Causal Mediation Circuit Maps for scGPT: Cross-Tissue Evidence and Limits

Ihor Kendiukhov
University of Tuebingen, Computer Science Department
kenduhov.ig@gmail.com

February 2026

Abstract

Mechanistic interpretability for single-cell foundation models is still dominated by aggregate edge-level metrics, leaving uncertain which internal components mediate specific regulatory effects. We present a mediation-based circuit mapping framework for scGPT that decomposes transcription-factor (TF) to target effects into attention-head and MLP mediator contributions using intervention-plus-patching experiments. We analyze Tabula Sapiens kidney and lung, plus an immune subset, with a discovery-to-refinement protocol that separates pair-support coverage from expensive tracing. Discovery coverage quantifies the bottleneck directly: in a random 132-pair panel, no TF-target pair has support in all three tissues at any tested cell-count threshold; in a fixed 40-pair replay, only one pair (KLF2→CXCR4) remains supported across tissues at thresholds up to $n \geq 10$ cells per tissue. On the refined pair set, mediation is concentrated (top-5 mass 0.63–0.66 for heads vs. 0.82–0.92 for MLP) and MLP mediator overlap is higher than head overlap (mean top-5 Jaccard up to 0.71 for lung–immune MLP). Overlap remains above random-null expectations across top-k definitions and is significant at top-5 for both heads ($p = 0.019$) and MLP ($p < 0.001$). These results support a narrow but reproducible claim: scGPT contains compact, partly conserved MLP mediation structure for at least one cross-tissue regulatory effect, while broader generalization is presently limited by sparse shared-pair evidence.

1 Introduction

Single-cell transformer models capture broad gene-expression dependencies and support transfer learning in cellular biology [6, 15]. However, most analyses still report aggregate predictions (for example, edge rankings or attention summaries) rather than mechanistic localization of model-internal computation. This leaves a central question open: when a TF perturbation changes a target prediction, which model components carry that effect?

This question is important for both interpretability validity and biological utility. Mechanistic claims are stronger when they are tied to internal causal pathways instead of post hoc narratives [1, 7]. In addition, cross-tissue biological claims should ideally be supported by conserved internal mediators, not only by similar edge-level outputs.

We adapt causal mediation analysis [12, 14] and transformer patching methods [9, 13] to scGPT and evaluate cross-tissue mediator conservation. The study is designed around one practical concern: candidate TF-target support is sparse across tissues, so coverage must be quantified explicitly before tracing.

Contributions.

1. A mediation-circuit pipeline for scGPT that estimates component-level restoration from TF ablations.
2. A discovery-coverage analysis showing where cross-tissue evidence is sufficient or insufficient for mechanistic tracing.
3. Evidence that, on the refined shared-pair set, MLP mediators are more concentrated and more cross-tissue conserved than attention-head mediators.

2 Related Work

Single-cell foundation models. scGPT introduced generative foundation modeling for single-cell omics [6]. Geneformer showed that transformer pretraining can support transfer-learning and network-level biological inference [15]. Our work focuses on mechanistic decomposition of specific TF-target effects within such models.

Mechanistic interpretability and intervention methods. Circuit-oriented transformer interpretability emphasizes internal pathways rather than global scores [7]. Activation patching and related targeted interventions are now standard tools for localizing computations across layers and modules [9, 13]. Sparse feature decomposition has also been used to improve interpretability granularity in large models [3].

Interpretability validity and pitfalls. Interpretability methods can produce plausible but weakly grounded explanations unless validated with controls. Probe interpretability caveats and methodological pitfalls are well documented [2, 11], and attribution methods require sanity checks to avoid spurious narratives [1]. We therefore emphasize explicit uncertainty reporting and negative-evidence cases.

