RegFormer: A Single-Cell Foundation Model Powered by Gene Regulatory Hierarchies

Luni Hu[1,†], Ping Qiu[1,†], Hua Qin[1,†], Lei Cao[1], Wenjian Jiang[1], Yilin Zhang[1], Qianqian Chen[1], Yanbang Shang[1], Tianyi Xia[1], Ziqing Deng[1], Xun Xu[2,*], Shuangsang Fang[1,*], Yuxiang Li[2,*], Yong Zhang[2,*]

1. BGI Research, Beijing 102601, China

2. BGI Research, Wuhan 430074, China

*Correspondence: zhangyong2@genomics.cn (Y.Z) liyuxiang@genomics.cn (X.L), fangshuangsang@genomics.cn (S.F.), xuxun@genomics.cn (X.X.)

† These authors contributed equally as the first authors.

Abstract

Single-cell RNA sequencing (scRNA-seq) has significantly advanced our understanding of cellular diversity and the molecular mechanisms underlying biological processes. However, existing computational models often struggle to incorporate essential biological knowledge, handle sparse and noisy data, and scale effectively across large datasets. To address these challenges, we introduce RegFormer, a novel foundation model specifically designed for scRNA-seq analysis. RegFormer integrates hierarchical relationships from gene regulatory networks (GRNs) through an innovative architecture based on Mamba Blocks, enabling more effective modeling of gene interactions and cellular states. Pretrained on a vast dataset of 22 million human cells and comprising approximately 50 million parameters, RegFormer employs dual embeddings to separately capture gene expression levels and gene identities. This approach enhances interpretability by aligning gene expression data with regulatory hierarchies, offering more precise biological insights. Extensive evaluations demonstrate that RegFormer outperforms existing state-of-the-art models, such as scGPT, Geneformer, scFoundation, and scBERT, across a wide range of tasks, including cell annotation, GRN construction, genetic perturbation prediction, and drug response prediction. By combining cutting-edge deep learning techniques with biological knowledge, RegFormer not only improves accuracy and interpretability but also provides deeper insights into cellular processes and regulatory mechanisms, positioning it as a powerful tool for advancing biological discovery.

Introduction

Single-cell RNA sequencing (scRNA-seq) has revolutionized our understanding of cellular heterogeneity, gene expression dynamics, and the complex regulatory networks that underpin cellular processes[1-3]. By enabling the detailed characterization of individual cells, scRNA-seq has unveiled key biological mechanisms involved in tissue development[4], disease progression[5], and responses to environmental stimuli[6]. Large-scale initiatives like the Human Cell Atlas exemplify the growing scope of single-cell transcriptomics, providing comprehensive maps of cellular diversity across tissues and disease contexts[7]. Despite these transformative insights, computational methods for analyzing scRNA-seq data remain fragmented, with models often tailored for specific tasks and lacking scalability across diverse datasets and cellular states[8-10], underscoring the need for a unified, integrated analytical framework capable of performing multiple tasks simultaneously with higher accuracy and consistency.

To address this gap, recent advances have led to the development of single-cell foundation models, such as scGPT[11], Geneformer[12], scFoundation[13], and scBERT[14], which aim to provide a general-purpose solution for the analysis of scRNA-seq data. These models leverage deep learning techniques, which have proven effective at capturing complex patterns within high-dimensional, sparse scRNA-seq datasets. By learning representations of gene expression profiles, these models can achieve strong performance in tasks like cell annotation and genetic perturbation prediction. However, despite their success, these models typically rely on natural language processing (NLP) algorithms[15, 16], which are inherently designed for sequential data, such as text or genomic sequences. While these approaches are beneficial in certain contexts, they do not fully address the unique challenges posed by scRNA-seq data, specifically the unordered nature of gene expression measurements across individual cells. Unlike traditional sequence data, where elements (such as nucleotides or words) follow a fixed order, scRNA-seq data lacks this inherent structure, which makes it difficult for sequence-based models to capture the complex, non-linear interactions between genes and the hierarchical relationships that define cellular states.

