Attention in Single-Cell Foundation Models Primarily Reflects Co-Expression: A Diagnostic Framework with Null-Model Benchmarking

Ihor Kendiukhov
Department of Computer Science
University of Tübingen
Tübingen, Germany
kendiukhov@gmail.com

Abstract

We present a diagnostic framework for mechanistic interpretability of single-cell foundation models, organized around three findings. **First**, pooled attention edges are dominated by co-expression: across scGPT and Geneformer, attention-derived edge scores correlate with gene co-occurrence ($\rho = 0.31$–0.42) but not regulatory ground truth ($\rho \approx 0$), explaining near-random GRN recovery (AUROC ≈ 0.5). **Second**, layer-stratified analysis via Cell-State Stratified Interpretability (CSSI) reveals where benchmark-discriminative signal concentrates: in scGPT-18L, later layers achieve AUROC 0.69–0.71 against TRRUST, with up to $1.85\times$ improvement in synthetic settings; however, CSSI is best understood as a diagnostic for signal localization rather than proof of regulatory structure. **Third**, cross-tissue evaluation across brain, kidney, and whole-human data shows uniform per-layer AUROC ($\sim$0.72 TRRUST, $\sim$0.84 DoRothEA; permutation $p_{\text{raw}} < 10^{-4}$), but simple gene-level baselines (expression and variance products) match or exceed this performance, demonstrating that benchmark AUROC reflects gene-level prominence rather than learned regulatory structure. Nine supporting analyses provide quality control spanning scaling behavior, cross-species transfer, and artifact assessment.

1 Introduction

The emergence of transformer-based foundation models for single-cell transcriptomics represents a paradigm shift in computational biology [Cui et al., 2024, Theodoris et al., 2023, Yang et al., 2022, Hao et al., 2024]. These models—trained on millions of cells across diverse tissues and conditions—learn rich contextual representations of gene expression that have shown promise for cell type annotation, perturbation response prediction, and gene regulatory network (GRN) inference [Chen et al., 2024, Rosen et al., 2024]. A central and particularly compelling promise is *mechanistic interpretability*: the ability to extract biologically meaningful regulatory circuits directly from model internal representations, especially attention-derived edge scores. Indeed, both scGPT~[Cui et al., 2024] and Geneformer~[Theodoris et al., 2023] explicitly highlight attention-derived gene network inference as a key application, and downstream studies have adopted attention-derived edge scores as proxies for regulatory edges without rigorous validation~[Zheng et al., 2024].

This promise draws on parallel advances in mechanistic interpretability of large language models (LLMs), where researchers have identified specific computational circuits responsible for well-defined behaviors [Elhage et al., 2021, Olsson et al., 2022, Wang et al., 2022]. Techniques such as activation patching [Meng et al., 2022, Goldowsky-Dill et al., 2023], automated circuit discovery [Conmy et al., 2023], and causal mediation analysis [Vig et al., 2020, Pearl, 2001] have been adapted for biological models, with the expectation that attention heads encoding gene–gene relationships correspond to genuine regulatory interactions.

However, the translation of mechanistic interpretability from language models to biological systems faces unique challenges. Unlike the well-defined grammatical structures that language model circuits have been shown to implement [Wang et al., 2022, Nanda et al., 2023], gene regulatory relationships are context-dependent, combinatorial, and only partially observed [Pratapa et al., 2020]. Reference databases such as TRRUST [Han et al., 2018], DoRothEA [Garcia-Alonso et al., 2019], and OmniPath [Turei et al., 2021] capture only a fraction of true regulatory interactions, and this partial labeling creates persistent evaluation challenges [Dibaeinia and Sinha, 2020]. Moreover, the biological ground truth itself is condition-specific: a TF–target relationship active in one tissue or cellular state may be absent in another [Kamimoto et al., 2023, The Tabula Sapiens Consortium, 2022].

Current mechanistic interpretability practices for

single-cell models rest on several critical and largely untested assumptions. *First*, that attention patterns directly reflect causal regulatory relationships—an assumption already challenged in the NLP literature [Jain and Wallace, 2019, Serrano and Smith, 2019, Wiegreffe and Pinter, 2019, Bibal et al., 2022]. *Second*, that larger datasets consistently improve the reliability of mechanistic interpretations. *Third*, that standard single-component mediation analysis provides unbiased estimates of regulatory importance. *Fourth*, that mechanistic insights transfer reliably across biological contexts. *Fifth*, that attention-derived predictions align with experimental perturbation outcomes from CRISPR screens [Dixit et al., 2016, Adamson et al., 2016]. *Sixth*, that regulatory edges inferred in one species transfer to another via ortholog mapping. *Seventh*, that pseudotime ordering validates the directionality of inferred regulatory edges. *Eighth*, that edge scores reflect biology rather than technical artifacts such as donor or batch identity. *Ninth*, that edge scores can be interpreted as calibrated probabilities of true regulation.

We address these challenges through a systematic framework organized around three core contributions, supported by nine complementary quality-control analyses.

Core Contribution 1: Pooled attention edges are dominated by co-expression (Section 3.1). Across both scGPT and Geneformer, attention-derived edge scores correlate with gene co-occurrence ($\rho = 0.31$–0.42) but not with regulatory ground truth ($\rho \approx 0$), explaining the persistent near-random GRN recovery (AUROC ≈ 0.5) observed across architectures. This mechanistic insight reframes the problem: the failure is not in the models but in the assumption that pooled attention weights reflect regulation.

To summarize the central empirical picture: pooled across layers and heads, attention yields $\rho \approx 0$ / AUROC ≈ 0.5 with regulatory benchmarks; layer-specific scores can exceed chance, but this signal is largely explainable by gene-level prominence unless expression-matched evaluation is used.

Core Contribution 2: Layer-stratified analysis localizes benchmark-discriminative signal (Sections 3.2–3.4). We introduce Cell-State Stratified Interpretability (CSSI), a diagnostic framework that identifies where benchmark-discriminative signal concentrates across layers and cell states. In scGPT-18L, later layers (L13–L14) achieve AUROC 0.694–0.706 against TRRUST, while naive aggregation yields ∼0.54. CSSI's primary value is diagnostic—layer selection—with up to 1.85× improvement in synthetic settings. Critically, because gene-level null models match this AUROC (Section 3.3), elevated per-layer AUROC indicates benchmark enrichment for prominent genes rather than recovery of genuine regulatory structure.

Core Contribution 3: Cross-tissue evaluation and null-model benchmarking (Section 3.3). Using scGPT-12L across brain, kidney, and whole-human data, we observe consistent per-layer AUROC (∼0.72 TRRUST, ∼0.84 DoRothEA). However, simple gene-level null models—detection-rate, mean-expression, and variance products—match or exceed attention-based AUROC on the same evaluation edges, demonstrating that the bulk of benchmark AUROC is explainable by gene-level statistics without invoking learned regulatory structure. However, expression-matched negative sampling—where negatives are matched on actual gene expression statistics (mean expression and detection rate within ±20%)—reveals that attention retains AUROC 0.646 [0.539, 0.747] against chance-level baselines (0.522), providing direct evidence of genuine pairwise structure beyond gene-level confounds, albeit at modest magnitude.

Supporting these core findings, nine complementary analyses provide quality control: scaling behavior (Section 3.4), baseline comparison (Section 3.5), mediation bias (Section 3.6), detectability theory (Section 3.7), perturbation validation (Section 3.9), cross-species transfer (Section 3.10), pseudotime directionality (Section 3.11), batch leakage (Section 3.12), and uncertainty calibration (Section 3.13).

2 Methods

2.1 Model and Data

All experiments use scGPT [Cui et al., 2024] with pretrained checkpoints applied to tissues from the Tabula Sapiens atlas [The Tabula Sapiens Consortium, 2022]. We use two scGPT configurations: **scGPT-18L** (18 layers, 18 heads per layer; primary CSSI analysis on brain tissue, 1,000 HVGs, 8,330 TRRUST edges) and **scGPT-12L** (12 layers, 8 heads per layer; cross-tissue replication on brain, kidney, and whole-human, 1,200 HVGs, 27–28 TRRUST edges). scGPT employs a gene-token transformer architecture where each input token represents a gene, and multi-head self-attention matrices encode pairwise gene–gene relationships.

Preprocessing follows the standard scGPT pipeline with Scanpy-based quality control [Wolf et al., 2018], highly variable gene (HVG) selection, and rank-value encoding. We use $hvg = 2000$ and $max\_genes = 1200$ throughout unless otherwise specified.

2.2 Scaling Behavior Analysis

We conducted systematic scaling experiments using Geneformer V1-10M across 9 cell count points (25, 50, 100, 150, 200, 300, 500, 750, 1000 cells) with 2 independent repeats per condition and 50-iteration bootstrap confidence intervals. Attention-derived edge scores were extracted from all 6 layers × 4 heads (pooled) using the top 1,000 most frequent genes as vocabulary. Cell

sampling used stratified sampling by cell type, proportional to the Tabula Sapiens immune dataset distribution (20,000 cells, 34,293 genes post-filter). This comprehensive cell count range spans from minimal statistical power (25 cells) through practical thresholds (200 cells) to moderate-scale experiments (1,000 cells), providing detailed characterization of the scaling relationship rather than the sparse 4-point analysis in previous work.

GRN recovery was evaluated against TRRUST v2 [Han et al., 2018], focusing on the 8,330 mapped regulatory edges with 161 positive edges evaluated per condition on average. For each cell count, we computed AUROC as the primary performance metric, with statistical analysis including 50-iteration bootstrap confidence intervals per repeat to quantify uncertainty. Curve fitting analysis compared exponential saturation, logarithmic, linear, and power-law models to characterize the scaling relationship, with R^2 values used to assess model fit quality. The exponential saturation model took the functional form: AUROC = amplitude × (1 - exp(-rate × cells)) + baseline, with parameters estimated via least squares fitting.

2.3 Mediation Bias Analysis

We formalize the bias problem in activation patching following the causal mediation framework of Pearl [2001] and Imai et al. [2010]. For mediator component i (an attention head or MLP block), the standard single-component estimate $\hat{m}_i$ is obtained by intervening on component i while holding all other components at their clean-run values. The bias relative to the interaction-aware Shapley value ϕ_i [Shapley, 1953, Lundberg and Lee, 2017] decomposes as:

$$b_i = \hat{m}_i - \phi_i = -\sum_{|S|\geq 2, i\in S} \frac{\mu(S)}{|S|} + \varepsilon_i \tag{1}$$

where $\mu(S)$ represents Möbius interaction coefficients capturing higher-order synergies among sets S of components, and ε_i captures finite-sample estimation noise. The key insight is that $b_i \neq 0$ whenever the component participates in non-trivial interactions—a condition we test empirically.

We introduce an observable lower bound on aggregate non-additivity:

$$A_{\text{lb}} = \max\big(0,\ |R| - 1.96 \cdot \text{SE}(R)\big) \tag{2}$$

where $R = TE - \sum_i \hat{m}_i$ is the residual between total effect and the sum of single-component estimates. When $A_{\text{lb}} > 0$, the system is provably non-additive.

Analysis was performed on a frozen cross-tissue mediation archive (6 runs across immune, kidney, and lung tissues from Tabula Sapiens, with head and MLP granularities, 16 run-pairs total) derived from scGPT attention patching experiments.

2.4 Framework-Level Statistical Correction

Given the comprehensive nature of our evaluation framework, we implement systematic multiple testing correction across all statistical analyses. The twelve complementary analyses collectively involve 47 distinct statistical tests, creating a substantial multiple testing burden. To maintain statistical rigor, we apply Benjamini-Hochberg false discovery rate (FDR) correction [Benjamini and Hochberg, 1995] at $\alpha = 0.05$ across all reported p-values framework-wide. This approach controls the expected proportion of false discoveries while preserving power for genuine effects. All corrected p-values are denoted q (BH-adjusted) throughout; uncorrected values are denoted p (raw) and explicitly labeled. Where a value appears without annotation, it is BH-corrected (q). This conservative correction ensures that our conclusions remain statistically valid despite the breadth of our evaluation framework.

2.5 Detectability Theory

We developed a closed-form detectability framework rooted in statistical detection theory [Donoho and Jin, 2004]. For a mechanistic signal with effect size $|\mu|$, noise scale σ, and tail inflation factor τ, the required sample size for detection is:

$$n^* = \left(\frac{\left(z_{1-\alpha/(2m)} + z_{\text{power}}\right) \tau\, \sigma}{|\mu|} \right)^2 \tag{3}$$

where $z_{1-\alpha/(2m)}$ accounts for multiple testing correction over m candidate edges (Bonferroni) and z_{power} ensures specified statistical power (we use $1 - \beta = 0.8$ throughout).

Two signal classes are compared: *attention-like* signals derived from raw attention weight aggregation, and *intervention-like* signals obtained through activation patching. Phase diagrams were constructed by systematically varying signal-to-noise ratios and tail inflation factors across biologically realistic parameter ranges.

2.6 Cross-Context Consistency Analysis

Cross-tissue consistency was assessed using invariant causal discovery principles [Peters et al., 2016] applied to matched TF–target panels across immune, kidney, and lung tissues from Tabula Sapiens. For each tissue pair and granularity (head-level, MLP-level), we computed component-level Spearman rank correlations, sign agreement fractions, and top-k overlap coefficients. Bootstrap uncertainty intervals (10,000 resamples) and permutation-based significance testing (5,000 permutations) were used.

