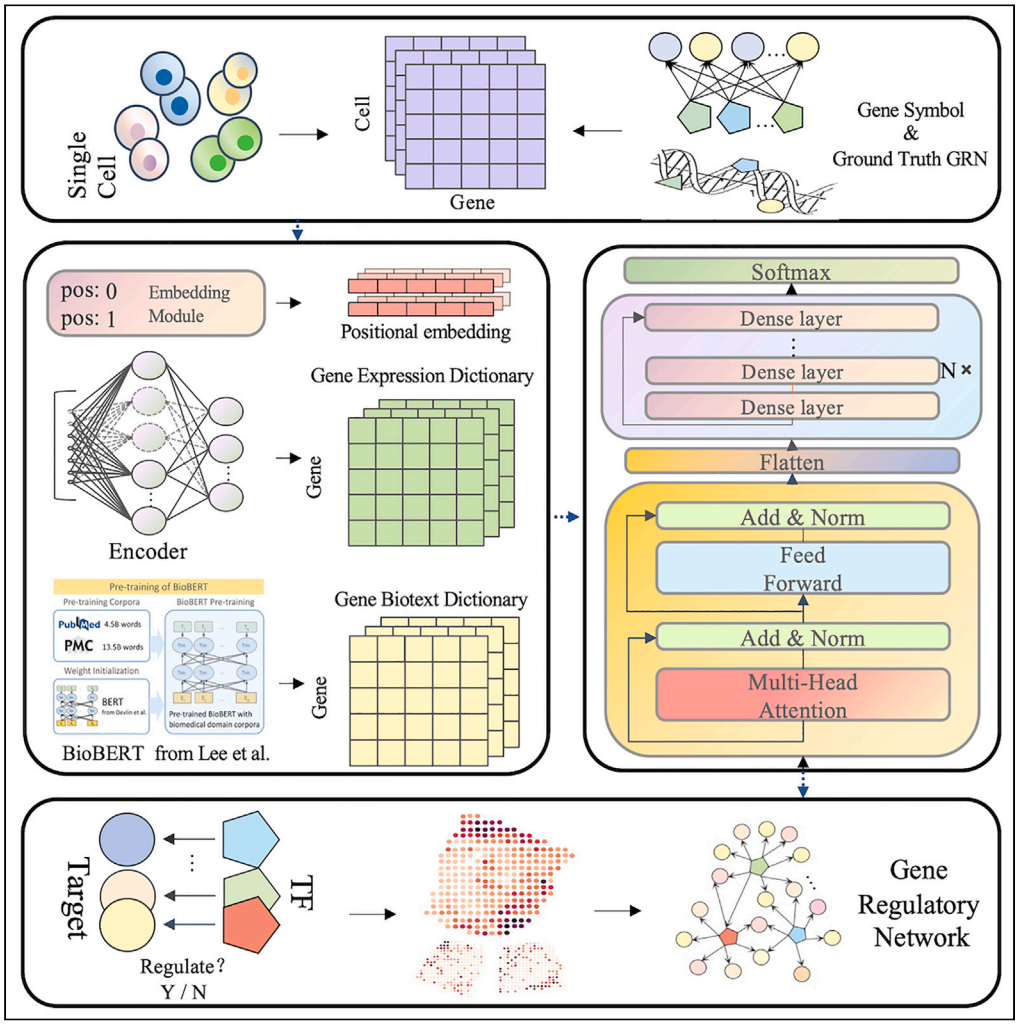
iScience

CellPress
OPEN ACCESS

Article

scGREAT: Transformer-based deep-language model for gene regulatory network inference from single-cell transcriptomics



Yuchen Wang, Xingjian Chen, Zetian Zheng, ..., Fuzhou Wang, Zhaolei Zhang, Ka-Chun Wong

kc.w@cityu.edu.hk

Highlights

Taking advantage of transformer backbone and biomedical language model

Outperforming SOTA models on 7 benchmark datasets 4 kinds of gene network platforms

Employing spatial transcriptomics data as external validation

Ability to uncover novel relationships between genes

Wang et al., iScience 27, 109352
April 19, 2024
https://doi.org/10.1016/j.isci.2024.109352

iScience

CellPress
OPEN ACCESS

Article

scGREAT: Transformer-based deep-language model for gene regulatory network inference from single-cell transcriptomics

Yuchen Wang,[1] Xingjian Chen,[1,2] Zetian Zheng,[1] Lei Huang,[1] Weidun Xie,[1] Fuzhou Wang,[1] Zhaolei Zhang,[4,5,6] and Ka-Chun Wong[1,3,7,*]

SUMMARY

Gene regulatory networks (GRNs) involve complex and multi-layer regulatory interactions between regulators and their target genes. Precise knowledge of GRNs is important in understanding cellular processes and molecular functions. Recent breakthroughs in single-cell sequencing technology made it possible to infer GRNs at single-cell level. Existing methods, however, are limited by expensive computations, and sometimes simplistic assumptions. To overcome these obstacles, we propose scGREAT, a framework to infer GRN using gene embeddings and transformer from single-cell transcriptomics. scGREAT starts by constructing gene expression and gene biotext dictionaries from scRNA-seq data and gene text information. The representation of TF gene pairs is learned through optimizing embedding space by transformer-based engine. Results illustrated scGREAT outperformed other contemporary methods on benchmarks. Besides, gene representations from scGREAT provide valuable gene regulation insights, and external validation on spatial transcriptomics illuminated the mechanism behind scGREAT annotation. Moreover, scGREAT identified several TF target regulations corroborated in studies.

INTRODUCTION

The rapid development of single-cell RNA sequencing (scRNA-seq), along with the exponential increase in genomic data, has expanded the frontiers of single-cell research and accentuated the need for the development of computational methods to interpret gene-gene interactions and relationships.[1] Gene regulatory network (GRN) illustrates the intricate interactions among genes, consisting of regulatory relationships among a variety of molecular entities.[2] Accurate reconstruction of GRN is essential for understanding the behavior of different genes,[3,4] such as gene expression mechanisms within cells, and advancing research in disease pathology.[5] Single-cell technology has brought opportunities for GRN inference but also unprecedented challenges, especially complexity and inherent noise in scRNA-seq data post unique challenges.[1] Fortunately, deep learning-based methods offer robust solutions for handling noisy data, integrating diverse knowledge sources, and learning complex relationships by its capabilities of feature extraction and optimization as exemplified in.[6–13]

Recently, numerous methods have been proposed to infer GRN. For instance, SCODE,[14] using ordinary differential equations (ODE), treats pseudotime as time information to reconstruct GRN during cell differentiation. GENIE3[15] and GRNBoost2[16] are both tree-based machine learning algorithms for inferring GRN, which was incorporated into the program of SCENIC.[17,18] Tree rules are utilized to learn regulatory relationships by leaving one gene out at a time to find its relationships with other genes. Boosting method is an approach to enhance the performance of trees.[16] Despite its success, the primary limitation of SCODE is its dependency on accurate pseudotime data and potential oversimplification of complex pathological processes using linear ordinary differential equations. Besides, the tree-based methods require segmenting input data for iteratively establishing multiple models, which is computationally expensive and less scalable for large datasets. Various deep learning methods have been described to overcome these limitations.[7] Shu et al. proposed DeepSEM,[6] a structural equation model (SEM) with a beta-variational autoencoder and neural network to predict regulatory relationships. However, the prior domain knowledge required and SEM assumptions about the underlying causal structure may not always hold in practice.[7] GNE[8] is a deep learning method based on multilayer perceptron (MLP) for GRN inference applied to microarray data. It utilized one-hot gene ID vectors from the gene topology to capture topological information, which is always inefficient due to the highly sparse nature of the resulting one-hot feature vector.[19] In

[1]Department of Computer Science, City University of Hong Kong, Kowloon Tong, Hong Kong SAR
[2]Cutaneous Biology Research Center, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA
[3]Shenzhen Research Institute, City University of Hong Kong, Shenzhen, China
[4]Department of Molecular Genetics, University of Toronto, Toronto, ON, Canada
[5]Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, ON, Canada
[6]Department of Computer Science, University of Toronto, Toronto, ON, Canada
[7]Lead contact
*Correspondence: kc.w@cityu.edu.hk
https://doi.org/10.1016/j.isci.2024.109352

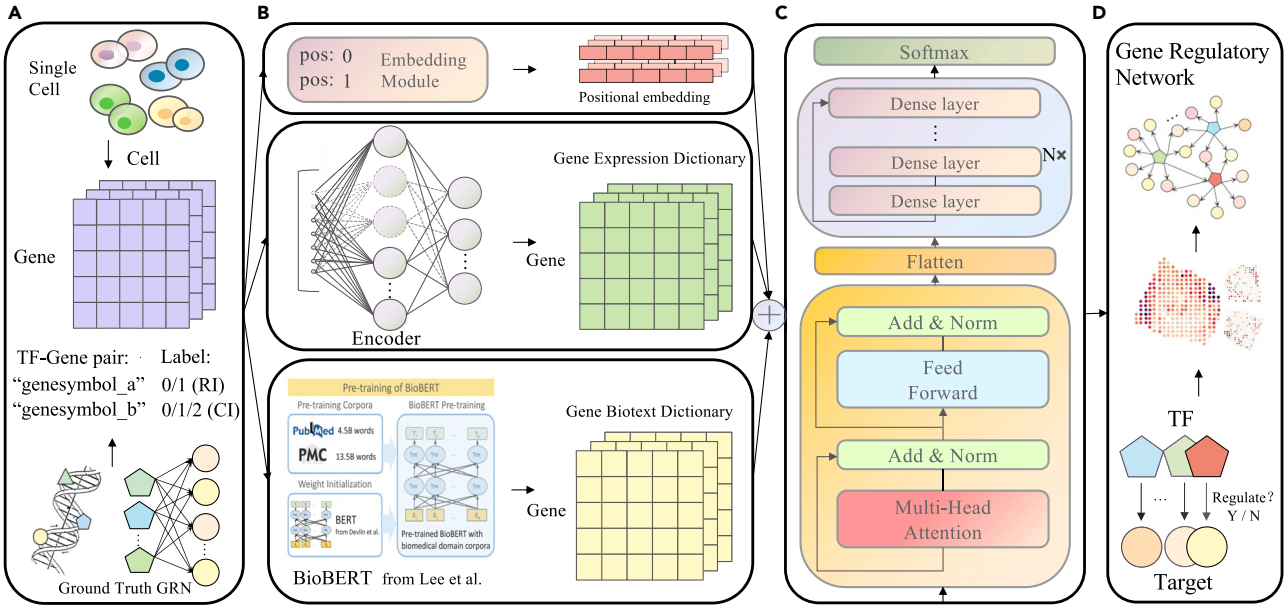


Figure 1. Overview of scGREAT framework

(A) Feature initialization.

(B) Gene dictionary construction.

(C) Inference engine.

(D) GRN construction. * RI, Regulatory Inference; CI, Causality Inference.

addition, graph models have gained attention in capturing the topology of gene networks. Chen and Liu introduced GENELink,[20] a deep learning framework with graph attention networks (GATs). However, the method relies on node feature quality and emphasizes local network information over global perspectives, which could lead to suboptimal or inadequate node feature representations, especially given the sparsity of known ground truth networks. Besides, convolutional neural network (CNN) has had great achievement in many computer vision tasks[21] and their applications in bioinformatics have also been quite successful.[22,23] Yuan and Bar-Joseph proposed convolutional neural network for coexpression (CNNC),[24] a deep learning method converting gene expression co-occurrence values into pixel values in images through normalized empirical probability distribution function and using CNN to extract the relationship. However, the method requires substantial computation to transform the transcriptomic data into images for every TF-gene pair, which can be computationally expensive for larger datasets and lacks the advantages of end-to-end deep learning models.[25] To address the above problems, we propose scGREAT, a transformer-based supervised deep learning model for GRN inference from single-cell transcriptomics data. Inspired by the application of computer vision techniques in CNNC,[24] we leverage state-of-the-art (SOTA) transformer model in natural language processing (NLP) to efficiently process and interpret complex scRNA-seq data and capture dependencies between genes. Besides, we utilized language models to construct a gene biotext dictionary and an expression dictionary for unraveling the complex regulatory relationships between Transcription Factors (TF) and genes. Our contributions can be summarized as follows.