To overcome these limitations, biological prior knowledge, particularly that derived from gene regulatory networks (GRNs), has been increasingly incorporated into scRNA-seq analysis[17-19]. GRNs represent the intricate regulatory relationships between genes and provide a framework for understanding how gene expression is regulated within cellular contexts. These networks capture key biological insights that are not immediately apparent from raw gene expression data alone, such as transcriptional hierarchical

dependencies, and spatial-temporal regulation of gene activity[20]. Integrating GRN-based prior knowledge into computational models improves both the interpretability and biological relevance of scRNA-seq analyses. This not only enhances the accuracy of predictions but also provides richer, more biologically grounded insights into the underlying regulatory mechanisms that govern cellular behavior and function. The integration of GRNs allows for the modeling of hierarchical dependencies between genes, which is critical for understanding cellular dynamics, differentiation, and disease mechanisms.

Here, we introduce RegFormer, a novel machine learning model specifically designed for scRNA-seq data analysis. RegFormer incorporates GRNs and features a unique architecture based on Mamba Blocks, optimized for handling high-dimensional and sparse data. This design enables RegFormer to capture gene interactions and cellular states at multiple scales, with a particular focus on the hierarchical relationships that regulate gene expression. Pretrained on a large-scale dataset of 22 million human cells, RegFormer employs a generative pretraining strategy that not only captures gene expression patterns but also models the hierarchical structure of gene regulation, improving biological interpretability. By integrating GRN-based prior knowledge, RegFormer effectively models transcriptional dependencies within this hierarchy, providing deeper insights into gene regulation and cellular behavior. This architecture allows RegFormer to navigate the complexities of scRNA-seq data, outperforming existing models like scGPT and GeneFormer in key tasks, such as cell annotation, GRN construction, genetic perturbation prediction, and drug response prediction. RegFormer's ability to combine high-dimensional data processing with biological insights makes it a powerful tool for advancing our understanding of cellular processes and accelerating biomedical research.

Results

RegFormer Leverages Gene Regulatory Networks and Mamba Blocks for Precise Single-Cell RNA Sequencing Interpretation.

RegFormer is an innovative foundation model tailored for single-cell RNA sequencing (scRNA-Seq) data analysis, designed to overcome significant challenges in this field by leveraging extensive datasets and the knowledge from gene regulatory networks (GRNs) (**Figure 1A**). Pretrained on a vast dataset encompassing 22 million human cells (**Supplementary Figure 1**), RegFormer boasts approximately 50 million parameters. The model's architecture integrates GRNs to enhance its biological interpretability (**Supplementary Figure 2**), with genes sorted topologically according to their regulatory relationships

within the network (**Figure 1B**). This sorting aligns expression data with regulatory hierarchies, ensuring biologically coherent order in the model's operations. The input to RegFormer consists of gene expression matrices (X), where each row represents a gene (var) and each column a cell (obs). Central to RegFormer's design is its integration of GRNs, which not only enriches the biological interpretability but also guides the model in understanding complex cellular processes.

During pretraining (**Figure 1C**), RegFormer adopts a dual-embedding approach to capture both quantitative and qualitative features of gene expression data: (1) Value Embeddings: These encode the levels of gene expression, providing the model with information about the intensity of expression; (2) Token Embeddings: These represent the identity of genes within the context of the GRN, enabling the model to learn the regulatory roles of individual genes and preserve the biological context throughout the modeling process. To improve robustness and ensure that the model can accurately reconstruct masked values, a portion of the expression levels are randomly concealed during training. This masking strategy challenges the model to predict these hidden values based on the available surrounding data, thereby enhancing its predictive capabilities.

These embeddings are then processed through Mamba Blocks, specialized computational units which can effectively handle the complexity of scRNA-Seq data **(Figure 1D)**. Each Mamba Block combines multi-layer perceptrons (MLP), convolutional layers (Conv), and state-space models (SSM) to effectively capture dependencies at various scales within the gene expression profiles—from detailed interactions between neighboring genes to broader trends across entire pathways or cellular states. The Mamba Block architecture supports parameter sharing, improving computational efficiency and ensuring that the learned features remain biologically relevant. This design facilitates scalable pretraining while maintaining focus on capturing meaningful biological representations. Training is guided by two complementary loss functions: (1)Masked-Value Prediction Loss: Focuses on restoring concealed gene expression levels, ensuring accurate reconstruction of the original data; (2) Next-Token Prediction Loss: Aims at predicting the next gene in the topologically sorted GRN sequence, helping the model to understand the hierarchical relationships among genes.

