Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Evaluation protocol choices dominate reported performance in gene regulatory network benchmarks for single-cell foundation models

Ihor Kendiukhov

kenduhov.ig@gmail.com

University of Tübingen

Article

Keywords: evaluation bias, gene regulatory network, single-cell foundation model, scGPT, benchmark methodology, universe-aware evaluation

Posted Date: May 15th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9283852/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Evaluation protocol choices dominate reported performance in gene regulatory network benchmarks for single-cell foundation models

Ihor Kendiukhov[1,*]

[1]Department of Computer Science, University of Tübingen,
Sand 14, 72076 Tübingen, Germany

[*]Corresponding author: kenduhov.ig@gmail.com

Abstract

Gene regulatory network (GRN) recovery claims from single-cell foundation models depend critically on evaluation protocol choices that are rarely reported. We decompose the evaluation pipeline for scGPT attention-derived networks into four dimensions—symbol mapping, gold-standard selection, candidate-set restriction, and label noise—and show that each shifts reported AUPR by one to three orders of magnitude, dwarfing the signal attributable to the model itself. We introduce a *universe-aware* protocol that scores predictions against all biologically plausible TF–target pairs ($\sim 3.6 \times 10^9$), rather than restricted candidate sets containing only known TF–target pairs ($\sim 1.9 \times 10^6$). Under restricted evaluation, scGPT attention achieves AUPR of 0.00738—a 295× inflation over the universe-aware AUPR of 2.5×10^{-5}—driven entirely by the mechanical increase in positive base rate. Under universe-aware evaluation, scGPT attention falls below a random baseline, while tree-based methods (GENIE3, GRNBoost2) consistently outperform attention-derived edges across three Tabula Sapiens tissues. These findings demonstrate that published GRN recovery claims may reflect evaluation artifacts rather than learned biological structure.

Keywords: evaluation bias; gene regulatory network; single-cell foundation model; scGPT; benchmark methodology; universe-aware evaluation

1 Introduction

Single-cell foundation models such as scGPT[1], Geneformer[2], scFoundation[3], and scBERT[4] apply transformer architectures[5] pretrained on large-scale single-cell transcriptomic atlases[6,7] to encode gene dependencies that can be probed for mechanistic insight. A recent review surveys the rapidly growing landscape of these models[8]. A natural next question is whether these models internalize the causal regulatory structure of the genome: do attention weights or hidden-state interactions reflect real transcription factor (TF)–target dependencies? Several recent studies answer affirmatively by showing that attention-derived edges overlap with curated gene regulatory network (GRN) databases[1,9]. However, the interpretability of attention weights remains contested in the broader NLP literature[10,11,12], and emerging evidence suggests that single-cell foundation model evaluations require deeper scrutiny[13].

GRN evaluation is a fragile pipeline whose results are acutely sensitive to protocol details that are frequently underreported. The space of all directed gene pairs in a typical single-cell dataset exceeds 10^9, of

which only a tiny fraction appear in any gold standard. This extreme class imbalance makes precision-recall metrics sensitive to how the candidate set is constructed[14,15]: restricting evaluation to pairs where both source and target appear in the gold standard shrinks the denominator by three orders of magnitude, mechanically inflating AUPR without any change in the model's predictions. Simultaneously, gene symbol inconsistencies between model outputs and reference databases—a pervasive problem in computational biology that has been shown to affect published results[16]—can silently reduce overlap to near zero if unhandled.

GRN evaluation commonly relies on curated gold standards such as TRRUST v2[17], DoRothEA[18], OmniPath[19], and experimentally derived resources from ENCODE[20] and ChIP-Atlas[21]. These databases differ in coverage, confidence stratification, and bias toward well-studied TFs. Benchmarks frequently compare against established inference methods: tree-based approaches such as GENIE3[22] and GRNBoost2[23], the SCENIC pipeline[24,25], information-theoretic methods such as PIDC[26], deep learning approaches such as DeepSEM[27], and perturbation-based methods such as CellOracle[28], following community benchmark standards[29,30,31,32].

