GRNFormer: Accurate Gene Regulatory Network Inference Using Graph Transformer

Akshata Hegde[1,2], Jianlin Cheng[1,2*]

[1] Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO, USA 65201

[2]Roy Blunt Nextgen Precision Health, University of Missouri, Columbia, MO, USA 65201

*Corresponding author

Abstract

We introduce GRNFormer, a generalizable graph transformer framework for accurate gene regulatory network (GRN) inference from transcriptomics data. Designed to work across species, cell types, and platforms without requiring cell-type annotations or prior regulatory information, GRNFormer integrates a transformer-based Gene Transcoder with a variational graph autoencoder GraViTAE combined with pairwise attention to learn the representations of GRNs. Leveraging TF-Walker, a transcription factor-anchored subgraph sampling strategy, it effectively captures gene regulatory interactions from single-cell or bulk RNA-seq data. Evaluated on standard benchmark datasets, GRNFormer outperforms existing traditional and deep learning state-of-the-art methods in blind evaluations, achieving 0.90-0.98 average Area Under the Receiver Operating characteristic Curve (AUROC) and Area Under the Precision-Recall Curve (AUPRC) as well as 0.87-0.98 average F1 score. It robustly recovers both known and novel regulatory networks, including pluripotency circuits in human embryonic stem cells (hESCs) and immune cell modules in Peripheral Blood Mononuclear Cells (PBMCs). Its architecture enables scalable, biologically interpretable GRN inference across various datasets, cell types, and species, establishing GRNFormer as a robust and transferable tool for network biology.

Main

Gene expression data provides critical insight into transcriptional activity across cell types, tissues, and conditions, reflecting the genome's dynamic regulation in response to intrinsic and extrinsic cues [1]. RNA sequencing, particularly at single-cell resolution, has transformed transcriptomics by enabling comprehensive, high-throughput profiling of entire transcriptomes [2,3]. Gene expression is regulated by a complex network of molecular mechanisms that control the process of gene transcription. Transcription factors (TFs), which bind to specific DNA motifs located near or within gene promoters, as well as at distal enhancer regions, play a significant role in regulating transcription by activating or repressing gene expression. These layered control systems maintain cellular identity, support developmental transitions, and enable responses to environmental stimuli [4].

Precise gene regulation is essential for biological processes. Dysregulation can result in pathological conditions including cancer, neurodegeneration, and developmental disorders. Understanding gene regulation at the network level—through gene regulatory networks (GRNs)—can illuminate the interactions among genes and proteins that orchestrate cellular function. GRN

inference enables the identification of key regulatory drivers and pathways, offering a foundation for mechanistic insights and therapeutic targeting [5].

However, inferring gene regulatory networks (GRNs) from gene expression data, particularly from high-throughput platforms like RNA-seq data, remains a major challenge. Traditional approaches, such as correlation-based or simple statistical models, often fall short in capturing the non-linear, context-specific nature of gene regulation, including feedback loops and combinatorial control. More advanced methods—such as information-theoretic frameworks like ARACNE and CLR [6,7], or Bayesian networks [8,9]—can model complex dependencies, but typically require large sample sizes and face scalability limitations. Dynamic models, including Boolean networks and differential equation-based approaches [10,11], offer temporal insights but depend on time-resolved data, which is often lacking. Compounding these challenges, RNA-seq data are high-dimensional, noisy, and sample-limited, making many inference methods prone to overfitting [12]. The BEELINE benchmarking study [13] underscores these issues, showing that conventional statistical and shallow learning approaches struggle to generalize across diverse regulatory contexts.

To address these challenges, newer approaches, particularly those based on deep learning, have been developed. These methods are capable of learning complex, non-linear patterns from high-dimensional data, making them better suited for inferring gene regulatory networks from RNA-Seq data [14 -17]. Latest deep learning models, such as transformers, can integrate diverse types of biological information, to provide more accurate and robust predictions of gene regulatory interactions [18][19]. Transformers, the backbone of high-end large language models like ChatGPT, have revolutionized the perspectives of analyzing the input information through the "attention" mechanism [20]. The attention mechanism captures both global and local information while emphasizing the significant features for learning. There have been a few transformer-based GRN inference methods developed, such as STGRNs [21] and scGREAT [22].

Graph Neural Networks (GNNs) also have emerged as a potential tool for inferring GRNs, complementing transformer-based methodologies. GNNs are designed to operate on graph-structured data where nodes represent entities and edges symbolize their relationships [23]. GNNs excel in modeling the complex configurations of gene interactions, characterized by regulatory connections between transcription factors and target genes. By propagating information through these graphs, GNNs can identify the topological features of GRNs, revealing key regulators and elucidating complex hierarchical regulatory interactions vital for gene expression. Furthermore, GNNs effectively capture both local interactions and global structures, establishing them as robust instruments for clarifying biological pathways and functional modules within GRNs [24].

However, most GRN inference methods are tailored to specific datasets, limiting their generalizability across cell types or conditions. Hence, there is a critical need for robust, transferable models that perform reliably beyond narrow contexts.

In this study, we introduce GRNFormer, a graph transformer framework designed for accurate and generalizable inference of GRNs from single-cell RNA-seq data. GRNFormer integrates the strengths of graph neural networks (GNNs) and transformer architectures to model both the

dynamic expression profiles of genes and the underlying topological structure of regulatory interactions. By combining local structural modeling from GNNs with the distillation and integration of relevant information through attention enabled by transformers, GRNFormer substantially improves inference performance over existing approaches. Although primarily developed for single-cell transcriptomics, we further demonstrate its capacity to generalize effectively to bulk RNA-seq data, highlighting the framework's robustness and scalability across data modalities and species.

Results

GRNFormer Overview

GRNFormer is a graph deep learning framework developed to infer GRNs from either single-cell or bulk transcriptomic data with high accuracy and generalizability (Figure 1). Designed to operate across diverse cell types, species, and regulatory contexts, GRNFormer addresses key challenges in GRN inference such as context-specificity, data sparsity, and model transferability—through a modular architecture comprising three main components.

First, TFWalker introduces a transcription factor-centered subgraph sampling approach that constructs localized gene co-expression subgraphs from a full gene co-expression network (GCEN) input, capturing the neighborhood context around each TF. This biologically informed strategy enhances the model's ability to learn meaningful regulatory patterns by focusing on TF-driven structures in the gene expression space.

Second, we implement end-to-end representation learning through two key modules: the Gene Transcoder, a transformer-based embedding module that captures context-aware gene representations across diverse datasets; and GraViTAE (Graph Variational Transformer Autoencoder), which jointly encodes node and edge features to reconstruct gene regulatory feature representations.

Third, a dedicated GRN inference module integrates node- and edge-level information to predict TF-target interactions. This inference strategy allows GRNFormer to generalize effectively across input sizes, data modalities, and regulatory frameworks, including both ChIP-seq-based and STRING[13,96]-derived networks.

By combining transformer-based contextual learning with graph-aware architecture, GRNFormer enables scalable, transferable, and interpretable inference of GRNs from either single-cell or bulk RNA-seq data. Its outstanding performance across several benchmark datasets highlights its utility for uncovering regulatory relationships in both well-studied and novel biological systems.

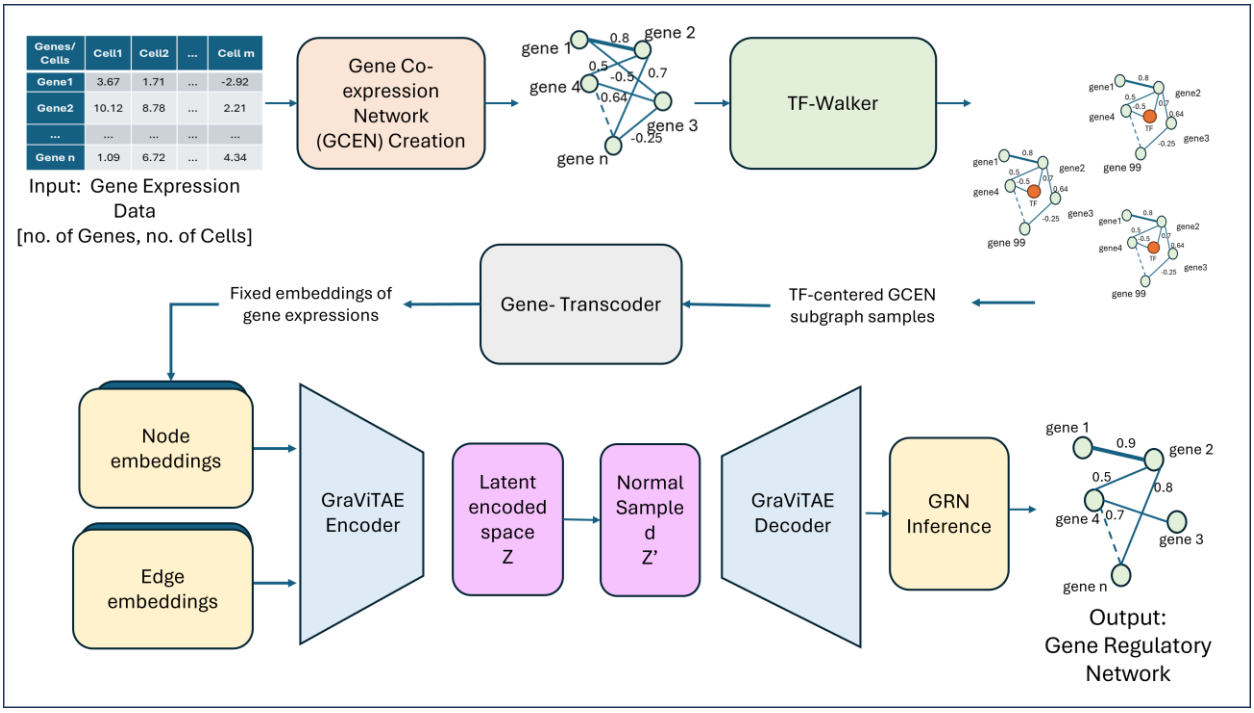

Figure 1: Overview of GRNFormer pipeline.

High-accuracy Inference of GRNs Across Regulatory Contexts

To evaluate the generalizability and contextual sensitivity of GRNFormer, we conducted a comprehensive benchmarking study using the BEELINE suite of single-cell gene expression datasets. These datasets span two species (*Homo sapiens* and *Mus musculus*), seven distinct cell types: human embryonic stem cells (hESC), human hepatocytes (hHep), mouse dendritic cells (mDC), and three mouse hematopoietic stem cell subtypes (mHSC-GM, mHSC-E, and mHSC-L) , and mouse embryonic stem cells (mESC). It contains three types of ground truth GRNs: non-cell-type-specific ChIP-seq networks, cell-type-specific ChIP-seq networks, and STRING-derived interaction networks. Our goal was to assess whether GRNFormer could robustly infer GRNs across species boundaries, cell identities, and diverse regulatory contexts.

GRNFormer was trained to infer GRNs from any set of gene and transcription factor (TF) expression profiles. Training was performed on randomly shuffled subgraphs drawn from large-scale expression matrices, paired with general regulatory interaction sets as supervision. No information about gene identity, cell type, species, or regulatory context was provided during training, encouraging inductive generalization across diverse biological settings. During evaluation, it reconstructs global GRNs by aggregating regulatory interactions inferred from locally sampled, TF-centered co-expression subgraphs. For aggregated GRNs, we perform 100 rounds of bootstrap negative sampling, each with a 1:1 ratio of positive to negative edges. Metrics are averaged across replicates to ensure statistical robustness and preserve evaluation sensitivity to class imbalance.

GRNFormer was first trained and internally evaluated on the data of five cell types (hESC, hHep, mDC, mHSC-E, mHSC-GM). Although individual genes may appear in both training and internal

evaluation phases, we performed localized z-score normalization across genes for each cell in every subgraph, ensuring that the model treats each instance independently and preventing memorization. To rigorously assess generalization and guard against information leakage, we blindly evaluated the trained GRNFormer on two held-out cell types - mESC and mHSC-L - not seen during training.

Table 1 summarizes GRNFormer's average performance on the two withheld cell types (mESC and mHSC-L) unseen in the training across regulatory settings, reporting the area under the receiver operating characteristic curve (AUROC), the area under the precision-recall curve (AUPRC), F1 score, precision, recall, and accuracy for each species-cell type-regulatory context combination. The results of the internal evaluation on the other five cell types are presented in Supplementary Table S1.

During the blind evaluation, we evaluated GRNFormer's performance on subsets of the dataset of each cell type (mESC and mHSC-L) containing the top 500 and 1,000 highly variable genes, both with and without the most significantly varying TFs. The test subsets were prepared according to BEELINE's standard data preprocessing and evaluation protocols.

