GENESCOPE: ADAPTIVE GENE RESOLUTION FOR CHEMICAL PERTURBATION PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Predicting how cells respond to drugs could help researchers decide which experiments to run across a vast space of compounds and cellular contexts. Current methods learn perturbation effects in latent spaces, model changes in cell populations, or draw on pretrained cellular representations. Yet benchmarks of genetic perturbation prediction show that complex deep models do not consistently outperform simple linear baselines. These findings encourage a closer look at what models preserve about cells before treatment. Global compression can blur gene-level differences, but choosing which details to retain is difficult when the responding genes are still unknown. We introduce **GeneScope**, an adaptive-resolution Transformer built on *full coverage with selective refinement*. GeneScope keeps a compact view of all input genes, then uses untreated cells to identify where additional detail is needed. Genes that are poorly reconstructed or vary across controls receive finer-resolution representations. The controls thus guide where to preserve detail, while the compound guides how the state changes. Every gene can still respond, whether or not it is selected for refinement. We evaluate GeneScope on Tahoe and two Sci-Plex3-derived benchmarks. Shared-condition comparisons show improvements in response correlation and high-response gene recovery over linear and multiple deep-learning baselines. These gains hold without a pretrained cellular encoder. Adding pretrained representations further improves response rank correlation on both SciPlex benchmarks. GeneScope therefore ties gene resolution to the cellular state we can observe, rather than the response we have yet to predict. Code and preprocessing pipelines are available at an anonymous repository: `https://anonymous.4open.science/r/GeneScope/`.

1 INTRODUCTION

Single-cell perturbation screens reveal how compounds reshape gene expression (Srivatsan et al., 2020; McFarland et al., 2020; Chang et al., 2022; McFaline-Figueroa et al., 2024; Zhang et al., 2025). Learning to predict these responses could extend experimental screens to unmeasured conditions, helping researchers prioritize experiments and develop hypotheses about how drugs act on cells (Hetzel et al., 2022; Lotfollahi et al., 2023; Adduri et al., 2026; Qi et al., 2024; Yu et al., 2025; Alsulami et al., 2026; Feng et al., 2026). Realizing this potential, however, requires understanding the cellular state on which a drug acts. The same compound can produce different responses in different contexts, and even large screens cover only a small part of this space. The goal is therefore to generalize to new compounds, new treated cellular contexts, or both, given untreated cells from the target context.

Existing methods pursue this goal through latent perturbation shifts, molecular representations, and population-level prediction (Lotfollahi et al., 2019; 2023; Hetzel et al., 2022; Adduri et al., 2026). In parallel, cellular foundation models provide a complementary source of knowledge by learning transferable representations from large expression corpora (Theodoris et al., 2023; Cui et al., 2024; Hao et al., 2024). Yet model capacity alone is not a reliable shortcut to better predictions. Recent benchmarks of genetic perturbation prediction show that deep models can struggle to outperform simple linear and additive predictors (Ahlmann-Eltze et al., 2025). These findings bring a more basic question into focus, i.e., *what should a model preserve about cells before predicting how they will change?* Global compression captures a broad cellular state, but can obscure distinctions between individual genes. Representing every gene at high resolution retains these distinctions, while

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.

leaving the model to separate useful variation from noise and redundancy with limited perturbation supervision. A natural alternative is to preserve global context and selectively retain finer detail. The key question is where that detail should go. Choosing genes expected to respond to a compound creates a circular problem, since their identity is part of the outcome we want to predict.

A key insight is that allocating resolution need not depend on anticipating the response. Untreated cells already tell us where a compact representation falls short. Reconstruction residuals highlight poorly retained features, while variation across controls reveals heterogeneity within the starting population. We use these observable signals to guide refinement. This gives a simple division of labor, with controls guiding where to retain detail and the compound conditioning how the state changes. To this end, we introduce **GeneScope**, an adaptive-resolution Transformer built on *full coverage with selective refinement* (Figure 1). A compact global path covers the full input gene panel, while a fine-resolution path adds detail for genes selected from controls. A population-aware residual predictor combines these representations with a direct shared-response branch. The shared branch captures broad response trends, providing a foundation for gene-specific nonlinear corrections. All input genes retain a prediction path, including those not selected for refinement. Since selection uses untreated cells, the same rule also applies in new cellular contexts where treated observations are unavailable.

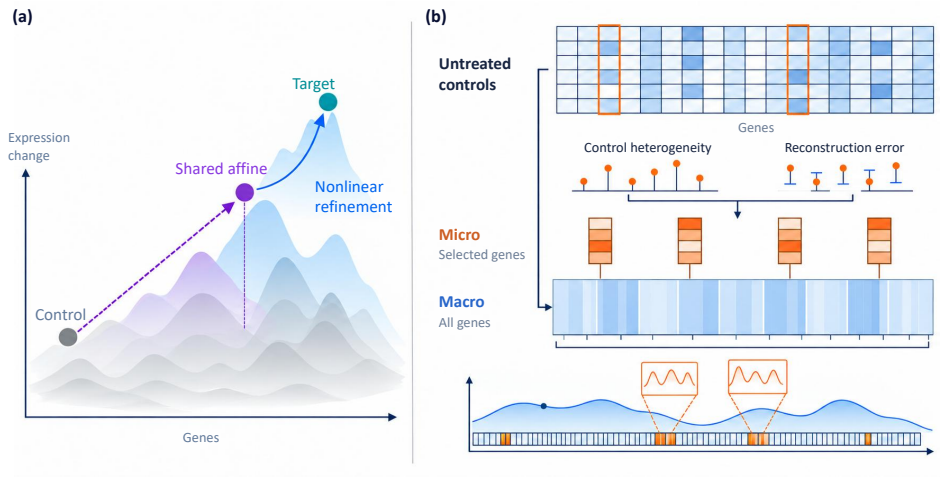


Figure 1: **Full coverage with selective refinement.** (a) A shared affine response provides a foundation for nonlinear refinement. (b) Macro features cover all input genes, while Micro features add detail at genes selected using control heterogeneity and reconstruction error. Selection allocates resolution without restricting which genes can respond. All visual examples are schematic.

We evaluate GeneScope on Tahoe and two Sci-Plex3-derived benchmarks. Shared-condition comparisons show improvements over ridge regression, chemCPA, and a reproduced STATE transition model (STATE-ST), including gains in response Pearson correlation and high-response gene recovery. On Tahoe, response Pearson reaches 0.522, compared with 0.431 for STATE-ST. On cellxgene–SciPlex3, high-response gene AUPRC reaches 0.107, compared with 0.073 for the same baseline. These gains do not require a pretrained cellular encoder, while adding pretrained representations further improves response rank correlation on both SciPlex tasks. The ablations also reveal a useful distinction between retaining more detail and retaining it effectively. These results support adaptive resolution as a useful inductive bias for perturbation prediction, i.e., preserve the global cellular context, and add gene-level detail where the observed starting state calls for it.

This work contributes *(i)* a control-guided approach to gene resolution, *(ii)* a residual architecture integrating global coverage, selective refinement, and population context, and *(iii)* a three-benchmark evaluation of predictive performance, generalization, and component contributions. We begin with related work in Section 2, before detailing the model and its training in Section 3. Section 4.1 introduces the experimental setting, followed by performance comparisons, transfer evaluations, and ablation analyses. Section 5 closes with the main findings and directions for future work.

2 Related Work

2.1 Cellular representations for virtual cell modeling

Single-cell foundation models support virtual cell modeling through reusable gene and cell representations. scBERT, Geneformer, scFoundation, and scGPT learn these from large expression corpora (Yang et al., 2022; Theodoris et al., 2023; Hao et al., 2024; Cui et al., 2024). Recent work enriches them with protein and regulatory knowledge (Rosen et al., 2026; Yang et al., 2024; Kalfon et al., 2025) alongside intercellular and spatial context (Wen et al., 2024; Tejada-Lapuerta et al., 2025). Others expand pretraining scale or gene-context length (Zeng et al., 2025; Bai et al., 2026). Their downstream benefits remain task- and data-dependent (Kedzierska et al., 2025; DenAdel et al., 2026). For perturbation prediction, these representations must preserve broad cellular variation and the gene-level distinctions needed to model treatment responses.

2.2 Predicting perturbation responses

scGen and CPA represent perturbations through latent shifts or composable effects (Lotfollahi et al., 2019; 2023), while chemCPA and PerturbNet incorporate molecular features to predict responses to unseen compounds (Hetzel et al., 2022; Yu et al., 2025). Other methods disentangle condition-associated variation or transfer effects across cellular states (Piran et al., 2024; Zhang et al., 2024; Kana et al., 2023; Jiang et al., 2024). To capture population heterogeneity, CellOT learns transport maps between unpaired populations, STATE models cell-set transitions, and PerturbDiff models response distributions through diffusion (Bunne et al., 2023; Adduri et al., 2026; Yuan et al., 2026). Genetic-perturbation prediction and effect identification address related questions under different intervention and input assumptions (Roohani et al., 2024; Wang et al., 2024; Dong et al., 2023).

Nevertheless, strong simple baselines remain an important reference. Ahlmann-Eltze et al. (2025) found that evaluated deep models did not consistently outperform linear or additive predictors on genetic-perturbation benchmarks. STATE likewise emphasizes linear and mean baselines for cross-context prediction (Adduri et al., 2026). Systema further shows that common metrics can reward systematic control–perturbation differences without establishing accurate recovery of perturbation-specific effects (Viñas Torné et al., 2025). Together, these studies motivate separating shared structure from condition-specific responses rather than equating model capacity with predictive accuracy.

2.3 Efficient and adaptive Transformers

Efficient Transformers compress representations or restrict token interactions to handle large inputs. Perceiver summarizes inputs through a latent bottleneck (Jaegle et al., 2021), while BigBird combines sparse local, random, and global connections (Zaheer et al., 2020). Routing Transformers make interaction patterns depend on token content (Roy et al., 2021). TokenLearner and DynamicViT learn which tokens to aggregate or retain for each input (Ryoo et al., 2021; Rao et al., 2021). Applying these ideas to perturbation prediction raises a question: what should guide refinement before the response is known? The observed cellular state reveals which details are difficult to represent, even when the responding genes remain unknown. Preserving those details and predicting which genes will change are distinct modeling decisions.

3 Method

Given untreated cells $X^0 = \{x_i^0\}_{i=1}^N$ over G genes and a chemical descriptor e_p encoding compound and available dose information, GeneScope predicts a treated population $\widehat{X}^p$. Control and treated cells are unpaired, so the model learns a population transformation rather than a supervised trajectory for each cell. Figure 2 shows the architecture. The Macro encoder gives every gene a compressed, drug-conditioned representation; the Micro encoder preserves additional detail for genes selected from untreated controls. A population encoder then contextualizes cell summaries. The output module combines these features with shared response terms and adds the predicted change to the original controls. Notably, GeneScope itself requires no pretrained cellular encoder, while the optional GeneScope + SE variant adds frozen state features through an output-scale adapter.

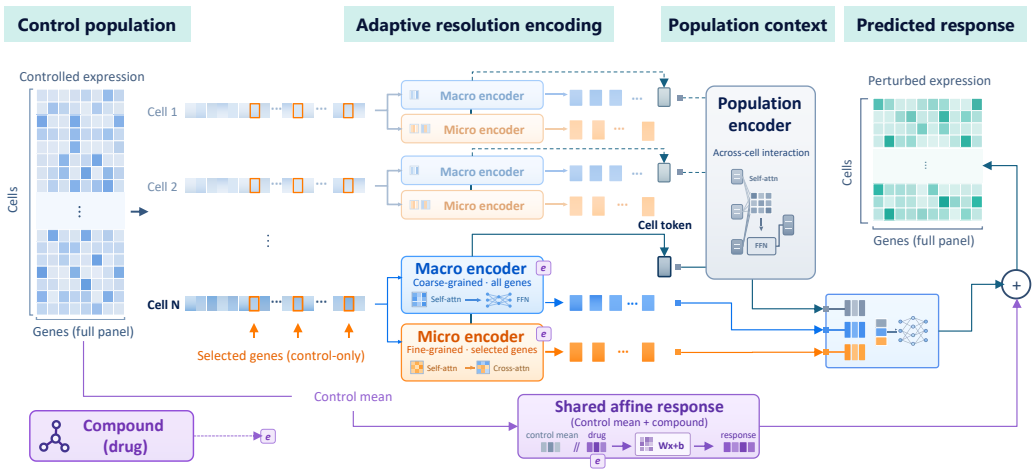


Figure 2: **GeneScope architecture.** Macro sketches cover all input genes; Micro tokens refine a control-selected subset using chemical and sketch context. Pooled Macro summaries enter a population Transformer. Macro and Micro gene features bypass pooling and join population features in the residual decoder. Shared response modeling comprises an affine branch and a separate nonlinear control-mean correction. Gated response composition adds the resulting change to the original controls. Repeated encoder diagrams denote shared processing across cells.

3.1 Macro Encoder: Full-Gene Coverage

We embed gene identity and expression as $h_{ig} = q_g + \phi(x^0_{ig})$, then average within R fixed partitions of the gene panel. For bin $B^{(r)}_m$ in view r, the compressed token and its gene-level reconstruction are

$$z^{(r)}_{im} = \psi_b\left(\frac{1}{|B^{(r)}_m|}\sum_{g\in B^{(r)}_m} h_{ig}\right), \qquad \widetilde{x}^{0,(r)}_{ig} = f_{\mathrm{rec}}([z^{(r)}_{i,b_r(g)}; q_g]), \tag{1}$$

where $b_r(g)$ denotes the bin assigned to gene g. Every view covers the complete panel. These partitions define computational groupings, without requiring them to correspond to biological pathways. The reconstruction head acts before drug conditioning and measures the information lost through compression.

A shared Transformer conditions each sketch on a projected chemical token, $U^{(r)}_i = T_{\mathrm{macro}}([\psi_p(e_p); Z^{(r)}_i])$. We aggregate views according to their control reconstruction quality:

$$\epsilon_{ir} = \frac{1}{G}\sum_g |\widetilde{x}^{0,(r)}_{ig} - x^0_{ig}|, \quad \alpha_{ir} = \mathrm{softmax}_r(-\epsilon_{ir}/\tau), \quad m_{ig} = \sum_r \alpha_{ir} U^{(r)}_{i,b_r(g)}. \tag{2}$$

Bin indexing excludes the chemical-token position. Thus each gene receives a global feature m_{ig} while attention operates on sketch tokens. We use three 300-bin views and $\tau = 0.25$, while the two additional partitions distribute genes evenly across bins.

