P2P: CROSS-VIEW POPULATION DENOISING FOR UNPAIRED SINGLE-CELL PERTURBATION RESPONSE PREDICTION

Anonymous authors
Paper under double-blind review

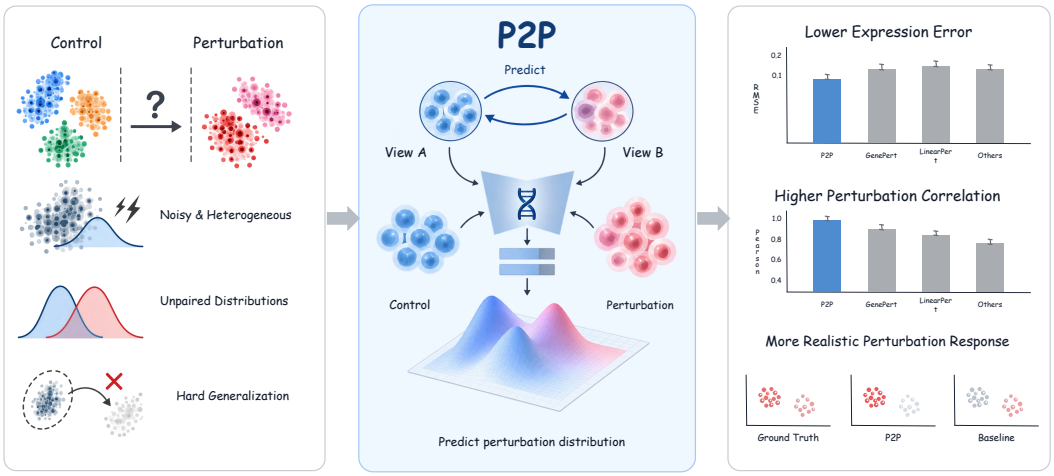


Figure 1: Cross-view population denoising. Perturbation assays provide noisy and heterogeneous cell collections without one-to-one control pairs. P2P samples two stochastic views from each population and learns the shared condition-level effect, which is then evaluated through population expression, effect recovery, and response structure.

ABSTRACT

AIVC (AI Virtual Cell) is a learned simulator of cellular behavior across conditions. Predicting how a cell population responds transcriptionally to a genetic perturbation is a core task. Perturb-seq records that response by destructive sequencing, so a control cell and a perturbed cell are never observed as a pair, and cells under one condition remain heterogeneous and noisy. Regression on individual cells absorbs sampling variation into the estimated effect, whereas interpretation requires the reproducible population effect. P2P (Perturbation-to-Perturbation) takes a stochastic cell-set view as its supervision unit. Two views drawn from the same condition share a reproducible population effect and differ by view-specific variation. A permutation-invariant set encoder summarizes the control population, a structured encoder represents perturbation tokens, cellular context, dose, and combination interactions, and a gate blends empirical condition-effect memory with a neural residual. A heteroscedastic head predicts the population mean and gene-wise response variance. Under one protocol and five seeds, P2P attains the lowest expression RMSE and the highest Effect Pearson, DEG F1, and DEG average precision on each of Adamson, Norman, Replogle K562, and Replogle RPE1 relative to GenePert, LinearPert, SLIM, Scouter, and scPILOT. On Replogle K562, Effect Pearson rises from 0.643 to 0.702 and DEG F1 rises from 0.067 to 0.178 relative to Scouter, the strongest baseline on both metrics.

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.

1 INTRODUCTION

Single-cell RNA sequencing measures expression variation across thousands of cells (Macosko et al., 2015; Zheng et al., 2017). Coupling these readouts to pooled genetic screens connects interventions to cellular responses (Dixit et al., 2016; Adamson et al., 2016; Datlinger et al., 2017; 2021). Combinatorial and genome-scale screens extend this approach to genetic interactions and broad genotype-phenotype landscapes (Norman et al., 2019; Replogle et al., 2022). A predictive model can estimate responses for unmeasured conditions and prioritize affected genes for follow-up experiments. The practical target is a population response, since biological conclusions usually rely on many cells.

The data structure complicates this target. RNA sequencing is destructive, so control cells and their perturbed counterparts are not observed before and after treatment. Multiplexed measurements and integration methods address sample identity and technical variation (Stoeckius et al., 2018; Haghverdi et al., 2018; Korsunsky et al., 2019), but cells under one condition still differ in cycle, stress state, and capture efficiency. Pairing an arbitrary control cell with an unrelated perturbed cell creates an unobserved counterfactual. The resulting target mixes the condition effect with sampling variation.

Figure 1 presents the supervision principle behind P2P. Two random sets sampled from one condition have different sampling variation but share a reproducible population response. One set can therefore supervise the population statistics predicted from the other. P2P constructs paired stochastic views of the control and perturbed populations, encodes each unordered cell set through shared functions and symmetric moments, and learns the condition effect through cross-view agreement.

The model combines empirical effect memory with a neural residual for context, dose, and combination structure. Its output contains a population mean and gene-wise variance, separating effect recovery from response heterogeneity. The central contribution is a population learning objective that uses stochastic cell sets as supervision units without requiring cell correspondences. The four-dataset evaluation tests this formulation through expression error, effect correlation, and DEG ranking, while component ablations isolate the roles of cross-view supervision, memory, and consistency constraints.

2 RELATED WORK

Perturbation models reconstruct different aspects of an intervention. Statistical models link pooled CRISPR assignments to transcriptional changes (Duan et al., 2019). scGen and CPA learn latent responses to perturbations and covariates (Lotfollahi et al., 2019; 2023), while GEARS incorporates gene relationships to predict multigene responses (Roohani et al., 2024). These methods motivate structured representations of the intervention. P2P addresses how the response target is constructed from finite cell samples and uses agreement between sampled populations as a learning signal.

Probabilistic representations capture variation beyond a population mean. scVI and scvi-tools provide latent-variable models and inference infrastructure for single-cell data (Lopez et al., 2018; Gayoso et al., 2022). totalVI and MultiVI extend this framework to multimodal observations (Gayoso et al., 2021; Ashuach et al., 2023), while scGPT learns gene representations through generative pretraining (Cui et al., 2024). P2P couples a mean and variance readout to an effect-centered objective, using repeated population views to supervise the reproducible response.

Integration and reference mapping establish comparable cell representations across datasets. Seurat and scArches align datasets or map queries to reference atlases (Stuart et al., 2019; Lotfollahi et al., 2022); Cell BLAST supports reference search through cell embeddings (Cao et al., 2020). Scanpy and UMAP provide analysis and visualization tools (Wolf et al., 2018; Becht et al., 2019). Integration benchmarks evaluate both technical alignment and conservation of biological structure (Luecken et al., 2022). Population perturbation prediction adds a distinct requirement: the representation must preserve the signed response relative to a matched control population.

Biological structure can also be represented through regulatory networks, propagation, and cell-state dynamics. SCENIC infers regulatory programs, and SCAVENGE propagates variant-associated signals across cell networks (Aibar et al., 2017; Yu et al., 2022). scVelo and CellRank use dynamical information to characterize state transitions and fate probabilities (Bergen et al., 2020; Lange et al.,

2022). Graph neural networks encode relational structure (Wu et al., 2021), while symmetric aggregation supports learning from unordered sets (Qi et al., 2017). P2P uses the latter principle to encode cell populations and evaluates recovered effects through population, gene, and graph-derived views.

3 METHOD

3.1 PROBLEM FORMULATION AND CROSS-VIEW SUPERVISION

Let G be the number of measured genes. For a context s, dose d, and perturbation token set p, let $X_{0,s}$ denote a control cell set and $X_{p,s,d}$ the corresponding perturbed set, with each cell in $\mathbb{R}^G$. The population mean and the condition effect are defined in Eq. 1.

$$\mu_{p,s,d} = \mathbb{E}[x \mid p, s, d], \qquad \delta_{p,s,d} = \mu_{p,s,d} - \mu_{0,s}. \tag{1}$$

The observations contain no natural mapping between elements of $X_{0,s}$ and $X_{p,s,d}$. Two stochastic views are therefore drawn for each set using Eq. 2,

$$(X^a_{0,s}, X^b_{0,s}, X^a_{p,s,d}, X^b_{p,s,d}) \sim \mathrm{ViewSplit}(X_{0,s}, X_{p,s,d}). \tag{2}$$

When enough cells are available, the views are disjoint samples. For smaller conditions, the implementation uses independent sampling with replacement so that the prescribed view size remains available.

A training prediction maps one control view and condition metadata to the mean of the other perturbed view as specified in Eq. 3,

$$\widehat{\mu}^{a \leftarrow b} = f_\theta(X^b_{0,s}, p, s, d), \tag{3}$$

with the reverse direction computed in the same update. The loss compares the prediction with the target view mean, its effect relative to the matching control view, and the within-view second moment. At inference, the target view is absent. The model receives a control set, perturbation tokens, context, dose, and an optional condition memory calculated from training-role cells.