(1) Drawing an analogy between genes and cells to words and sentences, taking advantage of the SOTA model backbone, transformer, in NLP field.
(2) Applying the biomedical language representation model abstracts contextual biotext vectors for gene symbols.
(3) Significantly outperforms other SOTA models, achieving up to 91.30% average AUROC on seven benchmark datasets across four kinds of gene network platforms.
(4) Employing spatial transcriptomics data as external validation to visualization and illuminate the mechanism behind model annotation for gene regulation.
(5) Ability to uncover novel relationships between genes. scGREAT discovered various unreported TF-target regulatory relationships corroborated in other published studies.

RESULTS

Design principles of scGREAT framework

We designed scGREAT as an end-to-end deep learning model based on gene embedding and transformer backbone[26] for inferring gene regulatory relationships. In detail, the scGREAT consists of four major components: feature initialization, gene dictionary construction, inference engine that is scGREAT network backbone, and GRN construction (see Figure 1). For feature initialization, after the train-test split to

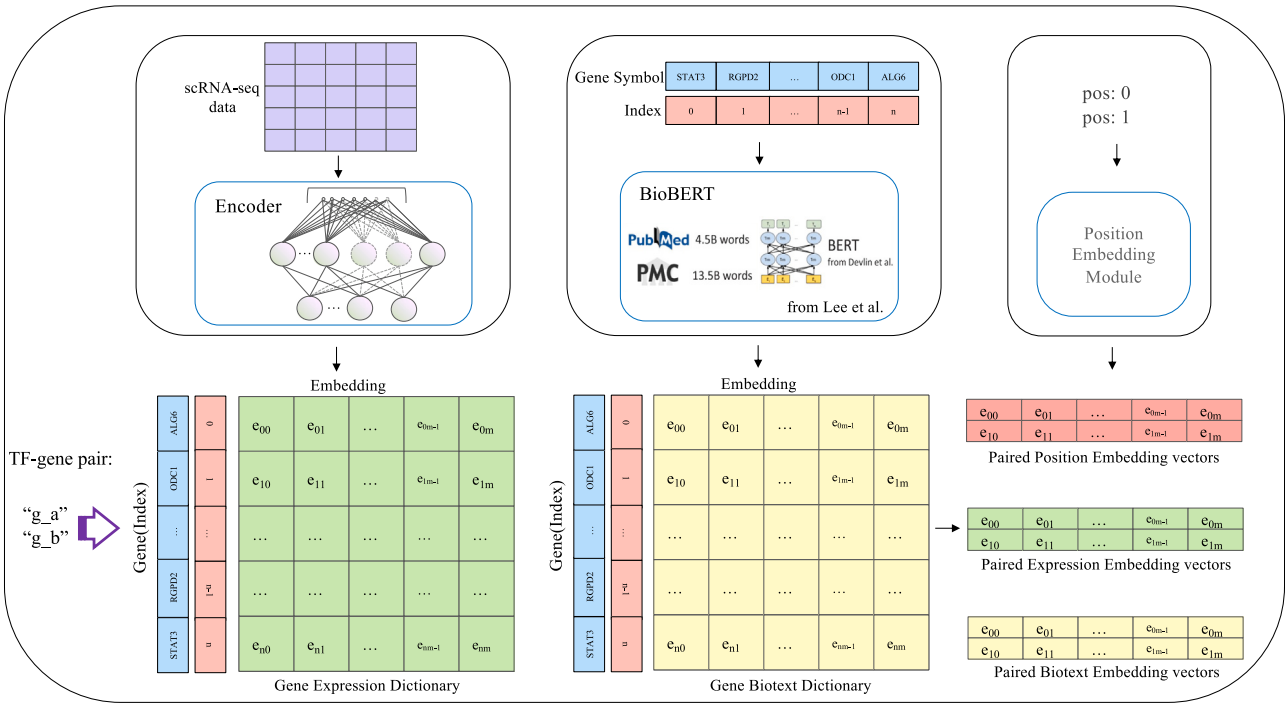


Figure 2. Dictionary Construction and Application

The *n* in e_{nm} represents the *n-th* gene symbol and *m* is the *m-th* dimension of embedding. The purple arrow on the left refers to the search process and the two embedding vectors on the right are the results of searching in the two dictionaries. The position embedding vector is the generated vector by position module.

avoid data leakage, we preprocessed the gene expression data and subsequently extracted gene pairs from the ground truth network along with generating negative samples. In the second component, gene expression dictionaries and gene biotext dictionaries are learned from the original gene expression data, gene symbols. We first adopted a neural network encoder to construct gene expression data into a gene expression dictionary (detailed in the STAR methods section). In parallel, we also extracted textual information of gene symbols from BioBERT[27] and combine the regulatory position order of TF and target to construct the principal feature representation (see Figure 2). In the scGREAT network backbone part, the transformer architecture was designed as the backbone framework and heavily modified to ensure meaningful feature representations for predicting regulatory relationships. By employing the backbone of the network, scGREAT analyzes single-cell data to uncover potential regulatory relationships among genes, predict associations between gene pairs, assess the loss between these predictions and actual observations, and iteratively adjust network parameters until convergence is achieved. For the final part, GRNs are constructed according to predicted regulatory relationships of candidate gene pairs and further validated through spatial transcriptomics information.

Performance on benchmark datasets

We evaluated the effectiveness and generalization ability of scGREAT on seven benchmark datasets (detailed in the STAR Methods section), including hESC, hHEP, mDC, mESC, mHSC-E, mHSC-GM, and mHSC-L across all four kinds of ground truth networks from STRING, non-specific ChIP-seq, LOF/GOF, and Cell-type-specific ChIP-seq (see Tables 1 and S1). In detail, the performance of scGREAT is compared with several state-of-the-art methods, including GENELink, GNE, CNNC, DeepSEM, PCC, MI, SCODE, GRNBoost2, and GENIE3[6,8,14–18,20,24] as well as two popular methods for gene–gene coexpression analysis Pearson correlation coefficient (PCC) and Mutual Information (MI).[24] Among these methods, GENIE3, GRNBoost2, SCODE, MI, PCC, and DeepSEM utilize unsupervised or self-supervised methods, whereas CNNC, GNE, GENELink, and scGREAT are based on supervised methods. In supervised learning, models are trained with labeled data, whereas unsupervised learning operates without. In order to guarantee fairness in evaluations, we maintained training and independent testing data consistency for all the supervised methods by following the data processing rules of BEELINE,[28] as well as the HNS selection method and train test data splitting method according to.[20]

The results showed that, on the Cell-type-specific ChIP-seq network type, scGREAT ultimately achieved the best performance on all (14/14) scRNA-seq datasets in terms of both the AUROC (see Figure 3 bottom) and the AUPRC (see Figure S1) evaluation metric, achieving average AUROC score of 90.5% (81.4%–95.0%) as increased by 6.3%, 15.5%, and 23.9% compared to the latest and famous methods: GENELink, GNE, and CNNC, respectively.

scGREAT also outperformed across the STRING, non-specific ChIP-seq, and LOF/GOF network types, which with extremely sparse connection density. The CNNC method necessitates the transformation of scRNA-seq data into image data, demanding significant

Table 1. The size of training sets about each ground-truth network with TFs and most-varying 500 (1000) genes

Datasets	Specific	STRING	Non-specific	LOF/GOF
hESC	20677 (32065)	208614 (331058)	172153 (275435)	–
hHEP	19002 (30026)	259147 (401011)	204039 (321581)	–
mDC	10969 (18556)	144820 (241221)	137156 (224444)	–
mESC	65895 (96460)	370740 (540893)	386511 (565848)	25459 (36848)
mHSC-E	13632 (26565)	73346 (129635)	67712 (118364)	–
mHSC-GM	9280 (17406)	38827 (75698)	34615 (66621)	–
mHSC-L	5976 (7392)	14573 (18487)	13081 (17096)	–

Specific represents Cell-type-specific.

computational resources on the extensive training data size in these three ground truth networks. Besides, it is based on the deep CNN network framework VGGnet, which, given the substantial processing power requirements, faced limitations in training on these three ground truth networks with one GPU. For the STRING network, scGREAT achieved an average AUROC score of 94.3% (85.2%–97.4%), for the non-specific ChIP-seq network, the average AUROC performance of scGREAT is 89.4% (81.1%–93.0%) and for LOF/GOF network, the average AUROC is 91.0% (see Figure 3 top We performed a detailed comparison with two-tailed paired t-tests and Mann-Whitney U test on the AUROC performance of scGREAT and the second-best method GENELink to clearly illustrate that the performance of scGREAT significantly outperforms and becomes the state-of-the-art method, with the t-tests p values of 0.0206, 0.0087, and 0.0023 and U tests p values of 0.001, 0.0023 and 0.0045 on seven datasets based on STRING, non-Specific and cell-type-specific ChIP-seq network (see Figures 4 and S2), respectively.

Furthermore, we evaluate the AUROC performance of these representative algorithm backbones in gene regulation prediction tasks. These backbones are commonly used in statistics and computer science research. The results are presented in Table 2. In general, supervised methods have significantly improved performance compared to unsupervised methods. Among supervised methods, scGREAT with transformer in the NLP field performs best, achieving an average AUROC value of 91.3% in seven datasets, and GENELink using graph performed second, reaching an average AUROC value of 86.7%. However, the result is not intended to judge the merits of methods in different fields. Rather, it provides some implications for model selection when solving gene regulation prediction problems. The performance of these algorithms is also affected by the setting of model parameters, feature extraction strategies, and supervision methods. Figure 5 displays the average AUROC with line chart accompanied by error bars. This visualization effectively illustrates the distinctions among various models, highlighting scGREAT's competitive performance.

Additionally, considering the sparsity and low density of gene regulatory relationships, we emphasized maintaining the predictive capability of scGREAT within the actual data distribution instead of strictly balancing positive and negative sample data. Interestingly, the results reflect that scGREAT also achieved outstanding performance in AUPRC (see Figures 6 and S1), which is the crucial indicator for imbalanced data. Take the Specific network as an example, scGREAT achieved an average AUPRC score of 76.65% with an average improvement of 34.1% (9.6%–42.0%) across seven datasets, compared to the other methods.

In general, scGREAT achieved the best AUROC performance on all (44/44) benchmark scRNA-seq datasets and the best AUPRC performance on 93% (42/44) benchmark datasets. The two exceptional cases can primarily be attributed to the extreme sparsity of the data, as the network density is merely 0.048 (0.045) with only 137 (154) positive samples in the case of an overall sample size of 14573 (18487), so for the AUPRC metric, scGREAT performance did not meet expectations. In a comparison against the second-best model GENELink, scGREAT demonstrated superior performance across all benchmark datasets in terms of both the AUROC and the AUPRC evaluation metric.

To illustrate the performance of finding true positives and avoiding true negatives, in Figure 7, we compared the positive sample prediction results of scGREAT and the second-best model GENELink on seven datasets with Cell-type-specific ChIP-seq network. The Venn diagrams reflect the superiority of scGREAT in predicting positive samples. To ensure fairness in comparison, we classify positive and negative samples according to the threshold under the optimal AUROC indicator for each model. On this basis, we obtained the final prediction labels of the compared models and labels of scGREAT. By drawing the Venn diagrams of the positive regulatory relationship predicted by scGREAT and GENELink, and the ground truth, we visually demonstrated the overlaps between the predicted and actual labels of the two models.