Less attention has been paid to evaluation protocol bias itself. Symbol mapping and gene identifier normalization are often underreported, despite their influence on overlap with gold standards. Candidate-set construction dominates reported metrics by changing the positive base rate, potentially inflating AUPR by orders of magnitude without altering predictions. Kernfeld et al.[33] recently demonstrated that transcriptome data alone are insufficient to control false discovery rates in GRN inference. Our study complements this perspective by showing that even when the inference method is held fixed, protocol choices alone can shift metrics by orders of magnitude, and we propose universe-aware evaluation as a controlled alternative to restricted scoring.

We present a protocol-focused evaluation for GRN benchmarking in mechanistic interpretability. Our central contribution is a *universe-aware* evaluation protocol that scores predictions against all directed gene pairs among expressed genes, rather than restricted candidate sets. We show that this distinction alone accounts for a 295× AUPR inflation for scGPT attention. Using Tabula Sapiens[34] kidney, lung, and immune tissues, we evaluate scGPT attention-derived edges and standard baselines against TRRUST v2 and DoRothEA references, systematically varying symbol mapping, candidate-set construction, gold-standard choice, and label noise.

2 Results

2.1 Symbol mapping is a binary gate on evaluation validity

The most striking finding is that symbol normalization acts as an all-or-nothing gate: without it, edge overlap between model outputs and gold standards is effectively zero, rendering any downstream metric meaningless. This occurs because scGPT's gene vocabulary uses a mix of current HGNC symbols[35] and historical aliases that do not match the naming conventions in TRRUST or DoRothEA without resolution. After mapping, average edge coverage reaches 0.997 (range 0.994–0.999) and average gene coverage reaches 0.997 (range 0.995–0.998). DoRothEA curated and ChIP-seq show higher mapped coverage (0.9983 and 0.9989) than TRRUST (0.9949), reflecting differences in gene coverage and curation depth (Supplementary Table S2).

Among mapping policies, practical differences in aggregate AUPR are modest (0.00679 for drop-ambiguous, 0.00663 for drop-unmapped, 0.00643 for lexicographic), but the choice matters in sparse-overlap regimes where a single misresolved alias can shift a TF from present to absent.

2.2 Candidate-set restriction inflates AUPR through base-rate mechanics

Restricting the candidate set is the most powerful lever in the evaluation protocol, capable of inflating AUPR by nearly three orders of magnitude without any change in model predictions (Table 2). For scGPT attention, median AUPR rises from 2.5×10^{-5} (universe-aware) through 1.41×10^{-3} (TF-source) to 7.38×10^{-3} (TF-source + target), a 295× lift. Across all methods, median AUPR rises from 3.16×10^{-4} to 1.40×10^{-2} (44×). These shifts track base-rate changes from 3.23×10^{-5} to 9.98×10^{-3}, confirming that the apparent improvement is an arithmetic consequence of evaluation design rather than evidence of better regulatory recovery. AUPR increases approximately linearly with base rate in log-log space across all methods (Fig. 1), a property of the evaluation metric under varying class balance[14,15], not of the models being evaluated. Table 1 summarizes AUPR across methods and protocols.

2.3 Tree-based baselines outperform scGPT attention across all conditions

Under controlled evaluation protocols, the method ranking is consistent and unambiguous (Table 1): GRNBoost2 and GENIE3 achieve the highest AUPR (0.0255 and 0.0225), followed by Spearman (0.00836), Pearson (0.00817), random (0.00223), and scGPT attention (0.00141). Under TF-source + target restriction, scGPT attention (0.00738) falls *below* the random baseline (0.0117), meaning that random guessing within the constrained candidate set outperforms attention-derived edges. This suggests that averaged attention weights do not capture directional regulatory structure recoverable by expression-based methods. Figure 2 summarizes method medians.

2.4 Tissue context modulates evaluation

Tissue-level differences in AUPR are substantial: kidney and lung show 2–7× higher AUPR than immune tissue across all methods (Table 3), reflecting denser gold-standard coverage of lineage-specific regulatory programs in kidney and lung versus the combinatorial, context-dependent programs of immune cells. Importantly, the tissue effect interacts with candidate-set restriction: the AUPR gap between tissues widens under universe-aware evaluation and narrows under TF-restricted evaluation, meaning tissue-level comparisons are only meaningful when the candidate-set protocol is held constant. In TF-source-target settings, scGPT remains below the random baseline in all tissues (Fig. 3a).