This formulation distinguishes three quantities that are often conflated. A cell-level state is one observed expression vector. A population mean is a statistic of a condition. The perturbation effect is a difference between two population means under a common context. P2P optimizes the latter two quantities and uses cell-level observations only to construct the set statistics.

3.2 P2P ARCHITECTURE

The complete data path is shown in Figure 2; the subsections below specify the components that determine its population effect prediction.

3.2.1 PERMUTATION-INVARIANT POPULATION ENCODING

For a cell set $X = \{x_i\}_{i=1}^n$, a shared projection maps each cell to a hidden state according to Eq. 4,

$$h_i = \mathrm{Dropout}[\mathrm{GELU}(\mathrm{LN}(W_x x_i + b_x))]\,. \tag{4}$$

A score network produces normalized cell weights a_i. The set state includes the attention-weighted summary and the unweighted dispersion summary in Eq. 5,

$$\bar{h}_a = \sum_i a_i h_i, \qquad \bar{h} = \frac{1}{n}\sum_i h_i, \qquad v = \frac{1}{n}\sum_i (h_i - \bar{h})^2, \tag{5}$$

$$h_{set} = \mathrm{MomentMLP}([\bar{h}_a, v]). \tag{6}$$

The final set state is given by Eq. 6. Every operation is shared across cells and followed by a symmetric sum, mean, or variance. The resulting encoder is invariant to permutations of the input cells, while v preserves a compact representation of population heterogeneity. The unweighted mean anchors the measured control baseline, while attention captures informative heterogeneity.

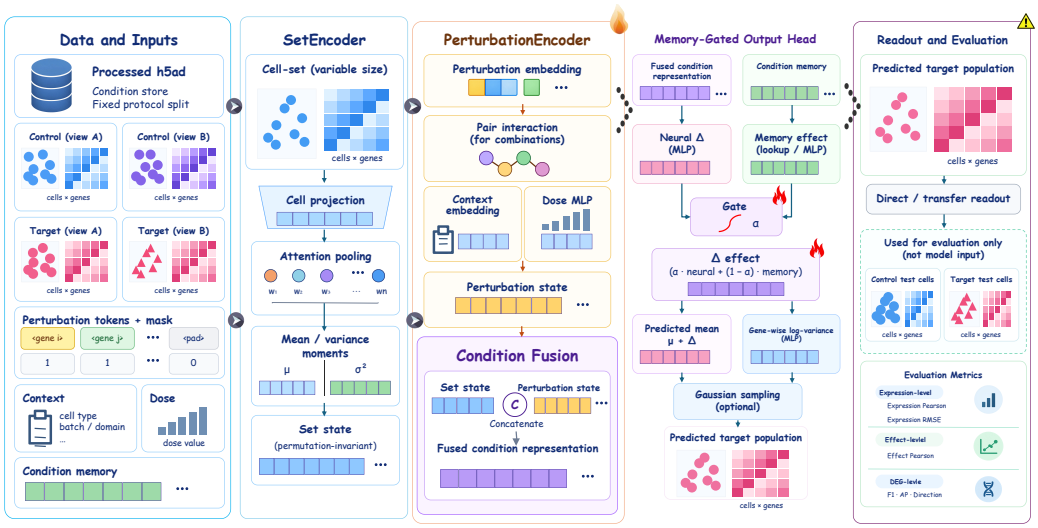


Figure 2: P2P architecture and readout. The permutation-invariant SetEncoder summarizes a control cell set, while the structured perturbation encoder combines perturbation tokens, context, dose, and pair interactions. A memory-gated effect head combines the empirical condition effect with a neural residual and predicts the population mean and gene-wise log variance. Target views supervise the cross-view objective during training.

3.2.2 Structured perturbation and context encoding

A perturbation can contain one token or several tokens. The token embeddings are pooled with a mask as in Eq. 7, so padding cannot contribute to the condition state,

$$e_p = \frac{\sum_j m_j E[p_j]}{\max(1, \sum_j m_j)}. \tag{7}$$

For a combination, a small pair-interaction MLP adds a nonlinear correction $q(e_p)$; the correction is disabled for a single token. Context and dose are encoded separately, and the condition state in Eq. 8 is

$$h_{pert} = \mathrm{LN}\left(e_p + \mathbb{1}_{|p|>1} q(e_p) + E_s[s] + r(d)\right). \tag{8}$$

This design exposes the factors that change a population response and supports both single-gene and combination conditions, while keeping context and dose available to the residual branch.

3.2.3 Memory-gated effect prediction

The control set state and structured condition state are fused by a two-layer MLP, producing the residual in Eq. 9,

$$h_{cond} = \mathrm{Fusion}(h_{set}, h_{pert}), \qquad \Delta_{neural} = \mathrm{DeltaHead}(h_{cond}). \tag{9}$$

For a condition with training-role cells, an empirical memory is computed as the population difference in Eq. 10,

$$m_{p,s,d} = \mathrm{mean}(X^{role}_{p,s,d}) - \mathrm{mean}(X^{role}_{0,s}). \tag{10}$$

A learned scalar gate combines this stable statistic with the neural residual according to Eq. 11,

$$g = \sigma(\alpha), \qquad \widehat{\delta} = g\, m_{p,s,d} + (1-g)\, \Delta_{neural}, qquad \widehat{\mu} = \mathrm{mean}(X_{0,s}) + \widehat{\delta}. \tag{11}$$

The gate is initialized with $\alpha = 1.1$, which gives the memory branch an initial weight of approximately 0.75 while retaining a trainable residual. The formulation makes the statistical prior explicit. The explicit prior lets the model interpolate between empirical condition effects and the learned residual. Transfer readouts use target training cells to construct the empirical memory and are reported under the memory-assisted protocol.

3.2.4 HETEROSCEDASTIC OUTPUT

A second head predicts a gene-wise log variance using Eq. 12,

$$\log \widehat{\sigma}^2 = \mathrm{clip}\left(\mathrm{LogVarHead}(h_{cond}), -8, 4\right). \tag{12}$$

Optional population samples are drawn as $\widetilde{Y} = \widehat{\mu} + \epsilon \exp(\frac{1}{2}\log \widehat{\sigma}^2)$ with standard normal ϵ. The sample mean is centered on $\widehat{\mu}$ before evaluation. This head is trained jointly with the mean and moment losses and is used to represent response variation; uncertainty quality is evaluated separately from point prediction.

3.2.5 EFFECT-CENTRIC CROSS-VIEW OBJECTIVE

For a target view with mean $\bar{y}$ and matching control mean $\bar{c}$, define the predicted effect as $\widehat{\delta} = \widehat{\mu} - \bar{c}$. The direct Gaussian term is defined in Eq. 13,

$$\mathcal{L}_{nll} = \frac{1}{2G}\sum_{g=1}^{G}\left[\log \widehat{\sigma}_g^2 + (\bar{y}_g - \widehat{\mu}_g)^2 \exp(-\log \widehat{\sigma}_g^2)\right]. \tag{13}$$

The effect term is a weighted Smooth L1 distance between $\widehat{\delta}$ and $\delta = \bar{y} - \bar{c}$, where the gene weight is $1 + |\delta_g| / [\mathrm{mean}_k\, |\delta_k| + \varepsilon]$. An effect-correlation term compares centered effect vectors. A moment term matches the predicted variance to the target-view variance in log space. A consistency term couples the two view predictions for the same condition.

The full objective combines these terms as shown in Eq. 14,

$$\mathcal{L} = 0.35\mathcal{L}_{nll} + 1.50\mathcal{L}_{\Delta} + 0.15\mathcal{L}_{corr} + 0.08\mathcal{L}_{mom} + 0.05\mathcal{L}_{cv}. \tag{14}$$

For each update, direct pairs and the crossed pairs $X_0^a \to X_p^b$ and $X_0^b \to X_p^a$ are evaluated. Crossed NLL and effect losses enter their direct counterparts with weight 0.25. The objective therefore gives the model two target views and two control views without treating a cell in one view as the counterpart of a cell in the other.

3.3 POPULATION INTERPRETATION

The cross-view objective has a simple population interpretation. Suppose that, for a fixed condition, a view mean can be written as $\bar{y}^v = \mu_{p,s,d} + \eta^v$ and a control view mean as $\bar{c}^v = \mu_{0,s} + \xi^v$. If the view errors have conditional mean zero and the two draws use the same sampling distribution, then the shared first moments satisfy Eq. 15,

$$\mathbb{E}[\bar{y}^v \mid p, s, d] = \mu_{p,s,d}, \qquad \mathbb{E}[\bar{y}^v - \bar{c}^v \mid p, s, d] = \delta_{p,s,d}. \tag{15}$$

These equalities identify the population first moment targeted by cross-view supervision. When the input is independent of the target-view sampling error, the conditional target has this first moment. The weighted Smooth L1 and correlation terms then shape optimization toward effect magnitude and rank agreement.

