific response predictor instead. Single-prior models use one prior at a time, whereas fusion methods combine multiple priors. Response-level fusion rows are shaded gray. Each entry is the mean across cross-validation folds, with the standard deviation across folds as a subscript. *Per-perturbation oracle* is a reference that selects the best single-prior predictor with the lowest MMD for each test perturbation using its observed response. Bold and underlining indicate the best and second-best mean in each column, respectively, excluding per-perturbation oracle.

	SciPlex3 K562			SciPlex3 MCF7		
Method	MMD ↓	Sign ↑	$\text{Pearson}_{\Delta}^{\text{pert}}$ ↑	MMD ↓	Sign ↑	$\text{Pearson}_{\Delta}^{\text{pert}}$ ↑
Learning-free baselines						
Random control cell	$0.291_{\pm.036}$	$\underline{0.556}_{\pm.001}$	$-0.024_{\pm.019}$	$0.328_{\pm.048}$	$0.553_{\pm.004}$	$-0.147_{\pm.010}$
Random perturbed cell	$\underline{0.284}_{\pm.035}$	$0.553_{\pm.002}$	$0.009_{\pm.028}$	$0.307_{\pm.046}$	$0.538_{\pm.009}$	$-0.032_{\pm.033}$
Single prior						
One-hot	$0.386_{\pm.025}$	$0.482_{\pm.021}$	$0.080_{\pm.050}$	$0.491_{\pm.084}$	$0.511_{\pm.045}$	$0.083_{\pm.063}$
ECFP	$0.369_{\pm.021}$	$0.535_{\pm.018}$	$0.132_{\pm.039}$	$0.320_{\pm.057}$	$0.576_{\pm.007}$	$0.229_{\pm.016}$
CLAMP	$0.376_{\pm.024}$	$0.532_{\pm.014}$	$0.109_{\pm.034}$	$0.335_{\pm.043}$	$0.576_{\pm.021}$	$0.219_{\pm.026}$
ChemBERTa	$0.364_{\pm.029}$	$0.541_{\pm.012}$	$\underline{0.145}_{\pm.057}$	$\underline{0.300}_{\pm.046}$	$0.569_{\pm.018}$	$0.231_{\pm.049}$
MoLFormer	$0.371_{\pm.021}$	$0.539_{\pm.012}$	$0.126_{\pm.038}$	$0.313_{\pm.043}$	$0.571_{\pm.021}$	$0.217_{\pm.065}$
MolT5	$0.373_{\pm.030}$	$0.542_{\pm.012}$	$0.109_{\pm.053}$	$0.314_{\pm.055}$	$0.567_{\pm.025}$	$0.228_{\pm.057}$
Uni-Mol	$0.369_{\pm.024}$	$0.538_{\pm.018}$	$0.115_{\pm.045}$	$0.308_{\pm.044}$	$0.591_{\pm.015}$	$0.209_{\pm.055}$
Multi-prior: representation-level fusion						
Concatenation	$0.367_{\pm.024}$	$0.541_{\pm.016}$	$0.137_{\pm.050}$	$0.323_{\pm.051}$	$0.596_{\pm.020}$	$0.239_{\pm.033}$
Soft-argmax	$0.386_{\pm.042}$	$0.525_{\pm.012}$	$0.116_{\pm.049}$	$0.312_{\pm.044}$	$0.599_{\pm.011}$	$0.231_{\pm.032}$
Perceiver-style	$0.383_{\pm.037}$	$0.534_{\pm.012}$	$0.120_{\pm.028}$	$0.305_{\pm.043}$	$0.622_{\pm.037}$	$0.236_{\pm.029}$
PRESAGE-style	$0.368_{\pm.021}$	$0.535_{\pm.018}$	$0.143_{\pm.031}$	$0.301_{\pm.041}$	$0.601_{\pm.022}$	$0.241_{\pm.060}$
