ATTRIBUTION-BASED DIAGNOSTICS FOR SINGLE-CELL FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Single-cell RNA foundation models (scFMs) are increasingly used for *in silico* biological discovery, yet predictive performance alone does not reveal whether their decisions rely on biologically meaningful signals or model-specific shortcuts. We present a systematic post-hoc attribution study of downstream predictions in two applications: cell-type annotation with pretrained Cell2Sentence and scGPT models, and a controlled scGPT+STATE perturbation-prediction case study. For cell annotation, our analyses connect predictions to established marker genes while examining how gene relevance varies with model architecture, expression rank, and the surrounding expression profile. In a synthetic perturbation prediction setting, we compare attribution patterns with known regulatory ground truth to assess what they reveal about the model's learned computations. Across these tasks, attribution exposes biological feature use, higher-order dependencies, pathway-specific computation, learning progression, prediction reliability, and potential failure modes that are not fully captured by aggregate output metrics. Experiments with complete and sparse regulatory knowledge further show that the conclusions drawn from attribution depend critically on the chosen biological reference, baseline, and null model. Together, these results establish attribution as a flexible post-hoc diagnostic for understanding, debugging, and guiding the development of scFMs.

1 INTRODUCTION

The development of single-cell foundation models (scFMs) is transforming biological research, with the promise of accelerating discovery through *in silico* experimentation (Jiang et al., 2026). However, extrapolation remains challenging. Generalizing beyond the training distribution requires models to capture biological mechanisms that remain stable across changing conditions rather than relying on spurious correlations (Tejada-Lapuerta et al., 2025). This motivates methods for assessing whether the internal computations underlying scFM predictions are biologically plausible, both to better understand model failures and to evaluate whether successful predictions arise for biologically meaningful reasons (Tejada-Lapuerta et al., 2025; Dimitrov et al., 2026).

Interpretability provides a natural framework for such analysis. Beyond supporting model debugging and trust (Räuker et al., 2023), inspecting learned representations and computations can provide evidence for learned internal models of the underlying domain (Li et al., 2025) and may itself serve as a tool for scientific discovery (Freiesleben et al., 2024). Existing interpretability studies of scFMs have primarily analyzed intermediate representation spaces, identifying biological information encoded in cell or gene embeddings (see Section 2). However, the presence of biological information in a representation does not reveal which input features support that information or how it is used by downstream predictors built on top of the representation. A more direct analysis therefore requires linking downstream predictions back to biologically meaningful input features.

In this work, we use attribution-based explainability to assess whether downstream predictors built on scFM representations rely on biologically meaningful signals consistent with established biological knowledge, as illustrated in Fig. 1. We study two common applications of increasing complexity. For cell type annotation, we assess whether predictions rely on established cell-type marker genes beyond confounding factors such as expression rank, and use symbolic explanations to examine how the relevance of marker genes shared by closely related subtypes changes in different contextual settings. For perturbation prediction, we assess how closely attribution patterns align with the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

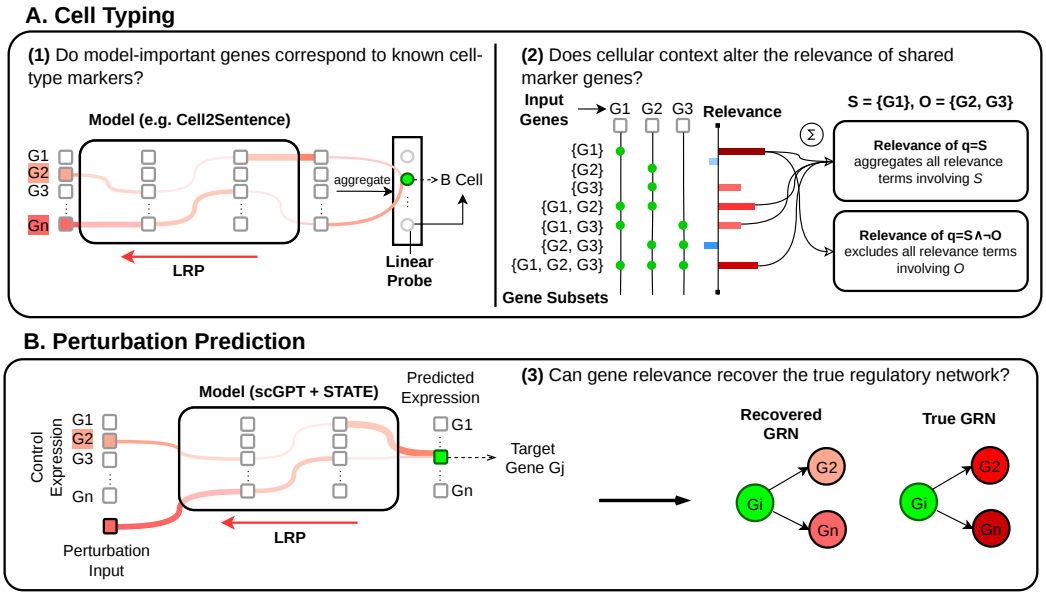

Figure 1: Overview of our framework for explaining downstream predictions of single-cell foundation models. **A.1.** Layer-wise Relevance Propagation (LRP) decomposes a cell-type probe logit into per-gene relevance scores, which are compared with known markers of the predicted cell type. **A.2.** Symbolic XAI decomposes the probe logit into relevance scores for gene subsets, which can be combined into logical queries to quantify how the relevance of marker genes changes in different contextual settings. **B.3.** For each perturbation of gene G_i and output gene G_j, LRP attributes the predicted expression of G_j to the model inputs, quantifying the contribution of G_i. Repeating this across perturbations yields a gene–gene relevance matrix that can be interpreted as an attribution-based estimate of regulatory influence and compared with the ground-truth gene regulatory network (GRN). Red paths in A and B indicate relevance propagation through the models, with line width and color intensity encoding relevance magnitude.

underlying gene regulatory network, providing a diagnostic of what the model has learned. Because reliable regulatory ground-truth is generally unavailable in real perturbation datasets, we construct a synthetic setting with a fully specified causal regulatory system, enabling direct comparison between model explanations and the mechanisms used to generate the data. Our contributions are threefold:

(1) We provide the first systematic study of post-hoc scFM explainability that focuses on dissecting downstream task predictions rather than only analyzing information encoded in intermediate representations or attention patterns.

(2) We develop attribution-based analyses that connect downstream predictions to established biological knowledge at both the level of individual genes and higher-order contextual interactions (Section 3).

(3) We demonstrate that attribution-based explainability can serve as a practical tool for analyzing gene regulatory structure learned by scFMs, enabling direct evaluation of whether model attributions recover known regulatory relationships in a controlled perturbation setting (Section 4).

2 Related Work

Post-hoc interpretability of biological foundation models. Interpretability has been increasingly used to investigate the biological information learned by foundation models. In protein language models, supervised probing of learned embeddings has been used to evaluate the decodability of biological attributes at both residue and protein levels (Li et al., 2024; Whitfield et al., 2026), while intervention-based analyses have revealed three-dimensional contacts even in models trained only

on protein sequences (Nainani et al., 2025; Candido et al., 2026). Unsupervised sparse autoencoders have further been used to decompose protein-model representations into more interpretable features (Rives et al., 2021; Adams et al., 2025; Simon & Zou, 2025; Gujral et al., 2025; Nainani et al., 2025; Candido et al., 2026), revealing features associated with structural, physicochemical, and functional properties. Beyond proteins, interpretability analyses of DNA foundation-model embeddings have also uncovered phylogenetic structure (Pearce et al., 2025).

Post-hoc interpretability of single-cell foundation models. Within transcriptomics, recent work has largely examined the representations learned by scFMs. Sparse autoencoders have uncovered cell- and gene-level features associated with known pathways and marker genes (Claye et al., 2025; Pedrocchi et al., 2025), while zero-shot analyses have studied how embeddings organize biological properties such as cell identity relative to technical factors (Kedzierska et al., 2025). Attribution methods have further estimated the influence of individual genes on learned representations (Zhao et al., 2026; Gold et al., 2026) or interpretable features derived from them (Claye et al., 2025). Attention weights have also been used to rank genes for annotated cell types (Yang et al., 2022) and to investigate recovery of regulatory interactions (Kendiukhov, 2026). However, attention is output-agnostic and captures only one step of the computation, so it need not reflect what drives a prediction (Jain & Wallace, 2019; Wiegreffe & Pinter, 2019).

These studies show that scFM representations encode biologically meaningful information, but largely focus on intermediate representations. In contrast, we explain *downstream task predictions* by propagating a chosen output through the full network with LRP (Bach et al., 2015), assigning each input feature a signed relevance for that prediction. While Keyl et al. (2023) use LRP on a shallow, task-specific network trained to impute masked gene expression to construct cell-specific regulatory networks, we use it to audit downstream predictions from pretrained scFMs. This allows us to test whether biological signals are not only represented, but actually used for prediction. We evaluate these explanations against explicit biological hypotheses, including known cell-type markers and the ground-truth gene regulatory network in a controlled perturbation setting.

3 Cell Type Classification

Cell typing is a fundamental task in single-cell analysis, where each cell is assigned to a biologically meaningful cell type based on its gene-expression profile. Although single-cell foundation models can achieve strong cell-type classification performance, predictive accuracy alone does not reveal what biological information the models use to distinguish cell populations. We therefore use cell typing as our first interpretability setting, asking whether model predictions are driven by biologically meaningful signals.

3.1 Experimental setup

Datasets. We curate two benchmarks from the CZ CELLxGENE Census (CZI Cell Science Program et al., 2025) with well-characterized molecular signatures, comprising peripheral blood cells and adult human brain nuclei. Each sample is represented by its gene expression profile. The blood dataset has a two-level label hierarchy, with four coarse cell types (T cells, B cells, natural killer cells, and monocytes), further divided into 22 subtypes. The brain dataset contains seven broad cell types, including excitatory and inhibitory neurons, astrocytes, oligodendrocytes, oligodendrocyte precursor cells, microglia/macrophages, and vascular/perivascular cells. Both datasets contain $\sim$61K samples, and are split into train, validation, and test sets using a 70/15/15 ratio. Further dataset details are provided in Appendix A.

Models and context settings. We evaluate two families of pretrained single-cell foundation models throughout this study, Cell2Sentence (C2S; Levine et al., 2024; Rizvi et al., 2026) and scGPT (Cui et al., 2024). C2S models are adapted from pretrained autoregressive language models (here, Pythia-410M, Pythia-1B, Gemma2-2B). They represent cells as sequences of top-K genes sorted by expression value, tokenized with the original text vocabulary. We use $K = 200$ for C2S (Pythia-410M) and both $K = 200$ and $K = 1000$ for the larger C2S-Scale models, Pythia-1B and Gemma2-2B, allowing us to examine the effects of model scale and input context. scGPT models are encoder-only Transformers that represent gene names and their expression values as combined

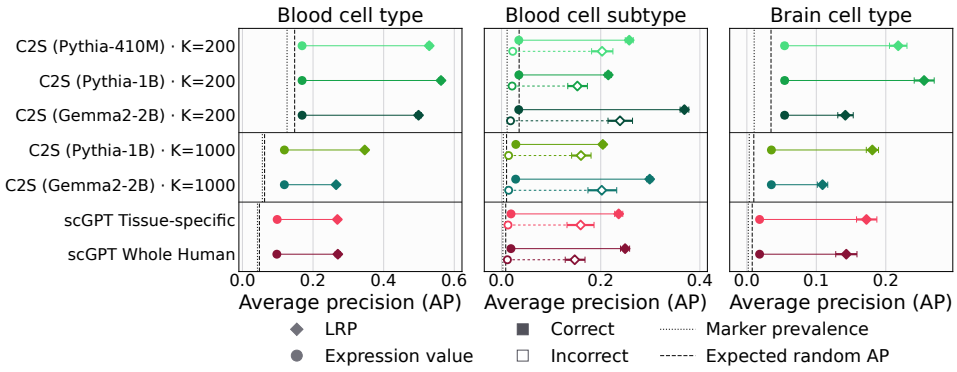

Figure 2: Mean marker-recovery average precision (AP) across tasks and models. Class-stratified bootstrapping 95% CIs are also shown, though many are too narrow to visualise at this scale. Relevance scores recover marker genes better than expression values and random ranking. For blood cell subtypes, AP is higher for correct than incorrect predictions, indicating that prediction errors are associated with weaker marker alignment. Incorrect predictions are reported only for blood subtypes, since each model makes at most five errors on the broader cell-type tasks. Marker prevalence and expected AP under random finite rankings are shown for reference.

token embeddings and were trained with a maximum context length of $K = 1200$. We consider scGPT models pretrained on blood, brain, and whole-human cells.

