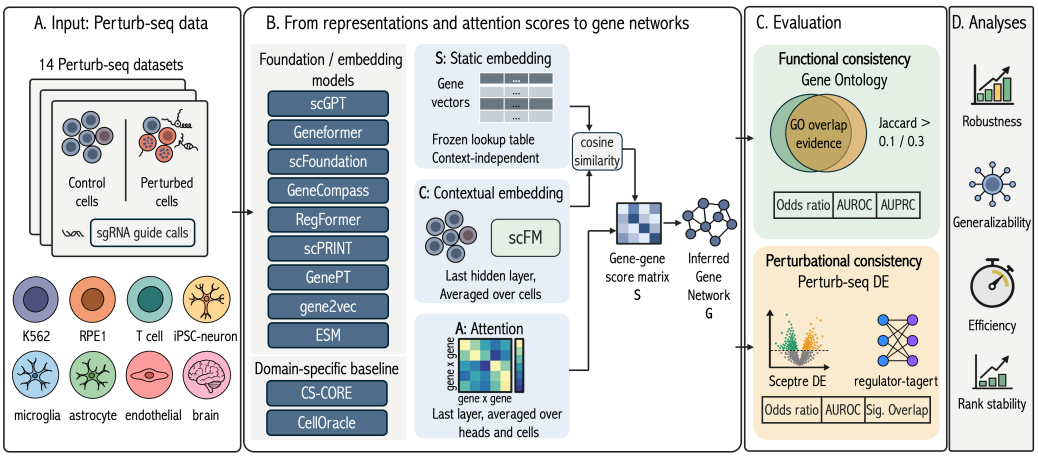
GNEVAL: BENCHMARKING OF CONTEXT-SPECIFIC FOUNDATION MODELS FOR GENE NETWORK INFERENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Gene networks encode regulatory and functional relationships among genes. Single-cell foundation models, pretrained on large and diverse datasets, provide a range of gene-gene association signals that offer a promising basis for gene network inference. However, it remains unclear which gene network properties these signals capture and how they compare with domain-specific methods. To address this gap, we introduce **GNEval**, a unified benchmark for evaluating pretrained model derived gene associations for network inference. We benchmark 21 gene representations and association scores from 11 models, including static gene embeddings (S), contextual embeddings (C), and gene-gene attention scores (A). GNEval spans 14 Perturb-seq datasets across diverse cell systems, perturbation types and data sizes. Inferred gene networks are assessed based on their consistency with functional and perturbational evidence. The results reveal substantial heterogeneity across evaluation tasks. Rankings based on Gene Ontology and Perturb-seq evidence were nearly uncorrelated (Spearman $\rho = 0.01$). Context-independent representations showed a substantial advantage in recovering functional relationships ($\Delta\text{AUROC} = 0.13$), whereas context-dependent gene associations improved recovery of perturbation-associated differential expression relationships by $4.9\times$. The domain-specific method achieves relatively strong performance on both evidence types, offering a better balance across the two evaluations than most foundation-model-derived networks. Overall, gene representation and association scores show task-specific strengths, highlighting the need for task-aware selection and integration with domain-specific methods.



1 INTRODUCTION

Gene network (GN) inference provides a valuable framework for identifying gene relationships and elucidating the biological processes and disease mechanisms underlying cellular behavior,

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

thereby informing the development of virtual cell models (Zhang & Horvath, 2005; Roohani et al., 2023). Numerous methods have been developed to infer GNs from small-scale single-cell RNA-seq (scRNA-seq) datasets collected within specific cellular contexts (Su et al., 2023; Kamimoto et al., 2023). However, these approaches often rely primarily on limited cells from a single study and do not fully leverage the extensive prior knowledge of gene relationships accumulated across existing datasets and biological resources.

Recent advances in single-cell foundation models (scFMs) (Cui et al., 2024; Theodoris et al., 2023; Yang et al., 2024; Hao et al., 2024; Hu et al., 2026; Kalfon et al., 2025) and large language models (LLMs) (Chen & Zou, 2025) provide a new opportunity to advance GN inference for small-scale, context-specific datasets (Ota et al., 2026). Specifically, these models encode rich prior knowledge from large-scale single-cell datasets and literature-curated resources in the learned gene representations, and similarity of gene representations inform similarity of gene function (Hu & Wei, 2026). For example, pretrained gene embeddings can be used directly in a zero-shot setting, contextual gene representations can be extracted from the final hidden layers, and attention scores can be leveraged as model-internal signals of gene-gene relationships.

However, it remains unclear what types of GN information are encoded by scFMs. A key challenge in interpreting inferred GNs is that gene relationships are highly context-dependent and heterogeneous, and different evaluation perspectives may capture distinct types of biological information. Existing studies often rely on a single source of evidence and rarely assess whether the inferred networks are supported by perturbational evidence (Wu et al., 2025; Wenteler et al., 2025).

Advances in genome editing and scRNA-seq, particularly Perturb-seq across diverse cellular contexts and perturbation types, provide new opportunities to study gene relationships at scale (Dixit et al., 2016). Perturb-seq measures cellular transcriptomic profiles upon perturbation of regulators, such that genes differentially expressed between perturbed and control cells are potential targets of the perturbed regulator in GNs. Furthermore, Perturb-seq captures both perturbation-induced gene effects and matched single-cell transcriptomic profiles from the same cellular context, allowing scFMs to be applied to the observed cells and their inferred gene relationships to be evaluated against experimentally measured perturbational effects. Such data have also accelerated the development of virtual-cell models by enabling systematic modeling of cellular responses to genetic perturbations (Roohani et al., 2025).

In recent years, several studies have benchmarked scFMs on gene-related tasks (Hou et al., 2026; Wenteler et al., 2025; Wu et al., 2025; Li et al., 2025; Radig et al., 2026; Zhong et al., 2025). However, existing evaluations often focus on context-independent representations, do not cover GN inference tasks, and provide limited comparisons with domain-specific GN inference methods. To our knowledge, no existing benchmark has systematically evaluated the utility of context-specific gene embeddings or attention-derived association scores from scFMs for GN inference.

To bridge these gaps, we introduce a comprehensive benchmarking framework **GNEval** for evaluating gene representations in GN inference. GNEval covers 14 datasets, 11 models, and three types of gene representations, and evaluates gene relationships using complementary evidence from gene function database and perturbational effects. GNEval assesses network quality from three perspectives: edge-set enrichment, edge ranking, and recovery of supported gene relationships. We employ unified data preprocessing, standardized representation extraction, consistent evaluation metrics, and sensitivity analyses to ensure fair and comprehensive comparisons. We additionally examine stability across benchmark settings, generalizability across biological contexts, and computational efficiency and scalability.

Through systematic experiments, we derive several key insights. First, context-independent and context-dependent representations capture distinct aspects of gene relationships. Context-independent representations generally better recover functional relationships defined by Gene Ontology (GO), whereas context-dependent representations ans association scores better recover perturbational evidence. Second, the domain-specific method CS-CORE achieves relatively strong performance on both evidence types, offering a better balance across the two evaluations than most foundation-model-derived networks.

Our main contributions are as follows:

- To our knowledge, GNEval is the first systematic benchmark of context-specific gene associations derived from scFMs for GN inference. We evaluate diverse scFMs and other pretrained gene representation models against domain-specific GN inference methods across multiple datasets and biological contexts.
- To our knowledge, GNEval introduces the first systematic use of Perturb-seq data for context-matched evaluation of GNs, combining perturbational and functional evidence across a curated collection of datasets with analyses of robustness, generalizability, and computational efficiency.
- We develop a standardized pipeline for harmonizing heterogeneous Perturb-seq datasets and deriving consistent perturbation-based evidence for gene relationships from an up-to-date collection of public datasets.
- We provide an extensible, reproducible implementation for benchmarking, supporting future evaluation and development of gene representation and network inference methods.

2 GNEVAL FRAMEWORK

2.1 TASK DEFINITION: GENE NETWORK INFERENCE

Let $\mathcal{V} = \{g_1, \ldots, g_N\}$ denote a set of genes. We propose to infer GN using a gene-gene association matrix

$$\mathbf{S} \in \mathbb{R}^{N \times N},$$

where S_{ij} quantifies the inferred association between genes g_i and g_j.

$\mathbf{S}$ can be derived from different model outputs. For gene embeddings, each gene g_i is associated with an embedding $\mathbf{z}_i \in \mathbb{R}^d$, and pairwise association scores are derived from the learned representations, using cosine similarity (Hu et al., 2026; Chen & Zou, 2025; Yang et al., 2024),

$$S_{ij} = \frac{\mathbf{z}_i^\top \mathbf{z}_j}{\|\mathbf{z}_i\|_2 \|\mathbf{z}_j\|_2}.$$

For pairwise gene-gene scores obtained directly from models, including attention scores and association scores from domain-specific methods, we use their inferred scores as $\mathbf{S}$ (Section 2.2).

To enable comparison across methods, gene pairs are ranked according to their inferred association scores, and the top K ranked pairs are retained to form the inferred network

$$\mathcal{G}_K = (\mathcal{V}, \mathcal{E}_K),$$

where $\mathcal{V}$ denotes the set of evaluated genes and $\mathcal{E}_K$ contains the K highest-ranked gene pairs according to their absolute association scores. We consider $K \in \{5{,}000, 10{,}000, 20{,}000\}$.

The benchmark evaluates whether the resulting networks recover biologically supported gene-gene relationships using independent functional and perturbational evidence.

2.2 MODELS

GNEval evaluates GNs inferred from 11 models spanning diverse sources (Table 1). We include four recent scFMs, **scGPT** (Cui et al., 2024), **Geneformer** (Theodoris et al., 2023), **scFoundation** (Hao et al., 2024), and **GeneCompass** (Yang et al., 2024), which were trained on atlas-scale single-cell data. We extract signals at different stages of the model, including context-independent pretrained token embeddings, contextualized final-layer gene embeddings, and attention scores. To provide broader baselines, we further consider gene representations learned from alternative data sources beyond single-cell transcriptomic data. **GenePT** (Chen & Zou, 2025) provides literature-derived gene embeddings obtained from a large language model, while **gene2vec** (Du et al., 2019) learns gene representations from co-expression pairs using a Word2Vec objective. We also include **ESM** (Lin et al., 2023), a protein language model from which gene-level token embeddings are constructed by averaging the pretrained representations of all proteins encoded by each gene, following the construction adopted in scFMs such as UCE (Rosen et al., 2026) and scPRINT (Kalfon et al., 2025). Beyond general-purpose representation models, we evaluate other two recent scFMs,

Method	Dim	Type	Training basis	Used as
GenePT	1536	Text embedding	GPT text of gene descriptions	S
gene2vec	200	Word2Vec	Bulk co-expression (GEO)	S
ESM	5120	Protein LM	ESM2 embeddings of gene products	S
scGPT	512	Transformer	scRNA-seq (∼33M cells)	S, C, A
scFoundation	768	Transformer	scRNA-seq (∼50M cells)	S, C, A
Geneformer	768	Transformer	scRNA-seq (∼95M cells, gc-104M)	S, C, A
GeneCompass	768	Transformer	scRNA-seq (∼120M human + mouse cells) + prior knowledge	S, C, A
RegFormer	512	Mamba (state-space)	scRNA-seq (∼26M human cells), GRN-ordered	S, C
scPRINT	256	Transformer	scRNA-seq (∼50M cells)	C, A
CS-CORE	—	Co-expression estimator	scRNA-seq (no pre-training)	Domain-specific
CellOracle	—	GRN inference	scRNA-seq + motif-based GRN prior (no pre-training)	Domain-specific

Table 1: Methods under comparison in GNEval. **Used as: S** static gene embedding, **C** contextual embedding, and **A** gene-gene attention read directly as edge scores.

RegFormer (Hu et al., 2026) and **scPRINT** (Kalfon et al., 2025), with architectures and training strategies tailored toward modeling gene regulatory relationships.

Finally, we include two domain-specific methods: **CS-CORE** (Su et al., 2023), which infers gene-gene relationships directly from observational scRNA-seq, with its test statistic used as the association score, and **CellOracle** (Kamimoto et al., 2023), which reconstructs context-specific gene regulatory networks (GRNs) by combining scRNA-seq with motif-based transcription factor (TF)-target priors and regularized regression, with the regression coefficient used as the association score.

2.3 Gene Association Scores and Network Construction

GNEval evaluate three broad classes of methods (Table 1): scFMs, other pretrained gene representation models, and domain-specific GN inference methods. Because a single scFM may provide gene-gene association scores from multiple stages of the model, the 11 evaluated models yield a total of 21 representations or association scores. Specifically, scFMs can provide static embeddings, contextual embeddings, and attention scores. In contrast, pretrained models such as GenePT, gene2vec, and ESM provide only static embeddings and do not produce dataset-specific contextualized representations. Finally, CS-CORE and CellOracle directly infer gene-gene associations from single-cell data without pretraining and are therefore included as a separate class of domain-specific baselines.

For each scFM, we extract gene representations or gene-gene association scores from different stages of the model and categorize them into three representation types: static embeddings (**S**), contextual embeddings (**C**), and attention scores (**A**) (Baek et al., 2025). Table 1 summarizes the representation or score types available from each evaluated model.

Static embeddings (S). Static embeddings are context-independent gene representations obtained without providing any dataset-specific single-cell observations to the model. Each gene is therefore associated with a single fixed vector for a given model, shared across all downstream datasets. For scFMs, these representations are extracted from the pretrained token embedding table when available. We additionally consider context-independent representations derived from other sources, including GenePT (Chen & Zou, 2025) embeddings generated from gene-description text, gene2vec (Du et al., 2019) embeddings learned from bulk gene co-expression, and ESM (Lin et al., 2023) protein embeddings aggregated to the gene level. All these representations do not depend on the single-cell datasets used for downstream evaluation. These embeddings are converted to pairwise gene-gene association scores using cosine similarity.

Contextual embeddings (C). Contextual embeddings are dataset-specific gene representations obtained by running each scFM in inference mode on single-cell expression profiles. For each dataset,

we extract gene representations from the final hidden layer and average the representations of each gene across cells in which that gene is observed. These embeddings are converted to pairwise gene-gene association scores using cosine similarity and treated as context-dependent.

Attention scores (A). For models with self-attention, we directly extract gene-gene attention scores by running each scFM in inference mode on single-cell expression profiles. We use the attention matrix from the final layer, average across attention heads, and then average each gene pair over cells in which both genes are observed. These scores are used directly as edge scores and treated as context-dependent.