The implementation follows this interpretation with a fixed role split and a fixed task list. Training-role cells are used for optimization and memory construction, validation-role cells select checkpoints, and support and test roles are reserved for readout and scoring. For small conditions, view overlap is possible because of replacement sampling; the protocol records overlap explicitly.

4 EXPERIMENTS

4.1 DATASETS AND PROTOCOL

The benchmark uses the processed AnnData files supplied with the project. Adamson is a CRISPRi screen focused on the unfolded protein response. Norman contains single and paired CRISPRa perturbations. Replogle K562 and RPE1 are essential-gene CRISPRi screens. The file-level metadata are retained throughout. The Norman processed file carries an A549 cell-type label, and all Norman analyses use that label.

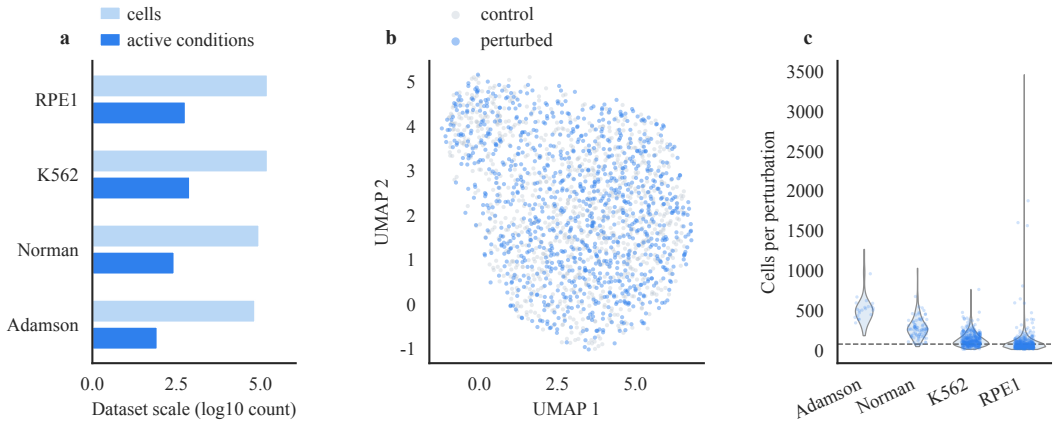


Figure 3: Benchmark scale and support. The four processed files differ in cell count and condition coverage, while the RPE1 embedding shows overlapping control and perturbed states. The dashed support reference marks 80 cells per condition.

The four files contain 68,603, 91,205, 162,751, and 162,733 cells, with 5,060, 5,045, 5,000, and 5,000 measured genes, respectively. Figure 3 visualizes these scale differences and the support distribution. After the 80-cell filter, they retain 86, 274, 794, and 593 active perturbation conditions. The fixed protocol uses the context field and condition label as the condition key, with an approximate 50 percent, 20 percent, 10 percent, and 20 percent role allocation for train, validation, support, and test. The task counts are 57 for Adamson, 128 for Norman, 611 for Replogle K562, and 463 for Replogle RPE1. The supplied summary reports means and sample standard deviations over five seeds. The supplied baseline outputs are evaluated under the fixed benchmark protocol, while the P2P runs use the same view size, population readout, and role assignments across datasets.

4.2 Baselines and Metrics

The comparison includes GenePert, LinearPert, SLIM, Scouter, and scPILOT, using the supplied evaluation outputs. Expression RMSE and expression Pearson are reported for the population mean. Effect Pearson measures the correlation between predicted and observed perturbation effects. DEG F1 and DEG average precision assess the ranking of differentially expressed genes, and DEG Direction measures sign agreement on the target DEG set. A gene is a DEG when a Welch test followed by Benjamini-Hochberg correction gives false discovery rate at most 0.05 and the absolute effect is at least 0.1. This evaluation makes effect recovery and DEG ranking visible alongside expression reconstruction.

4.3 Main Results

Table 1 reports the supplied benchmark. Figure 4 gives representative effect-level predictions. P2P obtains the best RMSE, Effect Pearson, DEG F1, and DEG AP on all four datasets. The absolute gains are most visible on the larger Replogle screens. On K562, RMSE is 0.0499 for P2P, compared with 0.0861 for the strongest baseline on that metric, while Effect Pearson rises to 0.7018 from 0.6432. DEG F1 increases from 0.0671 to 0.1784 and DEG AP reaches 0.2572. RPE1 shows the same pattern, with RMSE 0.0910, Effect Pearson 0.6609, DEG F1 0.2173, and DEG AP 0.3120. Adamson and Norman have smaller absolute errors and retain high expression correlation, while the effect metrics still favor P2P.

Expression Pearson scores are similar across methods on Adamson and Norman, where GenePert has the highest value. This metric primarily reflects the shared expression baseline. The effect metrics separate methods more clearly because they subtract the matched control population before scoring. That separation is the intended operating point of the cross-view objective.

Table 1: Main benchmark results. Values are mean plus or minus sample standard deviation across the supplied five seeds. Best values for the principal effect-recovery metrics are bold.

Dataset	Model	RMSE ↓	Expr. r ↑	Effect r ↑	DEG F1 ↑	DEG AP ↑	Direction ↑
Adamson	**P2P**	**0.0478 ± 0.0000**	0.9950 ± 0.0000	**0.8894 ± 0.0000**	**0.5306 ± 0.0000**	**0.6167 ± 0.0006**	**0.9991 ± 0.0002**
	GenePert	0.0534 ± 0.0002	**0.9959 ± 0.0000**	0.8653 ± 0.0004	0.5031 ± 0.0000	0.5812 ± 0.0000	0.9990 ± 0.0000
	LinearPert	0.0507 ± 0.0000	0.9955 ± 0.0000	0.8762 ± 0.0000	0.4950 ± 0.0000	0.5560 ± 0.0000	0.9943 ± 0.0000
	SLIM	0.0496 ± 0.0000	0.9958 ± 0.0000	0.8801 ± 0.0000	0.3428 ± 0.0015	0.5546 ± 0.0000	0.9961 ± 0.0000
	Scouter	0.0511 ± 0.0005	0.9954 ± 0.0001	0.8697 ± 0.0025	0.4275 ± 0.0057	0.5787 ± 0.0070	0.9973 ± 0.0000
	scPILOT	0.0546 ± 0.0005	0.9947 ± 0.0001	0.8521 ± 0.0019	0.4481 ± 0.0232	0.5235 ± 0.0073	0.9956 ± 0.0003
Norman	**P2P**	**0.0402 ± 0.0000**	0.9931 ± 0.0000	**0.8128 ± 0.0000**	**0.2810 ± 0.0003**	**0.6013 ± 0.0013**	**0.9830 ± 0.0006**
	GenePert	0.0454 ± 0.0001	**0.9945 ± 0.0000**	0.7603 ± 0.0011	0.2308 ± 0.0000	0.5731 ± 0.0000	0.9776 ± 0.0000
	LinearPert	0.0456 ± 0.0000	0.9931 ± 0.0000	0.7465 ± 0.0000	0.1996 ± 0.0000	0.4042 ± 0.0000	0.9673 ± 0.0000
	SLIM	0.0441 ± 0.0000	0.9936 ± 0.0000	0.7471 ± 0.0000	0.2137 ± 0.0000	0.4162 ± 0.0000	0.9727 ± 0.0000
	Scouter	0.0588 ± 0.0009	0.9891 ± 0.0003	0.6459 ± 0.0105	0.0620 ± 0.0014	0.5933 ± 0.0215	0.9789 ± 0.0007
	scPILOT	0.0471 ± 0.0005	0.9927 ± 0.0002	0.7053 ± 0.0051	0.1810 ± 0.0075	0.3733 ± 0.0075	0.9544 ± 0.0064
K562	**P2P**	**0.0499 ± 0.0000**	**0.9891 ± 0.0000**	**0.7018 ± 0.0003**	**0.1784 ± 0.0001**	**0.2572 ± 0.0001**	0.8414 ± 0.0004
	GenePert	0.0889 ± 0.0000	0.9851 ± 0.0000	0.6321 ± 0.0000	0.0452 ± 0.0000	0.2339 ± 0.0000	**0.9944 ± 0.0000**
	LinearPert	0.0920 ± 0.0000	0.9843 ± 0.0000	0.6119 ± 0.0000	0.0362 ± 0.0000	0.1450 ± 0.0000	0.9126 ± 0.0000
	SLIM	0.0956 ± 0.0000	0.9830 ± 0.0000	0.5803 ± 0.0000	0.0255 ± 0.0000	0.1251 ± 0.0000	0.9071 ± 0.0000
	Scouter	0.0861 ± 0.0003	0.9861 ± 0.0001	0.6432 ± 0.0025	0.0671 ± 0.0060	0.2426 ± 0.0130	0.9722 ± 0.0095
	scPILOT	0.0932 ± 0.0005	0.9837 ± 0.0002	0.5872 ± 0.0028	0.0448 ± 0.0026	0.1593 ± 0.0065	0.9242 ± 0.0074
RPE1	**P2P**	**0.0910 ± 0.0005**	**0.9829 ± 0.0002**	**0.6609 ± 0.0041**	**0.2173 ± 0.0002**	**0.3120 ± 0.0262**	0.8656 ± 0.0001
	GenePert	0.0960 ± 0.0000	0.9810 ± 0.0000	0.6426 ± 0.0000	0.0640 ± 0.0000	0.2907 ± 0.0000	**0.9886 ± 0.0000**
	LinearPert	0.0995 ± 0.0000	0.9797 ± 0.0000	0.6157 ± 0.0000	0.0324 ± 0.0000	0.1743 ± 0.0000	0.8868 ± 0.0000
	SLIM	0.1065 ± 0.0000	0.9764 ± 0.0000	0.5706 ± 0.0000	0.0209 ± 0.0000	0.1518 ± 0.0000	0.8631 ± 0.0000
	Scouter	0.1244 ± 0.0000	0.9686 ± 0.0000	0.5474 ± 0.0002	0.0978 ± 0.0083	0.2398 ± 0.0002	0.9861 ± 0.0034
	scPILOT	0.1030 ± 0.0014	0.9783 ± 0.0006	0.5789 ± 0.0088	0.0585 ± 0.0040	0.1785 ± 0.0043	0.9043 ± 0.0114