In summary, RegFormer's unique architecture, which includes the integration of GRNs and the use of Mamba Blocks, enables it to provide a powerful framework for analyzing scRNA-Seq data. It offers

precise predictions and deep insights into biological mechanisms, paving the way for advanced single-cell data interpretation.

RegFormer Enhances Single-Cell Data Embedding with Superior Performance Across Diverse Human Tissues

Single Cell Foundation Models are indispensable for accurately classifying and interpreting the intricate and multifaceted information embedded within single-cell datasets. RegFormer stands out as a leading model in this domain, offering superior performance across diverse human tissues compared to other state-of-the-art models such as scGPT, GeneFormer, scFoundation, and scBERT. This enhanced capability is demonstrated by higher Adjusted Silhouette Width (ASW) scores and clearer clustering of cell types in Uniform Manifold Approximation and Projection (UMAP) visualizations (**Figure 2**).

In human lung microenvironment dataset (**Figure 2A**), RegFormer achieves superior classification accuracy, effectively discerning subtle cellular heterogeneity. Similarly, it excels in distinguishing hematopoietic cell types within human bone marrow development dataset (**Figure 2B**). For human dendritic cells dataset (**Figure 2C**), RegFormer demonstrates its ability to classify distinct subtypes across batches, maintaining consistency despite batch-specific variations. Furthermore, in multiple batch data from human pancreas dataset (**Figure 2D**), RegFormer sustains high performance in classifying diverse cell types, highlighting its robustness against technical variability. Moreover, graph topological sorting ablation experiments conducted on each dataset revealed that the integration of biological prior knowledge from the gene regulatory network enhances the potential of RegFormer (**Suppplementary Figure 3**).

These results underscore RegFormer's capability to provide precise and biologically interpretable predictions from complex single-cell datasets. By leveraging gene regulatory networks and innovative computational architectures, RegFormer not only enhances the interpretability of single-cell data but also facilitates deeper insights into cellular heterogeneity and function.

RegFormer Achieves Superior Performance on Cell Annotation of scRNA-seq Data

RegFormer's advanced architecture significantly enhances the embedding of single-cell RNA sequencing (scRNA-seq) datasets, potentially leading to more accurate and biologically meaningful cell annotations, especially in complex tissue environments. Comprehensive evaluations show that RegFormer outperforms

other leading models such as scGPT, GeneFormer, scFoundation, and scBERT across various human tissues (**Figure 3A and Supplementary Figure 4**). Notably, on the Zheng68K dataset, RegFormer achieves an accuracy of 0.86 and a Macro-F1 score of 0.77, marking a significant improvement over competing models. UMAP visualizations highlight RegFormer's superior ability to produce clearer and more distinct clusters on the Zheng68K dataset, effectively delineating different cell types (**Figure 3B**). Confusion matrices further illustrate RegFormer's precision in classifying cell types, with higher values along the diagonal indicating fewer misclassifications. Specifically, for subsets of human CD4+ T cells, RegFormer demonstrates high specificity and sensitivity, as evidenced by the elevated diagonal values in the confusion matrix (**Figure 3C**). This enhanced performance facilitates more reliable and nuanced cell annotation, enabling discerning subtle cellular heterogeneity that might be overlooked by other methods.

When comparing models pretrained and fine-tuned against those trained from scratch on the Z68k and M.S. datasets, RegFormer's pretraining and fine-tuning approach notably surpasses training from scratch in terms of cell annotation performance. This superiority is evident in the accuracy and Macro-F1 scores, where RegFormer excels after being pretrained and subsequently fine-tuned (**Figure 3D**). Detailed confusion matrices for the Z68k and M.S. datasets provide a granular view of RegFormer's classification performance (**Figure 3E and 3F**). The matrices show high values along the diagonal, indicating accurate classification of various cell types such as CD14+ monocytes and CD19+ B cells. The minimal off-diagonal values represent low misclassification rates, underscoring the enhanced reliability of cell annotations achieved through RegFormer's pretraining and fine-tuning methodology. This level of detail is crucial for validating the superiority of RegFormer's approach over starting training from scratch.

Collectively, RegFormer's superior embedding capabilities translate into enhanced accuracy and precision in cell annotation. By capturing the intrinsic structure of cellular populations more faithfully, RegFormer facilitates clearer clustering and more reliable classification of cell types.