Cell-type classification with linear probes. To obtain task-aligned predictions without modifying the pretrained scFMs, we freeze each model and train linear probes on final-layer representations to predict blood cell types and subtypes, and brain cell types. This is necessary even for C2S models that can generate cell-type labels directly, since their generated labels may not match the annotation granularity of our datasets. For C2S, we compare mean pooling over non-padding tokens with the final-token representation, and for scGPT mean pooling with the global cell representation (`[CLS]`), selecting the better-performing option per model configuration and task based on validation accuracy. The probes achieve high accuracy overall, with near-perfect performance on coarse cell types. Hyperparameters and pooling/layer sweep results are detailed in Appendix B.

3.2 Marker recovery beyond expression value

To assess whether linear-probe predictions are supported by biologically meaningful evidence, we examine whether the genes most relevant to each prediction align with established cell-type markers rather than merely reflecting expression magnitude. We use Layer-wise Relevance Propagation (LRP; Bach et al., 2015) to assign each input gene a signed relevance score indicating how strongly it supports or opposes the probe prediction. LRP obtains these scores by propagating the model output backward through the computational graph.[1] Please refer to Appendix C for details.

We evaluate marker recovery using average precision (AP), which measures how strongly marker genes are concentrated near the top of a gene ranking. For each cell, genes in the marker set $\mathcal{M}$ are treated as positives and all remaining genes as negatives. Genes are ranked by their relevance scores, and AP is computed from the precision–recall curve as $\mathrm{AP} = \sum_i (R_i - R_{i-1}) P_i$, where P_i and R_i denote precision and recall at the i-th score threshold. Tied scores are evaluated jointly, making AP invariant to their input ordering. We compute AP per cell using signed gene-level relevance scores and, as a model-free baseline, gene-expression values. Scores are then macro-averaged across biological classes, with correct and incorrect predictions reported separately.

Figure 2 shows that relevance scores recover marker genes above chance across models and tasks. For blood cell subtyping, $\mathrm{AP}_{\mathrm{LRP}}$ is consistently higher for correct than incorrect predictions, suggesting that errors are associated with weaker alignment between relevance scores and known marker genes. The expression-value baseline varies little across context settings, suggesting that most marker genes are already concentrated among the top $K = 200$ highly expressed genes.

[1] For C2S models, the relevance of a gene is obtained by summing the scores over its name's sub-tokens.

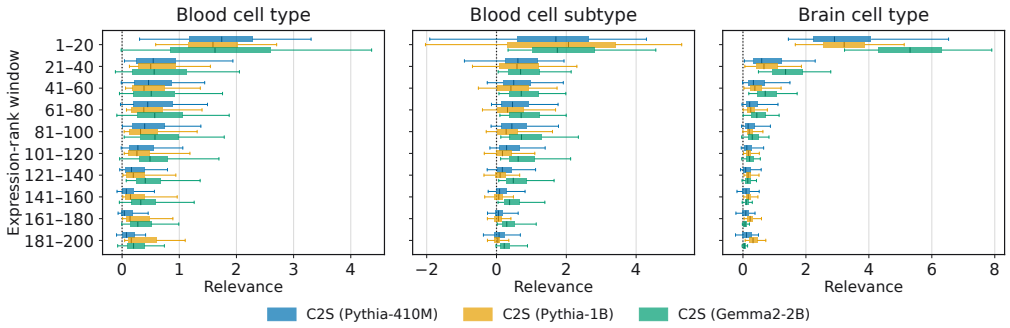

Figure 3: Distribution of relevance scores across gene-rank bins, ordered by decreasing expression. Box plots show the distribution across test cells. Relevance is concentrated among early, highly expressed genes and decreases with rank, most strongly for C2S (Gemma2-2B), although this pattern alone does not establish positional shortcut behavior.

Relevance scores for both C2S and scGPT models dramatically outperformed the expression-value baseline. C2S (Gemma2-2B) configurations show the strongest marker recovery for blood cell subtyping, but slightly weaker recovery than Pythia configurations for blood and brain coarse cell types, implying a nuanced effect of model scale and architecture.

Position bias in C2S. C2S represents each cell as a sequence of gene identifiers sorted by expression value, so token position reflects expression magnitude. Because the language model backbone also uses position through positional encodings and causal attention, biological relevance and sequence position are confounded, meaning a gene may be relevant because it is informative or because it appears early. To examine this effect, we use $K = 200$ genes and divide each input sequence into 10 equally-sized bins. For each cell, we normalize the gene-level relevance scores by the maximum absolute relevance score within that cell, sum the normalized scores within each bin, and then aggregate these bin-level scores across cells.

Figure 3 shows that relevance is strongly concentrated in the earliest positions and declines steadily with rank across all models, most prominently in C2S (Gemma2-2B). This pattern alone does not establish positional shortcut behaviour, since early positions correspond to highly expressed genes that may themselves be biologically informative.

3.3 Contextual relevance of shared marker genes via symbolic queries

Marker recovery shows whether predictions rely on biologically relevant genes, but not whether this relevance depends on the surrounding cellular context. This is important for closely related cell subtypes that share marker genes, whose discriminative value may emerge only through interactions with other genes. We use Symbolic XAI (Schnake et al., 2025), which quantifies the relevance of higher-order, human-interpretable logical queries, composed of input features and operators like $\neg$, $\vee$, and $\wedge$, to assess whether shared markers receive context-dependent relevance. Refer to Appendix D for further details.

Let $\mathcal{S}$ be the set of shared markers and $\mathcal{O}$ the remaining tokens in the input. We define two queries over the shared markers: $q_{\text{full}} = \mathcal{S}$, which measures the relevance of the shared markers in the full cellular context, and $q_{\text{minimal}} = \mathcal{S} \wedge \neg\mathcal{O}$, which measures their relevance when all other genes are absent from the context. The difference between the relevance assigned to these queries quantifies the extent to which the model is context-sensitive.

We study four pairs of closely related blood cell subtypes, each sharing two marker genes (Table A.2), using C2S (Gemma2-2B, $K = 1{,}000$) and scGPT (Whole Human). We restrict the analysis to correctly classified cells and explain the logit difference between each subtype pair (Eberle et al., 2023), such that positive relevance supports one subtype and negative relevance the other.

Figure 4 shows a clear separation between the two queries. The relevance of q_{minimal} remains near zero across subtype pairs, whereas q_{full} is substantially broader and often shifted. This suggests that shared marker genes provide little subtype-specific evidence on their own and become informative

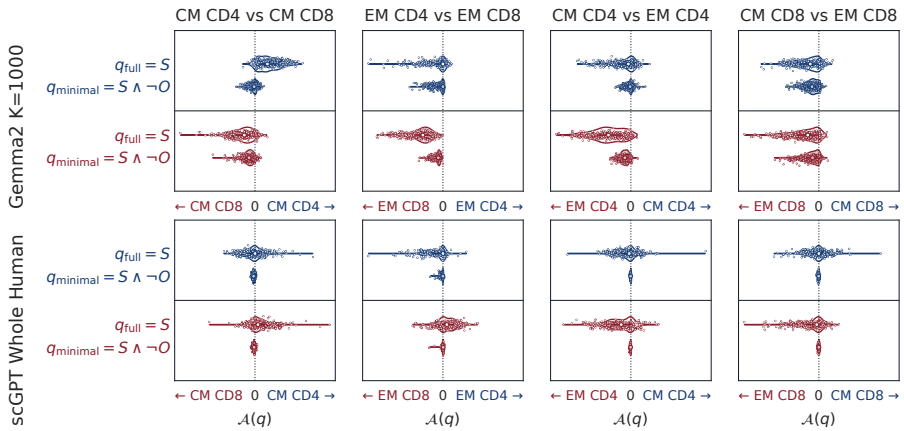

Figure 4: Context-sensitivity of C2S (Gemma2-2B) with $K = 1000$ and scGPT Whole Human using symbolic queries. The relevance of q_{minimal} is consistently concentrated near zero, whereas the full-context query q_{full} shows broader and often shifted relevance. This indicates that shared markers derive much of their relevance from interactions with the surrounding cellular context rather than from contributions confined to the shared-marker set.

mainly through interactions with the surrounding cellular context. For C2S, some shifts align with how strongly the shared markers are represented in the input. In EM CD4 versus EM CD8, *CCL5* and *GZMA* occur much more often in EM CD8 cell sentences, consistent with the shift of q_{full} toward EM CD8. Likewise, in CM CD4 versus EM CD4, *IL7R* is more abundant and higher ranked in EM CD4, matching the stronger full-context relevance toward EM CD4.

4 Perturbation Prediction

Having first examined cell typing as a relatively simple setting for assessing whether predictions rely on biologically meaningful signals, we extend our analysis to genetic perturbation prediction, a more complex task with important applications in areas such as drug discovery. This progression allows us to move from cell-identity signals to the regulatory dependencies that govern cellular responses to intervention. Recent benchmarks suggest that scFMs do not consistently outperform simpler baselines on perturbation prediction (Ahlmann-Eltze et al., 2025; Kernfeld et al., 2025; Bendidi et al., 2024), motivating a closer examination of what biological structure these models capture beyond predictive performance. We therefore use attribution-based analysis to assess whether the learned dependencies reflect the underlying gene regulatory network.

4.1 Experimental Setup

Synthetic dataset. Ground-truth regulatory relationships are often unavailable in real single-cell perturbation datasets, making direct evaluation of whether model predictions reflect the underlying gene regulatory network (GRN) difficult. We therefore use a synthetic setting with fully known regulatory structure. Following Qi & Chapfuwa (2026); Lorch et al. (2026), cells are generated from context-specific linear stochastic differential equations followed by a noisy emission process, yielding two cellular contexts over 128 genes. We generate 8M unperturbed control cells for pretraining and, for fine-tuning, an additional 5M controls and 1M perturbed cells across both contexts. The perturbed set contains 8,192 independently simulated cells for each of 128 randomly signed single-gene perturbations, covering all genes. Controls and perturbed cells are sampled independently from their respective stationary populations and thus have no one-to-one correspondence. Refer to Appendix A.3 for full details.

Model. For perturbation analysis, we use scGPT as the backbone in a two-stage training procedure. We first pretrain scGPT from scratch on control cells using masked gene-expression reconstruction (Cui et al., 2024). We then feed its cell embeddings to the State Transition architecture of STATE (Adduri et al., 2025) and fine-tune the joint model on perturbation data. STATE processes

sets of control cells and mixes their intermediate representations through self-attention to predict distributions of perturbed cell states. Details are provided in Appendix E. To distinguish learned behavior from architectural inductive biases, we also evaluate two random-weight baselines with five replicates each, fully random initialization and pretrained scGPT with an untrained STATE head. The model preserves perturbation identity (centroid accuracy (0.962)) and captures dominant response directions (Pearson $\Delta_{20} = 0.704$), despite worse all-gene RMSE than the perturbed-mean baseline (0.855 vs. 0.606). Full results and metric definitions are provided in Appendices F and H.