2.5 Gold-standard choice shifts the evaluation baseline

DoRothEA curated and ChIP-seq references yield systematically higher AUPR than TRRUST across all methods and tissues, even when predictions are identical. This difference arises because DoRothEA covers more TF–target pairs, producing a higher base rate under the same candidate-set protocol. Base rates range from 2.02×10^{-6} (universe-aware, TRRUST) to 2.60×10^{-2} (TF-source + target, DoRothEA curated). Studies reporting GRN recovery against DoRothEA will appear to perform better than equivalent studies reporting against TRRUST, even with identical models.

2.6 Evaluation stability degrades under candidate-set restriction

TF-restricted candidate sets, which produce the highest headline AUPR numbers, are also the least stable (Fig. 3b). Bootstrap CI widths increase from 1.02×10^{-4} (universe-aware) through 2.16×10^{-3} (TF-source) to 7.86×10^{-3} (TF-source + target), a 77× widening. Structured TF dropout increases mean AUPR standard

deviation to 1.42×10^{-3}, while random noise yields low variance under universe-aware evaluation (mean std 5.3×10^{-5}). This means that the evaluation settings most commonly used in published work are precisely the settings where results are least reproducible.

3 Discussion

Our central finding is that protocol choices account for more variance in reported GRN metrics than model differences do. Candidate-set restriction alone produces a 44× AUPR shift across methods, while the best-to-worst method difference within any fixed protocol is only 5–18×. This means that a poor model evaluated under a favorable (restricted) protocol will report higher AUPR than a strong model evaluated under a universe-aware protocol, making cross-study comparisons unreliable without protocol standardization. This observation parallels concerns raised in other benchmarking domains, where evaluation setup choices have been shown to dominate method rankings in trajectory inference[36] and single-cell data integration[37].

The consistent underperformance of scGPT attention relative to expression-based baselines challenges the premise that averaged attention weights encode regulatory relationships. Self-attention in transformer models serves multiple computational functions—positional encoding, expression-magnitude normalization, coexpression grouping[38]—and averaging across heads and layers dilutes any head-specific regulatory signal. Research in NLP has shown that attention weights do not straightforwardly correspond to feature importance[10,11], and that individual heads can encode specialized linguistic structure[12,39]. This does not mean scGPT lacks regulatory information entirely; rather, raw attention aggregation is the wrong extraction method, and more targeted probing strategies may be needed to access regulatory signals encoded in model internals[40].

We recommend that mechanistic interpretability studies report: (1) the exact symbol mapping pipeline with HGNC[35] and Ensembl[41] cross-references; (2) the candidate-set protocol with explicit sizes and base rates; (3) results against at least two gold standards; (4) comparison to expression-based and random baselines under identical protocols; (5) bootstrap confidence intervals; and (6) tissue-stratified results where applicable. Universe-aware evaluation should be reported alongside any restricted evaluation to quantify the base-rate contribution to apparent performance.

This study has several limitations. Gold standards are incomplete and biased toward well-studied TFs[17,18]. scGPT runs use reduced budgets (max_genes=1000, max_cells=2000) and CPU extraction. Tree-based baselines are run with reduced regulator/target sets for tractability. We evaluate a single attention aggregation and thresholding scheme; alternative strategies may change rankings.

4 Methods

4.1 Data and gene sets

We use Tabula Sapiens[34] scRNA-seq data for three tissues with distinct regulatory complexity: kidney (epithelial and mesenchymal lineages), lung (diverse epithelial, endothelial, and immune populations), and immune cells (hematopoietic lineages with extensively curated regulatory networks). This tissue selection tests whether evaluation biases are tissue-invariant or interact with regulatory structure. Standard quality control filters (min_genes=200, min_cells=3) are applied using Scanpy[42], following established best practices for single-cell analysis[43]. Library-size normalization to 10,000 counts and log1p transformation are applied[44,45]. We select 5,000 highly variable genes[46] for preprocessing. For scGPT attention extraction, we

further restrict to 1,000 genes and 2,000 cells sorted by expression, matching the model's input constraints. This subsetting is itself a source of evaluation bias: it limits the gene universe to highly expressed genes, which are more likely to overlap with well-studied TFs in gold standards.