RegFormer Facilitates Precise Construction of Gene Regulatory Networks

RegFormer integrates gene regulatory prior knowledge into its pretraining process, enabling the precise and efficient construction of high-resolution GRNs. By leveraging the RegFormer pretrained model to transform gene expression data into robust gene embeddings, we combined this approach with SCENIC[21] to infer cell-type-specific GRNs, effectively linking transcription factors (TFs) to their target genes

(**Figure 4A**).

When benchmarked against state-of-the-art models—including scFoundation, scBERT, scGPT, GeneFormer, and gene2vec—RegFormer demonstrated clear superiority on human lung datasets (**Figure 4B**). It achieved the highest Functional Similarity Index, a quantitative metric of functional similarity in Gene Ontology (GO) Biological Processes (BP), between genes within the gene regulatory network, underscoring its capacity to capture biologically meaningful gene regulatory roles during the embedding process (**Figure 4C**). Additionally, RegFormer excelled in identifying enriched Gene Ontology (GO) pathways across a wide range of Leiden clustering resolutions, underscoring its robustness and adaptability across diverse biological contexts (**Figure 4D**). These findings establish RegFormer as an unparalleled tool for extracting biologically relevant insights compared to competing methods.

The embeddings produced by RegFormer revealed transcriptional landscapes with remarkable clarity and granularity. Heatmaps of regulon scores and TF expression patterns further confirmed RegFormer's ability to delineate cell-type-specific regulatory programs with high precision (**Figure 4E**). To further illustrate the biological insights enabled by RegFormer, we examined the transcriptional regulation of FOSB, SOX7 and other classical TFs (**Figure 4F and Supplementary Figure 5**). In T and B cells[22-24], FOSB emerged as a key regulator with high regulon activity, governing an extensive network of target genes and highlighting its critical role in immune function (**Figure 4F, left panels**). Similarly, SOX7 exhibited strong, cell-specific activity in endothelial cells, orchestrating a detailed regulatory network[25, 26] **(Figure 4F, right panels**). These case studies exemplify RegFormer's ability to infer biologically grounded regulatory interactions with exceptional clarity and reliability.

RegFormer Accurately Predicts Gene Expression Changes Under Genetic Perturbation

The ability of RegFormer to model gene expression changes under genetic perturbations was evaluated using a framework that integrates foundational single-cell models with the GEARS[27] perturbation embedding approach (**Figure 5A**). In this framework, unperturbed gene expression states are transformed into GeneID embeddings using single-cell foundation models. Concurrently, perturbations are encoded into embeddings through the GEARS model, capturing the influence of genetic perturbations. These embeddings are integrated by RegFormer to predict perturbed gene expression levels, effectively capturing gene-specific upregulation and downregulation with high precision.

The performance of RegFormer was compared against state-of-the-art models, including GEARS, scGPT, and scFoundation, using two benchmark datasets: Adamson and Norman. On the Adamson dataset, RegFormer achieved the lowest Mean Squared Error (MSE) and the highest Pearson correlation, indicating its ability to minimize prediction error while maintaining strong agreement with experimental data (**Figure 5B**). Similarly, on the Norman dataset, RegFormer demonstrated superior performance, achieving consistent accuracy across both metrics compared to other methods (**Figure 5C**). These results highlight RegFormer's robustness and adaptability across diverse genetic perturbation datasets.

By integrating gene regulatory prior knowledge and perturbation embeddings, RegFormer effectively models complex transcriptional responses with exceptional accuracy. Its ability to outperform leading methods establishes RegFormer as a powerful tool for predicting gene expression changes under perturbative conditions, advancing the understanding of gene regulatory dynamics.

RegFormer Outperforms Current Single-Cell Foundation Models in Predicting Drug Responses.

Drug response analysis is vital for advancing personalized medicine, optimizing drug discovery, and uncovering mechanisms of action. It enables the development of tailored treatment strategies by predicting individual sensitivities to drugs, accelerates drug discovery by identifying promising compounds, and provides insights into resistance mechanisms and alternative therapeutic strategies. Computational models like RegFormer address scalability challenges in experimental methods, offering efficient and biologically interpretable solutions for large-scale drug and cell line datasets. RegFormer, now with the integration of DeepCDR[28], provides a comprehensive approach for IC50 prediction by combining drug structure information, cell line gene expression profiles, and graph convolutional networks. (**Figure 6A**).