Table 1: Summary of evaluation metrics of GRNFormer across BEELINE dataset

Network Type	Cell Type	Gene sets	#Genes	#True Positives	AUROC	AUPRC	Precision	Recall	F1Score	Accuracy
Non-specific ChIP-Seq	mESC	1000	1000	852	0.965	0.938	0.925	0.998	0.96	0.959
		500	500	264	0.97	0.953	0.92	0.999	0.958	0.956
		TF+1000	1000	853	0.965	0.939	0.926	0.998	0.961	0.959
		TF+500	500	264	0.965	0.943	0.918	0.999	0.957	0.955
	mHSC-L	1000	1000	589	0.984	0.974	0.956	0.999	0.978	0.977
		500	500	109	0.981	0.966	0.956	1	0.977	0.977
		TF+1000	1124	771	0.979	0.964	0.945	0.999	0.972	0.971
		TF+500	624	467	0.942	0.903	0.886	0.859	0.872	0.874
Cell Type Specific ChIP-Seq	mESC	1000	1000	5499	0.967	0.942	0.93	0.998	0.963	0.962
		500	500	1913	0.968	0.944	0.923	0.999	0.959	0.958
		TF+1000	1000	5500	0.968	0.943	0.93	0.999	0.963	0.962
		TF+500	500	1913	0.967	0.943	0.923	0.999	0.96	0.958
	mHSC-L	1000	1000	1325	0.979	0.959	0.956	0.999	0.978	0.977
		500	500	1295	0.979	0.954	0.96	0.999	0.98	0.979
		TF+1000	1124	7746	0.976	0.954	0.949	1	0.974	0.973
		TF+500	624	5182	0.943	0.898	0.89	1	0.942	0.938
STRING	mESC	500	500	264	0.957	0.928	0.916	0.996	0.954	0.952
		TF+1000	1000	661	0.964	0.937	0.925	0.991	0.957	0.955
		TF+500	500	263	0.959	0.931	0.917	0.992	0.953	0.951
	mHSC-L	500	500	14	0.982	0.973	0.957	1	0.977	0.976
		TF+1000	1124	114	0.981	0.972	0.941	1	0.969	0.968
		TF+500	624	67	0.944	0.913	0.89	0.855	0.872	0.874

As shown in Table 1, GRNFormer achieves consistent accurate performance across all the subsets spanning different network types (non-specific-ChIP-Seq network, cell-type-specific ChIP-Seq

network, and STRING-based network), cell types, and gene sets, with high AUROC and AUPRC scores between ~0.90 and 0.98 as well as high F1 scores between 0.87 and 0.98, indicating robust, accurate, generalized GRN inference across board.

GRNFormer outperforms state-of-the-art GRN inference methods across benchmarks under blind evaluation

To systematically evaluate the performance of GRNFormer, we conducted a comprehensive benchmarking analysis against nine widely used gene regulatory network (GRN) inference methods: with five deep-learning methods, CNNC[25], GNE[26], GNNLink[27], STGRNS, and scGREAT and four traditional methods, LEAP[28], PIDC[29], PPCOR[30], SINCERITIES[31]. Table S2 provides the brief summary of each method used for benchmarking. The benchmarking was performed across three independent gold-standard reference GRNs: (i) non-cell-type specific ChIP-seq, (ii) cell-type-specific ChIP-seq, and (iii) functional interaction networks from STRING. For each method and evaluation setting, we computed AUROC and AUPRC, with top varying 500 and 1000 gene sets with and without significant TFs. We considered cell types mESC and mHSC-L to ensure a blind, stringent assessment of the generalizability of the GRNFormer, as these two cell types are not used in training GRNFormer.

GRNFormer is trained once on a general-purpose dataset spanning five cell types (hESC, hHep, mDC, mHSC-E, mHSC-GM) and then applied to the two new cell types (mESC and mHSC-L) without task-specific fine-tuning, simulating a truly blind generalization scenario. In contrast, the baseline deep learning methods were trained and evaluated on the training and test data split from the same dataset of each of the two cell types (mESC and MHSC-L) respectively. This less stringent evaluation strategy was applied to the nine baseline methods, because, as described in their original publications, most of these models were optimized for within-dataset performance and did not perform cross-cell-type generalization. For consistency and fairness, we reproduced each method on our test datasets following the evaluation strategy reported in its original work, using the same training-validation splits for model development and the same held-out test split for final evaluation across all the baseline methods to reflect their best-case results.

Figure 2 shows the heatmap of AUROC and AUPRC scores on all the common test datasets across all the methods. For all the gene selection strategies, we followed the evaluation protocol defined in BEELINE, using four standard configurations: (1) the 500 most highly variable genes (500), (2) the 1,000 most highly variable genes(1000), (3) all significantly varying transcription factors along with the top 500 most variable genes (TF500), and (4) all significantly varying transcription factors along with the top 1,000 most variable genes (TF1000).

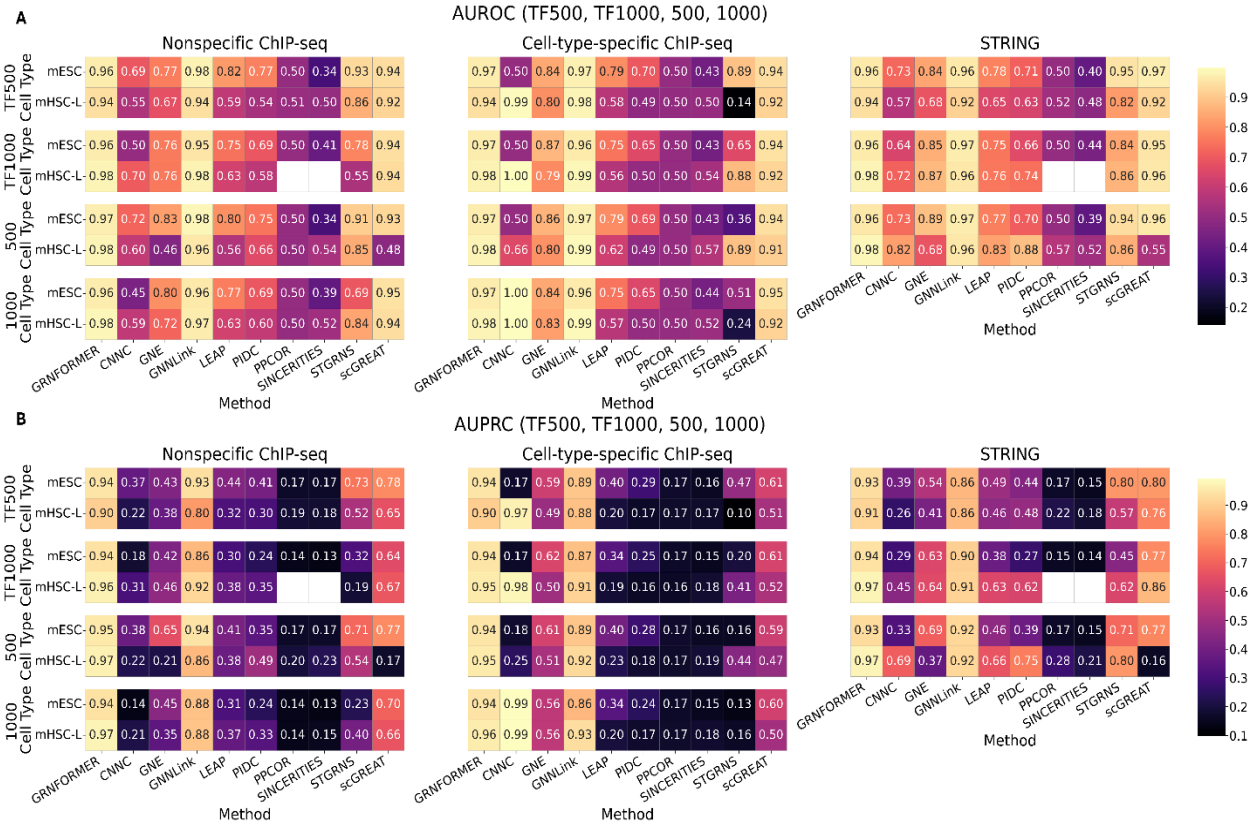

Figure 2. Comparison of GRNFormer and nine baseline methods in various BEELINE test datasets. (A) AUROC scores for 10 GRN inference methods evaluated on the BEELINE test set using three gold-standard benchmark GRNs: nonspecific ChIP-seq, cell-type-specific ChIP-seq, and STRING protein-protein interaction network. GRNFormer consistently ranks among the top performers across all evaluation scenarios. (B) AUPRC scores for the same evaluation settings. GRNFormer showed superior precision-recall trade-offs, maintaining robust performance across datasets where other methods experienced significant degradation. It particularly excels in STRING and cell-type-specific ChIP-seq, where the other methods exhibit significant performance drops.

Although GRNFormer was tested on the new cell types that it had never seen before while the other methods were trained and tested on the data split of the same cell types, GRNFormer consistently outperformed all other methods in almost all evaluation situations. In terms of AUROC (Fig. 2A), GRNFormer achieved the highest or near-highest scores in nearly all contexts, with particularly robust performance in the STRING benchmark (AUROC 0.95-0.98) and cell-type-specific ChIP-seq (AUROC >0.9 in most TF set sizes). This indicates that GRNFormer effectively recovers both direct TF-gene interactions and broader functional relationships.

Similarly, in AUPRC evaluations (Fig. 2B), GRNFormer showed superior precision-recall trade-offs, maintaining robust performance in all settings where other methods experienced significant degradation in all or some settings - especially in the cell-type-specific ChIP-seq and STRING settings. Notably, methods like PIDC, PPCOR, and SINCERITIES performed poorly on these benchmarks, highlighting their limited generalization.

Among the competing methods, GNNLink stood out as the second-best performer overall, with AUROC values closely trailing those of GRNFormer across multiple benchmarks even with AUROC surpassing GRNFormer in some datasets. Its performance in both nonspecific and cell-type-specific ChIP-seq evaluations was strong, suggesting it captures regulatory patterns with competitive sensitivity. However, in the AUPRC evaluation, GNNLink was consistently outperformed by GRNFormer in all the contexts, particularly in more complex benchmarks like STRING, where the distinction between true and false positives is most critical. This indicates that GRNFormer is much more effective in predicting true positives - true TF-gene regulatory relationships than GNNLink and other methods.

These results demonstrate that GRNFormer is capable of inferring high-confidence GRNs in a data-agnostic fashion, achieving state-of-the-art accuracy without requiring retraining or task-specific tuning. This robust generalization across cell types and evaluation criteria underscores the method's practical utility for single-cell regulatory genomics.

Ablation Study: Dissecting GRNFormer's Architecture via Blind Cross-Lineage Evaluation

Table 2: Ablation study of GRNFormer components in blind cross-lineage evaluation. Each model was trained on five cell types (hESC, hHep, mDC, mHSC-E, and mHSC-GM), and blindly assessed on two new cell types (mESC and mHSC-L). AUROC and AUPRC averaged across the test sets of the two cell types are reported.

Ablation Variant	TransConv Block	Subgraph Sampling	Embedding Type	Decoder	AUROC	AUPRC
Full GRNFormer (average)	**✓**	**TF-Walker**	**Gene Transcoder**	**✓**	**0.970**	**0.950**
No TF-Walker	✓	Random Sampling	Gene Transcoder	✓	0.562	0.562
GCN (Full GCN-based variant)	✗ (GCN)	TF-Walker	Gene Transcoder	✗	0.928	0.904
No Decoder (Edge Directly Predicted)	✓	TF-Walker	Gene Transcoder	✗	0.931	0.907
Simple Embedding Only	✓	TF-Walker	Linear Embedding	✓	0.931	0.906
No Gene Transcoder + No Decoder	✓	TF-Walker	Linear Embedding	✗	0.930	0.906
Plain Baseline (GCN only)	✗ (GCN)	Random Sampling	Linear Embedding	✗	0.562	0.562

To assess the contribution of each architectural component in GRNFormer, we performed a systematic ablation study using a cross-lineage blind evaluation. The model was trained on five diverse expression datasets (hESC, hHep, mDC, mHSC-E, mHSC-GM) and blindly evaluated on

two unseen cell types (mESC and mHSC-L) to rigorously test generalization across species and lineages. Ablation conditions and their results are shown in Table 2.

The full GRNFormer model - comprising the TF-Walker subgraph sampler, Gene Transcoder, Transformer Convolution (TransConv) block, and decoder - achieved the highest performance (AUROC = 0.97, AUPRC = 0.95), confirming its robustness across biological contexts.

Removing the TF-Walker, which samples TF-anchored co-expression neighborhoods, and replacing with random sampling caused a substantial performance drop (AUROC/AUPRC = 0.562/0.562), highlighting its critical role in focusing the model on biologically meaningful subgraphs. Replacing the Gene Transcoder with a linear embedding reduced contextual learning and led to a modest decline (AUROC = 0.931, AUPRC = 0.906). Excluding the decoder, which contributes edge-level specificity, produced similar effects (AUROC = 0.931, AUPRC = 0.907), with further reduction when both components were removed.

Replacing the TransConv block with a Graph Convolutional Network (GCN) [32] retained moderate performance (AUROC = 0.928, AUPRC = 0.904), but underperformed relative to the transformer-based architecture, indicating the value of pairwise attention for regulatory inference.

A minimal GCN-only baseline lacking all GRNFormer components yielded near-random results (AUROC = 0.562, AUPRC = 0.562), reinforcing that deep contextual modeling is essential for extracting regulatory signals from expression data.

Together, these results demonstrate that each GRNFormer module contributes distinct value: TF-Walker localizes transcriptional context, the Gene Transcoder and TransConv Block enrich representation learning, and the decoder sharpens edge inference. Their synergy enables accurate and generalizable GRN reconstruction across varied transcriptomic landscapes.

Cross-Species Generalization on DREAM5 Bulk RNA-Seq Datasets

To evaluate GRNFormer's ability to generalize beyond mammalian single-cell data, we conducted a blind assessment using bulk RNA-seq datasets from Escherichia coli and Saccharomyces cerevisiae curated by the DREAM5 challenge [33]. These benchmarks offer distinct regulatory architectures-prokaryotic and unicellular eukaryotic-and include experimentally validated GRNs derived from high-confidence ChIP datasets. Crucially, GRNFormer pretrained on the human and mouse single-cell RNA-seq datasets were blindly tested on the Escherichia coli and Saccharomyces cerevisiae datasets, without using any species-specific annotations or prior regulatory information.