Multi-prior: response-level fusion						
Mean	$0.364_{\pm.025}$	$0.539_{\pm.014}$	$0.144_{\pm.048}$	$0.303_{\pm.049}$	$0.577_{\pm.014}$	$0.246_{\pm.048}$
Median	$0.366_{\pm.024}$	$0.550_{\pm.014}$	$0.141_{\pm.046}$	$0.301_{\pm.049}$	$\underline{0.632}_{\pm.019}$	$\underline{0.260}_{\pm.050}$
Trimmed mean	$0.365_{\pm.024}$	$0.547_{\pm.013}$	$0.142_{\pm.047}$	$0.302_{\pm.048}$	$0.617_{\pm.016}$	$0.257_{\pm.050}$
Learned weighted mean	$0.364_{\pm.024}$	$0.539_{\pm.014}$	$0.144_{\pm.048}$	$0.303_{\pm.048}$	$0.577_{\pm.014}$	$0.246_{\pm.047}$
Shared MLP	$\mathbf{0.214}_{\pm.031}$	$\mathbf{0.956}_{\pm.021}$	$\mathbf{0.535}_{\pm.074}$	$\mathbf{0.291}_{\pm.046}$	$\mathbf{0.801}_{\pm.103}$	$\mathbf{0.323}_{\pm.042}$
Per-perturbation oracle	$0.337_{\pm.023}$	$0.565_{\pm.014}$	$0.261_{\pm.029}$	$0.259_{\pm.038}$	$0.612_{\pm.009}$	$0.492_{\pm.044}$

	Replogle K562			Replogle RPE1		
Method	MMD ↓	Sign ↑	$\text{Pearson}_{\Delta}^{\text{pert}}$ ↑	MMD ↓	Sign ↑	$\text{Pearson}_{\Delta}^{\text{pert}}$ ↑
Learning-free baselines						
Random control cell	$0.611_{\pm.019}$	$0.506_{\pm.003}$	$-0.141_{\pm.022}$	$1.062_{\pm.030}$	$0.512_{\pm.003}$	$-0.231_{\pm.013}$
Random perturbed cell	$0.556_{\pm.016}$	$0.578_{\pm.004}$	$0.023_{\pm.007}$	$0.795_{\pm.027}$	$0.692_{\pm.002}$	$-0.138_{\pm.012}$
Single prior						
One-hot	$0.604_{\pm.018}$	$0.516_{\pm.009}$	$-0.104_{\pm.018}$	$0.874_{\pm.029}$	$0.691_{\pm.006}$	$-0.132_{\pm.022}$
DepMap	$0.496_{\pm.011}$	$0.653_{\pm.008}$	$0.301_{\pm.010}$	$0.691_{\pm.032}$	$0.721_{\pm.010}$	$0.306_{\pm.019}$
GenePT gene	$0.510_{\pm.012}$	$0.634_{\pm.006}$	$0.238_{\pm.007}$	$0.733_{\pm.027}$	$0.706_{\pm.008}$	$0.213_{\pm.022}$
GenePT protein	$0.502_{\pm.013}$	$0.638_{\pm.006}$	$0.261_{\pm.016}$	$0.718_{\pm.028}$	$0.710_{\pm.010}$	$0.232_{\pm.020}$
STRING	$0.501_{\pm.015}$	$0.641_{\pm.006}$	$0.262_{\pm.016}$	$0.730_{\pm.014}$	$0.703_{\pm.013}$	$0.218_{\pm.031}$
ESM-2	$0.535_{\pm.016}$	$0.606_{\pm.003}$	$0.174_{\pm.030}$	$0.760_{\pm.018}$	$0.701_{\pm.006}$	$0.167_{\pm.025}$
Multi-prior: representation-level fusion						
Concatenation	$\underline{0.482}_{\pm.015}$	$0.657_{\pm.002}$	$\mathbf{0.325}_{\pm.018}$	$\underline{0.659}_{\pm.021}$	$0.731_{\pm.008}$	$\underline{0.348}_{\pm.023}$
Soft-argmax	$0.497_{\pm.013}$	$0.640_{\pm.017}$	$0.252_{\pm.044}$	$0.714_{\pm.034}$	$0.711_{\pm.013}$	$0.239_{\pm.050}$