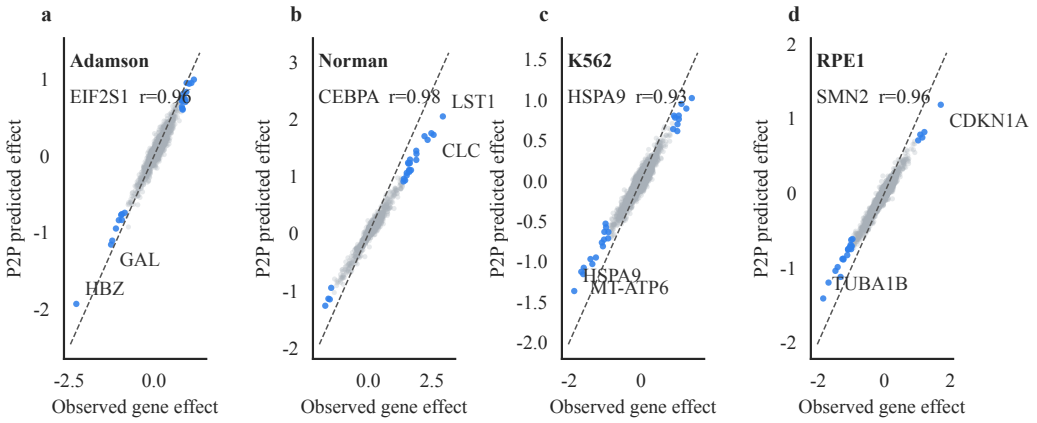


Figure 4: Representative gene-effect predictions. Observed and predicted effects align across Adamson EIF2S1, Norman CEBPA, Replogle K562 HSPA9, and Replogle RPE1 SMN2. The scatter views the quantity optimized by the effect-centered objective.

4.4 Ablation of cross-view supervision and effect components

The RPE1 ablation in Table 2 isolates the parts of the training signal. The robustness diagnostics in Appendix Figure 8 place these component changes beside seed and uncertainty diagnostics. Removing dual-view supervision increases RMSE from 0.0910 to 0.0987 and lowers Effect Pearson from 0.6609 to 0.6412. Removing the empirical memory raises expression error and reduces DEG ranking accuracy, with RMSE 0.1049 and DEG AP 0.2088. The variant labeled no pair interaction has RMSE 0.1123 and DEG F1 0.1742. This ablation quantifies the value of the pair-interaction block on the RPE1 protocol, and the same mechanism supplies the representation used for combination conditions. The gate also contributes as a controlled interpolation between empirical and learned effects.

The heteroscedastic head has a distinct function. Removing it gives a slightly higher Effect Pearson on this split, while DEG AP and expression metrics decrease relative to the full model. This

Table 2: RPE1 ablation. Each row removes one component from the supplied evaluation.

Variant	RMSE ↓	Expr. r	Effect r	DEG F1	DEG AP
Full P2P	**0.0910**	**0.9829**	0.6609	**0.2173**	**0.3120**
No dual-view loss	0.0987	0.9762	0.6412	0.1985	0.2211
No memory effect	0.1049	0.9694	0.6282	0.1865	0.2088
No learned gate	0.0965	0.9781	0.6473	0.2038	0.2265
No pair interaction	0.1123	0.9619	0.6096	0.1742	0.1953
No heteroscedastic head	0.0942	0.9795	**0.6681**	0.2097	0.2319
No consistency or corr.	0.1016	0.9723	0.6350	0.1917	0.2147

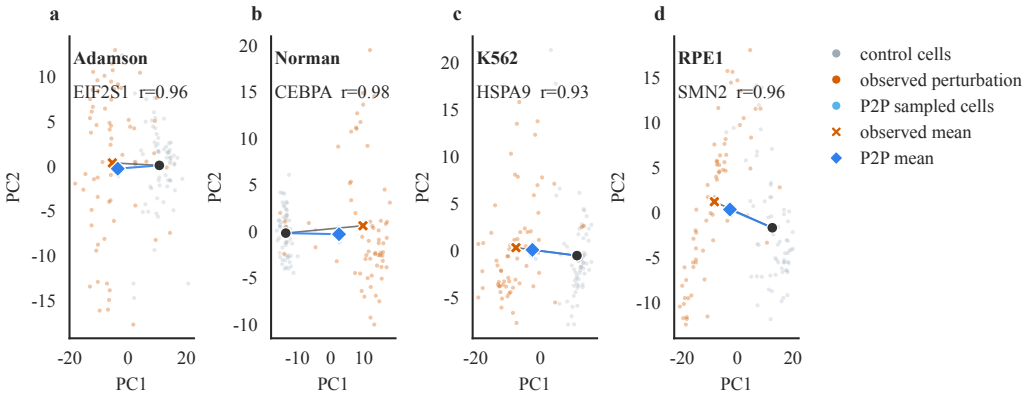


Figure 5: Population-space cases across the four benchmarks. Control cells, observed perturbed cells, and P2P samples are projected into a common PCA coordinate system for one representative task per dataset.

pattern coincides with a trade-off in the supplied ablation between point metrics and the variance-related objective. The cross-view consistency and effect-correlation terms provide a complementary constraint: removing them reduces all principal effect metrics and raises RMSE to 0.1016.

4.5 Population and Biological Structure

The population cases in Figure 5 show the geometric effect of the prediction. The P2P samples move from the control cloud toward the observed perturbed mean across all four datasets, including RPE1. The projection makes population displacement visible in two dimensions.

Figure 6 examines a graph-derived neighborhood for the Adamson EIF2S1 perturbation. The observed and predicted graphs preserve the sign pattern of the dominant local effects, while the residual graph concentrates on a smaller set of nodes. The module panel reports local effect structure from the supplied gene graph.

4.6 Effect Fidelity and Transfer Readout

Appendix Figure 9 relates gene ranking to signed response error through top-ranked overlap, direction agreement, residual distributions, and the Effect Pearson versus DEG AP frontier. Figure 7 resolves these aggregate trends at gene level: observed and predicted effects retain the response ordering, while the connecting segments expose the direction and magnitude of residual error for the largest response genes.

The evaluation also supports a memory-assisted transfer readout. In this setting a source support set is fed to the control-set slot, while target training-role cells provide the empirical memory and target test cells are used only for scoring. This readout tests whether the shared set representation and effect head can adapt to a different observed support population. This memory-assisted setting evaluates adaptation across observed support populations while retaining target training memory and the supplied perturbation vocabulary.

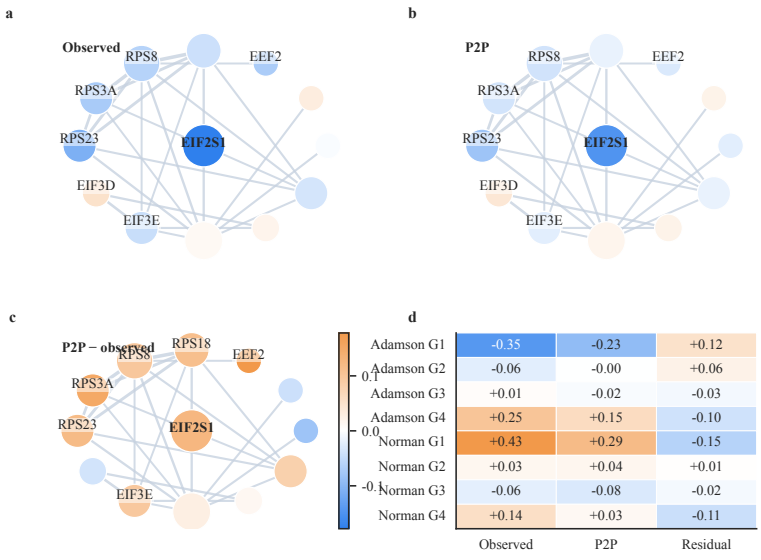


Figure 6: Graph-derived local effect structure. Observed, predicted, and residual networks are paired with module-level effects for Adamson and Norman.