This evaluation addressed two key questions: (1) Can GRNFormer generalize across distant species with divergent transcriptional logic? (2) Does it maintain accuracy when transitioning from single-cell data to bulk RNA-seq data without retraining or fine-tuning?

Applied directly to the DREAM5 gene expression matrices, GRNFormer achieved AUROC scores of 0.979 for E. coli and 0.977 for yeast, with AUPRCs of 0.955 and 0.957. Figure 3 A and 3B present the precision-recall (PR) and ROC curves for *E. coli*, and Figure 3C and 3D show the corresponding curves for *S. cerevisiae*. These results match or exceed the best-performing methods

from the original DREAM5 challenge, despite GRNFormer being trained exclusively on unrelated human single-cell datasets. To ensure robustness in performance estimates, we performed 100 rounds of bootstrap negative sampling for each aggregated GRN, maintaining a 1:1 ratio of positive to negative edges. Evaluation metrics were averaged across the samples to minimize sampling variance and provide statistically reliable comparisons.

These findings confirm GRNFormer's broad utility: it captures regulatory principles conserved across species, performs robustly across transcriptomic modalities, and scales up to genome-wide data without retraining. Its robust performance on DREAM5 underscores its potential as a general-purpose GRN inference engine, particularly in settings where ground-truth networks are limited or unavailable.

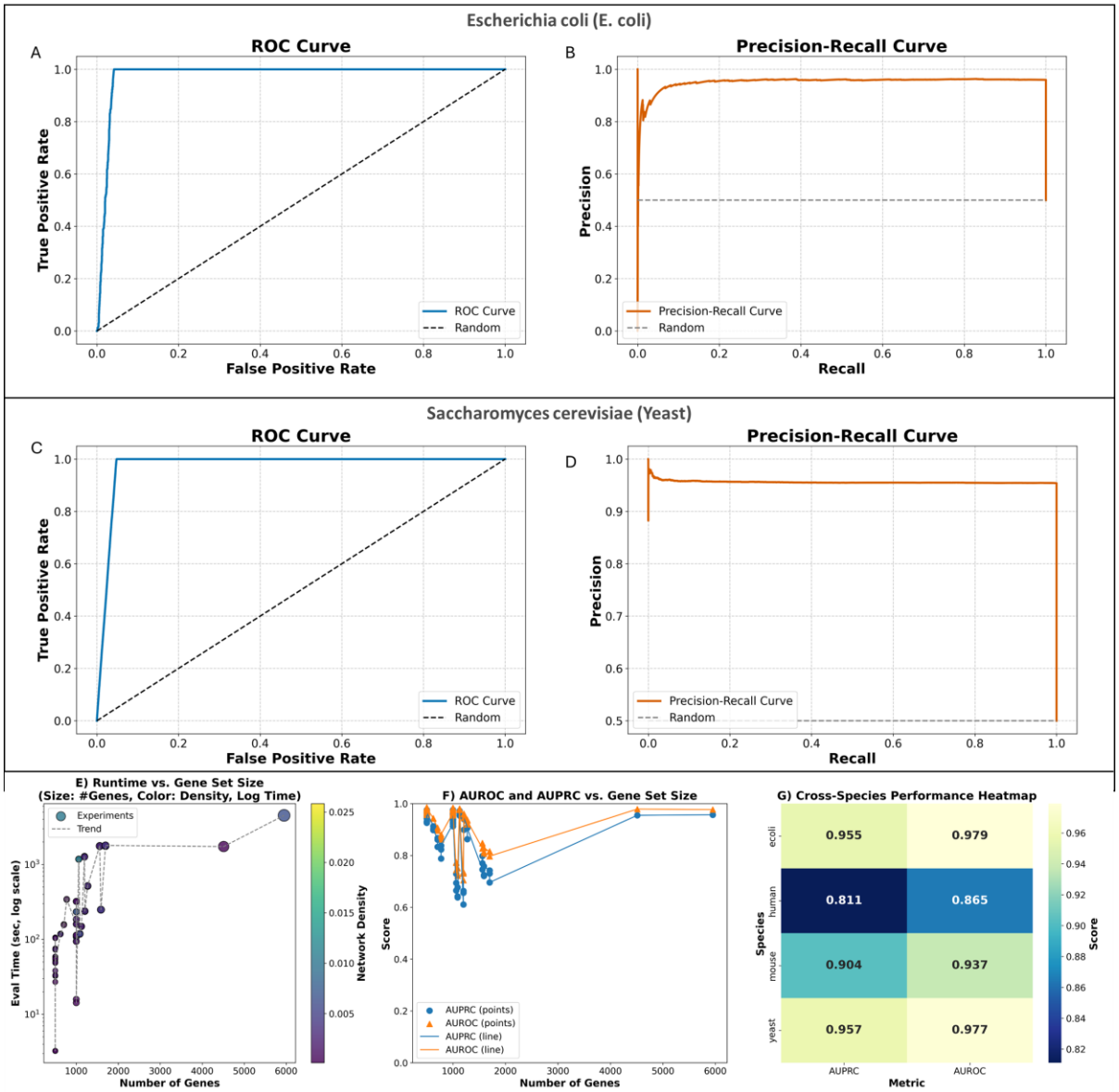

Figure 3: Blind evaluation of GRNFormer on DREAM5 bulk RNA-seq datasets for E. coli and S. cerevisiae and its scalability and cross-species generalization. (A) and (B). Precision-recall (PR)

curve and receiver operating characteristic (ROC) curve, respectively, for the E. coli gene regulatory network (GRN). (C) and (D). PR and ROC curves for the S. cerevisiae GRN prediction. (E). Runtime vs. number of genes. (F). AUROC or AUPRC vs. number of genes. (G). The average performance on the test datasets of four different species.

Scalability of GRNFormer across Gene Set Size, Dataset Complexity, and Species

We evaluated the computational and biological scalability of GRNFormer using a diverse set of expression datasets varying in gene number, cell type, species, and regulatory context. We use all the single-cell RNA-seq datasets from BEELINE suite and bulk RNA-seq datasets (E. Coli and yeast) from DREAM5 challenge for this analysis. As shown in Figure 3E, evaluation time scaled predictably with the number of input genes, ranging from ~500 to ~5,900 genes. Despite the high dimensionality, GRNFormer remained computationally efficient, with inference runtime increasing smoothly on a log scale. Notably, this trend held across a wide range of network densities and cell counts, reflecting the efficiency of the TF-Walker sampling strategy and the fixed-size subgraph processing pipeline.

The prediction performance remained consistently high across increasing gene set sizes. Figure 3F shows AUROC and AUPRC scores as a function of the number of genes per dataset. GRNFormer maintained robust performance—even on large-scale networks—with both AUROC and AUPRC exceeding 0.9 for most settings beyond 2,000 genes. These results demonstrate that the model does not overfit to low-dimensional settings and effectively captures regulatory signals in complex transcriptomic profiles.

To further assess generalizability across phylogenetic and regulatory variation, we benchmarked GRNFormer on gene expression datasets from *Homo sapiens*, *Mus musculus*, *Saccharomyces cerevisiae*, and *Escherichia coli*. As shown in Figure 3G, GRNFormer achieved high AUROC and AUPRC scores across all four species. This cross-species robustness highlights GRNFormer's capacity to learn transferable regulatory representations.

Together, these results affirm GRNFormer's suitability for scalable high-accuracy GRN inference—capable of generalizing across increasing gene dimensions, diverse species, and varied biological contexts without retraining or fine-tuning.

GRNFormer Recovers Core hESC Pluripotency and Predicts Novel Regulatory Networks

Human embryonic stem cells (hESCs) are pluripotent cells derived from early-stage embryos, characterized by their capacity for indefinite self-renewal and their ability to differentiate into all somatic lineages [34,35]. These unique properties make hESCs an ideal system for investigating the transcriptional regulation of cell fate, and for understanding how pluripotency is maintained or disrupted in development and disease. Given the overlap between regulatory programs in hESCs and those aberrantly activated in cancer-such as cell cycle control, apoptosis, and epigenetic modulation-hESCs also offer a powerful platform for identifying candidate regulators of oncogenesis [36].

To evaluate GRNFormer's ability to recover biologically meaningful transcriptional circuits, we applied it to an hESC dataset using all annotated transcription factors and the top 500 most variable genes. The predicted GRN was benchmarked against cell-type-specific ChIP-seq ground truth, and all interactions shown in Figure 4A correspond to predictions supported by this experimental ground truth reference. GRNFormer successfully reconstructed the core transcriptional architecture of hESCs, identifying master regulators of pluripotency, including POU5F1 (OCT4), SOX2, MYC, and NANOG, as central nodes in the network. These transcription factors are well established to operate in tightly interconnected autoregulatory circuits that coordinate gene expression programs essential for sustaining pluripotency and suppressing differentiation [37].

The inferred network included numerous interactions that align with known stem cell biology. SOX2 was predicted to regulate GATA3, EGR2, and POU2F1, consistent with its role in modulating transcriptional plasticity and epithelial-mesenchymal transition during reprogramming [38-40]. MYC was linked to cell cycle and biosynthetic regulators such as TP53, NFYB, and FOXM1, in line with its function in promoting proliferation and metabolic activity [41,42]. POU5F1 was associated with ETS2, ZBTB20, and PRDM1, reinforcing its involvement in epiblast specification and developmental gene silencing [43-45]. NANOG, another core regulator of naïve pluripotency, was connected to key early developmental factors including NODAL, GDF3, and KLF2[46-48]. GRNFormer also captured several well-supported regulatory links validated by ChIP-seq data, including SOX2-ZEB1, MYC-TCF7L1, MYC-SNAI1 and NANOG-DKK1, further confirming its ability to infer high-confidence regulatory edges from expression data alone [49-52].

In addition to predicting known pluripotency networks, GRNFormer revealed a previously uncharacterized transcriptional module centered on GATA6, HAND1, NRP1, HNF1B, and TET2 (Figure 4B). These transcription factors are not typically active in ground-state pluripotency but are known to mediate early lineage specification events, including mesendoderm formation, cardiac development, neurulation, and vascular morphogenesis [53,54]. TET2, in particular, plays a role in DNA demethylation and chromatin remodeling, and its inclusion suggests that this module may reflect an epigenetically primed state [55]. Importantly, this subnetwork was absent from the ChIP-seq-derived hESC gold-standard networks, implying that GRNFormer may capture regulatory heterogeneity or transient cell states that elude bulk profiling techniques.

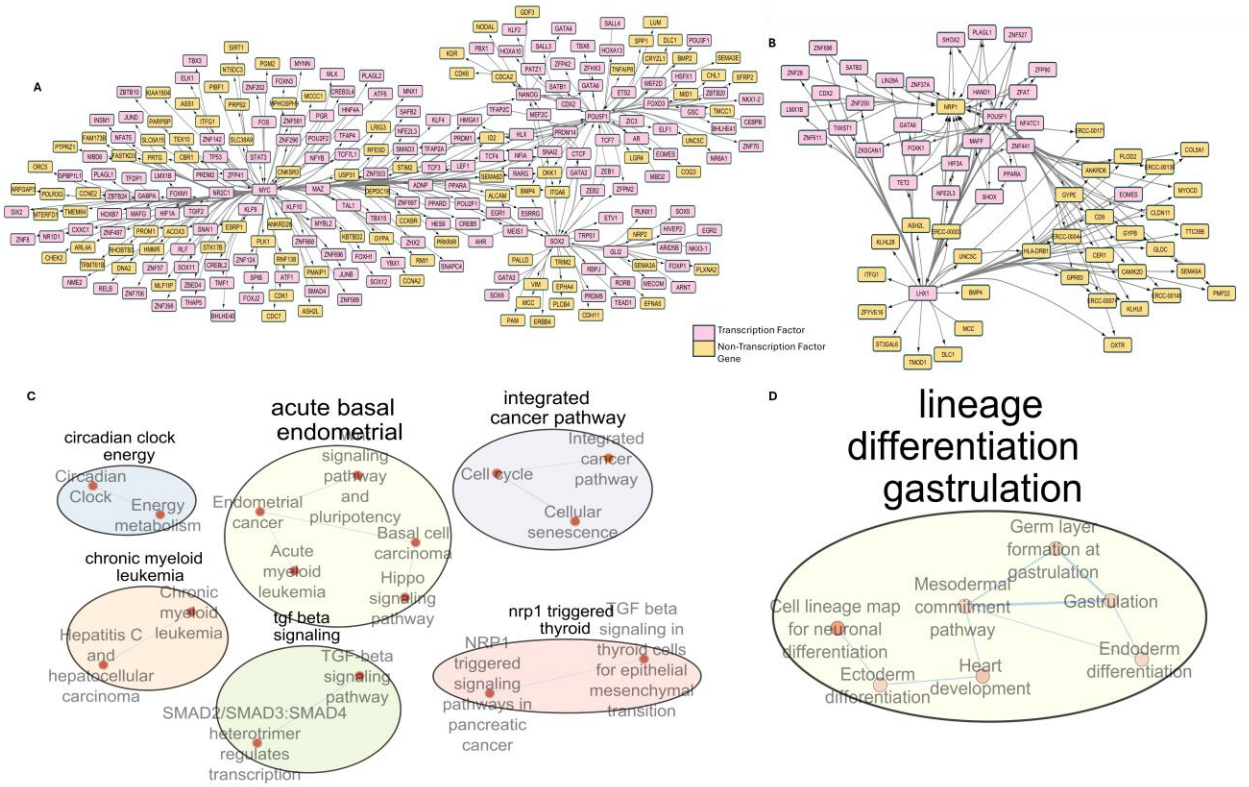

Figure 4: Application of GRNFormer to hESC cells. (A) GRNFormer predicted hESC-cell-type-specific GRN subgraph, centering around biomarkers of hESC cell types. The predicted GRN subgraph is verified against the ground truth network. (B) novel regulatory interactions of hESCs predicted by GRNFormer outside of ground truth networks. (C) Enrichment map of GRN shown in 4A. (D) Enrichment map of GRN shown in 4B

To characterize the functional relevance of these predictions in Figure 4A and 4B, we performed pathway enrichment analysis using g: Profiler[56] and visualized the results in Figure 4C and 4D respectively using Cytoscape[57]. The targets of core pluripotency TFs (Figure 4C) were enriched in well-known stem cell pathways, including TGF-β/SMAD signaling [58], Wnt and Hippo signaling [59], circadian rhythm[60], metabolic regulation[61], and chromatin remodeling[62]. In contrast, the targets of the novel regulatory module (Figure 4D) were enriched for developmental programs, such as gastrulation and specification of the endoderm, mesoderm, and ectoderm [63-67]. Additional pathways included cardiac development and neuronal differentiation [68-70]. These findings indicate that the network in Figure 4B reflects an early, lineage-primed transcriptional program-potentially marking the transition from naïve pluripotency to germ layer commitment.