Perceiver-style	$0.494_{\pm.012}$	$0.643_{\pm.005}$	$0.274_{\pm.018}$	$0.683_{\pm.030}$	$0.719_{\pm.009}$	$0.286_{\pm.026}$
PRESAGE-style	$0.491_{\pm.015}$	$0.649_{\pm.006}$	$0.290_{\pm.019}$	$0.672_{\pm.031}$	$0.721_{\pm.014}$	$0.317_{\pm.017}$
Multi-prior: response-level fusion						
Mean	$0.494_{\pm.014}$	$0.658_{\pm.006}$	$0.305_{\pm.014}$	$0.692_{\pm.021}$	$0.727_{\pm.010}$	$0.289_{\pm.005}$
Median	$0.494_{\pm.014}$	$\underline{0.666}_{\pm.006}$	$0.297_{\pm.015}$	$0.690_{\pm.021}$	$\underline{0.740}_{\pm.010}$	$0.284_{\pm.006}$
Trimmed mean	$0.494_{\pm.014}$	$0.664_{\pm.006}$	$0.299_{\pm.015}$	$0.691_{\pm.021}$	$0.737_{\pm.010}$	$0.285_{\pm.007}$
Learned weighted mean	$0.492_{\pm.014}$	$0.660_{\pm.006}$	$0.310_{\pm.014}$	$0.689_{\pm.022}$	$0.729_{\pm.010}$	$0.298_{\pm.006}$
Shared MLP	$\mathbf{0.469}_{\pm.012}$	$\mathbf{0.749}_{\pm.018}$	$\underline{0.312}_{\pm.021}$	$\mathbf{0.598}_{\pm.031}$	$\mathbf{0.834}_{\pm.020}$	$\mathbf{0.354}_{\pm.010}$
Per-perturbation oracle	$0.439_{\pm.011}$	$0.696_{\pm.005}$	$0.435_{\pm.010}$	$0.576_{\pm.017}$	$0.760_{\pm.009}$	$0.486_{\pm.014}$

Table 6: Mean-response prediction results, with one pseudo-bulk profile per perturbation. Each entry is the mean across cross-validation folds, with the standard deviation as a subscript. Per-perturbation oracle is a reference that selects the best single-prior predictor with the lowest RMSE for each test perturbation using its observed response. Response-level fusion rows are shaded gray. We include PRESAGE only for the genetic datasets, as the published model works on genetic data only. Bold and underlining indicate the best and second-best mean in each column, respectively, excluding per-perturbation oracle.

	SciPlex3 K562			SciPlex3 MCF7		
Method	RMSE ↓	Sign ↑	$\text{Pearson}_{\Delta}^{\text{pert}}$ ↑	RMSE ↓	Sign ↑	$\text{Pearson}_{\Delta}^{\text{pert}}$ ↑
Learning-free baselines						
Control mean	$0.097_{\pm.010}$	$0.000_{\pm.000}$	$-0.034_{\pm.035}$	$0.160_{\pm.025}$	$0.000_{\pm.000}$	$-0.174_{\pm.013}$
Training perturbed mean	$0.094_{\pm.010}$	$0.497_{\pm.007}$	$0.045_{\pm.039}$	$0.146_{\pm.025}$	$0.503_{\pm.009}$	$-0.189_{\pm.029}$
Single prior						
One-hot	$0.094_{\pm.010}$	$0.777_{\pm.010}$	$0.065_{\pm.013}$	$0.150_{\pm.023}$	$0.776_{\pm.006}$	$0.121_{\pm.031}$
ECFP	$0.083_{\pm.010}$	$0.797_{\pm.012}$	$0.168_{\pm.021}$	$0.128_{\pm.024}$	$0.819_{\pm.008}$	$0.198_{\pm.020}$