To extract contextual embeddings and attention scores from scFMs, we randomly subsample 1,000 cells from the full Perturb-seq population, without providing perturbation labels. The pretrained parameters remain fixed, and the expression data are used only to contextualize the representations.

For the domain-specific methods, we use the CS-CORE test statistic and the CellOracle absolute regression coefficient as gene-gene edge scores. We construct networks by retaining the top 5,000, 10,000, and 20,000 highest-scoring edges for each method when available. Because CellOracle estimates scores only for candidate TF-target pairs and therefore produces substantially sparser networks, it is evaluated using all edges available in the network.

2.4 Functional and Perturbational Evaluation

GNEval evaluated inferred GNs against two complementary sources of biological evidence: GO-based functional relationships and Perturb-seq-derived perturbational relationships. GO annotations (Gene Ontology Consortium, 2004) were used to define co-functional gene pairs among the top 1,000 highly variable genes in each dataset. Treating GO as a bipartite gene-term graph, each gene i was represented by its annotated GO-term set G_i, and pairwise functional similarity between genes i and j was quantified using the Jaccard index $J(i,j) = |G_i \cap G_j|/|G_i \cup G_j|$ (Roohani et al., 2023). Gene pairs with $J(i,j) \geq 0.1$ or 0.3 were considered GO-supported, with both thresholds evaluated separately to assess robustness.

For perturbational evaluation, Perturb-seq datasets were processed using a unified pipeline (see details in Appendix B). Differential expression (DE) analysis was performed with `sceptre` (Barry et al., 2024; 2021) using raw counts from single-guide cells, with non-targeting-guide cells serving as controls. Only the top 1,000 highly variable genes were considered in DE testing, with sequencing depth and detected-gene count included as covariates. An edge $i \to j$ was considered DE-supported when perturbation of regulator i produced a significant expression change in gene j with $\mathrm{FDR} < 0.05$ and $|\log_2 \mathrm{FC}| > 0.25$.

For both evidence sources, we selected three metrics to capture complementary aspects of network quality. Two metrics assess the identification and ranking of biologically supported edges. The third metric is evidence-specific: for GO, AUPRC evaluates the recovery of positive gene pairs under class imbalance, whereas for perturbational evidence, regulator-level enrichment measures agreement between inferred target genes and perturbation-derived DE genes. For identification, enrichment was quantified using the Fisher odds ratio computed from the contingency table comparing inferred edges with supported gene pairs. This metric measures whether supported edges are overrepresented among the inferred edges. For ranking, AUROC was used to assess whether supported gene pairs tend to receive higher association scores than other pairs. Because GO-supported pairs are typically sparse relative to the full set of gene pairs, GO-based evaluation additionally included AUPRC to assess positive-pair recovery under class imbalance. For Perturb-seq-based evaluation, we further used the significant-overlap rate, defined as $\#\{r \in R_{\mathrm{test}} : Q_{\mathrm{Fisher}}(r) < 0.05\}/|R_{\mathrm{test}}|$, where R_{test} is the set of regulators retained after filtering (Appendix A) and $Q_{\mathrm{Fisher}}(r)$ is the FDR-adjusted p-value from a one-sided Fisher's exact test assessing over-representation of perturbation-derived DE genes among the inferred target genes of regulator.

2.5 Comparison of Representation Types

We tested whether the use of target-dataset context was associated with systematic differences in functional and perturbational consistency. Pretrained representations and association scores were grouped into context-independent and context-dependent categories, with CS-CORE and CellOracle analyzed separately as a domain-specific observational baseline (Table 1; Section 2.3). For

each dataset, metric values were averaged within each representation group and compared across datasets using paired two-sided Wilcoxon signed-rank tests. Multiple-testing correction and additional statistical details are provided in Appendix H.

Additional analyses evaluated robustness to network size, GO similarity thresholds, cell subsampling, dataset characteristics, and computational scaling. (Appendices J and K).

3 RESULTS

3.1 GNEVAL DATASETS

We evaluate all GNs on 14 publicly available Perturb-seq datasets spanning diverse cell types, perturbation modalities, and dataset scales. To our knowledge, this is the first systematic curation of Perturb-seq datasets for benchmarking GN inference across diverse biological contexts. We obtain the raw count matrices and associated metadata from the Gene Expression Omnibus (GEO) and develop a consistent preprocessing pipeline across datasets (see Section 4 for code availability). Detailed dataset characteristics, including the numbers of cells, perturbations, cell types, and significant DE edges are summarized in Table 2.

Table 2: Perturb-seq datasets used for GNEval. "Sig. DE edges" is the number of significant regulator-gene pairs identified by SCEPTRE (FDR < 0.05 and $|\log_2 \mathrm{FC}| > 0.25$).

Data	Perturbation	Cell type	Cell number	Sig. DE edges
Adamson et al. (2016)	CRISPRi	K562	65,145	8,386
Norman et al. (2019)	CRISPRa	K562	69,686	9,775
Replogle et al. (2022)	CRISPRi	K562	310,385	11,781
Replogle et al. (2022)	CRISPRi	RPE1	247,914	24,615
Schmidt et al. (2022)	CRISPRa	primary T cells	61,041	6,220
Sunshine et al. (2023)	CRISPRi	Calu-3	35,729	1,334
Otto et al. (2023)	CRISPRko	MOLM-13	31,069	1,244
Schnitzler et al. (2024)	CRISPRi	endothelial	214,449	98
Tian et al. (2021)	CRISPRi	iPSC neuron	32,300	143
Tian et al. (2021)	CRISPRa	iPSC neuron	21,193	999
Dräger et al. (2022)	CRISPRi	iPSC microglia	10,377	1,375
Leng et al. (2022)	CRISPRi	iPSC astrocyte	7,933	168
Zhu et al. (2026)	CRISPRi	primary CD4 T cells	7,282,273	51,311
Shi et al. (2026)	CRISPR (*in vivo*, mouse)	excitatory neuron	2,246,807	186

Note: CRISPRi, CRISPRa, and CRISPRko denote CRISPR interference, CRISPR activation, and CRISPR knockout, respectively.

3.2 FUNCTIONAL AND PERTURBATIONAL EVALUATION PERFORMANCE

We evaluate gene association scores derived from scFMs and other pretrained gene representation models, together with CS-CORE and CellOracle, across 14 Perturb-seq datasets using complementary functional and perturbational evidence. GO-based evaluation measures consistency with functionally related gene pairs defined from GO annotations, with performance summarized in Fig. 1. Perturb-seq-based evaluation measures consistency with regulator-target relationships supported by DE following CRISPR perturbation, with results summarized in Fig. 2.

In Figure 1, a clear overall trend is evident. GenePT leads every GO metric (median OR 14.9, AUROC 0.85, AUPRC 0.32), followed by ESM (7.5, 0.79, 0.24). Next come the static scFM embeddings scGPT_S and Geneformer_S. All attention scores, CellOracle and scFoundation_S sit at chance. In Figure 2, performance against perturb-seq DE evidence is weak overall: median DE AUROC falls between 0.40 and 0.55 for every representation except CellOracle, the only method clearly above chance (median ≈ 0.68). By mean rank over the three metrics, the leaders are Geneformer_A, CS-CORE, Gene2vec, RegFormer_S and CellOracle. The attention scores (_A) and CS-CORE have the highest DE odds ratios (median 3.8 for Geneformer_A, 2.6 for CS-CORE) and the largest fraction of regulators with significant overlap (median $\approx 5\%$).

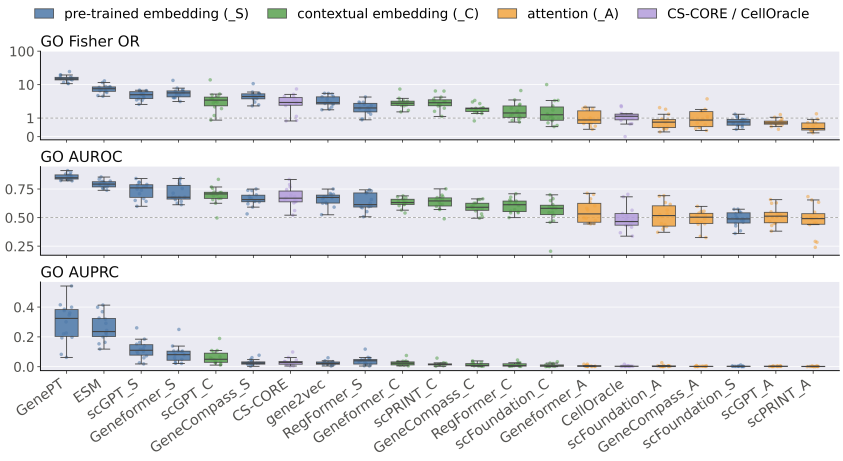


Figure 1: GO-based evaluation across 14 Perturb-seq datasets, measured by Fisher OR, AUROC, and AUPRC. Blue, green, orange, and gray denote static, contextual, attention-derived, and domain-specific methods, respectively; dashed lines indicate random expectation. Methods are ordered by the average rank of their median performance across the three metrics.

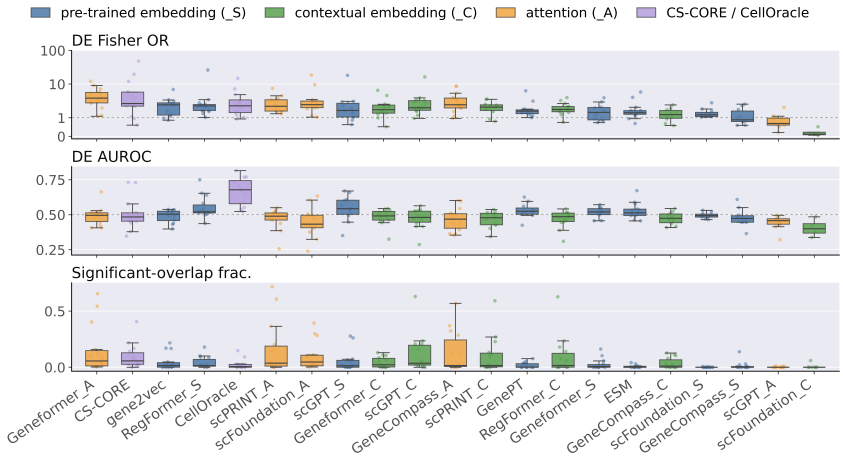


Figure 2: Perturb-seq-based evaluation across 14 Perturb-seq datasets, measured by Fisher OR, AUROC, and Significant Overlap Rate. Blue, green, orange, and gray denote static, contextual, attention-derived, and domain-specific methods, respectively; dashed lines indicate random expectation. Methods are ordered by the average rank of their median performance across the three metrics.

No representation ranks consistently high across both evaluation settings (Figure 3). For each evaluation setting, methods were first ranked separately on each of the three metrics based on their median performance across datasets. The three metric-specific ranks within each setting were then averaged to obtain the overall rank for that setting. The GO-based and perturbation-based rankings are essentially uncorrelated (Spearman $\rho = 0.01$, $n = 21$), and the methods group by how the representation is built. Static gene embeddings (_S) occupy the GO-favored side: GenePT, ESM and scGPT_S are the top three on GO but rank 8th-16th on perturb-seq. Attention scores (_A) and CellOracle occupy the opposite corner: Geneformer_A is the best method on perturb-seq, yet all _A scores rank 15th-21th on GO. Contextual embeddings (_C) fall in between. The near-zero correlation suggests that GO-based and perturbation-based evaluations capture different aspects of a GN, so neither can substitute for the other.

Methods built for expression data strike a better balance. CS-CORE is the only method ranked within the top seven on both benchmarks (GO rank 7, perturb-seq rank 5). Gene2vec, trained on bulk co-expression, sits just behind it (8 and 3) and is solid on both, though best on neither. Cel-

lOracle is the exception: it is the strongest method on DE AUROC but at chance on GO, potentially reflecting its restriction to a TF-target edge space. GenePT and ESM show the other extreme: they dominate GO but are mid-ranked on perturb-seq. Notably, GenePT embeds curated text summaries that describe gene function in GO-like terms, and ESM's protein-sequence similarity captures domain and family homology, which are often associated with shared GO annotations. Neither captures regulation in a specific cell type.

Averaging each method's GO and perturb-seq mean ranks yields a single overall score. The top third (7 of 21 methods) comprises GenePT (6.0), CS-CORE (6.2), scGPT_S (6.2), scGPT_C (7.3), ESM (7.5), gene2vec (7.5), and Geneformer_S (7.7). These methods draw on substantially different sources of gene relationships, including literature-derived knowledge, single-cell pretraining, observational single-cell expression, and protein-sequence information. This diversity suggests that no single representation paradigm consistently dominates GN inference. Notably, the conventional gene2vec representation ranks ahead of several scFM representations and association scores, while GRN-aware scFMs such as RegFormer and scPRINT do not consistently outperform more general-purpose models. Overall, scFM representations and association scores can achieve leading performance under specific evaluation criteria but their advantage is not universal. Their performance remains competitive with, rather than consistently superior to, conventional representation learning and gene relationships inferred directly from observational single-cell data.

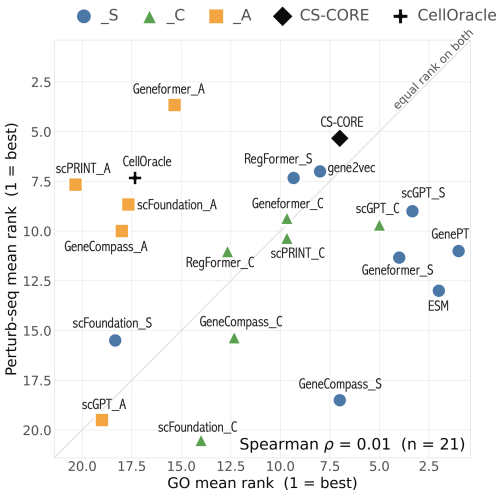


Figure 3: Method comparison by method ranks between GO and Perturb-seq evaluations, with the diagonal indicating equal performance.

3.3 Context-independent versus context-dependent representations and association scores

To assess how dependence on the target dataset affects GN inference, we grouped the evaluated representations and association scores into context-independent and context-dependent categories. The context-independent group contains eight static embeddings, in which each gene is represented by a fixed vector shared across datasets. The context-dependent group contains eleven representations or association scores derived from the target dataset, including contextual embeddings and attention scores. CS-CORE, which is the domain-specific method showing relatively strong performance across metrics, was analyzed separately as a domain-specific observational baseline inferred directly from the target dataset. Multiple-testing correction was performed using the Benjamini-Hochberg procedure across the six comparisons shown in Fig. 4, and adjusted significance was indicated by an asterisk ($q < 0.05$).