Notably, the developmental programs enriched in the novel subnetwork were absent from the core pluripotency GRN, supporting the view that GRNFormer distinguishes between stable and transitional regulatory states. These predictions suggest the presence of cryptic or low-frequency transcriptional programs within the hESC population that are not captured in curated ground truth but may represent early determinants of cell fate decisions. Altogether, this case study

demonstrates GRNFormer's ability to recover both well-established regulatory interactions and novel, biologically plausible transcriptional modules. Its performance in hESCs illustrates how the model can identify context-specific GRNs with developmental significance and provides a framework for uncovering hidden regulatory states relevant to differentiation and disease.

GRNFormer recovers lineage-resolved regulatory programs and functionally enriched modules in PBMCs via blind inference

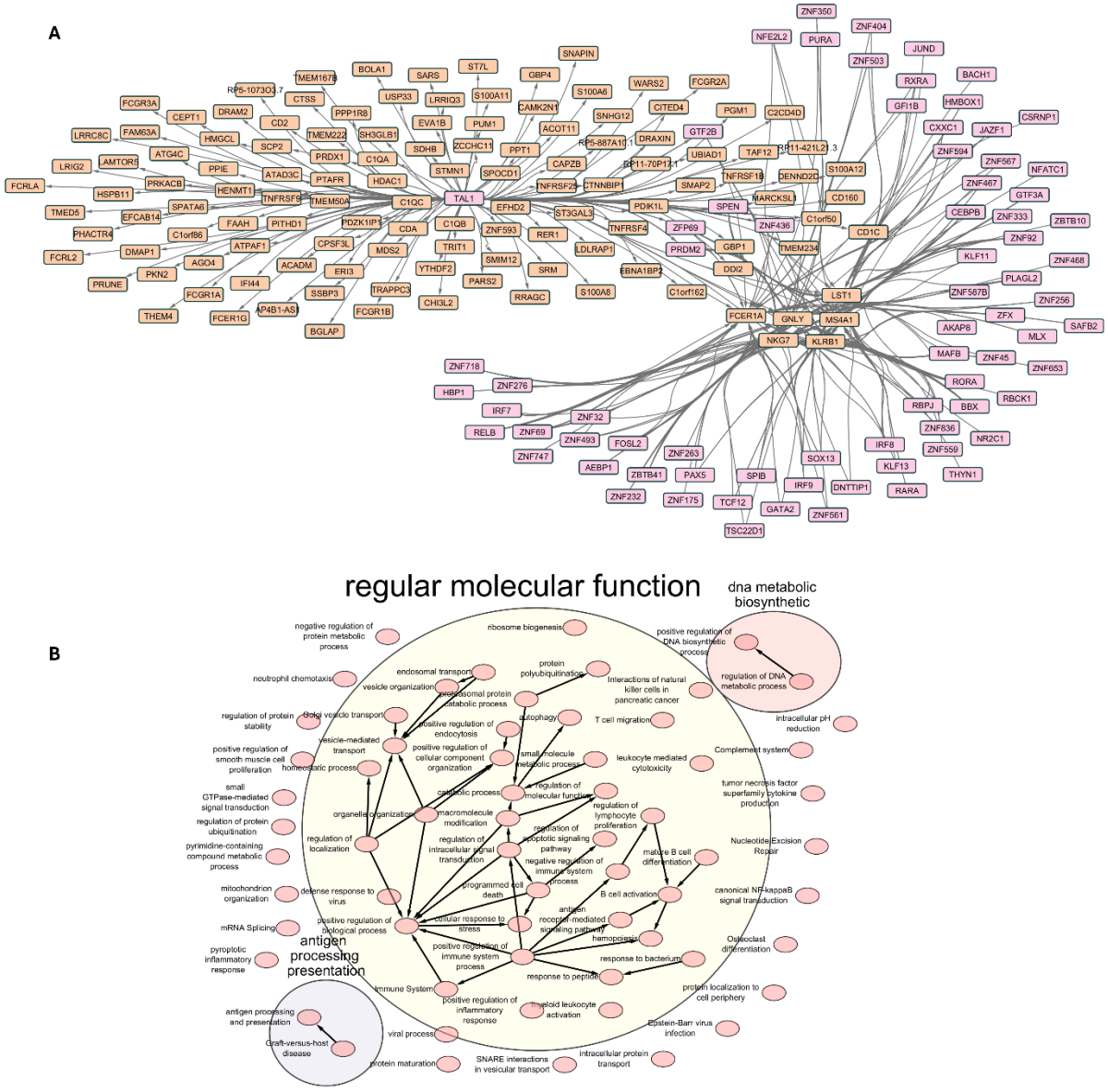

Figure 5: GRNFormer recovers cell-type-specific regulatory networks and functional modules from PBMCs via blind inference. (A) Subnetwork from the GRNFormer-predicted GRN on the

PBMC 3k dataset, highlighting six key regulators (MS4A1, GNLY, KLRB1, LST1, FCER1A, and TAL1) corresponding to major immune lineages. The pretrained GRNFormer was applied to the dataset blindly without access to cell-type labels or prior pathway information. (B) Pathway enrichment map derived from GRN targets in (A), using g: Profiler and visualized with Enrichment Map in Cytoscape. Clusters represent immune processes such as antigen presentation, interferon signaling, cytotoxic activity, and B-cell DNA repair. Several predicted edges are supported by prior literature, validating the model's biological relevance.

To evaluate the biological specificity of GRNFormer in resolving cell-type-resolved transcriptional programs, we applied it in a zero-shot setting to the 10x Genomics PBMC 3k dataset, a standard benchmark for immune single-cell profiling [71]. The dataset was obtained in preprocessed form via the Scanpy toolkit, with GRNFormer operating without any access to cell-type labels, clustering results, or pathway priors. Instead, the model inferred regulatory structure directly from gene expression [72].

A subnetwork extracted from the GRNFormer-predicted GRN (Fig. 5A) highlights high-centrality genes with strong immunological relevance, including MS4A1, GNLY, NKG7, KLRB1, LST1, FCER1A, and TAL1. These genes mark key immune lineages: MS4A1 (CD20) marks B cells [73], GNLY and NKG7 define cytotoxic T and NK cells[74], KLRB1 is linked to NK cell function[75], LST1 marks monocytes[76], FCER1A labels plasmacytoid dendritic cells[77], and TAL1 regulates early hematopoietic lineage commitment[78]. These predictions were made without cell-type labels, underscoring the model's capacity to capture underlying regulatory logic directly from transcriptomic signals.

To interpret the functional landscape of the inferred network, we also performed pathway enrichment analysis using g: Profiler and visualized the results in Cytoscape via Enrichment Map. The enriched pathways formed coherent modules aligning with known PBMC biology (Fig. 5B). A central cluster reflected general immune functions including vesicle trafficking, antigen presentation, and apoptotic signaling, suggesting basal regulatory programs across T cells and monocytes [79,80]. A distinct cluster enriched for "graft-versus-host disease" and "viral process" mapped to monocytes and dendritic cells, known for their roles in antigen processing and cytokine signaling [81,82]. Another module featured pathways like "DNA metabolic process," "complement activation," and "nucleotide excision repair," consistent with transcriptional activity in activated B cells and plasma cells [83-85].

Additional pathways such as "T cell migration," "immune effector regulation," and "leukocyte-mediated cytotoxicity" mapped to $CD8^+$ T cells and NK cells, reflecting cytotoxic immune functions [86-88]. Interaction module around TAL1 inferred by GRNFormer is supported by literature evidence, that TAL1 is a master regulator of hematopoiesis [89].

These findings demonstrate that GRNFormer can infer biologically coherent cell-type-specific functional modules from single-cell data without external supervision. The model not only recapitulates known transcriptional circuits in immune lineages but also organizes them into interpretable regulatory programs, validating its utility for de novo GRN reconstruction in complex tissues.

Discussion

In this study, we introduce GRNFormer, a transformer-based graph learning framework for reconstructing gene regulatory networks (GRNs) from gene expression data, capable of blind inference without requiring any prior annotations such as cell types and pathways. Through a unified architecture comprising the TransConv block, a gene-centric embedding module (Gene Transcoder), and an attention-based decoder, GRNFormer delivers accurate, biologically meaningful predictions across diverse organisms, transcriptomic modalities, and developmental contexts. Our method not only recovers known regulatory interactions but also identifies novel ones, enabling robust downstream biological insights.

One of the key innovations of GRNFormer lies in its contextual and scalable modeling of transcriptional dependencies. Unlike conventional GRN inference methods that rely heavily on pre-defined motifs, GRNFormer learns regulatory patterns via localized subgraphs centered on transcription factors (TFs), sampled using the TF-Walker strategy. This allows the model to focus on neighborhoods enriched with regulatory signals while preserving inductive generalization. Our ablation analysis revealed that replacing this targeted subgraph sampling with random selection significantly degraded performance, underscoring its critical role. The Gene Transcoder module further enables biologically meaningful representation learning by capturing functional latent features specific to gene regulatory dynamics, independent of explicit sequence or motif priors.

We demonstrate that GRNFormer generalizes effectively across species (human, mouse, yeast, and E. coli), cell types, data types (single-cell and bulk RNA-seq), and regulatory scales (non-specific, cell-type specific and protein-protein interactions). Notably, its performance on the *DREAM5* bacterial and yeast datasets - despite being trained without any prior knowledge of prokaryotic systems - approached theoretical upper bounds (AUROC ~0.97, AUPRC ~0.95), reflecting its strong cross-domain robustness. Furthermore, the model captured lineage-resolved regulatory programs in complex human tissues and predicted functional modules relevant to processes such as antigen presentation, pluripotency maintenance, and Epithelial–Mesenchymal Transition (EMT), even in blind inference settings. Particularly striking is the model's ability to uncover developmental pathways such as gastrulation and cross-context interactions (e.g., pluripotency-oncogenesis linkages), which are rarely captured by traditional models.

Beyond performance, GRNFormer offers practical advantages in data efficiency, scalability, and interpretability. Different from the existing supervised GRN inference methods that need to be trained on the same species that they are applied to, its ability to generalize to new species without the need of retraining makes it attractive for poorly annotated systems or understudied organisms. Importantly, its modular architecture allows for rapid adaptation to new tasks or data modalities. Moreover, through edge-centric decoding, the model integrates context-aware regulation without explicitly relying on full graph propagation, offering computational scalability to networks with thousands of genes.

Nevertheless, GRNFormer has certain limitations. Although GRNFormer performs well in zero-shot and cross-context settings, its accuracy may be reduced in datasets with extreme sparsity or weak transcription factor signal. While the model infers directional TF-target interactions based

on expression-driven subgraphs, with the TF-centered subgraph sampling strategy, which is computationally efficient and biologically motivated, but may also miss some long-range or distal regulatory interactions such as enhancer-promoter loops or inter-chromosomal contacts-that fall outside gene expression neighborhoods.

Looking forward, future developments could extend GRNFormer's capabilities in several directions. Integrating multi-omics data, such as chromatin accessibility (ATAC-seq), epigenetic modifications, and TF motif occupancy, could enrich the contextual learning of regulatory interactions. Similarly, adapting the architecture to model temporal dynamics would enable causal modeling of transcriptional cascades. Further, incorporating cross-cell alignment or contrastive learning objectives may improve generalization in datasets with limited signal or across species. Finally, GRNFormer opens up exciting avenues for hypothesis generation in developmental biology, cancer research, and synthetic gene circuit design, where interpretable and robust GRNs are crucial.

Online Methods

Construction of gene co-expression network

To construct the gene co-expression network (GCEN), raw single-cell RNA-seq (scRNA-seq) data were first normalized using the inverse hyperbolic sine (Arc Sinh) transformation to stabilize variance across highly skewed expression profiles [90]. Pairwise Pearson correlations were then computed between all genes across individual cells to quantify co-expression, and gene pairs with absolute correlation >0.1 were retained to form the GCEN. In this graph, nodes represent genes and edges denote co-expression relationships, forming a dense, high-dimensional network. Figure 6A shows the pipeline of GCEN construction.