4.2 Recovering Regulatory Relationships with LRP

We use LRP to test whether scGPT+STATE has learned the direct and indirect gene–gene relationships encoded in the synthetic GRN. For each perturbed gene, the ground-truth GRN specifies which other genes it directly regulates. We therefore ask whether highly relevant genes coincide with these known direct targets.

All quantities below are evaluated separately for each cellular context, and we omit the context index c for readability. The simulator provides a regulatory matrix $\mathbf{A}$, where off-diagonal entry $A_{j,i}$ represents direct regulation $i \to j$. A perturbation q targets gene i_q and is represented by a signed, sparse intervention vector $\boldsymbol{\gamma}_q \in \mathbb{R}^G$, non-zero only at index i_q. We consider two ground-truth perturbation effect vectors:

$$\mathbf{d}_q = \mathbf{A}\boldsymbol{\gamma}_q, \qquad \mathbf{t}_q = -\mathbf{A}^{-1}\boldsymbol{\gamma}_q, \qquad \mathbf{d}_q, \mathbf{t}_q \in \mathbb{R}^G \tag{1}$$

$\mathbf{d}_q$ encodes the direct effects of perturbation q on genes directly regulated by i_q. Under the simulator's linear dynamics, the total stationary effect $\mathbf{t}_q$ gives the expected change in latent cell state (Qi & Chapfuwa, 2026) and incorporates direct and indirect effects. For each perturbation q, a signature m yields a gene-level score vector $\mathbf{s}_{m,q} \in \mathbb{R}^G$, where entry j scores output gene j after averaging over cells and STATE batches. We quantify the *regulatory alignment* of each signature by comparing $\mathbf{s}_{m,q}$ with the ground-truth direct-effect vector $\mathbf{d}_q$ and total stationary-effect vector $\mathbf{t}_q$.

LRP signatures. We consider two LRP signatures. *Perturbation relevance* measures how much the STATE perturbation input contributes to each predicted expression value. *Same-cell gene LRP* propagates relevance backward through scGPT to the input genes of the corresponding control cell. The signatures are evaluated to be faithful to the model and robust to the choice of LRP rules (Appendix C). Same-cell gene LRP initially yields a $G \times G$ array per perturbation, indexed by input and output genes. To reduce it to a signature $\mathbf{s}_{m,q}$, we keep only the input entries corresponding to the perturbed gene i_q. These signatures measure contributions of that gene expression in the control state, rather than the effect of the perturbation.

Baselines. As baselines, *Empirical DE* is the differential expression between observed perturbed and control cells and serves as an oracle upper-bound for the regulatory signal present in the data. *Predicted DE* is the corresponding differential expression computed from the model predictions and serves as a black-box baseline for the regulatory information captured in the model output. We also evaluate two black-box counterfactuals: ablating the perturbation input and masking gene i_q (Appendix H.2).

Random control. For each context, we construct a null distribution from 999 randomly rewired versions of its GRN that preserve each gene's in- and out-degree. We keep the signatures fixed and recompute the same recovery metrics against each randomized network, macro-averaged over perturbations. One-sided empirical p-values quantify how often the randomized GRNs achieve a score at least as high as the ground-truth GRN. The exact definition is provided in Appendix G.

4.3 Regulatory Alignment Beyond Differential Expression

We evaluate each signature along three axes. For each perturbed gene i_q, target gene binary recovery is evaluated with AP (ranking genes $j \neq i_q$ with $\mathbf{s}_{m,q}$ and treating non-zero entries of $\mathbf{d}_q$ as positives), similarly to Section 3.2 and also adopted by Pratapa et al. (2020) for GRN recovery. Sign agreement evaluates accuracy between the signs of $\mathbf{s}_{m,q}$ and $\mathbf{d}_q$. Alignment of a candidate signature $\mathbf{s}_{m,q}$ with the total effect $\mathbf{t}_q$ is measured using Spearman correlation. Metrics are computed per perturbation, then macro-averaged, and are reported in Fig. 5.

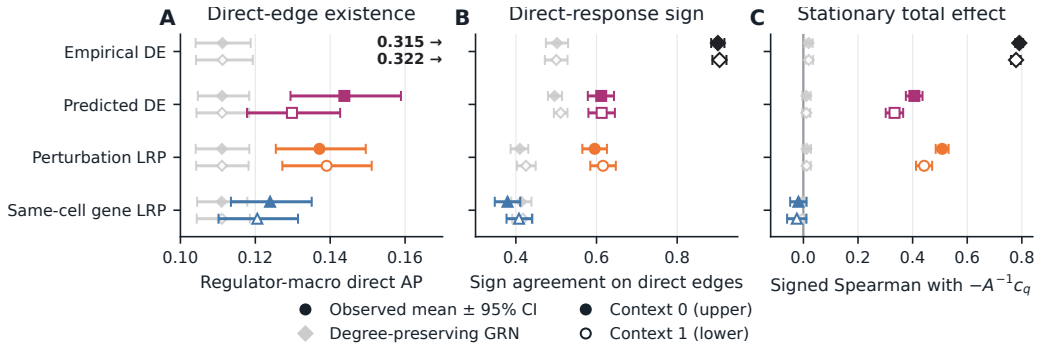

Figure 5: Evaluation of regulatory alignment in both in-distribution contexts (0 and 1). **(A)** Binary recovery of direct edge existence, measured as average precision (AP). Empirical DE is clipped for better visualisation. **(B)** Sign agreement with nonzero ground-truth direct responses. **(C)** Spearman correlation with stationary total effects. Error bars for the true measurements reflect 95% bootstrap CIs over the 128 perturbations. The grey control error bars are the central 95% range of the perturbation-average evaluated against the randomised-GRN controls.

All exploratory development used context 0 test cells, with context 1 evaluated only after the analysis protocol was fixed. Comparable results across contexts and all three metrics indicate that the findings are robust to this controlled mechanism shift rather than specific to the development context. Overall, empirical DE shows substantially stronger regulatory alignment than the model-derived signatures, consistent with the model's moderate perturbation-prediction performance. Perturbation-LRP performs similarly to predicted DE across all three metrics, with slightly stronger alignment to total stationary effects. Gene-LRP shows modest but significant direct-edge recovery compared with the randomized-GRN control ($p = 0.001$), but does not recover response signs or total effects. Complementary analyses with extended baselines can be found in Appendix H.2.

4.4 Attribution-based model diagnostics

Relevance concentration and prediction quality. We next ask whether the share of model relevance assigned to different input sources is associated with perturbation-prediction quality. For each perturbation, we compute the fraction of total absolute relevance assigned to each input source, averaged across cells and STATE batches, and correlate these fractions with Pearson-Δ using Spearman correlation. In-distribution, the relevance fraction assigned to the explicit perturbation input is strongly associated with prediction quality, reaching $\rho = 0.897$. Because this association could partly reflect properties of the black-box model output, we additionally control for predicted-DE dispersion, sparsity, and magnitude. The association remains substantial after adjustment, with $\rho = 0.666$, indicating that relevance assigned to the perturbation input provides information about prediction quality beyond these output statistics.

Learning progression. Regulatory alignment, as measured in Section 4.3, would be useful to evaluate while training models, to track acquisition of mechanistic biological knowledge. However, it is unrealistic to assume *full* regulatory knowledge in practical applications. Therefore, we simulate a scenario wherein an analyst would ascertain regulatory interactions for only the most differentially expressed genes (DEGs). For each perturbation, we select the top-six genes (excluding the perturbed gene) with the highest training-dataset empirical DE. The only revealed knowledge is which of the selected DEGs are directly regulated by the perturbed gene, with no interaction sign or magnitude. Because this is a positive-vs-unknown recovery problem and AP requires labelled negatives, we adopt recall@20 as the primary metric.

Figure 6A shows a consistent pattern across the sparse- and full-knowledge settings: recall increases with training for all signatures and is higher for perturbation-LRP than for gene-LRP. Further, perturbation-LRP displays a learned alignment signal by epoch 5, at which point predicted DE alignment was still not distinguishable from the random-model controls. This suggests that perturbation-specific information emerged in the perturbation pathway sooner than it was expressed in model outputs.

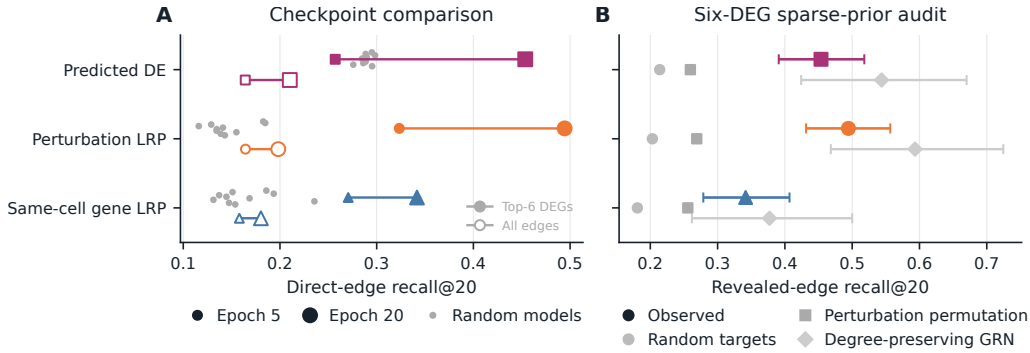

Figure 6: Regulator recovery under sparse prior knowledge for top-6 differentially expressed genes (DEGs), measured by recall@20. **(A)** Comparing the final model checkpoint (epoch 20) vs an early checkpoint (epoch 5) and random untrained model baselines. Recall increases with training, and at epoch 5 LRP signatures exceed the random untrained models while predicted DE does not. Results under full knowledge (all direct edges) are also shown for reference as hollow markers, though the numeric scales are not directly comparable as the denominators differ. **(B)** Recovering known regulatory edges, alongside randomised controls (details in Section 4.4).

Limits of sparse auditing. Though we have seen that attribution signals can serve practical diagnostic uses even with no or limited ground-truth knowledge, care must be taken not to over-interpret the results as true regulatory discovery. To illustrate this, we consider two typically well-motivated randomised controls (100 replicates each): random targets and perturbation permutation, respectively measuring the effect of DEG-panel selection and perturbation-specificity. Indeed, sparse regulator recovery results clearly surpass these controls for all evaluated signatures in Fig. 6B.

On the other hand, none of the signatures align with the sparse ground-truth significantly better than with the degree-preserving GRNs from Section 4.3. This implies that the signatures broadly prioritised the DEGs but did not distinguish which were direct targets of the perturbed gene, underscoring the need for careful experimental design and controls that preserve relevant graph structure.

5 Conclusion

Across cell-type classification and perturbation prediction, attribution analyses provided information complementary to predictive metrics. They revealed which marker genes, contextual interactions, input pathways, and potential shortcuts supported predictions, while also exposing learning progression and prediction reliability beyond output statistics. However, biological conclusions depended on appropriate references, informative baselines, and controls matched to the data-generating process. Together, these findings establish attribution-based explainability as a powerful post-hoc tool for debugging scFMs and guiding model development by revealing the biological signals, shortcuts, and computational pathways behind their successes and failures. Although our study shows a range of questions attribution can address, it covers selected pretrained models, linear-probe tasks, and a single perturbation model. More fundamentally, scFM interpretability is constrained by scarce biological ground-truth. Marker panels provide only partial references, while comprehensive GRNs are generally unavailable for real perturbation data, motivating our synthetic setting with fully known regulatory structure despite its limited biological complexity. Future work should test transfer across models, tasks, and datasets using complementary biological evidence, controlled interventions, and task-specific nulls. Broader validation could extend attribution beyond post-hoc analysis to training and model selection, supporting more robust and biologically grounded scFMs.