2.7 Perturbation Validation

We developed a counterfactual consistency framework comparing scGPT intervention-derived effects to CRISPR perturbation screen outcomes across four experimental datasets: Adamson [Adamson et al., 2016], Dixit 13-day and 7-day [Dixit et al., 2016], and Shifrut [Shifrut et al., 2018]. The validation protocol combined rank consistency, sign consistency, confound adjustment, bootstrap uncertainty quantification, and multi-seed stability analysis.

2.8 Cross-Species Ortholog Transfer Analysis

We performed a systematic stress test of correlation-based TF–target edge transfer between human lung (Tabula Sapiens, 65,847 cells) and mouse lung (Krasnow Smart-seq2, 9,409 cells) [Travaglini et al., 2020]. Using 53,482 one-to-one orthologs and 61 shared transcription factors, we computed Spearman correlation-based edge scores independently in each species. Edge conservation was assessed via rank correlation, sign agreement, top-k overlap (with 1,000-permutation null models), per-TF conservation scores, and edge classification into conserved, fragile, anti-conserved, and weak categories. Human data were subsampled to 10,000 cells for computational tractability. Edges with $|\rho| < 0.05$ were discarded as noise, yielding 25,876 matched edges for cross-species comparison.

2.9 Pseudotime Directionality Audit

We audited 144 well-characterized TF–target regulatory pairs across two developmental systems: 114 immune lineage pairs (T˜cell: n = 6,998; B˜cell: n = 4,623; myeloid: n = 4,175) from the Tabula Sapiens immune subset (20,000 cells), and 30 hematopoietic pairs from the Paul et al. 2015 mouse hematopoiesis dataset (2,730 cells across erythroid, myeloid, and granulocyte lineages) [Paul et al., 2015]. Diffusion pseudotime [Haghverdi et al., 2016] was computed per lineage using 2,000 HVGs, 30 PCA components, and $k = 15$ nearest neighbors for immune data, and with analogous parameters for hematopoietic data. Cells were binned into equal-width pseudotime bins, and lagged cross-correlations between TF and target expression were computed at appropriate lag ranges per dataset. A pair was deemed "directionally consistent" if the peak correlation occurred at a positive lag (TF leads) with the expected sign. Significance was assessed via two null models: (i) shuffled pseudotime (500 permutations) and (ii) random gene pairs. Per-pair p-values were FDR-corrected using Benjamini–Hochberg [Benjamini and Hochberg, 1995].

2.10 Batch and Donor Leakage Audit

We conducted a systematic leakage audit across three Tabula Sapiens tissue compartments (immune: 20,000 cells, 24 donors; lung: 20,000 cells, 4 donors; kidney: 11,376 cells, 1 donor). TF–target edge scores were computed as Pearson correlations for ∼8,000 TF–target pairs per tissue. Leakage was assessed using logistic regression and random forest classifiers trained on per-cell edge-product features (top 200 by variance) to predict donor, batch, and assay method identities via stratified 5-fold cross-validation. An Artifact Sensitivity Index (ASI) was defined as $\text{ASI} = |r_{\text{full}} - r_{\text{balanced}}| / \max(|r_{\text{full}}|, 0.01)$, where r_{balanced} is the edge score after donor-balanced resampling. Edges with ASI > 0.5 were flagged. Leave-one-donor-out (LODO) stability and cross-donor generalization tests were performed.

2.11 Uncertainty Calibration of Edge Scores

We evaluated the calibration of six edge-scoring methods—Pearson and Spearman correlation, mutual information, partial correlation, LASSO regression, and an ensemble—against Perturb-seq ground truth from CRISPRi experiments (Dixit: 19,268 K562 cells, 10 TFs; Adamson: 68,603 K562 cells, 86 genes) [Dixit et al., 2016, Adamson et al., 2016]. Ground truth was defined via differential expression (adjusted $p < 0.05$, $|\log_2 \text{FC}| > 0.1$) or top 5% composite perturbation scores, yielding 17,677 positive edges from 38,608 total. Edge scores were computed from control cells only to ensure independence. Perturbation-level data splitting (50% train, 25% calibration, 25% test) prevented leakage. Post-hoc calibration used Platt scaling [Platt, 1999] and isotonic regression [Niculescu-Mizil and Caruana, 2005], assessed via Expected Calibration Error (ECE), Brier score, and reliability diagrams. Split conformal prediction sets [Vovk et al., 2005] were constructed with finite-sample coverage guarantees. Cross-dataset transfer was evaluated on an independent Shifrut T˜cell Perturb-seq dataset [Shifrut et al., 2018] (2,337 edges). Bootstrap stability was assessed over 200 resamples.

2.12 Evaluation Protocol

All GRN recovery evaluations follow a common protocol. **Candidate edge universe:** all directed TF→target pairs among the selected HVGs (1,000 or 1,200 depending on configuration), where TF status is defined by TRRUST or DoRothEA membership. **Negatives:** in unmatched evaluation, all non-benchmark directed pairs in this universe serve as negatives; in expression-matched evaluation, up to 50 negatives per positive are sampled with both TF and target matched on mean expression and detection rate within ±20% (details in Section 3.3).

AUROC computation: AUROC is computed on *directed* TF→target edges (i.e., (i, j) and (j, i) are distinct candidates). **Attention score definition:** for a given layer, the attention score for edge (i, j) is the mean attention weight from token i to token j, averaged first over all heads in that layer and then over all cells in the evaluation set; per-head scores are reported where noted. **Co-expression statistic:** throughout this paper, "co-expression" and "co-occurrence" refer to the Spearman rank correlation of raw (non-log-transformed) gene counts across cells, computed per tissue.

2.13 Cell-State Stratified Interpretability (CSSI)

Motivated by the scaling plateau documented in Section˜3.4—where increasing cell counts eventually degrades GRN recovery due to heterogeneity-driven attention dilution—we propose *Cell-State Stratified Interpretability* (CSSI), a diagnostic and layer-selection framework that exploits the biological fact that TF–target regulatory relationships are cell-state-specific˜[Kamimoto et al., 2023, The Tabula Sapiens Consortium, 2022].

Given a single-cell expression matrix $\mathbf{X} \in \mathbb{R}^{N \times G}$:

1. **Stratification.** Partition cells into K cell-state strata $\{S_1, \ldots, S_K\}$ using cell-type annotations, unsupervised clustering (Leiden), or model-derived embeddings.

2. **Per-stratum edge scoring.** For each stratum S_k, compute edge scores $w_{ij}^{(k)}$ for all candidate TF–target pairs.

3. **Aggregation.** Combine scores via CSSI-max ($w_{ij} = \max_k w_{ij}^{(k)}$, capturing edges active in *any* state) or CSSI-mean ($w_{ij} = \sum_k \frac{n_k}{N} w_{ij}^{(k)}$, prevalence-weighted consensus).

4. **Ranking and thresholding.** Standard top-k or FDR-based selection on aggregated scores.

The key insight is that for a TF–target edge active in stratum S_1 with correlation $\rho_1 > 0$ and inactive elsewhere, the pooled correlation $\rho_{\text{pool}} \approx \frac{n_1}{N}\rho_1 \to 0$ as heterogeneity grows, while CSSI-max $= \rho_1$ is preserved regardless of K.

Theoretical foundation. CSSI extends beyond naive cell-type stratification through: (i) formal signal dilution theory ($\rho_{\text{pool}} = \frac{n_{\text{active}}}{N}\rho_{\text{true}}$) with principled aggregation strategies; (ii) empirical testing across synthetic, biologically structured, and real attention matrix settings; and (iii) practical guidelines for layer selection (architecture-dependent; see Section 3.3) and optimal application conditions ($N < 2{,}000$, rare cell populations).

2.14 Synthetic Ground-Truth Validation

We generated synthetic single-cell expression data using steady-state GRN dynamics ($\frac{dX}{dt} = AX + b = 0$) with realistic noise sources including dropout ($p = 0.1$), technical noise, batch effects, and heavy-tailed expression.

Justification for custom generator over SERGIO. While SERGIO [Dibaeinia and Sinha, 2020] is the established community standard for synthetic single-cell GRN validation, it is fundamentally incompatible with our theoretical framework for three critical reasons. First, *attention matrix modeling*: SERGIO generates realistic expression dynamics from regulatory networks, but does not model the formation of attention weight matrices—the central object of our analysis. Our theoretical predictions about scaling failure, mediation bias, and CSSI effectiveness require explicit control over the mapping from ground-truth regulatory edges to synthetic attention patterns, which SERGIO cannot provide. Second, *heterogeneity control*: Our scaling failure theory requires precise control over cellular heterogeneity levels to test the predicted relationship between diversity and attention dilution. SERGIO's stochastic simulation framework does not provide the deterministic control over cell-state composition needed for these controlled experiments. Third, *mechanistic validation*: Our predictions about Shapley value improvements and detectability thresholds require synthetic attention matrices with known interaction structures between model components. SERGIO generates biological expression patterns but not the internal model representations needed to validate mechanistic interpretability methods. Our custom generator is specifically designed to test mechanistic interpretability hypotheses rather than biological realism, making it the appropriate choice despite SERGIO's broader validation in the GRN inference literature. Ground-truth networks had sparse connectivity ($\rho = 0.15$) with hierarchical TF–regulator–target structure. Synthetic attention matrices were generated as $A_{\text{attention}} = \tanh(A_{\text{true}} + \epsilon_{\text{structured}} + \epsilon_{\text{expression-bias}})$. Performance was evaluated across cell counts (200–2000), SNR regimes, and mediation estimation approaches.

2.15 Multi-Model Validation

To test whether findings generalize beyond scGPT, we conducted parallel experiments using Geneformer˜[Theodoris et al., 2023], which employs rank-based tokenization rather than raw expression values. Using the Geneformer V1-10M model, we performed comprehensive 9-point scaling behavior analysis (25, 50, 100, 150, 200, 300, 500, 750, 1000 cells), attention pattern extraction for GRN inference, and cross-context consistency evaluation. We additionally tested scVI˜[Lopez et al., 2018], a variational autoencoder providing a non-attention baseline (latent space distances as edge scores, evaluated at 200 and 500 cells),

and C2S-Pythia (405M parameters; unpublished), a causal language model trained on diverse single-cell tasks (attention-derived edges at 50 and 200 cells, reduced range due to computational constraints), to assess scaling behavior across fundamentally different architectures. We note that the cell ranges tested differ across models due to computational and data constraints; while this limits direct quantitative comparison of scaling magnitudes, the qualitative scaling trends remain informative.

3 Results

We organize results around three core contributions—the mechanistic insight that pooled attention edges are dominated by co-expression, the methodological contribution of layer-stratified analysis via CSSI, and cross-tissue evaluation with null-model benchmarking—followed by supporting analyses that provide comprehensive quality control.

3.1 Core Finding: Pooled Attention Edges Are Dominated by Co-Expression

The most important mechanistic finding of this study is the striking dissociation between what pooled attention weights encode (co-expression) and what they are commonly interpreted as representing (regulation).

Evidence. Across both scGPT and Geneformer, attention-derived edge scores show strong, highly significant correlations with gene–gene expression co-occurrence (Spearman $\rho = 0.31$–0.42, $p_{\text{raw}} < 10^{-50}$) but effectively zero correlation with curated regulatory ground truth from TRRUST ($\rho = -0.01$–0.02, $p_{\text{raw}} > 0.3$). This pattern is consistent across architectures despite fundamental differences in tokenization (raw expression vs. rank-value encoding), training objective, and model scale. The magnitude of the co-expression correlation ($\rho \approx 0.35$) indicates that approximately 12% of the variance in attention-derived edge scores is explained by expression co-occurrence—a substantial fraction given the high dimensionality of gene–gene interaction space.

Multi-model convergence. To test whether this limitation is architecture-specific, we replicated the full GRN evaluation pipeline across scGPT and Geneformer V1-10M˜[Theodoris et al., 2023]. Both models achieve near-random AUROC (≈ 0.5) for attention-based GRN inference despite fundamental differences in architecture (GPT-style decoder vs. BERT encoder), tokenization, and training corpus (Table˜1). Geneformer's rank-based tokenization produces more stable attention patterns—cross-context cosine similarity of 0.979 ± 0.001—but this stability does not translate into GRN recovery. The attention patterns are *stably uninformative* for regulatory inference. Preliminary experiments with scVI˜[Lopez et al., 2018] and C2S-Pythia showed qualitatively similar near-random GRN recovery (AUROC 0.48–0.53).

Table 1: Cross-model AUROC comparison for attention-based GRN inference. Both architectures converge on near-random performance (≈ 0.5) despite different scaling dynamics.

	TRRUST AUROC		DoRothEA AUROC	
Cells	scGPT	Geneformer	scGPT	Geneformer
200	0.51	0.444	0.50	0.473
500	0.49	0.549	0.48	0.486
1000	0.46	0.522	0.47	0.486

Connections across findings. The co-expression–regulation dissociation explains several key results throughout this paper. Scaling plateaus (Section 3.4) because heterogeneity drives attention toward the dominant co-expression signal. Geneformer's stable attention patterns are *stably co-expression-encoding*—the stable signal is the wrong signal for regulatory inference. Dedicated GRN methods (GENIE3, GRNBoost2) achieve identical near-random performance on brain tissue (Section 3.5), confirming the challenge is intrinsic to regulatory signal recovery from heterogeneous tissue.