To evaluate the performance of RegFormer in drug response prediction, we compared it against baseline models, including deepCDR, scBERT, scGPT, Geneformer, and scFoundation. RegFormer outperformed all benchmarks, achieving the highest predictive accuracy, as measured by the Pearson Correlation Coefficient (PCC) and Spearman's Rank Correlation Coefficient (SRCC) (**Figure 6B and Supplementary Figure 6**). RegFormer showed a consistent and robust improvement in PCC values over deepCDR across all tested drugs (**Figure 6C**) and cancer types (**Figure 6D**). In a case analysis of the five most accurately

predicted drugs, RegFormer demonstrated strong concordance between observed and predicted lnIC50 values.

We further assessed RegFormer's generalizability through a leave-drug-out blind test, where it consistently surpassed deepCDR across 200 drugs, with PCC values remaining above baseline for the majority of drugs (**Figure 6E**). This highlights RegFormer's strong ability to generalize across unseen drug classes.

To explore the biological relevance of the predictions, we performed gene set enrichment analysis (GSEA) for drugs with well-characterized mechanisms of action. RegFormer accurately identified key pathways, such as TNF-alpha signaling via NF-kB for Bortezomib [29]and the p53 pathway for SN-38[30], aligning with existing literature (**Figure 6F**). These findings not only demonstrate RegFormer's superior predictive performance but also its capacity to provide biologically meaningful insights into drug response mechanisms.

Discussion

The introduction of RegFormer marks a significant advancement in single-cell RNA sequencing (scRNA-seq) analysis, addressing long-standing challenges related to data complexity, sparsity, and the incorporation of biological knowledge. By integrating gene regulatory networks (GRNs) into a novel architecture built on Mamba Blocks, RegFormer enhances the biological interpretability of scRNA-seq data. It sets new benchmarks in critical analytical tasks such as cell annotation, GRN construction, genetic perturbation prediction, and drug response prediction.

A key innovation of RegFormer is its use of GRNs, enabling the modeling of hierarchical gene regulation relationships crucial for cellular function and differentiation. Unlike prior models such as scBERT and GeneFormer, which either lacked biological priors or struggled with data sparsity and long-range dependencies, RegFormer explicitly incorporates these hierarchies. This alignment with known regulatory networks ensures that the model's operations are biologically coherent, improving both accuracy and interpretability. Additionally, RegFormer employs a dual-embedding strategy, using value embeddings for gene expression levels and token embeddings for gene identities. This design enhances the model's ability to capture both the quantitative and qualitative aspects of gene expression data.

The architecture of RegFormer, especially its use of Mamba Blocks, is optimized for handling the complexities of high-dimensional and sparse scRNA-seq data. By combining multi-layer perceptrons, convolutional layers, and state-space models, Mamba Blocks capture dependencies at various scales, from local gene interactions to broader cellular states. This approach improves computational efficiency through parameter sharing while ensuring the biological relevance of the learned features. Furthermore, the model supports scalable pretraining on large datasets, as demonstrated by its use of 22 million human cells in this study, allowing it to leverage vast biological data while preserving interpretability.

Empirical evaluations showcase RegFormer's superior performance across diverse human tissues and datasets. In cell annotation tasks, RegFormer consistently outperforms state-of-the-art models, achieving higher accuracy and Macro-F1 scores, as well as producing clearer clusters in UMAP visualizations. These improvements are essential for accurately identifying cellular heterogeneity and discerning subtle cell subtypes, which are critical for understanding complex biological processes and disease mechanisms. RegFormer's ability to maintain high performance across various datasets and batch conditions further highlights its robustness and generalizability, making it a versatile tool for single-cell genomics research. In GRN construction, RegFormer excels at generating high-resolution networks that closely align with known biological pathways and transcription factor activities. Its integration with SCENIC for inferring cell-type-specific GRNs further demonstrates its potential for revealing detailed regulatory programs. For instance, RegFormer’s precise identification of key transcription factors like FOSB and BCL11A underscores its ability to capture biologically meaningful gene interactions, offering valuable insights into the regulatory mechanisms that govern cellular functions.