TF-Walker - a de -novo subgraph sampling method

scRNA-seq data presents a fundamental challenge for machine learning: expression matrices are inherently high-dimensional and sparse, with expression values captured across thousands of genes and cells. However, these datasets often contain only a limited number of training examples. This imbalance results in under-constrained models, where the number of features far exceeds the number of samples suitable for learning regulatory patterns. Consequently, standard deep learning methods risk overfitting, fail to generalize, or learn spurious correlations when trained directly on full expression matrices or dense co-expression graphs.

To address challenges, GRNFormer incorporates a biologically motivated subgraph sampling strategy, termed **TF-Walker (**Figure 6B**)** that functions as a principled form of data augmentation. Instead of operating on the full GCEN, TF-Walker extracts local, transcription factor (TF)-centered subgraphs that capture meaningful co-expression contexts. These TF-anchored subgraphs act as modular training units, preserving regulatory context while increasing the number of training samples through structured subgraph sampling. During training, the algorithm begins with a TF as the root node and identifies its immediate neighbors. If the number of first-hop neighbors is greater than 100, a random subset of 99 nodes is selected to form the subgraph with TF as a center. If

fewer neighbors are available, the walker expands second and third hops iteratively until the 100-node threshold is met. This stochastic sampling encourages the model to generalize over diverse local contexts. The randomly sampled subgraphs can be used to train GRNFormer to predict TF-gene regulatory relationships within them, creating a large amount of data from each single GCEN for training.

In contrast, during inference, TF-Walker deterministically extends to all available neighbors of a given TF, traversing the GCEN in a sequential and exhaustive manner to generate subgraphs. This ensures that regulatory predictions are based on the full local topology of the expression network, without arbitrary truncation. Additional details on the inference strategy are provided in the Online Methods: GRN Inference section.

Each sampled subgraph includes node-level features (gene expression values and binary TF indicators) and edge features (co-expression correlation weights). Gene expression of each subgraph undergoes cellwise z-score normalization across genes. This normalization mitigates information leakage by preventing memorization of gene identities across subgraphs while preserving local expression dynamics.

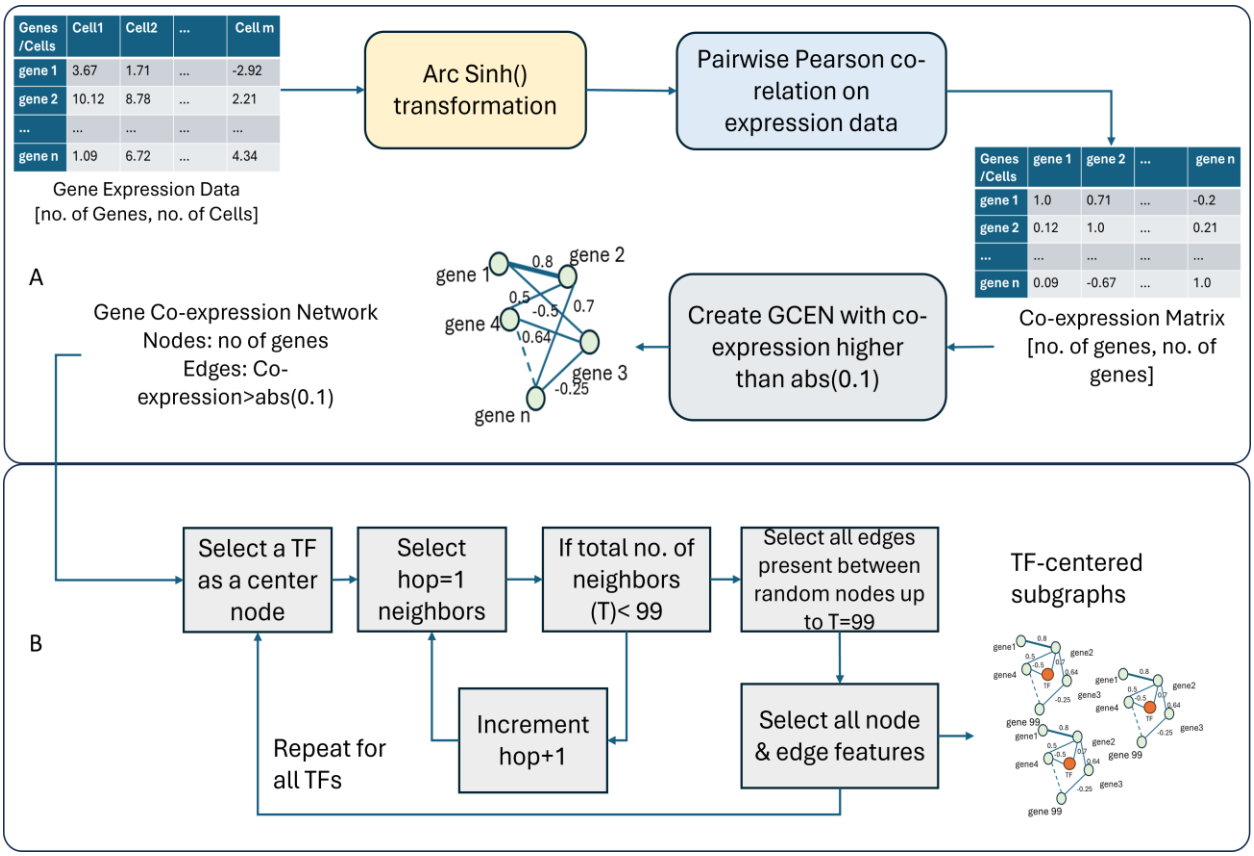

Figure 6. Input data preparation. (A) Gene expression preprocessing steps and gene co-expression network generation; (B) TF-walker methodology used for subgraph sampling during training of GRNFormer.

Ground Truth Network Sampling and Feature Representation

To supervise GRNFormer training, we constructed comprehensive ground truth regulatory networks by integrating three validated sources: cell-type-specific ChIP-seq data, non-specific ChIP-seq networks, and STRING-derived protein-protein interactions. This multi-source integration enhances biological coverage and supports cross-species generalization.

For each TF-centered subgraph generated by TF-Walker, a corresponding labeled regulatory subgraph was created by identifying all nodes (genes) in the subgraph and extracting known regulatory interactions among them from the global ground truth network. If no regulatory edges existed between nodes, the subgraph was discarded to ensure only biologically informative samples were used for training. This approach guarantees that the model learns from functionally relevant regulatory relationships, improving inference reliability.

Each subgraph is represented as a graph of 100 genes (nodes), where the features of each node consist of the expression values of the gene across all cells along with a binary indicator denoting TF identity (i.e., if a gene is a TF). Edge features are defined by pairwise Pearson correlation values computed from the full dataset, capturing global co-expression structure. While the edge weights reflect broader transcriptional associations, the node-level gene expression inputs provide localized information contextualized at the subgraph level.

Gene-Transcoder: Transformer encoder for gene expression representation learning

Figure 7A illustrates the architecture of the Gene-Transcoder, a transformer-based encoder designed for gene expression representation learning, which is the first stage in training GRNFormer. The primary challenge addressed here is the variability in gene expression data from scRNA-seq experiments, where the number of genes and cells can differ significantly across datasets. This variability makes it difficult for traditional models to learn and generalize effectively. To tackle this issue, the GENE-Transcoder processes gene expression data to produce fixed, representative embeddings that capture essential biological information.

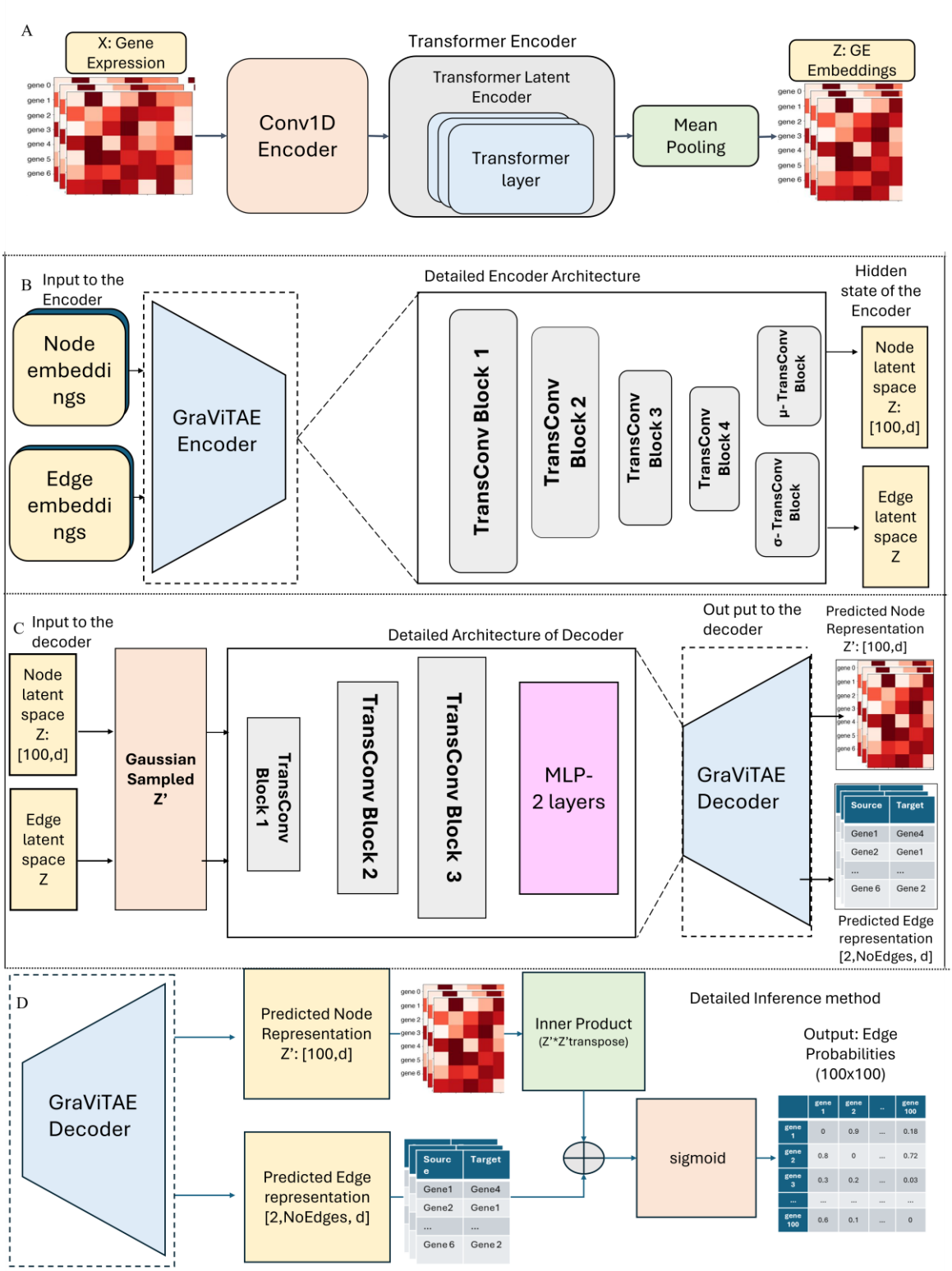

Figure 7: The four key sequentially connected components of GRNFormer. (A) Architecture of Gene Transcoder for fixed embeddings of gene expression. (B) Detailed architecture of GraViTAE Encoder. (C) Detailed architecture of GraViTAE variational decoder. (D) Edge probability prediction or GRN inference module

Gene-Transcoder begins with an input gene expression matrix $\in \mathbb{R}^{\text{batchsize}\times 100\times \text{noofcells}}$. A 1D convolutional layer processes the gene expression matrix. This convolution utilizes a sliding kernel of size 1 along the cell axis to capture local patterns of gene expression across cells. The output from the convolutional layer is then fed into a transformer encoder, where a multi-head attention mechanism is employed. This mechanism helps the model capture both local and global interactions between genes and cells by producing embeddings that represent gene expression data in an embedding feature space. Finally, a mean pooling operation is applied to the transformer output, averaging the gene embeddings over all cells. By lowering the dimensionality and providing each gene with a consistent, meaningful representation, this enables the model to oversee the inherent unpredictability in gene expression data. Detailed methodology and mathematical explanations for the Gene-Transcoder are provided in supplementary Method 1.

The resulting gene embeddings of dimension $\mathbb{R}^{\text{batchsize}\times 100\times 64}$are fixed, context-aware representations of a set robust features for downstream task such as GRN inference. These embeddings are independent of dataset-specific variations, enabling the model to generalize effectively across species, cell types, and experimental conditions.

Graph Variational Transformer Autoencoder (GraViTAE)

Following Gene-Transcoder is the Graph Variational Transformer Autoencoder (GraViTAE)which is the backbone or core of the pipeline. GraViTAE is a supervised variational graph transformer autoencoder composed of an encoder and decoder that operate on TF-centered gene co-expression subgraphs. The model integrates fixed-length node embeddings from the Gene-Transcoder with co-expression edge weights to jointly for further message passing.

GraViTAE Encoder: Learning Node and Edge Latent Representations

The GraViTAE encoder shown in Figure 7B, applies stacked Transformer Convolution (TransConv) blocks to capture both local (e.g., between closely related genes) and global gene-gene interactions (e.g., co-expression relationships across a subgraph). Each block uses multi-head attention and feedforward layers for message passing between node and edge features. To promote generalizability and to regularize the latent space, the encoder estimates distributions over latent variables by computing both the mean (μ) and variance (σ^2) of node and edge representations. These parameters are used to sample continuous latent embeddings via a Gaussian reparameterization trick.