3.2 Micro Encoder: Control-Guided Refinement

The Micro encoder allocates direct gene-level attention where the controls show heterogeneity or reconstruction error. Its scores are

$$v_g = \frac{1}{N}\sum_i (x^0_{ig} - \bar{x}^0_g)^2, \qquad u_g = \mathrm{median}_r\left[\frac{1}{N}\sum_i |\widetilde{x}^{0,(r)}_{ig} - x^0_{ig}|\right], \tag{3}$$

where $\bar{x}^0 = N^{-1}\sum_i x^0_i$. We select 60 genes by v_g, 100 by u_g, and 40 by $v_g + u_g$, excluding previously selected genes at each stage. The resulting set S is shared across cells in a population and

changes with the observed controls. For fixed controls and parameters, changing the compound does not change S and reconstruction error measures representation quality, not predicted drug sensitivity.

Projected embeddings of the selected genes undergo self-attention and cross-attention to the chemical token and the first complete Macro sketch. A learned elementwise gate interpolates between each projected input and its attention update, yielding f_{ig} for $g \in S$. This supplies a direct path for selected-gene detail while the other sketches continue to contribute through m_{ig}. Specifically, with $K = |S| = 200$ selected genes and $M = 300$ bins, Micro attention scales as $O(K^2 + KM)$, while Macro attention, population attention, and dense decoding contribute additional computational costs.

3.3 Population Encoder: Context Across Cells

Learned-query attention pools each drug-conditioned Macro sequence into a cell summary. Using the same view weights, we form $c_i = \sum_r \alpha_{ir} \operatorname{Pool}(U_i^{(r)})$ and apply a bidirectional population Transformer:

$$(o_1, \ldots, o_N) = T_{\text{pop}}(c_1, \ldots, c_N), \qquad p_i = o_i + h_{\text{basal}}(x_i^0). \tag{4}$$

The Transformer uses neither positional encodings nor learned cell-line identifiers, allowing cells to interact through their measured states with permutation-equivariant attention (Lee et al., 2019). The basal skip preserves a direct encoding of each control cell. Macro and Micro gene features bypass pooling, so the decoder retains gene-level detail alongside population context.

3.4 Shared Response and Residual Composition

Two complementary routes model changes shared across cells. An affine branch acts on the full control mean and chemical descriptor, while a bounded nonlinear branch acts on a fixed low-dimensional projection of the mean:

$$a = W_a[\bar{x}^0; e_p] + b_a, \qquad b = 0.25 \tanh\big(W_o h_\mu([\bar{x}^0 P; e_p])\big). \tag{5}$$

The affine map is learned jointly with the network. The separate control-mean correction b supplies nonlinear shared changes, with $P \in \mathbb{R}^{G \times 32}$ fixed at initialization.

The residual decoder predicts a dense response from Macro and population features and adds a correction at selected genes:

$$\begin{aligned} r_{ig} &= d_c([m_{ig}; p_i]) + \mathbf{1}[g \in S] d_f([f_{ig}; m_{ig}; p_i]), \\ \widehat{x}_{ig}^p &= \big[x_{ig}^0 + \gamma\{a_g + \eta(r_{ig} + b_g)\}\big]_+. \end{aligned} \tag{6}$$

Here $[t]_+ = \max(0, t)$. The scalar gate $\eta \in (0, 1)$ depends on the control mean and compound, while the scalar gain $\gamma \in [1/4, 4]$ depends on the compound and pooled population output. They regulate residual contribution and overall response magnitude, respectively. Every gene can change, including genes outside S, as well as selection limits refinement.

3.5 Optional Frozen State Features

GeneScope + SE uses the frozen STATE encoder (Adduri et al., 2026) to compute a 1,024-dimensional feature z_i^{SE} for each control cell. A trainable adapter combines their mean with the chemical descriptor:

$$s = 1 + 0.1 \tanh\big(f_{\text{SE}}([N^{-1} \textstyle\sum_i z_i^{\text{SE}}; e_p])\big), \qquad \widehat{x}_{ig}^{p,+} = \bar{x}_g^0 + s(\widehat{x}_{ig}^p - \bar{x}_g^0). \tag{7}$$

This condition-level scalar scales the complete backbone output around the control mean, including cell-level deviations, with mathematical range $[0.9, 1.1]$. The encoder remains frozen while the adapter and backbone are trained jointly. Cached features follow the sampled control rows. We evaluate its raw output consistently with the backbone comparison (Appendix C).

3.6 Training and Inference

Energy distance aligns predicted and observed treated populations without cell-level pairing. Defining $\widehat{\delta} = \operatorname{mean}(\widehat{X}^p) - \bar{x}^0$ and $\delta = \operatorname{mean}(X^p) - \bar{x}^0$, we optimize

$$\mathcal{L} = \mathcal{L}_{\text{energy}} + \frac{40}{G} \|\widehat{\delta} - \delta\|_2^2 + \mathcal{L}_{\text{auxiliary}} + \mathcal{L}_{\text{control}}. \tag{8}$$

For GeneScope + SE, prediction losses are computed from the final SE-calibrated expression profiles. The backbone and calibration module are trained jointly, while the pretrained cellular encoder remains frozen. The mean-response term directly constrains response magnitude, while auxiliary losses supervise correlation, direction, ranking, amplitude, and kernel distribution alignment. Meanwhile, control losses support sketch reconstruction and consistency of selection scores across two control subsets, while treated profiles supply training targets, never prediction-time routing inputs. Top-K indices are non-differentiable, while their shared input encoder is jointly optimized by prediction and control objectives.

Model training uses 256 cells per population and 5,000 optimizer updates. At inference, one forward pass on 256 sampled controls predicts 256 treated profiles across the fixed 2,000-gene panel. This is the population context size; prediction ensembles are specified separately. Appendix B gives optimization settings, complete loss definitions, and feature provenance.

4 EXPERIMENTS

We evaluate accuracy on shared conditions, generalization across complete test splits, and the contribution of individual modules.

4.1 DATASETS AND EVALUATION PROTOCOL

Tahoe is a Plate 3 subset of Tahoe-100M (Zhang et al., 2025), comprising 418,040 cells, 50 cell lines, and 92 non-control compounds. SciPlex-middle contains 354,640 cells in the chemCPA-distributed representation (Hetzel et al., 2022), while cellxgene–SciPlex3 contains 581,777 cells from CELLxGENE. Both derive from the same Sci-Plex3 screen (Srivatsan et al., 2020), testing representation robustness rather than independent biological replication. All models use a 2,000-gene panel; conditions identify cell lines and treatments, including dose for SciPlex. The test splits contain 646, 1,036, and 1,040 conditions, respectively, covering held-out compound–line pairs, unseen compounds, unseen treated cell lines, and double-unseen conditions, with target-context controls available throughout. Detailed counts and split construction appear in Appendix A.

Each GeneScope configuration uses three independent training runs of 5,000 updates, with validation-Pearson checkpoint selection. Inference uses fixed samples of 256 controls and never receives treated profiles; reference means use all native test controls and treated cells. We report condition-macro-averaged response Pearson Δr, Spearman $\Delta\rho$, Top-100 absolute-response overlap, sign accuracy on the observed top 100 (Dir.), and high-response gene average precision (AUPRC). Retrieval targets reflect response magnitude, not significance-tested differential expression. Appendix C additionally defines RMSE, top-response rank correlation, and AUROC.

Table 1 evaluates every method on identical condition identities: 529 on Tahoe, 536 on SciPlex-middle, and 248 on cellxgene–SciPlex3, determined by native support rather than metric values. Table 2 covers complete test splits and transfer strata, leaving unsupported outputs blank, and coverage-adjusted utilities are provided in Appendix H. In our experiments for model comparison, baselines include control mean, additive response, ridge, CPA, chemCPA, scGen, scVIDR, and STATE-ST. STATE-ST is a reproduced transition model trained from scratch with one-hot treatment identities, not the pretrained STATE encoder. Condition matching removes support mismatch but does not harmonize historical differences in normalization, reference sampling, or postprocessing. All scores come from retained predictions rather than published results; main tables report means, with sample standard deviations in the appendix.

4.2 PERFORMANCE AND GENERALIZATION

Under shared-condition evaluation, GeneScope combines strong response correlation with effective recovery of high-response genes, with particularly clear gains on Tahoe and cellxgene–SciPlex3. On Tahoe, response Pearson improves by approximately 21% over the reproduced STATE-ST model. The advantage in gene recovery is especially pronounced on cellxgene–SciPlex3, where AUPRC improves by 47% over STATE-ST and nearly doubles that of ridge regression. These gains show that a task-specific representation can capture useful response structure without a pretrained cellular

Table 1: Performance on exact shared condition identities. Higher is better; darker blue denotes a higher score within each dataset/metric. Bold/underline identify first/second distinct displayed means; rounded ties share rank. Means only; variability and all eight metrics are in Appendix D. Historical baseline preprocessing and reference differences remain (Section 4.1).

Method	Δr (↑)	$\Delta\rho$ (↑)	Top-100 (↑)	Dir. (↑)	AUPRC (↑)
Tahoe: 529 shared conditions					
Control mean	-0.002	0.113	0.467	0.390	0.442
Additive response	0.084	0.175	0.456	0.441	0.427
Ridge	0.124	0.188	0.446	0.448	0.419
CPA	0.168	0.213	0.396	0.593	0.337
chemCPA	0.163	0.215	0.398	0.591	0.343
scGen	0.401	**0.372**	0.401	0.760	0.365
scVIDR	0.228	0.242	0.392	0.622	0.349
STATE-ST	0.431	0.357	0.471	0.747	0.454
GeneScope	0.522	0.346	0.485	0.807	0.489
GeneScope + ensemble	**0.528**	0.355	0.489	**0.812**	0.494
GeneScope + SE	0.505	0.339	0.482	0.791	0.485
GeneScope + SE + ensemble	0.518	0.354	**0.490**	0.799	**0.496**
SciPlex-middle: 536 shared conditions					
Control mean	0.066	0.025	0.328	0.531	0.283
Additive response	0.075	0.044	0.328	**0.539**	0.287
Ridge	0.074	0.038	0.329	0.538	0.288
CPA	**0.086**	0.050	**0.420**	0.511	**0.394**
chemCPA	0.073	0.030	0.356	0.529	0.315
scGen	0.082	0.040	0.410	0.529	0.376
scVIDR	0.085	0.039	0.413	0.527	0.379
STATE-ST	0.078	0.058	0.360	0.528	0.326
GeneScope	0.084	0.064	0.414	0.533	0.368
GeneScope + ensemble	**0.086**	0.064	**0.420**	0.534	0.373
GeneScope + SE	0.082	0.085	0.413	0.514	0.379
GeneScope + SE + ensemble	0.084	**0.086**	**0.420**	0.515	0.385
cellxgene–SciPlex3: 248 shared conditions					
Control mean	0.028	0.028	0.055	0.557	0.053
Additive response	0.045	0.044	0.059	0.574	0.057
Ridge	0.042	0.041	0.058	0.573	0.055
chemCPA	0.016	0.012	0.019	0.548	0.040
scGen	0.028	0.028	0.017	0.548	0.039
scVIDR	0.028	0.032	0.017	0.545	0.038
STATE-ST	0.041	0.040	0.081	0.555	0.073
GeneScope	0.067	0.052	**0.145**	0.571	**0.107**
GeneScope + ensemble	0.068	0.053	**0.145**	0.572	**0.107**
GeneScope + SE	0.072	0.061	0.090	0.601	0.067
GeneScope + SE + ensemble	**0.073**	**0.062**	0.088	**0.603**	0.066

encoder. SciPlex-middle presents a closer comparison, i.e., GeneScope remains competitive with CPA in response correlation, while CPA leads in gene-recovery accuracy.

Pretrained control features offer a complementary source of information. On SciPlex-middle, GeneScope + SE improves both response rank correlation and gene recovery, with further gains from within-family ensembling. On cellxgene–SciPlex3, it instead favors correlation and direction accuracy at the expense of AUPRC. Thus, pretrained features can extend the base model's strengths, although their benefits depend on the prediction objective. Ensembling provides additional gains on Tahoe and SciPlex-middle; all reported ensembles average predictions from three runs within the same model family.

Table 2: Generalization evaluation: full-test Pearson/AUPRC and Pearson on disjoint unseen-compound (Drug), unseen-cell-line (Cell), and double-unseen (Both) strata. Each value covers its entire column-specific stratum; dashes denote missing predictions, not zero. No coverage adjustment is used. Baseline qualification and visual encoding follow Table 1; variability is in Appendix E.

Method	Full Δr (↑)	Full AP (↑)	Drug Δr (↑)	Cell Δr (↑)	Both Δr (↑)
Tahoe: full 646; drug 90, cell 249, both 27					
Ridge	0.133	0.421	0.185	0.080	0.154
scGen	—	—	—	0.293	—
STATE-ST	—	—	—	0.203	—
GeneScope	0.497	0.471	0.437	0.359	0.215
GeneScope + ensemble	**0.504**	0.476	0.445	**0.363**	0.218
GeneScope + SE	0.483	0.467	0.431	0.336	0.225
GeneScope + SE + ensemble	0.496	**0.477**	**0.446**	0.344	**0.232**
SciPlex-middle: full 1036; drug 152, cell 672, both 76					
Ridge	**0.135**	0.292	**0.287**	0.077	0.099
scGen	—	—	—	0.086	—
STATE-ST	—	—	—	0.081	—
GeneScope	0.103	0.359	0.136	0.086	0.109
GeneScope + ensemble	0.105	0.363	0.140	**0.088**	0.110
GeneScope + SE	0.105	0.369	0.149	0.085	0.109
GeneScope + SE + ensemble	0.108	**0.374**	0.153	0.087	**0.111**
cellxgene–SciPlex3: full 1040; drug 152, cell 676, both 76					
Ridge	**0.152**	0.087	**0.352**	0.073	0.081
scGen	—	—	—	0.061	—
STATE-ST	—	—	—	0.079	—
GeneScope	0.079	**0.115**	0.095	0.073	0.064
GeneScope + ensemble	0.080	**0.115**	0.096	0.074	0.065
GeneScope + SE	0.097	0.083	0.117	0.088	0.083
GeneScope + SE + ensemble	0.099	0.082	0.119	**0.089**	**0.084**

Beyond shared conditions, GeneScope provides predictions for every condition in all three test splits. Its accuracy advantage also extends to all three Tahoe generalization settings, outperforming ridge on unseen compounds, unseen treated cell lines, and double-unseen conditions. The separation is particularly pronounced when treated cell lines are held out, highlighting the value of conditioning predictions on the observed starting population. On cellxgene–SciPlex3, GeneScope + SE also surpasses ridge and STATE-ST in cell-only transfer, although ridge remains stronger for unseen-compound correlation on the SciPlex tasks. Drug-ID-based baselines lack trained representations for unseen compounds and therefore do not support zero-shot prediction for these compounds. Thus, we report each method's prediction coverage and compute metrics on supported test conditions, marking unsupported conditions as unavailable.