3 Methods

3.1 Datasets and preprocessing

We use Tabula Sapiens kidney and lung plus a 20k-cell immune subset [5]. Candidate regulatory pairs are sourced from TRRUST and DoRothEA-backed settings [8, 10]. Preprocessing maps Ensembl identifiers to HGNC symbols, applies normalization and HVG filtering, and restricts analysis to model-vocabulary genes.

3.2 Intervention and mediation estimation

For each TF-target pair and eligible cell, we compute a total ablation effect

$$\mathrm{TE} = y_{\mathrm{baseline}} - y_{\mathrm{ablated}}, \tag{1}$$

where y is the target readout and TF ablation is applied in the model input pipeline.

For each mediator component M (attention head or MLP block), we patch the clean activation into the ablated run and compute restoration

$$r_M = y_{\mathrm{patched}(M)} - y_{\mathrm{ablated}}. \tag{2}$$

We summarize pair-level mediator influence via normalized mediation score

$$\mathrm{med}_M = \frac{r_M}{\mathrm{TE}}. \tag{3}$$

At tissue level, we report:

- **Top-5 mediator mass**: fraction of absolute restoration captured by the five strongest mediators.
- **Top-5 Jaccard overlap**: cross-tissue overlap of top mediator sets for shared pairs.

3.3 Discovery coverage and refinement

To avoid tracing unsupported pairs, we run a staged protocol:

1. broad ablation-only discovery,
2. fixed-pair replay across tissues,
3. refined mediation tracing only on shared-support pairs.

Coverage is quantified at minimum cell thresholds ($n \geq 1, 3, 5, 10$ per tissue), producing a direct estimate of how many pairs are suitable for cross-tissue mechanistic comparison.

3.4 Uncertainty and control logic

For ablation effects we report mean, standard deviation, and approximate 95% confidence intervals (normal approximation). For mediator restorations we report component-level confidence intervals from cell-level variation. A built-in negative-evidence condition is the low-support pair `STAT4`→`S100A4` in lung ($n = 2$), where unstable estimates test whether the analysis overstates weak evidence.

To stress-test overlap-based interpretability claims, we add two controls. First, for each granularity and top-k setting, we compare observed mean Jaccard overlap against random-set expectations from the same component space size. Second, at top-5 we run permutation null tests over random top-k sets to estimate empirical p-values for mean overlap. We also evaluate sensitivity across $k \in \{3, 5, 8, 10\}$ to test whether conclusions depend on one arbitrary cutoff.

4 Experimental Setup

We run two tiers:

- **Initial tracing tier**: 5 curated pairs across kidney/lung/immune for pipeline validation.
- **Refined tier**: 2 cross-tissue pairs selected by discovery replay. Pair set: `KLF2`→`CXCR4`, `STAT4`→`S100A4`.

All refined analyses use both attention-head and MLP mediator tracing with the same pair set, enabling direct granularity comparisons.

5 Results

5.1 Coverage analysis identifies one robust shared pair

Coverage analysis shows that cross-tissue support, not tracing mechanics, is the immediate bottleneck. In the random discovery panel (132 unique pairs), zero pairs are shared across all three tissues at every threshold tested. In contrast, fixed-pair replay yields two shared pairs at $n \geq 1$, but only one pair remains at $n \geq 3$ and above.

Table 1: Cross-tissue support coverage from ablation-only discovery.

Dataset	Min cells	Kidney	Lung	Immune	Shared all tissues
Random discovery (132 pairs)	≥ 1	65	34	41	0
Random discovery (132 pairs)	≥ 3	32	20	14	0
Random discovery (132 pairs)	≥ 5	13	15	6	0
Random discovery (132 pairs)	≥ 10	2	9	1	0
Fixed replay (40 pairs)	≥ 1	40	5	10	2
Fixed replay (40 pairs)	≥ 3	34	4	7	1
Fixed replay (40 pairs)	≥ 5	20	4	4	1
Fixed replay (40 pairs)	≥ 10	2	3	3	1

Interpretation: robust cross-tissue mechanistic analysis is currently feasible for `KLF2`→`CXCR4`; broader claims are constrained by support sparsity.