This variational formulation allows GRNFormer to model uncertainty and variability in noisy single-cell expression data, while the joint encoding of expression and co-expression features ensures biologically grounded, context-aware representations.

GraViTAE Decoder

The GraViTAE decoder (Figure 7C) reconstructs node and edge representations from latent representations sampled by the encoder. These latent variables are drawn from Gaussian distributions parameterized by the encoder's outputs, allowing the model to capture uncertainty inherent in noisy single-cell gene expression data.

Using these samples, the decoder applies additional TransConv blocks followed by a lightweight multilayer perceptron (MLP) to produce two outputs: updated node embeddings, representing each gene's regulatory identity, and edge attentions, indicating the strength of potential regulatory interactions. These outputs are combined to reconstruct the GRN at the subgraph level. These outputs serve as intermediate proxy representations to guide the downstream GRN inference.

Graph Transformer Convolution Layer with Pairwise Attention and TransConv Block in GraViTAE

The TransConv block serves as the foundation for both the GraViTAE encoder and decoder modules (Figure 7B and 7C). TransConv block builds on *Transformer Convolution* layers by stacking multi-head pairwise attention with residual feedforward layers and nonlinear transformations (Figure S2 A). This structure allows GRNFormer to simultaneously model local dependencies, such as short-range transcriptional regulation, and global structures, such as co-regulation across transcriptional modules.

At the heart of the TransConv block lies the *Transformer Convolution* (TransConv) layer, a novel attention based a graph representation learning approach that extends traditional graph convolution by jointly updating node and edge features a pairwise attention mechanism. Unlike traditional graph convolutional networks that primarily focus on node attributes, TransConv explicitly integrates edge attributes such as co-expression into its pairwise attention computations, along with node features, allowing for context-aware message passing. This dual integration is critical in biological networks, where regulatory relationships are often mediated through complex combinations of gene expression and co-expression strength.

For each gene as a node, attention weights are computed by comparing the query representation of a target gene with key representations of its neighbors, modulated by their connecting edge features. This pairwise attention mechanism allows the model to selectively weigh incoming messages based not only on the features of neighboring genes but also on the biological plausibility of their interactions. The resulting attention scores guide message passing, enabling the model to distinguish between biologically meaningful and spurious links (see Supplementary Method 2 and Supplementary Fig. S2 B for formalization). The layer is adapted from a Transformer-based GNN architecture [91] and implemented using the pytorch_geometric framework.

Following the attention computation, the aggregated messages are passed through a lightweight multilayer perceptron (MLP) equipped with Leaky ReLU activations. This choice provides improved gradient flow and is particularly effective at learning negative co-expression signals - patterns often indicative of repressive regulatory mechanisms that standard ReLU activations tend to suppress [92] (see Supplementary Fig. S2 A for architectural diagram).

To further stabilize training and enhance representation fidelity, batch normalization is applied after each nonlinear transformation. Collectively, the components of the TransConv block offer a robust and generalizable architecture, effectively modeling the combinatorial and context-dependent nature of transcriptional regulation.

Adjacency Matrix Reconstruction and GRN Inference

The final step of GRNFormer infers a probabilistic GRN by reconstructing an adjacency matrix that captures the likelihood of regulatory interactions between genes (Fig. 7D). The decoder outputs two components: node embeddings $Z' \in R^{N\times d}$, where N stands for number of genes in the subgraph, encoding each gene's regulatory identity, and edge representations $E' \in R^{2\times M\times d}$,where M stands for number of edges present in input subgraph, capturing pairwise regulatory features. d is the dimension of decoder output. To estimate regulatory interactions, an initial adjacency matrix A_0 is computed via the inner product of node embeddings:

$$A_0 = Z'(Z')^T$$

This operation yields a symmetric interaction matrix, where each element $A_0[i,j]$ reflects the interaction strength between gene i and gene j, as inferred from their respective representations.

To enrich these pairwise estimates with contextual information captured at the edge level, the model applies mean pooling across the edge dimension of E' and integrates the resulting values into the corresponding positions of the adjacency matrix. Let $\bar{E} \in R^{N\times N}$ denote this pooled edge tensor, where each entry represents an aggregated edge feature score between gene pairs. The combined interaction matrix becomes:

$$A = A_0 + \bar{E}$$

Finally, a sigmoid activation function is applied elementwise to the combined matrix to transform the raw interaction scores into probabilistic values bounded between 0 and 1:

$$P = \mathrm{Sigmoid}(A),$$

where $P[i,j] \in [0,1]$ indicates the predicted probability of a regulatory relationship between gene i and gene j.

This subgraph-level inference captures modular gene regulation, while a dedicated strategy aggregates across subgraphs to reconstruct a coherent global GRN (see GRN Inference Section for details).

The final adjacency matrix, which encodes the predicted probabilities of regulatory interactions between gene pairs, is optimized using a binary cross-entropy loss that treats GRN inference as a binary link prediction task. Known regulatory links within each subgraph serve as supervision targets, enabling the model to learn to discriminate between interacting and non-interacting gene pairs. To promote smooth, disentangled latent representations and encourage generalization across diverse regulatory contexts, the objective is further regularized with a Kullback–Leibler (KL) divergence loss that penalizes deviation from a standard Gaussian prior (see Training section for details).

Datasets and Preprocessing

BEELINE data

To train and evaluate GRNFormer in biologically diverse settings, we used single-cell RNA-seq datasets curated by BEELINE, a comprehensive benchmarking framework for GRN inference. These datasets span two species (human and mouse), seven cell types, and include three types of regulatory ground truth networks, offering a broad foundation for testing model generalization across cell types, regulatory contexts, species, and evolutionary backgrounds.

Our analysis focused on five representative datasets: mouse hematopoietic stem cells (mHSC-E, mHSC-GM, mHSC-L), embryonic stem cells (mESC), dendritic cells (mDC), and human hepatocytes (hHep) and embryonic stem cells (hESC). Together, they reflect a wide range of biological processes from early development and lineage commitment to immune regulation and tissue-specific differentiation. All expression data were preprocessed using BEELINE's standardized workflow, including log transformation, gene filtering, and normalization, ensuring consistency and comparability across experiments.

Each expression dataset was paired with one or more regulatory reference networks for supervised training and validation. These included cell-type-specific transcription factor-target interactions, derived from experimental ChIP-seq datasets, non-cell-type specific ChIP-seq networks, as well as functional protein interaction networks from STRING. The combined use of regulatory evidence enabled robust benchmarking of GRNFormer's predictions in both precise and broad regulatory contexts.

DREAM5 Challenge and PBMC Datasets

We applied GRNFormer pretrained on the single-cell BEELINE datasets to bulk RNA-seq datasets from the DREAM5 challenge, including the data of *Escherichia coli* and *Saccharomyces cerevisiae*. These well-characterized prokaryotic and eukaryotic networks, with gold-standard ChIP-derived regulatory maps, provided an opportunity to evaluate the model's cross-species transferability and robustness to different transcriptomic data types. To further test generalizability, we conducted zero-shot inference using the 10x Genomics PBMC 3k dataset, which lacks any cell-type labels or prior regulatory annotations. This allowed us to assess GRNFormer's ability to recover meaningful regulatory structure de novo.

Training

To assess the generalizability of GRNFormer, we adopted a cross-lineage training and evaluation strategy. The model was trained and validated on subgraphs derived from five datasets - human embryonic stem cells (hESC), human hepatocytes (hHep), mouse dendritic cells (mDC), and two mouse hematopoietic stem cell subtypes (mHSC-GM, mHSC-E) - while holding out mouse embryonic stem cells (mESC) and an additional mHSC subtype (mHSC-L) for blind testing. This design enables evaluation across both species and developmental lineages. From each dataset, we first constructed full gene co-expression networks (GCENs) and then applied TF-Walker subgraph sampling to generate localized training instances. Ground-truth regulatory interactions were aggregated from three sources - cell-type-specific ChIP-seq, non-specific ChIP-seq, and STRING

networks - into a unified training label set, allowing the model to learn from heterogeneous regulatory modalities.

During training, GRNFormer samples TF–centered subgraphs by randomly selecting 99 neighboring genes from each TF's local co-expression neighborhood such that each subgraph has 100 genes including each TF itself. This strategy is designed to expose the model to diverse TF-centered regulatory contexts across multiple species and cell types. By training on shuffled subgraphs drawn from a broad collection of datasets, the model learns generalizable patterns of transcriptional regulation, rather than overfitting to any specific expression profile or network topology.

To prevent memorization and promote generalization, each subgraph is z-score normalized across genes present within subgraph, during both training and evaluation ensuring that even if a gene appears in multiple subgraphs, the model treats it as an independent node due to its context-specific feature values.

. The Gene-Transcoder module encodes variable-length gene expression profiles of subgraph genes into fixed-length node embeddings of dimension 64. These node embeddings, along with co-expression values used as edge features, are passed to the encoder implemented as a part of variational graph transformer autoencoder (GraViTAE). The encoder comprises four Transformer Convolution (TransConv) blocks, each with 4 attention heads, and projects the input features into a latent space of dimension 16.

The sampled latent node and edge representations are then passed through the GraViTAE decoder, which consists of three additional TransConv blocks followed by two fully connected (MLP) layers, reconstructing interaction-specific representations. These intermediate outputs are aggregated and transformed into probabilistic edge predictions used to infer the GRN.

GRNFormer was trained for 100 epochs with a batch size of 8, using a composite loss function that combines binary cross-entropy (BCE) for predicting regulatory edge presence with Kullback–Leibler (KL) divergence to regularize the variational latent space. The training objective is framed as a binary classification task: identifying whether an edge between any two genes within a subgraph corresponds to a known regulatory interaction.

The binary cross-entropy loss evaluates the discrepancy between the predicted edge probabilities $\widehat{y_{ij}}$ and the true binary labels $y_{ij} \in \{0,1\}$ for each gene pair (i,j), and is computed as:

$$\mathcal{L}_{\mathrm{BCE}} = -\sum_{(i,j)}\left[y_{ij} \cdot \log\left(\widehat{y_{ij}}\right) + \left(1 - y_{ij}\right) \cdot \log\left(1 - \widehat{y_{ij}}\right)\right]$$

To enforce regularization in the variational latent space, we incorporate the Kullback–Leibler divergence between the approximate posterior $q(z|x)$ (parameterized by the encoder) and a standard normal prior $p(z)$. This is given by:

$$\mathcal{L}_{\mathrm{KL}} = -\frac{1}{2}\sum_{i=1}^{n}(1 + \log(\sigma_i^2) - \mu_i^2 - \sigma_i^2)$$

where μ_i and σ_i^2 are the mean and variance of the latent embedding for gene i.

The total loss function is the weighted sum of these two terms:

$$\mathcal{L}_{total} = \mathcal{L}_{BCE} + \beta \cdot \mathcal{L}_{\mathrm{KL}}$$

where β is a scaling factor with value $0.01\ or\ \frac{1}{\#genes\ in\ subgraph}$, controlling the strength of KL regularization.

Optimization was performed using Adam (initial learning rate 0.001) with a ReduceLROnPlateau scheduler and early stopping. L1 regularization was applied to prevent overfitting.

To address extreme class imbalance-typical of GRN inference where true regulatory edges are sparse, we employed dynamic negative sampling during training. For each subgraph, we matched the number of negative (non-regulatory) edges to the number of positives, ensuring balanced and effective learning across highly imbalanced edge distributions.

Importantly, the model was trained without any explicit knowledge of species, cell type, or regulatory context, and therefore is expected to work for any species, cell type, and biological conditions. All subgraphs were randomly shuffled prior to training, forcing the model to rely solely on expression-derived contextual signals. This design allows GRNFormer to generalize across unseen biological systems, demonstrating its capacity to infer regulatory logic across species, tissues, and regulatory types. GRNFormer is trained on 2 GPUs of NVDIA A10, on national GPU clusters of Nautilus server.

GRN Inference and Evaluation on Test Datasets

During inference, GRNFormer applies an extended version of the TF-Walker strategy to generate high-coverage subgraphs for each test dataset (Figure 8). For every TF, subgraphs are sampled sequentially to include all co-expressed neighbors, rather than randomly selecting some as in training. This ensures maximal coverage of genes and local co-expression interactions across the expression space. Multiple subgraphs are generated per TF, with or without overlapping across gene content. Because each subgraph undergoes z-score normalization across genes, any gene appearing in multiple subgraphs is treated as independent instance in a context-specific manner, reducing bias and improving inference robustness. The predicted regulatory interactions from all subgraphs are then aggregated to reconstruct a comprehensive GRN for the dataset. To prioritize regulatory relevance, we focused on TF-gene and TF-TF interactions, which form the core of most GRNs. For each predicted subgraph, we retained interactions with a predicted probability >0.3. To weigh TF-centered predictions more strongly, interactions where the source gene was a TF were given higher priority during aggregation.

Each retained interaction was assigned a weighted score based on its predicted probability, TF involvement, and frequency across subgraphs. These scores were then summed and averaged over all occurrences of the interaction, followed by min-max normalization to yield a final adjacency matrix with values scaled between 0 and 1. This matrix represents the full predicted GRN, with entries reflecting the likelihood of regulatory relationships between gene pairs.

This aggregation framework above allows GRNFormer to scale beyond subgraph-based inference, capturing broader gene-gene dependencies while maintaining biological specificity. By emphasizing TF-driven predictions and accounting for sampling frequency, the strategy produces robust, interpretable GRNs that generalize well across large and complex datasets.