Taking the mESC dataset with 500 most-varying genes as an example (see Figure 7 middle of the first line), 5181 (75.6%) regulations were found by both methods. scGREAT failed to detect 1051 (15.3%) positive regulations while GENELink overlooked 1298 (18.9%) instances, which shows that scGREAT had a lower missed detection number of positive samples by 247 (3.6%) compared to the second-best model. For false positives samples, there were 2090 spotted by scGREAT while 2981 from GENELink, indicating that scGREAT reduced 891 (30.0%) false positives samples in comparison to GENELink. Unfortunately, on the mDC dataset, scGREAT did not demonstrate strong performance in terms of low false positives, which may be attributed to the characteristics of the dataset that various models showed suboptimal performance on this dataset, indicating significant noise and confounding information in the networks of specific cell types. However, despite this, scGREAT detected more positive samples compared to GENELink. Generally, scGREAT outperformed the second-best method by improving the identification rate of potential regulatory relationships by 2.3%–38.35% on 93% (13/14) of the datasets. It shows that scGREAT can effectively

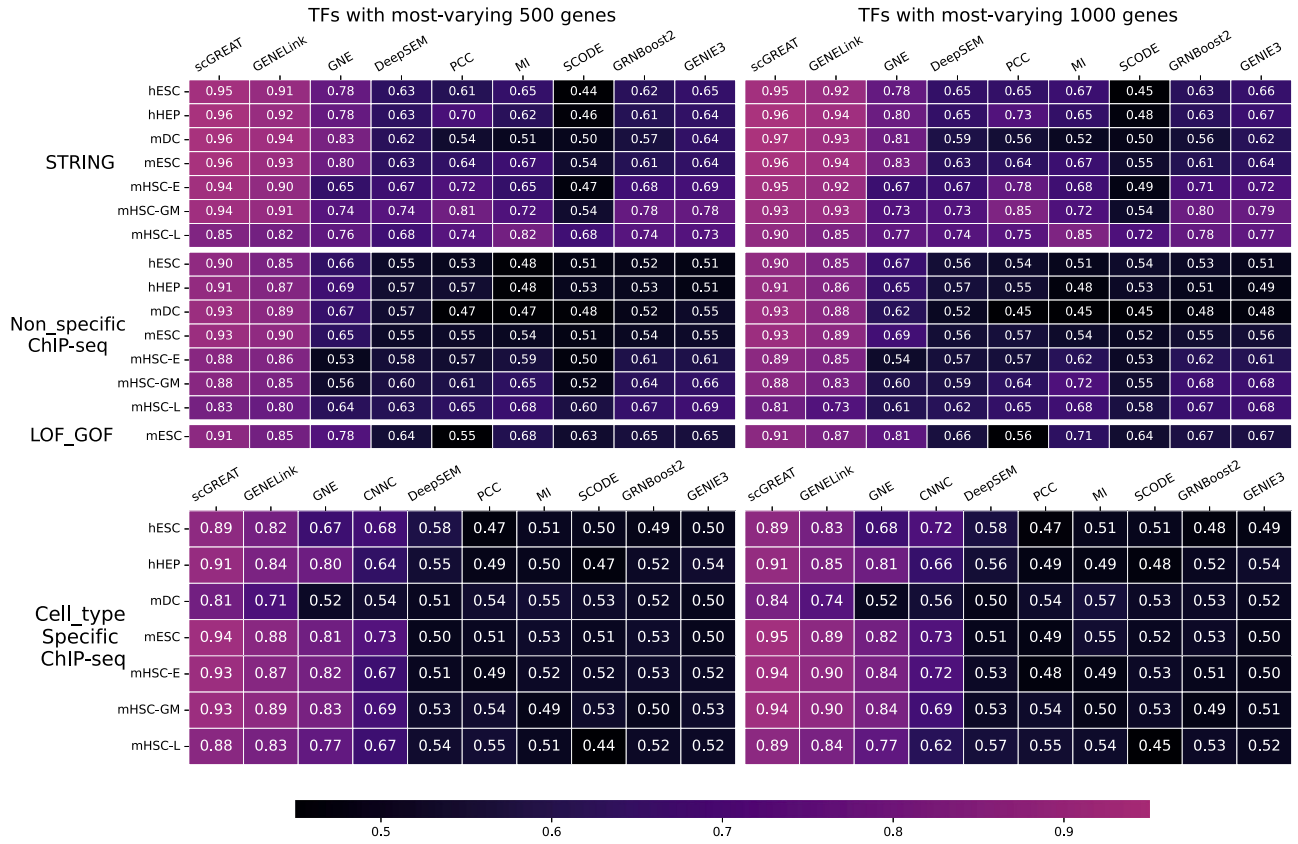


Figure 3. AUROC Heatmap

AUROC of scGREAT and other state-of-the-art methods on benchmark datasets across four types of networks.

identify potential regulatory relationships with high sensitivities. For false positives performance, scGREAT outperformed GENELink by reducing the rate of false positives by 3.7%–29.5% on 10/14 of the datasets.

We also evaluated the causality inference[20,24,29] ability of scGREAT on hESC and mESC datasets with cell-type-specific networks (see Figure 8). Comparing the results with the second-best model GENELink, for the causality inference task, scGREAT achieves an average of 9% and 2.9% improvements in AUROC, and 11.0% and 5.5% improvements in AUPRC.

Embedding clustering

To gain an in-depth understanding of gene regulatory information within scGREAT, we utilized the t-distributed Stochastic Neighbor Embedding (t-SNE) technique[30] to visualize both the original paired gene data and the embeddings after penultimate layer of scGREAT (see Figure 9 Left column). Original data, organized in a three-dimensional array with dimensions representing samples, gene pairs, and cells, respectively (see Figure S6), is constructed from gene expression data and network. The penultimate layer embeddings, derived from the original paired gene data through the optimization process of scGREAT, are structured similarly in terms of gene pairs but with a different third dimension 256, where 256 is the dimension of features out from the layer. These embeddings project comprehensive feature representation of the gene regulatory and expression information. In contrast to the original data, embeddings out from penultimate layer effectively replicated the class patterns that are highly discriminative for classification. This visualization demonstrated the capacity of scGREAT to capture and illustrate the distinct variations amongst classes.

Additionally, we generated projections of gene embeddings extracted from the scGREAT framework, which include the embeddings of TF, Target Genes (TGs), and Non-Target Genes (Non-TGs) (see Figure 9 Right two columns). These projections effectively portray the distinct regulatory statuses and relationships amongst these gene types. For each TF, multiple embeddings are generated as it involves various gene regulatory pairs. We computed the arithmetic mean of all generated TF embeddings to obtain a representative one, and picked its respective target and non-target gene embeddings for t-SNE clustering. The results show that the TF was centrally positioned amongst its target genes, with SIN3A TF serving as a representative example. We observed that these exhibited close associations with the TF, while the Non-TGs were located more distantly. The spatial arrangement observed suggests a correlation with the presence or absence of regulation, indicating that scGREAT may offer valuable perspectives on the dynamics of gene regulation. As an extension, we also performed gene embedding

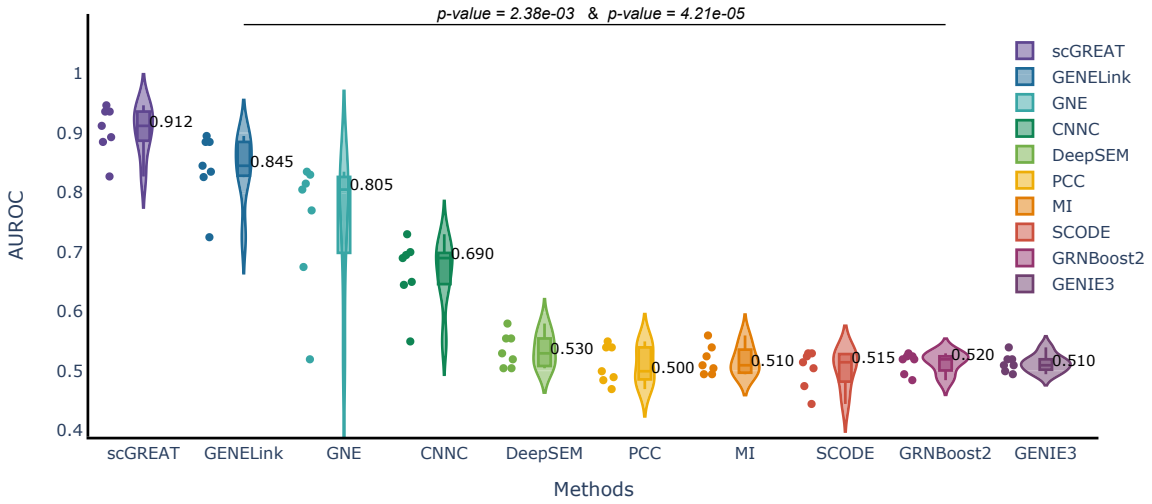


Figure 4. AUROC Violin

Violin plots on the AUROC performance of scGREAT and state-of-the-art methods across the Cell-type-specific ChIP-seq network, with p values of 2.38e-03 and 4.21e-05 compared with GENELink and GNE, respectively, from two-tailed paired t-tests and with p values of 2.38e-03 and 2.55e-05 from Mann-Whitney U test.

clustering for other TFs to validate the consistency of these spatial patterns and relationships (see Figure 9 Right two columns and Figure S5). The same tendency was observed across different TFs, further strengthening the reliability of scGREAT in deciphering gene regulatory patterns.

External validation on spatial transcriptomics data

According to[31] and,[32] spatial similarities in gene expression can effectively reflect gene regulatory relationships. Leveraging spatial transcriptomics data, we next attempted to find corroborative evidence on the TF-gene pairs in the hHEP testing dataset that were predicted by scGREAT but were not reported by other methods. This confirmation was achieved through external validation using spatial scRNA-seq of the human liver, with a focus on identifying and clustering hepatic stellate cells[33] (see Figures 10 and S3).

Specifically, ATF3 was analyzed as a representative TF, the spatial scRNA-seq data of ATF3 depicts a close alignment between the expression levels of ATF3 and its target genes. We calculated the Pearson correlation coefficient between the TF-gene pairs. The result show that the pcc between ATF3 and its target gene GTF2H3 is 0.30, while the it between ATF3 and the non-target gene IFITM3 is 0.09. Moreover, upon scrutinizing the cell-type clustering umap, a noticeable co-expression pattern was also discovered (see Figure 10 row), which suggested that ATF3 and its target genes are largely co-expressed within the same cluster of cell types. This observation reinforces scGREAT's dependability and effectiveness in identifying genuine TF-gene pairs. Contrastingly, this co-expression pattern is conspicuously missing when analyzing the expression map of ATF3's non-target genes. These non-target genes, also correctly identified by scGREAT but incorrectly predicted by other methods, exhibit a different expression level and pattern, which further emphasizes to scGREAT's robustness and precision in pinpointing valid TF-gene pairs. we (see Figure S4) also showed some regulations that are missed by scGREAT but pointed out by other methods. Taking

Table 2. Average AUROC results for the state-of-the-art models under various backbones on four ground truth networks

Method	Field	Backbone	AUROC
GENIE3	Machine Learning	Tree	0.610
GRNBoost2	Machine Learning	Boost-Tree	0.604
SCODE	Calculus	ODE	0.547
MI	Information Theory	Probability	0.612
PCC	Statistics	Correlation	0.581
DeepSEM	Data Analysis	SEM	0.605
CNNC	Computer Vision	CNN	0.666
GNE	Deep Learning	MLP	0.735
GENELink	Graph	GAT	0.867
scGREAT	NLP	Transformer	**0.913**

NLP represents Natural language processing.

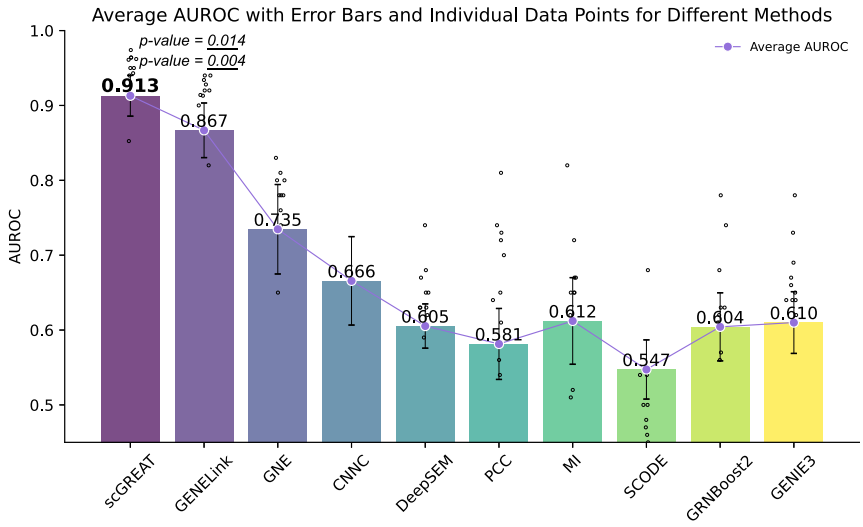


Figure 5. Average AUROC bars chart

Improvement with p value of 0.014 and 0.004 compared with GENELink and GNE, respectively.