5.2 Pair-level total effects are strongest for KLF2-to-CXCR4

Table 2: Refined pair-level ablation effects with approximate 95% confidence intervals.

Pair	Tissue	n_{cells}	Effect mean	95% CI
KLF2→CXCR4	Kidney	12	-0.0315	[-0.0469, -0.0160]
KLF2→CXCR4	Lung	11	-0.0818	[-0.1765, 0.0128]
KLF2→CXCR4	Immune	12	-0.1570	[-0.2376, -0.0764]
STAT4→S100A4	Kidney	6	0.0726	[0.0308, 0.1145]
STAT4→S100A4	Lung	2	-0.0009	[-0.0031, 0.0012]
STAT4→S100A4	Immune	9	0.0509	[-0.0636, 0.1654]

The low-support lung estimate for `STAT4`→`S100A4` is effectively null at this sample size, reinforcing the coverage-driven refinement strategy.

5.3 Mediator concentration is consistently higher for MLP blocks

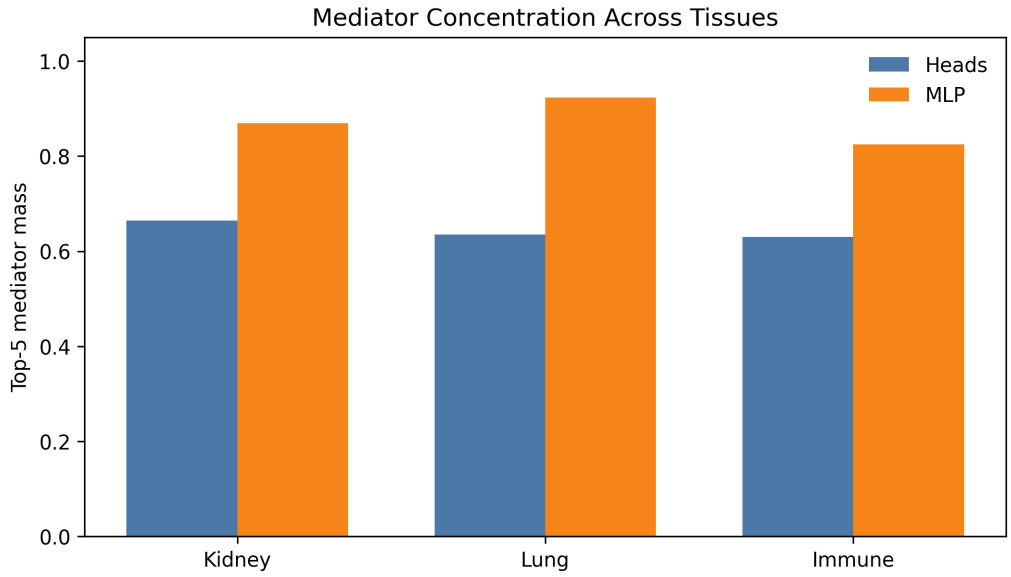

Figure 1: Top-5 mediator mass by tissue and granularity. Higher values indicate stronger concentration of restoration into a small mediator subset.

Table 3: Aggregate refined mediation metrics.

Tissue	Granularity	Pairs	Mean |effect|	Top-5 mediator mass
Kidney	Heads	2	0.0521	0.6647
Kidney	MLP	2	0.0521	0.8689
Lung	Heads	2	0.0414	0.6353
Lung	MLP	2	0.0414	0.9237
Immune	Heads	2	0.1040	0.6304
Immune	MLP	2	0.1040	0.8244

Across tissues, top-5 MLP mediators account for most restoration mass, while head mediation is less concentrated.