Across all three GO-based metrics, context-independent representations consistently outperformed context-dependent representations and association scores across all 14 datasets, indicating stronger alignment with GO-defined functional associations.

In contrast, the perturbation-based metrics revealed a divergence between edge-set enrichment and edge-ranking performance. DE Fisher OR and regulator-level enrichment were higher for context-dependent representations and association scores, with the regulator-level enrichment reaching statistical significance, indicating stronger recovery of perturbation-associated DE relationships. However, DE AUROC favored context-independent representations, indicating better ranking of perturbation-supported pairs among the inferred edges.

Taken together, these results highlight a distinction between edge selection and edge ranking: performance depends on both the reference evidence and the evaluation objective, and no representation performs consistently well across all evaluation settings.

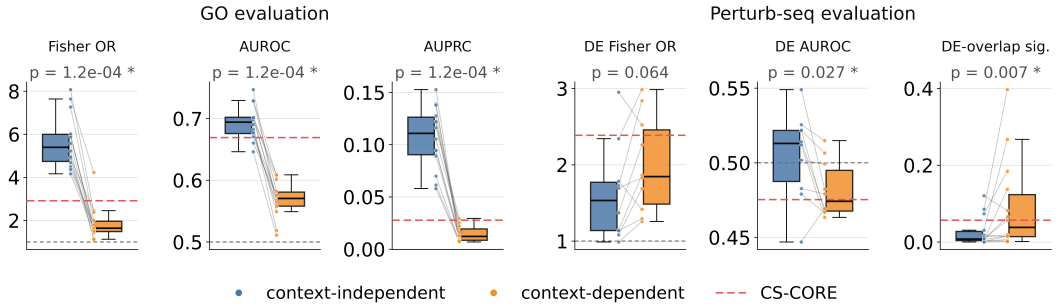


Figure 4: Comparison of context-independent and context-dependent representations or association scores across functional and perturbation-based metrics. For each dataset, each point shows the mean performance across gene representations or association scores within the corresponding group. Means from the same dataset are connected by lines and summarized by boxplots. Gray dashed lines indicate no-information baselines, whereas red dashed lines denote CS-CORE.

3.4 Robustness, Sensitivity, and Computational Cost

GNEval rankings were robust to network size, GO similarity thresholds, and cell subsampling (Appendix Fig. 5), with mean pairwise Spearman correlations of 0.93 across network sizes and 0.90 across 250-4,000-cell subsamples. GO-based rankings were similarly stable across thresholds, supporting the robustness of the main conclusions.

GNEval performance showed no significant systematic differences across broad tissue, cell-system, or perturbation-modality groups after multiple-testing correction (Appendix Table 1), suggesting that the observed patterns were not primarily driven by these dataset characteristics.

Finally, scFM embedding and attention score extraction remained computationally tractable at the evaluated scale, with contextual methods benchmarked on a single NVIDIA A30 GPU and CS-CORE on eight CPU cores (Appendix Fig. 6). Computational cost was therefore not a primary bottleneck of GN inference task.

4 Conclusion and Limitations

Our study highlights two main takeaways. First, gene representations should be selected in a task-aware manner rather than by a single leaderboard: context-independent representations better recover GO-defined functional relationships, whereas context-dependent representations better recovered regulator-level target-set enrichment, with perturbation performance remaining metric-dependent. Notably, rankings across the two benchmarks are nearly uncorrelated. Second, model rankings are highly sensitive to how GN quality is defined and evaluated. The distinct strengths of CS-CORE, gene2vec, and CellOracle across different metrics motivate task-aware representation selection and suggest potential benefits from integrating scFMs and domain-specific methods.

GNEval has several limitations. The analysis is restricted to approximately 1,000 highly variable genes, which improves computational tractability but narrows the evaluated gene space and may underrepresent lowly expressed regulators. Perturbation-derived DE evidence may be limited by statistical power, and only experimentally tested regulators can be evaluated. Because inferred networks are undirected, directed perturbation edges $i \rightarrow j$ are evaluated using symmetric scores S_{ij}, so GNEval measures association rather than regulatory directionality. In addition, overlap between public model pretraining corpora and benchmark datasets cannot be fully excluded, although recent Perturb-seq datasets were preferentially included to reduce this risk. The strong GO performance of GenePT may also partly reflect semantic overlap between biomedical text and ontology annotations. Future work should extend the benchmark to broader gene sets and contexts and more closely integrate domain-specific network inference with scFMs.

AI USE STATEMENT

(This section is **required** and does not count toward the page limit.)

In this work, we used generative AI tools to provide feedback on research methodology and experimental design, assist with the implementation of methods, and support the interpretation of experimental results. We have not used generative AI tools to generate synthetic datasets, develop theoretical models or conceptual frameworks, formulate mathematical claims, assist in the writing of proofs, or clean and reformat datasets, and the remaining tasks requiring disclosure are not applicable to this work. Additionally, we used generative AI tools to create and edit software code, draft and edit parts of the manuscript for readability, brainstorm research ideas, and suggest the structure and presentation of the paper. We have reviewed all AI-assisted work. AI-assisted code was manually reviewed, executed, and tested for correctness, and all reported experimental results were independently generated and verified by us using our computational pipelines. AI-assisted suggestions regarding experimental design and result interpretation were independently evaluated by us against the underlying data and results. All AI-assisted manuscript text was reviewed and revised by us for accuracy and consistency with the reported experiments. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

(This section is **recommended** and does not count toward the page limit.)

This study uses previously published datasets for computational benchmarking and does not involve the collection of new human-subject data or personally identifiable information.

REPRODUCIBILITY STATEMENT

(This section is **recommended** and does not count toward the page limit.)

To facilitate reproducibility, we provide detailed descriptions of the datasets, preprocessing procedures, evaluation protocols, and experimental settings in the Appendix. Additional implementation details and hyperparameter settings are documented where applicable. Code and scripts for data processing, representation extraction, network construction, model evaluation, and downstream analysis are available in an anonymous repository at `https://anonymous.4open.science/r/GNEval-7C8C/`.

REFERENCES

Britt Adamson, Thomas M Norman, Marco Jost, Min Y Cho, James K Nuñez, Yuwen Chen, Jacqueline E Villalta, Luke A Gilbert, Max A Horlbeck, George T Hess, et al. A multiplexed single-cell CRISPR screening platform enables systematic dissection of the unfolded protein response. *Cell*, 167(7):1867–1882, 2016. doi: 10.1016/j.cell.2016.11.048.

Seungbyn Baek, Kyungwoo Song, and Insuk Lee. Single-cell foundation models: bringing artificial intelligence into cell biology. *Experimental & Molecular Medicine*, 57(10):2169–2181, 2025. doi: 10.1038/s12276-025-01547-5. URL `https://doi.org/10.1038/s12276-025-01547-5`.

Timothy Barry, Xuran Wang, John A. Morris, Kathryn Roeder, and Eugene Katsevich. SCEPTRE improves calibration and sensitivity in single-cell CRISPR screen analysis. *Genome Biology*, 22: 344, 2021. doi: 10.1186/s13059-021-02545-2.

Timothy Barry, Kaishu Mason, Kathryn Roeder, and Eugene Katsevich. Robust differential expression testing for single-cell CRISPR screens at low multiplicity of infection. *Genome Biology*, 25: 124, 2024.

Yiqun Chen and James Zou. Simple and effective embedding model for single-cell biology built from ChatGPT. *Nature Biomedical Engineering*, 9:483–493, 2025.

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

Atray Dixit, Oren Parnas, Biyu Li, Jenny Chen, Charles P. Fulco, Livnat Jerby-Arnon, Nemanja D. Marjanovic, Danielle Dionne, Tyler Burks, Raktima Raychowdhury, Britt Adamson, Thomas M. Norman, Eric S. Lander, Jonathan S. Weissman, Nir Friedman, and Aviv Regev. Perturb-seq: Dissecting molecular circuits with scalable single-cell rna profiling of pooled genetic screens. *Cell*, 167(7):1853–1866.e17, 2026/09/22 2016. doi: 10.1016/j.cell.2016.11.038. URL https://doi.org/10.1016/j.cell.2016.11.038.

Nina M. Dräger, Sydney M. Sattler, Cindy Tzu-Ling Huang, Olivia M. Teter, Kun Leng, Sayed Hadi Hashemi, Jason Hong, Giovanni Aviles, Claire D. Clelland, Lihong Zhan, Joe C. Udeochu, Lay Kodama, Andrew B. Singleton, Mike A. Nalls, Justin Ichida, Michael E. Ward, Faraz Faghri, Li Gan, and Martin Kampmann. A CRISPRi/a platform in human iPSC-derived microglia uncovers regulators of disease states. *Nature Neuroscience*, 25(9):1149–1162, 2022. doi: 10.1038/s41593-022-01131-4.

Jingcheng Du, Peilin Jia, Yulin Dai, Cui Tao, Zhongming Zhao, and Degui Zhi. Gene2vec: Distributed representation of genes based on co-expression. *BMC Genomics*, 20:82, 2019.

Gene Ontology Consortium. The gene ontology (GO) database and informatics resource. *Nucleic Acids Research*, 32:D258–D261, 2004.

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

Tongxin He, Xiaoxiao Yang, Yang Tong, Xiaochuan Liu, Yihu Wei, Wenhui Wang, Jiapei Yuan, Yuting Wang, and Yang Yang. Perturbseq.db: An integrated repository for comprehensive analysis of single-cell perturbation data. *Journal of Molecular Biology*, 437(17):169209, 2025. doi: https://doi.org/10.1016/j.jmb.2025.169209. URL https://www.sciencedirect.com/science/article/pii/S002228362500275X.

Siyu Hou, Penghui Yang, Wenjing Ma, Jinxi Xiang, Jade Xiaoqing Wang, Hui Wan, Ying Ma, and Xiang Zhou. Accessible and reproducible deployment reveals the practical boundaries of single-cell foundation models. *bioRxiv*, 2026. doi: 10.64898/2026.01.06.698060. URL https://www.biorxiv.org/content/early/2026/09/02/2026.01.06.698060.

Luni Hu, Hua Qin, Yilin Zhang, Yi Lu, Ping Qiu, Zhihan Guo, Lei Cao, Wenjian Jiang, Yixin Shen, Qianqian Chen, Yanbang Shang, Tianyi Xia, Ziqing Deng, Xun Xu, Hansheng Zhao, Shuangsang Fang, Yuxiang Li, and Yong Zhang. RegFormer: A single-cell foundation model powered by gene regulatory hierarchies. *Nature Communications*, 17:6432, 2026.

Tianxing Hu and Zhi Wei. Transformers for single-cell rna sequencing: a survey. *Briefings in Bioinformatics*, 27(3):bbag332, 2026. doi: 10.1093/bib/bbag332. URL https://doi.org/10.1093/bib/bbag332.

Jérémie Kalfon, Jules Samaran, Gabriel Peyré, and Laura Cantini. scPRINT: Pre-training on 50 million cells allows robust gene network predictions. *Nature Communications*, 16:3607, 2025.

Kenji Kamimoto, Blerta Stringa, Christy M. Hoffmann, Kunal Jindal, Lilianna Solnica-Krezel, and Samantha A. Morris. Dissecting cell identity via network inference and in silico gene perturbation. *Nature*, 614(7949):742–751, 2023. doi: 10.1038/s41586-022-05688-9. URL https://doi.org/10.1038/s41586-022-05688-9.

Kun Leng, Indigo V. L. Rose, Hyosung Kim, Wenlong Xia, Wilber Romero-Fernandez, Brendan Rooney, Mark Koontz, Emmy Li, Yan Ao, Shinong Wang, Mitchell Krawczyk, Julia Tcw, Alison Goate, Yi Zhang, Erik M. Ullian, Michael V. Sofroniew, Stephen P. J. Fancy, Matthew S. Schrag, Ethan S. Lippmann, and Martin Kampmann. CRISPRi screens in human iPSC-derived astrocytes elucidate regulators of distinct inflammatory reactive states. *Nature Neuroscience*, 25(11):1528–1542, 2022. doi: 10.1038/s41593-022-01180-9.

Chen Li, Haoxiang Gao, Yuli She, Haiyang Bian, Qing Chen, Kai Liu, Lei Wei, and Xuegong Zhang. Benchmarking ai models for in silico gene perturbation of cells. *bioRxiv*, 2025. doi: 10.1101/2024.12.20.629581. URL https://www.biorxiv.org/content/early/2025/01/07/2024.12.20.629581.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379:1123–1130, 2023.

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

Mineto Ota, Jeffrey P. Spence, Tony Zeng, Emma Dann, Nikhil Milind, Alexander Marson, and Jonathan K. Pritchard. Causal modelling of gene effects from regulators to programs to traits. *Nature*, 650(8101):399–408, 2026. doi: 10.1038/s41586-025-09866-3. URL https://doi.org/10.1038/s41586-025-09866-3.

Jordan E. Otto, Oana Ursu, Allison P. Wu, Esther B. Winter, Michael S. Cuoco, Sai Ma, Kevin Qian, Brittany C. Michel, Jason D. Buenrostro, Bonnie Berger, Aviv Regev, and Cigall Kadoch. Structural and functional properties of mSWI/SNF chromatin remodeling complexes revealed through single-cell perturbation screens. *Molecular Cell*, 83(8):1350–1367, 2023. doi: 10.1016/j.molcel.2023.03.013.

Stefan Peidli, Tessa D. Green, Ciyue Shen, Torsten Gross, Joseph Min, Samuele Garda, Bo Yuan, Linus J. Schumacher, Jake P. Taylor-King, Debora S. Marks, Augustin Luna, Nils Blüthgen, and Chris Sander. scperturb: harmonized single-cell perturbation data. *Nature Methods*, 21(3):531–540, 2024. doi: 10.1038/s41592-023-02144-y. URL https://doi.org/10.1038/s41592-023-02144-y.

Jean Radig, Robin Droit, Daria Doncevic, Albert Li, Duc Thien Bui, Luis Herfurth, Thaddeus Kühn, and Carl Herrmann. scarchon: a scalable benchmarking framework for assessing single-cell perturbation models. *Genome Biology*, 27(1):162, 2026. doi: 10.1186/s13059-026-04104-z. URL https://doi.org/10.1186/s13059-026-04104-z.