HDAC2 as an example, a distinct difference can be observed between the distributions of missed and accurately identified samples, potentially attributed to various regulatory modes like promotion and inhibition; moreover, the direct inhibition or indirect inhibition could also lead to different expression of TG. For example, the expression status of the TF-gene pair HDAC2-MEP1A is the same as the status of the ATF3-APOC2 pair, where the two TFs were moderately expressed and TGs were densely and highly expressed, but the two pair were with completely opposite labels. This kind of situation may have had a negative impact on scGREAT.

Ablation experiments

We carried out two ablation experiments, the first involved contrasting scGREAT structure with a Multilayer Perceptron(MLP) model, while the second focused on evaluating the impact of gene biotext dictionary and gene expression dictionary on the performance of scGREAT.

Validate the role of the transformer backbone

In order to state the effectiveness of Transformer structure, we evaluated the AUROC and AUPRC performance of scGREAT with and without the Transformer architecture with the same parameter settings across the seven datasets with the Cell-type-specific ChIP-seq network types of ground truth.

In general, the average AUROC value for the scGREAT method is 90.2% (range from 81.2% to 94.4%), and the average AUPRC value is 76.7% (range from 19.5% to 95.2%). In contrast, the average AUROC value for the model without the Transformer architecture is 82.9% (range from 76.8% to 87.7%), and the average AUPRC value is 64.0% (range from 12.3% to 87.0%). The average AUROC difference in performance due to the Transformer architecture shows an improvement with 7.2%, which is varied from 4.4% to 11.4%. Similarly, the average AUPRC difference is 12.6%, with a range from 7.2% to 23.0%. Overall, the result indicates that the Transformer architecture consistently improves the performance in terms of both AUROC and AUPRC across all datasets, with notable improvements observed in both metrics. These discrepancies in performance justify the structural and computational characteristics and rationale behind scGREAT (see Table 3).

Validate the role of the biotext vectors

To assess the impact of incorporating a gene biotext dictionary, we compared the AUROC and AUPRC metrics with and without these vectors across seven datasets. As shown in Table 4, the average AUROC for the scGREAT with biotext vectors is 90.2%, and the average AUPRC is 76.7%. In contrast, without biotext vectors, these values drop to 89.4% and 75.0%, respectively. The differences in average AUROC and AUPRC are 0.75% and 1.68%, differs among the various datasets, ranging from 0% to 1.49% for AUROC and 0.24%–3.79% for AUPRC, respectively. Experiments confirm that biotext dictionary is of limitation to the model in some cases. Besides, without the use of biotext vectors, scGREAT still consistently outperforms other methods, indicating its effectiveness in identifying regulatory relationships between TFs and genes. In conclusion, the biotext dictionary, aggregating biological information of gene symbols, offers some support for gene regulation tasks.

Running time

We investigate the running time of each method on different sizes of hESC datasets with cell type-specific networks (see Table 5). Our method demonstrates competitive runtimes on the TFs+500 genes and TFs+1000 genes with around 10^1 minutes while achieving the best

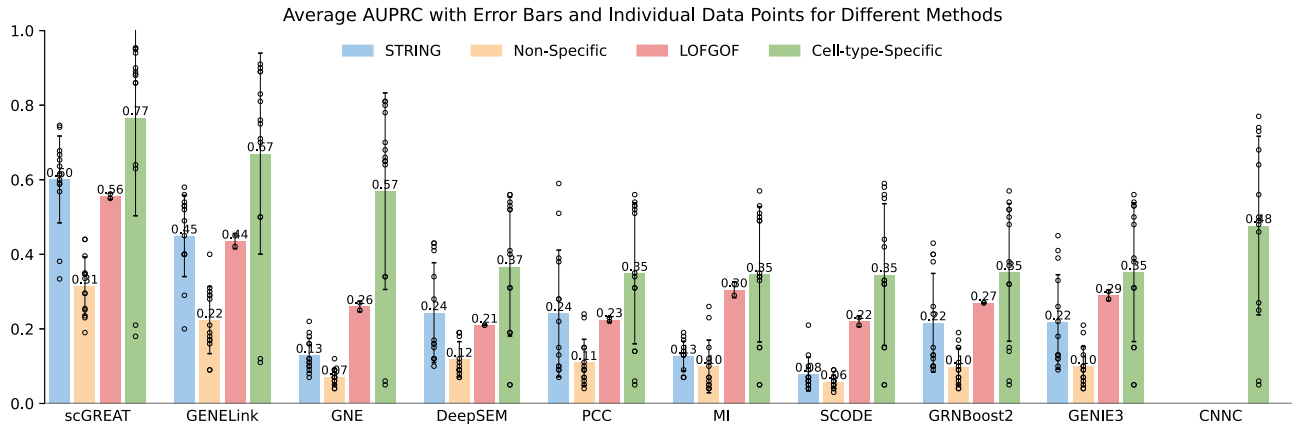


Figure 6. AUPRC Bar

Average AUPRC with error bars (mean ± s.d) and data points for scGREAT and the state-of-the-art methods across different platforms.

performance in AUROC and AUPRC. Regarding runtime, scGREAT might incur a slight sacrifice compared to the GENELink and DeepSEM models. However, it significantly excels in performance metrics that, compared to GENELink, scGREAT shows a 7% enhancement in AUROC and a 13% improvement in AUPRC. Further, in comparison to DeepSEM, scGREAT exhibits a 31% increase in AUROC and a 45% enhancement in AUPRC. The results were obtained using an Ubuntu 20.04 computer equipped with an 8-core 2.9 GHz processor, and an NVIDIA GeForce RTX 3080 with 10 GB of memory.

Unveiling unannotated regulatory relationships

To demonstrate the ability of the scGREAT in identifying potential gene regulatory relationships, in addition to statistical models for AUROC and AUPRC indicators,[72] we also performed a comprehensive literature search and external validation from spatial transcriptomics data to verify predicted relationships by scGREAT using the optimal threshold screening on the testing set of benchmark datasets for confirming the authenticity. We found that various potential positive regulatory relationships predicted by scGREAT, which were not reflected in the ground truth, have rich literature support (see Table S2 and Figure 11).

Specifically, in the mESC500 dataset (see Figure 11 Left), the predicted relationship between E2F1 and TOP2A was proved by[34] that after the induction of E2F-1, the expression of TOP2A increased, which contributed to the assembly and function of the DNA replication mechanism in the process of the cell cycle. There is also pathway evidence from Wikipath (https://www.wikipathways.org/instance/WP3206) to support the illustration. Furthermore, the regulatory between SOX2 and CCND1 predicted was proved by.[35] Specifically, SOX2 facilitates the G1/S transition in the cell cycle and regulates the transcription of the CCND1 gene, which encodes cyclin D1.[35] And the NANOG - CCND1 predicted regulatory relationship has also been reported. The study[36] demonstrates that when the expression of NANOG was silenced in MCF-7 breast cancer cells, the expression of cyclin D1 (encoded by the CCND1 gene) and c-myc were significantly downregulated, and cell cycle progression was blocked in G0/G1 phase. Additionally, chromatin immunoprecipitation experiments revealed that NANOG protein could bind directly to the promoter region of the CCND1 gene, thereby regulating its expression and influencing cell cycle progression.[36]

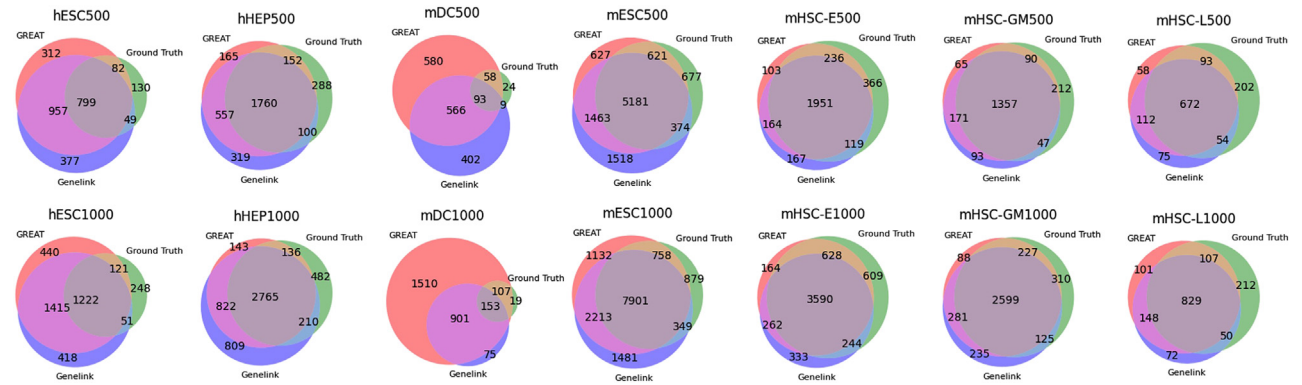


Figure 7. Venn

Venn diagram of the positive regulatory relationship predicted by scGREAT, GENELink, and the Ground Truth of Cell-type-specific ChIP-seq network.

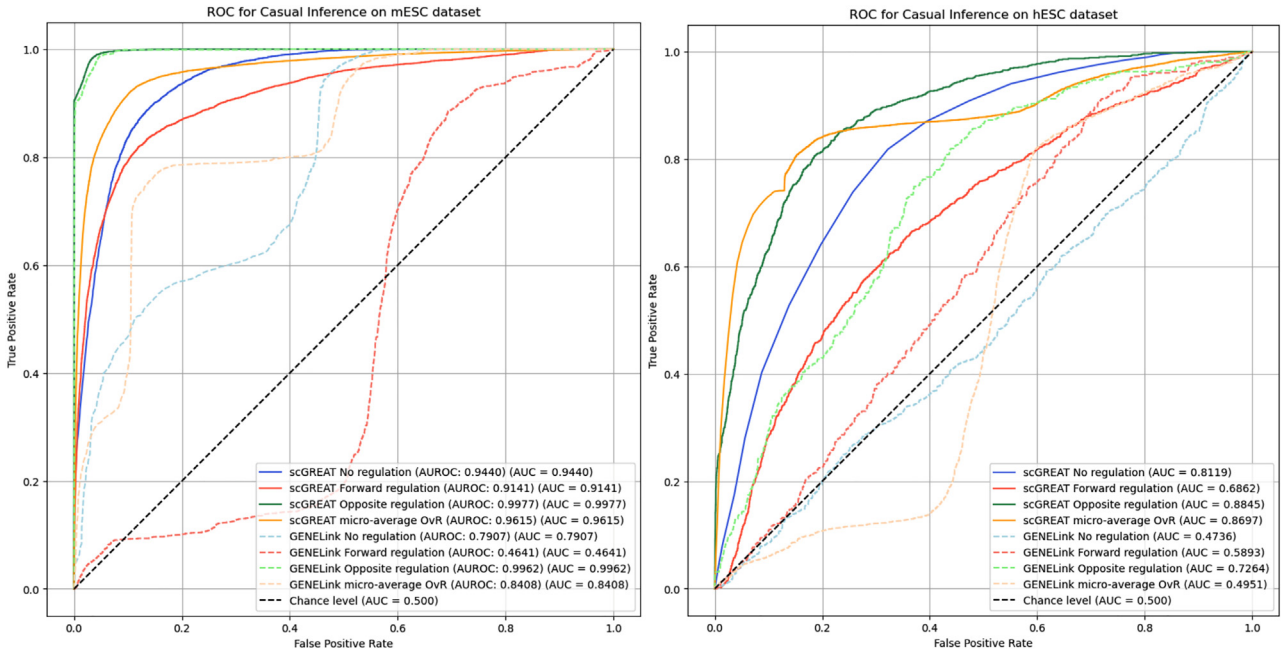


Figure 8. Causality Inference

Comparison of scGREAT and GENELink ROC curves for task of causality inference in mESC500 dataset (left) and hESC500 dataset (right) on the cell-type-specific ChIP-seq network.