4.3 Ablation Study

We examine what each component contributes through three paired runs, matching each ablation to its full-model counterpart while holding the split, training budget, and evaluation samples fixed. Table 3 summarizes the resulting changes.

The shared affine branch provides a complementary foundation, i.e., it supports response correlation on Tahoe and cellxgene–SciPlex3 and improves expression-scale accuracy on SciPlex-middle, where removing it nevertheless benefits correlation and retrieval. The stream reductions further clarify how global coverage and local detail work together. On both SciPlex benchmarks, retaining only Macro or only Micro features reduces correlation and gene recovery, favoring their combined representation. Micro-only remains competitive on Tahoe, suggesting that the benefit of combining resolutions depends on the dataset. Population encoding contributes most consistently to expression-scale

Table 3: Ablation study: paired changes relative to full GeneScope, averaged across three runs under matched test conditions. Positive Pearson/AP changes and negative RMSE changes indicate improvements. Darker red indicates greater performance degradation, normalized within each column; improvements and displayed zeros are unshaded. Bold and underlined values mark the best and second-best changes, respectively.

	Tahoe			SciPlex-middle			cellxgene–SciPlex3		
Intervention	Δr (↑)	AP (↑)	RMSE (↓)	Δr (↑)	AP (↑)	RMSE (↓)	Δr (↑)	AP (↑)	RMSE (↓)
No affine response	-0.108	-0.030	+0.016	+0.009	+0.026	+0.030	-0.010	-0.012	+0.003
One sketch	-0.012	-0.009	+0.008	-0.001	+0.009	+0.004	-0.002	+0.000	+0.000
No population encoder	+0.006	-0.006	+0.010	+0.000	+0.015	+0.004	+0.000	+0.013	+0.002
Macro only	-0.027	-0.025	+0.004	-0.009	-0.004	-0.006	-0.010	-0.008	-0.001
Micro only	+0.005	+0.000	-0.005	-0.005	-0.014	-0.004	-0.009	-0.006	-0.001

accuracy, as removing it increases RMSE across all three datasets, while its effects on correlation and retrieval are less uniform. Exact intervention definitions, including changes to dependent selection and pooling operations, and complete results are provided in Appendix F.

Table 4: Resolution budget on Tahoe validation (362 conditions): means over three runs, 5,000 updates each. The full output panel stays fixed at 2,000 genes.

K	Δr (↑)	Top-100 (↑)	AUPRC (↑)
100	0.653	<u>0.527</u>	<u>0.540</u>
200	0.665	**0.528**	**0.545**
300	0.663	0.523	0.539
400	0.665	0.522	0.536
500	**0.670**	0.521	0.532
600	<u>0.667</u>	0.519	0.530
700	0.660	0.517	0.525

The resolution sweep complements these module ablations by examining how much fine-grained detail to retain. We vary K from 100 to 700 in three-run Tahoe validation experiments and use $K = 200$ as a common operating point across datasets.

5 Conclusion

GeneScope builds on a simple yet effective principle, namely allocating gene-level detail according to the observed cellular state rather than anticipating which genes will respond. By pairing global coverage with control-guided refinement, it preserves gene-level detail where needed while leaving every gene free to respond. This design delivers strong predictive performance on Tahoe and two SciPlex benchmarks, with matched interventions highlighting the value of its direct fine-resolution prediction path. Building on these findings, we see opportunities to make adaptive gene resolution more transferable and better understand where it brings the greatest benefit. First, extending adaptive resolution to cellular pretraining could provide transferable starting-state representations, helping downstream models learn treatment-induced changes from limited perturbation data. Second, evaluation on independent biological screens would test whether the current gains extend across experimental settings. Finally, extending the framework to time-resolved perturbation modeling could allow gene resolution to evolve with cellular state, preserving different gene-level details as responses unfold and moving beyond endpoint prediction toward response trajectories.

AI USE STATEMENT

Generative AI tools assisted with manuscript revision, implementation of partial experimental code and scripts, review of code and experimental records, aggregation of evaluation results into tables, and consistency checks. The authors reviewed the AI-assisted outputs and take full responsibility for the manuscript, code, and reported results.

ETHICS STATEMENT

This study analyzes public cell-line perturbation datasets through computational prediction. Predicted responses are research hypotheses, not validated intervention outcomes or clinical recommendations. Expression prediction does not establish compound safety or efficacy.

REPRODUCIBILITY STATEMENT

Our implementation is available at the code link provided earlier in the paper. Section 3 defines the model, while Appendices A and B detail preprocessing, data splits, architecture, and optimization. Appendices C and D provide metric definitions, coverage handling, native-support scores, and run-to-run variation. Appendices F and G document component ablations, pretraining, and within-family prediction ensembles, identifying analyses conducted only on validation data.

REFERENCES

Abhinav K. Adduri, Dhruv Gautam, Beatrice Bevilacqua, Mohsen Naghipourfar, Alishba Imran, Rohan Shah, Noam Teyssier, Rishi Verma, Christopher Carpenter, Basak Eraslan, Francis Chalissery, Rajesh Ilango, Vishvak Subramanyam, Chiara Ricci-Tam, Sanjay Nagaraj, Aidan Winters, Mingze Dong, Stefanie Fellinger, Adam Krejci, Tilmann Burckstummer, Sravya Tirukkovular, Jeremy Sullivan, Brian S. Plosky, Nicholas D. Youngblut, Jure Leskovec, Luke A. Gilbert, Silvana Konermann, Patrick D. Hsu, Alexander Dobin, Dave P. Burke, Hani Goodarzi, and Yusuf H. Roohani. Predicting cellular responses to perturbation across diverse contexts with State. *Cell*, 189(19):5914–5931.e20, 2026. doi: 10.1016/j.cell.2026.07.052.

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

Reem Alsulami, Robert Lehmann, Sumeer A Khan, Vincenzo Lagani, Alberto Maillo, David Gomez-Cabrero, Narsis A Kiani, and Jesper Tegner. Predicting and interpreting cell-type-specific drug responses in the small-data regime using inductive priors. *Nature Machine Intelligence*, 8(3): 461–473, 2026.

Ding Bai, Shentong Mo, Ruiyi Zhang, Yingtao Luo, Jiahao Gao, et al. scLong: a billion-parameter foundation model for capturing long-range gene context in single-cell transcriptomics. *Nature Communications*, 17:2380, 2026. doi: 10.1038/s41467-026-69102-y.

Charlotte Bunne, Stefan G. Stark, Gabriele Gut, et al. Learning single-cell perturbation responses using neural optimal transport. *Nature Methods*, 20:1759–1768, 2023. doi: 10.1038/s41592-023-01969-x.

Matthew T Chang, Frances Shanahan, Thi Thu Thao Nguyen, Steven T Staben, Lewis Gazzard, Sayumi Yamazoe, Ingrid E Wertz, Robert Piskol, Yeqing Angela Yang, Zora Modrusan, et al. Identifying transcriptional programs underlying cancer drug response with trace-seq. *Nature Biotechnology*, 40(1):86–93, 2022.

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scGPT: toward building a foundation model for single-cell multi-omics using generative ai. *Nature Methods*, 21:1470 – 1480, 2024.

Alan DenAdel, Madeline Hughes, Akshaya Thoutam, Anay Gupta, Andrew W. Navia, Nicolo Fusi, Srivatsan Raghavan, Peter S. Winter, Ava P. Amini, and Lorin Crawford. Evaluating the role of pretraining dataset size and diversity on single-cell foundation model performance. *Nature Methods*, 23:1447–1457, 2026. doi: 10.1038/s41592-026-03120-y.

Mingze Dong, Bao Wang, Jessica Wei, Antonio H. de O. Fonseca, Curtis J. Perry, Alexander Frey, Feriel Ouerghi, Ellen F. Foxman, Jeffrey J. Ishizuka, Rahul M. Dhodapkar, and David van Dijk. Causal identification of single-cell experimental perturbation effects with CINEMA-OT. *Nature Methods*, 20:1769–1779, 2023. doi: 10.1038/s41592-023-02040-5.

Jinghao Feng, Ziheng Zhao, Xiaoman Zhang, Mingfei Liu, Jingyi Chen, Xingran Quan, Boyang Fu, Jian Zhang, Yanfeng Wang, Ya Zhang, et al. A knowledge-driven framework for predicting single-cell responses for unprofiled drugs. *Nature Machine Intelligence*, pp. 1–14, 2026.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Scholkopf, and Alex Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13:723–773, 2012.

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

Leon Hetzel, Simon Böhm, Niki Kilbertus, Stephan Günnemann, Mohammad Lotfollahi, and Fabian J. Theis. Predicting cellular responses to novel drug perturbations at a single-cell resolution. *Advances in Neural Information Processing Systems 35*, 2022.

Andrew Jaegle, Felix Gimeno, Andrew Brock, Andrew Zisserman, Oriol Vinyals, and João Carreira. Perceiver: General perception with iterative attention. In *International Conference on Machine Learning*, pp. 4651–4664. PMLR, 2021.

Qun Jiang, Shengquan Chen, Xiaoyang Chen, and Rui Jiang. scPRAM accurately predicts single-cell gene expression perturbation response based on attention mechanism. *Bioinformatics*, 40(5): btae265, 2024. doi: 10.1093/bioinformatics/btae265.

Jérémie Kalfon, Jules Samaran, Gabriel Peyré, and Laura Cantini. scPRINT: pre-training on 50 million cells allows robust gene network predictions. *Nature Communications*, 16:3607, 2025. doi: 10.1038/s41467-025-58699-1.

Omar Kana, Rance Nault, David Filipovic, Daniel Marri, Tim Zacharewski, and Sudin Bhattacharya. Generative modeling of single-cell gene expression for dose-dependent chemical perturbations. *Patterns*, 4(8):100817, 2023. doi: 10.1016/j.patter.2023.100817.

Kasia Z. Kedzierska, Lorin Crawford, Ava P. Amini, and Alex X. Lu. Zero-shot evaluation reveals limitations of single-cell foundation models. *Genome Biology*, 26:101, 2025. doi: 10.1186/s13059-025-03574-x.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set Transformer: A framework for attention-based permutation-invariant neural networks. In *International Conference on Machine Learning*, pp. 3744–3753. PMLR, 2019.

Mohammad Lotfollahi, F. Alexander Wolf, and Fabian J Theis. scGen predicts single-cell perturbation responses. *Nature Methods*, 16:715 – 721, 2019.

Mohammad Lotfollahi, Anna Klimovskaia Susmelj, Carlo De Donno, Leon Hetzel, Yuge Ji, Ignacio L Ibarra, Sanjay R Srivatsan, Mohsen Naghipourfar, Riza M. Daza, Beth K. Martin, Jay A. Shendure, José L. McFaline-Figueroa, Pierre Boyeau, F. Alexander Wolf, Nafissa Yakubova, Stephan Günnemann, Cole Trapnell, David Lopez-Paz, and Fabian J Theis. Predicting cellular responses to complex perturbations in high-throughput screens. *Molecular Systems Biology*, 19, 2023.

José L McFaline-Figueroa, Sanjay Srivatsan, Andrew J Hill, Molly Gasperini, Dana L Jackson, Lauren Saunders, Silvia Domcke, Samuel G Regalado, Paul Lazarchuck, Sarai Alvarez, et al. Multiplex single-cell chemical genomics reveals the kinase dependence of the response to targeted therapy. *Cell genomics*, 4(2), 2024.

James M McFarland, Brenton R Paolella, Allison Warren, Kathryn Geiger-Schuller, Tsukasa Shibue, Michael Rothberg, Olena Kuksenko, William N Colgan, Andrew Jones, Emily Chambers, et al. Multiplexed single-cell transcriptional response profiling to define cancer vulnerabilities and therapeutic mechanism of action. *Nature communications*, 11(1):4296, 2020.

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

Xiaoning Qi, Lianhe Zhao, Chenyu Tian, Yueyue Li, Zhen-Lin Chen, Peipei Huo, Runsheng Chen, Xiaodong Liu, Baoping Wan, Shengyong Yang, et al. Predicting transcriptional responses to novel chemical perturbations using deep generative model for drug discovery. *Nature Communications*, 15(1):9256, 2024.

Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. *Advances in neural information processing systems*, 34:13937–13949, 2021.

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

Yanay Rosen, Yusuf Roohani, Ayush Agrawal, Leon Samotorčan, Tabula Sapiens Consortium, Stephen R. Quake, and Jure Leskovec. Universal cell embeddings: A foundation model for cell biology. *bioRxiv*, 2026. doi: 10.1101/2023.11.28.568918. Version 3; first posted in 2023.

Aurko Roy, Mohammad Taghi Saffar, Ashish Vaswani, and David Grangier. Efficient content-based sparse attention with routing transformers. *Transactions of the Association for Computational Linguistics*, 9:53–68, 2021.

Michael Ryoo, AJ Piergiovanni, Anurag Arnab, Mostafa Dehghani, and Anelia Angelova. Tokenlearner: Adaptive space-time tokenization for videos. *Advances in neural information processing systems*, 34:12786–12797, 2021.

Sanjay R. Srivatsan, Jose L. McFaline-Figueroa, Vijay Ramani, et al. Massively multiplex chemical transcriptomics at single-cell resolution. *Science*, 367(6473):45–51, 2020. doi: 10.1126/science.aax6234.

Alejandro Tejada-Lapuerta, Anna C. Schaar, Robert Gutgesell, Giovanni Palla, Lennard Halle, Mariia Minaeva, Larsen Vornholz, Leander Dony, Francesca Drummer, Till Richter, Mojtaba Bahrami, and Fabian J. Theis. Nicheformer: a foundation model for single-cell and spatial omics. *Nature Methods*, 22:2525–2538, 2025. doi: 10.1038/s41592-025-02814-z.

Christina V. Theodoris, Ling Xiao, Anant Chopra, Mark D. Chaffin, Zeina R. Al Sayed, Matthew C. Hill, Helene Mantineo, Elizabeth M. Brydon, Zexian Zeng, X. Shirley Liu, and Patrick T. Ellinor. Transfer learning enables predictions in network biology. *Nature*, 618:616–624, 2023.

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*, 2025. doi: 10.1038/s41587-025-02777-8. Advance online publication.