Joseph M. Replogle, Reuben A. Saunders, Angela N. Pogson, Jeffrey A. Hussmann, Alexander Lenail, Alina Guna, Lauren Mascibroda, Eric J. Wagner, Karen Adelman, Gila Lithwick-Yanai, Nika Iremadze, Florian Oberstrass, Doron Lipson, Jessica L. Bonnar, Marco Jost, Thomas M. Norman, and Jonathan S. Weissman. Mapping information-rich genotype-phenotype landscapes with genome-scale perturb-seq. *Cell*, 185(14):2559–2575.e28, 2026/05/12 2022. doi: 10.1016/j.cell.2022.05.013.

Yusuf Roohani, Kexin Huang, and Jure Leskovec. Predicting transcriptional outcomes of novel multigene perturbations with GEARS. *Nature Biotechnology*, 2023.

Yusuf H. Roohani, Tony J. Hua, Po-Yuan Tung, Lexi R. Bounds, Feiqiao B. Yu, Alexander Dobin, Noam Teyssier, Abhinav Adduri, Alden Woodrow, Brian S. Plosky, Reshma Mehta, Benjamin Hsu, Jeremy Sullivan, Chiara Ricci-Tam, Nianzhen Li, Julia Kazaks, Luke A. Gilbert, Silvana Konermann, Patrick D. Hsu, Hani Goodarzi, and Dave P. Burke. Virtual cell challenge: Toward a turing test for the virtual cell. *Cell*, 188(13):3370–3374, 2026/09/25 2025. doi: 10.1016/j.cell.2025.06.008. URL https://doi.org/10.1016/j.cell.2025.06.008.

Yanay Rosen, Yusuf Roohani, Ayush Agrawal, Leon Samotorčan, Stephen R. Quake, Jure Leskovec, and Tabula Sapiens Consortium. Universal cell embedding provides a foundation model for cell biology. *Nature*, 656:183–191, 2026.

Ralf Schmidt, Zachary Steinhart, Madeline Layeghi, Jacob W. Freimer, Raymund Bueno, Vinh Q. Nguyen, Franziska Blaeschke, Chun Jimmie Ye, and Alexander Marson. Crispr activation and interference screens decode stimulation responses in primary human t cells. *Science*, 375(6580):eabj4008, 2022. doi: 10.1126/science.abj4008.

Gavin R. Schnitzler, Helen Kang, Shi Fang, Ramcharan S. Angom, Vivian S. Lee-Kim, X. Rosa Ma, Ronghao Zhou, Tony Y. Zhao, Mehul Mistry, et al. Convergence of coronary artery disease genes onto endothelial cell programs. *Nature*, 626(8000):799–807, 2024. doi: 10.1038/s41586-024-07022-x.

Tuo Shi, Maria Korshunova, Seoyeon Kim, David DeTomaso, Xinhe Zheng, Lavanya Vishvanath, Thokozile Nyasulu, Nhan Huynh, Alexander Sun, Patrick C. Thompson, Yifan Zhang, Emilie M. Wigdor, Narjes Rohani, Salma Ali, Huixian Qiu, Michael Geralt, Ziyan Zhao, Sara Rabhi, Zizhen Yao, Cindy TJ van Velthoven, Joseph R. Nery, Rosa Gomez Castanon, Severin Dicks, Tiffany J. Chen, Joseph R. Ecker, Hongkui Zeng, Grace XY Zheng, Stephan J. Sanders, Laksshman Sundaram, and Xin Jin. Genome-scale functional mapping of the mammalian whole brain with in vivo perturb-seq. *bioRxiv*, 2026. doi: 10.64898/2026.03.16.711480. URL https://www.biorxiv.org/content/early/2026/03/18/2026.03.16.711480.

Chang Su, Zichun Xu, Xinning Shan, Biao Cai, Hongyu Zhao, and Jingfei Zhang. Cell-type-specific co-expression inference from single cell RNA-sequencing data. *Nature Communications*, 14: 4846, 2023.

Sara Sunshine, Andreas S. Puschnik, Joseph M. Replogle, Matthew T. Laurie, Jamin Liu, Beth Shoshana Zha, James K. Nuñez, Jennifer R. Byrum, Andrew H. McCarthy, Calla Schmidt, Rosalvina M. Burrell, Lauren Roberts, Christina Bhatt, et al. Systematic functional interrogation of SARS-CoV-2 host factors using Perturb-seq. *Nature Communications*, 14:6245, 2023. doi: 10.1038/s41467-023-41788-4.

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.

Ruilin Tian, Anthony Abarientos, Jason Hong, Sayed Hadi Hashemi, Rui Yan, Nina Dräger, Kun Leng, Mike A. Nalls, Andrew B. Singleton, Ke Xu, Faraz Faghri, and Martin Kampmann. Genome-wide CRISPRi/a screens in human neurons link lysosomal failure to ferroptosis. *Nature Neuroscience*, 24(7):1020–1034, 2021. doi: 10.1038/s41593-021-00862-0.

Zhiting Wei, Duanmiao Si, Bin Duan, Yicheng Gao, Qian Yu, Zhenbo Zhang, Ling Guo, and Qi Liu. Perturbase: a comprehensive database for single-cell perturbation data analysis and visualization. *Nucleic Acids Research*, 53(D1):D1099–D1111, 2025. doi: 10.1093/nar/gkae858. URL https://doi.org/10.1093/nar/gkae858.

Aaron Wenteler, Martina Occhetta, Nikhil Branson, Victor Curean, Magdalena Huebner, William Dee, William Connell, Siu Pui Chung, Alex Hawkins-Hooker, Yasha Ektefaie, César Miguel Valdez Córdova, and Amaya Gallagher-Syed. Perteval-scfm: Benchmarking single-cell foundation models for perturbation effect prediction. *bioRxiv*, 2025. doi: 10.1101/2024.10.02.616248. URL https://www.biorxiv.org/content/early/2025/05/01/2024.10.02.616248.

Jialu Wu, Qing Ye, Yilin Wang, Renling Hu, Yiheng Zhu, Mingze Yin, Tianyue Wang, Jike Wang, Chang-Yu Hsieh, and Tingjun Hou. Biology-driven insights into the power of single-cell foundation models. *Genome Biology*, 26(1):334, 2025. doi: 10.1186/s13059-025-03781-6. URL https://doi.org/10.1186/s13059-025-03781-6.

Bing Yang, Man Zhang, Yanmei Shi, Bing-Qi Zheng, Chuanping Shi, Daning Lu, Zhi-Zhi Yang, Yi-Ming Dong, Liwen Zhu, Xingyu Ma, Jingyuan Zhang, Jiehua He, Yin Zhang, Kaishun Hu, Haoming Lin, Jian-You Liao, and Dong Yin. Perturbdb for unraveling gene functions and regulatory networks. *Nucleic Acids Research*, 53(D1):D1120–D1131, 01 2025. ISSN 1362-4962. doi: 10.1093/nar/gkae777. URL https://doi.org/10.1093/nar/gkae777.

Xiaodong Yang, Guole Liu, Guihai Feng, Dechao Bu, Pengfei Wang, Jie Jiang, Shubai Chen, Qinmeng Yang, Hefan Miao, Yiyang Zhang, Zhenpeng Man, Zhongming Liang, Zichen Wang, Yaning Li, Zheng Li, Yana Liu, Yao Tian, Wenhao Liu, Cong Li, Ao Li, Jingxi Dong, Zhilong Hu, Chen Fang, Lina Cui, Zixu Deng, Haiping Jiang, Wentao Cui, Jiahao Zhang, Zhaohui Yang, Handong Li, Xingjian He, Liqun Zhong, Jiaheng Zhou, Zijian Wang, Qingqing Long, Ping Xu,

Xin Li, Hongmei Wang, Baoyang Hu, Wei Li, Fei Gao, Jingtao Guo, Leqian Yu, Qi Gu, Weiwei Zhai, Zhengting Zou, Jingqi Yu, Wenhui Wu, Xinxin Lin, Yu Zou, Yongshun Ren, Fan Li, Yixiao Zhao, Yike Xin, Longfei Han, Shuyang Jiang, Kai Ma, Qicheng Chen, Haoyuan Wang, Huanhuan Wu, Chaofan He, Yilong Hu, Shuyu Guo, Yiyun Li, Yuanchun Zhou, Yangang Wang, Xuezhi Wang, Fei Li, Zhen Meng, Zaitian Wang, Huimin He, Shan Zong, Jiajia Wang, Yan Chen, Chunyang Zhang, Chengrui Wang, Ran Zhang, Meng Xiao, Yining Wang, Yiqiang Chen, Yi Zhao, Xin Qin, Jiaxin Qin, Chenhao Li, Zhufeng Xu, Zeyuan Zhang, Xiaoning Qi, Wuliang Huang, Ge Yang, Jing Liu, Yaoru Luo, Qinxuan Luo, Ziwen Liu, Teng Wang, Yiming Huang, Yong Wang, Shihua Zhang, Shirui Li, Kangning Dong, Qunlun Shen, and The X-Compass Consortium. GeneCompass: Deciphering universal gene regulatory mechanisms with a knowledge-informed cross-species foundation model. *Cell Research*, 34:830–845, 2024.

Bin Zhang and Steve Horvath. A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology*, 4(1), 2005. doi: 10.2202/1544-6115.1128.

Jeffrey Zhong, Lechuan Li, Ruth Dannenfelser, and Vicky Yao. Benchmarking gene embeddings from sequence, expression, network, and text models for functional prediction tasks. *bioRxiv*, 2025. doi: 10.1101/2025.01.29.635607. URL https://www.biorxiv.org/content/early/2025/11/10/2025.01.29.635607.

Ronghui Zhu, Emma Dann, Jun Yan, Justine Reyes Retana, Ryunosuke Goto, Reese C. Guitche, Lillian Brixi, Mineto Ota, Austin Hartman, Theodore L. Roth, Ansuman T. Satpathy, Jonathan K. Pritchard, and Alexander Marson. Genome-scale perturb-seq in primary human cd4$^+$ t cells maps context-specific regulators of t cell programs and human immune traits. *Cell*, 2026/09/22 2026. doi: 10.1016/j.cell.2026.08.002. URL https://doi.org/10.1016/j.cell.2026.08.002.

A Complete dataset description

Table 3: Additional information on the Perturb-seq datasets used for evaluation. "Cells × genes" reports the input matrix size; "Response genes" denotes the per-dataset response universe, consisting of the top 1,000 highly variable genes plus force-included targeted regulators; "Tested Regulators" is the number of regulators retained after filtering; "ctrl cells" is the number of non-targeting control cells; and "Sig. DE edges" is the number of significant regulator–gene pairs identified by SCEPTRE (BH-FDR < 0.05 and $|\log_2 \mathrm{FC}| > 0.25$).

Data	Perturbation	Cell type	Cells × genes	Response genes	Tested Regulators	ctrl cells	Sig. DE edges
Adamson et al. (2016)	CRISPRi	K562	65,145 × 32,643	1083	87	7,629	8,386
Norman et al. (2019)	CRISPRa	K562	69,686 × 33,694	1073	102	11,855	9,775
Replogle et al. (2022)	CRISPRi	K562	310,385 × 8,563	1000	135	10,691	11,781
Replogle et al. (2022)	CRISPRi	RPE1	247,914 × 8,749	1000	150	11,485	24,615
Schmidt et al. (2022)	CRISPRa	primary T cells	61,041 × 36,601	1020	72	4,000	6,220
Sunshine et al. (2023)	CRISPRi	Calu-3	35,729 × 33,549	1174	174	3,541	1,334
Otto et al. (2023)	CRISPRko	MOLM-13	31,069 × 36,601	1023	24	1,367	1,244
Schnitzler et al. (2024)	CRISPRi	endothelial	214,449 × 36,601	1000	72	5,506	98
Tian et al. (2021)	CRISPRi	iPSC neuron	32,300 × 33,514	1177	181	437	143
Tian et al. (2021)	CRISPRa	iPSC neuron	21,193 × 33,514	1081	95	434	999
Dräger et al. (2022)	CRISPRi	iPSC microglia	10,377 × 33,514	1035	36	817	1,375
Leng et al. (2022)	CRISPRi	iPSC astrocyte	7,933 × 33,514	1025	28	273	168
Zhu et al. (2026)	CRISPRi	primary CD4 T cells	7,282,273 × 18,130	1000	680	309,554	51,311
Shi et al. (2026)	CRISPR (*in vivo*, mouse)	excitatory neuron	2,246,807 × 16,684	1000	180	199,649	186

B Data preprocessing

B.1 Overview

We apply two principles consistently across all datasets:

1. **Single perturbations only.** Combination and multi-gene perturbations are excluded.
2. **Low-MOI representation with non-targeting controls.** A control cell is defined as a cell whose detected guides are all non-targeting. Cells with no detected guide are excluded rather than treated as controls.

B.2 COMMON PREPROCESSING PIPELINE

For each dataset, SCEPTRE DE testing, GN inference, and downstream evaluation are all restricted to the same filtered gene universe. (Peidli et al., 2024; Wei et al., 2025; Yang et al., 2025; He et al., 2025)

B.2.1 GUIDE CALLING AND SINGLE-PERTURBATION SELECTION

The datasets differ in the extent to which guide assignments are provided by the original authors. We therefore use three guide-processing strategies, all of which produce the same final representation: one retained cell associated with at most one perturbation target.

Author-provided per-cell calls. For datasets with author-provided perturbation assignments, we use those calls directly and retain only unambiguous single-guide or singlet cells. Cells annotated as multiplets, ambiguous assignments, or combinatorial perturbations are excluded.

Guide-capture UMI tables without per-cell calls. For datasets providing only guide-by-cell UMI counts, perturbation assignments are inferred using a shared procedure. A guide is considered present only if it has at least 3 UMIs, duplicate cell–guide entries are collapsed by retaining the maximum UMI count, and a cell is retained only if exactly one guide remains after filtering. Cells with no qualifying guide or multiple qualifying guides are excluded, and retained barcodes are intersected with the dataset's filtered cell set.

High-MOI screens reduced to low-MOI subsets. For high-MOI screens, we retain only cells compatible with a single-perturbation representation and satisfying the available singlet-quality criteria. Cells targeting multiple genes are excluded, whereas cells containing only non-targeting guides are retained as controls. When multiple guides target the same gene, they are collapsed into a single construct-level perturbation for downstream SCEPTRE analysis.