AI USE STATEMENT

In this work, we used generative AI tools for polishing the writing, identifying related work, and supporting research execution. We have not used generative AI tools to draft substantive sections of the paper, generate synthetic datasets, or make scientific conclusions. We have reviewed all AI-assisted work. In particular, author review verified that LLM-assisted writing accurately reflected our ideas, while literature retrieval was independently validated through manual literature review. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

For reproducibility, the CELLxGENE subsets used in this study, together with the external databases used to obtain marker-gene annotations, are described in detail in Appendix A.1 and Appendix A.2. The synthetic dataset generated for the perturbation-prediction experiments is described in Appendix A.3. For the cell-type classification experiments, the complete probe-training procedure is provided in Appendix B. The architecture and training procedure used for the perturbation-prediction model are detailed in Appendix E. The methodology and implementation details of our attribution-based analyses are provided in Appendix C and Appendix D. Additional information required to reproduce the experimental results is provided in the main paper, including the experimental setup in Section 3 and Section 4.

REFERENCES

Shiyu Jiang, Zhaoyu Fang, Yujie Zhang, Xuting Zhang, Jérémie Kalfon, Weixu Wang, Xi Fu, Aakash Patel, Syed Rizvi, Yuancheng Ryan Lu, Siyu He, Yixin Wang, Kejun Ying, Peter Pao-Huang, Yifan Lu, Weizu Xu, Mengchen Wang, Ziyang Miao, Jianhui Lin, Jimmy Ding, Zitong Jerry Wang, Wei Ouyang, Tianlong Chen, Guoxian Yu, Min Li, Jiayi Ma, Fei Wang, Yuying Xie, Jiliang Tang, Raul Rabadan, David van Dijk, Pengtao Xie, Peng He, Emily B. Fox, Le Song, Fabian J. Theis, Eric Xing, Christina V. Theodoris, Xiaojie Qiu, and Jiayuan Ding. The landscape of single-cell foundation models: Design principles, applications, and open challenges, August 2026. doi: 10.20944/preprints202608.1166.v1.

Alejandro Tejada-Lapuerta, Paul Bertin, Stefan Bauer, Hananeh Aliee, Yoshua Bengio, and Fabian J. Theis. Causal machine learning for single-cell genomics. *Nature Genetics*, 57(4):797–808, April 2025. doi: 10.1038/s41588-025-02124-2.

Daniel Dimitrov, Stefan Schrod, Martin Rohbeck, and Oliver Stegle. Interpretation, extrapolation and perturbation of single cells. *Nature Reviews Genetics*, pp. 1–22, January 2026. doi: 10.1038/s41576-025-00920-4.

Tilman Räuker, Anson Ho, Stephen Casper, and Dylan Hadfield-Menell. Toward transparent AI: A survey on interpreting the inner structures of deep neural networks. In *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pp. 464–483, February 2023. doi: 10.1109/SaTML54575.2023.00039.

Kenneth Li, Fernanda Viégas, and Martin Wattenberg. What does it mean for a neural network to learn a "world model"? arXiv:2507.21513, July 2025. doi: 10.48550/arXiv.2507.21513.

Timo Freiesleben, Gunnar König, Christoph Molnar, and Álvaro Tejero-Cantero. Scientific inference with interpretable machine learning: Analyzing models to learn about real-world phenomena. *Minds and Machines*, 34(3):32, July 2024. doi: 10.1007/s11023-024-09691-z.

Francesca-Zhoufan Li, Ava P. Amini, Yisong Yue, Kevin K. Yang, and Alex Xijie Lu. Feature reuse and scaling: Understanding transfer learning with protein language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 27351–27375. PMLR, July 2024. URL https://proceedings.mlr.press/v235/li24a.html.

Shawn Whitfield, Tom Marty, Robert M. Vernon, Christopher James Langmead, Dhanya Sridhar, and Quentin Fournier. High-resolution dissection of concept acquisition in different families of

protein language models. bioRxiv preprint 2026.07.20.739599, July 2026. doi: 10.64898/2026.07.20.739599.

Jatin Nainani, Bryn Marie Reimer, Connor Watts, David Jensen, and Anna G. Green. Mechanistic evidence that motif-gated domain recognition drives contact prediction in protein language models. bioRxiv preprint 2025.08.22.671739, August 2025. doi: 10.1101/2025.08.22.671739.

Salvatore Candido, Thomas Hayes, Alexander Derry, Roshan Rao, Zeming Lin, Robert Verkuil, Bryan Z. Wu, Jin Sub Lee, Elise S. Bruguera, Jehan A. Keval, Mykhailo Kopylov, John E. Pak, Wesley Wu, Neil Thomas, Samson Mataraso, Alvin Hsu, Ashton C. Trotman-Grant, Kilian Fatras, Allan dos Santos Costa, Rohil Badkundri, Halil Akın, Deniz Oktay, Jonathan Deaton, Elizabeth Montabana, Hrishita Sitwala, Yue Yu, Marius Wiggert, Dylan Alexander Carlin, Anthony W. Goering, Tomasz Blazejewski, McCullen Sandora, Michael Hla, Tina Z. Jia, Leon H. Kloker, Nicholas J. Sofroniew, Masatoshi Uehara, Jassi Pannu, Sharrol Bachas, Daniel S. Liu, Tom Sercu, and Alexander Rives. Language modeling materializes a world model of protein biology. bioRxiv preprint 2026.06.03.729735, June 2026. doi: 10.64898/2026.06.03.729735.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jerry Ma, and Rob Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, April 2021. doi: 10.1073/pnas.2016239118.

Etowah Adams, Liam Bai, Minji Lee, Yiyang Yu, and Mohammed AlQuraishi. From mechanistic interpretability to mechanistic biology: Training, evaluating, and interpreting sparse autoencoders on protein language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 460–476. PMLR, June 2025. URL https://proceedings.mlr.press/v267/adams25a.html.

Elana Simon and James Zou. InterPLM: Discovering interpretable features in protein language models via sparse autoencoders. *Nature Methods*, 22(10):2107–2117, October 2025. doi: 10.1038/s41592-025-02836-7.

Onkar Gujral, Mihir Bafna, Eric Alm, and Bonnie Berger. Sparse autoencoders uncover biologically interpretable features in protein language model representations. *Proceedings of the National Academy of Sciences*, 122(34):e2506316122, August 2025. doi: 10.1073/pnas.2506316122.

Michael Pearce, Elana Simon, Michael Byun, and Daniel Balsam. Finding the tree of life in Evo 2. *Goodfire Research*, August 2025. URL https://www.goodfire.ai/papers/phylogeny-manifold.

Charlotte Claye, Pierre Marschall, Wassila Ouerdane, Céline Hudelot, and Julien Duquesne. Discovering interpretable biological concepts in single-cell RNA-seq foundation models. arXiv:2510.25807, October 2025. doi: 10.48550/arXiv.2510.25807.

Flavia Pedrocchi, Florian Barkmann, Amir Joudaki, and Valentina Boeva. Sparse autoencoders reveal interpretable features in single-cell foundation models. bioRxiv preprint 2025.10.22.681631, October 2025. doi: 10.1101/2025.10.22.681631.

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(1):101, April 2025. doi: 10.1186/s13059-025-03574-x.

Yubo Zhao, Duozhi Sun, Minsheng Hao, Yifan Xiong, Chen Li, Tieliang Gong, Lei Wei, and Xuegong Zhang. How different ai models understand cells differently. bioRxiv preprint 2026.01.29.702682, January 2026. doi: 10.64898/2026.01.29.702682.

Maxwell P. Gold, Miguel Reyes, Nathaniel Diamant, Tony Kuo, Ehsan Hajiramezanali, Jane W. Newburger, Mary Beth F. Son, Pui Y. Lee, Gabriele Scalia, Aicha BenTaieb, Sharookh B. Kapadia, Anupriya Tripathi, Héctor Corrada Bravo, Graham Heimberg, and Tommaso Biancalani. Scoring gene importance by interpreting single-cell foundation models. *Nature Biotechnology*, May 2026. doi: 10.1038/s41587-026-03112-5.

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(10):852–866, October 2022. doi: 10.1038/s42256-022-00534-z.

Ihor Kendiukhov. Systematic evaluation of single-cell foundation model interpretability: Attention-derived edge scores add no incremental value over gene-level features for perturbation-target prediction. *BMC Genomics*, 27(1):634, July 2026. doi: 10.1186/s12864-026-12965-8.

Sarthak Jain and Byron C. Wallace. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019.

Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 11–20, Hong Kong, China, 2019. Association for Computational Linguistics.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

Philipp Keyl, Philip Bischoff, Gabriel Dernbach, Michael Bockmayr, Rebecca Fritz, David Horst, Nils Blüthgen, Grégoire Montavon, Klaus-Robert Müller, and Frederick Klauschen. Single-cell gene regulatory network prediction by explainable AI. *Nucleic Acids Research*, 51(4):e20, 2023. doi: 10.1093/nar/gkac1212. URL https://doi.org/10.1093/nar/gkac1212.

CZI Cell Science Program, Shibla Abdulla, Brian Aevermann, Pedro Assis, Seve Badajoz, Sidney M Bell, Emanuele Bezzi, Batuhan Cakir, Jim Chaffer, Signe Chambers, J Michael Cherry, Tiffany Chi, Jennifer Chien, Leah Dorman, Pablo Garcia-Nieto, Nayib Gloria, Mim Hastie, Daniel Hegeman, Jason Hilton, Timmy Huang, Amanda Infeld, Ana-Maria Istrate, Ivana Jelic, Kuni Katsuya, Yang Joon Kim, Karen Liang, Mike Lin, Maximilian Lombardo, Bailey Marshall, Bruce Martin, Fran McDade, Colin Megill, Nikhil Patel, Alexander Predeus, Brian Raymor, Behnam Robatmili, Dave Rogers, Erica Rutherford, Dana Sadgat, Andrew Shin, Corinn Small, Trent Smith, Prathap Sridharan, Alexander Tarashansky, Norbert Tavares, Harley Thomas, Andrew Tolopko, Meghan Urisko, Joyce Yan, Garabet Yeretssian, Jennifer Zamanian, Arathi Mani, Jonah Cool, and Ambrose Carr. CZ CELLxGENE Discover: A single-cell data platform for scalable exploration, analysis and modeling of aggregated data. *Nucleic Acids Research*, 53(D1):D886–D900, January 2025. doi: 10.1093/nar/gkae1142.

Daniel Levine, Syed A. Rizvi, Sacha Lévy, Nazreen Pallikkavaliyaveetil, David Zhang, Xingyu Chen, Sina Ghadermarzi, Ruiming Wu, Zihe Zheng, Ivan Vrkic, Anna Zhong, Daphne Raskin, Insu Han, Antonio Henrique De Oliveira Fonseca, Josue Ortega Caro, Amin Karbasi, Rahul Madhav Dhodapkar, and David van Dijk. Cell2Sentence: Teaching large language models the language of biology. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 27299–27325. PMLR, July 2024. URL https://proceedings.mlr.press/v235/levine24a.html.