Gefei Wang, Tianyu Liu, Jia Zhao, Youshu Cheng, and Hongyu Zhao. Modeling and predicting single-cell multi-gene perturbation responses with scLAMBDA. *bioRxiv*, 2024. doi: 10.1101/2024.12.04.626878.

Hongzhi Wen, Wenzhuo Tang, Xinnan Dai, Jiayuan Ding, Wei Jin, Yuying Xie, and Jiliang Tang. CellPLM: Pre-training of cell language model beyond single cells. In *International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=BKXvPDekud`.

Fan Yang, Wenchuan Wang, Fang Wang, Yuan Fang, Duyu Tang, Junzhou Huang, Hui Lu, and Jianhua Yao. scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell rna-seq data. *Nature Machine Intelligence*, 4:852–866, 2022. doi: 10.1038/s42256-022-00534-z.

Xiaodong Yang, Guole Liu, Guihai Feng, Dechao Bu, Pengfei Wang, et al. GeneCompass: deciphering universal gene regulatory mechanisms with a knowledge-informed cross-species foundation model. *Cell Research*, 34:830–845, 2024. doi: 10.1038/s41422-024-01034-y.

Hengshi Yu, Weizhou Qian, Yuxuan Song, and Joshua D. Welch. PerturbNet predicts single-cell responses to unseen chemical and genetic perturbations. *Molecular Systems Biology*, 2025. doi: 10.1038/s44320-025-00131-3.

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

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, et al. Big bird: Transformers for longer sequences. In *Advances in Neural Information Processing Systems*, volume 33, pp. 17283–17297, 2020.

Yuansong Zeng, Jiancong Xie, Ningyuan Shangguan, Zhuoyi Wei, Wenbing Li, et al. CellFM: a large-scale foundation model pre-trained on transcriptomics of 100 million human cells. *Nature Communications*, 16:4679, 2025. doi: 10.1038/s41467-025-59926-5.

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

Ziqi Zhang, Xinye Zhao, Mehak Bindra, Peng Qiu, and Xiuwei Zhang. scDisInFact: disentangled learning for integration and prediction of multi-batch multi-condition single-cell RNA-sequencing data. *Nature Communications*, 15:912, 2024. doi: 10.1038/s41467-024-45227-w.

A Dataset Construction and Statistical Units

Source representations. Tahoe is a curated Plate 3 subset of Tahoe-100M, with 50 cell lines and 92 non-control compounds. SciPlex-middle is the chemCPA-distributed representation, containing 187 base compounds and 748 compound–dose labels. The CELLxGENE representation contains 188 compounds and 752 labels. Both originate from the same Sci-Plex3 screen: A549, MCF7, and K562 cells exposed for 24 hours at 0.01, 0.1, 1, and $10\,\mu\mathrm{M}$, with two biological replicates (Srivatsan et al., 2020). Replicate cells are pooled into cell-line–compound–dose conditions. Cells, conditions, biological replicates, and fitted models are distinct statistical units. The SciPlex representations are not independent biological replications.

Table 5: Dataset inventory before internal validation. Cell counts include controls; condition counts include treated conditions only. All predictors use a 2,000-gene output panel. Source genes refer to the supplied representation.

	Tahoe	SciPlex-middle	cellxgene–SciPlex3
Source cells	418,040	354,640	581,777
Source genes	62,710	2,000	54,780
Outer-training cells	352,464	219,187	364,633
Test cells	65,576	135,453	217,144
Outer-training conditions	3,621	1,208	1,216
Test conditions	646	1,036	1,040

Outer splits. Tahoe withholds treated responses in CVCL_0459, CVCL_1055, and CVCL_1635. Nine compounds are absent from training responses. Held-out pairs comprise 56 seen compounds in five seen lines; compound-only transfer comprises nine unseen compounds in ten seen lines. The inclusive unseen-line subset contains all 92 compounds in the three held-out lines. Both SciPlex representations withhold treated K562 cells and 19 base compounds at every dose. Another 34 compounds are held out in A549 but retained in MCF7. All K562 controls and 20% of A549/MCF7 controls belong to the test partition. Untreated cells from a target context remain available to the predictor.

Table 6: Test cells / treated conditions. The first three rows are mutually exclusive. Double-unseen conditions are nested within the inclusive unseen-line row and separated from cell-only transfer in the main text.

Subset	Tahoe	SciPlex-middle	cellxgene–SciPlex3
Held-out pair	27,951 / 280	15,085 / 136	25,030 / 136
Unseen compound only	8,197 / 90	29,240 / 152	43,433 / 152
Unseen line, inclusive	27,554 / 276	85,840 / 748	143,393 / 752
Test controls (cells)	1,874	5,288	5,288
Double unseen (conditions)	27	76	76

Internal validation and preprocessing. A training condition needs at least ten treated cells and a control pool. This retains 3,620/1,205/1,216 conditions. Reserving $\max(1, \lfloor 0.1C \rfloor)$ gives internal train/validation counts of 3,258/362, 1,085/120, and 1,095/121. The split seed is 913 and validation sampling seed 914, fixed across runs. Checkpoints maximize validation response Pearson. Newly fitted preprocessing uses the outer training partition; the SciPlex-middle panel is inherited from released preprocessing. SciPlex-middle expression is retained as supplied; aligned Tahoe and cellxgene counts use the configured normalization. The test split was inspected in earlier development, not newly blinded.

B Architecture and Optimization Details

B.1 Encoder Configuration and Control-Side Selection

All primary GeneScope and GeneScope + SE models use a fixed panel of 2,000 genes, 64-dimensional identity/expression embeddings, 128-dimensional Macro and Micro tokens, and 512-dimensional cell representations. Each Macro Transformer has two layers and four attention heads, with parameters shared across the three sketches. The bin projection ψ_b is linear followed by LayerNorm; the chemical projection ψ_p additionally applies GELU. Micro uses one Transformer decoder layer with four heads: selected gene tokens attend to each other, then to the projected chemical token and the 300 drug-conditioned tokens from the first sketch. The population encoder has four bidirectional layers, eight heads, and an intermediate width of 2,048. Its LLaMA-style blocks use no rotary positional embeddings or causal mask. Dropout is 0.1 and the chemical descriptor has 768 dimensions. No learned cell-line embedding is used.

The SciPlex chemical caches concatenate 704 Morgan fingerprint bits and 64 radial-basis dose features. Tahoe uses its supplied 768-dimensional chemical cache, not a newly trained chemical encoder. The descriptor projections inside GeneScope are learned.

The first sketch retains the dataset's cached gene partition. Two additional partitions use fixed random permutations with seeds 914 and 915. In each additional partition, 200 bins contain seven genes and 100 bins contain six genes; every gene belongs to exactly one bin per view. The seed used to construct these views is independent of the training seed. Reliability weights use the mean absolute control reconstruction error and temperature 0.25.

Gene selection uses population variance with denominator N and the median, across views, of per-gene mean absolute reconstruction error. Selection proceeds sequentially with quotas 60, 100, and 40 for dispersion, reconstruction error, and their sum. Occupied indices are masked before each selection. The primary configurations provide no separate gene-coverage prior, so the third score is exactly $v_g + u_g$. The selector has no trainable routing parameters. Its discrete indices do not receive gradients, although the reconstruction scores depend on the jointly trained gene and reconstruction encoders.

Let t_{ig} be the projected embedding of a selected gene and $\widetilde{f}_{ig}$ its Micro attention output. The implemented update is

$$f_{ig} = t_{ig} + \sigma\big(h_{\mathrm{gate}}([t_{ig}; \psi'_p(e_p)])\big) \odot (\widetilde{f}_{ig} - t_{ig}). \tag{9}$$

The gate is elementwise and drug-conditioned; selection itself remains control-only.

B.2 Shared Terms and Response Gates

The shared affine branch is one learned linear layer with bias from $[\bar{x}^0; e_p]$ to the complete gene panel. It is initialized to zero. The control-mean correction uses a fixed Gaussian matrix $P \in \mathbb{R}^{2000\times 32}$ with unit-normalized columns and seed 913. Only the projected mean is active in the primary model. The implementation also reserves slots for variance and skewness, but sets these slots to zero. A two-layer SiLU network of widths 128 and 32 feeds a bias-free 2,000-gene output layer, initialized to zero, followed by $0.25\tanh$. Thus this correction is nonlinear and shared across cells; it is not part of the affine map.

Writing $\bar{o} = N^{-1}\sum_i o_i$, the condition-level composition parameters in Equation 6 are

$$\eta = \sigma\big(h_\eta([h_{\mathrm{ctrl}}(\bar{x}^0); e_p])\big), \qquad \gamma = \exp(\mathrm{clip}(h_\gamma([e_p; \bar{o}]), -\log 4, \log 4)). \tag{10}$$

The final layer for η has zero weights and bias -1, so η initially equals $\sigma(-1)$. The gain head has zero final-layer weights and bias, giving $\gamma = 1$ at initialization. The two terms shared across cells and the cell-specific residual are learned together; no orthogonality or zero-population-mean constraint is imposed on them.

B.3 Training Objective

All losses are computed per condition and averaged over a batch. Let $\widehat{Y} = \{\widehat{y}_i\}_{i=1}^n$ and $Y = \{y_j\}_{j=1}^m$ denote predicted and observed treated populations. For GeneScope + SE, $\widehat{Y}$ is the output after the SE

scale adapter. The energy objective uses the GeomLoss energy-kernel convention:

$$\mathcal{L}_{\text{energy}} = \frac{1}{nm}\sum_{i,j}\|\widehat{y}_i - y_j\|_2 - \frac{1}{2n^2}\sum_{i,i'}\|\widehat{y}_i - \widehat{y}_{i'}\|_2 - \frac{1}{2m^2}\sum_{j,j'}\|y_j - y_{j'}\|_2. \tag{11}$$

The condition means define the control-relative responses $\widehat{\delta} = n^{-1}\sum_i \widehat{y}_i - \bar{x}^0$ and $\delta = m^{-1}\sum_j y_j - \bar{x}^0$. Their mean squared error is $\mathcal{L}_{\text{mean}} = G^{-1}\|\widehat{\delta} - \delta\|_2^2$. With ρ denoting Pearson correlation across genes, the remaining response losses are

$$\mathcal{L}_{\text{Pearson}} = 1 - \rho(\widehat{\delta}, \delta), \tag{12}$$

$$\mathcal{L}_{\text{direction}} = \frac{1}{G}\sum_g \text{BCEWithLogits}(10\widehat{\delta}_g, \mathbf{1}[\delta_g > 0]), \tag{13}$$

$$\mathcal{L}_{\text{rank}} = 1 - \rho(R_{0.1}(\widehat{\delta}_T), R_{0.1}(\delta_T)), \tag{14}$$

$$\mathcal{L}_{\text{amplitude}} = \text{SmoothL1}\Big(\log(A(\widehat{\delta}) + 10^{-6}), \log(A(\delta) + 10^{-6})\Big), \tag{15}$$

where T contains the 100 genes with largest $|\delta_g|$, $R_\tau(v)_g = \sum_h \sigma((v_g - v_h)/\tau)$, and $A(v) = \sqrt{G^{-1}\sum_g v_g^2 + 10^{-8}}$. The target-derived set T is used only in the training loss and never supplies Micro routing.

The Gaussian-kernel MMD loss (Gretton et al., 2012) uses $k(u, v) = \exp(-\|u - v\|_2^2/2)$ and the biased estimator $\text{mean}(k(\widehat{Y}, \widehat{Y})) + \text{mean}(k(Y, Y)) - 2\,\text{mean}(k(\widehat{Y}, Y))$, clamped below at zero. For efficiency, distances use the 500 genes with highest summed predicted and observed within-population variances, averaged over the batch. This gene subset also belongs only to the training objective.

The reconstruction loss averages SmoothL1 error over all control cells, views, and genes. For selector stability, split the input control population into two equal halves, recompute their scores $q^{(a)}$ and $q^{(b)}$, subtract each score vector's gene mean, and normalize its Euclidean norm with a 10^{-6} floor. The stability loss is $1 - \langle \widetilde{q}^{(a)}, \widetilde{q}^{(b)} \rangle$. In the primary configuration, $q = 2(v + u)$ because the coverage score falls back to $v + u$. The complete objective is

$$\begin{aligned}\mathcal{L} = \mathcal{L}_{\text{energy}} &+ 40\mathcal{L}_{\text{mean}} + 0.2\mathcal{L}_{\text{Pearson}} + 0.25\mathcal{L}_{\text{direction}} + 0.1\mathcal{L}_{\text{MMD}} \\ &+ 0.35\mathcal{L}_{\text{rank}} + 0.12\mathcal{L}_{\text{amplitude}} + 0.03\mathcal{L}_{\text{reconstruction}} + 0.02\mathcal{L}_{\text{stability}}.\end{aligned} \tag{16}$$

During the first two epochs, the Pearson, direction, and MMD weights are set to zero. The energy, mean-response, ranking, amplitude, reconstruction, and stability terms remain active.

B.4 Optimization and Inference

We use AdamW with learning rate 2.5×10^{-4}, weight decay 0.01, gradient-norm clipping at 1.0, and BF16 mixed precision. The learning rate warms up for two epochs and then follows cosine decay over a budget of 5,000 optimizer updates. Each training sample contains 256 control and 256 treated cells from a condition; the configured per-process batch size is two conditions across four distributed processes, giving eight conditions per optimizer update. Primary runs use seeds 1000, 1001, and 1002. These settings apply to the current GeneScope comparison and its matched component interventions.

The final native evaluator draws 256 controls with replacement for each condition and performs one model forward pass. The same sampled rows select the SE features when required. Historical configuration fields specifying a different inference context do not override this evaluator. Predictions cover all 2,000 genes. Within each condition, the evaluator records both the raw population mean and the mean after clamping each predicted cell entry to zero. GeneScope already applies this clamp inside its backbone. GeneScope + SE can produce negative entries after scaling, so its two recorded means can differ. The metric definitions in Appendix C specify which stored output is used for each reported score. Projection is applied before population averaging and does not clamp the signed control-relative response.

B.5 Frozen SE Features and Provenance

The optional feature source is the STATE SE-100M checkpoint at asset revision `bc72702320639217128df42673e94a9658f67d24`. Its model weights, protein-embedding

table, and configuration are held fixed. Control-cell features have 1,024 dimensions and are detached from GeneScope's training graph. The scale head normalizes the mean SE vector, projects it to 64 dimensions with a SiLU activation, concatenates the 768-dimensional chemical descriptor, applies another 64-dimensional SiLU layer, and predicts one scalar without output bias. Zero initialization of that output layer gives $s = 1$ initially. The bound in Equation 7 is 0.1; finite-precision rounding can shift the represented endpoints slightly beyond their real-valued limits.