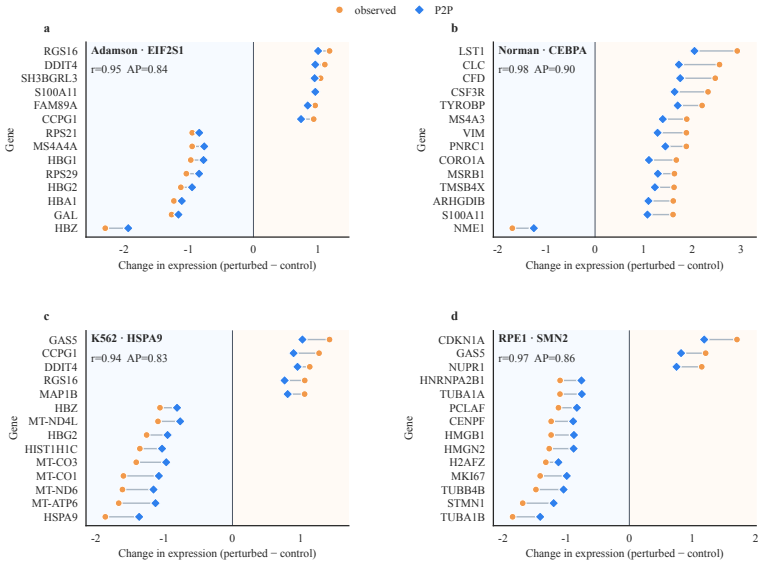


Figure 7: Observed and predicted effects for representative genes. Dumbbells show the signed error for Adamson EIF2S1, Norman CEBPA, Replogle K562 HSPA9, and Replogle RPE1 SMN2.

5 Conclusion

P2P targets population-level perturbation response prediction from unpaired control and perturbed cell sets. Cross-view supervision turns stochastic cell-set views into a direct learning signal, and the memory-gated set model recovers condition effects while retaining a neural residual for context, dose, and combination structure. Across the fixed four-dataset benchmark, the method delivers its clearest gains on effect correlation and DEG ranking, the measures that determine whether a predicted response supports biological prioritization. The transfer readout demonstrates adaptation across observed support populations under the stated memory-assisted protocol.

AI USE STATEMENT

We used generative AI tools to improve the readability and organization of the manuscript. The authors reviewed the edited text, checked every scientific claim against the supplied implementation and evaluation files, and take responsibility for the final manuscript. Generative AI tools were not used to generate experimental data, alter measurements, select favorable seeds, or replace author verification of the reported results.

ETHICS STATEMENT

P2P improves the fidelity of population-level perturbation response prediction. This capability could be misused to infer sensitive biological responses or to construct misleading claims about identifiable biological samples without appropriate authorization. We encourage the use of single-cell expression data only when the relevant rights and permissions are available. The work uses de-identified public or project-provided single-cell datasets and does not recruit human participants or process personally identifying information.

REPRODUCIBILITY STATEMENT

The paper provides the details needed to reproduce P2P. Section 3 defines the stochastic population views, set and condition encoders, memory-gated residual, heteroscedastic head, and cross-view objective. Section 4 specifies the four datasets, role protocol, baselines, metrics, and supplied five-seed evaluation. All parameters remain frozen at inference, and the method is applied to the observed expression matrices and condition metadata. Experiments use an NVIDIA RTX 4090 GPU with 24 GB of memory. The reported training time is approximately three hours. Code, model configurations, evaluation scripts, protocol files, and instructions for reproducing the reported results will be released.

REFERENCES

Britt Adamson, Thomas M. Norman, Marco Jost, et al. A multiplexed single-cell CRISPR screening platform enables systematic dissection of the unfolded protein response. *Cell*, 167(7):1867–1882.e21, 2016. doi: 10.1016/j.cell.2016.11.048.

Sara Aibar, Carmen Bravo González-Blas, Thomas Moerman, et al. SCENIC: single-cell regulatory network inference and clustering. *Nature Methods*, 14(11):1083–1086, 2017. doi: 10.1038/nmeth.4463.

Tal Ashuach, Mariano I. Gabitto, Rohan V. Koodli, et al. MultiVI: deep generative model for the integration of multimodal data. *Nature Methods*, 20(8):1222–1231, 2023. doi: 10.1038/s41592-023-01909-9.

Etienne Becht, Leland McInnes, John Healy, et al. Dimensionality reduction for visualizing single-cell data using UMAP. *Nature Biotechnology*, 37(1):38–44, 2019. doi: 10.1038/nbt.4314.

Volker Bergen, Marius Lange, Stefan Peidli, et al. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nature Biotechnology*, 38(12):1408–1414, 2020. doi: 10.1038/s41587-020-0591-3.

Zhi-Jie Cao, Lin Wei, Shen Lu, et al. Searching large-scale scRNA-seq databases via unbiased cell embedding with Cell BLAST. *Nature Communications*, 11(1):3458, 2020. doi: 10.1038/s41467-020-17281-7.

H. Cui et al. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nature Methods*, 21(8):1470–1480, 2024. doi: 10.1038/s41592-024-02201-0.

Paul Datlinger, André F Rendeiro, Christian Schmidl, et al. Pooled CRISPR screening with single-cell transcriptome readout. *Nature Methods*, 14(3):297–301, 2017. doi: 10.1038/nmeth.4177.

Paul Datlinger, André F. Rendeiro, Thorina Boenke, et al. Ultra-high-throughput single-cell RNA sequencing and perturbation screening with combinatorial fluidic indexing. *Nature Methods*, 18 (6):635–642, 2021. doi: 10.1038/s41592-021-01153-z.

Atray Dixit, Oren Parnas, Biyu Li, et al. 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.

Bin Duan, Chi Zhou, Chengyu Zhu, et al. Model-based understanding of single-cell CRISPR screening. *Nature Communications*, 10(1):2233, 2019. doi: 10.1038/s41467-019-10216-x.

Adam Gayoso, Zoë Steier, Romain Lopez, et al. Joint probabilistic modeling of single-cell multi-omic data with totalVI. *Nature Methods*, 18(3):272–282, 2021. doi: 10.1038/s41592-020-01050-x.

Adam Gayoso, Romain Lopez, Galen Xing, et al. A Python library for probabilistic analysis of single-cell omics data. *Nature Biotechnology*, 40(2):163–166, 2022. doi: 10.1038/s41587-021-01206-w.

Laleh Haghverdi, Aaron T L Lun, Michael D Morgan, and John C Marioni. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nature Biotechnology*, 36(5):421–427, 2018. doi: 10.1038/nbt.4091.

Ilya Korsunsky, Nghia Millard, Jean Fan, et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature Methods*, 16(12):1289–1296, 2019. doi: 10.1038/s41592-019-0619-0.

Marius Lange, Volker Bergen, Michal Klein, et al. CellRank for directed single-cell fate mapping. *Nature Methods*, 19(2):159–170, 2022. doi: 10.1038/s41592-021-01346-6.

Romain Lopez, Jeffrey Regier, Michael B. Cole, et al. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15(12):1053–1058, 2018. doi: 10.1038/s41592-018-0229-2.

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, Mohsen Naghipourfar, Malte D. Luecken, et al. Mapping single-cell data to reference atlases by transfer learning. *Nature Biotechnology*, 40(1):121–130, 2022. doi: 10.1038/s41587-021-01001-7.

Mohammad Lotfollahi, Anna Klimovskaia Susmelj, Carlo De Donno, et al. Predicting cellular responses to complex perturbations in high-throughput screens. *Molecular Systems Biology*, 19 (6):e11517, 2023. doi: 10.15252/msb.202211517.

Malte D. Luecken, M. Büttner, K. Chaichoompu, et al. Benchmarking atlas-level data integration in single-cell genomics. *Nature Methods*, 19(1):41–50, 2022. doi: 10.1038/s41592-021-01336-8.

Evan Z. Macosko, Anindita Basu, Rahul Satija, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214, 2015. doi: 10.1016/j.cell.2015.05.002.

Thomas M. Norman, Max A. Horlbeck, Joseph M. Replogle, et al. Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. *Science*, 365(6455):786–793, 2019. doi: 10.1126/science.aax4438.

Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. PointNet: Deep learning on point sets for 3D classification and segmentation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 77–85, 2017. doi: 10.1109/CVPR.2017.16.

Joseph M. Replogle, Ryan A. Saunders, Andrew N. Pogson, et al. 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.

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

Marlon Stoeckius, Shiwei Zheng, Brian Houck-Loomis, et al. Cell hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. *Genome Biology*, 19(1):224, 2018. doi: 10.1186/s13059-018-1603-1.

Tim Stuart, Andrew Butler, Paul Hoffman, et al. Comprehensive integration of single-cell data. *Cell*, 177(7):1888–1902.e21, 2019. doi: 10.1016/j.cell.2019.05.031.