B.2.2 GENE AND CELL QUALITY CONTROL

Quality control is applied to the raw count matrix produced by each dataset adapter. Duplicated gene symbols are removed first, because duplicated symbols cannot be mapped unambiguously to gene embeddings, GO annotations, or SCEPTRE responses. The remaining QC steps are applied in the following order:

1. **filter_genes(min_cells = 3).** Genes detected in fewer than 3 cells are removed.
2. **filter_cells(min_genes = 200).** Cells expressing fewer than 200 genes are removed.
3. **Depth covariates.** Before any gene subsetting, each cell's total UMI count and number of detected genes are recorded from the full transcriptome.
4. **Mitochondrial fraction.** The mitochondrial UMI fraction is computed over the full gene set using genes matching ^MT- (case-insensitive). Cells above the threshold are removed. The default threshold is 0.1; it is increased to 0.2 for Replogle_K562 and 0.25 for Schmidt because, in these datasets, a threshold of 0.1 falls within the main body of the observed distribution rather than its upper tail.
5. **Depth quantile clipping.** Cells outside the $[0.01, 0.99]$ quantile range of either depth covariate are removed, matching the intent of SCEPTRE's response_n_umis_range and response_n_nonzero_range defaults. These internal SCEPTRE filters are disabled because they would otherwise be applied after variable-gene selection. Both QC masks are computed on the original distribution and then combined, avoiding sequential clipping of an already-trimmed population.

B.2.3 NORMALIZATION AND VARIABLE-GENE SELECTION

Counts from retained cells are normalized to 10^4 UMIs per cell and $\log(1 + x)$-transformed, after which the top 1,000 highly variable genes are selected using the Seurat procedure. Perturbation target genes are force-included when feasible, so the response universe consists of the top-1,000 highly variable genes together with QC-passing targets. For screens with more than 500 targets,

force-inclusion is omitted for computational tractability, and regulators are restricted to targets overlapping the top-1,000-gene universe.

B.2.4 TEST GRID SUPPLIED TO SCEPTRE

The response set is the gene universe defined in Section B.2.3. Raw counts are supplied to SCEPTRE; the normalized expression matrix is used only for variable-gene selection, whereas SCEPTRE fits its own negative-binomial model to the count data. Regulators are restricted to perturbation targets that belong to the response universe and have at least **20** retained cells after QC. Cells with zero counts across all response genes are excluded because the model includes the log response depth. The discovery set is the full regulator $\times$ response grid. Covariates include the two log-transformed depth measures defined in Section B.2.2 and, for datasets assembled from multiple 10x libraries or samples, a batch factor derived from sample identity.

B.3 DATASET-SPECIFIC PREPROCESSING

All datasets follow the shared preprocessing pipeline described above. Table 4 summarizes only dataset-specific processing decisions.

C DIFFERENTIAL-EXPRESSION REFERENCE

We constructed the perturbational reference for each dataset using SCEPTRE (Barry et al., 2024) (v0.10.3). All differential-expression tests were performed on raw UMI counts; normalized and log-transformed expression values were used only for highly variable gene selection (Section B.2.3).

C.1 SCEPTRE ANALYSIS

We analyzed all datasets under the low-MOI setting. Perturbation assignments were taken from the single-guide calls defined in Section B.2.1; guides targeting the same gene were combined using the SCEPTRE `union` strategy. Non-targeting-guide cells served as controls, and cells without a guide call were excluded upstream.

Tests were adjusted for log-transformed sequencing depth and detected-gene count, both computed from the full transcriptome before gene selection. A sample or library batch covariate was additionally included where applicable. Because cell-level QC had already been performed using the full transcriptome (Section B.2.2), SCEPTRE's response-matrix-based cellwise QC thresholds were disabled. We retained its pairwise sparsity filter, requiring at least seven cells with nonzero expression of the response gene in both the treatment and non-targeting control groups.

All regulator–response pairs in the analysis universe were tested using two-sided low-MOI SCEPTRE tests. We additionally ran SCEPTRE's negative-control calibration check for each dataset.

C.2 DEFINITION OF PERTURBATIONAL EDGES

We did not use SCEPTRE's default significance calls. Instead, p-values were adjusted by the Benjamini–Hochberg procedure separately within each regulator across its tested response genes. A regulator–response pair (r, g) was considered a DE edge when

$$q_{rg} < 0.05 \qquad \text{and} \qquad |\log_2 \mathrm{FC}_{rg}| > 0.25.$$

Self-pairs ($r = g$) were excluded.

Because the inferred networks evaluated in this study are undirected and unsigned, DE edges were converted to unordered gene pairs irrespective of effect direction or sign. Evaluation was restricted to the testable subgraph: pairs for which at least one endpoint was a regulator subjected to perturbational testing. Pairs between two non-regulators were therefore treated as untested rather than as negative examples.

Table 4: Dataset-specific preprocessing of the Perturb-seq datasets used for evaluation.

Dataset	Accession	Dataset-specific preprocessing
Adamson et al. (2016)	GSE90546	Only the largest single batch (approximately 65k cells) is used; the three available batches are highly imbalanced and are not combined.
Norman et al. (2019)	GSE133344	Cells containing only non-targeting guides are treated as controls. Cells targeting exactly one gene are retained, whereas two-gene combinatorial perturbations are excluded.
Replogle et al. (2022)	Figshare 20029387	The authors' per-cell guide and target assignments are used directly, with non-targeting guides mapped to NT. Each retained cell carries a single guide.
Replogle et al. (2022)	Figshare 20029387	The authors' per-cell guide and target assignments are used directly, with non-targeting guides mapped to NT. Each retained cell carries a single guide.
Schmidt et al. (2022)	GSE190604	Only single-guide cells are retained, and non-targeting guides are mapped to NT.
Sunshine et al. (2023)	GSE208240	Cells without a guide assignment are excluded. Non-targeting-only cells are controls, single-target cells are retained, and multi-target cells are excluded to obtain a low-MOI-compatible subset.
Otto et al. (2023)	GSE200201	Only the single-guide RNA-seq samples are used, and cells are retained only when assigned exactly one guide.
Schnitzler et al. (2024)	GSE210681	Only high-confidence single-guide cells are retained. Non-targeting and safe-targeting controls are mapped to NT.
Tian et al. (2021)	GSE152988	The four CRISPRi libraries are analyzed separately from CRISPRa. Author-provided single-guide assignments are used directly, non-targeting guides are mapped to NT, and library identity is included as the batch covariate.
Tian et al. (2021)	GSE152988	The two CRISPRa libraries are analyzed separately from CRISPRi. Author-provided single-guide assignments are used directly, non-targeting guides are mapped to NT, and library identity is included as the batch covariate.
Dräger et al. (2022)	GSE178317	Author-provided per-cell guide assignments are used, retaining only cells classified as singlets. Non-targeting controls are mapped to NT.
Leng et al. (2022)	GSE182308	Only the two vehicle-condition wells are used. Guide assignments are inferred from guide-capture UMI counts using the shared UMI-based procedure, and non-targeting guides are mapped to NT; well identity is used as the batch covariate.
Zhu et al. (2026)	—	Only resting-condition, single-guide cells are retained across the four donors, with donor identity used as the batch covariate. Because of the dataset size, cell QC and highly variable gene selection are performed using a streaming implementation equivalent to the shared pipeline.
Shi et al. (2026)	—	Only excitatory neurons with exactly one guide and a singlet call are retained. Non-targeting and safe-targeting guides are mapped to NT. Mouse genes and targets are mapped to human orthologs, and the large expression matrix is processed using the streaming implementation of the shared pipeline.

D REPRESENTATION EXTRACTION

For each method, we use the released checkpoint and follow the original representation-extraction procedure whenever available. Table 5 summarizes the checkpoint, readout, aggregation, dimensionality, and coverage of every representation.

Shared conventions. All representations are aligned to the dataset-specific gene universe defined in Appendix B. Genes without a valid representation are flagged and excluded from downstream network construction rather than treated as zero vectors. Contextual and attention representations use the same stratified 1,000-cell post-QC subsample per dataset; sample-size experiments use nested subsets of 250–4,000 cells. Static representations require no cell input.

Table 5: Representations used in the benchmark. **S**: static embedding; **C**: contextual embedding; **A**: attention. Coverage is the median fraction of the dataset-specific gene panel represented across the evaluated datasets.

Representation	Checkpoint	Layer / tensor	Aggregation	Dim	Coverage
GenePT (S)	GenePT v2, `ada_text`	text-encoder output	—	1536	84.0%
gene2vec (S)	`dim_200_iter_9`	word2vec vector	—	200	80.5%
ESM-2 (S)	ESM-2 15B (via UCE)	final repr. layer	mean residues, then isoforms	5120	77.9%
scGPT (S)	`scGPT_human`	token emb. + `enc_norm`	—	512	87.9%
scGPT (C)	`scGPT_human`	layer 12 hidden state	mean over cells	512	87.9%
scGPT (A)	`scGPT_human`	layer 12 q, k	mean over heads/cells	—	87.9%
Geneformer (S)	V2-316M	input token emb. table	—	1152	74.6%
Geneformer (C)	V2-104M	layer 12 hidden state	mean over detecting cells	768	56.9%
Geneformer (A)	V2-104M	layer 12 attention	mean over heads/co-detecting cells	—	56.9%
scFoundation (S)	`models.ckpt`	`pos_emb` table	—	768	74.7%
scFoundation (C)	`models.ckpt`	decoder output	mean over cells	512	74.7%
scFoundation (A)	`models.ckpt`	encoder layer 12 attention	mean over heads/co-expressing cells	—	56.8%
GeneCompass (S)	GeneCompass_Base	word-emb. table	—	768	77.1%
GeneCompass (C)	GeneCompass_Base	layer 12 hidden state	mean over expressing cells	768	57.0%
GeneCompass (A)	GeneCompass_Base	layer 12 attention	mean over heads/co-expressing cells	—	57.0%
RegFormer (S)	RegFormer-10k	encoder emb. table	—	512	87.9%
RegFormer (C)	RegFormer-10k	final Mamba block	mean over expressing cells	512	54.2%
scPRINT (C)	`medium-v1.5`	final block gene emb.	mean over expressing cells	256	58.2%
scPRINT (A)	`medium-v1.5`	layer 8 shared-QK	official GNInfer aggregation	—	58.2%
CS-CORE	—	test statistic	all control cells	—	100%

For attention-based networks, gene–gene scores are symmetrized by averaging the two directions. When a model tokenizes only expressed genes, each pair is averaged over cells in which both genes are present, and pairs observed in fewer than 10 cells are discarded. Embeddings are otherwise used as extracted, with no additional centering, whitening, or normalization before cosine similarity.

D.1 Static gene representations

GenePT. We use the released GenePT v2 gene embeddings (Chen & Zou, 2025), generated by applying `text-embedding-ada-002` to NCBI RefSeq gene summaries. The released 1,536-dimensional vectors are used directly, with no additional preprocessing.

gene2vec. We use the released `gene2vec_dim_200_iter_9.txt` embeddings (Du et al., 2019), trained by skip-gram on gene co-expression relationships derived from bulk GEO data. The 200-dimensional vectors are used as published.

ESM-2. We use the gene-level ESM-2 15B representations distributed with UCE (Lin et al., 2023; Rosen et al., 2026). Protein representations are obtained from the final ESM-2 layer, mean-pooled over residues and then averaged across protein isoforms to obtain one vector per gene.

scGPT. For the static scGPT representation (Cui et al., 2024), we read the gene encoder from the released `scGPT_human` checkpoint. The representation is the token-embedding table followed by the encoder's own layer normalization.

Geneformer. Static Geneformer representations (Theodoris et al., 2023) are read directly from the V2-316M input embedding table. Special tokens are removed and Ensembl identifiers are mapped to gene symbols using the released dictionary.

scFoundation. For scFoundation (Hao et al., 2024), the static representation is the released `pos_emb` gene-identity table aligned to the model's fixed 19,264-gene vocabulary.

GeneCompass. Static GeneCompass representations (Yang et al., 2024) are taken from the GeneCompass_Base word-embedding table, restricted to human gene tokens and mapped to gene symbols using the repository-provided dictionary.

RegFormer. Static RegFormer representations (Hu et al., 2026) are taken directly from the RegFormer-10k gene-token embedding table. Its vocabulary matches that of the scGPT whole-human checkpoint.

D.2 CONTEXTUAL REPRESENTATIONS AND ATTENTION

scGPT. Raw counts from the shared cell subsample are processed with scGPT's `Preprocessor`: library-size normalization to 10^4, $\log(1 + x)$ transformation, and 51-bin discretization. The shared gene panel is retained rather than re-selecting HVGs. Contextual gene embeddings are the final-layer hidden states averaged across cells. Attention is reconstructed from the final layer's queries and keys using the model's released gene network procedure, then averaged across heads and cells.

Geneformer. Cells are tokenized with the official V2 `TranscriptomeTokenizer`, which ranks detected genes after normalization by pretraining median expression. Contextual representations are final-layer hidden states averaged over cells in which each gene is detected. Attention uses the final-layer post-softmax attention and is averaged over heads and co-detecting cells.

scFoundation. Inputs follow the released inference workflow: $\log(1 + x/\text{total} \times 10^4)$ on the model's fixed gene axis together with the two resolution tokens used during pretraining. Contextual representations are taken from the decoder output and averaged across cells. Attention is taken from the final encoder layer and averaged across heads and cells in which both genes are expressed.

GeneCompass. GeneCompass inputs consist of nonzero genes ordered by normalized expression using the model's released preprocessing. Contextual representations are final-layer hidden states averaged across cells expressing each gene. Attention is the model's final-layer softmax attention, averaged over heads and co-expressing cells. We follow the repository's released embedding-generation procedure for loading its prior-knowledge embeddings.

RegFormer. We use the repository's `infer_dataset` pipeline with its released preprocessing and graph-based token ordering. Contextual gene embeddings are taken from the final Mamba block and averaged over cells in which each gene is expressed. RegFormer provides no attention representation because its encoder uses Mamba blocks rather than self-attention.

scPRINT. Raw counts are aligned to the checkpoint's human gene axis and passed through the released scPRINT inference pipeline (Kalfon et al., 2025). Contextual embeddings are the final block's gene representations averaged over expressing cells. Attention-based networks use the official `GNInfer` shared-query/key procedure, which averages queries and keys across cells before computing the gene–gene score.

D.3 CS-CORE BASELINE

CS-CORE (Su et al., 2023) is fit directly to raw UMI counts from all post-QC non-targeting control cells on the shared gene panel. The signed CS-CORE test statistic is used directly as the gene–gene edge score, without an intermediate embedding or similarity calculation.

D.4 CELLORACLE BASELINE

CellOracle (Kamimoto et al., 2023) infers directed TF–target edges by combining a promoter motif-based prior with bagging ridge regression on imputed single-cell expression, and retains significant edges ranked by the absolute regression coefficient. We reproduce this procedure using the same significance filtering and edge ranking, then symmetrize directed edges by keeping the stronger direction and evaluate the resulting network at the available sizes.