1. **Initialize**
   - Input: Graph G = (V, E), transcription factors (TFs) as a list, target subgraph size T = 99.
   - Output: Subgraphs S1, S2, ..., Sk.

2. **For each TF v ∈ TF list:**
   Step 2.1: Identify 1-hop neighbors.
   - Find all 1-hop neighbors N1(v) of v, v is a TF node, and N1(v) is a 1-hop neighbors of v.

   Step 2.2: Check the size of N1(v):

   - **Case 1: |N1(v)| < T:**
     - Increment hops by 1.
     - Include neighbors from hop + 1 until |N(v)| ≥ T (where N(v) is the set of unique neighbors).
     - Proceed to Step 2.4.

   - **Case 2: |N1(v)| ≥ T**:
     - **Subcase 2.1**: N1(v) is too large to fit into a single subgraph:
       - Divide N1(v) into multiple subgraphs S1, S2, ..., such that each subgraph contains ≤ T nodes.
       - If any subgraph Si has fewer than T nodes, add nodes from hop + 1 neighbors to reach T.
     - **Subcase 2.2**: If a single subgraph is to be created with a mix of neighbors from different depths:
       - Assign the first T/2 nodes to S1 from 1-hop neighbors.
       - Assign the remaining T/2 nodes from hop + 1 neighbors to S1.
       - If N1(v) and Nk(v) (incrementally) still do not meet the target size T:
         - Increment the hop and include additional k-hop neighbors iteratively until T-sized subgraphs are satisfied.
         - For each depth, maintain balanced selection by ensuring T/2 allocation from N1(v) and subsequent depths.

3. **Construct subgraphs for each TF v:**
   - Include all edges Es ⊆ E where both endpoints belong to the subgraph nodes.
   - Include all node features and edge features for the selected nodes and edges.

4. **Process remaining uncovered nodes in V:**
   - Identify nodes Vr that are not included in any subgraph.
   - Select |Vr| as center nodes and repeat the process until all nodes are covered.
   - For subgraphs with fewer than T nodes, increment the hop or randomly add nodes to meet the target size T.

5. **Output the final subgraphs**:

Subgraphs S1, S2, ..., Sk, where each subgraph has exactly T nodes (if possible), and all edges and features within the subgraphs are included.

Figure 8. Algorithm of real-time subgraph sampling for GRN inference.

We evaluated GRNFormer's performance using a suite of standard classification metrics: AUROC (Area Under the Receiver Operating Characteristic Curve), AUPRC (Area Under the Precision-Recall Curve), F1 score, precision, recall, and accuracy. Each metric provides a complementary perspective on prediction quality, particularly in the context of class imbalance inherent to GRNs. We also performed regulatory context-specific evaluations on gene subsets (e.g., top 500 or 1,000 variable genes, with or without TFs), derived from the unseen cell-type gene expression matrices, designed to probe performance under constrained and biologically meaningful feature sets.

AUROC quantifies the model's ability to distinguish regulatory from non-regulatory interactions across all thresholds. AUPRC emphasizes the model's accuracy in detecting true regulatory edges, making it especially informative for imbalanced datasets.

F1 score balances precision and recall, offering a single summary of model performance when both false positives and false negatives matter. Precision reflects the fraction of predicted regulatory edges that are correct, while recall captures the proportion of true regulatory interactions successfully recovered. F1 score, precision, and recall depend on the decision threshold used to select positive predictions. The threshold leading to the maximum F1 score is used. Although accuracy (percentage of correct predictions in all predictions) is also reported, it is less reliable in imbalanced settings and used primarily for completeness.

Because true negative regulatory interactions are rarely annotated, several GRN inference methods adopted random negative sampling during evaluation to construct balanced datasets. Deep learning approaches such as GRLGRN [93], CNNC, and scGREAT sampled non-interacting gene pairs as negatives to compute binary classification metrics. Following this precedent, to further ensure robust and unbiased evaluation, we employed bootstrapped negative sampling during performance assessment. For each test dataset, positive regulatory interactions were defined by known TF–target edges from the corresponding ground-truth network. However, because true negative interactions are not explicitly annotated in most GRN benchmarks, we generated negative samples by randomly selecting an equal number of gene pairs that do not appear in the ground-truth network. This process was repeated for 100 independent sampling rounds, ensuring that evaluation results were not biased by any particular set of negative edges. For each round, AUROC, AUPRC, precision, recall, and F1 scores were computed by comparing the model's predicted edge probabilities against the union of known positives and the sampled negatives. Final metrics were averaged across all 100 replicates to provide a statistically stable estimate of model performance and minimize the influence of sampling variance.

In benchmarking analyses, we primarily report AUROC, AUPRC score, as they offer the most reliable assessment of performance in sparse and imbalanced GRN settings.

Benchmark Setup

For comparative benchmarking, we evaluated GRNFormer against a representative set of classical and deep learning-based GRN inference methods. To ensure reproducibility and ease of deployment, all the methods were run using Dockized environments provided by BEELINE, where available.

Among classical approaches, we selected four widely used and computationally efficient methods: PIDC, PPCOR, LEAP, and SCENIC-RITES. These were chosen for their speed and scalability, as they completed inference on 500-gene inputs in under one hour. More computationally intensive methods (e.g., GENIE3[94], GRNVBM [95]) were excluded due to runtime limitations in large-scale evaluations.

For deep learning baselines, we included five recent models: GNE, GNNLink, CNNC, STGRNs, and scGREAT. Each was trained and evaluated using its default parameters for both training and inference. To ensure fairness, we used the same train-validation-test data splits across all baseline methods, for every cell types, and performed dataset-specific training as described in their original publications GRNFormer, in contrast, was subjected to a more stringent blind evaluation on the same test split because it was trained on the data of cell types different from that of the test data, which can more rigorously assess generalization across unseen cell types.

Data availability

The data used in this study are publicly available. BEELINE benchmark datasets can be accessed at [https://zenodo.org/records/3701939]. The DREAM5 challenge datasets (E. coli and S. cerevisiae) are available at [https://www.synapse.org/Synapse:syn2787209/wiki/70351]. PBMC 3k data are available via 10x Genomics [https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.1.0/pbmc3k].

Code availability

The GRNFormer codebase, including scripts for model training, evaluation, and benchmarking, is available at [https://github.com/BioinfoMachineLearning/GRNformer]. The code is released under the MIT License and includes documentation and python environments to ensure reproducibility.

Acknowledgements

We thank the University of Missouri High Performance Computing Center for computational resources. This work was supported by an NSF grant (#: CCF2343612).

Author contributions

A.H. designed the model, performed experiments, collected the data, and wrote the manuscript. **J.C.** conceived the project, supervised the study, contributed to the design and interpretation, and revised the manuscript. All authors reviewed and approved the final version of the manuscript.

Competing interests

The authors declare no competing interests.

Ethics declaration

No human or animal subjects were involved in this study, and no ethical approval was required.

References

1. Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K., & Watson, J. D. (1994). Molecular biology of the cell (Vol. 3, pp. 966-996). New York: Garland.
2. Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, *10*(1), 57-63.
3. Ozsolak, F., & Milos, P. M. (2011). RNA sequencing: advances, challenges and opportunities. *Nature reviews genetics*, *12*(2), 87-98.
4. Alon, U. (2007). Network motifs: theory and experimental approaches. *Nature Reviews Genetics*, *8*(6), 450-461.
5. De Smet, R., & Marchal, K. (2010). Advantages and limitations of current network inference methods. *Nature Reviews Microbiology*, *8*(10), 717-729.
6. Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Favera, R. D., & Califano, A. (2006, March). ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. In *BMC bioinformatics* (Vol. 7, pp. 1-15). BioMed Central.
7. Faith, J. J., Hayete, B., Thaden, J. T., Mogno, I., Wierzbowski, J., Cottarel, G., ... & Gardner, T. S. (2007). Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS biology*, *5*(1), e8.
8. Shermin, A., & Orgun, M. A. (2009, March). Using dynamic Bayesian networks to infer gene regulatory networks from expression profiles. In *Proceedings of the 2009 ACM symposium on Applied Computing* (pp. 799-803).
9. Liu, F., Zhang, S. W., Guo, W. F., Wei, Z. G., & Chen, L. (2016). Inference of gene regulatory network based on local Bayesian networks. *PLoS computational biology*, *12*(8), e1005024.
10. Lu, T., Liang, H., Li, H., & Wu, H. (2011). High-dimensional ODEs coupled with mixed-effects modeling techniques for dynamic gene regulatory network identification. *Journal of the American Statistical Association*, *106*(496), 1242-1258.
11. Hickman, G. J., & Hodgman, T. C. (2009). Inference of gene regulatory networks using boolean-network inference methods. *Journal of bioinformatics and computational biology*, *7*(06), 1013-1029.
12. Chen, G., Ning, B., & Shi, T. (2019). Single-cell RNA-seq technologies and related computational data analysis. *Frontiers in genetics*, *10*, 317.
13. Pratapa, A., Jalihal, A. P., Law, J. N., Bharadwaj, A., & Murali, T. M. (2020). Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nature methods*, *17*(2), 147-154.
14. Chen, Y., Li, Y., Narayan, R., Subramanian, A., & Xie, X. (2016). Gene expression inference with deep learning. *Bioinformatics*, *32*(12), 1832-1839.
15. Huynh-Thu, V. A., & Sanguinetti, G. (2018). Gene regulatory network inference: an introductory survey. In *Gene regulatory networks: Methods and protocols* (pp. 1-23). New York, NY: Springer New York.
16. Yang, Y., Fang, Q., & Shen, H. B. (2019). Predicting gene regulatory interactions based on spatial gene expression data and deep learning. *PLoS computational biology*, *15*(9), e1007324.

17. Mochida, K., Koda, S., Inoue, K., & Nishii, R. (2018). Statistical and machine learning approaches to predict gene regulatory networks from transcriptome datasets. *Frontiers in Plant Science*, *9*, 1770.
18. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436-444.
19. Ma, A., Wang, X., Wang, C., Li, J., Xiao, T., Wang, J., ... & Ma, Q. (2021). Deepmaps: Single-cell biological network inference using heterogeneous graph transformer. *BioRxiv*, 2021-10.
20. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.
21. Xu, J., Zhang, A., Liu, F., & Zhang, X. (2023). STGRNS: an interpretable transformer-based method for inferring gene regulatory networks from single-cell transcriptomic data. *Bioinformatics*, *39*(4), btad165.
22. Wang, Y., Chen, X., Zheng, Z., Huang, L., Xie, W., Wang, F., ... & Wong, K. C. (2024). scGREAT: transformer-based deep-language model for gene regulatory network inference from single-cell transcriptomics. *Iscience*, *27*(4).
23. Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., ... & Sun, M. (2020). Graph neural networks: A review of methods and applications. *AI open*, *1*, 57-81.
24. Kipf, T. N., & Welling, M. (2016). Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907.
25. Yuan, Y., & Bar-Joseph, Z. (2019). Deep learning for inferring gene relationships from single-cell expression data. *Proceedings of the National Academy of Sciences*, *116*(52), 27151-27158.
26. Kc, K., Li, R., Cui, F., Yu, Q., & Haake, A. R. (2019). GNE: a deep learning framework for gene network inference by aggregating biological information. *BMC systems biology*, *13*, 1-14.
27. Yuan, Y., & Bar-Joseph, Z. (2019). Deep learning for inferring gene relationships from single-cell expression data. *Proceedings of the National Academy of Sciences*, *116*(52), 27151-27158.
28. Jiang, X., & Neapolitan, R. E. (2015). LEAP: biomarker inference through learning and evaluating association patterns. *Genetic epidemiology*, *39*(3), 173-184.
29. Chan, T. E., Stumpf, M. P., & Babtie, A. C. (2017). Gene regulatory network inference from single-cell data using multivariate information measures. *Cell systems*, *5*(3), 251-267.
30. Kim, S. (2015). ppcor: an R package for a fast calculation to semi-partial correlation coefficients. *Communications for statistical applications and methods*, *22*(6), 665.
31. Papili Gao, N., Ud-Dean, S. M., Gandrillon, O., & Gunawan, R. (2018). SINCERITIES: inferring gene regulatory networks from time-stamped single cell transcriptional expression profiles. *Bioinformatics*, *34*(2), 258-266.
32. Wu, F., Souza, A., Zhang, T., Fifty, C., Yu, T., & Weinberger, K. (2019, May). Simplifying graph convolutional networks. In *International conference on machine learning* (pp. 6861-6871). Pmlr.
33. Marbach, D., Costello, J. C., Küffner, R., Vega, N. M., Prill, R. J., Camacho, D. M., ... & Stolovitzky, G. (2012). Wisdom of crowds for robust gene network inference. *Nature methods*, *9*(8), 796-804.
34. Vazin, T., & Freed, W. J. (2010). Human embryonic stem cells: derivation, culture, and differentiation: a review. *Restorative neurology and neuroscience*, *28*(4), 589-603.
35. Gepstein, L. (2002). Derivation and potential applications of human embryonic stem cells. *Circulation research*, *91*(10), 866-876.
36. Blum, B., & Benvenisty, N. (2008). The tumorigenicity of human embryonic stem cells. *Advances in cancer research*, *100*, 133-158.
37. Yeo, J. C., & Ng, H. H. (2013). The transcriptional regulation of pluripotency. *Cell research*, *23*(1), 20-32.