Syed Asad Rizvi, Daniel Levine, Aakash Patel, Shiyang Zhang, Eric Wang, Curtis Jamison Perry, Ivan Vrkic, Nicole Mayerli Constante, Zirui Fu, Sizhuang He, David Zhang, Cerise Tang, Zhuoyang Lyu, Rayyan Darji, Chang Li, Emily Sun, David Jeong, Lawrence Zhao, Jennifer Kwan, David Braun, Brian Hafler, Hattie Chung, Rahul M. Dhodapkar, Paul Jaeger, Bryan Perozzi, Jeffrey Ishizuka, Shekoofeh Azizi, and David van Dijk. Scaling large language models for next-generation single-cell analysis. bioRxiv preprint 2025.04.14.648850, January 2026. doi: 10.1101/2025.04.14.648850.

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

Thomas Schnake, Farnoush Rezaei Jafari, Jonas Lederer, Ping Xiong, Shinichi Nakajima, Stefan Gugler, Grégoire Montavon, and Klaus-Robert Müller. Towards symbolic XAI — explanation

through human understandable logical relationships between features. *Information Fusion*, 118: 102923, June 2025. doi: 10.1016/j.inffus.2024.102923.

Oliver Eberle, Ilias Chalkidis, Laura Cabello, and Stephanie Brandl. Rather a nurse than a physician - contrastive explanations under investigation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

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

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

Ihab Bendidi, Shawn Whitfield, Kian Kenyon-Dean, Hanene Ben Yedder, Yassir El Mesbahi, Emmanuel Noutahi, and Alisandra K. Denton. Benchmarking Transcriptomics Foundation Models for Perturbation Analysis : One PCA still rules them all. arXiv:2410.13956, November 2024. doi: 10.48550/arXiv.2410.13956.

Shi-ang Qi and Paidamoyo Chapfuwa. Mechanisms matter: Transportability of cellular perturbation effects. bioRxiv preprint 2026.05.08.723625, May 2026. doi: 10.64898/2026.05.08.723625.

Lars Lorch, Jiaqi Zhang, Charlotte Bunne, Andreas Krause, Bernhard Schölkopf, and Caroline Uhler. Latent causal diffusions for single-cell perturbation modeling. arXiv:2601.15341, January 2026. doi: 10.48550/arXiv.2601.15341.

Abhinav K. Adduri, Dhruv Gautam, Beatrice Bevilacqua, Alishba Imran, Rohan Shah, Mohsen Naghipourfar, Noam Teyssier, Rajesh Ilango, Sanjay Nagaraj, Mingze Dong, Chiara Ricci-Tam, Christopher Carpenter, Vishvak Subramanyam, Aidan Winters, Sravya Tirukkovular, Jeremy Sullivan, Brian S. Plosky, Basak Eraslan, 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. bioRxiv preprint 2025.06.26.661135, July 2025. doi: 10.1101/2025.06.26.661135.

Aditya Pratapa, Amogh P. Jalihal, Jeffrey N. Law, Aditya Bharadwaj, and T. M. Murali. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nature Methods*, 17(2):147–154, February 2020. doi: 10.1038/s41592-019-0690-6.

Claire E. Gustafson, Peter J. Skene, Ananda W. Goldrath, Xiao-Jun Li, Troy R. Torgerson, Lynne A. Becker, et al. AIFI immune health atlas. doi: 10.57785/e9e1-wh09. URL https://apps.allenimmunology.org/aifi/resources/imm-health-atlas/.

Qiuyu Gong, Mehul Sharma, Marla C. Glass, Emma L. Kuan, Aishwarya Chander, Mansi Singh, et al. Multi-omic profiling reveals age-related immune dynamics in healthy adults. *Nature*, 648 (8094):696–706, 2025. doi: 10.1038/s41586-025-09686-5.

Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, William M. Mauck III, Shiwei Zheng, Andrew Butler, Maddie J. Lee, Aaron J. Wilk, Charlotte Darby, Michael Zager, Paul Hoffman, Marlon Stoeckius, Efthymia Papalexi, Eleni P. Mimitou, Jaison Jain, Avi Srivastava, Tim Stuart, Lamar M. Fleming, Bertrand Yeung, Angela J. Rogers, Juliana M. McElrath, Catherine A. Blish, Raphael Gottardo, Peter Smibert, and Rahul Satija. Integrated analysis of multimodal single-cell data. *Cell*, 184(13):3573–3587.e29, 2021. doi: 10.1016/j.cell.2021.04.048. URL https://azimuth.hubmapconsortium.org/references/human_pbmc/.

Oscar Franzén, Li-Ming Gan, and Johan L. M. Björkegren. PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. *Database*, 2019:baz046, 2019. doi: 10.1093/database/baz046. URL https://panglaodb.se/.

Congxue Hu, Tengyue Li, Yingqi Xu, Xinxin Zhang, Feng Li, Jing Bai, Jing Chen, Wenqi Jiang, Kaiyue Yang, Qi Ou, Xia Li, Peng Wang, and Yunpeng Zhang. CellMarker 2.0: an updated database of manually curated cell markers in human/mouse and web tools based on scRNA-seq data. *Nucleic Acids Research*, 51(D1):D870–D876, 2023. doi: 10.1093/nar/gkac947.

Chan Zuckerberg Initiative. CZ CELLxGENE Cell Guide. URL https://cellxgene.cziscience.com/cellguide. Cell-type descriptions and marker genes.

Allen Institute for Brain Science. Allen brain cell atlas: Human brain taxonomy, 2023. URL https://celltypes.brain-map.org/atlas?dataset=human. Human brain cell-type taxonomy and marker-gene resource.

Kimberly Siletti, Rebecca Hodge, Alejandro Mossi Albiach, Ka Wai Lee, Song-Lin Ding, Lijuan Hu, Peter Lönnerberg, Trygve Bakken, Tamara Casper, Michael Clark, Nick Dee, Jessica Gloe, Daniel Hirschstein, Nadiya V. Shapovalova, C. Dirk Keene, Julie Nyhus, Herman Tung, Anna Marie Yanny, Ernest Arenas, Ed S. Lein, and Sten Linnarsson. Transcriptomic diversity of cell types across the adult human brain. *Science*, 382(6667):eadd7046, 2023. doi: 10.1126/science.add7046.

Thomas M. Norman, Max A. Horlbeck, Joseph M. Replogle, Alex Y. Ge, Albert Xu, Marco Jost, Luke A. Gilbert, and Jonathan S. Weissman. Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. *Science*, 365(6455):786–793, 2019. doi: 10.1126/science.aax4438. URL https://www.science.org/doi/abs/10.1126/science.aax4438.

Grégoire Montavon, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. Layer-wise relevance propagation: An overview. In Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller (eds.), *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, volume 11700 of *Lecture Notes in Computer Science*, pp. 193–209. Springer International Publishing, Cham, 2019. doi: 10.1007/978-3-030-28954-6_10.

Reduan Achtibat, Sayed Mohammad Vakilzadeh Hatefi, Maximilian Dreyer, Aakriti Jain, Thomas Wiegand, Sebastian Lapuschkin, and Wojciech Samek. AttnLRP: Attention-aware layer-wise relevance propagation for transformers. In *International Conference on Machine Learning (ICML)*, 2024.

Ameen Ali, Thomas Schnake, Oliver Eberle, Grégoire Montavon, Klaus-Robert Müller, and Lior Wolf. XAI for transformers: Better explanations through conservative propagation. In *International Conference on Machine Learning (ICML)*, 2022.

Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. Evaluating the visualization of what a deep neural network has learned. *IEEE Transactions on Neural Networks and Learning Systems*, 28(11):2660–2673, 2017.

Stefan Blücher, Johanna Vielhaben, and Nils Strodthoff. Decoupling pixel flipping and occlusion strategy for consistent XAI benchmarks. *arXiv:2401.06654*, 2024.

Farnoush Rezaei Jafari, Grégoire Montavon, Klaus-Robert Müller, and Oliver Eberle. MambaLRP: Explaining selective state space sequence models. In *Advances in neural information processing systems*, 2024.

Leila Arras, José Arjona-Medina, Michael Widrich, Grégoire Montavon, Michael Gillhofer, Klaus-Robert Müller, Sepp Hochreiter, and Wojciech Samek. Explaining and interpreting LSTMs. *Explainable AI: Interpreting, explaining and visualizing deep learning*, pp. 211–238, 2019.

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

A Dataset Description

A.1 Blood PBMC Immune-Cell Benchmark

The blood benchmark contains 61,350 cells and 24,071 genes from 17 source datasets and 106 donors. Each cell has two labels, a coarse immune cell type and a fine-grained cell subtype. The final hierarchy contains 4 cell types and 22 subtypes.

Extraction and pre-processing. We constructed the blood benchmark from CZ CELLxGENE Census (CZI Cell Science Program et al., 2025), release `2025-11-08`. The goal was to obtain a clean PBMC-like immune-cell dataset in which models could be evaluated at both cell type and subtype resolution. To reduce technical and biological heterogeneity, we restricted the source pool to human primary cells from normal blood generated with the same 10x Genomics 3′ v3 assay. We also required each candidate cell to have at least 500 detected genes in the Census metadata.

From this source pool, we retained only cells with informative immune-cell annotations. Broad labels that did not specify a subtype were removed, and the remaining labels were mapped into a two-level hierarchy: 4 major immune cell types and 22 nested subtypes. We then applied standard cell-level quality control to raw counts, removing cells with high mitochondrial fraction, predicted doublet status, or unusually low or high numbers of detected genes. Finally, we sub-sampled the filtered cells to make the subtype distribution as balanced as possible, targeting approximately 3,000 cells per subtype. The extracted cell types and subtypes can be seen in Table A.1.

Table A.1: Hierarchy of cell types and cell subtypes in the dataset.

Cell Type	Cell Subtype
T cell	Naive $CD4^+$ T, Central memory $CD4^+$ T, Effector memory $CD4^+$ T, $CD4^+$ helper T, Cytotoxic $CD4^+$ T, Naive $CD8^+$ T, Central memory $CD8^+$ T, Effector memory $CD8^+$ T, Effector $CD8^+$ T, MAIT, $\gamma\delta$ T, Memory Treg, Naive Treg
B cell	Naive B, Memory B, Transitional B, Plasma cell
NK cell	$CD56^{dim}$ NK, $CD56^{bright}$ NK
Monocyte	Classical, Intermediate, Non-classical

Marker-gene extraction. For the blood benchmark, marker genes were curated independently of model predictions and used only as biological reference panels for downstream interpretability analyses. Candidate markers were collected from five external resources: AIFI (Gustafson et al.; Gong et al., 2025), Azimuth PBMC markers (Hao et al., 2021), PanglaoDB (Franzén et al., 2019), CellMarker 2.0 (Hu et al., 2023), and the CELLxGENE Cell Guide (Chan Zuckerberg Initiative). The candidate list was then evaluated directly on the final benchmark rather than accepted purely from the external databases.

Each candidate gene was assessed using expression patterns in the curated blood dataset, including linear counts-per-10k expression, the fraction of cells with non-zero expression, tau specificity, and one-vs-rest Wilcoxon differential-expression support. This produced two marker panels: a cell-type-level panel for the four coarse immune cell types and a subtype-level panel for the 22 fine-grained immune populations.

Shared cell subtype marker genes. Table A.2 lists the four pairs of T cell subtypes and their shared markers, as used in the context-sensitivity analysis of Section 3.3.

A.2 Brain Broad-Class Benchmark

This benchmark contains 61,250 nuclei and 38,019 genes from 102 source datasets and 623 donors. It is balanced across 7 broad classes, with 8,750 nuclei per class.

Extraction and pre-processing. We constructed the brain benchmark from the same CZ CELLxGENE Census (CZI Cell Science Program et al., 2025) release, `2025-11-08`. The goal was to

Table A.2: Blood cell subtype pairs and shared marker genes used in the context analysis.