In addition, in the hESC500 dataset (see Figure 11 Right), the predicted potential regulatory relationship (TRIM28(aka KAP1)-ID1) has been verified by.[37] They explored the regulatory interactions between KAP1 and ID1, which are influenced by MAGE I proteins (MAGE-A3 and MAGE-C2) and a KZNF family member, ZNF382.[37] Typically, ZNF382 directs KAP1 to repress the ID1 gene by causing localized heterochromatin changes, evident through histone 3 lysine 9 trimethylations (H3me3K9).[37] When MAGE I proteins are expressed, they bind to KAP1, reducing its repression of ID1, leading to increased ID1 mRNA expression and chromatin relaxation. Simultaneously, these proteins cause

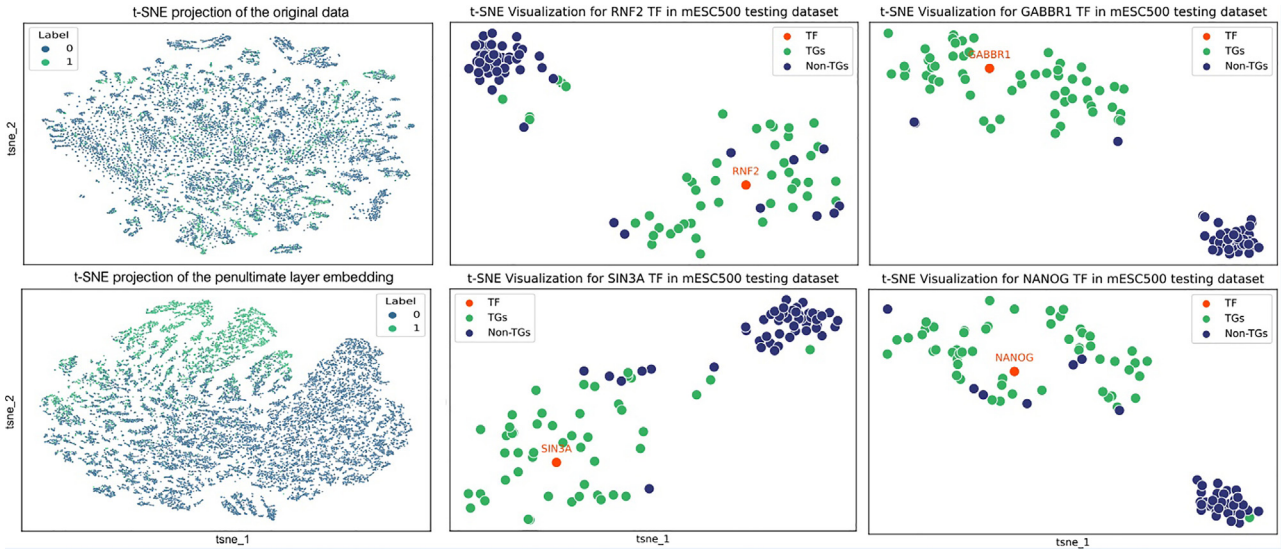


Figure 9. Embedding

t-SNE projection of all input pairs of TFs-gene vectors and the penultimate layer embedding (Left column). TF, TGs, and Non-TGs embedding t-SNE projections in mESC testing dataset with most-varying 500 genes on the Cell-type-specific ChIP-seq network (right two columns).

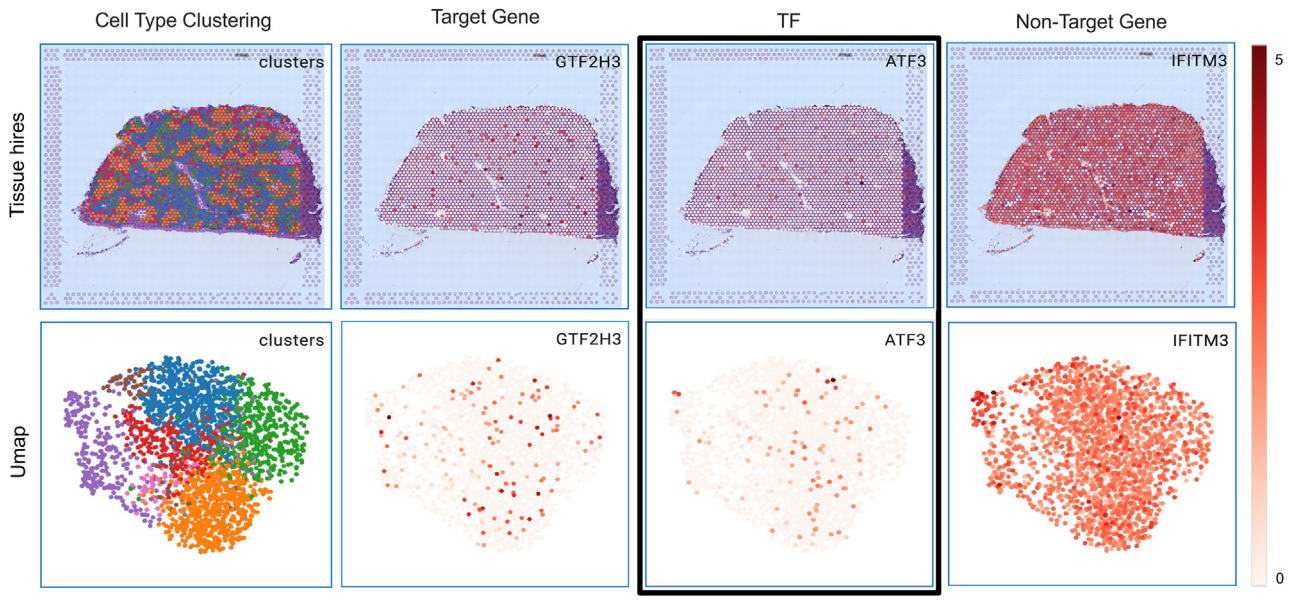


Figure 10. Spatial validation

External validation on spatial single-cell RNA sequencing of the human liver with emphasis on hepatic stellate cell identification and clustering. The figures with black boxes are the expression level and distribution of ATF3 TF in the tissue. The left panels illustrate target gene expressions and spatial distribution, and the right displays non-target gene expressions and spatial distribution of the respective TF. The color of the point in figures represents the expression level of the gene in the slice and the position of the point represents the expression of the gene in one of the seven categories of cells.

ZNF382 degradation, further reducing KAP1's binding to ID1.[37] In addition, the predicted relationship between TFAP2A and CRX can be confirmed in the study of[38] and.[39] Similarly, the inferred STAT3-HIFIA interaction has been proved by several studies. According to,[40] transducer and activator of transcription 3 (STAT3) have been found to upregulate the expression of EPO induced by hypoxia-inducible factor 1 alpha (HIF-1alpha). This process involves the transcription activation function of STAT3, enabling HIF-1alpha to induce the expression of the EPO gene more effectively. The correlation between the expression of STAT3 and HIF-1alpha was also found in breast cancer samples (correlation coefficient r = 0.4012, p value<0.0001), which reflects functional dependencies among STAT3, HIF-1alpha, EPO, and EPOR in cellular signal conduction.[41] also discovered the novel regulatory mechanism between signal transducer and STAT3 and HIF-1 under both hypoxia and growth signaling conditions. STAT3 is found to be necessary for HIF-1alpha RNA expression in both cancer cells and myeloid cells within the tumor microenvironment. Tumor-derived myeloid cells express elevated levels of HIF-1alpha mRNA in a STAT3-dependent manner. These findings suggest that HIF-1alpha regulation occurs not only in cancer cells but also in tumor-associated inflammatory cells, indicating STAT3 as a significant molecular target for inhibiting the oncogenic potential of HIF-1alpha induced by both hypoxia and overactive growth signaling pathways in cancer. The other study also revealed that hypoxia-induced STAT3 phosphorylation and HIF-1alpha are functionally interconnected in the context of a tumor microenvironment.[42] These findings indicate a cooperative relationship between STAT3 and HIF-1alpha in modulating the immune response within the tumor microenvironment, which could be exploited for immunotherapeutic interventions.

Table 3. Performance of scGREAT with and without the Transformer architecture on benchmark datasets with specific ChIP-seq network in both AUROC and AUPRC evaluation metric

	w Transformer		w/o Transformer		Difference	
Datasets	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
hESC	0.893	0.635	0.830	0.440	6.3%	19.5%
hHEP	0.912	0.860	0.798	0.630	11.4%	23.0%
mDC	0.812	0.195	0.768	0.123	4.4%	7.2%
mESC	0.944	0.895	0.877	0.760	6.7%	13.5%
mHSC-E	0.933	0.952	0.854	0.870	7.9%	8.2%
mHSC-GM	0.934	0.947	0.859	0.850	7.5%	9.7%
mHSC-L	0.882	0.882	0.82	0.810	6.2%	7.2%

Table 4. Performance of scGREAT with and without biotext vectors on benchmark datasets with specific ChIP-seq network in both AUROC and AUPRC evaluation metric

	w/BioVec		w/o BioVec		Difference	
Datasets	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
hESC	0.893	0.635	0.879	0.619	1.45%	1.65%
hHEP	0.912	0.860	0.903	0.841	0.91%	1.92%
mDC	0.812	0.195	0.803	0.157	0.91%	3.79%
mESC	0.944	0.895	0.930	0.863	1.49%	3.16%
mHSC-E	0.933	0.952	0.932	0.949	0.12%	0.34%
mHSC-GM	0.934	0.947	0.934	0.944	0.00%	0.24%
mHSC-L	0.882	0.882	0.878	0.876	0.39%	0.65%

BioVec represents biotext vectors.

In summary, these new findings illustrate the capacity of scGREAT to discover unannotated gene regulatory relationships, many of which are supported by existing literature. This highlights the value of the scGREAT tool in uncovering potential regulatory relationships.

DISCUSSION

Single-cell technology has brought opportunities for GRN inference but also challenges. In order to address the sparsity and platform noises as well as to enrich the contextual information of genes, we proposed scGREAT, a transformer-based supervised deep language model, taking advantage of end-to-end deep learning and large language models for GRN inference from single-cell transcriptomic data. By establishing an analogy between the relationships of genes and cells and those of words and sentences, we effectively adapted state-of-the-art natural language processing methods to genomics. In addition, we employed a biomedical language representation model to generate tokenized biotext vectors for each gene symbol, which encapsulates valuable gene information derived from biomedical corpora.

The results have reflected that scGREAT can significantly achieve competitive performance with an average AUROC of 91.30% and an average AUPRC of 55.97% across seven benchmark datasets encompassing four types of networks. These findings indicate that the integration of transcriptomic data and language embeddings substantially improves the actual performance in GRN inference tasks. Notably, the scGREAT model identified unannotated regulatory relationships in the benchmark datasets. These relationships were subsequently confirmed and verified by various related studies and papers, demonstrating the model's ability to uncover gene relationships.

Conclusion

In our research, we introduced scGREAT, a innovative transformer-based supervised deep language model tailored for GRN inference from single-cell transcriptomic data. Drawing inspiration from NLP, we likened the relationships between genes and cells to those of words and sentences, combining with the application of a biomedical language model, allowed us to abstract meaningful biotext vectors. Our findings highlight that scGREAT not only surpasses the performance of other SOTA methods, but also excels in uncovering unannotated regulatory relationships that were not previously identified. These newly discovered gene relationships received validation from published studies, underscore scGREAT's capacity in unveiling novel gene dynamics. Our research offers a promising avenue for leveraging the power of NLP in genomics, with the potential to revolutionize our understanding of GRNs.