For the two SciPlex representations, current caches encode actual control rows from the corresponding source and split. SciPlex-middle supplies its existing log-transformed expression; cellxgene–SciPlex3 supplies raw counts. Native case-sensitive gene mapping finds 1,324 and 1,918 supported genes, respectively, within their 2,000-gene panels. The SE tokenizer consumes the explicitly declared representation rather than inferring it from floating-point values. Feature generation uses a frozen encoder in evaluation mode, deterministic row-keyed tokenization, and a fixed forward batch size of eight; final incomplete batches are padded and padding outputs discarded. One all-zero SciPlex-middle test control is handled through the encoder's native zero-token path. Caches retain source row indices, and binding checks match dataset, split, source, manifest, and complete row coverage before training or inference. Tahoe uses its original frozen control-feature caches.

Feature extraction uses controls only. This establishes the downstream input boundary, but does not establish that external SE pretraining excluded these studies; pretraining provenance is discussed in Appendix G. GeneScope + SE is therefore identified explicitly wherever compared with GeneScope.

C Evaluation Metrics and Statistical Reporting

For condition c, let μ_c and $\widehat{\mu}_c$ be observed and predicted treated means, and $\mu^0_{\ell(c)}$ the held-out control mean. Define $\delta_c = \mu_c - \mu^0_{\ell(c)}$ and $\widehat{\delta}_c = \widehat{\mu}_c - \mu^0_{\ell(c)}$. Let T_c and $\widehat{T}_c$ contain their 100 largest absolute entries. All current GeneScope, ensemble, and intervention results use the stored *raw* output, including GeneScope + SE. The separately recorded cellwise nonnegative projection is not mixed into these tables.

Response correlations.
Δr and $\Delta\rho$ are Pearson and Spearman correlations of the complete response vectors.

Top-100 overlap and direction.
Overlap is $|T_c \cap \widehat{T}_c|/100$. Direction accuracy is the fraction of genes in T_c with matching signs.

High-response gene retrieval.
Genes at or above the 95th percentile of $|\delta_c|$ are positives, ranked by $|\widehat{\delta}_c|$. We report AUROC and average precision (AUPRC). The positive fraction is approximately 5%, with possible threshold ties. These are magnitude labels, not significance-tested DE calls.

Top-response rank correlation.
ρ_{100} is Spearman correlation of signed responses restricted to T_c. The stored field `lfc_spearman` uses differences of mean log-normalized expression, not fitted count-scale log fold changes.

Root mean squared error.
RMSE is $\sqrt{G^{-1}\sum_g(\widehat{\mu}_{cg} - \mu_{cg})^2}$. Unlike the other metrics, smaller values are better.

Metrics are computed per condition and macro-averaged, without weighting by cell count. Every current GeneScope family and ensemble has finite values on its full test split. Means and sample standard deviations summarize three training runs, with variance denominator two. These describe run variation on a fixed split, not confidence intervals over compounds or cell lines. Main tables show means alone. A fixed prediction ensemble is scored once and has no estimated ensemble standard deviation.

The control-mean baseline uses a training-derived mean, whereas evaluation subtracts held-out context controls. Its response is not identically zero, explaining nonzero correlation and potentially strong magnitude retrieval. It is not a zero-response oracle.

D COMPLETE COMPARISONS AND VARIABILITY

Reproductions. Statistical baselines are training-derived control mean, additive response, and ridge. Neural comparisons use retained reproductions of scGen, CPA, chemCPA, scVIDR, and STATE-ST, not published-paper numbers. STATE-ST is trained from scratch with expression input and one-hot treatment identities, without the pretrained STATE encoder. Non-STATE neural baselines have one retained run and STATE-ST has two per dataset; their compute budgets are not matched to GeneScope. Missing CPA results on cellxgene remain absent. Methods without the requested completed dataset coverage are omitted from the comparative inventory.

Support and reference matching. Shared-condition tables use exact identities, including SciPlex dose, with 529/536/248 conditions. The intersections were frozen before omitting incomplete methods from the displayed inventory; removing a row does not enlarge the set. Every displayed method uses these same identities. Current GeneScope predictions share fixed 256-cell input samples and full native test reference means. Baselines retain earlier normalization, reference sampling, and output processing. Support matching addresses condition coverage, not those remaining protocol differences. A controlled SOTA claim would require re-exporting all baselines under the same reference contract.

The shared-support tables give all eight metrics and available variability; native-support tables retain each method's own condition count. Blue shading is normalized within each dataset and metric, with darker values indicating better scores. Bold/underline mark the first and second distinct rounded means; displayed ties share rank. Scores use three decimals. Positive standard deviations below 0.0005 are printed as < 0.001, not zero.

Table 7: Tahoe: all eight metrics on the fixed shared-condition set. Means ± sample standard deviations; ensemble deviations are undefined. Darker blue denotes a better value within each column; RMSE is lower-is-better. Bold/underline mark first/second distinct displayed means. Tiny positive deviations are printed as < 0.001, not zero.

Method	Δr	$\Delta\rho$	Top-100	Dir.	AUPRC	ρ_{100}	RMSE ↓	AUROC
Control mean	-0.002	0.113	0.467	0.390	0.442	0.194	0.434	**0.942**
Additive response	0.084	0.175	0.456	0.441	0.427	0.307	0.412	<u>0.930</u>
Ridge	0.124	0.188	0.446	0.448	0.419	0.359	0.395	0.924
CPA	0.168	0.213	0.396	0.593	0.337	0.274	0.352	0.874
chemCPA	0.163	0.215	0.398	0.591	0.343	0.267	0.366	0.878
scGen	0.401	**0.372**	0.401	0.760	0.365	0.477	0.206	0.873
scVIDR	0.228	0.242	0.392	0.622	0.349	0.328	0.342	0.887
STATE-ST	0.431 ± 0.023	<u>0.357</u> ± 0.010	0.471 ± 0.002	0.747 ± 0.026	0.454 ± 0.002	0.521 ± 0.009	0.246 ± 0.005	0.912 ± 0.001
GeneScope	<u>0.522</u> ± 0.008	0.346 ± 0.016	0.485 ± 0.003	<u>0.807</u> ± 0.007	0.489 ± 0.005	<u>0.595</u> ± 0.006	0.171 ± 0.005	0.927 ± 0.002
GeneScope + ensemble	**0.528**	0.355	<u>0.489</u>	**0.812**	<u>0.494</u>	**0.601**	<u>0.168</u>	0.928
GeneScope + SE	0.505 ± 0.007	0.339 ± 0.029	0.482 ± 0.005	0.791 ± 0.009	0.485 ± 0.006	0.580 ± 0.003	0.170 ± 0.001	0.925 ± 0.004
GeneScope + SE + ensemble	0.518	0.354	**0.490**	0.799	**0.496**	0.591	**0.166**	0.927

Table 8: SciPlex-middle: all eight metrics on the fixed shared-condition set. Means ± sample standard deviations; ensemble deviations are undefined. Darker blue denotes a better value within each column; RMSE is lower-is-better. Bold/underline mark first/second distinct displayed means. Tiny positive deviations are printed as < 0.001, not zero.

Method	Δr	$\Delta\rho$	Top-100	Dir.	AUPRC	ρ_{100}	RMSE ↓	AUROC
Control mean	0.066	0.025	0.328	0.531	0.283	0.151	0.139	0.886
Additive response	0.075	0.044	0.328	**0.539**	0.287	0.158	0.139	0.886
Ridge	0.074	0.038	0.329	<u>0.538</u>	0.288	0.158	0.138	0.887
CPA	**0.086**	0.050	**0.420**	0.511	**0.394**	**0.182**	0.124	0.909
chemCPA	0.073	0.030	0.356	0.529	0.315	0.159	0.131	0.893
scGen	0.082	0.040	0.410	0.529	0.376	0.155	0.116	0.907
scVIDR	<u>0.085</u>	0.039	0.413	0.527	0.379	0.163	0.119	0.908
STATE-ST	0.078 ± 0.001	0.058 ± 0.004	0.360 ± 0.011	0.528 ± 0.004	0.326 ± 0.002	0.162 ± 0.001	0.132 ± 0.001	$0.902 \pm < 0.001$
GeneScope	0.084 ± 0.002	0.064 ± 0.002	$0.414 \pm < 0.001$	0.533 ± 0.005	0.368 ± 0.004	0.164 ± 0.003	<u>0.065</u> ± 0.002	0.902 ± 0.004
GeneScope + ensemble	**0.086**	0.064	**0.420**	0.534	0.373	0.168	**0.064**	0.903
GeneScope + SE	0.082 ± 0.002	<u>0.085</u> ± 0.006	0.413 ± 0.006	0.514 ± 0.005	0.379 ± 0.007	0.165 ± 0.001	0.082 ± 0.001	<u>0.917</u> ± 0.005
GeneScope + SE + ensemble	0.084	**0.086**	**0.420**	0.515	<u>0.385</u>	<u>0.169</u>	0.081	**0.919**

Table 9: cellxgene–SciPlex3: all eight metrics on the fixed shared-condition set. Means ± sample standard deviations; ensemble deviations are undefined. Darker blue denotes a better value within each column; RMSE is lower-is-better. Bold/underline mark first/second distinct displayed means. Tiny positive deviations are printed as < 0.001, not zero.

Method	Δr	$\Delta \rho$	Top-100	Dir.	AUPRC	ρ_{100}	RMSE ↓	AUROC
Control mean	0.028	0.028	0.055	0.557	0.053	0.101	0.503	0.483
Additive response	0.045	0.044	0.059	0.574	0.057	0.130	0.509	0.491
Ridge	0.042	0.041	0.058	0.573	0.055	0.127	0.506	0.489
chemCPA	0.016	0.012	0.019	0.548	0.040	0.056	1.001	0.391
scGen	0.028	0.028	0.017	0.548	0.039	0.095	0.994	0.353
scVIDR	0.028	0.032	0.017	0.545	0.038	0.083	1.071	0.340
STATE-ST	0.041 ± 0.001	0.040 ± 0.001	0.081 ± 0.005	0.555 ± 0.001	$\underline{0.073} \pm 0.001$	$0.106 \pm < 0.001$	0.469 ± 0.010	$\underline{0.575} \pm 0.005$
GeneScope	0.067 ± 0.001	0.052 ± 0.001	$\mathbf{0.145} \pm 0.003$	0.571 ± 0.004	$\mathbf{0.107} \pm 0.002$	0.138 ± 0.001	$\underline{0.126} \pm 0.001$	$\mathbf{0.646} \pm 0.007$
GeneScope + ensemble	0.068	0.053	**0.145**	0.572	**0.107**	0.140	**0.125**	**0.646**
GeneScope + SE	$\underline{0.072} \pm 0.002$	$\underline{0.061} \pm 0.003$	$\underline{0.090} \pm 0.014$	$\underline{0.601} \pm 0.002$	0.067 ± 0.006	$\underline{0.158} \pm 0.004$	0.162 ± 0.009	0.490 ± 0.020
GeneScope + SE + ensemble	**0.073**	**0.062**	0.088	**0.603**	0.066	**0.161**	0.161	0.485

Table 10: Tahoe: full native-support scores. Parentheses give native conditions/total; all GeneScope variants cover the complete split. Means ± sample standard deviations where available. Supports differ, so this is not a matched ranking; use Table 1 for fixed-support comparisons.

Method	Δr	$\Delta\rho$	Top-100	Dir.	AUPRC	ρ_{100}	RMSE ↓	AUROC
Control mean (646/646)	0.019	0.122	0.467	0.408	0.442	0.208	0.422	**0.941**
Additive response (646/646)	0.089	0.172	0.457	0.450	0.430	0.301	0.405	0.931
Ridge (646/646)	0.133	0.193	0.447	0.461	0.421	0.363	0.386	0.924
CPA (529/646)	0.168	0.213	0.396	0.593	0.337	0.274	0.352	0.874
chemCPA (529/646)	0.163	0.215	0.398	0.591	0.343	0.267	0.366	0.878
scGen (529/646)	0.401	**0.372**	0.401	0.760	0.365	0.477	0.206	0.873
scVIDR (529/646)	0.228	0.242	0.392	0.622	0.349	0.328	0.342	0.887
STATE-ST (529/646)	0.431 ± 0.023	0.357 ± 0.010	0.471 ± 0.002	0.747 ± 0.026	0.454 ± 0.002	0.521 ± 0.009	0.246 ± 0.005	0.912 ± 0.001
GeneScope	0.497 ± 0.007	0.339 ± 0.015	0.471 ± 0.003	0.795 ± 0.007	0.471 ± 0.004	0.570 ± 0.005	0.173 ± 0.005	0.922 ± 0.002
GeneScope + ensemble	**0.504**	0.347	0.475	**0.800**	0.476	**0.577**	0.170	0.923
GeneScope + SE	0.483 ± 0.006	0.332 ± 0.029	0.469 ± 0.005	0.780 ± 0.008	0.467 ± 0.006	0.559 ± 0.006	$0.172 \pm < 0.001$	0.921 ± 0.004
GeneScope + SE + ensemble	0.496	0.347	**0.476**	0.789	**0.477**	0.570	**0.168**	0.923

Table 11: SciPlex-middle: full native-support scores. Parentheses give native conditions/total; all GeneScope variants cover the complete split. Means ± sample standard deviations where available. Supports differ, so this is not a matched ranking; use Table 1 for fixed-support comparisons.