4.2 Symbol mapping policies

Gene symbol inconsistency is pervasive in computational biology: the same gene may appear under different HGNC-approved symbols or historical aliases across databases[16]. Our mapping pipeline resolves these inconsistencies using HGNC alias and previous-symbol tables[35] plus Ensembl cross-references[41], applying case normalization and punctuation stripping before alias lookup. We compare three ambiguity policies for cases where a single alias maps to multiple current symbols: (1) lexicographic tie-breaking (deterministic but arbitrary), (2) drop ambiguous (conservative, reduces false matches at the cost of coverage), and (3) drop unmapped (retains ambiguous resolutions but removes symbols with no HGNC match). Mapping outcomes (unchanged, alias-resolved, ambiguous dropped, unmapped) are logged by category and tissue to quantify how much of the evaluation depends on the mapping step.

4.3 Gold standards

We evaluate against three reference databases with different curation philosophies: TRRUST v2[17] (literature-only, high confidence, ∼8,000 interactions covering ∼800 TFs), DoRothEA curated[18] (multi-evidence integration, confidence-stratified A–D), and DoRothEA ChIP-seq (experimentally derived binding evidence, broader but noisier). These resources represent complementary approaches to cataloguing regulatory interactions, ranging from manual literature curation to large-scale experimental profiling[19,20,21]. All references are mapped to a common symbol namespace before evaluation, ensuring that coverage differences reflect genuine database content rather than naming artifacts.

4.4 Candidate-set construction and the universe-aware protocol

The choice of which gene pairs to evaluate is the single most consequential protocol decision in GRN benchmarking, because it directly controls the positive base rate, which in turn determines the AUPR floor[14]. We define three candidate-set protocols of increasing restrictiveness:

Universe-aware (all pairs). Every ordered pair (g_i, g_j) among the N expressed genes forms a candidate edge, yielding $N \times (N-1) \approx 3.6 \times 10^9$ candidates for $N \approx 60{,}000$ genes. With ∼116,000 gold-standard edges, this produces a positive base rate of $\sim 3.2 \times 10^{-5}$. This protocol treats the full set of biologically plausible TF–target interactions as the evaluation universe, making no assumptions about which genes can act as regulators or targets.

TF-source restriction. Only pairs where the source gene is a known TF from the gold standard. This reduces candidates to $\sim 4.4 \times 10^7$ and raises the base rate to $\sim 2.0 \times 10^{-3}$, a 60× increase that mechanically lifts AUPR even for random predictions.

TF-source + TF-target restriction. Both source and target must appear in the gold standard. This further reduces candidates to $\sim 1.9 \times 10^6$ with a base rate of $\sim 1.0 \times 10^{-2}$, a 300× increase over the universe-aware protocol. Under this protocol, even a random predictor achieves non-trivial AUPR simply by guessing within the constrained space.

Candidate sizes and base rates are reported for every evaluation configuration, enabling readers to distinguish genuine model performance from base-rate artifacts.

4.5 Prediction methods

scGPT attention. We extract multi-head attention weights from the scGPT whole-human pretrained checkpoint[1], aggregate across heads and layers by averaging, and threshold at the 95th percentile with top_k=50 to produce directed edges. This aggregation collapses the multi-head structure, which may discard head-specific regulatory signals[12,39], but mirrors the approach used in prior work[1,9]. Runs are tissue-matched across kidney, lung, and immune gene sets.

Tree-based baselines. GENIE3[22] and GRNBoost2[23] via Arboreto use random-forest and gradient-boosting regression to predict each target gene's expression from candidate regulators. These methods model nonlinear expression dependencies and have consistently ranked among top performers in community benchmarks[29,30]. We use 200 estimators, 2,000 sampled cells, 500 regulators, and 1,000 targets per tissue.

Correlation baselines. Pearson and Spearman correlations between gene expression vectors capture linear and rank-order coexpression, providing a lower bound on what coexpression alone can recover.

Random baseline. Edges sampled uniformly within each candidate set, providing the expected AUPR under the null hypothesis of no regulatory information.

4.6 Metrics and stability

We report AUROC, AUPR, precision, recall, and F1, with AUPR as the primary metric because it is more informative than AUROC under extreme class imbalance[14,15]. We evaluate label-noise robustness under two regimes: random edge perturbation (5–30% of gold-standard labels flipped) simulates annotation noise, while structured TF/target dropout (10–30% of TFs or targets removed) simulates systematic gaps in gold-standard coverage. Each noise setting uses five repeats. Bootstrap confidence intervals (200 resamples) quantify sampling uncertainty.