Limitations of the study

The drawback of supervised learning in biological network inference is that reliable ground truth and sufficient negative samples are assumed. In this condition, we adopted the uniformly negative sampling strategy to select negative targets for TFs from massive genes. In addition, ablation experiments implied that the improvement provided by BioBERT in performance is limited. This limitation arises because the nature of gene regulatory inference tasks is intrinsically tied to specific datasets; it is important to clarify that, although BioBERT aggregates information from a wide array of text sources, its training on a vast text database primarily equips itself with the ability to assimilate general information instead of specifics pertinent to GRN inference tasks, ensuring that there is no information leakage regarding the intricacies of individual GRNs. While, in other words, without biotext vectors, the results of scGREAT still outperformed other methods. For future

Table 5. Running time of methods

Running time	scGREAT	GENELink	GNE	CNNC	DeepSEM	SCODE	GRNBoost2	GENIE3	PCC	MI
TF+500 genes	5m	2m30s	6m20s	2h15m	1m	5m	20m	2h15m	20s	20s
TF+1000 genes	12m	6m	30m	20h40m	2m	5m	50m	3h50m	32s	32s

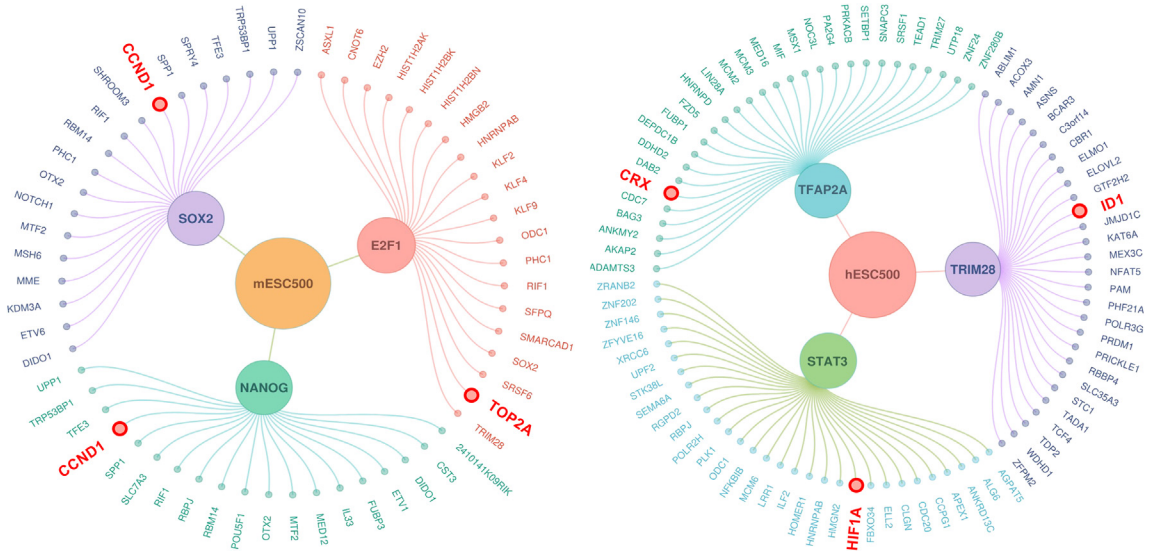


Figure 11. GRN

Potential regulatory relationships in mESC500 dataset (Left) and hESC500 dataset (Right) on the cell-type-specific ChIP-seq network. The central circle represents the dataset, the intermediate circles indicate TFs, and the outer nodes are target genes. The red nodes are the predicted Target genes that have been proven to be regulated by certain TFs.

research, we aim to introduce semi-supervised learning approaches to mitigate the reliance on reliable prior knowledge of regulatory relationships. Additionally, we plan to integrate the varied representation of gene symbols in literature, such as different symbols, names, and numbers, to increase the contribution of biotext to the model.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - Benchmark scRNA-seq datasets and preprocessing
  - Training, validation, and testing datasets
  - The scGREAT framework
  - GRN construction
  - Training setting
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Evaluation metrics

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2024.109352.

ACKNOWLEDGMENTS

This research was substantially sponsored by the research projects (Grant No. 32170654 and Grant No. 32000464) supported by the National Natural Science Foundation of China and was substantially supported by the Shenzhen Research Institute, City University of Hong Kong. The work described in this paper was substantially supported by the grant from the Research Grants Council of the Hong Kong Special Administrative Region [CityU 11203723]. This project was substantially funded by the Strategic Interdisciplinary Research Grant of City University of Hong Kong (Project No. 2021SIRG036). The work described in this paper was partially supported by the grant from City University of Hong

Kong (CityU 9667265). The three anonymous reviewers are thanked for their time and efforts, improving numerous aspects of the current study.

AUTHOR CONTRIBUTIONS

W.Y. and C.X. conceived the scGREAT Architecture. W.Y., C.X., and Z.Z. conceived the experiments. W.Y. carried out the experiments, wrote and revised the manuscript. C.X., Z.Z., H.L., X.W., W.F. proofread the manuscript. W.K.-C. and Z.Z. guided and proofread the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interest.

Received: September 1, 2023
Revised: December 29, 2023
Accepted: February 23, 2024
Published: February 28, 2024

REFERENCES

1. Hwang, B., Lee, J.H., and Bang, D. (2018). Single-cell RNA sequencing technologies and bioinformatics pipelines. Exp. Mol. Med. *50*, 1–14.
2. Delgado, F.M., and Gómez-Vela, F. (2019). Computational methods for Gene Regulatory Networks reconstruction and analysis: A review. Artif. Intell. Med. *95*, 133–145.
3. Alawad, D.M., Katebi, A., Kabir, M.W.U., and Hoque, M.T. (2023). AGRN: accurate gene regulatory network inference using ensemble machine learning methods. Bioinform. Adv. *3*, vbad032.
4. Li, L., Sun, L., Chen, G., Wong, C.W., Ching, W.K., and Liu, Z.P. (2023). LogBTF: gene regulatory network inference using Boolean threshold network model from single-cell gene expression data. Bioinformatics *39*, btad256.
5. Zhao, M., He, W., Tang, J., Zou, Q., and Guo, F. (2021). A comprehensive overview and critical evaluation of gene regulatory network inference technologies. Briefings Bioinf. *22*, bbab009.
6. Shu, H., Zhou, J., Lian, Q., Li, H., Zhao, D., Zeng, J., and Ma, J. (2021). Modeling gene regulatory networks using neural network architectures. Nat. Comput. Sci. *1*, 491–501.
7. Prosperi, M., Guo, Y., Sperrin, M., Koopman, J.S., Min, J.S., He, X., Bian, J., Rich, S., Wang, M., Buchan, I.E., and Bian, J. (2020). causal inference and counterfactual prediction in machine learning for actionable healthcare. Nat. Mach. Intell. *2*, 369–375.
8. Kc, K., Li, R., Cui, F., Yu, Q., and Haake, A.R. (2019). GNE: a deep learning framework for gene network inference by aggregating biological information. BMC Syst. Biol. *13*, 38.
9. Chen, X., Zhu, Z., Zhang, W., Wang, Y., Wang, F., Yang, J., and Wong, K.C. (2022). Human disease prediction from microbiome data by multiple feature fusion and deep learning. iScience *25*, 104081.
10. Wang, Q., Guo, M., Chen, J., and Duan, R. (2023). A gene regulatory network inference model based on pseudo-siamese network. BMC Bioinf. *24*, 163–218.
11. Chen, J., Cheong, C., Lan, L., Zhou, X., Liu, J., Lyu, A., Zhang, L., Cheung, W.K., and Zhang, L. (2021). DeepDRIM: a deep neural network to reconstruct cell-type-specific gene regulatory network using single-cell RNA-seq data. Briefings Bioinf. *22*, bbab325.
12. Zhao, M., He, W., Tang, J., Zou, Q., and Guo, F. (2022). A hybrid deep learning framework for gene regulatory network inference from single-cell transcriptomic data. Briefings Bioinf. *23*, bbab568.
13. Shrivastava, H., Zhang, X., Song, L., and Aluru, S. (2022). GRNUlar: A Deep Learning Framework for Recovering Single-Cell Gene Regulatory Networks. J. Comput. Biol. *29*, 27–44.
14. Matsumoto, H., Kiryu, H., Furusawa, C., Ko, M.S.H., Ko, S.B.H., Gouda, N., Nikaido, I., Hayashi, T., and Nikaido, I. (2017). SCODE: an efficient regulatory network inference algorithm from single-cell RNA-Seq during differentiation. Bioinformatics *33*, 2314–2321.
15. Huynh-Thu, V.A., Irrthum, A., Wehenkel, L., and Geurts, P. (2010). Inferring regulatory networks from expression data using tree-based methods. PLoS One *5*, e12776.
16. Moerman, T., Aibar Santos, S., Bravo González-Blas, C., Simm, J., Moreau, Y., Aerts, J., and Aerts, S. (2019). GRNBoost2 and Arboreto: efficient and scalable inference of gene regulatory networks. Bioinformatics *35*, 2159–2161.
17. Aibar, S., González-Blas, C.B., Moerman, T., Huynh-Thu, V.A., Imrichova, H., Hulselmans, G., Aerts, S., Rambow, F., Marine, J.C., Geurts, P., Aerts, J., et al. (2017). SCENIC: single-cell regulatory network inference and clustering. Nat. Methods *14*, 1083–1086.
18. Van de Sande, B., Flerin, C., Davie, K., De Waegeneer, M., Hulselmans, G., Aibar, S., Aerts, S., Seurinck, R., Saelens, W., Cannoodt, R., Rouchon, Q., et al. (2020). A scalable SCENIC workflow for single-cell gene regulatory network analysis. Nat. Protoc. *15*, 2247–2276.
19. He, X., and Chua, T.S. (2017, August). Neural factorization machines for sparse predictive analytics. In Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval, pp. 355–364.
20. Chen, G., and Liu, Z.P. (2022). Graph attention network for link prediction of gene regulations from single-cell RNA-sequencing data. Bioinformatics *38*, 4522–4529.
21. Khan, A., Sohail, A., Zahoora, U., and Qureshi, A.S. (2020). A survey of the recent architectures of deep convolutional neural networks. Artif. Intell. Rev. *53*, 5455–5516.
22. Lan, K., Wang, D.T., Fong, S., Liu, L.S., Wong, K.K.L., and Dey, N. (2018). A survey of data mining and deep learning in bioinformatics. J. Med. Syst. *42*, 139–220.
23. Alzubaidi, L., Zhang, J., Humaidi, A.J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Farhan, L., Santamaría, J., Fadhel, M.A., Al-Amidie, M., and Farhan, L. (2021). Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. J. Big Data *8*, 53–74.
24. Yuan, Y., and Bar-Joseph, Z. (2019). Deep learning for inferring gene relationships from single-cell expression data. Proc. Natl. Acad. Sci. USA *116*, 27151–27158.
25. ElAbd, H., Bromberg, Y., Hoarfrost, A., Lenz, T., Franke, A., and Wendorff, M. (2020). Amino acid encoding for deep learning applications. BMC Bioinf. *21*, 235–314.
26. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., and Polosukhin, I. (2017). Attention is all you need. Adv. Neural Inf. Process. Syst. *30*.
27. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., and Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics *36*, 1234–1240.
28. Pratapa, A., Jalihal, A.P., Law, J.N., Bharadwaj, A., and Murali, T.M. (2020). Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. Nat. Methods *17*, 147–154.
29. Qiu, X., Rahimzamani, A., Wang, L., Ren, B., Mao, Q., Durham, T., Kannan, S., McFaline-Figueroa, J.L., Saunders, L., Trapnell, C., and Kannan, S. (2020). Inferring causal gene regulatory networks from coupled single-cell expression dynamics using scribe. Cell Syst. *10*, 265–274.e11.
30. Kobak, D., and Berens, P. (2019). The art of using t-SNE for single-cell transcriptomics. Nat. Commun. *10*, 5416.
31. Puniyani, K., and Xing, E.P. (2013). GINI: from ISH images to gene interaction networks. PLoS Comput. Biol. *9*, e1003227.
32. Yang, Y., Fang, Q., and Shen, H.B. (2019). Predicting gene regulatory interactions based on spatial gene expression data and deep learning. PLoS Comput. Biol. *15*, e1007324.
33. Andrews, T.S., Atif, J., Liu, J.C., Perciani, C.T., Ma, X.Z., Thoeni, C., Slyper, M., Eraslan, G.,

Segerstolpe, A., Manuel, J., et al. (2021). Single-cell, single-nucleus, and spatial RNA sequencing of the human liver identifies cholangiocyte and mesenchymal heterogeneity. Hepatol. Commun. *6*, 821–840.