38. Nishimura, K., Noda, T., & Dabdoub, A. (2017). Dynamic expression of Sox2, Gata3, and Prox1 during primary auditory neuron development in the mammalian cochlea. *PloS one*, *12*(1), e0170568.
39. Sarkar, A., & Hochedlinger, K. (2013). The sox family of transcription factors: versatile regulators of stem and progenitor cell fate. *Cell stem cell*, *12*(1), 15-30.
40. Donner, A. L., Episkopou, V., & Maas, R. L. (2007). Sox2 and Pou2f1 interact to control lens and olfactory placode development. *Developmental biology*, *303*(2), 784-799.
41. Ahmadi, S. E., Rahimi, S., Zarandi, B., Chegeni, R., & Safa, M. (2021). MYC: a multipurpose oncogene with prognostic and therapeutic implications in blood malignancies. *Journal of hematology & oncology*, *14*, 1-49.
42. Stine, Z. E., Walton, Z. E., Altman, B. J., Hsieh, A. L., & Dang, C. V. (2015). MYC, metabolism, and cancer. *Cancer discovery*, *5*(10), 1024-1039.
43. Gupta, R., Ezashi, T., & Roberts, R. M. (2012). Squelching of ETS2 transactivation by POU5F1 silences the human chorionic gonadotropin CGA subunit gene in human choriocarcinoma and embryonic stem cells. *Molecular endocrinology*, *26*(5), 859-872.
44. Fang, F., Angulo, B., Xia, N., Sukhwani, M., Wang, Z., Carey, C. C., ... & Reijo Pera, R. A. (2018). A PAX5-OCT4-PRDM1 developmental switch specifies human primordial germ cells. *Nature cell biology*, *20*(6), 655-665.
45. Ermakova, V. V., Fokin, N. P., Aksenov, N. D., Bakhmet, E. I., Aleksandrova, E. V., Kuzmin, A. A., & Tomilin, A. N. (2023). Regulatory Elements Outside Established Pou5f1 Gene Boundaries Are Required for Multilineage Differentiation of Embryonic Stem Cells. *International Journal of Molecular Sciences*, *24*(20), 15434.
46. Papanayotou, C., Benhaddou, A., Camus, A., Perea-Gomez, A., Jouneau, A., Mezger, V., ... & Collignon, J. (2014). A novel nodal enhancer dependent on pluripotency factors and smad2/3 signaling conditions a regulatory switch during epiblast maturation. *PLoS biology*, *12*(6), e1001890.
47. Duggal, G., Warrier, S., Ghimire, S., Broekaert, D., Van der Jeught, M., Lierman, S., ... & Heindryckx, B. (2015). Alternative routes to induce naive pluripotency in human embryonic stem cells. *Stem cells*, *33*(9), 2686-2698.
48. Mulas, C., Kalkan, T., & Smith, A. (2017). NODAL secures pluripotency upon embryonic stem cell progression from the ground state. *Stem cell reports*, *9*(1), 77-91.
49. Herreros-Villanueva, M., Zhang, J. S., Koenig, A., Abel, E. V., Smyrk, T. C., Bamlet, W. R., ... & Billadeau, D. D. (2013). SOX2 promotes dedifferentiation and imparts stem cell-like features to pancreatic cancer cells. *Oncogenesis*, *2*(8), e61-e61.
50. Morrison, G., Scognamiglio, R., Trumpp, A., & Smith, A. (2016). Convergence of cMyc and β-catenin on Tcf7l1 enables endoderm specification. *The EMBO journal*, *35*(3), 356-368.
51. Baulida, J., Díaz, V. M., & García de Herreros, A. (2019). Snail1: a transcriptional factor controlled at multiple levels. *Journal of Clinical Medicine*, *8*(6), 757.
52. Zhu, Y., Sun, Z., Han, Q., Liao, L., Wang, J., Bian, C., ... & Zhao, R. C. (2009). Human mesenchymal stem cells inhibit cancer cell proliferation by secreting DKK-1. *Leukemia*, *23*(5), 925-933.
53. Cirio, M. C., Hui, Z., Haldin, C. E., Cosentino, C. C., Stuckenholz, C., Chen, X., ... & Hukriede, N. A. (2011). Lhx1 is required for specification of the renal progenitor cell field. *PloS one*, *6*(4), e18858.
54. Lynch, A. T., Phillips, N., Douglas, M., Dorgnach, M., Lin, I. H., Adamson, A. D., ... & Birket, M. J. (2025). HAND1 level controls the specification of multipotent cardiac and extraembryonic progenitors from human pluripotent stem cells. *The EMBO Journal*, *44*(9), 2541-2565.

55. Eyres, M., Lanfredini, S., Xu, H., Burns, A., Blake, A., Willenbrock, F., ... & O'Neill, E. (2021). TET2 drives 5hmc marking of GATA6 and epigenetically defines pancreatic ductal adenocarcinoma transcriptional subtypes. *Gastroenterology*, *161*(2), 653-668.
56. Raudvere, U., Kolberg, L., Kuzmin, I., Arak, T., Adler, P., Peterson, H., & Vilo, J. (2019). g: Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic acids research*, *47*(W1), W191-W198.
57. Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., ... & Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, *13*(11), 2498-2504.
58. Attisano, L., & Wrana, J. L. (2013). Signal integration in TGF-β, WNT, and Hippo pathways. *F1000prime reports*, *5*, 17.
59. Li, N., Lu, N., & Xie, C. (2019). The Hippo and Wnt signalling pathways: crosstalk during neoplastic progression in gastrointestinal tissue. *The FEBS Journal*, *286*(19), 3745-3756.
60. Gao, W., Li, R., Ye, M., Zhang, L., Zheng, J., Yang, Y., ... & Zhao, Q. (2022). The circadian clock has roles in mesenchymal stem cell fate decision. *Stem Cell Research & Therapy*, *13*(1), 200.
61. Burgess, R. J., Agathocleous, M., & Morrison, S. J. (2014). Metabolic regulation of stem cell function. *Journal of internal medicine*, *276*(1), 12-24.
62. Martinez-Sarmiento, J. A., Cosma, M. P., & Lakadamyali, M. (2024). Dissecting gene activation and chromatin remodeling dynamics in single human cells undergoing reprogramming. *Cell reports*, *43*(5).
63. Muncie, J. M., Ayad, N. M., Lakins, J. N., Xue, X., Fu, J., & Weaver, V. M. (2020). Mechanical tension promotes formation of gastrulation-like nodes and patterns mesoderm specification in human embryonic stem cells. *Developmental cell*, *55*(6), 679-694.
64. Lawson, A., & Schoenwolf, G. C. (2003). Epiblast and primitive-streak origins of the endoderm in the gastrulating chick embryo.
65. Britton, G., Heemskerk, I., Hodge, R., Qutub, A. A., & Warmflash, A. (2019). A novel self-organizing embryonic stem cell system reveals signaling logic underlying the patterning of human ectoderm. *Development*, *146*(20).
66. Yiangou, L., Grandy, R. A., Osnato, A., Ortmann, D., Sinha, S., & Vallier, L. (2019). Cell cycle regulators control mesoderm specification in human pluripotent stem cells. *Journal of Biological Chemistry*, *294*(47), 17903-17914.
67. Muhr, J., & Ackerman, K. M. (2023). Embryology, gastrulation. In *StatPearls [Internet]*. StatPearls Publishing.
68. Li, Y., Du, J., Deng, S., Liu, B., Jing, X., Yan, Y., ... & She, Q. (2024). The molecular mechanisms of cardiac development and related diseases. *Signal Transduction and Targeted Therapy*, *9*(1), 368.
69. Mensah, I. K., & Gowher, H. (2024). Signaling Pathways Governing Cardiomyocyte Differentiation. *Genes*, *15*(6), 798.
70. Yao, Z., Mich, J. K., Ku, S., Menon, V., Krostag, A. R., Martinez, R. A., ... & Ramanathan, S. (2017). A single-cell roadmap of lineage bifurcation in human ESC models of embryonic brain development. *Cell stem cell*, *20*(1), 120-134.
71. PBMCs from healthy donor (v3, 3k cells), Single Cell Immune Profiling Dataset by Cell Ranger v3.0.0, 10x Genomics, (2018, February 8).
72. Wolf, F. A., Angerer, P., & Theis, F. J. (2018). SCANPY: large-scale single-cell gene expression data analysis. *Genome biology*, *19*, 1-5.
73. Stamenkovic, I. V. A. N., & Seed, B. R. I. A. N. (1988). Analysis of two cDNA clones encoding the B lymphocyte antigen CD20 (B1, Bp35), a type III integral membrane protein. *The Journal of experimental medicine*, *167*(6), 1975-1980.

74. Krensky, A. M., & Clayberger, C. (2005). Granulysin: a novel host defense molecule. *American Journal of Transplantation*, *5*(8), 1789-1792.
75. Fang, S., & Zhou, Y. (2024). Deciphering the role of KLRB1: a novel prognostic indicator in hepatocellular carcinoma. *BMC gastroenterology*, *24*(1), 210.
76. Rollinger-Holzinger, I., Eibl, B., Pauly, M., Griesser, U., Hentges, F., Auer, B., ... & Zwierzina, H. (2000). LST1: a gene with extensive alternative splicing and immunomodulatory function. *The Journal of Immunology*, *164*(6), 3169-3176.
77. Reshetnikova, E., Guselnikov, S., Volkova, O., Baranov, K., Taranin, A., & Mechetina, L. (2018). B cell-specific protein FCRLA is expressed by plasmacytoid dendritic cells in humans. *Cytometry Part B: Clinical Cytometry*, *94*(4), 683-687.
78. Real, P. J., Ligero, G., Ayllon, V., Ramos-Mejia, V., Bueno, C., Gutierrez-Aranda, I., ... & Menendez, P. (2012). SCL/TAL1 regulates hematopoietic specification from human embryonic stem cells. *Molecular Therapy*, *20*(7), 1443-1453.
79. Onnis, A., Finetti, F., & Baldari, C. T. (2016). Vesicular trafficking to the immune synapse: how to assemble receptor-tailored pathways from a basic building set. *Frontiers in immunology*, *7*, 50.
80. Xu, G., & Shi, Y. (2007). Apoptosis signaling pathways and lymphocyte homeostasis. *Cell research*, *17*(9), 759-771.
81. Liu, J., Guo, Q., Liu, G., Wang, W., Jin, X., Hao, B., & Lei, B. (2025). Immune pathogenic response landscape of acute posterior multifocal placoid pigment epitheliopathy revealed by scRNA sequencing. *Genes & Immunity*, 1-16.
82. Marongiu, L., Valache, M., Facchini, F. A., & Granucci, F. (2021). How dendritic cells sense and respond to viral infections. *Clinical Science*, *135*(19), 2217-2242.
83. Marongiu, L., Valache, M., Facchini, F. A., & Granucci, F. (2021). How dendritic cells sense and respond to viral infections. *Clinical Science*, *135*(19), 2217-2242.
84. Dunkelberger, J. R., & Song, W. C. (2010). Complement and its role in innate and adaptive immune responses. *Cell research*, *20*(1), 34-50.
85. Spivak, G. (2015). Nucleotide excision repair in humans. *DNA repair*, *36*, 13-18.
86. Rosenberg, J., & Huang, J. (2018). CD8+ T cells and NK cells: parallel and complementary soldiers of immunotherapy. *Current opinion in chemical engineering*, *19*, 9-20.
87. Prager, I., & Watzl, C. (2019). Mechanisms of natural killer cell-mediated cellular cytotoxicity. *Journal of leukocyte biology*, *105*(6), 1319-1329.
88. Uzhachenko, R. V., & Shanker, A. (2019). CD8+ T lymphocyte and NK cell network: circuitry in the cytotoxic domain of immunity. *Frontiers in immunology*, *10*, 1906.
89. Sanda, T., Lawton, L. N., Barrasa, M. I., Fan, Z. P., Kohlhammer, H., Gutierrez, A., ... & Look, A. T. (2012). Core transcriptional regulatory circuit controlled by the TAL1 complex in human T cell acute lymphoblastic leukemia. *Cancer cell*, *22*(2), 209-221.
90. Johnson, K. A., & Krishnan, A. (2022). Robust normalization and transformation techniques for constructing gene coexpression networks from RNA-seq data. *Genome biology*, *23*, 1-26.
91. Shi, Y., Huang, Z., Feng, S., Zhong, H., Wang, W., & Sun, Y. (2020). Masked label prediction: Unified message passing model for semi-supervised classification. *arXiv preprint arXiv:2009.03509*.
92. Xu, J., Li, Z., Du, B., Zhang, M., & Liu, J. (2020, July). Reluplex made more practical: Leaky ReLU. In *2020 IEEE Symposium on Computers and communications (ISCC)* (pp. 1-7). IEEE
93. Wang, K., Li, Y., Liu, F., Luan, X., Wang, X., & Zhou, J. (2025). GRLGRN: graph representation-based learning to infer gene regulatory networks from single-cell RNA-seq data. *BMC bioinformatics*, *26*(1), 108.

94. Huynh-Thu, V. A., Irrthum, A., Wehenkel, L., & Geurts, P. (2010). Inferring regulatory networks from expression data using tree-based methods. *PloS one*, *5*(9), e12776.
95. Sanchez-Castillo, M., Blanco, D., Tienda-Luna, I. M., Carrion, M. C., & Huang, Y. (2018). A Bayesian framework for the inference of gene regulatory networks from time and pseudo-time series data. *Bioinformatics*, *34*(6), 964-970.
96. Szklarczyk, D., Kirsch, R., Koutrouli, M., Nastou, K., Mehryary, F., Hachilif, R., ... & Von Mering, C. (2023). The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic acids research*, *51*(D1), D638-D646.