Data availability

The Tabula Sapiens dataset is publicly available at https://tabula-sapiens-portal.ds.czbiohub.org/. TRRUST v2 is available at https://www.grnpedia.org/trrust/. DoRothEA is available via the decoupler Python package. The evaluation outputs and configurations used in this study are available from the corresponding author upon reasonable request.

Code availability

The evaluation pipeline code, configurations, and analysis scripts are available at [repository URL to be provided upon acceptance].

Funding

Not applicable.

Acknowledgements

The author acknowledges support from the University of Tübingen.

Author contributions

I.K. conceived the study, designed the evaluation protocol, implemented the analysis pipeline, performed all experiments, and wrote the manuscript.

Competing interests

The author declares no competing interests.

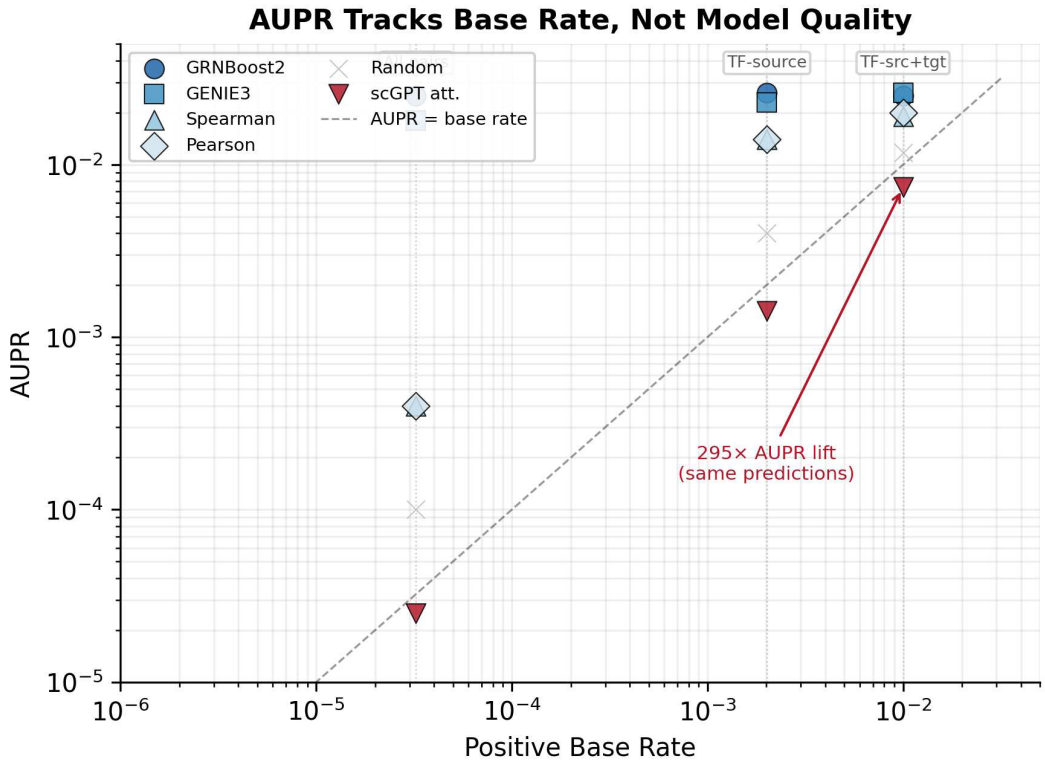


Figure 1: **AUPR tracks positive base rate, not model quality.** Scatter plot of AUPR versus positive base rate across all methods and candidate-set protocols. The 295× AUPR lift for scGPT (red arrow) is driven entirely by the mechanical increase in base rate from universe-aware to TF-restricted evaluation. The dashed line indicates AUPR equal to the base rate (random-predictor floor). All methods cluster near this line, confirming that candidate-set restriction inflates metrics without improving predictions.