CLAMP	$0.085_{\pm.010}$	$0.767_{\pm.006}$	$0.156_{\pm.032}$	$0.141_{\pm.021}$	$0.791_{\pm.010}$	$0.228_{\pm.053}$
ChemBERTa	$\underline{0.079}_{\pm.013}$	$0.798_{\pm.010}$	$\underline{0.242}_{\pm.061}$	$0.129_{\pm.024}$	$0.818_{\pm.010}$	$0.221_{\pm.017}$
MoLFormer	$0.084_{\pm.008}$	$0.786_{\pm.010}$	$0.198_{\pm.032}$	$0.132_{\pm.023}$	$0.815_{\pm.009}$	$0.151_{\pm.027}$
MolT5	$0.085_{\pm.014}$	$0.793_{\pm.004}$	$0.132_{\pm.064}$	$0.134_{\pm.030}$	$0.821_{\pm.005}$	$0.188_{\pm.025}$
Uni-Mol	$0.086_{\pm.011}$	$0.783_{\pm.013}$	$0.147_{\pm.038}$	$0.131_{\pm.025}$	$0.809_{\pm.014}$	$0.219_{\pm.049}$
Multi-prior: representation-level fusion						
Concatenation	$0.081_{\pm.010}$	$0.831_{\pm.006}$	$0.233_{\pm.049}$	$0.129_{\pm.026}$	$0.864_{\pm.011}$	$0.202_{\pm.030}$
Soft-argmax	$0.082_{\pm.010}$	$\underline{0.841}_{\pm.004}$	$0.211_{\pm.056}$	$0.131_{\pm.024}$	$0.863_{\pm.022}$	$0.208_{\pm.034}$
Perceiver-style	$0.081_{\pm.010}$	$0.833_{\pm.009}$	$0.206_{\pm.060}$	$0.129_{\pm.023}$	$0.859_{\pm.012}$	$0.209_{\pm.038}$
PRESAGE-style	$0.081_{\pm.010}$	$0.833_{\pm.004}$	$0.220_{\pm.060}$	$0.129_{\pm.023}$	$0.863_{\pm.022}$	$0.206_{\pm.033}$
Multi-prior: response-level fusion						
Mean	$0.080_{\pm.011}$	$0.691_{\pm.004}$	$0.226_{\pm.030}$	$\mathbf{0.122}_{\pm.026}$	$0.776_{\pm.005}$	$0.276_{\pm.019}$
Median	$0.081_{\pm.011}$	$0.832_{\pm.005}$	$\mathbf{0.244}_{\pm.048}$	$0.125_{\pm.026}$	$\underline{0.889}_{\pm.006}$	$0.231_{\pm.018}$
Trimmed mean	$0.080_{\pm.011}$	$0.776_{\pm.003}$	$\mathbf{0.244}_{\pm.045}$	$0.124_{\pm.026}$	$0.856_{\pm.006}$	$0.255_{\pm.020}$
Learned weighted mean	$0.081_{\pm.011}$	$0.690_{\pm.004}$	$0.226_{\pm.029}$	$\mathbf{0.122}_{\pm.026}$	$0.775_{\pm.006}$	$\underline{0.278}_{\pm.020}$
Shared MLP	$\mathbf{0.075}_{\pm.011}$	$\mathbf{0.916}_{\pm.017}$	$0.221_{\pm.016}$	$\underline{0.123}_{\pm.026}$	$\mathbf{0.933}_{\pm.007}$	$\mathbf{0.349}_{\pm.057}$
Per-perturbation oracle	$0.066_{\pm.011}$	$0.835_{\pm.007}$	$0.515_{\pm.027}$	$0.097_{\pm.024}$	$0.855_{\pm.003}$	$0.610_{\pm.029}$

	Replogle K562			Replogle RPE1		
Method	RMSE ↓	Sign ↑	$\text{Pearson}_{\Delta}^{\text{pert}}$ ↑	RMSE ↓	Sign ↑	$\text{Pearson}_{\Delta}^{\text{pert}}$ ↑
Learning-free baselines						
Control mean	$0.221_{\pm.005}$	$0.000_{\pm.000}$	$-0.161_{\pm.024}$	$0.313_{\pm.006}$	$0.000_{\pm.000}$	$-0.262_{\pm.016}$
Training perturbed mean	$0.208_{\pm.004}$	$0.527_{\pm.012}$	$0.114_{\pm.027}$	$0.259_{\pm.006}$	$0.740_{\pm.005}$	$-0.248_{\pm.018}$