T cell subtypes		Shared markers
CM CD4	CM CD8	*CCR7*, *IL7R*
EM CD4	EM CD8	*CCL5*, *GZMA*
CM CD4	EM CD4	*CD4*, *IL7R*
CM CD8	EM CD8	*CD8A*, *CD8B*

obtain a clean adult human brain single-nucleus dataset for broad cell-class evaluation. To reduce technical and biological heterogeneity, we restricted the source pool to human primary brain nuclei from exact-normal samples generated with the 10x Genomics 3′ v3 assay. Because the benchmark is intended to represent adult/postnatal brain, prenatal or fetal samples were excluded before class mapping.

We then mapped the remaining nuclei into seven broad classes: excitatory neurons, inhibitory neurons, astrocytes, oligodendrocytes, oligodendrocyte precursor cells, microglia/macrophages, and vascular/perivascular cells. Standard cell-level quality control was applied to raw counts, removing nuclei with insufficient detected genes, unusually high detected-gene counts within their class, or high mitochondrial fraction. The final dataset was exactly balanced across classes, with 8,750 nuclei per class and 61,250 nuclei in total. Because the available CELLxGENE brain data are not uniformly sampled across regions, the benchmark should be interpreted as cortex-weighted rather than as a whole-brain atlas.

Marker-gene extraction. For the brain benchmark, marker genes were curated independently of model predictions and used as biological reference panels for interpretability analyses. Candidate markers were collected from Allen Brain Atlas canonical markers (Allen Institute for Brain Science, 2023; Siletti et al., 2023), CELLxGENE Cell Guide markers (Chan Zuckerberg Initiative), and literature-derived canonical brain markers.

Each candidate gene present in the final dataset was evaluated using its class-level expression profile, including mean linear counts-per-10k expression, detection fraction, tau specificity, one-vs-rest Wilcoxon differential-expression rank, source provenance, primary brain-class assignment, and trust tier. Primary-class assignment was based on a robust Wilcoxon-rank rule rather than raw mean expression alone. The final marker panel contains 55 genes across the seven broad brain classes.

A.3 Synthetic Dataset

For the perturbation prediction experiment, we generate a synthetic dataset based on Qi & Chapfuwa (2026); Lorch et al. (2026). Each cell in this dataset is represented by the expression of $G = 128$ genes. The data is generated from a stochastic linear dynamical system whose interactions are defined by an explicit gene regulatory network (GRN). For a cellular context c, the latent expression state $\mathbf{x} \in \mathbb{R}^G$ evolves as

$$\mathrm{d}\mathbf{x}(t) = (\mathbf{A}_c\mathbf{x}(t) + \mathbf{b}_c + \boldsymbol{\gamma}_q)\,\mathrm{d}t + \sigma\,\mathrm{d}\mathbf{w}(t), \tag{2}$$

where $\mathbf{A}_c$ denotes the context-specific GRN, $\mathbf{b}_c$ the context-specific baseline, and $\mathbf{w}(t)$ is a standard Wiener process. Perturbation q is encoded as a sparse vector $\boldsymbol{\gamma}_q$, with non-zero entries reflecting the sign and strength of the intervention on the target genes. Also following Qi & Chapfuwa (2026), the observed expression value for each gene g is sampled from a negative binomial likelihood, with mean $\mathrm{softplus}(x_g)$ and dispersion estimated from the Norman et al. (2019) dataset.

We generate two cellular contexts over the same 128 genes. The first context 0 is defined by a sparse signed regulatory network, where each gene selects 10 regulators uniformly (without replacement, excluding self), and assigns each edge a random sign and magnitude uniformly between 1 and 3. The second is obtained from the first one by rewiring 50% of the regulatory edges in each row and modifying the context-specific baseline. The two contexts therefore share part of their regulatory structure while differing in the causal mechanisms governing gene expression. For control cells, $\boldsymbol{\gamma}_q = \mathbf{0}$. For perturbed cells, the intervention directly targets one of the 128 genes and its effect propagates through the corresponding GRN. The raw expression values are subsequently normalized to a library size of 10,000 and log-transformed.

We generate 8M unperturbed control cells for pretraining and, for fine-tuning, an additional 5M controls and 1M perturbed cells across both contexts. The perturbed set contains 8,192 independently simulated cells for each of 128 randomly signed single-gene perturbations, covering all genes. Controls and perturbed cells are sampled independently from their respective stationary populations and thus have no one-to-one correspondence. Perturbations are instantiated with random sign and random magnitude in $[2 \log G, 6 \log G] = [9.7, 29.1]$, bounded away from zero to prevent undetectable effects. Both datasets are divided into 80%, 10%, and 10% training, validation, and test splits, respectively, stratified by context and perturbation.

B CELL-TYPE CLASSIFICATION PROBES

Single-cell foundation models are designed to encode cellular information in their representations, but it remains unclear where cell identity becomes linearly accessible across the network and how this depends on gene-level representation aggregation and cellular context. Understanding this is important both for characterizing the learned representations and for selecting suitable prediction targets for our attribution analyses.

Training details. We train linear probes on frozen representations extracted at approximately 25%, 50%, 75%, and 100% of the network and evaluate them on the same held-out cells. This tests whether cell identity emerges progressively or is already linearly accessible in early layers. For C2S, we construct cell-level representations using either mean pooling over non-padding tokens or the final-token representation, and for scGPT we analogously compare mean pooling with the `[CLS]` token representation. Each blood probe is a single linear layer predicting either four cell types or 22 cell subtypes, while brain probes predict one of seven cell types.

Probes are trained with AdamW and class-balanced cross-entropy. We use a separate validation split for early stopping and model selection, evaluating the test split only after selecting the final probe. The learning rate is linearly warmed up over the first 10% of training and then follows cosine annealing. To reduce sensitivity to optimization choices, we perform a systematic grid search for each model, representation, and prediction task, evaluating every configuration across three random seeds and selecting the final probe based on validation macro-F1. The search spaces are reported in Tables B.1 and B.2.

Table B.1: Shared training and hyper-parameter search settings for linear probes.

Setting	Value
Optimizer	AdamW
Objective	Class-balanced cross-entropy
Selection metric	Macro-F1
Learning-rate schedule	10% linear warm-up, then cosine annealing
Weight decay	$10^{-5}, 10^{-4}, 10^{-3}$
Dropout	0.0, 0.1
Standardization	None
Random seeds	0, 1, 2

Table B.2: Model-specific probe hyper-parameter search configurations.

Model	Layers	Learning rate	Batch size	Epochs	Patience
Pythia 410M	6, 12, 18, 24	$3\times10^{-4}, 10^{-3}, 3\times10^{-3}$	64	200	15
Pythia 1B	4, 8, 12, 16	$10^{-4}, 3\times10^{-4}, 10^{-3}$	128	100	10
Gemma2 2B	6, 13, 20, 26	$10^{-4}, 3\times10^{-4}, 10^{-3}$	256	60	10
scGPT	3, 6, 9, 12	$10^{-4}, 3\times10^{-4}, 10^{-3}$	64	30	7

Selecting probes for attribution. For attribution analyses, we use probes trained on final-layer representations and select the pooling strategy separately for each model configuration and predic-

Table B.3: Pooling strategies selected for the final-layer probes used in attribution analyses. Each strategy is selected separately for each model, context length, and prediction task based on validation performance. A dash denotes an unevaluated combination. Context length K applies only to C2S.

Model	K	Blood type	Blood subtype	Brain type
C2S Pythia 410M	200	`last_token`	`last_token`	`last_token`
C2S Pythia 1B	200	`last_token`	`mean_non_pad`	`last_token`
C2S Pythia 1B	1000	`last_token`	`last_token`	`last_token`
C2S Gemma2 2B	200	`last_token`	`last_token`	`mean_non_pad`
C2S Gemma2 2B	1000	`mean_non_pad`	`last_token`	`mean_non_pad`
scGPT blood	—	`cls`	`cls`	–
scGPT brain	—	–	–	`cls`
scGPT whole-human	—	`cls`	`cls`	`cls`

tion task based on validation accuracy. The selected pooling strategies are reported in Table B.3. LRP is then applied to the corresponding probe logits.

Results. Figure B.1 shows that, as expected, fine-grained cell-subtype classification is consistently more challenging than coarse-grained cell-type classification across models and layers. Cell identity also appears initially distributed across token positions, with mean-pooled representations generally providing more accessible information than single-token representations in early and intermediate layers. Toward the final layers, however, cell identity becomes increasingly concentrated in a designated summary token, namely the last token in autoregressive C2S and the `[CLS]` token in scGPT. Finally, increasing the cellular context from $K = 200$ to $K = 1000$ genes also generally improves probe accuracy, suggesting that broader cellular context yields richer representations of cell identity.

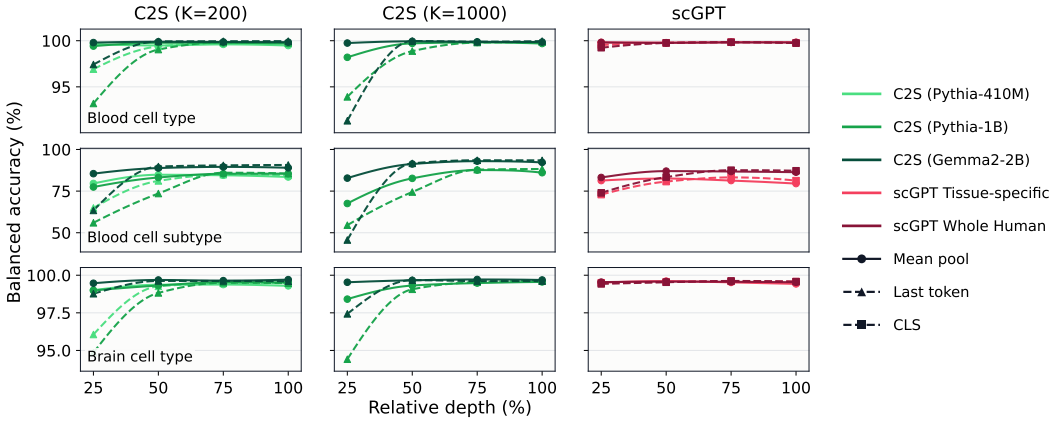

Figure B.1: Probe accuracies across model depths and pooling strategies. Broad cell types are predicted more accurately than fine-grained cell subtypes. Single-token representations become increasingly predictive toward the final layers, while longer C2S input contexts generally improve accuracy. Here, K denotes the number of input genes, and `[CLS]` denotes the classification token in scGPT.

C Layer-wise Relevance Propagation (LRP) Details

Layer-wise Relevance Propagation (LRP; Bach et al., 2015) is a framework for interpreting neural network predictions by attributing the model's output, or a quantity derived from it, to the input features. LRP assigns a relevance score $\mathcal{R}$ to each neuron that reflects its contribution to the quantity of interest and propagates these scores backward through the network using layer-specific rules. The propagation typically follows the principle of relevance conservation (Bach et al., 2015; Montavon et al., 2019), which requires the total relevance to be preserved as it is redistributed from one layer

Table C.1: Evaluating explanation methods. Larger values indicate greater attribution faithfulness.

Method	C2S (Pythia-410M)	C2S (Gemma2-2B)	scGPT Blood
LRP (AH-rule)	1.198	**1.085**	4.948
LRP (AttnLRP-rule)	**1.283**	0.945	**5.140**
Gradient × Input	-0.600	-0.121	0.268
Integrated Gradients	1.160	0.068	-1.576
Attention Rollout	0.017	-0.002	0.885
Random	0.013	0.005	-0.008

to the preceding layer. This process ultimately yields an attribution of the quantity of interest to the input features.