While RegFormer demonstrates notable advancements, several limitations remain. The computational complexity of training large-scale models, such as RegFormer, poses a challenge, particularly as the number of genes and cells grows. Additionally, although RegFormer effectively integrates GRNs, the quality and completeness of the underlying regulatory network data can influence model performance. Future research could explore the integration of dynamic and context-specific GRNs to further improve model accuracy and applicability. Furthermore, expanding RegFormer to incorporate multi-omics data could offer a more comprehensive understanding of cellular states and interactions, facilitating deeper biological interpretations.

In conclusion, RegFormer represents a major leap forward in single-cell genomics analysis by effectively integrating biological priors and utilizing advanced machine learning architectures to address key challenges in scRNA-seq data interpretation. Its exceptional performance across a variety of tasks affirms its design and opens new possibilities for biological discovery and translational research. As the field of single-cell genomics evolves, models like RegFormer will be crucial in unraveling the complexities of cellular heterogeneity and regulatory mechanisms, deepening our understanding of biology and informing future therapeutic advancements.

Reference

1. Choi, Y.H. & Kim, J.K. Dissecting Cellular Heterogeneity Using Single-Cell RNA Sequencing. *Mol Cells* **42**, 189-199 (2019).
2. Papalexi, E. & Satija, R. Single-cell RNA sequencing to explore immune cell heterogeneity. *Nat Rev Immunol* **18**, 35-45 (2018).
3. Gawad, C., Koh, W. & Quake, S.R. Single-cell genome sequencing: current state of the science. *Nat Rev Genet* **17**, 175-188 (2016).
4. He, P. et al. The changing mouse embryo transcriptome at whole tissue and single-cell resolution. *Nature* **583**, 760-767 (2020).
5. van Galen, P. et al. Single-Cell RNA-Seq Reveals AML Hierarchies Relevant to Disease Progression and Immunity. *Cell* **176**, 1265-1281 e1224 (2019).
6. Sheu, K.M., Pimplaskar, A. & Hoffmann, A. Single-cell stimulus-response gene expression trajectories reveal the stimulus specificities of dynamic responses by single macrophages. *Mol Cell* **84**, 4095-4110 e4096 (2024).
7. Regev, A. et al. The Human Cell Atlas. *Elife* **6** (2017).
8. Ryu, Y., Han, G.H., Jung, E. & Hwang, D. Integration of Single-Cell RNA-Seq Datasets: A Review of Computational Methods. *Mol Cells* **46**, 106-119 (2023).
9. Heumos, L. et al. Best practices for single-cell analysis across modalities. *Nat Rev Genet* **24**, 550-572 (2023).
10. Saliba, A.E., Westermann, A.J., Gorski, S.A. & Vogel, J. Single-cell RNA-seq: advances and future challenges. *Nucleic Acids Res* **42**, 8845-8860 (2014).
11. Cui, H. et al. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nat Methods* **21**, 1470-1480 (2024).
12. Theodoris, C.V. et al. Transfer learning enables predictions in network biology. *Nature* **618**, 616-624 (2023).
13. Hao, M. et al. Large-scale foundation model on single-cell transcriptomics. *Nature Methods* **21**, 1481-1491 (2024).
14. Yang, F. et al. scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. *Nature Machine Intelligence* **4**, 852-866 (2022).
15. Bian, H. et al. General-purpose pre-trained large cellular models for single-cell transcriptomics. *National Science Review* **11**, nwae340 (2024).
16. Ma, Q., Jiang, Y., Cheng, H. & Xu, D. Harnessing the deep learning power of foundation models in single-cell omics. *Nat Rev Mol Cell Biol* **25**, 593-594 (2024).
17. Schafer, P.S.L., Dimitrov, D., Villablanca, E.J. & Saez-Rodriguez, J. Integrating single-cell multi-omics and prior biological knowledge for a functional characterization of the immune system. *Nat Immunol* **25**, 405-417 (2024).
18. Doddahonnaiah, D. et al. A Literature-Derived Knowledge Graph Augments the Interpretation of Single Cell RNA-seq Datasets. *Genes (Basel)* **12** (2021).
19. Kim, D. et al. Gene regulatory network reconstruction: harnessing the power of single-cell multi-omic data. *NPJ Syst Biol Appl* **9**, 51 (2023).
20. Emmert-Streib, F., Dehmer, M. & Haibe-Kains, B. Gene regulatory networks and their applications: understanding biological and medical problems in terms of networks. *Front Cell Dev Biol* **2**, 38 (2014).
21. Kumar, N., Mishra, B., Athar, M. & Mukhtar, S. Inference of Gene Regulatory Network from Single-Cell Transcriptomic Data Using pySCENIC. *Methods Mol Biol* **2328**, 171-182 (2021).