Method	Δr	$\Delta\rho$	Top-100	Dir.	AUPRC	ρ_{100}	RMSE ↓	AUROC
Control mean (1036/1036)	0.127	0.090	0.340	0.573	0.294	0.208	0.108	0.879
Additive response (1036/1036)	0.125	0.093	0.328	0.575	0.285	0.210	0.108	0.876
Ridge (1036/1036)	**0.135**	**0.097**	0.336	**0.585**	0.292	**0.231**	0.108	0.880
CPA (808/1036)	0.109	0.064	0.403	0.534	0.373	0.209	0.108	0.892
chemCPA (808/1036)	0.070	0.039	0.366	0.530	0.319	0.156	0.127	0.893
scGen (808/1036)	0.101	0.070	0.406	0.552	0.372	0.178	0.102	0.904
scVIDR (536/1036)	0.085	0.039	**0.413**	0.527	**0.379**	0.163	0.119	**0.908**
STATE-ST (808/1036)	0.093 ± 0.003	0.074 ± 0.003	0.365 ± 0.007	0.536 ± 0.003	0.329 ± 0.001	0.176 ± 0.005	0.117 ± 0.001	0.899 ± 0.001
GeneScope	0.103 ± 0.001	0.070 ± 0.002	0.401 ± 0.001	0.548 ± 0.004	0.359 ± 0.004	0.177 ± 0.001	0.055 ± 0.002	0.893 ± 0.003
GeneScope + ensemble	0.105	0.071	0.406	0.550	0.363	0.181	**0.054**	0.893
GeneScope + SE	0.105 ± 0.002	0.092 ± 0.005	0.404 ± 0.005	0.534 ± 0.005	0.369 ± 0.006	0.182 ± 0.002	0.067 ± 0.001	0.904 ± 0.004
GeneScope + SE + ensemble	0.108	0.094	0.411	0.536	0.374	0.186	0.066	0.906

Table 12: cellxgene–SciPlex3: full native-support scores. Parentheses give native conditions/total; all GeneScope variants cover the complete split. Means ± sample standard deviations where available. Supports differ, so this is not a matched ranking; use Table 1 for fixed-support comparisons.

Method	Δr	$\Delta\rho$	Top-100	Dir.	AUPRC	ρ_{100}	RMSE ↓	AUROC
Control mean (1040/1040)	0.143	0.144	0.094	0.629	0.078	0.205	0.395	0.545
Additive response (1040/1040)	0.149	0.146	0.097	0.640	0.082	0.227	0.399	0.550
Ridge (1040/1040)	**0.152**	**0.151**	0.106	**0.642**	0.087	**0.229**	0.394	0.556
chemCPA (384/1040)	0.041	0.041	0.042	0.557	0.056	0.096	0.956	0.449
scGen (812/1040)	0.063	0.064	0.023	0.549	0.040	0.136	0.981	0.366
scVIDR (540/1040)	0.055	0.061	0.018	0.523	0.038	0.119	1.078	0.333
STATE-ST (812/1040)	0.112 ± 0.001	0.110 ± 0.002	0.115 ± 0.003	0.602 ± 0.003	$0.096 \pm < 0.001$	0.176 ± 0.001	0.411 ± 0.007	0.612 ± 0.001
GeneScope	0.079 ± 0.001	$0.065 \pm < 0.001$	$\mathbf{0.153} \pm 0.002$	0.573 ± 0.001	$\mathbf{0.115} \pm 0.002$	0.140 ± 0.002	$\mathbf{0.134} \pm 0.001$	$\mathbf{0.647} \pm 0.007$
GeneScope + ensemble	0.080	0.065	**0.153**	0.574	**0.115**	0.141	**0.134**	**0.647**
GeneScope + SE	0.097 ± 0.001	0.085 ± 0.001	0.112 ± 0.014	0.595 ± 0.002	0.083 ± 0.007	0.171 ± 0.005	0.157 ± 0.007	0.527 ± 0.020
GeneScope + SE + ensemble	0.099	0.086	0.110	0.597	0.082	0.174	0.155	0.523

E DISJOINT GENERALIZATION STRATA

Four disjoint strata distinguish held-out pairs, compound-only, cell-only, and double-unseen transfer. Their sizes are 280/90/249/27 for Tahoe, 136/152/672/76 for SciPlex-middle, and 136/152/676/76 for cellxgene. Condition-count-weighted stratum means reproduce every full-test GeneScope metric. Strata share the same checkpoint and prediction export; no subgroup-specific selection is performed.

The main baseline transfer comparison reports a value only with complete native coverage of the stated stratum. Missing values are not zeros or extrapolated full-split scores. Reproduced one-hot transition adapters lack some unseen-compound outputs. The retained chemCPA adapter filters unseen compounds and uses compound-specific dose parameters; this is not a general limitation of molecular conditioning. Classical adapters can preserve coverage through fallbacks: additive response assigns unknown compounds zero drug effect, and unknown contexts may use global training controls. Coverage alone therefore does not establish learned chemical transfer.

Table 13: Tahoe: all current-model transfer metrics on disjoint strata. Three fitted models per family; fixed within-family ensembles have no estimated ensemble deviation. Means ± sample standard deviations; lower RMSE is better.

Method	Δr	$\Delta\rho$	Top-100	Dir.	AUPRC	ρ_{100}	RMSE ↓	AUROC
Tahoe: Held-out pair (280)								
GeneScope	0.666 ± 0.017	0.428 ± 0.028	0.543 ± 0.008	$\underline{0.892} \pm 0.012$	0.568 ± 0.011	0.717 ± 0.010	$\underline{0.133} \pm 0.001$	$\underline{0.941} \pm 0.003$
GeneScope + ensemble	**0.675**	**0.439**	**0.551**	**0.898**	**0.576**	**0.724**	**0.130**	**0.943**
GeneScope + SE	0.655 ± 0.005	0.420 ± 0.032	0.535 ± 0.001	0.881 ± 0.003	0.559 ± 0.002	0.704 ± 0.003	0.134 ± 0.001	0.937 ± 0.002
GeneScope + SE + ensemble	$\underline{0.672}$	$\underline{0.436}$	$\underline{0.545}$	0.891	$\underline{0.573}$	$\underline{0.718}$	**0.130**	0.940
Tahoe: Unseen compound (90)								
GeneScope	0.437 ± 0.002	0.341 ± 0.013	$\underline{0.413} \pm 0.004$	0.773 ± 0.006	$\underline{0.397} \pm 0.004$	0.513 ± 0.003	0.168 ± 0.002	$\underline{0.902} \pm 0.002$
GeneScope + ensemble	$\underline{0.445}$	**0.349**	$\underline{0.413}$	**0.778**	**0.401**	$\underline{0.519}$	$\underline{0.166}$	**0.903**
GeneScope + SE	0.431 ± 0.003	0.330 ± 0.032	0.412 ± 0.008	0.764 ± 0.004	0.394 ± 0.009	0.508 ± 0.020	$0.168 \pm < 0.001$	0.901 ± 0.004
GeneScope + SE + ensemble	**0.446**	$\underline{0.347}$	**0.415**	$\underline{0.775}$	**0.401**	**0.521**	**0.165**	**0.903**
Tahoe: Unseen cell line (249)								
GeneScope	$\underline{0.359} \pm 0.002$	0.255 ± 0.003	0.419 ± 0.007	$\underline{0.711} \pm 0.003$	0.400 ± 0.005	$\underline{0.457} \pm 0.003$	0.213 ± 0.012	$\underline{0.911} \pm 0.001$
GeneScope + ensemble	**0.363**	$\underline{0.260}$	0.420	**0.715**	$\underline{0.402}$	**0.462**	$\underline{0.211}$	$\underline{0.911}$
GeneScope + SE	0.336 ± 0.009	0.249 ± 0.026	$\underline{0.423} \pm 0.008$	0.689 ± 0.015	0.401 ± 0.009	0.440 ± 0.004	$\underline{0.211} \pm < 0.001$	$\underline{0.911} \pm 0.006$
GeneScope + SE + ensemble	0.344	**0.262**	**0.428**	0.696	**0.408**	0.449	**0.207**	**0.912**
Tahoe: Double unseen (27)								
GeneScope	0.215 ± 0.002	0.183 ± 0.002	0.392 ± 0.004	0.628 ± 0.005	0.363 ± 0.006	0.290 ± 0.004	0.229 ± 0.012	$0.902 \pm < 0.001$
GeneScope + ensemble	0.218	0.188	0.395	$\underline{0.633}$	0.363	0.294	0.227	0.902
GeneScope + SE	$\underline{0.225} \pm 0.007$	$\underline{0.193} \pm 0.025$	$\underline{0.405} \pm 0.019$	$\underline{0.633} \pm 0.017$	$\underline{0.375} \pm 0.013$	$\underline{0.313} \pm 0.022$	$\underline{0.225} \pm 0.001$	$\underline{0.904} \pm 0.005$
GeneScope + SE + ensemble	**0.232**	**0.201**	**0.408**	**0.637**	**0.380**	**0.320**	**0.222**	**0.906**

Table 14: SciPlex-middle: all current-model transfer metrics on disjoint strata. Three fitted models per family; fixed within-family ensembles have no estimated ensemble deviation. Means ± sample standard deviations; lower RMSE is better.

Method	Δr	$\Delta\rho$	Top-100	Dir.	AUPRC	ρ_{100}	RMSE ↓	AUROC
SciPlex-middle: Held-out pair (136)								
GeneScope	0.142 ± 0.003	0.073 ± 0.002	0.370 ± 0.007	0.585 ± 0.004	0.333 ± 0.003	0.184 ± 0.008	$0.032 \pm < 0.001$	0.872 ± 0.001
GeneScope + ensemble	0.145	0.077	0.372	**0.588**	0.334	0.189	0.031	0.872
GeneScope + SE	0.155 ± 0.002	0.100 ± 0.004	0.386 ± 0.004	0.583 ± 0.002	0.346 ± 0.002	0.196 ± 0.002	$0.031 \pm < 0.001$	$\mathbf{0.879} \pm 0.002$
GeneScope + SE + ensemble	**0.159**	**0.104**	**0.390**	0.586	**0.348**	**0.201**	**0.030**	0.878
SciPlex-middle: Unseen compound (152)								
GeneScope	0.136 ± 0.002	0.081 ± 0.001	0.351 ± 0.004	0.579 ± 0.002	0.323 ± 0.003	0.204 ± 0.008	$0.031 \pm < 0.001$	$\mathbf{0.863} \pm 0.001$
GeneScope + ensemble	0.140	0.083	0.353	**0.581**	0.325	0.209	0.031	0.862
GeneScope + SE	0.149 ± 0.003	0.105 ± 0.003	0.364 ± 0.002	0.576 ± 0.004	0.330 ± 0.003	0.218 ± 0.003	$\mathbf{0.030} \pm < 0.001$	$\mathbf{0.863} \pm 0.001$
GeneScope + SE + ensemble	**0.153**	**0.109**	**0.367**	0.580	**0.332**	**0.223**	**0.030**	0.862
SciPlex-middle: Unseen cell line (672)								
GeneScope	0.086 ± 0.002	0.065 ± 0.002	$0.416 \pm < 0.001$	$\mathbf{0.534} \pm 0.005$	0.369 ± 0.004	0.170 ± 0.003	0.065 ± 0.002	0.903 ± 0.004
GeneScope + ensemble	**0.088**	0.065	**0.422**	**0.534**	0.375	0.173	**0.064**	0.904
GeneScope + SE	0.085 ± 0.002	0.086 ± 0.006	0.414 ± 0.007	0.515 ± 0.005	0.380 ± 0.008	0.171 ± 0.002	0.082 ± 0.001	0.918 ± 0.005
GeneScope + SE + ensemble	0.087	**0.088**	**0.422**	0.516	**0.386**	**0.175**	0.081	**0.919**
SciPlex-middle: Double unseen (76)								
GeneScope	0.109 ± 0.004	0.078 ± 0.005	0.432 ± 0.002	$\mathbf{0.550} \pm 0.005$	0.384 ± 0.006	0.179 ± 0.005	0.064 ± 0.002	0.905 ± 0.004
GeneScope + ensemble	0.110	0.079	0.437	**0.550**	0.389	0.183	**0.063**	0.905
GeneScope + SE	0.109 ± 0.002	0.102 ± 0.006	0.430 ± 0.008	0.529 ± 0.006	0.393 ± 0.008	0.186 ± 0.002	0.080 ± 0.001	0.917 ± 0.005
GeneScope + SE + ensemble	**0.111**	**0.104**	**0.438**	0.530	**0.399**	**0.191**	0.079	**0.919**

Table 15: cellxgene–SciPlex3: all current-model transfer metrics on disjoint strata. Three fitted models per family; fixed within-family ensembles have no estimated ensemble deviation. Means ± sample standard deviations; lower RMSE is better.

Method	Δr	$\Delta\rho$	Top-100	Dir.	AUPRC	ρ_{100}	RMSE ↓	AUROC
cellxgene–SciPlex3: Held-out pair (136)								
GeneScope	0.099 ± 0.002	0.081 ± 0.001	0.152 ± 0.002	0.595 ± 0.003	0.116 ± 0.001	0.145 ± 0.005	0.131± < 0.001	0.641 ± 0.002
GeneScope + ensemble	0.100	0.082	0.151	0.596	0.116	0.147	0.130	0.640
GeneScope + SE	0.129 ± 0.002	0.113 ± 0.001	**0.153** ± 0.009	0.622 ± 0.010	0.117 ± 0.005	0.178 ± 0.005	0.127 ± 0.001	**0.644** ± 0.009
GeneScope + SE + ensemble	**0.132**	**0.115**	0.152	**0.626**	**0.118**	**0.181**	**0.126**	**0.644**
cellxgene–SciPlex3: Unseen compound (152)								
GeneScope	0.095 ± 0.001	0.082 ± 0.001	0.133 ± 0.002	0.577 ± 0.002	0.103± < 0.001	0.172 ± 0.003	0.135± < 0.001	0.610± < 0.001
GeneScope + ensemble	0.096	0.083	0.133	0.577	0.103	0.174	0.134	0.609
GeneScope + SE	0.117 ± 0.003	0.104 ± 0.003	**0.136** ± 0.004	0.595 ± 0.004	**0.106** ± 0.002	0.195 ± 0.012	0.132 ± 0.001	**0.614** ± 0.006
GeneScope + SE + ensemble	**0.119**	**0.106**	0.135	**0.597**	0.105	**0.199**	**0.131**	**0.614**
cellxgene–SciPlex3: Unseen cell line (676)								
GeneScope	0.073± < 0.001	0.059± < 0.001	**0.158** ± 0.003	0.569 ± 0.002	**0.117** ± 0.003	0.134 ± 0.001	0.135 ± 0.001	**0.654** ± 0.010
GeneScope + ensemble	0.074	0.059	**0.158**	0.570	**0.117**	0.135	**0.134**	**0.654**
GeneScope + SE	0.088 ± 0.001	0.076 ± 0.002	0.099 ± 0.018	0.591 ± 0.002	0.072 ± 0.009	0.167 ± 0.004	0.167 ± 0.010	0.489 ± 0.026
GeneScope + SE + ensemble	**0.089**	**0.077**	0.097	**0.593**	0.071	**0.170**	0.166	0.483
cellxgene–SciPlex3: Double unseen (76)								
GeneScope	0.064 ± 0.001	0.053 ± 0.001	0.157 ± 0.003	0.565 ± 0.003	**0.116** ± 0.003	0.116 ± 0.002	0.136 ± 0.001	**0.662** ± 0.009
GeneScope + ensemble	0.065	0.053	**0.158**	0.566	**0.116**	0.115	**0.135**	**0.662**
GeneScope + SE	0.083 ± 0.001	0.073 ± 0.002	0.096 ± 0.018	0.575 ± 0.002	0.069 ± 0.008	**0.149** ± 0.002	0.167 ± 0.009	0.481 ± 0.028
GeneScope + SE + ensemble	**0.084**	**0.074**	0.095	**0.576**	0.068	**0.149**	0.165	0.474

F Component Interventions and Resolution Budget

Matched design. Eight interventions cover all three formal test splits, with three paired seeds per dataset. Each uses GeneScope without SE, 5,000 updates, the same split, and fixed final control samples. The 72 intervention fits comprise 67 new models and five exact compatible Micro-only fits retained from the same protocol. The primary comparison's full models are the anchors. Differences are computed within seed before averaging; their deviations are computed from these paired differences. Matching seeds does not imply identical initialization when the graph changes. Functional removals can retain allocated parameters and do not establish parameter or FLOP savings.