F. Alexander Wolf, Philipp Angerer, and Fabian J. Theis. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology*, 19(1):15, 2018. doi: 10.1186/s13059-017-1382-0.

Zonghan Wu, Shirui Pan, Fengwen Chen, et al. A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 32(1):4–24, 2021. doi: 10.1109/tnnls.2020.2978386.

Fulong Yu, Liam D. Cato, Chen Weng, et al. Variant to function mapping at single-cell resolution through network propagation. *Nature Biotechnology*, 40(11):1644–1653, 2022. doi: 10.1038/s41587-022-01341-y.

Grace X. Y. Zheng, Jessica M. Terry, Phillip Belgrader, et al. Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8(1):14049, 2017. doi: 10.1038/ncomms14049.

A Implementation details

The set projection uses a linear layer, layer normalization, GELU, and dropout. Attention scores are normalized within each set. The perturbation vocabulary is masked before pooling, and the pair-interaction block is enabled only when more than one perturbation token is active. The fusion block maps the concatenated set and condition states through layer normalization, two hidden layers, GELU, and dropout. The delta and log-variance heads are linear projections from the fused state. Log variance is clipped to the interval from negative eight to four before the Gaussian likelihood and sampling operations.

Training uses AdamW, a plateau-based learning-rate schedule, automatic mixed precision on CUDA, gradient clipping at one, early stopping on validation loss, and atomic checkpoint writes. The checkpoint stores model parameters, dimensions, optimizer configuration, vocabulary, gene names, the fixed protocol audit, and the best validation epoch. Input matrices are read in backed mode and control means are cached by context to reduce repeated I/O.

B Evaluation audit

The direct readout supplies a target condition's control test cells and uses target train-role memory. The memory-assisted transfer readout supplies source support cells to the control slot and uses target train-role memory. Target test cells are never used as model inputs. The vocabulary is constructed from the current processed file, so the supplied benchmark is transductive with respect to token identity, and this choice defines the scope of the evaluation.

C Cross-view objective and implementation map

The implementation uses the same control and perturbed roles to construct two stochastic views. When a condition has at least twice the requested view size, the views are sampled without replacement and are disjoint. Smaller conditions use independent sampling with replacement. This distinction matters for interpreting the cross-view estimator in Eq. 15: the conditional expectation remains the target population mean under the sampling model, while the two views remain samples from the same finite collection rather than independent experimental replicates.

The gene-weighted effect loss gives larger influence to genes with a measurable response. For a target-view effect δ_g, the weight used by the implementation is

$$w_g = 1 + \frac{|\delta_g|}{\mathrm{mean}_k\,|\delta_k| + \varepsilon}, \tag{16}$$

where ε avoids division by zero. Equation 16 is applied inside Smooth L1 before averaging over genes. It prevents a large collection of unchanged genes from dominating the direction of the effect objective.

The variance head is trained with a moment target in addition to the Gaussian term. Let s_g^2 be the empirical target-view variance and let $\widehat{\sigma}_g^2 = \exp(\log \widehat{\sigma}_g^2)$. The moment penalty is

$$\mathcal{L}_{mom} = \frac{1}{G}\sum_{g=1}^{G} \mathrm{SmoothL1}\big(\log(\widehat{\sigma}_g^2 + \varepsilon), \log(s_g^2 + \varepsilon)\big)\,. \tag{17}$$

The log-space comparison in Eq. 17 keeps genes with different expression scales comparable. It also separates the role of the variance head from the point prediction metrics reported in the main text.

For the two directions of a training update, the consistency term couples the predicted means and log variances,

$$\mathcal{L}_{cv} = \mathrm{SmoothL1}(\widehat{\mu}^{a\leftarrow b}, \widehat{\mu}^{b\leftarrow a}) + 0.1\,\mathrm{SmoothL1}(\widehat{\ell}^{a\leftarrow b}, \widehat{\ell}^{b\leftarrow a}), \tag{18}$$

where $\widehat{\ell}$ denotes the predicted log variance. Equation 18 is computed before population sampling. The cross-view term therefore constrains the deterministic prediction and does not depend on a particular draw of ϵ.

The condition memory is computed from role means, while the neural branch receives the encoded control set and condition state. At direct readout, the memory is

$$m_{p,s,d}^{direct} = \mathrm{mean}(X_{p,s,d}^{train}) - \mathrm{mean}(X_{0,s}^{test}). \tag{19}$$

At memory-assisted transfer readout, the current implementation uses

$$m_{p,s,d}^{transfer} = \mathrm{mean}(X_{p,s,d}^{train}) - \mathrm{mean}(X_{source,s}^{support}), \tag{20}$$

so source-support sampling noise enters the transfer memory. Equations 19 and 20 define the two readouts used in the supplied evaluation and explain why the transfer result is reported as memory-assisted.

D PROTOCOL, DATA INVENTORY, AND METRICS

The processed files contain 485,292 cells and 1,747 active perturbation conditions after the 80-cell filter. The fixed evaluation uses 1,259 condition tasks across the four datasets. These totals separate the inventory of eligible conditions from the smaller set of tasks retained by the fixed train, validation, support, and test roles.

The loss weights and readout choices are fixed before evaluation. Table 3 makes the weighting used in Eq. 14 explicit. NLL and effect losses include a direct term and a crossed term with multiplier 0.25; the table reports the outer weights.

For each test task, the target DEG set is defined by a Welch test followed by Benjamini-Hochberg correction and an absolute-effect threshold. In compact form,

$$\mathrm{DEG}(g) = \mathbb{1}[\mathrm{BH}(p_g) \le 0.05 \ \wedge\ |\delta_g| \ge 0.1]\,. \tag{21}$$

Equation 21 is used for DEG F1, DEG average precision, and DEG Direction. Expression RMSE and Pearson are computed on the population mean, while Effect Pearson is computed after subtracting the matching control mean.

Training uses AdamW with automatic mixed precision, gradient clipping at one, a plateau scheduler, early stopping, and atomic checkpoint writes. The reported hardware is an NVIDIA RTX 4090 GPU with 24 GB of memory, and the reported training time is approximately three hours.

Table 3: Fixed objective weights and readout settings.

Component	Setting
Gaussian NLL	0.35
Effect Smooth L1	1.50
Effect correlation	0.15
Moment matching	0.08
Cross-view consistency	0.05
Crossed NLL and effect multiplier	0.25
Minimum condition support	80 cells
Default set size	64 cells
Reported seeds	42, 123, 456, 789, 1024

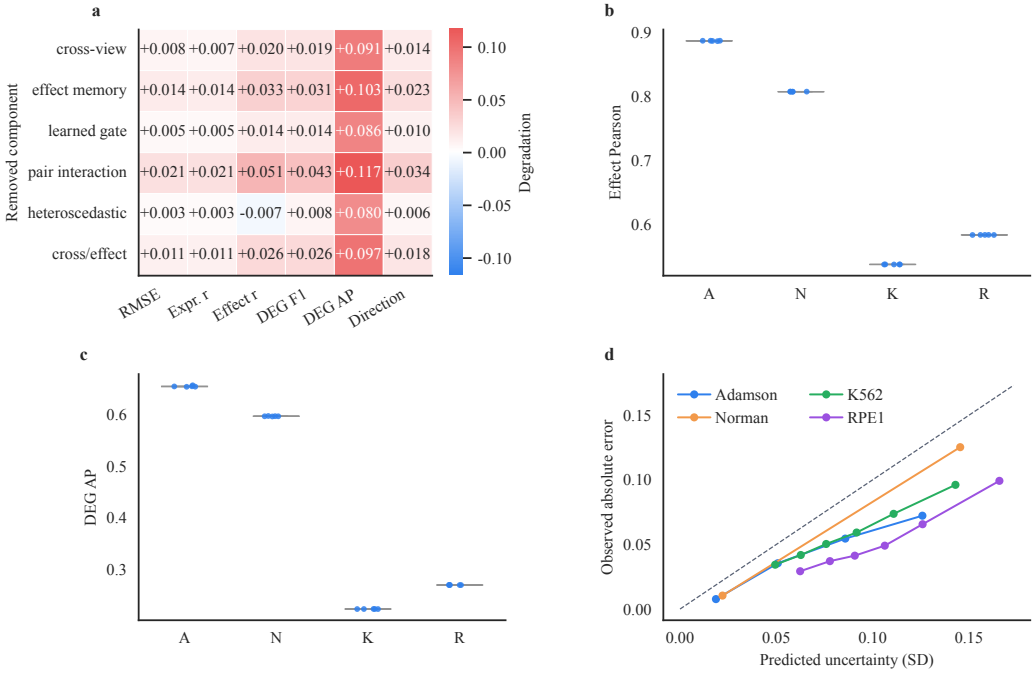


Figure 8: Robustness and uncertainty. The complete model is stable across the supplied seeds, component removals reduce effect and ranking quality, and predicted dispersion tracks absolute error across tasks.