E NETWORK CONSTRUCTION

Each representation in Appendix D is converted into an undirected gene network by defining a gene–gene score, restricting the gene space to genes represented by the method, and selecting the strongest edges.

E.1 GENE–GENE SCORES

For static and contextual gene embeddings, edge scores are cosine similarities:

$$S_{ij} = \frac{\mathbf{x}_i^\top \mathbf{x}_j}{\|\mathbf{x}_i\| \, \|\mathbf{x}_j\|}.$$

Representations are used as extracted, without additional centering, standardization, or whitening. Attention matrices and CS-CORE already provide gene–gene scores and are therefore used directly without an additional similarity calculation.

Alternative similarity constructions, including principal-component residualization, rank-normalized cosine similarity, and graphical lasso, are evaluated separately as robustness analyses.

E.2 GENE SPACE AND NETWORK CONVENTIONS

For each method and dataset, the network is constructed on the intersection between the dataset's response-gene universe and the genes for which that method provides a valid representation. Genes outside this intersection are not imputed or treated as zero vectors.

Self-pairs are excluded. All evaluated networks are undirected; asymmetric attention matrices are symmetrized as

$$\tilde{A} = \frac{1}{2}(A + A^\top).$$

For signed scores, edge selection uses absolute magnitude, $|S_{ij}|$, while the original sign is retained in the stored edge list.

Because representation coverage differs across methods, evaluation metrics are computed on the common dataset-level response universe rather than separately on each method's represented gene set.

E.3 EDGE SELECTION AND NETWORK SIZE

For a target network size K, we select the K gene pairs with the largest absolute scores from the strict upper triangle:

$$E_K = \mathrm{TopK}_{i<j}\{|S_{ij}|\}.$$

This rank-based construction allows representations with different numerical score scales, including cosine similarity, attention, and CS-CORE statistics, to be compared at the same edge count.

The main benchmark evaluates

$$K \in \{5{,}000,\ 10{,}000,\ 20{,}000\},$$

with reported results summarized across these network sizes. Because methods represent different numbers of genes, a fixed edge count implies different network densities. We therefore repeat the benchmark using fixed densities

$$\rho \in \{0.05,\ 0.10,\ 0.15\}$$

as a robustness analysis.

Exact-zero or otherwise unavailable gene pairs are not promoted to edges. All methods have sufficient valid candidate pairs to reach the three network sizes used in the main benchmark.

F GO-BASED EVALUATION

We evaluate whether predicted gene pairs share functional annotation using the gene-to-GO mapping distributed with GEARS (Roohani et al., 2023). The same fixed annotation snapshot is used for all datasets and methods.

F.1 GO REFERENCE

For each dataset, let $\mathcal{G}$ denote its response-gene universe and $\mathrm{GO}(g)$ the set of GO terms annotated to gene g. For every annotated gene pair, we compute the Jaccard similarity

$$J(g_1, g_2) = \frac{|\mathrm{GO}(g_1) \cap \mathrm{GO}(g_2)|}{|\mathrm{GO}(g_1) \cup \mathrm{GO}(g_2)|}.$$

A pair is considered a GO-positive if $J > \tau$. We use

$$\tau \in \{0.1,\, 0.3\},$$

with $\tau = 0.1$ as the primary reference and $\tau = 0.3$ as a stricter robustness analysis. Genes absent from the annotation source are excluded from the GO reference.

F.2 EVALUATION METRICS

For each predicted network, AUROC and AUPRC are computed over the selected edges, using $|S_{ij}|$ as the ranking score and GO membership as the binary label. AUROC has chance level 0.5; for AUPRC we additionally record the positive prevalence among the selected edges.

Set-level enrichment is measured with a one-sided Fisher exact test comparing selected versus unselected pairs and GO-positive versus GO-negative pairs. The reported effect size is the Fisher odds ratio.

Because methods differ in gene coverage, the Fisher background is defined over the dataset's full response-gene universe rather than each method's represented gene subset. This keeps the reference prevalence fixed across methods within a dataset. AUROC and AUPRC are unaffected by this background choice because they are computed only over selected edges.

All GO evaluations use unordered, non-self gene pairs, consistent with the undirected networks defined in Appendix E.

G PERTURBATIONAL EVALUATION

We evaluate whether predicted gene pairs are supported by perturbation-induced differential expression. Unlike the GO evaluation, this reference is based on experimental intervention: one gene is perturbed and the transcriptional response of other genes is measured.

G.1 DIFFERENTIAL-EXPRESSION REFERENCE

For each dataset, tested regulators are perturbation targets that belong to the response-gene universe and have at least 20 retained perturbed cells. Each regulator is tested against all response genes using SCEPTRE (Barry et al., 2024) on raw UMI counts.

Tests are two-sided and use non-targeting cells as controls where available. Covariates include log sequencing depth, log detected-gene count, and batch where applicable; depth covariates are computed from the full transcriptome before gene selection. A regulator–response pair is tested only when the response gene is detected in at least seven treatment and seven control cells.

A pair is called differentially expressed when

$$q < 0.05 \qquad \text{and} \qquad |\log_2 \mathrm{FC}| > 0.25,$$

where Benjamini–Hochberg correction is applied separately within each regulator. Self-pairs are excluded.

Because the evaluated networks are undirected, a significant (regulator, response) result is converted to the unordered pair $\{r, g\}$.

G.2 TESTABLE PAIR SPACE

Perturbational evaluation is restricted to the *testable subgraph*: gene pairs for which at least one endpoint is a tested regulator. Pairs between two non-regulators cannot be supported by the perturbation experiment and are therefore treated as untested rather than as negatives.

For a response universe of size p and regulator set R, the testable pair space contains

$$N_t = \binom{p}{2} - \binom{p - |R|}{2}$$

unordered pairs.

G.3 Evaluation metrics

We evaluate perturbational agreement at three complementary levels.

Set-level enrichment. Predicted testable edges are compared with DE-supported pairs using a one-sided Fisher exact test over the testable subgraph. We report the odds ratio, with an odds ratio of 1 corresponding to random enrichment.

Edge-level ranking. AUROC is computed over predicted testable edges, using the network edge score as the ranking variable and DE support as the binary label. Chance performance is 0.5.

Regulator-level enrichment. For each tested regulator r, we compare its predicted neighborhood $\mathcal{N}(r)$ with the set of genes $D(r)$ that respond significantly to perturbation of r. A one-sided Fisher test is performed for each regulator, followed by BH correction across regulators. We report

$$\text{enrichment rate} = \frac{\#\{r : \mathrm{FDR}(r) < 0.05\}}{\#\{r \text{ tested}\}},$$

the fraction of regulators whose predicted neighborhoods are significantly enriched for their measured transcriptional responses.

G.4 Filtering and aggregation

Method–dataset–network-size combinations containing fewer than 10 DE-supported testable edges are excluded from the network-level AUROC and odds ratio summaries. Regulator-level evaluation requires at least 10 tested regulators.

For each metric, values are first averaged across the three network sizes within each method–dataset pair. Method groups are then compared across datasets using two-sided paired Wilcoxon signed-rank tests. A dataset enters a group comparison only when at least 75% of methods in both groups have valid values. Multiple comparisons are corrected using Benjamini–Hochberg where applicable.

G.5 Interpretation

These metrics provide complementary views of perturbational agreement: the Fisher odds ratio measures whether the predicted edge set is enriched for perturbation-supported pairs, AUROC measures whether the network ranks such pairs highly, and the regulator-level enrichment rate measures whether individual predicted neighborhoods recover experimentally responsive genes.

H Statistical analysis

The dataset is the unit of statistical inference throughout. For each method–dataset pair, metrics are first averaged across the three network sizes. For representation-class comparisons, method-level values are then averaged within each comparison group, yielding one paired value per dataset. Methods and network sizes are therefore not treated as independent replicates.

Paired comparisons, including context-independent versus context-dependent representations, use two-sided Wilcoxon signed-rank tests. We report the median paired difference and matched-pairs rank-biserial correlation,

$$r = \frac{W^+ - W^-}{W^+ + W^-}.$$

Comparisons between disjoint dataset groups, such as tissue, cell system, or perturbation type, use two-sided Mann–Whitney U tests with Cliff's δ as the effect size. Associations between method rankings are summarized using Spearman correlation and are treated as exploratory.

Benjamini–Hochberg correction is applied separately within each predefined family of related hypothesis tests, with $q < 0.05$ considered significant. All tests are two-sided.

The default paired sample size is $n = 14$ datasets. For perturbational pair-level metrics, method–dataset cells containing fewer than 10 DE-supported testable edges are excluded. A dataset contributes to a group comparison only when at least 75% of methods in both groups retain valid values; consequently, the effective sample size is reported separately for each test.

Because the number of datasets is small, non-significant results are interpreted as insufficient evidence for a detectable effect rather than as evidence of equivalence.

I PER-DATASET BENCHMARK RESULTS

Tables 6–11 report the per-dataset results underlying Figures 1 and 2. Each value is averaged over networks with $K \in \{5{,}000, 10{,}000, 20{,}000\}$ edges; "–" denotes an unscoreable method–dataset combination. The final column reports the median across datasets. Rows are ordered by mean rank separately within the GO and perturbational blocks.

GO results use the Jaccard > 0.3 reference. Dataset abbreviations are: Adam (Adamson), Norm (Norman), K562/RPE1 (Replogle), Schm (Schmidt), SARS (Sunshine SARS-CoV-2), MOLM (Otto MOLM-13), Endo (Schnitzler endothelial), Ti-i/Ti-a (Tian2020 CRISPRi/CRISPRa), Drag (Drager2021 microglia), Leng (Leng2021 astrocyte), GWCD (GWCD4i), and WBrn (WholeBrain excitatory).

Origin of missing perturbation metrics. Missing entries ("–") in Tables 9– 11 indicate cases in which the corresponding perturbation-based metric is undefined, rather than a score of zero. This occurs primarily when a dataset contains too few DE relationships within the evaluated gene set. Of the 46 missing entries, 39 arise from the four datasets with the lowest DE density, and 35 contain only 1–9 DE-supported pairs while the remaining 11 contain none. MOLM-13 represents a related case: although its overall DE density is moderate, only 24 perturbed regulators remain in the evaluated gene set, leaving just 4.6% of gene pairs testable.

Missingness also depends on whether an inferred network contains edges incident to experimentally perturbed regulators. Across representations, the median fraction of such edges ranges from 0.04 to 0.42; representations with lower regulator coverage therefore produce fewer evaluable DE-supported pairs and more undefined metrics. Consequently, missing entries are concentrated in low-signal datasets and low-coverage representations rather than occurring at random. In Table 11, a reported value of 0 is distinct from a missing entry: it indicates that the metric was evaluable but no regulator satisfied the enrichment criterion.

J ROBUSTNESS ANALYSES

We evaluated the sensitivity of representation rankings to four benchmark choices: network size, GO similarity threshold, contextual cell-subsample size, and similarity definition. Rankings were recomputed under each setting and compared using Spearman correlation. Table 12 summarizes the results, and Figure 5 visualizes the first three analyses.

Table 12: Stability of representation rankings across benchmark settings. Agreement is the mean pairwise Spearman correlation between rankings; rank span summarizes movement across settings.

Axis	Settings	Representations	Agreement	Rank span (med / max)
Network size	5,000 / 10,000 / 20,000 edges	20	0.927 [0.918]	2.2 / 6
GO reference	Jaccard > 0.1 / > 0.3	20	0.912	0.8 / 6
Cells subsampled	250 / 500 / 1,000 / 2,000 / 4,000	11	0.947 [0.917]	2.0 / 3
Similarity	cosine / $-$PC1 / gaussianized	8	0.996 [0.994]	0.0 / 0.5

Table 6: Per-dataset GO Fisher odds ratio (Jaccard > 0.3).

Representation	Adam	Norm	K562	RPE1	Schm	SARS	MOLM	Endo	Ti-i	Ti-a	Drag	Leng	GWCD	WBrn	Median
GenePT	10.68	14.42	15.69	10.97	13.94	24.56	19.45	11.67	16.40	14.79	13.87	16.56	15.03	19.43	**14.91**
ESM	5.82	6.85	4.48	4.70	6.20	8.23	12.94	6.56	11.55	8.21	8.77	6.69	8.23	8.64	**7.53**
scGPT_S	4.20	3.48	4.02	5.24	5.16	3.80	6.41	4.79	6.83	3.37	6.64	6.42	5.47	2.57	**4.97**
Geneformer_S	5.13	4.10	3.13	4.12	6.54	7.10	6.15	4.88	13.35	5.58	6.35	5.73	7.86	4.03	**5.65**
scGPT_C	3.20	2.17	3.67	3.82	2.84	2.99	13.87	5.23	1.92	1.19	4.38	4.90	3.72	0.89	**3.44**
GeneCompass_S	4.49	4.30	2.31	3.00	5.54	5.24	5.34	2.47	10.68	3.73	5.10	4.17	6.01	3.83	**4.39**
CS-CORE	2.60	2.37	4.20	2.90	4.42	7.40	2.79	5.11	1.12	0.85	3.89	2.93	3.50	0.87	**2.92**
gene2vec	2.67	2.13	2.63	2.73	3.70	5.44	5.50	5.17	2.84	1.83	4.35	4.28	2.93	1.77	**2.89**
RegFormer_S	2.04	1.31	1.96	1.71	2.96	2.80	4.27	3.00	1.65	0.96	2.61	2.35	1.49	0.91	**2.00**
Geneformer_C	3.18	3.27	2.96	2.82	3.21	2.72	7.35	3.90	2.33	1.70	2.37	2.46	2.56	1.53	**2.77**
scPRINT_C	2.47	2.68	3.04	3.19	2.54	6.41	6.45	4.23	2.30	1.59	3.11	3.65	1.97	1.12	**2.86**
GeneCompass_C	1.57	1.37	1.90	1.95	1.44	1.85	3.03	2.71	3.37	0.85	2.01	1.94	2.00	1.61	**1.92**
RegFormer_C	2.43	0.88	1.16	0.78	1.50	0.93	6.64	3.49	1.62	1.17	1.02	2.58	1.34	1.98	**1.42**
scFoundation_C	1.31	1.11	0.60	0.96	0.74	1.74	1.30	0.54	10.02	3.38	0.86	1.22	3.33	2.29	**1.26**
Geneformer_A	1.94	1.86	0.92	0.73	1.00	2.13	0.54	1.90	0.71	0.74	0.94	0.87	0.56	0.40	**0.90**
CellOracle	1.17	2.19	0.98	1.31	1.02	1.28	1.08	0.00	1.31	2.33	0.68	0.70	0.88	1.37	**1.13**
scFoundation_A	1.31	0.78	0.94	0.42	0.48	0.86	2.05	1.83	0.78	0.26	0.51	0.53	0.40	0.89	**0.78**
GeneCompass_A	1.57	1.82	0.87	0.64	0.91	1.57	3.74	1.09	1.68	0.51	0.33	0.43	0.73	0.49	**0.89**
scFoundation_S	0.93	0.88	0.72	0.85	0.51	0.93	1.06	0.61	1.30	1.10	0.39	0.73	0.65	0.59	**0.79**
scGPT_A	0.67	0.68	1.27	0.54	0.77	0.40	0.75	0.73	0.70	0.97	0.63	0.84	1.06	0.83	**0.74**
scPRINT_A	0.57	0.93	0.73	0.34	0.39	0.40	0.85	1.34	0.74	0.21	0.45	0.29	0.42	0.35	**0.43**

Table 7: Per-dataset GO AUROC (Jaccard > 0.3).