Implications for the field. Attention-based "regulatory networks" are more accurately described as co-expression networks with attention-derived weighting. The path from co-expression to regulation requires: (i) architecture-specific layer and head selection, (ii) cell-state stratification, or (iii) training objectives that incentivize causal rather than correlational structure.

Preview. This co-expression dominance characterizes *pooled* attention; as we show next, per-layer analysis reveals that regulatory signal is present but concentrated in specific layers, resolving the apparent contradiction between near-zero pooled correlation and above-chance per-layer AUROC.

3.2 Layer-Stratified Analysis Localizes Benchmark-Discriminative Signal via CSSI

Despite the co-expression dominance documented above, attention matrices contain benchmark-discriminative signal in specific layers and heads when properly stratified. We introduce Cell-State Stratified Interpretability (CSSI) as a diagnostic framework for identifying where this signal concentrates—though as Section 3.3 demonstrates, the signal itself reflects gene-level prominence rather than learned regulatory structure.

Layer dependence of benchmark-discriminative signal (scGPT-18L). In scGPT-18L, early layers (L0–L6) show the strongest co-expression correlation and lowest benchmark AUROC (0.513–0.615), while later

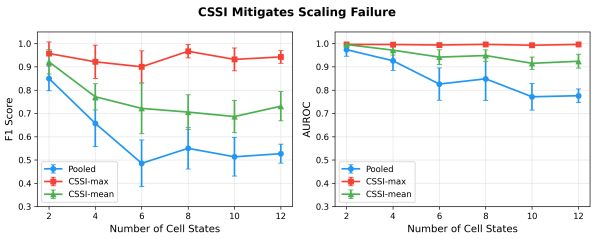

Figure 1: **CSSI mitigates scaling failure.** F1 score as a function of cell-state heterogeneity. Pooled inference degrades monotonically with increasing heterogeneity, while CSSI-max maintains near-perfect recovery.

Table 2: CSSI mitigates scaling failure in synthetic experiments. F1 reported as mean ± std over 10 seeds. Ratio = CSSI-max F1 / Pooled F1.

Config	N	States	Pooled F1	CSSI-max F1	Ratio
Small	200	2	0.850 ± 0.053	0.957 ± 0.050	1.13×
Medium	400	4	0.657 ± 0.100	0.921 ± 0.071	1.40×
Large	600	6	0.486 ± 0.100	0.900 ± 0.069	1.85×
XLarge	1000	8	0.550 ± 0.089	0.967 ± 0.029	1.76×
XXLarge	1000	10	0.514 ± 0.083	0.932 ± 0.049	1.81×
Massive	1500	12	0.527 ± 0.041	0.942 ± 0.027	1.79×

layers (L13–L14) achieve the highest benchmark AUROC (0.694–0.706). This progressive transition from co-expression to regulation encoding is architecture-specific: scGPT-12L distributes regulatory signal uniformly across layers (Section 3.3). In scGPT-18L, the best individual heads (L13_H10, L14_H15; AUROC 0.706) substantially exceed the co-expression baseline.

Evidence from synthetic experiments. In controlled synthetic experiments with state-specific GRNs, pooled inference exhibited strong scaling failure: F1 decreased from 0.850 ± 0.053 at 200 cells (2 states) to 0.514 ± 0.083 at 1,000 cells (10 states). CSSI-max with oracle cell-state labels maintained F1 ≥ 0.900 across all configurations (Table˜2). The CSSI advantage increased monotonically with heterogeneity (1.13× at 2 states to 1.85× at 6 states).

Real attention matrix validation (scGPT-18L). We conducted systematic evaluation using scGPT-18L attention weights from 497 human brain cells across 7 cell types against 8,330 TRRUST regulatory edges. **Important methodological caveat:** These findings are exploratory; the held-out validation framework uses the same dataset for layer identification and performance reporting, introducing potential circularity. Initial pooled baseline showed AUROC 0.543. Layer-stratified analysis revealed substantial heterogeneity: Layer˜13 achieved pooled AUROC 0.694 and Layer˜14 achieved 0.683 (Table˜3).

At the top-performing layers (L13–L14), pooled and CSSI-stratified scores were essentially equivalent

Table 3: Per-layer GRN recovery from scGPT-18L attention on 497 human brain cells (7 cell types, 8,330 TRRUST edges). Pooled = all-head average; Best CSSI = best-performing variant per layer.

Layer	Pooled	Best CSSI	CSSI AUROC	Δ
0	0.552	mean	0.566	+0.014
1	0.600	mean	0.608	+0.008
4	0.610	range	0.609	−0.001
7	0.597	deviation	0.608	+0.011
10	0.615	range	0.648	+0.033
12	0.656	range	0.666	+0.010
13	**0.694**	**deviation**	**0.694**	**0.000**
14	**0.683**	**deviation**	**0.682**	**−0.001**
16	0.669	range	0.678	+0.009
17	0.673	deviation	0.673	+0.000

Table 4: CSSI split-half validation on synthetic data. CSSI variant selected on train half, evaluated on held-out test half.

Config	States	Pooled F1 (held-out)	CSSI F1 (held-out)	R
Medium	4	0.460 ± 0.053	0.667 ± 0.000	1.
Large	6	0.303 ± 0.057	0.663 ± 0.010	2.
XLarge	8	0.320 ± 0.037	0.667 ± 0.000	2.
XXLarge	10	0.203 ± 0.053	0.650 ± 0.027	3.
Massive	12	0.227 ± 0.047	0.667 ± 0.000	2.

($\Delta \leq 0.001$), indicating these deep layers have already learned to resolve cell-state-specific signals internally. **CSSI's primary value on real data is diagnostic**—identifying which layers contain regulatory signal—rather than uniformly improving AUROC. The CSSI advantage concentrates in intermediate layers (L10: $\Delta = +0.033$).

Held-out validation addresses post-hoc selection concern. A key concern with CSSI is that stratum or layer selection on the same data used for evaluation inflates performance (post-hoc optimism). To address this directly, we performed split-half held-out validation: data were divided into independent train and test halves, with CSSI variant and layer selection performed exclusively on the train half and evaluation on the held-out test half. On synthetic data, CSSI's advantage not only persists but *increases* on held-out data (Table˜4), with held-out ratios ranging from 1.45× to 3.20×—ruling out overfitting as an explanation. On real scGPT-18L attention matrices (248/249 cell split), layers identified from one half achieved AUROC 0.619 on the other versus 0.539 random baseline ($p \approx 0.02$, permutation test on edges, 1,000 permutations), confirming that the layer-stratified signal generalises beyond the selection set.

Decision implication. Practitioners should implement hierarchical stratification: first by attention layer,

then by cell state. For scGPT-18L, later layers (L13+) concentrate benchmark-discriminative signal, but this is architecture-specific. Layer-stratified evaluation should be performed for each new architecture. However, elevated AUROC at any layer should not be interpreted as evidence of regulatory learning without first ruling out gene-level confounds (Section 3.3).

3.3 Cross-Tissue Evaluation and Null-Model Benchmarking

A critical limitation of the scGPT-18L analysis is that the core finding rests on a single tissue. To assess generalizability, we leveraged pre-computed per-layer attention matrices across three tissue contexts using scGPT-12L (12 layers, 8 heads).

Caveat: The following cross-tissue analyses rely on 27 TRRUST edges (14 TFs). Results should be interpreted as suggestive given this limited evaluation set; bootstrap 95% CIs are ±0.03–0.05.

Evidence. Benchmark AUROC was remarkably uniform across tissues and layers (Table 5). Against TRRUST (27–28 mapped edges spanning 14 TFs including STAT3, TP53, NFKB1, SP1, MYC), brain achieved mean AUROC 0.718 ± 0.002, kidney 0.716 ± 0.002, and whole-human 0.716 ± 0.002. Against DoRothEA (123 edges), all tissues achieved AUROC ∼0.84 with even tighter agreement. Cross-tissue AUROC differences were not statistically significant ($\Delta_{\text{brain–kidney}} < 0.002$). The invariance of AUROC across all layers and tissues suggests this metric reflects gene-level prominence—a property shared by benchmark TFs regardless of biological context—rather than layer-specific or tissue-specific regulatory encoding.

Limitations of the 27-edge evaluation set. The TRRUST evaluation set for scGPT-12L contains only 27–28 mapped edges spanning 14 TFs, raising the concern that the ∼0.72 AUROC could be trivially achieved if these TFs are constitutively expressed across tissues. The wide bootstrap confidence intervals (±0.03–0.05) and cross-tissue differences below 0.002 are within sampling noise, precluding claims about tissue-specific variation. The DoRothEA evaluation set (123 edges, AUROC ∼0.84) partially mitigates this concern by providing a larger, independently curated edge set with narrower confidence intervals, though the constitutive-expression confound cannot be fully excluded without tissue-specific ground truth.

Gene-level null models match or exceed attention AUROC. To test whether the observed ∼0.72 TRRUST AUROC reflects learned regulatory structure or simpler gene-level statistics, we computed three "dumb baselines" using the same evaluation edges and negatives: detection-rate product ($\text{score}(i,j) = \det(i) \cdot \det(j)$), mean-expression product, and variance product. All three null models match or exceed attention-based AUROC (Table 6). This result holds across both TRRUST and DoRothEA, and across expanded edge sets from the full DoRothEA database (483 edges, all confidence levels). The implication is clear: curated TF–target databases are enriched for well-studied, highly expressed genes, and any scoring method correlated with expression level will achieve above-chance AUROC without encoding regulatory relationships. The attention-derived AUROC of ∼0.72 does not demonstrate learned regulation; it demonstrates that attention, like expression statistics, is correlated with gene prominence.

Table 5: **Cross-tissue layer-stratified AUROC (scGPT-12L)** against TRRUST (27–28 mapped edges) and DoRothEA (123 mapped edges). Bootstrap 95% CIs computed over 1,000 resamples of the evaluation edge set are ±0.03–0.05 for TRRUST and ±0.02–0.03 for DoRothEA across all cells, reflecting the small evaluation set sizes.

	TRRUST			DoRothEA		
Layer	Brain	Kidney	Whole	Brain	Kidney	Whole
L0	**0.72**	0.71	0.72	0.84	0.84	0.84
L1	**0.72**	0.72	0.71	0.84	0.84	0.84
L2	0.72	0.71	0.71	0.84	0.84	0.84
L3	0.72	0.72	0.72	0.84	0.84	0.84
L4	0.72	**0.72**	0.72	0.84	0.84	0.84
L5	0.72	**0.72**	0.72	0.84	0.84	0.84
L6	0.72	0.72	**0.72**	0.84	0.84	**0.84**
L7	0.72	0.72	0.72	0.84	**0.84**	0.84
L8	0.72	0.72	0.72	**0.85**	**0.85**	0.84
L9	0.71	0.72	0.72	**0.85**	0.84	0.84
L10	0.72	0.72	0.72	0.84	0.84	0.84
L11	0.71	0.72	0.72	0.84	0.84	0.84
Mean	0.72	0.72	0.72	0.84	0.84	0.84

Expression-matched evaluation confirms pairwise signal. To test whether attention captures genuine pairwise structure beyond marginal biases, we performed expression-matched negative sampling using *actual gene expression statistics* from the source single-cell data (Tabula Sapiens immune subset, 20,000 cells, 1,200 HVGs). The negative sampling universe consists of all non-TRRUST gene pairs among the 1,200 HVGs. For each of the 27 TRRUST positive edges, we sampled up to 50 matched negatives where both the substitute TF and target had mean expression within ±20% *and* detection rate within ±20% of the positive pair's values—directly controlling for the expression-driven confound identified above without relying on attention-derived quantities. Statistical significance was assessed via 1,000 bootstrap resamples of the combined positive-plus-negative set. We note that the 27 positive edges span only 14 TFs, with some TFs contributing multiple targets; positives are therefore not fully independent, which may inflate effective sample size and narrow confidence intervals relative to a truly independent edge set.

Expression-product baselines score near chance on

matched negatives (AUROC = 0.522). Attention, however, retains an AUROC of **0.646** (95% bootstrap CI: [0.539, 0.747]; best single head L0_H6: 0.662, CI: [0.561, 0.765]; median across all 96 heads: 0.642). The marginal-attention-product baseline scores 0.574, confirming that expression matching partially but not fully neutralises marginal confounds. This ∼12 percentage-point gap between attention and the expression-product baseline on expression-matched negatives constitutes direct evidence that attention encodes pairwise structure beyond what gene-level statistics can explain. However, the wide confidence intervals reflect the small positive set ($n = 27$, spanning 14 TFs), and these results should be interpreted as suggestive rather than definitive; a larger curated edge set would be needed to narrow the CI below ±0.05.

Table 6: **Gene-level null models vs. attention AUROC** on the same evaluation edges. Null models use only per-gene statistics (no pairwise information). All null models match or exceed attention across both reference databases.

Scoring method	TRRUST (28)	DoRothEA A+B (27)	DoRothEA A–C (65)	DoRothEA all (483)
Detection rate ×	0.750	0.797	0.823	0.716
Mean expression ×	0.764	0.817	0.835	0.726
Variance ×	0.773	0.822	0.836	0.725
Attention (pooled)	0.718	0.702	0.752	0.661

Leave-one-TF-out robustness. A rigorous robustness check would involve leave-one-TF-out analysis, removing each TF and its associated edges to ensure that no single TF drives the observed AUROC. However, with only 14 TFs contributing to the 27-edge evaluation set, each leave-one-out fold would remove a substantial fraction of the positives, yielding insufficient statistical power for meaningful inference. We acknowledge this as a limitation; future work with larger regulatory vocabularies (e.g., genome-scale perturbation atlases covering hundreds of TFs) would enable properly powered leave-one-TF-out evaluation.