LRP is model-specific, as the relevance propagation rules must be adapted to the operations in specific layers. In this study, we use LRP to interpret transformer-based architectures and therefore employ LRP rules that are applicable to transformer models (Achtibat et al., 2024; Ali et al., 2022). To determine whether LRP provides faithful explanations for our models, we compare it with Integrated Gradients, Gradient × Input, attention rollout, and a random baseline. We evaluate the faithfulness of these methods using an input perturbation approach that ranks input features according to their attribution scores (Samek et al., 2017). The ranked features are then iteratively perturbed according to either a *Most Relevant First* (MoRF) or *Least Relevant First* (LeRF) strategy, while tracking the output logit f_c of the predicted class c. We denote the areas under the resulting perturbation curves by A^F_{MoRF} and A^F_{LeRF}, respectively.

A faithful explanation should yield a low A^F_{MoRF}, as perturbing the most relevant features is expected to strongly affect the model output. In contrast, A^F_{LeRF} should be high, as perturbing less relevant features should have a smaller effect. Following Blücher et al. (2024); Achtibat et al. (2024); Rezaei Jafari et al. (2024), we combine these two measures to reduce the influence of out-of-distribution perturbations and define $\Delta A^F = A^F_{\text{LeRF}} - A^F_{\text{MoRF}}$. A higher ΔA^F indicates a larger separation between the effects of relevant and less relevant features and therefore a more faithful explanation. We use this metric to compare the different explanation methods and select the method that most faithfully captures the behavior of the models in our case studies. We then use the same evaluation to select the most faithful combination of LRP propagation rules for each architecture.

Cell typing LRP. The results of this analysis are reported in Table C.1. LRP achieves the highest faithfulness across all evaluated architectures, outperforming the other explanation methods. We therefore use LRP as our primary attribution method. The propagation rules used for each model are reported in Table C.2.

Table C.2: Propagation rules used in our LRP implementation for the three single-cell architectures considered in this study. A green tick (✓) indicates that the rule was applied, while a red cross (×) indicates that it was not used.

Propagation Rule	**C2S (Gemma2)**	**C2S (Pythia)**	**scGPT**
LN-rule (Ali et al., 2022)	✓	✓	✓
Identity-rule (Rezaei Jafari et al., 2024)	✓	✓	×
ϵ-rule (Montavon et al., 2019)	✓	✓	×
Half-rule (Rezaei Jafari et al., 2024; Arras et al., 2019)	✓	×	×
AH-rule (Ali et al., 2022)	✓	×	×
AttnLRP (Achtibat et al., 2024)	×	✓	✓

Perturbation prediction LRP. For the joint scGPT+STATE perturbation model, we used the LN-, identity-, and AttnLRP rules. Faithfulness was evaluated by masking genes ranked using relevance from independent cell sets. Most-relevant-first masking produced larger prediction-shift AUCs than least-relevant-first masking ($\Delta A = 0.491$ for same-cell and 0.508 for all-cell relevance). We additionally ablated the model's perturbation channel: condition-level prediction Pearson-Δ dropped

from 0.506 to 0.035, and perturbation-source LRP agreed with the resulting output changes (absolute Spearman $\rho = 0.814$). Replacing AttnLRP with the AH-rule in STATE produced similar rankings ($\rho = 0.999$ for same-cell and 0.946 for all-cell relevance).

D SYMBOLIC XAI DETAILS

Symbolic XAI (Schnake et al., 2025) extends conventional feature-wise attribution to explanations defined over interactions and logical relationships between input features. Let $\mathcal{N}$ denote the set of input features. The framework first represents the model prediction using a multi-order decomposition

$$f(X) = \sum_{L \subseteq \mathcal{N}} \mu_L, \tag{3}$$

where μ_L denotes the contribution associated with the subset of features L. The decomposition therefore contains both first-order terms, such as $\mu_{\{i\}}$, and higher-order terms, such as $\mu_{\{i,j\}}$, which capture contributions involving multiple input features.

In the propagation-based formulation used in this work, the multi-order terms are obtained through higher-order relevance propagation. Rather than assigning relevance only to individual input features, relevance is propagated along ordered sequences of interacting features, referred to as walks. If R_W denotes the relevance associated with a walk W, walks involving the same set of input features are aggregated as

$$\mu_L = \sum_{W\colon \mathrm{set}(W) = L} R_W. \tag{4}$$

This produces relevance terms associated with individual features as well as their higher-order interactions while preserving an additive decomposition of the model prediction.

Symbolic queries are subsequently used to select and aggregate relevant terms from this decomposition. For a query q, the framework defines a binary function $\lambda_L(q)$ indicating whether the feature subset L satisfies the logical condition expressed by q. The relevance associated with the query is

$$A(q) = \sum_{L \subseteq \mathcal{N}} \eta_L \, \mu_L \, \lambda_L(q), \tag{5}$$

where η_L controls the weighting of the individual multi-order terms. A symbolic query can therefore be interpreted as a Boolean filter over the higher-order relevance decomposition, followed by aggregation of the selected terms.

In our analysis, let S denote the set of marker genes shared between two related cell subtypes, and let $O = \mathcal{N} \backslash S$ denote all other input genes in the cellular context. We consider two symbolic queries that differ in whether interactions between the shared marker genes and the remaining cellular context are included.

The first query,

$$q_{\text{full}} = S, \tag{6}$$

selects every relevance term involving at least one shared marker gene,

$$\lambda_L(S) = \mathbb{1}[L \cap S \neq \emptyset]. \tag{7}$$

Thus, q_{full} includes both relevance terms formed entirely within S and higher-order terms that involve shared marker genes together with genes from O. Its relevance score therefore measures the contribution of the shared marker genes in the full cellular context.

The second query,

$$q_{\text{minimal}} = S \wedge \neg O, \tag{8}$$

additionally requires that no gene outside the shared marker set participates in the relevance term. Specifically,

$$\lambda_L(S \wedge \neg O) = \mathbb{1}[L \cap S \neq \emptyset]\mathbb{1}[L \cap O = \emptyset]. \tag{9}$$

Since $O = \mathcal{N} \setminus S$, this query retains only relevance terms composed exclusively of genes in S. Hence, q_{minimal} provides a context-isolated measure of the relevance of the shared marker genes, excluding their interactions with all other genes in the cell.

Comparing q_{full} and q_{minimal} therefore allows us to distinguish relevance arising from interactions within the shared marker set from relevance that depends on interactions between the shared markers and the broader cellular context.

E scGPT Training Details

E.1 Pre-training

We pretrain the scGPT backbone from scratch on the synthetic control-cell dataset described in Appendix A.3.

Architecture. We use the original scGPT backbone. Each input cell contains expression values for 128 genes together with a `[CLS]` token, resulting in a sequence length of 129. Gene-expression values are represented as continuous normalized-log1p values, and genes with zero expression are retained explicitly. The model consists of 12 Transformer layers with an embedding dimension of 512, 8 attention heads, a feed-forward dimension of 512, and a dropout rate of 0.1. The complete pretraining model contains approximately 20.3M trainable parameters.

Masked expression prediction. Pretraining follows the masked expression-prediction setup of scGPT (Cui et al., 2024). For each batch, we sample a masking ratio uniformly from $\{0.25, 0.50, 0.75\}$, corresponding to 32, 64, or 96 masked genes, respectively. Masked genes are sampled uniformly without replacement, and the model is trained to reconstruct their continuous normalized-log1p expression values from the remaining genes.

Reconstruction objective. The standard reconstruction loss is the mean-squared error over the masked genes:

$$\mathcal{L}_{\text{micro}} = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} (\hat{y}_i - y_i)^2, \tag{10}$$

where $\mathcal{M}$ denotes the set of masked genes, y_i is the target expression value of gene i, and $\hat{y}_i$ is the corresponding prediction.

Since training examples are unevenly distributed across genes and expression ranges, we additionally use a balanced reconstruction loss, $\mathcal{L}_{\text{balanced}}$, that first balances reconstruction errors across genes within each expression stratum and then averages the resulting stratum-level losses. The primary reconstruction objective is

$$\mathcal{L}_{\text{primary}} = \alpha \mathcal{L}_{\text{micro}} + (1 - \alpha)\mathcal{L}_{\text{balanced}}, \tag{11}$$

with $\alpha = 0.75$.

We further use the GEPC objective of scGPT as an auxiliary training signal. The GEPC decoder predicts masked gene-expression values from the learned cell representation. The total pretraining objective is

$$\mathcal{L}_{\text{pre}} = \mathcal{L}_{\text{primary}} + \beta \mathcal{L}_{\text{GEPC}}, \tag{12}$$

with $\beta = 0.5$.

Model selection is based only on $\mathcal{L}_{\text{primary}}$, excluding the auxiliary GEPC loss.

Optimization. We optimize the model using AdamW with a learning rate of 10^{-4} and weight decay of 0.01. The batch size is 32, with gradients accumulated over four batches to obtain an effective batch size of 128. Training runs for 100K optimization steps. The learning rate is linearly warmed up during the first 5K steps and subsequently follows a cosine decay schedule, reaching

10% of its initial value at the end of training. Gradients are clipped to a maximum norm of 1.0, and mixed-precision training is enabled.

Validation is performed every 5K optimization steps using all three masking ratios. Early stopping is enabled after 50K steps with a patience of six validation evaluations, although training completes all 100K steps. The model with the best validation performance under $\mathcal{L}_{\text{primary}}$ is used to initialize fine-tuning. Test evaluation is performed only after training.

E.2 Fine-tuning

For perturbation prediction, we augment the pretrained scGPT backbone with the State Transition architecture of STATE (Adduri et al., 2025) and jointly train the resulting model on the synthetic control–perturbed data.

Architecture. Each control-cell expression profile is processed by the scGPT gene encoder, continuous value encoder, and 12-layer Transformer backbone to obtain a `[CLS]` cell representation. This representation is projected through a two-layer basal encoder and combined with a separately projected one-hot perturbation embedding. The resulting cell representations are processed jointly by a bidirectional, non-causal STATE set Transformer. We omit the batch encoding, as the dataset contains homogeneous cell populations for each condition.

The STATE module has hidden dimension 64, two Transformer layers, four attention heads, and dropout of 0.1. Both the basal-state and perturbation encoders contain two layers, followed by a two-layer output decoder. We additionally use a learned context token together with an auxiliary context classifier. The decoder predicts post-perturbation expression values for the same 128 genes, with a ReLU activation applied to the output.

Fine-tuning objective. Since predicted and observed perturbed cells do not have a one-to-one correspondence, we follow STATE Adduri et al. (2025) and optimize the model at the population level using an Energy distance between the predicted and observed perturbed cell sets. The objective additionally regularizes the predicted change relative to the control population and includes an auxiliary context-classification loss:

$$\mathcal{L}_{\text{post}} = \mathcal{L}_{\text{Energy}} + \lambda_{\Delta}\mathcal{L}_{\Delta} + \lambda_{\text{context}}\mathcal{L}_{\text{context}}. \tag{13}$$

Here, $\mathcal{L}_{\text{context}}$ denotes the cross-entropy loss of the auxiliary context classifier. We set

$$\lambda_{\Delta} = 10^{-5} \quad \text{and} \quad \lambda_{\text{context}} = 0.1.$$

The Energy distance is evaluated over four randomized gene chunks with a blur parameter of 0.05.

Optimization. Fine-tuning proceeds in two stages. During the first five epochs, the pretrained scGPT backbone is frozen and only the STATE parameters are optimized. We then unfreeze the complete scGPT gene encoder, value encoder, and Transformer backbone and jointly optimize scGPT and STATE for up to 15 additional epochs.