Representation	Adam	Norm	K562	RPE1	Schm	SARS	MOLM	Endo	Ti-i	Ti-a	Drag	Leng	GWCD	WBrn	Median
GenePT	0.881	0.838	0.875	0.856	0.824	0.911	0.822	0.830	0.899	0.838	0.872	0.870	0.840	0.842	**0.849**
ESM	0.819	0.769	0.749	0.803	0.739	0.835	0.761	0.816	0.788	0.770	0.855	0.796	0.790	0.810	**0.793**
scGPT_S	0.771	0.749	0.775	0.708	0.790	0.807	0.810	0.842	0.650	0.599	0.783	0.740	0.667	0.642	**0.760**
Geneformer_S	0.755	0.677	0.642	0.613	0.792	0.843	0.799	0.676	0.823	0.626	0.713	0.657	0.677	0.678	**0.678**
scGPT_C	0.719	0.709	0.693	0.667	0.722	0.668	0.835	0.768	0.497	0.625	0.756	0.718	0.704	0.657	**0.707**
GeneCompass_S	0.739	0.638	0.647	0.591	0.667	0.737	0.687	0.532	0.750	0.620	0.699	0.642	0.646	0.667	**0.657**
CS-CORE	0.695	0.641	0.675	0.638	0.834	0.683	0.745	0.789	0.521	0.650	0.768	0.561	0.663	0.629	**0.669**
gene2vec	0.711	0.702	0.626	0.682	0.680	0.624	0.678	0.750	0.525	0.632	0.704	0.667	0.674	0.591	**0.676**
RegFormer_S	0.586	0.599	0.725	0.658	0.734	0.712	0.717	0.743	0.626	0.552	0.601	0.595	0.592	0.507	**0.613**
Geneformer_C	0.664	0.651	0.690	0.618	0.682	0.672	0.639	0.604	0.631	0.540	0.614	0.565	0.633	0.620	**0.632**
scPRINT_C	0.682	0.598	0.614	0.660	0.751	0.656	0.675	0.668	0.489	0.570	0.696	0.613	0.635	0.577	**0.646**
GeneCompass_C	0.626	0.495	0.559	0.578	0.664	0.657	0.532	0.585	0.577	0.508	0.626	0.603	0.658	0.595	**0.590**
RegFormer_C	0.679	0.556	0.653	0.569	0.645	0.607	0.708	0.645	0.541	0.500	0.619	0.625	0.542	0.551	**0.613**
scFoundation_C	0.581	0.458	0.613	0.581	0.470	0.699	0.543	0.205	0.649	0.585	0.521	0.545	0.593	0.665	**0.581**
Geneformer_A	0.713	0.664	0.566	0.632	0.451	0.473	0.445	0.582	0.448	0.603	0.703	0.497	0.454	0.494	**0.532**
CellOracle	0.367	0.465	0.416	0.338	0.521	0.703	0.535	–	0.682	0.564	0.452	0.436	0.485	0.433	**0.465**
scFoundation_A	0.530	0.438	0.687	0.371	0.566	0.398	0.447	0.692	0.421	0.420	0.631	0.509	0.525	0.616	**0.517**
GeneCompass_A	0.489	0.598	0.336	0.456	0.446	0.460	0.517	0.517	0.554	0.533	0.325	0.538	0.442	0.556	**0.503**
scFoundation_S	0.572	0.485	0.573	0.491	0.361	0.506	0.553	0.427	0.459	0.533	0.384	0.450	0.479	0.545	**0.488**
scGPT_A	0.507	0.381	0.550	0.657	0.625	0.629	0.430	0.535	0.448	0.452	0.522	0.464	0.517	0.504	**0.512**
scPRINT_A	0.491	0.489	0.654	0.547	0.291	0.239	0.634	0.438	0.451	0.284	0.684	0.452	0.498	0.496	**0.490**

Table 8: Per-dataset GO AUPRC (Jaccard > 0.3).

Representation	Adam	Norm	K562	RPE1	Schm	SARS	MOLM	Endo	Ti-i	Ti-a	Drag	Leng	GWCD	WBrn	Median
GenePT	0.194	0.226	0.399	0.301	0.063	0.361	0.198	0.083	0.542	0.220	0.346	0.376	0.386	0.415	**0.324**
ESM	0.243	0.163	0.153	0.226	0.118	0.399	0.209	0.228	0.326	0.200	0.414	0.369	0.291	0.319	**0.236**
scGPT_S	0.163	0.141	0.104	0.117	0.119	0.151	0.261	0.059	0.019	0.026	0.185	0.102	0.090	0.073	**0.110**
Geneformer_S	0.101	0.060	0.026	0.021	0.105	0.250	0.139	0.023	0.105	0.042	0.082	0.048	0.080	0.104	**0.081**
scGPT_C	0.026	0.053	0.100	0.046	0.028	0.036	0.189	0.031	0.011	0.016	0.066	0.109	0.109	0.059	**0.049**
GeneCompass_S	0.023	0.017	0.024	0.014	0.033	0.027	0.024	0.003	0.047	0.015	0.032	0.020	0.061	0.077	**0.024**
CS-CORE	0.026	0.032	0.099	0.028	0.027	0.021	0.034	0.044	0.005	0.004	0.061	0.011	0.035	0.015	**0.028**
gene2vec	0.023	0.060	0.017	0.030	0.010	0.017	0.034	0.019	0.004	0.020	0.031	0.040	0.032	0.016	**0.022**
RegFormer_S	0.006	0.046	0.118	0.045	0.044	0.016	0.062	0.049	0.057	0.033	0.020	0.017	0.036	0.005	**0.040**
Geneformer_C	0.037	0.018	0.075	0.033	0.010	0.014	0.031	0.005	0.028	0.038	0.011	0.011	0.032	0.016	**0.023**
scPRINT_C	0.012	0.012	0.018	0.020	0.015	0.023	0.026	0.012	0.003	0.006	0.018	0.015	0.016	0.057	**0.016**
GeneCompass_C	0.006	0.003	0.029	0.029	0.006	0.005	0.011	0.005	0.007	0.003	0.011	0.013	0.024	0.039	**0.009**
RegFormer_C	0.022	0.005	0.009	0.004	0.027	0.002	0.045	0.007	0.003	0.002	0.004	0.015	0.007	0.025	**0.007**
scFoundation_C	0.008	0.002	0.004	0.005	0.001	0.013	0.003	0.000	0.024	0.011	0.002	0.005	0.021	0.033	**0.005**
Geneformer_A	0.008	0.017	0.006	0.008	0.001	0.003	0.001	0.004	0.001	0.002	0.017	0.004	0.003	0.002	**0.003**
CellOracle	0.002	0.004	0.003	0.003	0.002	0.003	0.003	–	0.016	0.006	0.002	0.002	0.004	0.006	**0.003**
scFoundation_A	0.003	0.002	0.027	0.001	0.001	0.001	0.005	0.008	0.001	0.001	0.006	0.002	0.003	0.006	**0.002**
GeneCompass_A	0.003	0.005	0.002	0.002	0.001	0.002	0.008	0.001	0.003	0.001	0.001	0.002	0.003	0.003	**0.002**
scFoundation_S	0.002	0.002	0.003	0.003	0.001	0.001	0.004	0.001	0.004	0.004	0.001	0.002	0.003	0.012	**0.002**
scGPT_A	0.002	0.001	0.005	0.004	0.001	0.001	0.001	0.001	0.001	0.002	0.005	0.002	0.005	0.004	**0.002**
scPRINT_A	0.001	0.002	0.003	0.001	0.000	0.000	0.003	0.001	0.001	0.000	0.004	0.002	0.003	0.002	**0.002**

Table 9: Per-dataset perturbational Fisher odds ratio.

Representation	Adam	Norm	K562	RPE1	Schm	SARS	MOLM	Endo	Ti-i	Ti-a	Drag	Leng	GWCD	WBrn	Median
Geneformer_A	3.59	5.19	1.17	1.10	3.33	5.77	2.84	9.12	12.19	4.07	1.96	7.20	2.73	5.29	**3.83**
CS-CORE	3.88	2.49	1.07	2.29	2.21	5.80	2.64	11.95	19.56	4.08	0.95	–	0.61	48.10	**2.64**
RegFormer_S	1.66	2.30	1.34	1.64	2.18	2.39	1.02	2.51	26.03	2.35	2.73	3.44	1.65	1.53	**2.24**
CellOracle	1.50	2.91	1.02	0.94	2.82	7.14	2.41	–	–	2.10	1.22	4.86	1.73	14.79	**2.26**
gene2vec	2.45	1.57	0.97	2.52	2.37	3.39	1.24	2.69	6.94	3.09	0.87	–	1.10	–	**2.41**
scPRINT_A	3.81	3.42	1.41	1.31	1.58	3.37	2.37	7.45	2.07	4.50	2.19	1.68	1.54	–	**2.19**
scGPT_S	2.96	1.05	1.05	2.43	1.38	3.08	–	–	18.15	1.64	0.63	–	0.87	1.86	**1.64**
scFoundation_A	2.03	3.47	3.20	2.54	1.99	2.98	1.15	18.37	9.51	2.63	1.04	1.83	2.02	2.34	**2.44**
Geneformer_C	2.34	2.08	0.99	1.39	1.29	2.52	1.44	1.47	6.51	4.58	0.52	–	1.99	–	**1.73**
scGPT_C	3.90	1.82	1.19	1.97	1.87	2.57	2.00	16.34	3.30	1.61	1.56	3.44	0.97	3.04	**1.99**
scPRINT_C	3.59	1.77	0.81	2.24	2.49	2.06	2.35	–	3.55	1.52	0.93	–	2.04	–	**2.06**
GeneCompass_A	3.95	3.55	0.97	1.12	3.18	2.52	2.08	8.68	8.62	5.38	1.91	1.53	2.34	2.30	**2.43**
GenePT	1.81	1.62	1.01	1.41	1.57	6.30	1.14	–	–	1.58	1.43	3.12	1.24	–	**1.57**
RegFormer_C	3.25	1.65	0.75	1.24	1.75	2.16	2.11	–	3.96	2.63	1.58	1.92	1.51	1.44	**1.75**
Geneformer_S	1.64	1.25	0.94	2.05	1.94	2.95	0.85	–	–	3.91	0.74	–	0.86	–	**1.44**
ESM	1.26	1.16	0.69	1.40	1.48	1.63	1.63	–	5.82	2.05	0.95	1.40	1.28	4.04	**1.40**
scFoundation_S	1.08	1.22	1.10	1.07	1.57	1.33	1.31	–	2.78	1.82	1.12	–	1.04	–	**1.22**
GeneCompass_C	1.72	1.32	1.08	1.41	1.89	0.69	1.14	–	–	0.93	0.58	–	2.41	–	**1.23**
GeneCompass_S	0.94	0.80	0.81	1.76	1.06	2.53	0.84	–	–	2.29	0.60	–	0.59	–	**0.89**
scGPT_A	0.21	0.51	0.99	0.73	0.61	0.60	1.11	–	–	2.02	0.66	–	0.95	–	**0.69**
scFoundation_C	–	0.10	–	0.52	0.13	–	–	–	–	–	–	–	0.06	–	**0.12**

Table 10: Per-dataset perturbational AUROC.

Representation	Adam	Norm	K562	RPE1	Schm	SARS	MOLM	Endo	Ti-i	Ti-a	Drag	Leng	GWCD	WBrn	Median
Geneformer_A	0.487	0.496	0.474	0.495	0.511	0.444	0.529	0.406	0.408	0.513	0.514	0.663	0.518	0.423	**0.496**
CS-CORE	0.514	0.441	0.467	0.500	0.485	0.577	0.459	0.379	0.729	0.484	0.347	–	0.454	0.731	**0.484**
RegFormer_S	0.504	0.526	0.580	0.512	0.539	0.521	0.437	0.654	0.750	0.515	0.521	0.500	0.520	0.640	**0.521**
CellOracle	0.551	0.736	0.523	0.538	0.678	0.767	0.677	–	–	0.705	0.588	0.769	0.645	0.815	**0.678**
gene2vec	0.510	0.520	0.497	0.535	0.522	0.494	0.431	0.398	0.538	0.535	0.439	–	0.464	–	**0.504**
scPRINT_A	0.513	0.501	0.489	0.512	0.438	0.462	0.550	0.256	0.386	0.471	0.487	0.532	0.523	–	**0.489**
scGPT_S	0.600	0.566	0.511	0.543	0.539	0.660	–	–	0.669	0.495	0.351	–	0.447	0.607	**0.543**
scFoundation_A	0.477	0.432	0.605	0.634	0.492	0.498	0.405	0.376	0.324	0.415	0.431	0.240	0.503	0.433	**0.432**
Geneformer_C	0.535	0.491	0.516	0.485	0.464	0.520	0.539	0.325	0.452	0.544	0.443	–	0.492	–	**0.492**
scGPT_C	0.535	0.471	0.564	0.523	0.526	0.485	0.477	0.287	0.416	0.416	0.489	0.530	0.439	0.470	**0.481**
scPRINT_C	0.537	0.503	0.430	0.524	0.506	0.515	0.478	–	0.343	0.402	0.470	–	0.426	–	**0.478**
GeneCompass_A	0.601	0.467	0.365	0.470	0.509	0.597	0.362	0.504	0.354	0.501	0.544	0.393	0.437	0.436	**0.468**
GenePT	0.560	0.502	0.516	0.529	0.525	0.626	0.424	–	–	0.536	0.505	0.596	0.495	–	**0.525**
RegFormer_C	0.541	0.489	0.515	0.512	0.466	0.512	0.451	–	0.389	0.459	0.445	0.310	0.487	0.486	**0.486**
Geneformer_S	0.572	0.523	0.457	0.511	0.530	0.516	0.548	–	–	0.551	0.457	–	0.498	–	**0.519**
ESM	0.513	0.533	0.512	0.540	0.495	0.502	0.461	–	0.577	0.534	0.456	0.672	0.490	0.589	**0.513**
scFoundation_S	0.469	0.520	0.494	0.506	0.492	0.465	0.502	–	0.531	0.484	0.483	–	0.501	–	**0.494**
GeneCompass_C	0.508	0.459	0.544	0.523	0.481	0.468	0.408	–	–	0.417	0.438	–	0.497	–	**0.474**
GeneCompass_S	0.476	0.481	0.447	0.497	0.470	0.609	0.551	–	–	0.447	0.365	–	0.446	–	**0.473**
scGPT_A	0.415	0.449	0.489	0.473	0.433	0.322	0.466	–	–	0.496	0.432	–	0.465	–	**0.457**
scFoundation_C	–	0.379	–	0.420	0.337	–	–	–	–	–	–	–	0.485	–	**0.399**

Table 11: Per-dataset regulator-level enrichment rate.