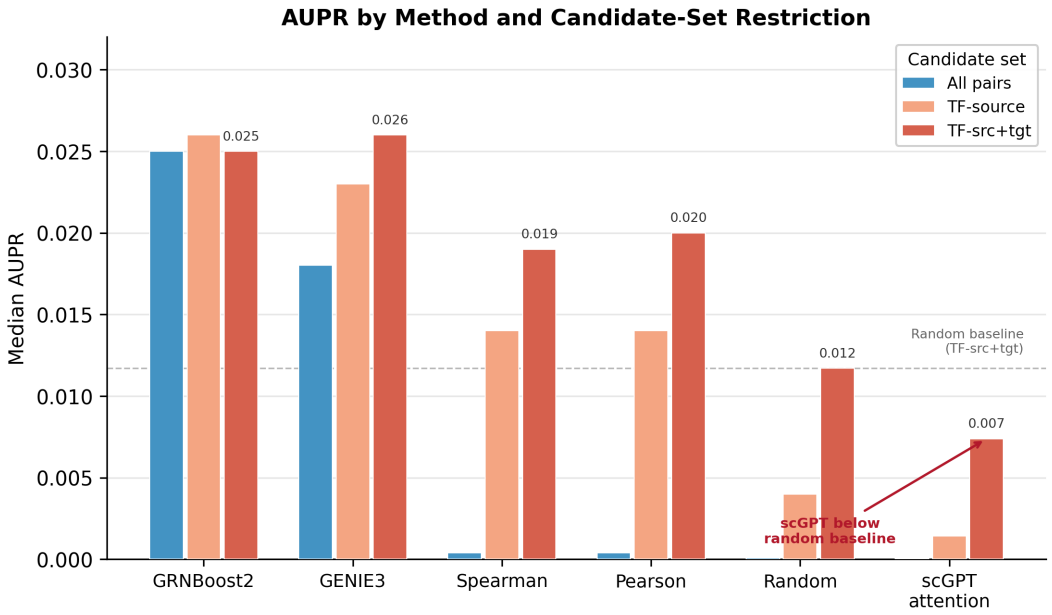


Figure 2: **AUPR by method and candidate-set restriction.** Median AUPR for each prediction method under three candidate-set protocols: all pairs (universe-aware), TF-source, and TF-source+target. GRNBoost2 and GENIE3 consistently outperform all other methods. scGPT attention (rightmost) falls below the random baseline (dashed line) under TF-source+target restriction, indicating that averaged attention weights do not recover directional regulatory structure.

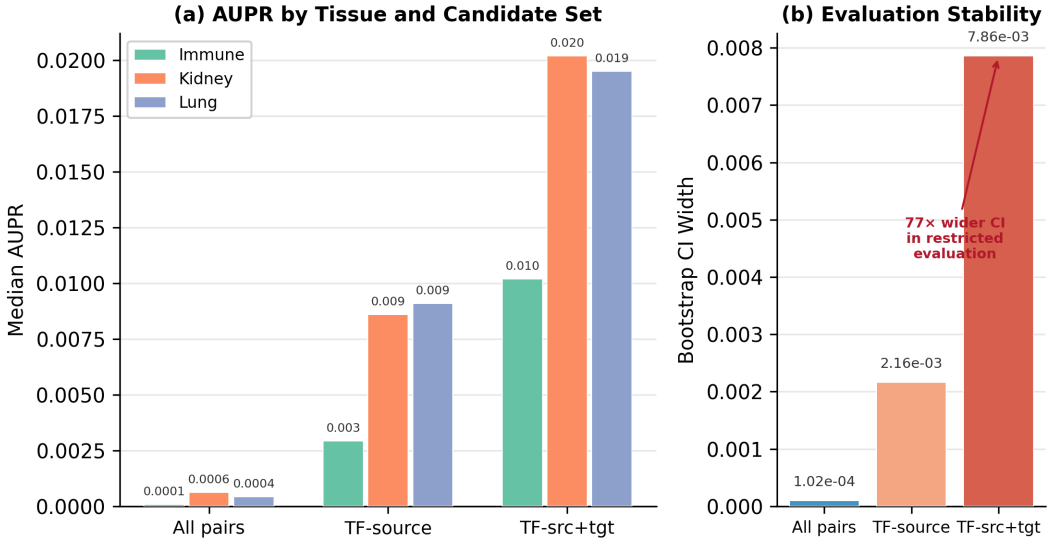


Figure 3: **Tissue context modulates AUPR and evaluation stability degrades under restriction.** (**a**) Median AUPR by tissue and candidate-set protocol. Kidney and lung show 2–7× higher AUPR than immune tissue, but the gap narrows under TF-restricted evaluation. (**b**) Bootstrap confidence interval width by candidate-set protocol, showing a 77× widening from universe-aware to TF-source+target restriction.