Response-path interventions. No affine response sets $a = 0$, retaining the nonlinear control-mean correction and learned gate. No selected readout zeros the complete selected-gene decoder output while retaining Micro attention; it is not complete Micro removal. No composition gate fixes $\eta = 1$ but retains γ. No basal bypass removes $h_{\mathrm{basal}}(x_i^0)$, not the final control-expression skip. No population encoder replaces its output by input cell tokens. One sketch retains the first view.

Functional stream reductions. Macro-only bypasses Micro attention and disables the selected-gene decoder, including bias and population-dependent terms; Macro features, reconstruction, and pooling remain. Micro-only bypasses Macro attention and reconstruction, zeros Macro gene features, and pools selected Micro tokens. Cross-attention memory contains only the chemical token. Selection replaces reconstruction error with mean absolute control deviation, retaining dispersion and the disjoint quotas. Shared affine, basal, population, and remaining decoder paths are retained. This tests a coherent reduced predictor, not an isolated routing intervention.

Table 16: Tahoe: absolute formal-test scores over all 646 conditions, means ± sample standard deviations. Lower RMSE is better. Seed pairing does not assert identical initialization for all functional interventions.

Intervention	Δr	$\Delta\rho$	Top-100	Dir.	AUPRC	ρ_{100}	RMSE ↓	AUROC
Full GeneScope	0.497 ± 0.007	0.339 ± 0.015	$\underline{0.471} \pm 0.003$	0.795 ± 0.007	$\mathbf{0.471} \pm 0.004$	$\underline{0.570} \pm 0.005$	$\underline{0.173} \pm 0.005$	$\mathbf{0.922} \pm 0.002$
No affine response	0.389 ± 0.005	0.179 ± 0.002	0.453 ± 0.001	0.691 ± 0.006	0.441 ± 0.002	0.474 ± 0.009	0.189 ± 0.005	0.911 ± 0.001
No basal bypass	0.489 ± 0.003	0.300 ± 0.018	$\underline{0.471} \pm 0.004$	$\underline{0.796} \pm 0.003$	$\underline{0.469} \pm 0.005$	$\underline{0.570} \pm 0.003$	$\mathbf{0.168} \pm 0.007$	$\mathbf{0.922} \pm 0.002$
No selected readout	0.466 ± 0.021	0.305 ± 0.020	0.451 ± 0.011	0.730 ± 0.042	0.445 ± 0.014	0.562 ± 0.003	0.179 ± 0.007	0.906 ± 0.009
No composition gate	0.458 ± 0.006	0.305 ± 0.004	0.452 ± 0.005	0.721 ± 0.010	0.444 ± 0.006	0.562 ± 0.002	0.181 ± 0.003	0.904 ± 0.003
One sketch	0.485 ± 0.015	0.338 ± 0.025	0.465 ± 0.008	0.776 ± 0.036	0.462 ± 0.011	0.567 ± 0.003	0.180 ± 0.007	0.919 ± 0.005
No population encoder	$\mathbf{0.503} \pm 0.002$	$\underline{0.351} \pm < 0.001$	0.463 ± 0.003	$\mathbf{0.801} \pm 0.002$	0.465 ± 0.002	$\mathbf{0.571} \pm 0.001$	0.183 ± 0.004	0.920 ± 0.001
Macro only	0.470 ± 0.026	0.310 ± 0.025	0.452 ± 0.012	0.734 ± 0.050	0.446 ± 0.016	0.565 ± 0.004	0.177 ± 0.009	0.907 ± 0.010
Micro only	$\underline{0.502} \pm 0.002$	$\mathbf{0.388} \pm 0.003$	$\mathbf{0.473} \pm 0.001$	0.794 ± 0.005	$\mathbf{0.471} \pm 0.001$	0.565 ± 0.002	$\mathbf{0.168} \pm 0.001$	$\underline{0.921} \pm 0.002$

Table 17: SciPlex-middle: absolute formal-test scores over all 1036 conditions, means ± sample standard deviations. Lower RMSE is better. Seed pairing does not assert identical initialization for all functional interventions.

Intervention	Δr	$\Delta\rho$	Top-100	Dir.	AUPRC	ρ_{100}	RMSE ↓	AUROC
Full GeneScope	0.103 ± 0.001	0.070 ± 0.002	0.401 ± 0.001	0.548 ± 0.004	0.359 ± 0.004	0.177 ± 0.001	0.055 ± 0.002	0.893 ± 0.003
No affine response	$\mathbf{0.111} \pm 0.003$	0.055 ± 0.005	$\mathbf{0.425} \pm 0.001$	$\underline{0.555} \pm 0.006$	$\mathbf{0.385} \pm 0.002$	$\mathbf{0.190} \pm 0.004$	0.085 ± 0.005	0.897 ± 0.002
No basal bypass	0.101 ± 0.004	0.052 ± 0.006	0.361 ± 0.009	$\mathbf{0.565} \pm 0.001$	0.320 ± 0.011	0.172 ± 0.004	0.065 ± 0.001	0.876 ± 0.003
No selected readout	$0.093 \pm < 0.001$	$\mathbf{0.091} \pm 0.001$	0.390 ± 0.006	0.520 ± 0.004	0.355 ± 0.007	0.158 ± 0.005	$\underline{0.050} \pm 0.003$	$\underline{0.901} \pm 0.004$
No composition gate	$\underline{0.105} \pm 0.002$	0.070 ± 0.007	0.407 ± 0.007	0.545 ± 0.004	0.367 ± 0.006	$\underline{0.184} \pm 0.003$	0.057 ± 0.001	0.896 ± 0.003
One sketch	0.102 ± 0.001	0.078 ± 0.003	0.407 ± 0.002	0.537 ± 0.002	0.368 ± 0.002	0.175 ± 0.001	0.060 ± 0.002	0.899 ± 0.001
No population encoder	0.102 ± 0.001	0.079 ± 0.002	$\underline{0.411} \pm 0.002$	0.537 ± 0.001	$\underline{0.374} \pm 0.002$	0.182 ± 0.003	0.059 ± 0.002	$\mathbf{0.902} \pm 0.001$
Macro only	0.094 ± 0.002	$\mathbf{0.091} \pm 0.001$	0.390 ± 0.005	0.521 ± 0.004	0.355 ± 0.006	0.158 ± 0.004	$\mathbf{0.049} \pm 0.002$	$\underline{0.901} \pm 0.003$
Micro only	0.098 ± 0.002	$\underline{0.088} \pm 0.008$	0.379 ± 0.020	0.537 ± 0.007	0.345 ± 0.022	0.165 ± 0.011	0.051 ± 0.003	0.895 ± 0.010

Table 18: cellxgene–SciPlex3: absolute formal-test scores over all 1040 conditions, means ± sample standard deviations. Lower RMSE is better. Seed pairing does not assert identical initialization for all functional interventions.

Intervention	Δr	$\Delta\rho$	Top-100	Dir.	AUPRC	ρ_{100}	RMSE ↓	AUROC
Full GeneScope	$\underline{0.079} \pm 0.001$	$\underline{0.065} \pm < 0.001$	0.153 ± 0.002	$\underline{0.573} \pm 0.001$	0.115 ± 0.002	0.140 ± 0.002	$\underline{0.134} \pm 0.001$	0.647 ± 0.007
No affine response	0.069 ± 0.001	0.056 ± 0.001	0.134 ± 0.001	0.571 ± 0.002	0.102 ± 0.001	0.120 ± 0.002	$0.137 \pm < 0.001$	0.616 ± 0.002
No basal bypass	$\underline{0.079} \pm 0.002$	0.063 ± 0.002	$\underline{0.156} \pm 0.007$	0.572 ± 0.008	$\underline{0.117} \pm 0.006$	0.143 ± 0.004	$0.135 \pm < 0.001$	$\underline{0.654} \pm 0.015$
No selected readout	$0.068 \pm < 0.001$	0.059 ± 0.001	0.137 ± 0.002	0.562 ± 0.002	0.104 ± 0.002	0.111 ± 0.001	$\mathbf{0.133} \pm < 0.001$	0.634 ± 0.005
No composition gate	$\mathbf{0.081} \pm 0.001$	$\mathbf{0.066} \pm < 0.001$	0.152 ± 0.004	$\mathbf{0.577} \pm 0.001$	0.113 ± 0.003	$\mathbf{0.146} \pm 0.002$	0.135 ± 0.001	0.640 ± 0.008
One sketch	0.077 ± 0.002	0.063 ± 0.002	0.152 ± 0.003	0.570 ± 0.004	0.114 ± 0.004	0.131 ± 0.005	0.135 ± 0.001	0.648 ± 0.012
No population encoder	$\underline{0.079} \pm 0.001$	0.064 ± 0.001	$\mathbf{0.168} \pm 0.001$	$0.563 \pm < 0.001$	$\mathbf{0.128} \pm 0.001$	$\underline{0.144} \pm 0.001$	$0.137 \pm < 0.001$	$\mathbf{0.678} \pm 0.001$
Macro only	0.069 ± 0.002	0.059 ± 0.001	0.141 ± 0.005	0.560 ± 0.001	0.107 ± 0.004	0.112 ± 0.003	$\mathbf{0.133} \pm < 0.001$	0.644 ± 0.008
Micro only	0.070 ± 0.004	0.060 ± 0.003	0.143 ± 0.002	0.562 ± 0.007	0.109 ± 0.003	0.116 ± 0.008	$\mathbf{0.133} \pm 0.001$	0.642 ± 0.016

Table 19: Tahoe: seed-paired intervention-minus-full differences over all 646 conditions, means ± sample standard deviations. Lower RMSE is better. Seed pairing does not assert identical initialization for all functional interventions.

Intervention	Δr	$\Delta \rho$	Top-100	Dir.	AUPRC	ρ_{100}	RMSE ↓	AUROC
No affine response	-0.108 ± 0.005	-0.160 ± 0.013	-0.018 ± 0.003	-0.104 ± 0.004	-0.030 ± 0.003	-0.096 ± 0.011	$+0.016 \pm 0.008$	-0.011 ± 0.001
No basal bypass	-0.008 ± 0.004	-0.039 ± 0.012	$\underline{+0.000} \pm 0.007$	$\underline{+0.001} \pm 0.004$	$\underline{-0.002} \pm 0.009$	$\underline{+0.000} \pm 0.008$	$\mathbf{-0.005} \pm 0.012$	$\mathbf{-0.001} \pm 0.004$
No selected readout	-0.031 ± 0.027	-0.033 ± 0.033	-0.020 ± 0.009	-0.065 ± 0.049	-0.026 ± 0.014	-0.008 ± 0.002	$+0.006 \pm 0.003$	-0.017 ± 0.010
No composition gate	-0.039 ± 0.009	-0.034 ± 0.015	-0.019 ± 0.002	-0.074 ± 0.014	-0.027 ± 0.003	-0.008 ± 0.004	$+0.008 \pm 0.005$	-0.018 ± 0.002
One sketch	-0.012 ± 0.019	-0.001 ± 0.035	-0.006 ± 0.005	-0.019 ± 0.041	-0.009 ± 0.009	-0.004 ± 0.003	$+0.008 \pm 0.006$	-0.003 ± 0.006
No population encoder	$\mathbf{+0.006} \pm 0.005$	$\underline{+0.012} \pm 0.014$	-0.008 ± 0.005	$\mathbf{+0.006} \pm 0.007$	-0.006 ± 0.006	$\mathbf{+0.001} \pm 0.004$	$+0.010 \pm 0.004$	$\underline{-0.002} \pm 0.002$
Macro only	-0.027 ± 0.033	-0.029 ± 0.039	-0.019 ± 0.011	-0.060 ± 0.057	-0.025 ± 0.016	-0.005 ± 0.003	$\underline{+0.004} \pm 0.004$	-0.016 ± 0.011
Micro only	$\underline{+0.005} \pm 0.009$	$\mathbf{+0.049} \pm 0.013$	$\mathbf{+0.002} \pm 0.003$	-0.001 ± 0.010	$\mathbf{+0.000} \pm 0.004$	-0.005 ± 0.006	$\mathbf{-0.005} \pm 0.006$	$\underline{-0.002} \pm 0.003$

Table 20: SciPlex-middle: seed-paired intervention-minus-full differences over all 1036 conditions, means ± sample standard deviations. Lower RMSE is better. Seed pairing does not assert identical initialization for all functional interventions.