34. Kalma, Y., Marash, L., Lamed, Y., and Ginsberg, D. (2001). Expression analysis using DNA microarrays demonstrates that E2F-1 up-regulates expression of DNA replication genes including replication protein A2. Oncogene *20*, 1379–1387.
35. Chen, Y., Shi, L., Zhang, L., Li, R., Liang, J., Yu, W., Sun, L., Yang, X., Wang, Y., Zhang, Y., and Shang, Y. (2008). The molecular mechanism governing the oncogenic potential of SOX2 in breast cancer. J. Biol. Chem. *283*, 17969–17978.
36. Han, J., Zhang, F., Yu, M., Zhao, P., Ji, W., Zhang, H., Wu, B., Wang, Y., and Niu, R. (2012). RNA interference-mediated silencing of NANOG reduces cell proliferation and induces G0/G1 cell cycle arrest in breast cancer cells. Cancer Lett. *321*, 80–88.
37. Xiao, T.Z., Bhatia, N., Urrutia, R., Lomberk, G.A., Simpson, A., and Longley, B.J. (2011). MAGE I transcription factors regulate KAP1 and KRAB domain zinc finger transcription factor mediated gene repression. PLoS One *6*, e23747.
38. Han, H., Cho, J.W., Lee, S., Yun, A., Kim, H., Bae, D., Lee, I., Yang, S., Kim, C.Y., Lee, M., Kim, E., et al. (2018). TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. Nucleic Acids Res. *46*, D380–D386.
39. Hodges, M.D., Vieira, H., Gregory-Evans, K., and Gregory-Evans, C.Y. (2002). Characterization of the genomic and transcriptional structure of the CRX gene: substantial differences between human and mouse. Genomics *80*, 531–542.
40. Wincewicz, A., Sulkowska, M., Koda, M., Leśniewicz, T., Kanczuga-Koda, L., and Sulkowski, S. (2007). STAT3, HIF-1alpha, EPO and EPOR - signaling proteins in human primary ductal breast cancers. Folia Histochem. Cytobiol. *45*, 81–86.
41. Niu, G., Briggs, J., Deng, J., Ma, Y., Lee, H., Kortylewski, M., Kujawski, M., Kay, H., Cress, W.D., Jove, R., and Yu, H. (2008). Signal transducer and activator of transcription 3 is required for hypoxia-inducible factor-1alpha RNA expression in both tumor cells and tumor-associated myeloid cells. Mol. Cancer Res. *6*, 1099–1105. https://doi.org/10.1158/1541-7786.MCR-07-2177.
42. Noman, M.Z., Buart, S., Van Pelt, J., Richon, C., Hasmim, M., Leleu, N., Suchorska, W.M., Jalil, A., Lecluse, Y., El Hage, F., et al. (2009). The cooperative induction of hypoxia-inducible factor-1 alpha and STAT3 during hypoxia induced an impairment of tumor susceptibility to CTL-mediated cell lysis. J. Immunol. *182*, 3510–3521.
43. Szklarczyk, D., Gable, A.L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Mering, C.V., Simonovic, M., Doncheva, N.T., Morris, J.H., Bork, P., et al. (2019). STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. *47*, D607–D613.
44. Garcia-Alonso, L., Holland, C.H., Ibrahim, M.M., Turei, D., and Saez-Rodriguez, J. (2019). Benchmark and integration of resources for the estimation of human transcription factor activities. Genome Res. *29*, 1363–1375.
45. Liu, Z.P., Wu, C., Miao, H., and Wu, H. (2015). RegNetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse. Database *2015*.
46. ENCODE Project Consortium, Moore, J.E., Purcaro, M.J., Pratt, H.E., Epstein, C.B., Shoresh, N., Weng, Z., Adrian, J., Kawli, T., Davis, C.A., Dobin, A., et al. (2020). Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature *583*, 699–710.
47. Oki, S., Ohta, T., Shioi, G., Hatanaka, H., Ogasawara, O., Okuda, Y., Meno, C., Kawaji, H., Nakaki, R., Sese, J., and Meno, C. (2018). Ch IP-Atlas: a data-mining suite powered by full integration of public Ch IP-seq data. EMBO Rep. *19*, e46255.
48. Xu, H., Baroukh, C., Dannenfelser, R., Chen, E.Y., Tan, C.M., Kou, Y., Kim, Y.E., Lemischka, I.R., and Ma'ayan, A. (2013). ESCAPE: database for integrating high-content published data collected from human and mouse embryonic stem cells. Database *2013*, bat045.
49. Sun, S., Zhu, J., Ma, Y., and Zhou, X. (2019). Accuracy, robustness and scalability of dimensionality reduction methods for single-cell RNA-seq analysis. Genome Biol. *20*, 269–321.
50. Blatti, C., Kazemian, M., Wolfe, S., Brodsky, M., and Sinha, S. (2015). Integrating motif, DNA accessibility and gene expression data to build regulatory maps in an organism. Nucleic Acids Res. *43*, 3998–4012.
51. Marbach, D., Costello, J.C., Küffner, R., Vega, N.M., Prill, R.J., Camacho, D.M., Stolovitzky, G., Allison, K.R., DREAM5 Consortium, Kellis, M., Collins, J.J., and Stolovitzky, G. (2012). Wisdom of crowds for robust gene network inference. Nat. Methods *9*, 796–804.
52. De Smet, R., and Marchal, K. (2010). Advantages and limitations of current network inference methods. Nat. Rev. Microbiol. *8*, 717–729.
53. Dundar, M., Krishnapuram, B., Rao, R., and Fung, G. (2006). Multiple instance learning for computer aided diagnosis. Adv. Neural Inf. Process. Syst. *19*.
54. Chandrasekaran, S., and Price, N.D. (2010). Probabilistic integrative modeling of genome-scale metabolic and regulatory networks in Escherichia coli and Mycobacterium tuberculosis. Proc. Natl. Acad. Sci. USA *107*, 17845–17850.
55. Thabtah, F., Hammoud, S., Kamalov, F., and Gonsalves, A. (2020). Data imbalance in classification: Experimental evaluation. Inf. Sci. *513*, 429–441.
56. Radenović, F., Tolias, G., and Chum, O. (2016). CNN image retrieval learns from BoW: Unsupervised fine-tuning with hard examples. In Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14 (Springer International Publishing), pp. 3–20.
57. Zhu, X., Jing, X.Y., Zhang, F., Zhang, X., You, X., and Cui, X. (2019). Distance learning by mining hard and easy negative samples for person re-identification. Pattern Recogn. *95*, 211–222.
58. Yang, Z., Ding, M., Zou, X., Tang, J., Xu, B., Zhou, C., and Yang, H. (2022). Region or global a principle for negative sampling in graph-based recommendation. IEEE Trans. Knowl. Data Eng. *35*, 6264–6277.
59. Suh, Y., Han, B., Kim, W., and Lee, K.M. (2019). Stochastic class-based hard example mining for deep metric learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7251–7259.
60. Huynh-Thu, V.A., and Sanguinetti, G. (2019). Gene regulatory network inference: an introductory survey. Methods Mol. Biol. *1883*, 1–23.
61. Mercatelli, D., Scalambra, L., Triboli, L., Ray, F., and Giorgi, F.M. (2020). Gene regulatory network inference resources: A practical overview. Biochim. Biophys. Acta. Gene Regul. Mech. *1863*, 194430.
62. Wang, Y., Joshi, T., Zhang, X.S., Xu, D., and Chen, L. (2006). Inferring gene regulatory networks from multiple microarray datasets. Bioinformatics *22*, 2413–2420.
63. Zhang, S., Li, X., Lin, Q., Lin, J., and Wong, K.C. (2020). Uncovering the key dimensions of high-throughput biomolecular data using deep learning. Nucleic Acids Res. *48*, e56.
64. Yu, J., Li, J., Yu, Z., and Huang, Q. (2020). Multimodal transformer with multi-view visual representation for image captioning. IEEE Trans. Circ. Syst. Video Technol. *30*, 4467–4480.
65. Ba, J.L., Kiros, J.R., and Hinton, G.E. (2016). Layer normalization. Preprint at arXiv. https://doi.org/10.48550/arXiv.1607.06450.
66. Veit, A., Wilber, M.J., and Belongie, S. (2016). Residual networks behave like ensembles of relatively shallow networks. Adv. Neural Inf. Process. Syst. *29*.
67. Gholamalinezhad, H., and Khosravi, H. (2020). Pooling methods in deep neural networks, a review. Preprint at arXiv. https://doi.org/10.48550/arXiv.2009.07485.
68. Noh, H., Hong, S., and Han, B. (2015). Learning deconvolution network for semantic segmentation. In Proceedings of the IEEE international conference on computer vision, pp. 1520–1528.
69. Joseph, A.A., Abdullahi, M., Junaidu, S.B., Ibrahim, H.H., and Chiroma, H. (2022). Improved multi-classification of breast cancer histopathological images using handcrafted features and deep neural network (dense layer). Intelligent Systems with Applications *14*, 200066.
70. Jin, S., Zeng, X., Xia, F., Huang, W., and Liu, X. (2021). Application of deep learning methods in biological networks. Briefings Bioinf. *22*, 1902–1917.
71. Ozenne, B., Subtil, F., and Maucort-Boulch, D. (2015). The precision–recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases. J. Clin. Epidemiol. *68*, 855–859.
72. Fan, Z., Chen, R., and Chen, X. (2020). SpatialDB: a database for spatially resolved transcriptomes. Nucleic Acids Res. *48*, D233–D237.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
scRNA-seq dataset	BEELINE	https://doi.org/10.5281/zenodo.3701939
Software and algorithms		
BioBERT: a biomedical language model for biomedical vector extracting	BioBERT	https://doi.org/10.1093/bioinformatics/btz682
scGREAT	This paper	https://doi.org/10.5281/zenodo.10646474

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Prof. Ka-Chun Wong (kc.w@cityu.edu.hk).

Materials availability

This study did not generate new biological data.

Data and code availability

- All relevant data have been deposited at Github and is publicly available as of the date of publication. DOI is listed in the key resources table.
- All original code has been deposited at Github and is publicly available as of the date of publication. DOI is listed in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

This paper analyzes existing, publicly available data. The study does not use experimental models typical in life sciences.

METHOD DETAILS

Benchmark scRNA-seq datasets and preprocessing

We evaluate the performance of scGREAT on seven cell types of scRNA-seq datasets in BEELINE[28]: (i) human embryonic stem cells (hESC); (ii) human mature hepatocytes (hHEP); (iii) mouse dendritic cells (mDC); (iv) mouse embryonic stem cells (mESC); (v) mouse hematopoietic stem cells with erythroid-lineage (mHSC-E); (vi) mouse hematopoietic stem cells with granulocyte-monocyte-lineage (mHSC-GM); (vii) mouse hematopoietic stem cells with lymphoid-lineage (mHSC-L). Those datasets are all with three ground-truth networks from functional interaction networks documented in STRING database,[43] non-specific ChIP-seq[38,44,45] and cell-type-specific ChIP-seq.[46–48] Harnessing cell-type-specific ChIP-seq enables the analysis of regulatory information focused on a single specific cell type, while non-specific ChIP-seq is performed on a mixed population of cells and thus reflects average and comprehensive regulatory events across all the cell types present. The mESC dataset is also with the ground truth from loss-/gain-of-function (LOF/GOF).[48] The seven scRNA-seq datasets can be downloaded from Gene Expression Omnibus with the accession numbers: GSE81252 (hHEP), GSE75748 (hESC), GSE98664 (mESC), GSE48968 (mDC) and GSE81682 (mHSC).

We pre-processed each scRNA-seq dataset by following[28] and specifically focused on inferring the interactions outgoing from TFs as previously outlined by Pratapa et al. on *Nature Methods* in 2020.[28] The top 500 and 1000 most differential expressed genes (DEGs) were selected according to their sorted FDR adjusted p values for GRN inference.[28] Subsequently, scRNA-seq data is normalized using the *Z* score method outlined by Sun et al. to standardize gene expression features by removing the mean and scaling to unit variance.[49] The formula as follows:

$$z = \frac{x - \mu}{\sigma} \quad \text{(Equation 1)}$$

where μ and σ are the mean and standard deviation of the scRNA-seq data distribution, respectively. Besides, gene names are represented as embedded vectors from pre-trained language models, which allows to link the scRNA-seq data and the gene contextual information.[27]

Training, validation, and testing datasets

To construct positive and negative samples, we consider the regulatory relationships within ground truth networks as positive samples and all remaining TF-gene pairs as candidate negative samples.[50] These negative samples have a certain probability of incorporating true regulatory relationships, as they are yet undiscovered.[51] Given the fact that the total count of candidate regulatory relationships significantly surpasses that of actual regulatory relationships between TFs and genes,[52] resulting in a substantial imbalance in the sample distribution.[53] It has been suggested that regulatory networks between TFs and genes are sparsely populated in the real world,[54] therefore an overwhelming number of negative samples can significantly compromise the model's ability to focus on learning positive regulation effectively.[55] To alleviate this situation, we employ the hard negative samples (HNS) methodology, selecting negative samples that share a common TF with the positive samples to serve as HNS.[56] According to Zhu and Yang et al.,[57,58] HNS contains more discriminative information because these samples demonstrate a high degree of similarity but bear opposite labels to their positive counterparts. Comparing to other negative samples, it is particularly challenging for a model to predict them correctly since they are referred to as 'hard' negative samples.[59] We implemented HNS selection approach following GENELink,[20] in which for each positive TF-gene pair, a corresponding negative TF-gene pair involving the same TF is uniformly-random sampled from the pool of negative samples. Then we select 2/3 relationships from positive and hard negative samples for training and validation uniformly and randomly (the proportion of training and validation data is 9:1), and the remaining 1/3 for isolated and independent testing.[20] Table 1 tabulates the final training data statistical information used for training. In general, the proportion of positive samples is approximately equal to the network density based on the genes among the scRNA-seq datasets. The detailed statistics of each scRNA-seq dataset with four ground-truth networks with TFs and 500 (1000) most-varying genes are shown in Supplementary Table LABEL:NAR-TS1.