E Additional visual evidence

The main text uses the motivation, architecture, benchmark overview, gene-effect, population, graph-network, and dumbbell figures. The robustness analysis and the remaining supplied figures are retained here so that the full set of visual diagnostics can be inspected. Figure 8 summarizes component ablations, seed variation, and predictive dispersion. Figure 9 summarizes ranking overlap, direction agreement, residual distributions, and the Effect Pearson versus DEG AP frontier.

The appendix cases probe agreement from complementary statistical views. The scatter-residual plot in Figure 10 provides a parity view with explicit residuals, and the Bland-Altman plot in Figure 11 shows the mean-dependent error and limits of agreement. These views are useful because a high correlation can coexist with a scale-dependent bias.

The cell-population case in Figure 12 complements the main population projection by showing the control, observed perturbation, and P2P sample clouds with their means. Figure 13 focuses on Norman combinations, which directly exercises the masked token pooling and pair-interaction block. Figure 14 gives a gene-level ranking view for high-quality single perturbations.

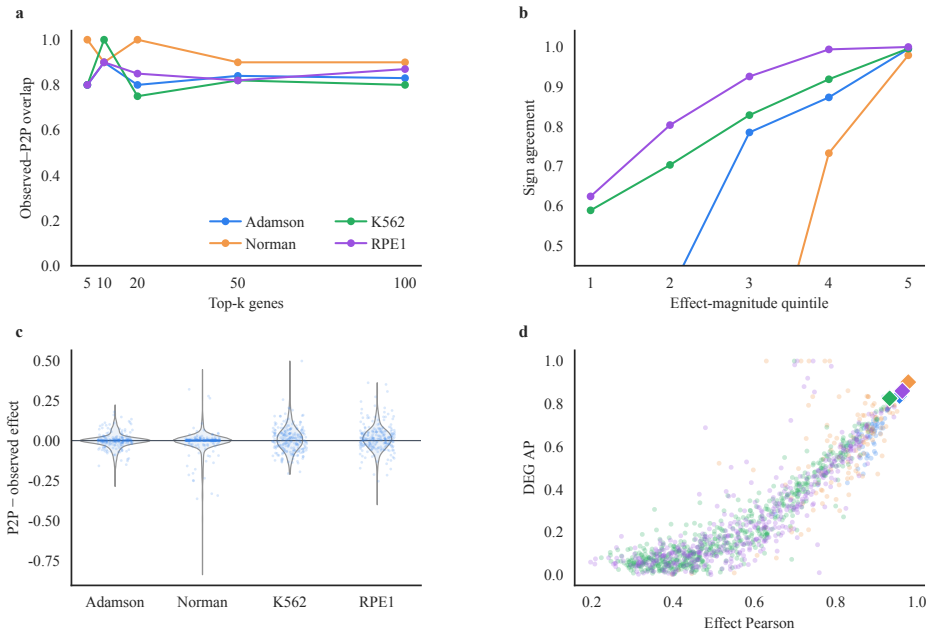


Figure 9: Effect-fidelity summary. The panels report top-k overlap, direction agreement by effect magnitude, residual distributions, and the task-level Effect Pearson versus DEG AP frontier.

The uncertainty diagnostic in Figure 15 displays observed effects, P2P means, and 90 percent predictive intervals for representative genes. Together with the robustness panel, it reports interval coverage as a distributional property alongside point metrics.

The appendix figures are diagnostic views of the same fixed evaluation. They expose population displacement, signed gene effects, combination behavior, agreement structure, and predictive dispersion using outputs from the aggregate evaluation.

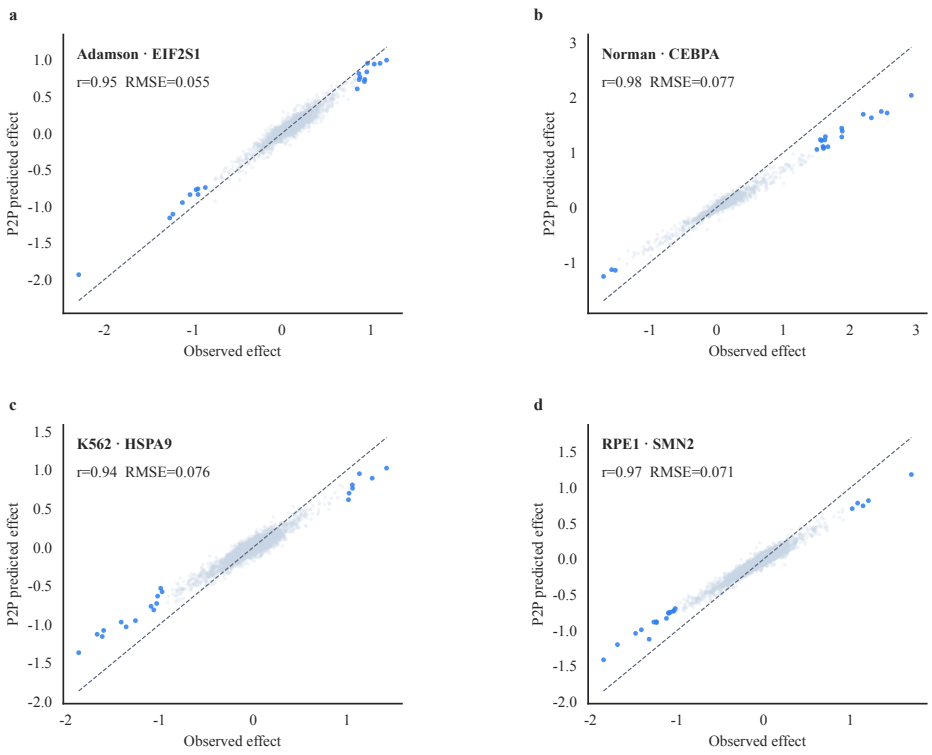


Figure 10: Gene-level parity and residual diagnostics for representative perturbations. The diagonal identifies exact agreement, while darker points highlight genes with large absolute effects.

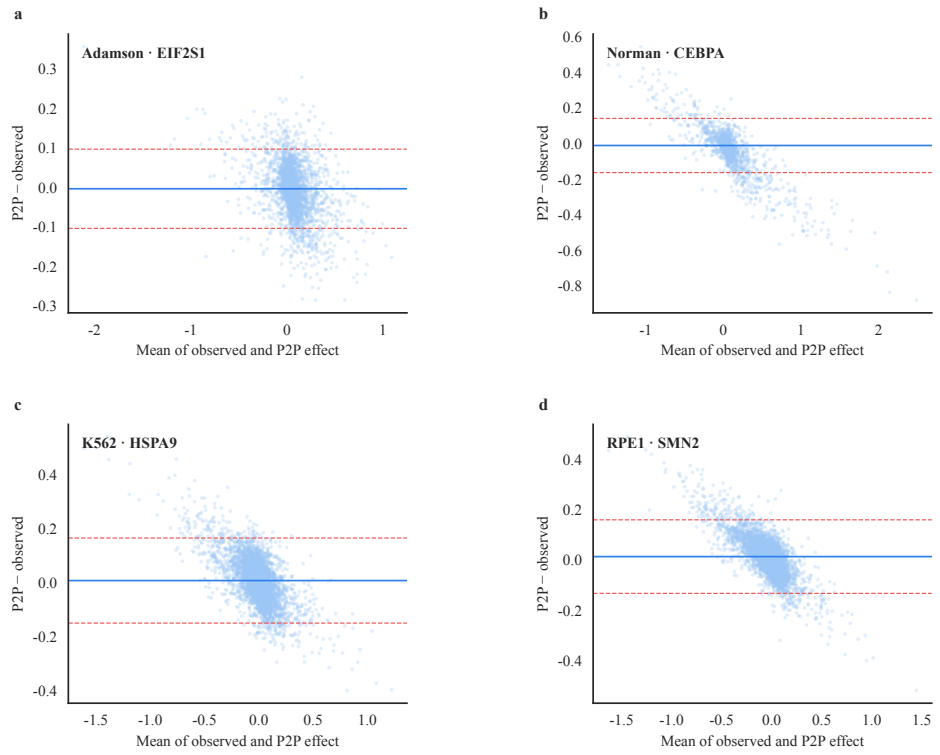


Figure 11: Bland-Altman agreement analysis. The horizontal axis is the observed and predicted effect mean, and the vertical axis is the P2P minus observed effect. The solid line marks the mean difference, and dashed lines mark the 95 percent limits of agreement.