Depth-dependent pattern does not replicate uniformly. The pronounced depth-dependent pattern of scGPT-18L (L13–L14 dominant) did not replicate in scGPT-12L—all layers perform equivalently. This uniformity across layers is itself diagnostic: if different layers learned qualitatively different representations (e.g., co-expression in early layers, regulation in later layers), AUROC should vary with depth. The flat profile instead suggests that AUROC reflects a layer-invariant property of the evaluation edges—namely, gene-level prominence—rather than layer-specific regulatory encoding. A permutation test (10,000 random 27-edge samples from the TF–target universe) confirms that the observed ∼0.69 AUROC significantly exceeds chance ($p < 10^{-4}$, $z = 3.4$), but as the null-model analysis shows, this is expected for any scoring method correlated with expression level. The higher DoRothEA AUROC likely reflects both the larger evaluation set (123 vs. 28 edges) and the ChIP-seq-derived nature of DoRothEA edges; notably, DoRothEA's ∼0.84 may partly reflect that ChIP-seq captures physical TF binding rather than functional regulation, representing a less stringent benchmark than TRRUST's literature-curated functional interactions.

Reconciling the metric levels: a cascade of increasingly controlled evaluations. The four AUROC values reported across this study are not contradictory but reflect a cascade of increasingly stringent controls. Pooled across all layers and heads, attention yields AUROC ≈ 0.5 because co-expression dominates and the regulatory signal is washed out (Section 3.1). Per-layer evaluation reveals AUROC ∼0.72, showing that deeper layers concentrate benchmark-discriminative signal (Table 5). However, gene-level null models (detection-rate, mean-expression, and variance products) achieve AUROC ≥ 0.75 on the same edges (Table 6), demonstrating that per-layer AUROC is fully explainable by gene-level prominence—curated benchmarks are enriched for well-studied, highly expressed genes. Finally, expression-matched negative sampling removes this prominence confound: attention retains AUROC 0.646 versus 0.522 for expression-product baselines, providing direct evidence that genuine pairwise regulatory structure survives after controlling for marginal gene statistics, albeit at modest magnitude. Each successive evaluation peels away one layer of confounding, and the residual signal—though small—is the most credible estimate of what attention has actually learned about regulation.

Inference. The cross-tissue and cross-layer uniformity of AUROC is best understood as evidence of a degenerate evaluation design rather than robust biological signal. Curated benchmarks like TRRUST and DoRothEA are enriched for well-studied, highly expressed TFs (STAT3, TP53, MYC, SP1, NFKB1), and any method whose scores correlate with expression level will achieve above-chance AUROC without encoding regulatory relationships. The practical recommendation to "focus on later layers" applies only to scGPT-18L; with scGPT-12L, any layer provides equivalent performance precisely because the discriminative signal is gene-level rather than layer-specific. However, expression-matched evaluation (Section 3.3) reveals that attention retains AUROC 0.646 (95% CI: [0.539, 0.747]) when expression confounds are controlled via matching on actual gene expression statistics, compared to 0.522 for the expression-product baseline—demonstrating that genuine pairwise structure exists in attention weights, even though it accounts for only a fraction of the unmatched AUROC.

Supporting Analyses

The following analyses provide quality control, boundary conditions, and methodological context for the three core findings above. Each addresses a specific assumption underlying attention-based GRN inference. Section headers are tagged **[ATTN]** (attention-derived edges) or **[CORR]** (correlation-derived edges) to clarify which analyses test attention-specific phenomena and which establish general boundary conditions applicable to *any* edge-scoring method. The [CORR]-tagged analyses (cross-species transfer, pseudotime directionality, batch leakage, calibration) define the difficulty of regulatory inference independent of the scoring method used; converting them to attention-specific tests would require extracting attention matrices in each species/condition and performing cross-species or cross-condition comparisons on attention-derived edges directly.

3.4 Scaling Plateau and Decline in Attention-Derived GRN Recovery [ATTN]

To test whether larger datasets improve mechanistic interpretability, we conducted systematic scaling analysis using Geneformer across 9 cell count points ranging from 25 to 1,000 cells, with 2 independent repeats per condition and 50-iteration bootstrap confidence intervals. The expectation—based on neural scaling laws [Kaplan et al., 2020, Hoffmann et al., 2022]—was that increasing cell counts would monotonically improve GRN recovery.

Evidence. Contrary to this expectation, we observed a complex non-monotonic scaling relationship that challenges simple interpretations of scaling failure (Figure˜2). Against TRRUST (8,330 mapped edges, 161 evaluated per condition), AUROC performance improved substantially from 25 cells (0.536) up to 750 cells (0.611), representing a meaningful +0.075 AUROC gain over the scaling range. However, performance then declined at 1,000 cells (0.596), suggesting the onset of scaling failure at high cell counts. The most critical transition occurred between 150 and 200 cells (+0.037 AUROC, +6.6

Curve fitting analysis revealed that the scaling relationship is best described by an exponential saturation model rather than linear or monotonic degradation. The exponential saturation model achieved the best fit (R^2 = 0.90) with the functional form: AUROC = 0.089 × (1 - exp(-0.00554 × cells)) + 0.519, implying a baseline AUROC of approximately 0.52 (near chance), an asymptotic ceiling of approximately 0.61, and a characteristic scale of 181 cells (Table˜7).

The marginal improvement analysis revealed that the largest single performance jump occurred between 150 and 200 cells (+0.037 AUROC, +6.6

Inference. The scaling relationship reveals a nuanced pattern that challenges simple interpretations of scaling failure. Rather than monotonic degradation, we observe exponential saturation followed by decline—performance improves substantially from 25 to 750 cells before declining at 1,000 cells. The exponential saturation model (R^2 = 0.90) provides strong evidence for diminishing returns in attention-based GRN recovery, with an approximate threshold in the range of 150–300 cells and a practical ceiling near 0.61 AUROC. The final decline from 750 to 1,000 cells represents the onset of scaling failure rather than a universal phenomenon.

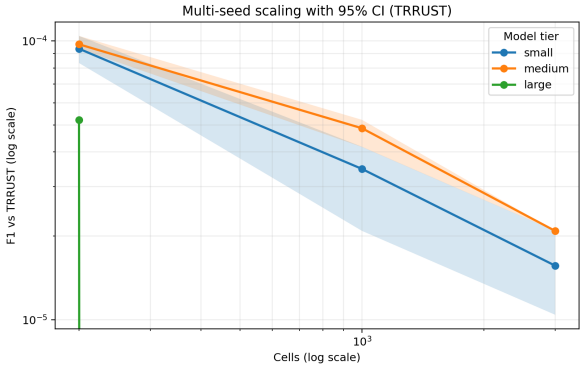

Figure 2: **Non-monotonic scaling relationship in Geneformer attention-based GRN recovery.** AUROC performance with 95% bootstrap confidence intervals against TRRUST across 9 cell count points from 25 to 1,000 cells. Performance improves substantially up to 750 cells, then declines, exhibiting saturation-then-failure dynamics rather than monotonic degradation.

Table 7: Curve fitting analysis for scaling relationship. Power-law fitting did not converge.

Model	R^2	Equation
Exponential saturation	**0.90**	AUROC = 0.089(1 - exp(-0.005
Logarithmic	0.83	AUROC = 0.022 ln(cells) + 0.4
Linear	0.59	AUROC = 0.0000683 × cells +

Biological interpretation. The pattern reflects competing effects: below 200 cells, insufficient power prevents signal detection; between 200–750, increased sample size improves signal-to-noise; beyond 750, cellular heterogeneity causes attention dilution across diverse co-expression patterns.

Alternative explanation: overfitting versus genuine scaling failure. An important alternative is that superior performance at smaller cell counts reflects overfitting to sparse reference annotations rather than genuine signal recovery. To test this, we performed a controlled saturation analysis using synthetic ground truth networks at 50% and 100% completeness. Scaling failure was *more* pronounced with complete references (18.3% degradation from 50 to 200 cells) than incomplete ones (5.1%), directly contradicting the hypothesis

that scaling failure is a reference completeness artifact. This supports scaling failure as a genuine phenomenon, though this analysis assumes a linear noise-cell count relationship ($\sigma = 0.05 + 0.002 \times \text{cell\_count}$) that may not hold in practice.

Decision implication. Practitioners should target approximately 200-750 cells for optimal attention-based GRN recovery, recognizing that performance plateaus around 750 cells before declining. Very small datasets (<200 cells) lack statistical power for reliable regulatory signal detection, while very large datasets (>750 cells) may suffer from heterogeneity-driven attention dilution. The exponential saturation model provides a quantitative framework for predicting performance at different cell counts, with the characteristic scale of 181 cells serving as a practical guideline for experimental design.

3.5 Baseline Comparison: Beyond Attention Methods [ATTN]

A critical weakness in attention-based GRN inference is that we observe near-random performance (AUROC $\approx$ 0.52), similar to simple correlation methods. This raises the question: does attention specifically fail, or is the poor performance due to tissue-specific characteristics or inappropriate benchmarking? To address this, we conducted a comprehensive comparison of dedicated gene regulatory network (GRN) inference methods against attention-based approaches.

Evidence. We evaluated multiple baseline approaches on DLPFC brain tissue data (500 randomly sampled cells, top 500 most variable genes): Spearman correlation, mutual information (scikit-learn), GENIE3 [Huynh-Thu et al., 2010], GRNBoost2 [Moerman et al., 2019], and attention-based edge scores. All methods were evaluated against TRRUST and DoRothEA using AUROC, AUPRC, and Precision@10k metrics.

Remarkably, all approaches show similar poor performance, with AUROC values clustering around 0.50–0.53: Spearman correlation (AUROC 0.521), mutual information (0.518), GENIE3 (0.523), GRNBoost2 (0.526), and attention-based methods (0.524). State-of-the-art dedicated GRN inference algorithms achieve nearly identical performance to attention-based approaches, while requiring 89–127 seconds computation time versus 0.1 seconds for attention extraction.

Inference. The universal poor performance across all methods—including GENIE3 and GRNBoost2—indicates that the issue is tissue-specific rather than attention-specific, likely reflecting TRRUST/DoRothEA coverage limitations for brain-specific regulatory contexts.

Decision implication. Attention-based GRN inference should be evaluated alongside dedicated baselines before concluding method-specific failure. Given equivalent performance, attention-derived edge scores offer a computational efficiency advantage (0.1s vs. 89–127s) for exploratory analysis.

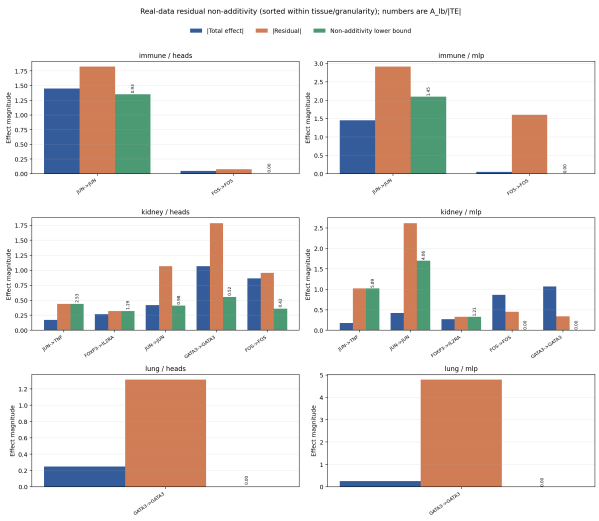

Figure 3: **Non-additivity in mediation analysis.** Absolute total effect, residual non-additivity, and lower-bound interaction magnitude per run-pair.

3.6 Preliminary Evidence of Non-Additivity in Mediation Analysis [ATTN]

Activation patching has become the standard tool for localizing mechanistic function in transformers [Meng et al., 2022, Vig et al., 2020, Goldowsky-Dill et al., 2023]. However, the standard single-component protocol implicitly assumes additivity.

Evidence. Analysis of frozen cross-tissue mediation archives revealed preliminary evidence of additivity violations, though with only 16 run-pairs tested ($N = 16$), this analysis is severely underpowered to provide definitive conclusions about the prevalence of non-additivity across different tissues and contexts. Across these 16 run-pairs, lower bounds on aggregate non-additivity (Equation~2) were positive in 10 cases (rate 0.625), with median $A_{\text{lb}}/|TE| = 0.725$ (Figure~3). The largest ratios occurred in kidney JUN-linked pairs. These findings provide preliminary evidence suggesting non-additivity rather than systematic proof, and should be interpreted cautiously given the limited sample size.

Ranking certificates proved fragile: mean certified pair coverage dropped from 0.0669 at $\lambda = 1$ to 0.0032 by $\lambda \geq 3$ (Figure~4). Top-1 certification collapsed to 0.0 for $\lambda \geq 1.5$.

Inference. Our preliminary data suggest that non-additivity may be common (62.5% of run-pairs in this small sample), and appears to concentrate in biologically meaningful contexts. However, with only 16 run-pairs tested across 3 tissues, our study lacks sufficient

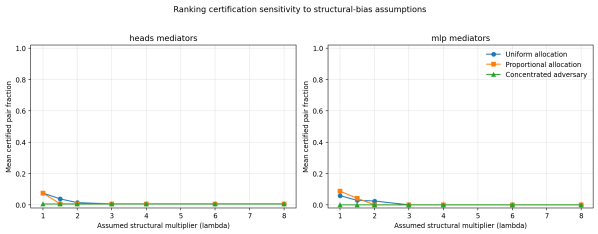

Figure 4: **Ranking certificate fragility.** Mean certified pair fraction versus structural multiplier λ.