STATE parameters are optimized with a learning rate of 10^{-4}, while the pretrained scGPT parameters use a smaller learning rate of 10^{-6}. We use the Adam optimizer. Gradients are clipped to a maximum norm of 1.0. Each training example comprises a set of 64 cells, and each training step processes 2 sets. After unfreezing the scGPT backbone, early stopping with a patience of five validation evaluations is applied, and the model with the best validation performance is used for final evaluation.

F Perturbation prediction metrics

We evaluate the predictive performance of our perturbation model using metrics from Viñas Torné et al. (2026). For a fixed cellular context, let G denote the number of genes and let $\mathcal{P}$ denote the set of evaluated perturbations. For perturbation $p \in \mathcal{P}$, we define

- $\boldsymbol{\mu}_p \in \mathbb{R}^G$: the observed mean post-perturbation expression profile,
- $\widehat{\boldsymbol{\mu}}_p \in \mathbb{R}^G$: the predicted mean post-perturbation expression profile,
- $\boldsymbol{\mu}_0 \in \mathbb{R}^G$: the observed control centroid.

All metrics are calculated separately for each perturbation and context and are then averaged across the evaluated conditions.

Perturbed reference. To reduce the influence of a generic control-to-perturbed expression shift, we use the equal-weighted mean of the training perturbation centroids as reference:

$$\mathbf{r} = \frac{1}{|\mathcal{P}_{\text{train}}|} \sum_{q \in \mathcal{P}_{\text{train}}} \boldsymbol{\mu}_q^{\text{train}}. \tag{14}$$

The observed and predicted perturbation-specific shifts are therefore

$$\boldsymbol{\Delta}_p = \boldsymbol{\mu}_p - \mathbf{r}, \qquad \widehat{\boldsymbol{\Delta}}_p = \widehat{\boldsymbol{\mu}}_p - \mathbf{r}. \tag{15}$$

Perturbed-reference Pearson Δ. Pearson Δ measures whether the predicted and observed response vectors have the same relative pattern across genes:

$$\text{Pearson}\Delta_p = \frac{\sum_{g=1}^{G}(\Delta_{p,g} - \overline{\Delta}_{p,g})(\widehat{\Delta}_{p,g} - \overline{\widehat{\Delta}}_{p,g})}{\sqrt{\sum_{g=1}^{G}(\Delta_{p,g} - \overline{\Delta}_{p,g})^2}\sqrt{\sum_{g=1}^{G}(\widehat{\Delta}_{p,g} - \overline{\widehat{\Delta}}_{p,g})^2}}. \tag{16}$$

It ranges from -1 to 1, with larger values indicating better recovery of the direction and relative structure of the perturbation response. Because correlation is invariant to rescaling, it does not assess whether the predicted magnitudes are correct.

Perturbed-reference Pearson Δ_{20}. For each perturbation, let $\mathcal{D}_p$ be the 20 genes with the largest observed absolute changes from control:

$$\mathcal{D}_p = \text{Top20}_g \left|\mu_{p,g} - \mu_{0,g}\right|. \tag{17}$$

Pearson Δ_{20} is the same Pearson correlation restricted to $g \in \mathcal{D}_p$:

$$\text{Pearson}\Delta_{20,p} = \text{corr}\left(\boldsymbol{\Delta}_{p,\mathcal{D}_p}, \widehat{\boldsymbol{\Delta}}_{p,\mathcal{D}_p}\right). \tag{18}$$

This metric focuses on the strongest and most biologically prominent expression changes, which can otherwise be diluted by many weakly affected genes.

Root mean-squared error. RMSE measures the absolute numerical discrepancy between predicted and observed post-perturbation expression:

$$\text{RMSE}_p = \sqrt{\frac{1}{G}\sum_{g=1}^{G}(\widehat{\mu}_{p,g} - \mu_{p,g})^2}. \tag{19}$$

The corresponding top-20 metric is

$$\text{RMSE}_{20,p} = \sqrt{\frac{1}{|\mathcal{D}_p|}\sum_{g \in \mathcal{D}_p}(\widehat{\mu}_{p,g} - \mu_{p,g})^2}. \tag{20}$$

Lower values are better. Unlike Pearson correlation, RMSE is sensitive to the magnitude of the prediction and is independent of the choice of expression reference. RMSE therefore prevents a model from receiving a high score merely for predicting the correct direction with an incorrect response magnitude.

Centroid accuracy. Centroid accuracy evaluates whether a predicted profile is closer to its correct observed perturbation centroid than to the centroids of other perturbations. Using Euclidean distance, it is defined as

$$\mathrm{CA}_p = \frac{1}{|\mathcal{P}| - 1} \sum_{q \in \mathcal{P},\, q \neq p} \mathbb{I}\left[\|\widehat{\boldsymbol{\mu}}_p - \boldsymbol{\mu}_p\|_2 < \|\widehat{\boldsymbol{\mu}}_p - \boldsymbol{\mu}_q\|_2\right], \tag{21}$$

where $\mathbb{I}[\cdot]$ is the indicator function. A score of 1 means that the correct perturbation centroid is closer than every competing centroid, whereas a score near 0.5 corresponds to approximately random ranking. This metric assesses whether the model reconstructs the geometry and identity of the perturbation landscape.

Systematic variation. Let the true control-referenced shift of perturbation p be

$$\mathbf{s}_p = \boldsymbol{\mu}_p - \boldsymbol{\mu}_0, \tag{22}$$

and let the average perturbation shift be

$$\bar{\mathbf{s}} = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \mathbf{s}_p. \tag{23}$$

Systematic variation is the mean cosine similarity between each perturbation-specific shift and the average shift:

$$\mathrm{SV} = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \frac{\mathbf{s}_p^{\mathsf{T}} \bar{\mathbf{s}}}{\|\mathbf{s}_p\|_2 \, \|\bar{\mathbf{s}}\|_2}. \tag{24}$$

A high value indicates that many perturbations share one generic response direction, which can inflate conventional correlation scores. A value near zero indicates more heterogeneous, perturbation-specific responses. Systematic variation is a dataset diagnostic rather than a direct model-performance metric.

Perturbed-mean baseline. We compare the model with a non-parametric baseline that predicts the same cell-weighted mean training perturbation profile for every perturbation:

$$\mathbf{b} = \frac{1}{N_{\mathrm{pert}}^{\mathrm{train}}} \sum_{i \in \text{perturbed train cells}} \mathbf{x}_i, \qquad \widehat{\boldsymbol{\mu}}_p^{\mathrm{baseline}} = \mathbf{b}. \tag{25}$$

The model and baseline are evaluated with identical RMSE, RMSE_{20}, and centroid-accuracy calculations. Beating this baseline demonstrates that the model captures perturbation-specific information rather than only the average perturbed-cell state.

G Degree-preserving random GRN control.

To test whether regulatory alignment exceeds what could arise from network topology alone, we generate, for each context c, 999 randomized versions of its GRN A_c by rewiring edges while preserving each gene's in- and out-degree. Candidate signatures are kept fixed and evaluated against each randomized GRN using the same recovery metrics as for the ground-truth network. Each metric is computed per perturbation and macro-averaged across perturbations.

For each signature and metric, we compute a one-sided empirical p-value as

$$p = \frac{1 + \sum_{r=1}^{999} \mathbb{I}[v_r \geq v_{\mathrm{obs}}]}{1000}, \tag{26}$$

where v_r is the score obtained for randomized network r and v_{obs} is the score for the ground-truth GRN. Thus, a small p-value indicates that the observed alignment is rarely matched or exceeded by degree-preserving randomized networks. With 999 randomizations, the minimum attainable empirical p-value is 0.001.

H Extended results

H.1 Perturbation prediction results

The scGPT+STATE model achieves a centroid accuracy of 0.962, indicating strong preservation of perturbation identity. It also reaches a Pearson Δ_{20} of 0.704, showing good recovery of the dominant gene-response directions. Across all genes, the model obtains an RMSE of 0.855, compared with 0.606 for the baseline, while on the 20 most strongly affected genes it improves RMSE_{20} over baseline from 1.256 to 1.123. The systematic variation score of 0.147 is low, suggesting that these results are not driven solely by a shared generic perturbation response. Overall, the model recovers perturbation identity and dominant response patterns well, but exact expression values are not consistently well recovered.

H.2 Regulatory alignment

In this section, we provide complementary results for the regulatory-alignment analysis using the previously introduced baselines together with additional variants. Beyond same-cell gene LRP, *all-cell gene LRP* propagates relevance to input genes across all control cells in the STATE input set and aggregates their contributions.

Counterfactual baselines. As counterfactual baselines, we directly modify the model inputs and measure how much the predicted expression profile changes. We either mask a gene embedding or ablate the perturbation embedding, providing black-box measures of how strongly the model depends on each input. Gene-level counterfactuals are reduced in a similar manner to gene-LRP signatures, by keeping only the intervention on input gene i_q.

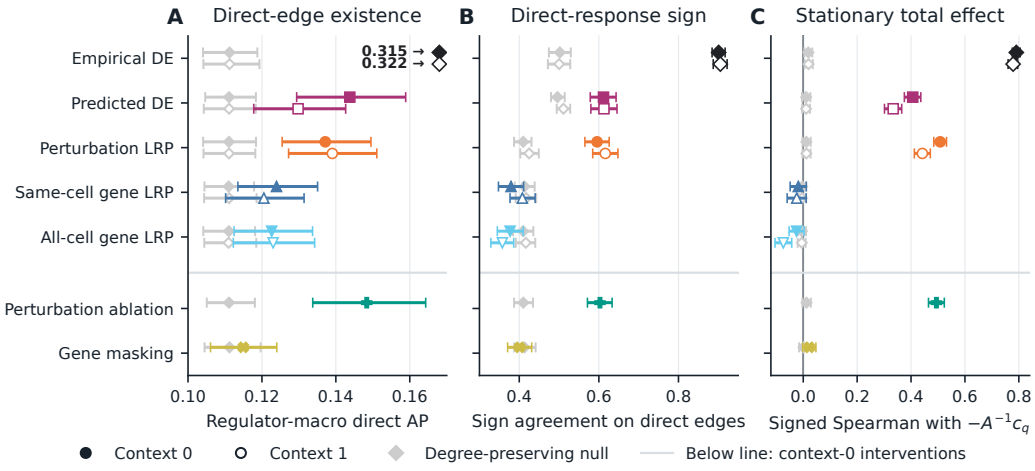

Figure H.1: Evaluation of regulatory alignment in both in-distribution contexts (0 and 1). **(A)** Binary recovery of direct edge existence, measured as average precision (AP). Empirical DE is clipped for better visualisation. **(B)** Sign agreement with nonzero ground-truth direct responses. **(C)** Spearman correlation with stationary total effects. Error bars for the true measurements reflect 95% bootstrap CIs over the 128 perturbations. The grey control error bars are the central 95% range of the perturbation-average evaluated against the randomised-GRN controls. This is an extension of Fig. 5, including all-cell gene LRP as well as interventional signatures: perturbation ablation and gene masking.

The extended analysis in Fig. H.1 largely confirms the main findings. Same-cell and all-cell gene LRP behave similarly, showing modest direct-edge recovery but little alignment with response signs or stationary total effects, indicating that aggregating relevance across all control cells provides no clear advantage. In contrast, perturbation ablation closely tracks perturbation LRP across all three metrics and achieves strong model-derived regulatory alignment, supporting the relevance assigned to the explicit perturbation input as a meaningful measure of model dependence. Gene masking

remains close to the degree-preserving null, with weak direct-edge recovery and little sign or total-effect alignment. These trends are consistent across the two evaluated contexts where both are available.