Representation	Adam	Norm	K562	RPE1	Schm	SARS	MOLM	Endo	Ti-i	Ti-a	Drag	Leng	GWCD	WBrn	Median
Geneformer_A	0.543	0.655	0.016	0.007	0.404	0.037	0.159	0.011	0.008	0.102	0.039	0.073	0.118	0.002	**0.056**
CS-CORE	0.217	0.168	0.095	0.406	0.141	0.069	0.029	0.010	0.009	0.068	0.046	0.000	0.026	0.026	**0.057**
RegFormer_S	0.100	0.091	0.006	0.085	0.180	0.011	0.000	0.010	0.014	0.026	0.009	0.026	0.014	0.000	**0.014**
CellOracle	0.012	0.148	0.000	0.005	0.093	0.030	0.021	0.000	0.000	0.000	0.000	0.027	0.007	0.000	**0.006**
gene2vec	0.170	0.047	0.000	0.217	0.167	0.019	0.000	0.005	0.004	0.025	0.000	0.012	0.018	0.000	**0.015**
scPRINT_A	0.717	0.607	0.003	0.010	0.202	0.044	0.365	0.011	0.000	0.147	0.037	0.012	0.035	0.000	**0.036**
scGPT_S	0.262	0.056	0.064	0.278	0.072	0.009	0.000	0.000	0.004	0.000	0.000	0.019	0.017	0.000	**0.013**
scFoundation_A	0.283	0.393	0.010	0.097	0.299	0.026	0.111	0.011	0.004	0.090	0.019	0.036	0.058	0.002	**0.047**
Geneformer_C	0.105	0.130	0.024	0.109	0.051	0.020	0.000	0.005	0.004	0.065	0.000	0.000	0.084	0.002	**0.022**
scGPT_C	0.630	0.208	0.058	0.235	0.208	0.027	0.161	0.015	0.000	0.023	0.019	0.026	0.044	0.000	**0.036**
scPRINT_C	0.592	0.109	0.012	0.229	0.270	0.006	0.098	0.000	0.004	0.018	0.010	0.000	0.132	0.000	**0.015**
GeneCompass_A	0.370	0.569	0.000	0.000	0.324	0.018	0.286	0.011	0.006	0.111	0.009	0.012	0.117	0.000	**0.015**
GenePT	0.077	0.030	0.023	0.052	0.032	0.040	0.000	0.000	0.000	0.000	0.000	0.000	0.002	0.000	**0.001**
RegFormer_C	0.628	0.235	0.000	0.077	0.140	0.012	0.179	0.000	0.004	0.059	0.010	0.000	0.017	0.000	**0.015**
Geneformer_S	0.057	0.010	0.029	0.162	0.103	0.012	0.000	0.000	0.000	0.007	0.000	0.000	0.004	0.000	**0.006**
ESM	0.011	0.003	0.000	0.030	0.000	0.010	0.042	0.000	0.000	0.004	0.000	0.000	0.001	0.000	**0.000**
scFoundation_S	0.000	0.000	0.000	0.000	0.005	0.000	0.000	0.000	0.000	0.004	0.000	0.000	0.000	0.000	**0.000**
GeneCompass_C	0.104	0.030	0.029	0.079	0.127	0.000	0.014	0.000	0.000	0.000	0.000	0.000	0.114	0.006	**0.010**
GeneCompass_S	0.004	0.007	0.021	0.138	0.029	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.003	0.000	**0.000**
scGPT_A	0.000	0.004	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.011	0.009	0.000	0.000	0.000	**0.000**
scFoundation_C	–	0.000	0.000	0.059	0.000	–	–	–	–	0.000	–	–	0.000	0.000	**0.000**

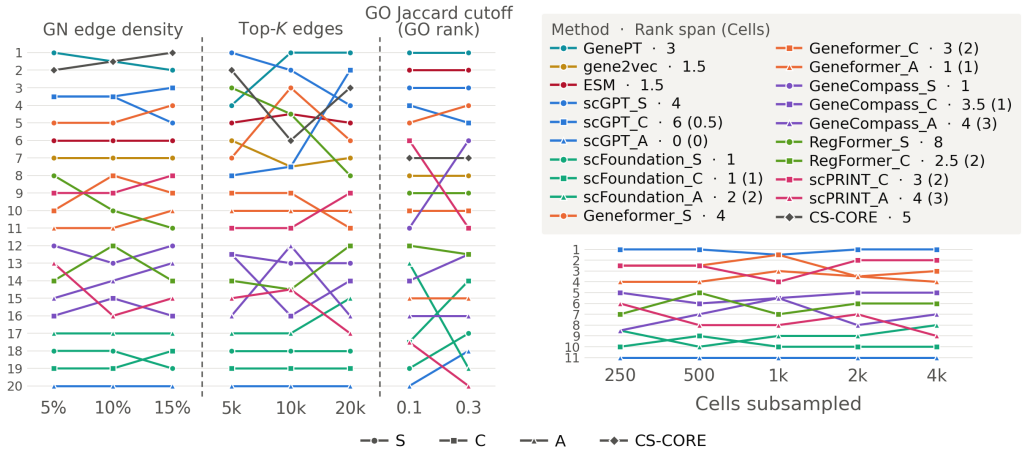


Figure 5: Rank stability across network size, GO Jaccard threshold, and contextual cell-subsample size. Lower ranks indicate better performance; the GO-threshold panel uses GO-specific ranks, and the cell-subsampling analysis includes only contextual and attention-based representations.

Network size. Rankings were stable across networks with 5,000, 10,000, and 20,000 edges (mean pairwise Spearman $\rho = 0.927$). The median rank span was 2.2 positions, although individual representations moved by up to six positions, suggesting that broad trends are stable while small rank differences should not be over-interpreted.

GO similarity threshold. Using GO Jaccard thresholds of 0.1 and 0.3 produced similar GO rankings ($\rho = 0.912$), despite the stricter reference at 0.3. Both thresholds are therefore reported as a robustness check.

Cell-subsample size. Contextual and attention representations were recomputed from nested subsamples of 250, 500, 1,000, 2,000, and 4,000 cells. Rankings remained stable across this range (mean pairwise $\rho = 0.947$), with no representation moving by more than three positions, indicating that the 1,000-cell default is sufficient for stable relative comparisons.

Similarity definition. For vector representations, we compared plain cosine similarity with cosine after removing the leading principal component and with an additional rank-inverse-normal transformation. Rankings were nearly unchanged (mean pairwise $\rho = 0.996$), so the main benchmark reports the unmodified cosine similarity.

Fixed edge count versus fixed density. We additionally repeated the benchmark at fixed network densities $d \in \{0.05, 0.10, 0.15\}$. Rank-based metrics were largely stable across the two conventions, whereas count-based perturbational metrics were more sensitive. We therefore report fixed edge count as the primary analysis and fixed density as a robustness check.

Overall, the broad representation ordering and group-level conclusions are stable across these design choices, whereas differences of only a few rank positions should not be interpreted as robust differences between individual representations.

J.1 Sensitivity to Dataset Characteristics

We next tested whether broad biological characteristics of the Perturb-seq datasets systematically influenced benchmark scores. For each dataset and metric, scores were first averaged across the three network sizes (5k, 10k, and 20k edges) within each representation and then across the 20 representations, yielding one dataset-level value per metric. Datasets were stratified by tissue (brain, $n = 5$; blood, $n = 6$), cell system (cell line, $n = 6$; primary/iPSC, $n = 8$), and perturbation type (CRISPRi, $n = 9$; CRISPRa, $n = 3$). Group differences were assessed using two-sided Mann–Whitney U tests, with Cliff's δ as the effect size.

None of the comparisons remained significant after Benjamini–Hochberg correction (Table 13). The regulator-level perturbation enrichment metric showed nominal differences between brain and blood datasets ($\delta = -0.93$, $p = 0.009$) and between cell-line and primary/iPSC datasets ($\delta = 0.67$, $p = 0.043$), but neither survived multiple-testing correction. Overall, we found no statistically significant evidence that the benchmark metrics were systematically driven by the tested tissue, cell-system, or perturbation-type differences.

Table 13: Association between dataset-level biological characteristics and benchmark metrics. Effect sizes are reported as Cliff's $\delta = P(X > Y) - P(X < Y)$, where X and Y denote the first and second groups listed, respectively. Bold values indicate nominal $p < 0.05$; no comparison remained significant after Benjamini–Hochberg correction.

Metric	**Tissue**		**Cell system**		**Perturbation**	
	brain ($n = 5$) vs. blood ($n = 6$)		cell line ($n = 6$) vs. primary/iPSC ($n = 8$)		CRISPRi ($n = 9$) vs. CRISPRa ($n = 3$)	
	δ	p	δ	p	δ	p
GO, Jaccard > 0.3						
Fisher OR	-0.07	0.931	-0.12	0.755	$+0.56$	0.209
AUROC	-0.53	0.177	$+0.54$	0.108	$+0.48$	0.282
AUPRC	$+0.20$	0.662	$+0.21$	0.573	$+0.78$	0.064
Perturb-seq						
DE Fisher OR	$+0.67$	0.082	-0.38	0.282	$+0.04$	1.000
DE AUROC	$+0.07$	0.931	$+0.42$	0.228	$+0.33$	0.482
Regulator-level enrichment	-0.93	**0.009**	$+0.67$	**0.043**	-0.63	0.145

K Computational cost

All runtimes are wall-clock measurements from the benchmark cluster. Contextual representation extraction is the only GPU stage and uses one NVIDIA A30 (24 GB) with 8 CPU cores and 64 GB host memory. Differential-expression testing, CS-CORE, network construction, and evaluation are CPU-only and use 8 cores with up to 128 GB memory. No stage requires multiple GPUs or distributed execution.

Static representations are read directly from released embedding files or checkpoint token tables and therefore incur negligible inference cost. For contextual representations, we measured runtime across six foundation models using 250–4,000 subsampled cells and approximately 1,000 genes.

Table 14: Runtime of contextual representation extraction on one NVIDIA A30.

	Checkpoint	**Load**	**Seconds per dataset**					**Slope**
Model	(MB)	(s)	**250**	**500**	**1,000**	**2,000**	**4,000**	
scPRINT	211	30	3.3	3.7	2.7	6.6	10.8	0.43
GeneCompass	1099	< 5	4.5	4.8	8.8	17.0	33.2	0.76
scGPT	196	127	11.7	14.4	14.4	33.9	59.8	0.59
Geneformer	398	< 5	12.9	14.6	20.0	26.9	42.9	0.43
RegFormer	189	6	7.0	12.1	24.1	47.3	93.5	0.94
scFoundation	1366	< 5	43.6	85.9	170.0	342.2	680.0	0.99
All six, 14 datasets			**24 min**	**38 min**	**68 min**	**2.2 h**	**4.3 h**	

Runtime increases with cell count and varies substantially across models (Fig. 6). At 4,000 cells, extraction ranges from 10.8 s for scPRINT to 680 s for scFoundation, while all models complete within one hour per dataset. At the default 1,000-cell setting, running all six contextual models over the 14 processed datasets requires approximately 68 minutes of A30 time; the full cell-subsampling sweep requires about 8.6 GPU-hours.

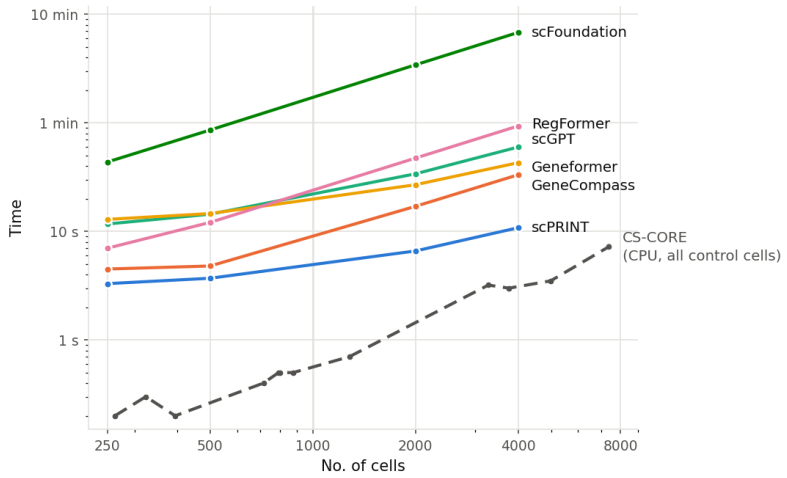


Figure 6: Runtime scaling of contextual foundation models and CS-CORE. Solid lines show mean runtime for six contextual models on one NVIDIA A30; the dashed line shows CS-CORE using eight CPU cores and all available control cells. Both axes are logarithmic.

The SCEPTRE differential-expression reference requires approximately 17 CPU-hours in total but is computed once per dataset and reused by all methods. CS-CORE is substantially cheaper, requiring 0.2–234 s per dataset on eight CPU cores. Network construction and evaluation are also CPU-only, requiring approximately 0.6 minutes per representation per dataset across the three network sizes in the reported cosine arm.

Overall, the main benchmark requires approximately 1.1 GPU-hours for contextual representation extraction, while static representations require no GPU inference. The remaining computational cost is dominated by CPU-based differential-expression testing and network evaluation.