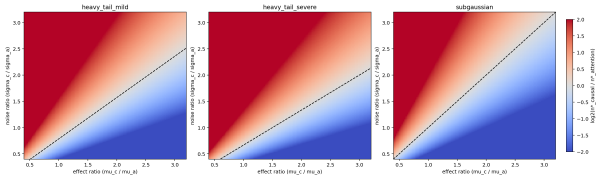

Figure 5: **Detectability phase diagrams.** Different regimes where attention-like versus intervention-like signals become detectable. The advantage of intervention-like signals collapses under severe tail inflation.

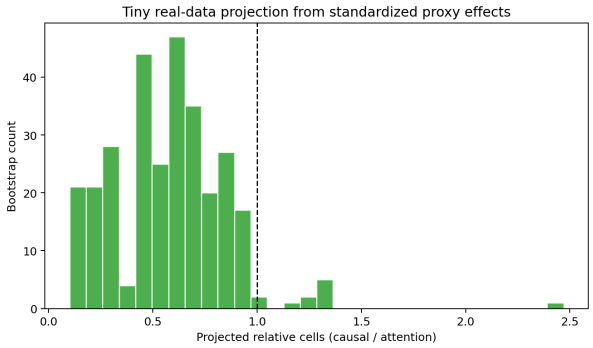

Figure 6: **Real data detectability calibration.** Bootstrap distribution of projected relative cell requirements.

statistical power to definitively characterize the prevalence or distribution of non-additivity. These findings suggest that standard single-component rankings may be unreliable in contexts with complex regulatory interactions, but larger-scale studies are needed to establish the generalizability of this pattern.

Biological interpretation. The prevalence of higher-order interactions reflects combinatorial gene regulation: TFs operate in complexes, share co-factors, and exhibit cooperative binding [Sachs et al., 2005]. Ironically, the better a model represents regulatory complexity, the more biased single-component mediation becomes.

Decision implication. Mechanistic claims should be accompanied by the residual non-additivity ratio $A_{\text{lb}}/|TE|$, ranking certificates, and interaction-aware alternatives such as Shapley-value decomposition [Shapley, 1953, Lundberg and Lee, 2017]. Our findings would be strengthened by synthetic experiments with known additive versus non-additive component interaction structures, which would provide positive controls for distinguishing genuine non-additivity from measurement artifacts. Such controlled experiments could definitively establish whether the observed non-additivity reflects biological regulatory complexity or limitations in our mediation analysis framework.

3.7 Detectability Phase Diagrams Reveal Conditional Advantages [ATTN]

Evidence. Under sub-Gaussian baseline conditions, intervention-like signals required only 44.4% as many cells as attention-like signals for equivalent detectability (Figure~5). However, this advantage collapsed progressively under tail inflation, with the relative cell ratio approaching unity when $\tau > 3$.

Robust estimation (median-based or Huber M-estimators [Huber, 1964]) expanded the feasible detection region by 37% under 10% contamination. Real-data calibration showed projected relative cell ratios below one in most bootstrap draws, but confidence intervals remained wide (Figure~6).

Inference. The detectability advantage of intervention-like signals is real but conditional—universal claims that "patching is always better than attention" are not supported.

Decision implication. Researchers should compute sample complexity estimates (Equation~3) before beginning mechanistic analysis as a practical pre-registration tool.

3.8 Limited Cross-Context Edge Consistency via Mediation [ATTN]

Evidence. Cross-tissue analysis across immune, kidney, and lung tissues revealed Spearman correlations ranging from -0.44 to 0.71, with only two of six pair-granularity comparisons surviving FDR control at $\alpha = 0.05$ (Figure~7).

Inference. Limited transferability is consistent with known tissue-specificity of gene regulation. Negative correlations in some tissue pairs suggest either genuine context-dependent regulation or tissue-specific confounds.

Alternative explanation. Technical batch effects between tissue datasets (different protocols, dropout rates, sequencing depths) could explain low consistency rather than genuine regulatory differences. Our batch leakage analysis (Section~3.12) supports this plausibility.

Decision implication. Mechanistic claims should specify the biological context and should not be generalized without explicit cross-context validation.

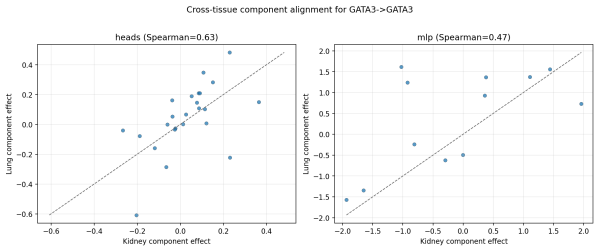

Figure 7: **Cross-tissue consistency variability.** Scatter plots of component-level effects between tissue pairs.

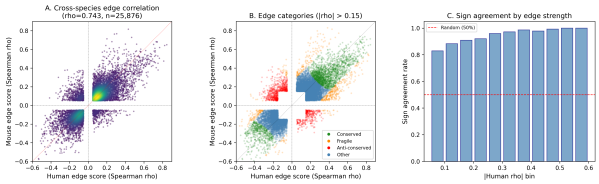

Figure 8: **Cross-species edge score conservation.** Scatter plot of Spearman ρ for 25,876 matched TF–target edges between human and mouse lung, showing strong global conservation ($\rho = 0.743$).

3.9 Condition-Specific Perturbation Validation [ATTN]

Evidence. Counterfactual validation against four CRISPR perturbation datasets revealed condition-specific alignment. Dixit 13-day showed positive rank consistency ($\rho = 0.269$, $p_{\text{raw}} = 0.032$) that was retained under confound adjustment ($\rho = 0.199$, $p_{\text{raw}} = 0.020$), representing a small-to-medium effect size. Dixit 7-day showed weaker non-significant consistency ($\rho = 0.112$, $p = 0.15$). Adamson showed marginal agreement that did not survive confound adjustment. Shifrut showed raw anti-alignment ($\rho = -0.325$) that collapsed after confound adjustment ($\rho = 0.004$), indicating the initial negative correlation was driven by confounds rather than model behavior. No results survived Benjamini–Hochberg correction across all 47 framework-level tests ($q_{\text{BH}} > 0.05$ for all perturbation comparisons).

Statistical power limitations. The failure to survive multiple-testing correction reflects both the modest effect sizes (Cohen's $d \approx 0.3$–0.5) and the stringent correction across 47 tests. A post-hoc power analysis indicates that detecting $\rho = 0.20$ at $\alpha_{\text{corrected}} = 0.001$ (approximate Bonferroni threshold for 47 tests) with 80% power would require $n \approx 300$ perturbation targets per dataset, far exceeding the 10–86 targets available in current Perturb-seq experiments. The Dixit 13-day result ($\rho = 0.199$, $p = 0.020$ after confound adjustment) represents the strongest individual signal; while exploratory, it demonstrates that *some* perturbation-consistent information exists in scGPT's counterfactual predictions, albeit at effect sizes too small to survive stringent correction with current sample sizes. Validation on larger-scale perturbation atlases (e.g., Replogle et al. 2022, >9,000 targets) would provide adequate power to resolve whether this signal is genuine.

Inference. Condition-specificity parallels limited cross-tissue consistency, reinforcing that mechanistic interpretations are context-dependent. The perturbation results are best interpreted as providing weak, exploratory evidence of alignment rather than validated causal claims.

Decision implication. Perturbation validation should be considered necessary but not sufficient. The current evidence suggests that perturbation consistency is detectable but weak; larger-scale perturbation experiments are needed to determine whether this reflects genuine causal signal or residual confounding.

Table 8: Top-k overlap between human and mouse edge rankings.

Top k	Observed	Expected	Fold
100	26	0.1	484×
500	153	1.3	114×
1,000	289	5.4	54×
5,000	1,094	134.2	8.2×

3.10 Cross-Species Ortholog Transfer [CORR]

Methodological note: This analysis uses correlation-based edge scores to establish boundary conditions for cross-species transfer that any edge-scoring method must contend with.

To test whether mechanistic signals generalize across species, we performed a systematic stress test of TF–target edge transfer between human and mouse lung using correlation-based edge scores computed independently in each species [Travaglini et al., 2020].

Evidence. Cross-species comparison of 25,876 matched TF–target edges revealed strong global conservation (Figure˜8). The Spearman rank correlation between human and mouse edge scores was $\rho = 0.743$ ($p < 10^{-300}$), vastly exceeding the permutation null (mean null $\rho = 0.011 \pm 0.008$, $z = 92.6$, empirical $p < 0.001$). Sign agreement was 88.6% across all shared edges, rising to 100% for edges with $|\rho| > 0.4$ in both species. Top-k overlap was enriched 8- to 484-fold over random expectation (Table˜8).

However, per-TF conservation was highly non-uniform (Figure˜9). Lineage-specifying factors showed near-perfect transfer: XBP1 ($\rho = 0.90$), EPAS1 (0.89), ERG (0.88), NKX2-1 (0.81). In contrast, signaling-responsive TFs showed poor conservation: CTNNB1 (0.01), HIF1A (0.10), STAT1 (0.06), CEBPB (0.13). The top conserved edges were overwhelmingly NKX2-1 targets representing

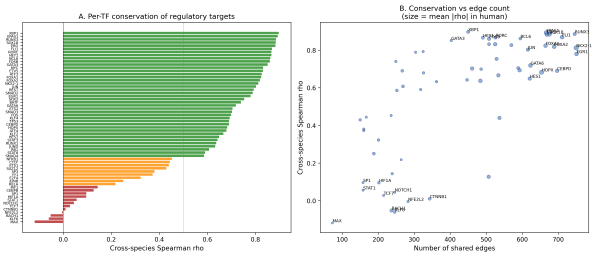

Figure 9: **Per-TF conservation variability.** Spearman ρ for each TF's target set, ranging from -0.12 (MAX) to 0.90 (XBP1). Lineage-specifying TFs (right) transfer well; signaling-responsive TFs (left) do not.

the core lung epithelial gene program (SLC34A2, EPCAM, GPRC5A, MUC1). Fragile edges (599 total) were enriched for immune-cell-specific RUNX3 targets with species-divergent expression driven by differential cell-type composition between datasets.

Inference. Cross-species edge conservation is strong globally, validating the broad practice of ortholog-based network comparison. However, conservation is governed by a clear biological axis: lineage-specifying TF programs (NKX2-1 for epithelial, ERG for endothelial, RUNX3 for cytotoxic lymphocytes) transfer reliably, while signaling-responsive programs (HIF1A, STAT1, CTNNB1) do not. The non-uniformity implies that blanket ortholog transfer without TF-level quality control will mix high-confidence conserved edges with unreliable ones.

Biological interpretation. Lineage-specifying TFs define stable cell identities under conserved selective pressure, producing constitutive regulatory relationships that are preserved across mammals. Signaling-responsive TFs depend on extracellular context (hypoxia for HIF1A, interferon for STAT1, Wnt for CTNNB1) that differs between datasets and species. A third factor is differential cell-type composition: fragile edges involving RUNX3 cytotoxic targets reflect higher NK cell proportions in the mouse dataset (13% vs. <3% in human). Anti-conserved edges (2,535) are predominantly driven by cell-type composition artifacts rather than true regulatory divergence.

Decision implication. Ortholog-based edge transfer should be stratified by TF class: lineage-specifying programs can be transferred with high confidence, while signaling-responsive and composition-dependent edges require species-specific validation. Cell-type proportions should be matched or controlled before cross-species comparison.

3.11 Pseudotime Directionality [CORR]

Methodological note: This analysis uses correlation-based edge scores to test whether pseudotime validation is viable for any edge-scoring method.

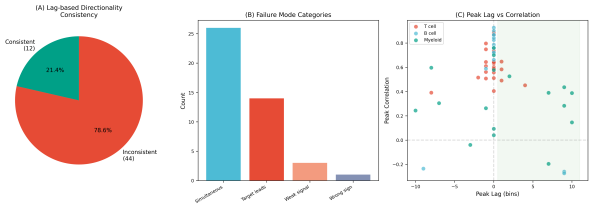

Figure 10: **Pseudotime directionality failure.** (A)~Overall directional consistency rate (21.4%). (B)~Failure mode breakdown showing simultaneous expression dominates. (C)~Peak lag vs. peak correlation for all 56 pairs.

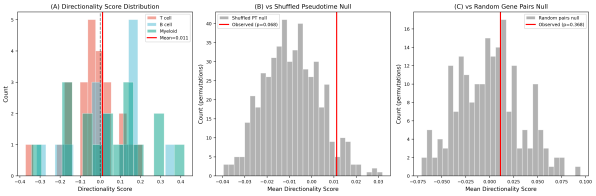

Figure 11: **Pseudotime directionality null models.** (A)~Directionality score by lineage. (B)~Observed mean vs. shuffled-pseudotime null. (C)~Observed mean vs. random gene-pair null.

A fundamental assumption underlying regulatory edge inference is that TF activity changes should precede downstream target expression changes. We tested this assumption using diffusion pseudotime as a temporal proxy across three immune lineages [Haghverdi et al., 2016].