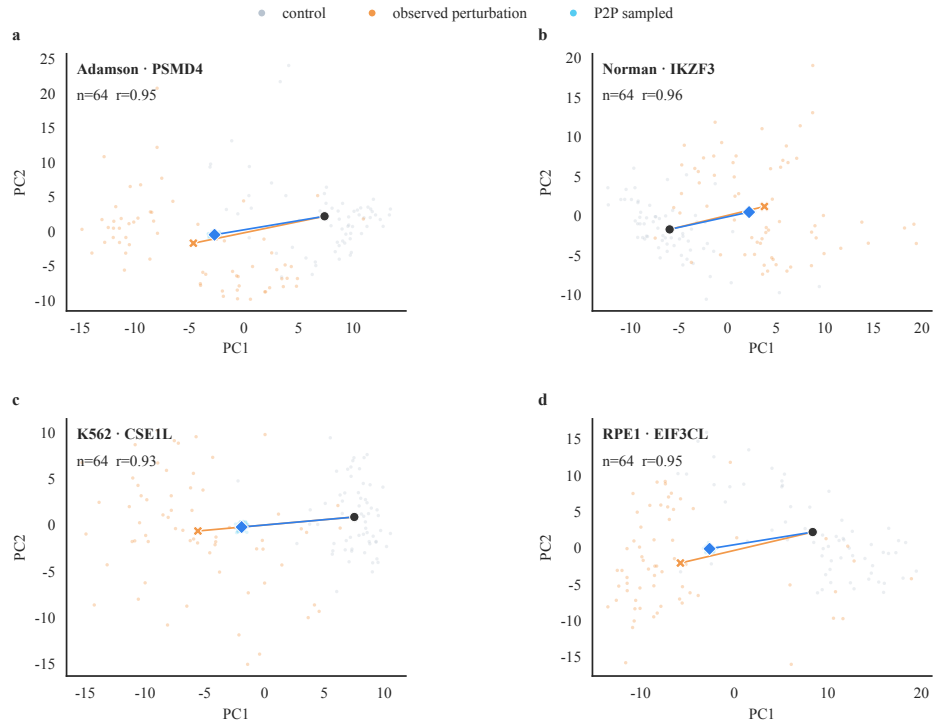


Figure 12: Cell-population reconstruction cases. The control, observed perturbation, and P2P sample clouds are shown in PCA space with their population means.

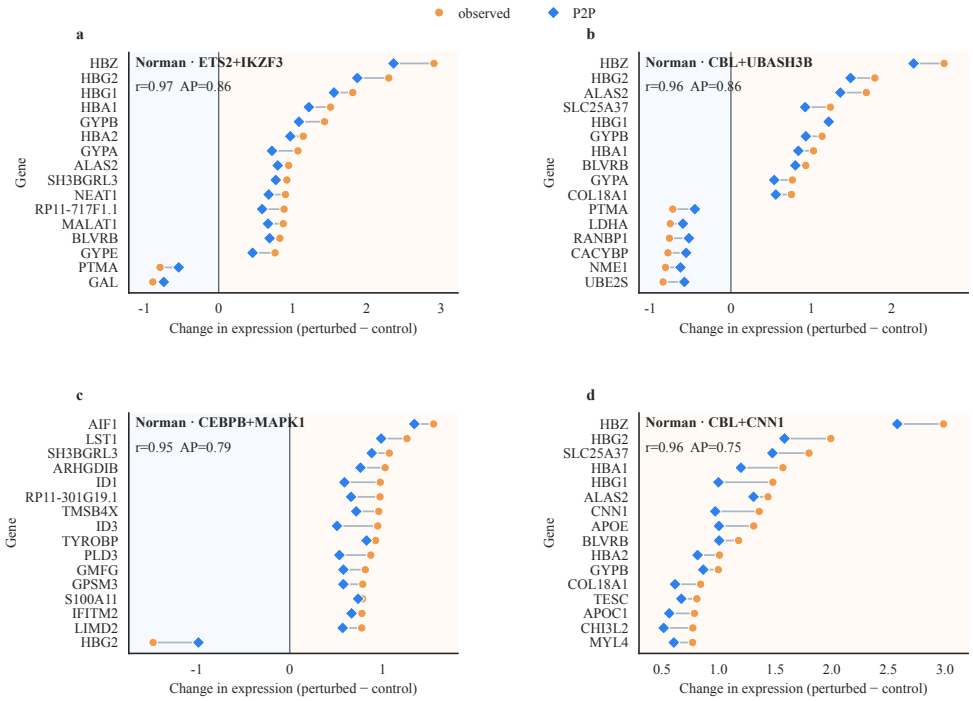


Figure 13: Combination-perturbation cases from Norman. The examples cover ETS2 plus IKZF3, CBL plus UBASH3B, CEBPB plus MAPK1, and CBL plus CNN1.

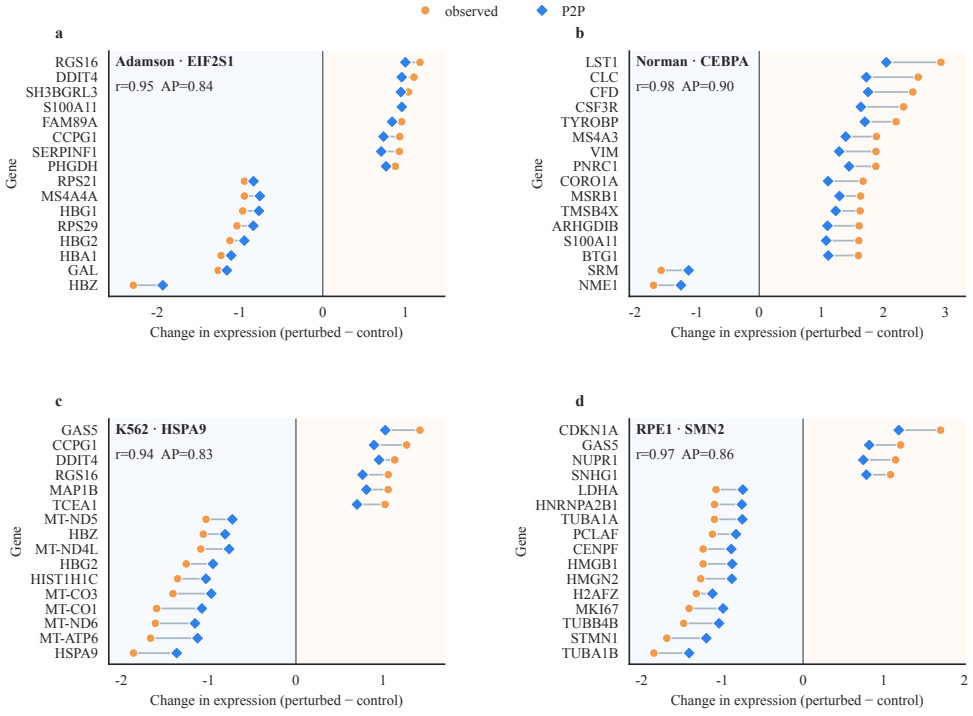


Figure 14: Ranked gene-effect comparisons for representative cases. The dot and line representation exposes the signed discrepancy for the largest observed effects.

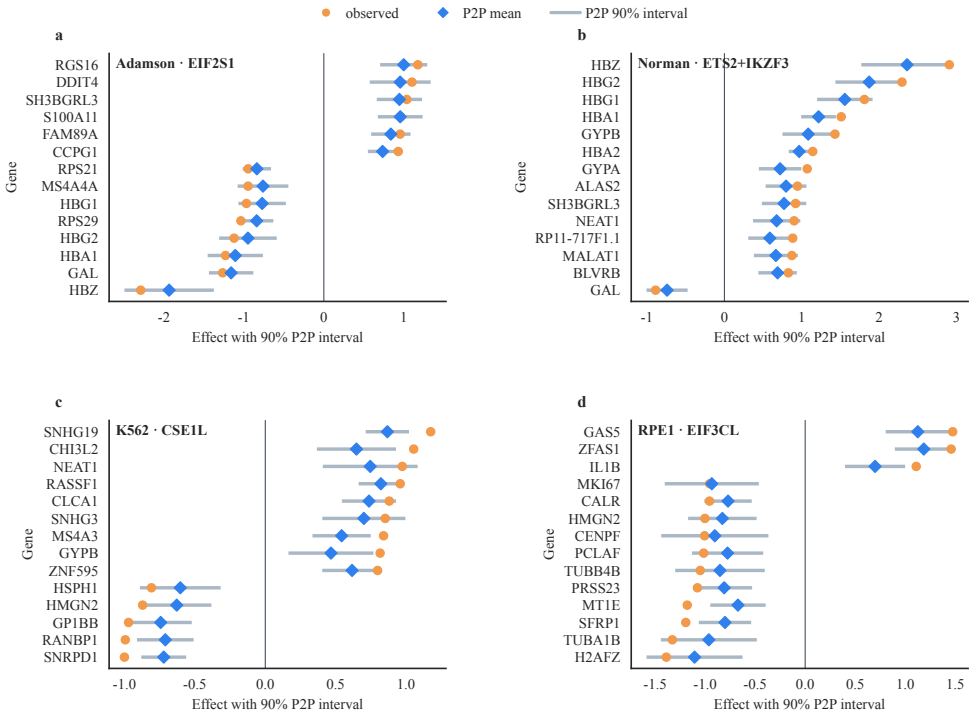


Figure 15: Gene-level predictive intervals. Observed effects are compared with P2P means and 90 percent predictive intervals for representative genes in four datasets.