5.4 MLP mediator overlap exceeds head overlap across tissues

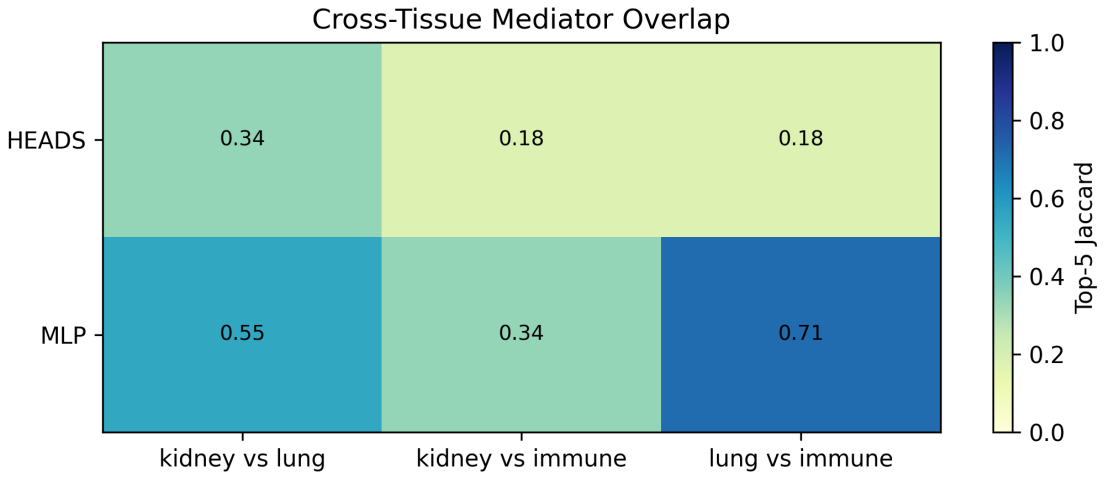

Figure 2: Top-5 mediator Jaccard overlap across tissue pairs. MLP overlap is consistently higher than head overlap in the refined setting.

Table 4: Mean top-5 mediator overlap (Jaccard) for shared refined pairs.

Granularity	Tissue pair	Mean Jaccard
Heads	Kidney-Lung	0.3393
Heads	Kidney-Immune	0.1806
Heads	Lung-Immune	0.1806
MLP	Kidney-Lung	0.5476
MLP	Kidney-Immune	0.3393
MLP	Lung-Immune	0.7143

These overlap patterns support an MLP-dominant conservation signal within the current evidence envelope.

5.5 Overlap signal exceeds random-null expectations and is top-k robust

Table 5: Null-control and sensitivity analysis for mediator overlap. Random expectation is based on top-k set overlap in the same component space (heads: 24 components; MLP: 12 components).

Granularity	Top-k	Obs. Jaccard	Rand. Jaccard	Enrich.	Perm. p (k=5)
Heads	3	0.1667	0.0667	2.50	–
Heads	5	0.2335	0.1163	2.01	0.019
Heads	8	0.4828	0.2000	2.41	–
Heads	10	0.6095	0.2632	2.32	–
MLP	3	0.4333	0.1429	3.03	–
MLP	5	0.5337	0.2632	2.03	< 0.001
MLP	8	0.7852	0.5000	1.57	–
MLP	10	0.8788	0.7143	1.23	–

Across all tested top-k values, overlap remains above random expectation, and top-5 overlap is statistically significant for both granularities. This reduces the risk that reported conservation is only a random finite-component artifact.

5.6 Pair-level heterogeneity check to limit cherry-picking risk

Table 6: Pair-level robustness diagnostics. Sign consistency is the fraction of tissues agreeing on effect direction after sign-coding (1.0 indicates complete agreement).

Pair	Sign consistency	Min cells	Mean top-5 Jaccard (heads)	Mean top-5 Jaccard (MLP)
KLF2→CXCR4	1.00	11	0.3095	0.6190
STAT4→S100A4	0.33	2	0.1574	0.4484

The stronger pair (`KLF2→CXCR4`) combines high support, consistent directionality, and higher overlap, while the weaker pair is directionally unstable and low-support. This supports reporting both pairs transparently while restricting stronger conservation claims to the robust pair.