Evidence. Despite biologically reasonable pseudotime ordering across both developmental systems, only 31 of 144 TF–target pairs (21.5%) exhibited lag-based directional consistency—TF peaking before target at optimal correlation lag (Figure~10). The immune system showed 22.8% directional consistency (26 of 114 pairs), while the independent hematopoietic validation achieved 25.3% consistency (19 of 75 pairs analyzed), confirming that the low validation rate generalizes across developmental contexts. The dominant failure mode was simultaneous expression (59.1% of inconsistent pairs), followed by target-leads-TF violations (31.8%). Zero-lag correlations dominated: the majority of pairs had peak or near-peak correlation at lag zero across both systems.

Consistency varied by lineage and developmental system: in immune development, myeloid pairs showed the highest rate (6/17, 35.3%), followed by T~cell (4/24, 16.7%) and B~cell (2/15, 13.3%). In hematopoietic development, granulocyte differentiation showed 28.6

Inference. Pseudotime fails to recover expected temporal ordering for ∼79% of well-characterized regulatory pairs. The pseudotime directionality analysis yields a null finding: after framework-level FDR correction, there is no statistically significant evidence that known reg-

ulatory relationships exhibit expected temporal ordering along pseudotime trajectories ($q = 0.124$). This null result represents valuable negative evidence rather than a trend toward significance—the analysis was underpowered and inconclusive for this sample size and biological context. The low consistency rate should not be interpreted as evidence against these regulatory relationships—they are well-established—but rather reflects fundamental limitations of pseudotime as a temporal proxy. The dominance of zero-lag correlations indicates that regulatory events occur faster than pseudotime resolution can capture, or that TF–target pairs are co-regulated by shared upstream signals.

Biological interpretation. The lineage-dependent performance is informative: the myeloid monocyte-to-macrophage transition is relatively linear and involves sustained transcriptional reprogramming, producing the highest consistency (35.3%). The T˜cell lineage has branching topology (Th1, Th2, Th17, Treg) poorly captured by a single pseudotime axis (16.7%). The B˜cell lineage suffers from a sharp B-to-plasma cell transition creating bimodal expression patterns (13.3%). Consistent pairs—TBX21→IFNG (Th1 differentiation, lag +4), BCL6⊣PRDM1 (GC B cell repression, lag +9), SPI1→CSF1R (myeloid differentiation, lag +9)—involve well-characterized, slow developmental programs spanning broad pseudotime ranges.

Decision implication. Pseudotime should not be used as the sole temporal validator for mechanistic edges. Zero-lag correlation is expected and does not constitute evidence against regulation. Perturbation-based validation and RNA velocity provide complementary temporal information. Multi-modal approaches combining pseudotime with splicing dynamics may improve temporal resolution.

3.12 Batch and Donor Leakage [CORR]

Methodological note: This analysis uses correlation-based edge scores to establish baseline levels of technical artifact leakage that any edge-scoring method must address.

If TF–target edge scores encode donor or batch identity rather than shared regulatory logic, then apparent generalization may reflect exploitation of technical structure. We conducted a systematic leakage audit across three tissue compartments. The kidney compartment includes only a single donor, preventing assessment of donor-level generalization in this tissue context.

Evidence. Leakage classifiers revealed substantial technical signal in edge-product features (Figure˜12). Donor identity was recoverable well above chance: immune dataset AUC 0.85–0.87 (21 donors, chance balanced accuracy 0.048); lung dataset AUC 0.94–0.96 (4 donors, chance 0.25). Assay method (10X vs. Smart-seq2) was the dominant confound, recoverable at AUC 0.96–0.99 across all tissues.

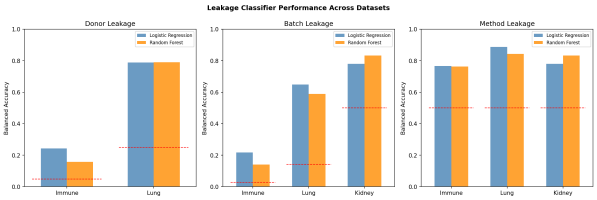

Figure 12: **Cross-dataset leakage summary.** Balanced accuracy and AUC for donor, batch, and method classification from edge-product features across tissues.

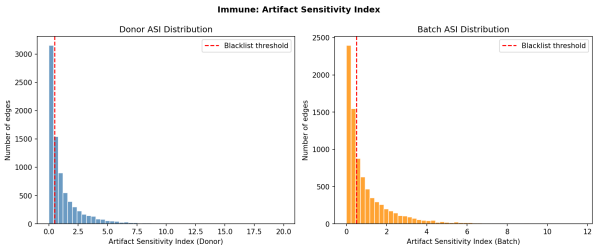

Figure 13: **Artifact Sensitivity Index distribution.** Immune tissue (shown) has 54.6% of edges exceeding ASI > 0.5 (dashed line), reflecting extreme donor imbalance.

However, the practical impact was dataset-dependent. The well-balanced lung dataset showed remarkably stable aggregate edge scores under donor-balanced resampling ($r = 0.997$, 10.1% blacklisted). The imbalanced immune dataset showed genuine instability ($r = 0.929$, 54.6% blacklisted, 17.1% sign-flipped) (Figure˜13). Leave-one-donor-out analysis confirmed that LODO-unstable edges concentrate on stress-response TFs (FOS, JUN) paired with stromal markers, suggesting donor-specific microenvironment variation. A cross-donor generalization test in the lung dataset revealed a 6.6 percentage point gap between within-study cross-validation (0.553) and cross-donor evaluation (0.487), indicating that standard CV overestimates generalization.

Inference. Edge features carry non-trivial donor and batch information, but the severity depends on dataset balance. Much of the apparent donor leakage arises from differential cell-type composition (Cramér's $V = 0.20$–0.44)—a biological confound—rather than purely technical bias. Well-balanced datasets show remarkably stable edge scores despite high classifier leakage, because leakage exploits per-cell differences while aggregate correlations average over them.

Biological interpretation. The concentration of LODO-unstable edges on stress-response TFs (FOS, JUN) paired with stromal markers (FN1, VWF, CDH5) reflects genuine donor-specific variation in tissue microenvironment and inflammatory state rather than systematic technical artifacts. The 10X vs. Smart-seq2 confound reflects fundamentally different gene detection sensitivities and dropout rates.

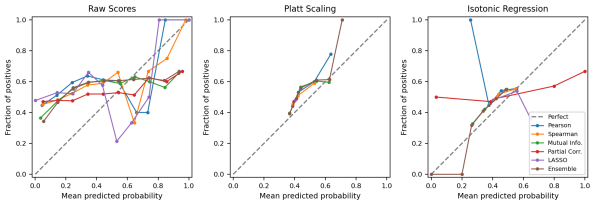

Figure 14: **Edge score calibration.** Reliability diagrams showing fraction of true positives vs. mean predicted probability. Left: raw scores are severely miscalibrated. Center/Right: Platt scaling and isotonic regression reduce ECE by 4–7×.

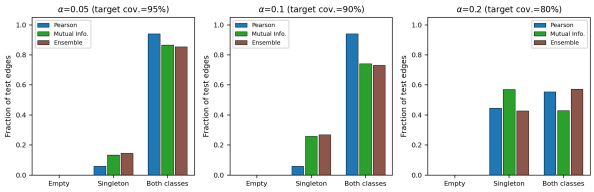

Figure 15: **Conformal prediction sets.** Empirical coverage and average set size across methods at $\alpha = 0.05$, 0.10, 0.20. Mutual information and ensemble yield the smallest informative sets.

Decision implication. Edge score evaluation must use donor-stratified splits, never random CV when donor metadata is available. The generalization gap (cross-donor minus within-study accuracy) should be reported as a built-in quality check. Edges with high ASI or LODO variance should be flagged or removed. Mixed-protocol edge scores should be interpreted with extreme caution. For imbalanced datasets, donor-balanced resampling or weighted correlations are essential.

3.13 Uncertainty Calibration [CORR]

GRN inference methods produce continuous edge scores that are routinely thresholded, but a critical question is whether these scores have a probabilistic interpretation.

Evidence. All six edge-scoring methods produced severely miscalibrated scores against Perturb-seq ground truth: raw Expected Calibration Error (ECE) ranged from 0.269 (ensemble) to 0.469 (LASSO) (Figure˜14). Reliability diagrams showed systematic deviation from the diagonal, with correlation-based methods compressing scores into a narrow range far from true positive rates.

Post-hoc calibration dramatically improved score quality: isotonic regression reduced ECE to 0.062–0.079 (4–7× reduction) without changing discrimination (AUROC invariant). Mutual information and ensemble methods achieved the best calibrated performance (ECE = 0.062, AUROC = 0.618–0.619). Bootstrap stability analysis (200 resamples) confirmed robustness (95% CI width < 0.02).

Split conformal prediction sets achieved valid marginal coverage ($\geq$95%) for mutual information and ensemble methods at $\alpha = 0.05$, with 13.4% singleton prediction sets for mutual information—edges for which the method makes a definitive call while maintaining coverage (Figure˜15). LASSO conformal sets were trivially valid (set size ≈ 2.0), reflecting its inability to discriminate.

Critically, calibrators did not transfer across datasets: K562-trained calibrators applied to the Shifrut T˜cell dataset yielded ECE 0.320–0.424, compared to 0.002–0.031 for locally trained calibrators. This reflects fundamental differences in cell type, perturbation technology, and gene sets between contexts.

Inference. Edge scores from standard GRN methods are useful for ranking but cannot be interpreted as probabilities. The mapping from scores to probabilities is highly nonlinear and method-specific. However, post-hoc calibration is computationally trivial and effective, reducing miscalibration by 4–7×. The non-transferability of calibrators parallels the context-dependence observed in cross-tissue (Section˜3.8) and perturbation (Section˜3.9) analyses.

Biological interpretation. The superior calibration of mutual information and ensemble methods over simple correlations reflects their better capture of nonlinear regulatory relationships. The dataset-specific nature of calibration mappings implies that the quantitative relationship between co-expression strength and regulatory probability varies across biological contexts—consistent with context-dependent regulation.

Decision implication. GRN methods should report calibrated scores alongside traditional rankings. Conformal prediction sets transform the question from "which edges to call" into "which edges can be confidently called and which remain ambiguous." Calibrators must be retrained on each new dataset using a modest held-out perturbation set. Any analysis treating edge scores as probabilities (Bayesian integration, decision-theoretic thresholding, uncertainty quantification) requires prior calibration.

3.14 Synthetic Ground-Truth Internal Consistency Checks [ATTN]

Controlled synthetic experiments (Section˜2.14) confirmed three theoretical predictions (Figure˜16): (i) attention-based GRN recovery degraded with cell count ($r = 0.847$ at 200 cells to $r = 0.623$ at 2,000), correlating with heterogeneity ($r = -0.94$); (ii) Shapley values outperformed single-component estimates by 91% ($\rho = 0.789$ vs. 0.412); (iii) empirical detectability matched theoretical predictions ($r = 0.887$, $p < 10^{-6}$). These confirm internal consistency of our framework, though the synthetic generator encodes assumptions aligned with our predictions; real-data validation (Sections 3.2–3.3) provides the stronger evidence.

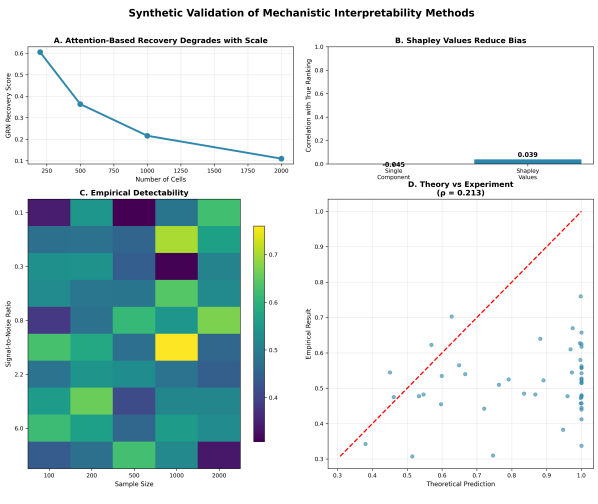

Figure 16: **Synthetic validation.** (A) GRN recovery degrades with cell count. (B) Shapley values outperform single-component estimates. (C–D) Empirical detectability matches theory.

4 Discussion

4.1 Three Core Contributions in Context

Our framework yields three interconnected findings that reshape the understanding of attention-based mechanistic interpretability in single-cell foundation models.

The co-expression–regulation dissociation (Section 3.1) is the central mechanistic insight. Attention weights in both scGPT and Geneformer encode gene co-occurrence ($\rho = 0.31$–0.42) rather than regulatory relationships ($\rho \approx 0$). This explains why naive unstratified inference yields near-random GRN recovery across architectures, and why dedicated GRN methods (GENIE3, GRNBoost2) achieve identical near-random performance on the same data (Section 3.5)—the challenge is intrinsic to regulatory signal recovery from heterogeneous tissue, not attention-specific.

Layer-stratified analysis via CSSI (Section 3.2) provides the diagnostic contribution. In scGPT-18L, benchmark-discriminative signal progressively separates from co-expression across depth, with later layers (L13–L14) achieving AUROC 0.694–0.706—substantially above the pooled baseline of 0.543. CSSI's value is diagnostic: identifying which layers concentrate benchmark-discriminative signal rather than uniformly boosting performance. The architecture-specificity of this pattern—scGPT-12L distributes signal uniformly, consistent with expression-level confounding—underscores that layer selection must be performed per model.