Table 1: **Median AUPR by method.** Overall median AUPR and AUPR under the most restrictive (TF-source + target) candidate-set protocol. scGPT attention (0.00738) falls below the random baseline (0.0117) under TF-source + target restriction.

Method	AUPR (all)	AUPR (TF-src+tgt)
GRNBoost2	0.0255	0.0255
GENIE3	0.0225	0.0256
Spearman	0.00836	0.0195
Pearson	0.00817	0.0204
Random	0.00223	0.0117
scGPT attention	0.00141	0.00738

Table 2: **Candidate-set sizes, base rates, and median AUPR.** Median across all methods. AUPR tracks base rate rather than model quality: a 300× reduction in candidate size produces a 44× lift in median AUPR.

Candidate set	Candidate size	Base rate	Median AUPR
Universe-aware	3.59e+09	3.23e-05	3.16e-04
TF-source	4.35e+07	2.01e-03	4.37e-03
TF-src+tgt	1.94e+06	9.98e-03	1.40e-02

Table 3: **Candidate-set median AUPR by tissue.** Immune tissue consistently shows lower AUPR, reflecting sparser gold-standard coverage of immune regulatory programs.

Tissue	Universe-aware	TF sources	TF src+tgt
Immune	9.10e-05	2.93e-03	1.02e-02
Kidney	6.24e-04	8.59e-03	2.02e-02
Lung	4.19e-04	9.09e-03	1.95e-02

References

[1] Cui, H., Wang, C., Maan, H., Pang, K., Luo, F. & Wang, B. scGPT: towards building a foundation model for single-cell multi-omics using generative AI. *Nat. Methods* **21**, 1470–1480 (2024).

[2] Theodoris, C. V. et al. Transfer learning enables predictions in network biology. *Nature* **618**, 616–624 (2023).

[3] Hao, M., Gong, J., Zeng, X. et al. Large-scale foundation model on single-cell transcriptomics. *Nat. Methods* **21**, 1481–1491 (2024).

[4] Yang, F., Wang, W., Wang, F. et al. scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. *Nat. Mach. Intell.* **4**, 852–866 (2022).

[5] Vaswani, A. et al. Attention is all you need. In *Advances in Neural Information Processing Systems* **30**, 5998–6008 (2017).

[6] Regev, A. et al. The Human Cell Atlas. *eLife* **6**, e27041 (2017).

[7] Zheng, G. X. Y. et al. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).

[8] Szalata, A. et al. Transformers in single-cell omics: a review and new perspectives. *Nat. Methods* **21**, 1430–1443 (2024).

[9] Xu, J., Zhang, A., Liu, F. & Zhang, X. STGRNS: an interpretable transformer-based method for inferring gene regulatory networks from single-cell transcriptomic data. *Bioinformatics* **39**, btad165 (2023).

[10] Jain, S. & Wallace, B. C. Attention is not explanation. In *Proc. NAACL-HLT* 3543–3556 (2019).

[11] Wiegreffe, S. & Pinter, Y. Attention is not not explanation. In *Proc. EMNLP-IJCNLP* 11–20 (2019).

[12] Clark, K., Khandelwal, U., Levy, O. & Manning, C. D. What does BERT look at? An analysis of BERT's attention. In *Proc. ACL Workshop BlackboxNLP* 276–286 (2019).

[13] Boiarsky, R. et al. Deeper evaluation of a single-cell foundation model. *Nat. Mach. Intell.* **6**, 1443–1446 (2024).

[14] Davis, J. & Goadrich, M. The relationship between precision-recall and ROC curves. In *Proc. 23rd International Conference on Machine Learning* 233–240 (ACM, 2006).

[15] Saito, T. & Rehmsmeier, M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE* **10**, e0118432 (2015).

[16] Ziemann, M., Eren, Y. & El-Osta, A. Gene name errors are widespread in the scientific literature. *Genome Biol.* **17**, 177 (2016).