5.7 Component-level uncertainty is tissue dependent

Table 7: Top-10 components with confidence intervals excluding zero.

Tissue	Granularity	Significant components (/10)	Fraction
Kidney	Heads	0/10	0.00
Kidney	MLP	0/10	0.00
Lung	Heads	5/10	0.50
Lung	MLP	5/10	0.50
Immune	Heads	4/10	0.40
Immune	MLP	7/10	0.70

The kidney refined run shows wider uncertainty at component level despite stable pair-level support, indicating that mediator ranking confidence remains heterogeneous across tissues.

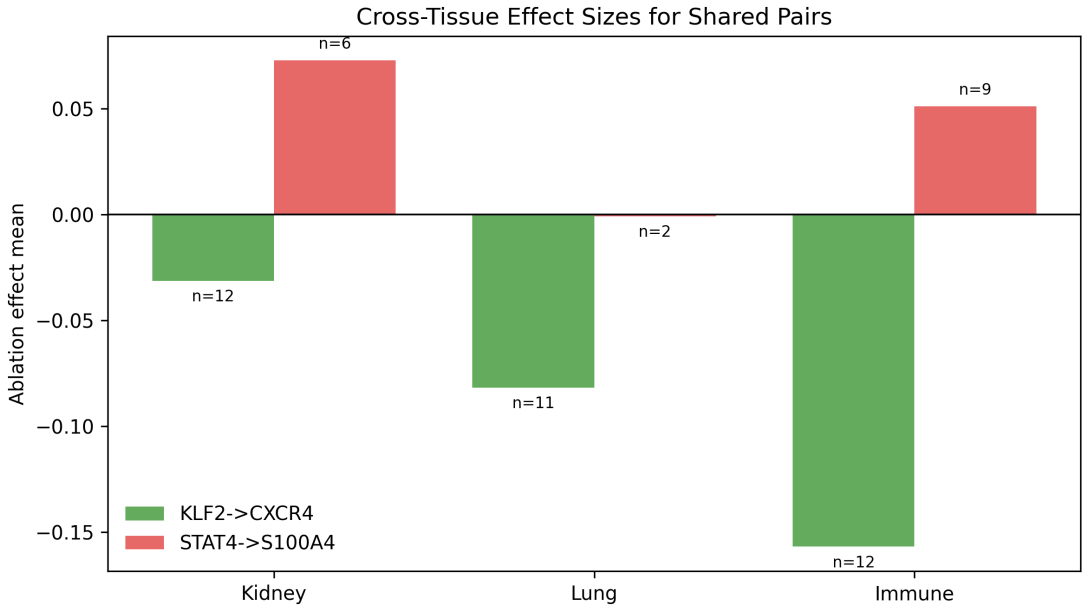

Figure 3: Cross-tissue ablation effects for refined shared pairs with per-bar cell-count annotations.

6 Biological Interpretation

The strongest shared pair, `KLF2`→`CXCR4`, is compatible with known roles of KLF2 in lymphocyte trafficking and chemokine-receptor-associated state programs [4]. Our claim is not that the model has recovered a proven direct regulatory edge; rather, the internal mediation pattern suggests that scGPT uses a compact feed-forward transformation for this TF-target context across tissues.

For `STAT4`→`S100A4`, evidence is insufficient for comparable mechanistic interpretation in lung because of minimal support ($n = 2$). This contrast between pairs is informative: biological narrative quality tracks support coverage.

7 Discussion

7.1 What is supported

Direct evidence supports three bounded conclusions: (1) component-level mediation tracing is feasible in a single-cell transformer, (2) restoration mass is concentrated in small mediator subsets, and (3) MLP mediator overlap exceeds head overlap for the refined shared-pair setting.