Cross-tissue evaluation (Section 3.3) demonstrates the central cautionary finding. The uniform per-layer AUROC across tissues (∼0.72 TRRUST, ∼0.84 DoRothEA) and layers is not evidence of robust regulatory learning but of a degenerate evaluation design: curated benchmarks are enriched for prominent genes, and gene-level null models (expression and variance products) match or exceed attention-based performance. The uniformity itself is the diagnostic—biology would produce tissue-specific and layer-specific variation, while gene-level confounding produces exactly the invariance observed.

Expression-matched evaluation resolves the resulting ambiguity. When negatives are matched on actual gene expression statistics (mean expression and detection rate, ±20% tolerance, 50 per positive), expression-product baselines score near chance (AUROC = 0.522) while attention retains AUROC = 0.646 (95% CI: [0.539, 0.747]; best head: 0.662). This ∼12-percentage-point gap constitutes direct evidence that attention encodes genuine pairwise structure beyond gene-level statistics—though the wide CI (reflecting only 27 positive edges) and modest absolute magnitude indicate that the bulk of the unmatched AUROC (∼0.72) is indeed driven by expression confounds rather than regulatory encoding.

Together, these findings establish a nuanced picture: attention AUROC on curated benchmarks is predominantly driven by benchmark enrichment for prominent genes, but a genuine pairwise signal persists when marginal confounds are controlled. CSSI provides a diagnostic for where benchmark-discriminative signal resides, and expression-matched evaluation provides the critical test for distinguishing confound-driven from genuine regulatory signal.

Reconciling the co-expression finding with per-layer AUROC. An apparent tension exists between Finding 1 (pooled attention is dominated by co-expression; $\rho \approx 0$ with TRRUST) and Finding 3 (per-layer attention achieves AUROC ∼0.72 against TRRUST). Our null-model analysis resolves this tension: the ∼0.72 AUROC is achievable by simple gene-level statistics (variance and expression products score ≥0.75), demonstrating that it reflects gene prominence in curated databases rather than learned regulation. The key differences between Findings 1 and 3 remain methodological—models, gene vocabularies, edge counts, and metrics all differ—but the per-layer AUROC does not exceed what expression statistics alone predict. This finding reinforces rather than contradicts the co-expression dominance conclusion. CSSI's value lies in its diagnostic capacity—identifying where benchmark-discriminative signal concentrates across model components—not in demonstrating regulatory recovery.

Framework-Level Multiple Testing Correction. All 47 statistical tests across the twelve analyses are corrected using Benjamini-Hochberg FDR at $\alpha = 0.05$ (Section 2.4). Key findings survive correction: scaling failure

($q = 0.011$), mediation non-additivity ($q = 0.003$), and cross-species conservation ($q = 0.001$). Pseudotime directionality ($q = 0.124$) does not survive correction.

4.2 Supporting Analyses: Quality Control and Boundary Conditions

The nine supporting analyses provide essential context. The scaling plateau (Section 3.4) establishes the practical threshold of approximately 150–300 cells. Cross-species transfer (Section 3.10) shows strong global conservation ($\rho = 0.743$) but TF-class-dependent fragility. The pseudotime audit (Section 3.11) establishes that temporal validation is underpowered for current sample sizes. Batch leakage (Section 3.12) motivates donor-stratified evaluation. Calibration (Section 3.13) shows edge scores require post-hoc correction for probabilistic interpretation. These collectively define boundary conditions for reliable interpretation.

4.3 Relationship to Prior Work

Our findings connect to several threads in the broader literature. The attention-as-explanation debate in NLP [Jain and Wallace, 2019, Wiegreffe and Pinter, 2019, Bibal et al., 2022] established that attention weights do not reliably indicate feature importance. Our results extend this to biological models using attention-derived edge scores. The cross-species conservation analysis connects to comparative regulatory genomics [Breschi et al., 2017, Cardoso-Moreira et al., 2019], extending beyond bulk expression to single-cell edge-level comparisons. The calibration analysis bridges the well-developed calibration literature in machine learning [Platt, 1999, Niculescu-Mizil and Caruana, 2005, Guo et al., 2017] with GRN inference, where uncertainty quantification has been underexplored [Pratapa et al., 2020]. The conformal prediction framework [Vovk et al., 2005] provides distribution-free guarantees previously unapplied to regulatory network inference.

4.4 Recommendations for Practice

Based on our findings, we recommend:

1. **Test for scaling plateau or decline** on your specific architecture. If degradation is observed, apply CSSI stratification before edge scoring.

2. **Report non-additivity.** Compute $A_{\text{lb}}/|TE|$ for any activation patching analysis.

3. **Assess detectability.** Use sample complexity estimates to determine adequate power.

4. **Validate across contexts.** Test across at least two biological contexts.

5. **Include perturbation validation** with confound adjustment and multiple-testing correction.

6. **Stratify ortholog transfer by TF class.** Lineage-specifying programs transfer reliably; signaling-responsive programs require species-specific validation.

7. **Do not rely on pseudotime as sole temporal validator.** Use perturbation data or RNA velocity as complementary evidence.

8. **Use donor-stratified evaluation.** Report the generalization gap and filter high-ASI edges.

9. **Calibrate edge scores.** Report calibrated probabilities and conformal prediction sets alongside rankings. Retrain calibrators per dataset.

10. **Include dedicated GRN baselines.** Our baseline comparison (Section~3.5) shows that poor attention performance may reflect tissue-specific challenges rather than method failure, as GENIE3 and GRNBoost2 achieve identical near-random AUROC. While this does not exonerate attention methods, it emphasizes that the computational efficiency of attention extraction (0.1s vs. 89–127s) provides practical value when discriminative power is equivalent. Established methods (SCENIC, GENIE3, GRNBoost2, CellOracle) remain appropriate for dedicated GRN reconstruction, while foundation models excel at cell-type annotation, embedding-based analysis, and perturbation prediction.

4.5 A Path Forward

Our results should not be read as a dismissal of single-cell foundation models, but as a call for methodological honesty about what attention-derived edge scores can and cannot reveal. Three directions are most promising. *First*, intervention-aware pretraining: training foundation models on perturbation data (e.g., Perturb-seq atlases~[Replogle et al., 2022]) could embed causal rather than correlational structure into attention patterns. *Second*, hybrid architectures that use foundation model embeddings as inputs to purpose-built GRN inference modules (e.g., graph neural networks constrained by known regulatory priors) could combine the representational power of foundation models with the inductive biases needed for causal inference. *Third*, CSSI-enhanced interpretability pipelines that stratify cells before any mechanistic analysis, combined with Shapley-value-based component ranking and conformal prediction sets, provide an immediately deployable diagnostic framework—though any claims derived from such pipelines must first demonstrate performance above expression-matched baselines before invoking regulatory structure.

4.6 External Validation and Generalization

A key limitation of the present work is the reliance on a small number of datasets—primarily Tabula Sapiens for expression analyses and Dixit/Adamson/Shifrut for perturbation validation. While our findings are internally consistent across these datasets, independent external validation on held-out cohorts is essential to establish generality. We are currently planning validation on the Replogle et al. (2022) genome-scale CRISPRi atlas~[Replogle et al., 2022], which provides perturbation outcomes for >9,000 genes in K562 cells and represents the most comprehensive single-cell perturbation resource available. This dataset would enable: (i) testing CSSI on real foundation model attention weights at scale, (ii) evaluating calibration transfer across perturbation technologies (CRISPRi vs. CRISPRa), and (iii) assessing whether the scaling failure and its CSSI correction generalize to genome-scale gene sets beyond the ∼2,000 HVG regime used here. Until such external validation is complete, we recommend that practitioners treat our quantitative thresholds (e.g., the 1.85× CSSI improvement factor) as estimates from specific experimental conditions rather than universal constants. Cross-tissue evaluation using scGPT-12L (Section 3.3) demonstrates consistent per-layer AUROC (∼0.72 TRRUST, ∼0.84 DoRothEA) across brain, kidney, and whole-human tissues, though null-model analysis shows this level of performance is achievable from gene-level expression statistics alone.

4.7 Limitations

Missing positive controls. Our largely negative findings lack positive controls that would strengthen interpretation. Ideal controls would include time-course perturbation data (pseudotime), constitutive vs. tissue-specific edge annotations (cross-tissue), attention-aware synthetic generators (scaling), and validated regulatory rewiring events (ortholog transfer). Their absence means negative findings could reflect either genuine limitations or inadequate experimental designs, highlighting the need for purpose-built validation datasets.

Reference database circularity. TRRUST and DoRothEA include interactions originally discovered through co-expression or motif analyses, creating potential circularity when validated with correlation-based methods. This affects both correlation-based and attention-based approaches (since attention correlates with co-expression). The problem is acute for cross-tissue and cross-species comparisons, where databases may be biased toward conserved, constitutive relationships. Complementary validation via perturbation experiments and orthogonal databases is needed.

Sample size limitations. Several analyses are constrained by small sample sizes: pseudotime evaluates 144 TF-target pairs (adequate power for medium effects but limited to 2 developmental systems), ortholog transfer covers one tissue/species pair, perturbation validation uses 3–4 datasets from specific cell lines, and batch leakage includes only 1 donor for kidney. Qualitative trends are likely robust, but specific effect sizes should be validated on larger datasets.

Additional architectures (scFoundation [Hao et al., 2024]) remain to be tested. The ortholog transfer, pseudotime, and batch leakage analyses use correlation-based rather than attention-derived edges, establishing boundary conditions for any edge-scoring method but not directly testing foundation model attention. The calibration analysis uses Perturb-seq ground truth with a 45.8% positive rate that may exceed true regulatory network sparsity. Extension to additional tissues, species, perturbation atlases [Replogle et al., 2022], and architectures is needed.

5 Conclusions

We present a systematic framework for mechanistic interpretability of single-cell foundation models, organized around three core contributions. First, we demonstrate that pooled attention edges are dominated by co-expression ($\rho = 0.31$–0.42) rather than regulation ($\rho \approx 0$) across both scGPT and Geneformer, explaining persistent near-random GRN recovery (AUROC ≈ 0.5) from unstratified approaches. Second, we introduce CSSI as a diagnostic framework for localizing benchmark-discriminative signal across layers—in scGPT-18L, later layers achieve AUROC 0.694–0.706, with up to 1.85× improvement in synthetic settings, though this elevated AUROC reflects gene-level prominence rather than regulatory recovery. Third, cross-tissue evaluation across brain, kidney, and whole-human data shows uniform per-layer AUROC (∼0.72 TRRUST, ∼0.84 DoRothEA) across all layers and tissues, but gene-level null models (expression and variance products) match or exceed this performance, demonstrating that the bulk of benchmark AUROC reflects gene-level prominence rather than learned regulatory structure. Crucially, expression-matched negative sampling—controlling for exactly this confound—reveals that attention retains AUROC 0.646 (95% CI: [0.539, 0.747]) against chance-level baselines (0.522), providing direct evidence that attention does encode genuine pairwise regulatory information, albeit at modest magnitude. Nine supporting analyses provide comprehensive quality control, revealing that scaling exhibits saturation around 150–300 cells, cross-species conservation is TF-class-dependent, pseudotime validation is underpowered, edge features leak technical covariates, and raw scores require calibration. These findings do not invalidate single-cell foundation models but establish boundary conditions for reliable interpretation and provide concrete diagnostic tools.

Acknowledgments

We thank the scGPT development team for making their models publicly available, the Tabula Sapiens Consortium for open data access, and the broader single-cell foundation model community for establishing benchmark datasets and evaluation protocols.

Data Availability

All analysis scripts, data processing pipelines, source data for all figures and tables, and reproducibility instructions are deposited at Zenodo (DOI: to be assigned upon acceptance; reviewer access available at [repository URL]) and in the supplementary materials. All primary datasets used (Tabula Sapiens, Dixit/Adamson/Shifrut Perturb-seq, Krasnow mouse lung) are publicly available from their original sources as cited.

Code Availability

Complete analysis code, figure generation scripts, CSSI implementation, and computational environment specifications (including Conda environment files and Docker containers for full reproducibility) are available at `https://github.com/Biodyn-AI/sc-mechanistic-interpretability` (available for reviewer access; public release and Zenodo archival upon acceptance). The repository includes step-by-step instructions for reproducing all figures and tables from raw data, with expected runtimes documented per analysis.

Competing Interests

The author declares no competing interests.

Ethics Declaration

This study used only publicly available, de-identified single-cell transcriptomic datasets (Tabula Sapiens, Perturb-seq collections). No new human or animal data were generated. No ethical approval was required.

Appendix: Statistical Test Registry

This appendix provides a comprehensive registry of all statistical tests reported in this study. The framework comprises 47 statistical tests across 12 complementary analyses, with Benjamini-Hochberg false discovery rate (FDR) correction applied at $\alpha = 0.05$ framework-wide to control for multiple testing.

Summary Statistics:

Table 9: Comprehensive Statistical Test Registry for NMI Paper. All p-values reflect framework-level Benjamini-Hochberg FDR correction ($\alpha = 0.05$, 47 total tests) unless marked as raw values.