Intervention	Δr	$\Delta\rho$	Top-100	Dir.	AUPRC	ρ_{100}	RMSE ↓	AUROC
No affine response	$+\mathbf{0.009} \pm 0.003$	-0.014 ± 0.005	$+\mathbf{0.024} \pm 0.002$	$\underline{+0.007} \pm 0.007$	$+\mathbf{0.026} \pm 0.004$	$+\mathbf{0.012} \pm 0.005$	$+0.030 \pm 0.007$	$+0.004 \pm 0.003$
No basal bypass	-0.002 ± 0.004	-0.018 ± 0.008	-0.040 ± 0.011	$+\mathbf{0.017} \pm 0.005$	-0.039 ± 0.013	-0.005 ± 0.005	$+0.009 \pm 0.003$	-0.017 ± 0.005
No selected readout	-0.010 ± 0.001	$+\mathbf{0.022} \pm 0.002$	-0.011 ± 0.007	-0.028 ± 0.004	-0.003 ± 0.008	-0.019 ± 0.006	$-\mathbf{0.006} \pm 0.004$	$\underline{+0.008} \pm 0.004$
No composition gate	$\underline{+0.003} \pm 0.001$	$+0.001 \pm 0.009$	$+0.005 \pm 0.008$	-0.003 ± 0.008	$+0.009 \pm 0.009$	$\underline{+0.007} \pm 0.003$	$+0.002 \pm 0.002$	$+0.003 \pm 0.006$
One sketch	$-0.001 \pm < 0.001$	$+0.008 \pm 0.004$	$+0.005 \pm 0.001$	-0.011 ± 0.007	$+0.009 \pm 0.002$	-0.002 ± 0.001	$+0.004 \pm < 0.001$	$+0.006 \pm 0.004$
No population encoder	$+0.000 \pm 0.001$	$+0.009 \pm 0.003$	$\underline{+0.010} \pm 0.003$	-0.011 ± 0.005	$\underline{+0.015} \pm 0.005$	$+0.005 \pm 0.003$	$+0.004 \pm 0.002$	$+\mathbf{0.009} \pm 0.003$
Macro only	-0.009 ± 0.002	$+\mathbf{0.022} \pm < 0.001$	-0.012 ± 0.006	-0.028 ± 0.004	-0.004 ± 0.006	-0.019 ± 0.004	$-\mathbf{0.006} \pm 0.003$	$\underline{+0.008} \pm 0.003$
Micro only	-0.005 ± 0.003	$\underline{+0.018} \pm 0.010$	-0.023 ± 0.021	-0.012 ± 0.011	-0.014 ± 0.025	-0.013 ± 0.011	$\underline{-0.004} \pm 0.005$	$+0.002 \pm 0.012$

Table 21: cellxgene–SciPlex3: seed-paired intervention-minus-full differences over all 1040 conditions, means ± sample standard deviations. Lower RMSE is better. Seed pairing does not assert identical initialization for all functional interventions.

Intervention	Δr	$\Delta \rho$	Top-100	Dir.	AUPRC	ρ_{100}	RMSE ↓	AUROC
No affine response	-0.010 ± 0.002	-0.009 ± 0.001	-0.019 ± 0.003	-0.002 ± 0.003	-0.012 ± 0.003	-0.020 ± 0.002	$+0.003 \pm < 0.001$	-0.030 ± 0.010
No basal bypass	$\underline{+0.000} \pm 0.002$	$\underline{-0.001} \pm 0.002$	$\underline{+0.002} \pm 0.009$	$\underline{-0.001} \pm 0.009$	$\underline{+0.002} \pm 0.008$	$+0.003 \pm 0.004$	$+0.001 \pm < 0.001$	$\underline{+0.008} \pm 0.022$
No selected readout	-0.011 ± 0.001	-0.006 ± 0.001	-0.016 ± 0.003	-0.011 ± 0.002	-0.011 ± 0.003	-0.029 ± 0.002	$\mathbf{-0.001} \pm 0.001$	-0.013 ± 0.011
No composition gate	$\mathbf{+0.002} \pm 0.001$	$\mathbf{+0.001} \pm < 0.001$	-0.001 ± 0.002	$\mathbf{+0.004} \pm 0.001$	-0.001 ± 0.002	$\mathbf{+0.006} \pm 0.003$	$+0.001 \pm < 0.001$	-0.006 ± 0.003
One sketch	-0.002 ± 0.003	$\underline{-0.001} \pm 0.002$	-0.001 ± 0.005	-0.004 ± 0.005	$+0.000 \pm 0.004$	-0.009 ± 0.007	$\underline{+0.000} \pm 0.001$	$+0.001 \pm 0.012$
No population encoder	$\underline{+0.000} \pm 0.001$	$\underline{-0.001} \pm 0.001$	$\mathbf{+0.015} \pm 0.003$	-0.010 ± 0.002	$\mathbf{+0.013} \pm 0.003$	$\underline{+0.004} \pm 0.002$	$+0.002 \pm 0.001$	$\mathbf{+0.031} \pm 0.008$
Macro only	-0.010 ± 0.002	-0.006 ± 0.002	-0.013 ± 0.004	$-0.013 \pm < 0.001$	-0.008 ± 0.003	-0.027 ± 0.005	$\mathbf{-0.001} \pm 0.001$	-0.003 ± 0.007
Micro only	-0.009 ± 0.003	-0.005 ± 0.002	-0.010 ± 0.001	-0.011 ± 0.006	-0.006 ± 0.001	-0.024 ± 0.009	$\mathbf{-0.001} \pm < 0.001$	-0.004 ± 0.008

Resolution sensitivity. The sweep evaluates $K \in \{100, 200, 300, 400, 500, 600, 700\}$ on the 362-condition Tahoe validation set, using three runs per point and 5,000 updates per run. The 2,000-gene output panel and three 300-bin Macro views remain fixed. There are 21 completed fits; no SciPlex or SE sweep is inferred. $K = 200$ is the common operating point, not the highest validation Pearson. $K = 500$ has the highest observed mean, but its paired improvement is small relative to run variation. The complete scores and paired changes below remain separate from formal-test component interventions.

Table 22: Tahoe validation K sweep: absolute scores. Means ± sample standard deviations over three runs; 362 validation conditions and 5,000 updates per model. These are not test scores.

K	Δr	$\Delta\rho$	Top-100	Dir.	AUPRC	ρ_{100}	RMSE ↓	AUROC
100	0.653 ± 0.011	0.502 ± 0.005	$\underline{0.527} \pm 0.005$	0.894 ± 0.008	$\underline{0.540} \pm 0.009$	$\underline{0.712} \pm 0.005$	0.148 ± 0.002	$\underline{0.935} \pm 0.005$
200	0.665 ± 0.012	0.511 ± 0.005	$\mathbf{0.528} \pm 0.007$	$\mathbf{0.902} \pm 0.009$	$\mathbf{0.545} \pm 0.008$	$\mathbf{0.715} \pm 0.006$	$\underline{0.145} \pm 0.002$	$\mathbf{0.937} \pm 0.002$
300	0.663 ± 0.003	0.509 ± 0.003	0.523 ± 0.002	0.900 ± 0.003	0.539 ± 0.003	0.709 ± 0.003	$\underline{0.145} \pm 0.001$	$\underline{0.935} \pm < 0.001$
400	0.665 ± 0.004	0.509 ± 0.008	0.522 ± 0.004	$\underline{0.901} \pm 0.001$	0.536 ± 0.006	0.709 ± 0.002	$\underline{0.145} \pm 0.001$	0.934 ± 0.003
500	$\mathbf{0.670} \pm 0.006$	0.517 ± 0.004	0.521 ± 0.004	0.899 ± 0.007	0.532 ± 0.008	0.711 ± 0.003	$\mathbf{0.144} \pm 0.001$	0.925 ± 0.006
600	$\underline{0.667} \pm 0.004$	$\mathbf{0.523} \pm 0.005$	0.519 ± 0.004	0.896 ± 0.004	0.530 ± 0.004	0.710 ± 0.003	$\underline{0.145} \pm 0.001$	0.923 ± 0.003
700	0.660 ± 0.016	$\underline{0.518} \pm 0.014$	0.517 ± 0.009	0.889 ± 0.013	0.525 ± 0.014	0.707 ± 0.007	0.147 ± 0.004	0.923 ± 0.003

Table 23: Tahoe validation K sweep: paired changes relative to K=200. Means ± sample standard deviations over three runs; 362 validation conditions and 5,000 updates per model. These are not test scores.

K	Δr	$\Delta\rho$	Top-100	Dir.	AUPRC	ρ_{100}	RMSE ↓	AUROC
100	-0.012 ± 0.013	-0.010 ± 0.008	$\underline{-0.001} \pm 0.003$	-0.008 ± 0.001	$\underline{-0.005} \pm 0.002$	$\underline{-0.003} \pm 0.004$	$+0.003 \pm 0.003$	-0.003 ± 0.003
200	$+0.000 \pm 0.000$	$+0.000 \pm 0.000$	$+\mathbf{0.000} \pm 0.000$	$+\mathbf{0.000} \pm 0.000$	$+\mathbf{0.000} \pm 0.000$	$+\mathbf{0.000} \pm 0.000$	$\underline{+0.000} \pm 0.000$	$+\mathbf{0.000} \pm 0.000$
300	-0.002 ± 0.010	-0.002 ± 0.008	-0.005 ± 0.006	-0.002 ± 0.007	-0.006 ± 0.007	-0.006 ± 0.003	$\underline{+0.000} \pm 0.002$	$\underline{-0.002} \pm 0.001$
400	$+0.000 \pm 0.008$	-0.003 ± 0.002	-0.006 ± 0.003	$\underline{-0.001} \pm 0.009$	-0.009 ± 0.002	-0.006 ± 0.003	$\underline{+0.000} \pm 0.002$	$\underline{-0.003} \pm 0.001$
500	$+\mathbf{0.005} \pm 0.007$	$+0.006 \pm 0.002$	-0.007 ± 0.004	-0.003 ± 0.002	-0.013 ± 0.003	-0.004 ± 0.005	$-\mathbf{0.001} \pm 0.001$	-0.012 ± 0.004
600	$\underline{+0.002} \pm 0.010$	$+\mathbf{0.012} \pm 0.008$	-0.009 ± 0.003	-0.006 ± 0.006	-0.015 ± 0.007	-0.005 ± 0.005	$\underline{+0.000} \pm 0.002$	-0.014 ± 0.004
700	-0.006 ± 0.004	$\underline{+0.007} \pm 0.010$	-0.011 ± 0.003	-0.013 ± 0.005	-0.020 ± 0.007	-0.008 ± 0.002	$+0.002 \pm 0.001$	-0.014 ± 0.005

G Frozen Features and Prediction Ensembles

GeneScope requires no pretrained cellular encoder. GeneScope + SE adds frozen STATE control features and the scale adapter of Section 3.5. Both families have three fitted models per dataset, mean-response weight 40, and refinement budget 200. Features use controls, not held-out treated profiles. External SE pretraining may nonetheless include related study data; control-only downstream extraction does not establish pretraining exclusion. The +SE comparison measures access to an external representation, not an overlap-free pretraining test.

Each within-family ensemble averages three aligned prediction tensors with equal weights before computing means and metrics, rather than averaging reported scores. Input controls and row order are shared. GeneScope + ensemble and GeneScope + SE + ensemble are reported separately, without a mixed-family ensemble. Each is a single fixed predictor; the component-model deviation is not its uncertainty. Only the current recipe, SE adapter, and within-family averaging are reported; unrelated earlier tuning inventories are excluded.

H Coverage-Adjusted Utility

For native support n among C conditions and native macro score m, define the retrospective utility

$$U = \frac{n}{C} m + \left(1 - \frac{n}{C}\right) b. \tag{17}$$

The reference b is zero for correlations, 0.05 for Top-100 overlap and average precision, and 0.5 for direction accuracy. The retrieval reference approximates the 5% positive fraction, without adjusting threshold ties. This is a transparent scoring convention, not a new prediction, unbiased estimate, or worst-case bound: zero correlation is not its minimum possible value. RMSE is not adjusted because it has no common chance-scale reference. Complete GeneScope scores are unchanged. Neither main comparison uses this adjustment.

Table 24: Tahoe: retrospective full-denominator coverage utility, Eq. 17. Fractions identify native support; current GeneScope variants cover all 646 conditions. Unsupported conditions receive a fixed reference score, not predicted expression.

Method	Δr	$\Delta\rho$	Top-100	Dir.	AUPRC
Control mean (646/646)	0.019	0.122	0.467	0.408	0.442
Additive response (646/646)	0.089	0.172	0.457	0.450	0.430
Ridge (646/646)	0.133	0.193	0.447	0.461	0.421
CPA (529/646)	0.138	0.174	0.333	0.576	0.285
chemCPA (529/646)	0.134	0.176	0.335	0.574	0.290
scGen (529/646)	0.329	0.305	0.338	0.713	0.308
scVIDR (529/646)	0.186	0.198	0.330	0.600	0.295
STATE-ST (529/646)	0.353	0.292	0.395	0.702	0.381
GeneScope	0.497	0.339	0.471	0.795	0.471
GeneScope + ensemble	**0.504**	**0.347**	0.475	**0.800**	0.476
GeneScope + SE	0.483	0.332	0.469	0.780	0.467
GeneScope + SE + ensemble	0.496	**0.347**	**0.476**	0.789	**0.477**

Table 25: SciPlex-middle: retrospective full-denominator coverage utility, Eq. 17. Fractions identify native support; current GeneScope variants cover all 1036 conditions. Unsupported conditions receive a fixed reference score, not predicted expression.

Method	Δr	$\Delta\rho$	Top-100	Dir.	AUPRC
Control mean (1036/1036)	0.127	0.090	0.340	0.573	0.294
Additive response (1036/1036)	0.125	0.093	0.328	0.575	0.285
Ridge (1036/1036)	**0.135**	**0.097**	0.336	**0.585**	0.292
CPA (808/1036)	0.085	0.050	0.325	0.526	0.302
chemCPA (808/1036)	0.055	0.030	0.296	0.523	0.260
scGen (808/1036)	0.079	0.054	0.328	0.541	0.301
scVIDR (536/1036)	0.044	0.020	0.238	0.514	0.220
STATE-ST (808/1036)	0.073	0.057	0.296	0.528	0.268
GeneScope	0.103	0.070	0.401	0.548	0.359
GeneScope + ensemble	0.105	0.071	0.406	0.550	0.363
GeneScope + SE	0.105	0.092	0.404	0.534	0.369
GeneScope + SE + ensemble	0.108	0.094	**0.411**	0.536	**0.374**

Table 26: cellxgene–SciPlex3: retrospective full-denominator coverage utility, Eq. 17. Fractions identify native support; current GeneScope variants cover all 1040 conditions. Unsupported conditions receive a fixed reference score, not predicted expression.

Method	Δr	$\Delta\rho$	Top-100	Dir.	AUPRC
Control mean (1040/1040)	0.143	0.144	0.094	0.629	0.078
Additive response (1040/1040)	0.149	0.146	0.097	0.640	0.082
Ridge (1040/1040)	**0.152**	**0.151**	0.106	**0.642**	0.087
chemCPA (384/1040)	0.015	0.015	0.047	0.521	0.052
scGen (812/1040)	0.049	0.050	0.029	0.539	0.043
scVIDR (540/1040)	0.028	0.032	0.033	0.512	0.044
STATE-ST (812/1040)	0.087	0.086	0.101	0.579	0.086
GeneScope	0.079	0.065	**0.153**	0.573	**0.115**
GeneScope + ensemble	0.080	0.065	**0.153**	0.574	**0.115**
GeneScope + SE	0.097	0.085	0.112	0.595	0.083
GeneScope + SE + ensemble	0.099	0.086	0.110	0.597	0.082