Single prior						
One-hot	$0.212_{\pm.005}$	$0.649_{\pm.004}$	$-0.089_{\pm.013}$	$0.238_{\pm.006}$	$0.790_{\pm.004}$	$0.037_{\pm.030}$
DepMap	$0.172_{\pm.005}$	$0.768_{\pm.006}$	$0.389_{\pm.031}$	$0.201_{\pm.007}$	$0.830_{\pm.003}$	$0.391_{\pm.026}$
GenePT gene	$0.179_{\pm.004}$	$0.738_{\pm.005}$	$0.295_{\pm.014}$	$0.220_{\pm.008}$	$0.807_{\pm.007}$	$0.253_{\pm.013}$
GenePT protein	$0.176_{\pm.004}$	$0.740_{\pm.005}$	$0.312_{\pm.023}$	$0.215_{\pm.010}$	$0.809_{\pm.006}$	$0.266_{\pm.037}$
STRING	$0.176_{\pm.004}$	$0.739_{\pm.009}$	$0.313_{\pm.014}$	$0.217_{\pm.007}$	$0.807_{\pm.005}$	$0.258_{\pm.019}$
ESM-2	$0.191_{\pm.006}$	$0.705_{\pm.007}$	$0.210_{\pm.024}$	$0.230_{\pm.007}$	$0.793_{\pm.006}$	$0.188_{\pm.025}$
Multi-prior: representation-level fusion						
Concatenation	$\underline{0.164}_{\pm.004}$	$0.781_{\pm.006}$	$0.423_{\pm.016}$	$\underline{0.196}_{\pm.006}$	$0.835_{\pm.007}$	$\mathbf{0.422}_{\pm.011}$
Soft-argmax	$0.172_{\pm.004}$	$0.757_{\pm.006}$	$0.334_{\pm.020}$	$0.203_{\pm.007}$	$0.835_{\pm.009}$	$0.342_{\pm.034}$
Perceiver-style	$0.166_{\pm.004}$	$0.784_{\pm.005}$	$0.405_{\pm.018}$	$0.199_{\pm.010}$	$0.837_{\pm.007}$	$0.370_{\pm.035}$
Multi-prior: existing representation-level fusion						
PRESAGE (our 5 priors)	$0.171_{\pm.012}$	$0.687_{\pm.016}$	$0.401_{\pm.040}$	$0.212_{\pm.005}$	$0.760_{\pm.014}$	$0.324_{\pm.042}$
PRESAGE (40 sources)	$0.167_{\pm.005}$	$0.691_{\pm.010}$	$0.431_{\pm.019}$	$0.202_{\pm.008}$	$0.772_{\pm.011}$	$0.385_{\pm.021}$
Multi-prior: response-level fusion						
Mean	$0.170_{\pm.004}$	$0.721_{\pm.006}$	$0.430_{\pm.014}$	$0.202_{\pm.008}$	$0.798_{\pm.007}$	$0.383_{\pm.008}$
Median	$0.172_{\pm.004}$	$\underline{0.788}_{\pm.005}$	$0.378_{\pm.021}$	$0.206_{\pm.009}$	$\underline{0.843}_{\pm.006}$	$0.337_{\pm.013}$
Trimmed mean	$0.171_{\pm.004}$	$0.761_{\pm.006}$	$0.405_{\pm.018}$	$0.204_{\pm.009}$	$0.823_{\pm.006}$	$0.357_{\pm.013}$
Learned weighted mean	$0.169_{\pm.004}$	$0.723_{\pm.006}$	$\underline{0.435}_{\pm.016}$	$0.202_{\pm.008}$	$0.799_{\pm.007}$	$\underline{0.394}_{\pm.012}$
Shared MLP	$\mathbf{0.153}_{\pm.005}$	$\mathbf{0.844}_{\pm.010}$	$\mathbf{0.459}_{\pm.011}$	$\mathbf{0.189}_{\pm.006}$	$\mathbf{0.889}_{\pm.009}$	$\mathbf{0.422}_{\pm.027}$
Per-perturbation oracle	$0.145_{\pm.004}$	$0.816_{\pm.005}$	$0.573_{\pm.010}$	$0.163_{\pm.007}$	$0.858_{\pm.003}$	$0.589_{\pm.017}$