The scGREAT framework

GRN inference involves accurately predicting each regulatory relationship, with nodes representing genes (including TFs and their target genes in a cyclic GRN) and edges signifying their connections. After that, GRN is then built based on those identified relationships.[60] As such, the scGREAT framework interprets the reconstruction of GRN as a classification task that determines if a regulatory relationship exists between a TF and a target gene.[61] scGREAT consists of four major components: feature initialization, gene dictionary construction, inference engine, and GRN construction (see Figure 1).

Feature initialization

For Network data processing, according to ground-truth network and negative samples from HNS, we randomly constructed training, validation, and testing samples with paired gene symbols and labels according to proportion mentioned before. For scRNA-seq data processing, we standardized gene expression data after train test splitting.

Gene dictionary construction

The gene dictionary operates primarily as a lookup table. Each gene is associated with a distinct index, allowing to efficiently retrieve its corresponding embedding using this index. This section is comprised of three parts: Positional embedding part (see Figure 1B top), followed by the construction of gene expression dictionary and the gene biotext dictionary (see Figure 1B middle & bottom).

For positional embedding part, the position vectors are used to determine the regulation of a gene from a TF. They are generated by position embedding with input $pos \in R^2$ number of 0 or 1. The formula is as follows, where $\partial \in R^{2*embed\_size}$ represents the parameters of the positional encoding *Embedding*, and $PE_{(pos)} \in R^{n*embed\_size}$. n is the number of input samples and $embed\_size$ in this paper is 768.

$$PE_{(pos)} = Embedding_{\partial}(pos) \quad \text{(Equation 2)}$$

Furthermore, the gene expression dictionary serves as an abstraction, capturing essential information from original data,[62] which is constructed with embedding transformed from the original gene expression data through the encoder (see Figure 2 left). It is referred to as a dictionary as it allows querying the encoded information of any gene according to the index. The TF or gene is encoded into embedding vectors with the length of 768, which represents the gene status within the context of gene expression data. Given the gene expression matrix $X \in R^{gn*cn}$, gn and cn represent the number of genes and cells, the encoder of scGREAT is expected to learn the mapping function f that can develop the gene representation as supported by a previously published study from our group,[63] $\theta \in R^{cn*embed\_size}$ is the set of parameters of the neural decoder f and $Dict_{expression} \in R^{gn*embed\_size}$.

$$Dict_{expression} = f_{\theta}(X) \quad \text{(Equation 3)}$$

The gene biotext dictionary is formed with embedding generated by gene symbols through the biomedical pre-trained model BioBERT (see Figure 2 middle). Similarly, it also enables querying the encoded information of any gene using its index. The encoded information represents the broad meaning of the gene in biological texts. The gene biotext dictionary provides access to contextual text meaning about genes for the model to look up, represented by biotext vector embedding sets, with the same length of gene expression dictionary for each gene symbol. Given the gene symbols list $S \in R^{embed\_size}$, the BioBERT denoted as V. Mathematically, $\omega \in R^{gn}$ is the set of parameters of V, and $Dict_{biotext} \in R^{gn*embed\_size}$.

$$Dict_{biotext} = V_{\omega}(S) \quad \text{(Equation 4)}$$

Inference engine

The primary function of inference engine is to understand and capture the characteristics and contextual information of the input sequence from the aggregated input (positional encoding, gene expression embedding, and gene biotext embedding) to extract multimodal feature representations (see Figure 1C). In practice, given a gene pair as represented by $genesymbol_a$ and $genesymbol_b$, they are ordered by serial numbers, which are the gene indices in the two dictionaries generated in the previous part. scGREAT searches the embedding vectors for each gene from the dictionaries according to the serial numbers and adds them together with the positional vectors (see Figure 2 right). Then the added vector gv is transferred into the attention module of the transformer backbone and represented by three informative vectors, q_g, k_g, and v_g acquired through the dot product with the learnable weight vectors w^v, w^k, and w^v.

$$q_g = gv \cdot wq$$

$$k_g = gv \cdot wk$$

$$v_g = gv \cdot wv$$

In practice, the mutual attention weight is computed by the following formula:

$$att(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad \text{(Equation 5)}$$

where $Q = \begin{pmatrix} q_{g_a} \\ q_{g_b} \end{pmatrix}$, $K = \begin{pmatrix} k_{g_a} \\ k_{g_b} \end{pmatrix}$, $V = \begin{pmatrix} v_{g_a} \\ v_{g_b} \end{pmatrix}$. d represents the dimension of vector k_g, which equals 768 in this study. q_{g_a} and q_{g_b} denote the q_g vectors generated by the two genes, $genesymbol_a$ and $genesymbol_b$ respectively, in a pair of genes. The *softmax* function[26] is used to transform the similarity scores between queries and keys into a probability distribution. The vectors gv' with rich interactions information are calculated in accordance with the formula:

$$gv' = \sum_{i=1}^{N} \frac{\exp\left(q_g \cdot k_{g_i}^T\right)}{\sum_{j=1}^{N} \exp\left(q_g \cdot k_{g_j}^T\right)} v^i \quad \text{(Equation 6)}$$

Moreover, we employed the multi-head attention that learns comprehensive information from diverse representation subspaces at different perspectives, which improves sensitivity to various patterns in the sequence.[64] The specific calculation is shown in the following formula:

$$MultiHead(Q, K, V) = Concat\left(att\left(QW_1^Q, KW_1^K, VW_1^V\right)^1, \ldots, att\left(QW_M^Q, KW_M^K, VW_M^V\right)^h\right)W^O \quad \text{(Equation 7)}$$

where $W^Q \in R^{d_m \times d_k}$, $W^K \in R^{d_m \times d_k}$, $W^V \in R^{d_m \times d_v}$, and $W^O \in R^{hd_v \times d_m}$ are trainable parameter matrices of attention module, M is the number of Heads.

$$gv'' = LayerNorm(gv' + Sublayer_{\delta}(gv')) \quad \text{(Equation 8)}$$

where $Sublayer(x)$ is the function of feedforward network (FFN) with RELU activation function implemented by the sub-layer itself. *LayerNorm* is the method used to normalizes the inputs across features in each layer.[65] δ is the parameter of FFN. The representation of the N-th layer is as follows:

$$Sublayer_{\delta N}(gv'') = \max\left(0, Sublayer_{\delta(N-1)}(gv'') \cdot W_N + b_N\right) \quad \text{(Equation 9)}$$

Subsequently, the output of the transformer encoder is as follows:

$$Transf\_e(gv'') = concat(LayerNorm(gv'') + \text{Sublayer}_{\delta N}(gv'')) \quad \text{(Equation 10)}$$

Following the transformer encoder backbone, we flatten the represented vector $Transf_{e(gva'')}$ and $Transf_{e(gvb'')}$, which then serves as inputs to the subsequent model structure. It comprises multi-layer perceptrons, constructed with N hidden layers for the transformation, with *PReLU* activations in between and a random dropout probability of 0.2.

$$PReLU(x) = \begin{cases} x, & \text{if } x \geq 0 \\ ax, & \text{otherwise} \end{cases} \quad \text{(Equation 11)}$$

where a is a learnable parameter. Additionally, we utilize the residual connection, a type of shortcut or skip connection helping in mitigating the vanishing gradient problem,[66] around the sub-dense layers (see Equation 8). Since the output dimension of dense network decrease,

average pooling operation[67] is implemented to ensure that the input from the earlier stage maintains the same dimension as the input for the subsequent layer.[68] The ultimate output of scGREAT is presented as follows:

$$scGREAT_{\Theta}(gx) = Softmax(Pooling(gx) + Dense_{\Phi}(gx')) \quad \text{(Equation 12)}$$

where $gx = flatten(Transf_e(gva''), Transf_e(gvb''))$ and gx' is the output after several dense layer[69] transformations and Φ involves the set of parameters of dense layers. Inspired by CNNC,[24] we further empower scGREAT with the capacity of causality inference in gene regulation. We flipped the TFs and genes in the network and set the third-class label with '0', '1', '2', where '0' means there is no regulatory relationship; '1' means gene1 regulate gene2; '2' means gene2 regulate gene1. Specifically the penultimate layer embedding is transformed into three output nodes to encode the causality by dense layer. The object of scGREAT is to optimize the difference between the predicted labels and the ground-truth by cross-entropy loss:

$$Loss = \sum_{j=1}^{N} -\left(y_j \log\left(scGREAT_{\Theta}(x_j)\right) + \left(1 - y_j\right)\log\left(1 - scGREAT_{\Theta}(x_j)\right)\right) \quad \text{(Equation 13)}$$

For causality inference:

$$Loss = \sum_{j=1}^{N}\sum_{j=1}^{K} -\left(y_j \log\left(scGREAT_{\Theta}(x_j)\right)\right) \quad \text{(Equation 14)}$$

where N denotes the number of samples, K is the number of classes, x_j represents the j_{th} TF-gene pair, and y_j and $scGREAT_{\Theta}(x_j)$ are the j_{th} true label and the predicted probability of regulatory relationship for the j_{th} TF-gene pair, respectively. And the Θ represents all the parameters of scGREAT.

GRN construction

After obtaining regulatory relationships of candidate gene pairs through scGREAT prediction, these relationships would be further validated through visualization of spatial transcriptome data of the same cell type. Finally the GRN is constructed based on the principle that: a link is established if positive regulation exists from a TF to its target gene; otherwise, no connection (see Figure 1D). The resultant GRN can serve as an auxiliary relationship table, offering a reference for biological researchers to comprehend and evaluate regulatory relationships and downstream tasks such as bioengineering.[70]

Training setting

To optimize scGREAT, the goal is to minimize the loss function, which encompasses all trainable parameters, denoted as Θ including the parameters θ for the decoder, ϑ for positional embedding, matrices W^Q, W^K, W^V, and W^O for the attention module, δ for the FFN layer, and Φ for the Dense layer. We trained scGREAT with the Adam stochastic gradient descent method. Taking the cell-type-specific ChIP-seq network ground truth as an example we employed default hyperparameters with a 0.00001 learning rate and a 0.999 weight decay rate every 10 steps. Training continues until either convergence or after completing 80 epochs. We utilize a mini-batch approach with a batch size of 32. The model training was conducted on an NVIDIA GeForce RTX 3080 GPU with 10 GB of memory.

QUANTIFICATION AND STATISTICAL ANALYSIS

Evaluation metrics

For evaluation, to facilitate the comparisons with other SOTA predictors, we follow the same evaluation standard as Chen and Liu[20] to examine the effectiveness of our models. We calculated the AUROC and AUPRC as the primary metrics to account for in situations where the data is imbalanced; AUROC alone might not provide a comprehensive performance assessment. AUPRC summarizes the trade-off between precision and recall, providing a more comprehensive reflection of the model's performance, especially when positive samples are few.[71] AUROC and AUPRC are commonly used metrics to evaluate the performance of a classifier and able to reflect both sensitivity and specificity. Generally, a higher score signifies better performance. An AUROC score of 0.5 corresponds to performance no better than random guessing, whereas a score of 1 indicates perfect classification. Similarly, for AUPRC, although the interpretation of a "superior" score depend more heavily on the class distribution, especially in imbalanced datasets where positive cases are rare, higher values are generally indicative of better performance.