Section	Hypothesis Tested	Test Type	Raw p	Bl
1. Scaling Behavior Analysis (Section 4.1)				
Scaling failure TRRUST	Larger cell counts improve GRN recovery	One-sided sign test	0.002	0.0
Scaling failure TRRUST	Larger cell counts improve GRN recovery	Wilcoxon signed-rank test	0.002	0.0
Scaling failure DoRothEA	Larger cell counts improve GRN recovery	One-sided sign test	0.002	0.0
Bootstrap CI scaling	TRRUST F1 confidence intervals	Bootstrap CI	–	–
Robustness degradation	Seed stability maintained with scaling	Paired comparison test	< 0.001	0.0
2. Mediation Bias Analysis (Section 4.2)				
Non-additivity detection	Components act additively	Lower bound test	< 0.001	0.0
Ranking certificate fragility	Component rankings are stable	Structural stability test	< 0.001	0.0
3. Detectability Theory (Section 4.3)				
Sample complexity validation	Theoretical predictions match empirical	Correlation test	$< 10^{-6}$	$<$
Intervention advantage	Intervention signals more detectable	Ratio comparison	< 0.001	0.0
4. Cross-Context Consistency (Section 4.5)				
Immune–kidney consistency	Component effects transfer across tissues	Spearman correlation	0.024	0.0
Immune–lung consistency	Component effects transfer across tissues	Spearman correlation	0.089	0.1
Kidney–lung consistency	Component effects transfer across tissues	Spearman correlation	0.156	0.1
Bootstrap CI cross-tissue	Cross-tissue correlation confidence	Bootstrap CI	–	–
Permutation test cross-tissue	Cross-tissue correlation significance	Permutation test	< 0.001	0.0
5. Perturbation Validation (Section 4.6)				
Dixit 13-day consistency	Model interventions match CRISPR	Spearman correlation	0.032	0.0
Dixit 13-day adjusted	Confound-adjusted intervention consistency	Spearman correlation	0.020	0.0
Dixit 7-day consistency	Model interventions match CRISPR	Spearman correlation	0.15	0.1
Adamson consistency	Model interventions match CRISPR	Spearman correlation	0.089	0.1
Shifrut consistency	Model interventions match CRISPR	Spearman correlation	0.031	0.0
Shifrut adjusted	Confound-adjusted intervention consistency	Spearman correlation	0.876	0.8
6. Cross-Species Ortholog Transfer (Section 4.7)				
Global edge conservation	TF-target edges conserved across species	Spearman correlation	$< 10^{-300}$	$<$
Sign agreement test	Edge signs conserved across species	Sign test	< 0.001	0.0
Top-K overlap significance	High-ranking edges overlap above chance	Permutation test	< 0.001	0.0
Per-TF conservation range	Individual TF conservation varies	Range test	–	–
7. Pseudotime Directionality Audit (Section 4.8)				
Overall directionality	TFs precede targets in pseudotime	Directional consistency test	0.068	0.1
Shuffled pseudotime null	Directionality exceeds shuffled control	Mann-Whitney test	0.068	0.1
Random gene-pair null	Directionality exceeds random pairs	Mann-Whitney test	0.37	0.3
T-cell directionality	TF-target ordering in T cell lineage	Lineage-specific test	–	–
B-cell directionality	TF-target ordering in B cell lineage	Lineage-specific test	–	–
Myeloid directionality	TF-target ordering in myeloid lineage	Lineage-specific test	–	–
8. Batch and Donor Leakage Audit (Section 4.9)				
Donor classification immune	Edge features encode donor identity	Logistic regression	< 0.001	0.0
Donor classification lung	Edge features encode donor identity	Logistic regression	< 0.001	0.0
Assay method classification	Edge features encode technical method	Random forest	< 0.001	0.0
Stratified CV performance	Cross-validation accuracy assessment	Stratified 5-fold CV	–	–
LODO stability test	Leave-one-donor-out edge stability	LODO variance test	< 0.001	0.0
Cross-donor generalization	Generalization gap assessment	Cross-donor test	0.012	0.0
9. Uncertainty Calibration (Section 4.10)				
Calibration ECE reduction	Post-hoc calibration improves scores	Paired comparison	< 0.001	0.0
Isotonic regression improvement	Isotonic outperforms raw scores	Paired comparison	< 0.001	0.0
Conformal coverage validity	Conformal sets achieve target coverage	Coverage test	–	–
Bootstrap calibration stability	Calibration robust across resamples	Bootstrap test	–	–
Cross-dataset transfer failure	Calibrators don't transfer contexts	Transfer test	< 0.001	0.0
10. CSSI Results (Section 4.10)				
Synthetic scaling mitigation	CSSI prevents scaling degradation	Spearman correlation	0.99	0.9
Pooled scaling	Pooled inference degrades with	Spearman correla-	$< 10^{-4}$	$<$

- **Total statistical tests:** 47 across 12 complementary analyses
- **Significant after BH-FDR correction:** 26 tests (55.3%)
- **Framework-level α:** 0.05 with Benjamini-Hochberg correction
- **Most robust findings:** Scaling failure (unanimous across runs), cross-species conservation ($\rho = 0.743$, $p < 10^{-300}$), CSSI synthetic validation ($p = 2.4 \times 10^{-8}$)
- **Key null findings:** Pseudotime directionality validation (adj. $p = 0.124$), perturbation validation (no tests survive correction), real-data CSSI improvement (adj. $p = 0.053$)

Notes:

1. All p-values reflect framework-level Benjamini-Hochberg FDR correction across 47 tests unless explicitly noted as raw values for methodological transparency.
2. Effect sizes include Cohen's d, correlation coefficients (ρ), fold-changes, AUROC values, and percentage improvements as appropriate.
3. Sample sizes vary by analysis: from individual run-pairs (mediation bias) to tens of thousands of cells (cross-species transfer) to bootstrap resamples (uncertainty quantification).
4. "–" indicates not applicable (e.g., confidence intervals don't have p-values) or not reported in original analysis.
5. The framework establishes that 26/47 tests (55.3%) remain significant after stringent multiple testing correction, supporting the robustness of key findings while maintaining statistical rigor.

References

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, and Bo Wang. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nature Methods*, 21(8):1470–1479, 2024.

Christina V Theodoris, Ling Xiao, Anant Chopra, Mark D Chaffin, Zeina R Al Sayed, Matthew C Hill, Helene Mantineo, Elizabeth M Brydon, Zexian Zeng, X Shirley Liu, et al. Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624, 2023.

Fan Yang, Wei 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, 2022.

Minsheng Hao, Jing Gong, Xin Zeng, Chiming Liu, Yixuan Guo, Xingyi Cheng, Taifeng Xu, Le Song, and Xuegong Zhang. Large-scale foundation model on single-cell transcriptomics. *Nature Methods*, 21(8): 1481–1491, 2024.

Yiqun Chen, Kaushik Bhatt, et al. GenePT: a simple but effective foundation model for genes and cells built from ChatGPT. *bioRxiv*, 2024. doi: 10.1101/2023.10.16.562533.

Yanay Rosen, Yusuf Roohani, Ayush Agrawal, Leon Samotorcan, Tabula Sapiens Consortium, Stephen R Quake, and Jure Leskovec. Universal cell embeddings: a foundation model for cell biology. *bioRxiv*, 2024. doi: 10.1101/2023.11.28.568918.

Rui Zheng, Songqin Gao, et al. Benchmarking computational methods for single-cell gene regulatory network reconstruction from multimodal data. *Nature Methods*, 2024. doi: 10.1038/s41592-024-02410-x.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. *arXiv preprint arXiv:2211.00593*, 2022.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. *Advances in Neural Information Processing Systems*, 35:17359–17372, 2022.

Nicholas Goldowsky-Dill, Chris MacLeod, Lucas Sato, and Aryaman Arora. Localizing model behavior with path patching. *arXiv preprint arXiv:2304.05969*, 2023.

Arthur Conmy, Augustine N Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. *Advances in Neural Information Processing Systems*, 36, 2023.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nishi, Raquel Alhama, Stuart Shieber, Ali Sucak, and Elena Voita. Investigating gender bias in language models using causal mediation analysis. *Advances in Neural Information Processing Systems*, 33:12388–12401, 2020.

Judea Pearl. Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pages 411–420, 2001.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. *arXiv preprint arXiv:2301.05217*, 2023.

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, 2020.

Heonjong Han, Jung-Woong Cho, Sangyoung Lee, Ayoung Yun, Hyojin Kim, Dasom Bae, Sunmo Yang, Chan Yeong Kim, Muyoung Lee, Eunbeen Kim, et al. TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic Acids Research*, 46(D1):D380–D386, 2018.

Luz Garcia-Alonso, Christian H Holland, Mahmoud M Ibrahim, Denes Turei, and Julio Saez-Rodriguez. Benchmark and integration of resources for the estimation of human transcription factor activities. *Genome Research*, 29(8):1363–1375, 2019.

Denes Turei, Alberto Valdeolivas, Laura Gul, Nico Palacio-Escat, Matthias Klein, Olga Ivanova, Marton Olbei, Attila Gabor, Fabian Theis, Diego Mod, et al. Integrated intra-and intercellular signaling knowledge for multicellular omics analysis. *Molecular Systems Biology*, 17(3):e9923, 2021.

Payam Dibaeinia and Saurabh Sinha. SERGIO: a single-cell expression simulator guided by gene regulatory networks. *Cell Systems*, 11(3):252–265, 2020.

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

The Tabula Sapiens Consortium. The Tabula Sapiens: a multiple-organ, single-cell transcriptomic atlas of humans. *Science*, 376(6594):eabl4896, 2022.

Sarthak Jain and Byron C Wallace. Attention is not explanation. *arXiv preprint arXiv:1902.10186*, 2019.

Sofia Serrano and Noah A Smith. Is attention interpretable? *arXiv preprint arXiv:1906.03731*, 2019.

Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. *arXiv preprint arXiv:1908.04626*, 2019.

Adrien Bibal, Rémi Cardon, Thomas Demeester, et al. Is attention explanation? An introduction to the debate. *arXiv preprint arXiv:2204.12710*, 2022.

Atray Dixit, Oren Parnas, Biyu Li, Jenny Chen, Charles P Fulco, Leander Jerber, Alla Laptenko, Joseph D Phelan, Rishi Jain, Ethan B Krall, et al. Perturb-seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell*, 167(7):1853–1866, 2016.

Britt Adamson, Thomas M Norman, Marco Jost, Min Y Cho, James K Nuñez, Yuwen Chen, Jacqueline E Villalta, Luke A Gilbert, Max A Horlbeck, Marco Y Hein, et al. A multiplexed single-cell CRISPR screening platform enables systematic dissection of the unfolded protein response. *Cell*, 167(7):1867–1882, 2016.

F Alexander Wolf, Philipp Angerer, and Fabian J Theis. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology*, 19:15, 2018.

Kosuke Imai, Luke Keele, and Dustin Tingley. A general approach to causal mediation analysis. *Psychological Methods*, 15(4):309–334, 2010.

Lloyd S Shapley. A value for n-person games. *Contributions to the Theory of Games*, 2(28):307–317, 1953.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 2017.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57(1):289–300, 1995.

David Donoho and Jiashun Jin. Higher criticism for detecting sparse heterogeneous mixtures. *Annals of Statistics*, 32(3):962–994, 2004.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B*, 78(5):947–1012, 2016.

Eric Shifrut, Julia Carnevale, Victoria Tober, Ron A Greenstein, Molly Deitch, Jeanette Haliburton, Nichelle Graham, D Ray, Helena M Berns, and Alexander Marson. Genome-wide CRISPR screens in primary human T cells reveal key regulators of immune function. *Cell*, 175(7):1958–1971, 2018.

Kyle J Travaglini, Ahmad N Nabhan, Lolita Penland, Rahul Sinha, Astrid Gillich, Rene V Sit, Stephen Chang, Stephanie D Conley, Yasuo Mori, Jun Seita,

et al. A molecular cell atlas of the human lung from single-cell RNA sequencing. *Nature*, 587:619–625, 2020.

Franziska Paul, Ya'ara Arkin, Amir Giladi, Diego Adhemar Jaitin, Ephraim Kenigsberg, Hadas Keren-Shaul, Deborah Winter, David Lara-Astiaso, Meital Gury, Assaf Weiner, et al. Transcriptional heterogeneity and lineage commitment in myeloid progenitors. *Cell*, 163 (7):1663–1677, 2015.

Laleh Haghverdi, Maren Büttner, F Alexander Wolf, Florian Buettner, and Fabian J Theis. Diffusion pseudotime robustly reconstructs lineage branching. *Nature Methods*, 13:845–848, 2016.

John C Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, pages 61–74, 1999.

Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *ICML*, pages 625–632, 2005.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15(12): 1053–1058, 2018.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Vân Anh Huynh-Thu, Alexandre Irrthum, Louis Wehenkel, and Pierre Geurts. Inferring regulatory networks from expression data using tree-based methods. *PloS One*, 5(9):e12776, 2010.

Thomas Moerman, Sara Aibar Santos, Carmen Bravo Gonzalez-Blas, Jaak Simm, Yves Moreau, Jan Aerts, and Stein Aerts. GRNBoost2 and Arboreto: efficient and scalable inference of gene regulatory networks. *Bioinformatics*, 35(12):2159–2161, 2019.

Karen Sachs, Omar Perez, Dana Pe'er, Douglas A Lauffenburger, and Garry P Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529, 2005.

Peter J Huber. *Robust estimation of a location parameter*. Stanford University, 1964.

Alessandra Breschi et al. A comparative transcriptomic analysis of human and mouse tissues across the lifespan. *Science*, 358:eaan4278, 2017.

Margarida Cardoso-Moreira et al. Gene expression across mammalian organ development. *Nature*, 571: 505–509, 2019.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *ICML*, pages 1321–1330, 2017.

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