7.2 What is not yet supported

The current data do not support broad claims about universal cross-tissue conservation across many TF-target pairs. Coverage analysis shows that most candidate pairs fail shared-support criteria even before tracing. In addition, wide component-level confidence intervals in kidney indicate that local mediator rankings can remain unstable.

7.3 Alternative explanations and falsification pressure

Observed mediator overlap could still be inflated by shared dataset structure rather than conserved regulatory computation. Random-null controls and top-k sensitivity reduce this concern but do not eliminate it. Stronger falsification would require larger shared-pair panels, batch-aware stratification, and targeted perturbation validation. We therefore frame current conclusions as mechanistic hypotheses with one robust exemplar pair.

8 Conclusion

This study provides a reproducible mediation-circuit workflow for scGPT and identifies a clear practical bottleneck: cross-tissue shared-pair support. Within that constraint, `KLF2`→`CXCR4` shows concentrated and partially conserved MLP mediation across kidney, lung, and immune settings. The next step is scaling shared-support discovery so that conservation claims can be tested on larger TF-target panels.

Data and Code Availability

All code, analysis scripts, configuration files, and figure-generation pipelines are publicly available at https://github.com/Biodyn-AI/scgpt-causal-mediation-circuit-map. The repository README provides end-to-end instructions to reproduce all main figures and tables and includes dataset access notes for Tabula Sapiens and reference-network resources.

Supplementary Notes

Supplementary tables, expanded component-level outputs, and discovery-coverage artifacts are provided in the public repository. Operational run instructions are documented in the repository README rather than this manuscript.

References

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[2] Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022. doi: 10.1162/coli_a_00422.

[3] Trenton Bricken, Adly Templeton, Joshua Batson, et al. Towards monosemanticity: Decomposing language models with dictionary learning. Transformer Circuits Thread, 2023. URL https://transformer-circuits.pub/2023/monosemantic-features/index.html.

[4] Cynthia M. Carlson, Benjamin T. Endrizzi, Jun Wu, Xueyan Ding, Mark A. Weinreich, Elizabeth R. Walsh, Meera A. Wani, Jerry B. Lingrel, Kristin A. Hogquist, and Stephen C. Jameson. Kruppel-like factor 2 regulates thymocyte and t-cell migration. *Nature*, 442(7105):299–302, 2006. doi: 10.1038/nature04882.

[5] The Tabula Sapiens Consortium, R. C. Jones, et al. The tabula sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science*, 376(6594):eabl4896, 2022.

[6] Haotian Cui, Chloe Wang, Hamza Maan, Kuan Pang, Feng Luo, and Bo Wang. scgpt: Towards building a foundation model for single-cell multi-omics using generative ai. *Nature Methods*, 2024. Preprint originally posted on bioRxiv in 2023.

[7] Nelson Elhage, Neel Nanda, Catherine Olsson, et al. A mathematical framework for transformer circuits. Transformer Circuits Thread, 2021. URL https://transformer-circuits.pub/2021/framework/index.html.

[8] Luz Garcia-Alonso, C. H. Holland, M. M. Ibrahim, D. 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.

[9] Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. *Proceedings of EMNLP*, 2021.

[10] Hao Han, Jin-Wu Cho, Suna Lee, et al. Trrust v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic Acids Research*, 46(D1):D380–D386, 2018.

[11] John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 2733–2743, 2019. doi: 10.18653/v1/D19-1275.

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

[13] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems (NeurIPS)*, 35, 2022.

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

[15] Christina V. Theodoris, Ling Xiao, Anant Chopra, Mark D. Chaffin, Zeina R. Al Sayed, Matthew C. Hill, Helene Mantineo, Elizabeth M. Brydon, Zexian Zeng, X. Shirley Liu, and Patrick T. Ellinor. Transfer learning enables predictions in network biology. *Nature*, 618(7965): 616–624, 2023. doi: 10.1038/s41586-023-06139-9.