[17] Han, H., Cho, J.-W., Lee, S. et al. TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic Acids Res.* **46**, D199–D202 (2018).

[18] Garcia-Alonso, L., Holland, C. H., Ibrahim, M. M., Turei, D. & Saez-Rodriguez, J. DoRothEA: a resource of transcription factor activities and regulons for human and mouse. *Genome Res.* **29**, 1363–1375 (2019).

[19] Türei, D. et al. Integrated intra- and intercellular signaling knowledge for multicellular omics analysis. *Mol. Syst. Biol.* **17**, e9923 (2021).

[20] ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).

[21] Oki, S. et al. ChIP-Atlas: a data-mining suite powered by full integration of public ChIP-seq data. *EMBO Rep.* **19**, e46255 (2018).

[22] Huynh-Thu, V. A., Irrthum, A., Wehenkel, L. & Geurts, P. Inferring regulatory networks from expression data using tree-based methods. *PLoS ONE* **5**, e12776 (2010).

[23] Moerman, T., Aibar, S., Gonzalez-Blas, C. B. et al. GRNBoost2 and Arboreto: efficient and scalable inference of gene regulatory networks. *Bioinformatics* **35**, 2159–2161 (2019).

[24] Aibar, S. et al. SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods* **14**, 1083–1086 (2017).

[25] Van de Sande, B. et al. A scalable SCENIC workflow for single-cell gene regulatory network analysis. *Nat. Protoc.* **15**, 2247–2276 (2020).

[26] Chan, T. E., Stumpf, M. P. H. & Babtie, A. C. Gene regulatory network inference from single-cell data using multivariate information measures. *Cell Syst.* **5**, 251–267.e3 (2017).

[27] Shu, H. et al. Modeling gene regulatory networks using neural network architectures. *Nat. Comput. Sci.* **1**, 491–501 (2021).

[28] Kamimoto, K. et al. Dissecting cell identity via network inference and in silico gene perturbation. *Nature* **614**, 742–751 (2023).

[29] Marbach, D., Costello, J. C., Küffner, R. et al. Wisdom of crowds for robust gene network inference. *Nat. Methods* **9**, 796–804 (2012).

[30] Pratapa, A., Jalihal, A. P., Law, J. N., Bharadwaj, A. & Murali, T. M. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nat. Methods* **17**, 147–154 (2020).

[31] Stolovitzky, G., Monroe, D. & Califano, A. Dialogue on reverse-engineering assessment and methods: the DREAM of high-throughput pathway inference. *Ann. N. Y. Acad. Sci.* **1115**, 1–22 (2007).

[32] Chicco, D. & Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* **21**, 6 (2020).

[33] Kernfeld, E., Keener, R., Cahan, P. & Battle, A. Transcriptome data are insufficient to control false discoveries in regulatory network inference. *Cell Syst.* **15**, 709–722 (2024).

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

[35] Braschi, B. et al. Genenames.org: the HGNC and VGNC resources in 2019. *Nucleic Acids Res.* **47**, D786–D792 (2019).

[36] Saelens, W., Cannoodt, R., Todorov, H. & Saeys, Y. A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.* **37**, 547–554 (2019).

[37] Luecken, M. D. et al. Benchmarking atlas-level data integration in single-cell genomics. *Nat. Methods* **19**, 41–50 (2022).

[38] Elhage, N. et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread* (2021). https://transformer-circuits.pub/2021/framework/index.html

[39] Vig, J. A multiscale visualization of attention in the transformer model. In *Proc. 57th Annual Meeting of the ACL: System Demonstrations* 37–42 (2019).

[40] Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proc. NAACL-HLT* 4171–4186 (2019).

[41] Cunningham, F. et al. Ensembl 2022. *Nucleic Acids Res.* **50**, D988–D995 (2022).

[42] Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).

[43] Luecken, M. D. & Theis, F. J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.* **15**, e8746 (2019).

[44] Lun, A. T. L., Bach, K. & Marioni, J. C. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol.* **17**, 75 (2016).

[45] Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol.* **20**, 296 (2019).

[46] Brennecke, P. et al. Accounting for technical noise in single-cell RNA-seq experiments. *Nat. Methods* **10**, 1093–1095 (2013).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplementaryinformation.pdf