22. Ohkubo, Y. et al. A role for c-fos/activator protein 1 in B lymphocyte terminal differentiation. *J Immunol* **174**, 7703-7710 (2005).
23. Mondino, A. et al. Defective transcription of the IL-2 gene is associated with impaired expression of c-Fos, FosB, and JunB in anergic T helper 1 cells. *J Immunol* **157**, 2048-2057 (1996).
24. Liu, H. et al. Functional studies of BCL11A: characterization of the conserved BCL11A-XL splice variant and its interaction with BCL6 in nuclear paraspeckles of germinal center B cells. *Mol Cancer* **5**, 18 (2006).
25. Klomp, J. et al. Comprehensive transcriptomic profiling reveals SOX7 as an early regulator of angiogenesis in hypoxic human endothelial cells. *J Biol Chem* **295**, 4796-4808 (2020).
26. Yao, Y., Yao, J. & Bostrom, K.I. SOX Transcription Factors in Endothelial Differentiation and Endothelial-Mesenchymal Transitions. *Front Cardiovasc Med* **6**, 30 (2019).
27. Roohani, Y., Huang, K. & Leskovec, J. Predicting transcriptional outcomes of novel multigene perturbations with GEARS. *Nature Biotechnology* **42**, 927-935 (2024).
28. Liu, Q., Hu, Z., Jiang, R. & Zhou, M. DeepCDR: a hybrid graph convolutional network for predicting cancer drug response. *Bioinformatics* **36**, i911-i918 (2020).
29. Sung, M.H. et al. Dynamic effect of bortezomib on nuclear factor-kappaB activity and gene expression in tumor cells. *Mol Pharmacol* **74**, 1215-1222 (2008).
30. Liu, Y. et al. Inhibition of Akt signaling by SN-38 induces apoptosis in cervical cancer. *Cancer Lett* **274**, 47-53 (2009).

Method

RegFormer

In comparison with Transformer, which is commonly used as the backbone model in the single-cell domain, Mamba features a completely different model in architecture design. In this section, we will delve into the challenges encountered when deploying Mamba in the single-cell domain, along with our proposed recipe.

scRNA-seq Modeling

Single-cell RNA sequencing data is typically represented as a cell-gene matrix $X \in R^{N \times G}$, where each element x_{ij} reflects the raw expression level of gene j in cell i. To simplify downstream analysis, raw expression values are binned into B intervals, and x_{ij} is assigned to bin k if it falls within the corresponding range or set to 0 if the gene is not expressed. This process allows the expression profile of a cell i to be modeled as a sequence $X_i = [x_{i1}, x_{i2}, \dots, x_{iG}]$, with the corresponding gene sequence represented as $G_i = [g_{i1}, g_{i2}, \dots g_{iG}]$.

State Space Model

State Space Models (SSMs) process sequential data by mapping 1-D input sequences $x(t)$ to outputs $y(t)$ via a hidden state $h(t)$. The model's evolution is determined by parameters A, B, and C, which govern state transitions, input influence, and output projection, respectively. For scRNA-seq data, SSMs are discretized using the Zero-order hold technique, leading to the formulation $h_t = \bar{A} h_{t-1} + \bar{B} x_t$ and

$y_t = \bar{C}h_t$. This enables a sequence-to-sequence transformation $x_1, x_2, \ldots, x_t \rightarrow y_1, y_2, \ldots y_t$, making SSMs well-suited for capturing temporal dynamics in gene expression data.

Sorting

The majority of current sequence models require token ordering that allows the model to take advantage of the provided positional information as much as possible, and so does Mamba. However, compared to sequences, scRNA data is more like a set without order. Thus, defining the order of scRNA data has become a major challenge. To bridge this gap, we propose an effective ordering scheme for scRNA sequences via a knowledge base.

Graph Construction: At the first stage, we retrieve the regulatory graph of genes expressed in certain cells from the knowledge base. To perform topological sorting, we removed all cycles (approximately 2% of edges in the graph) from the graph previously. Let's assign this regulatory DAG from the knowledge base as a very big graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ denotes the node set containing all the genes while $\mathcal{E}$ represents all the regulatory relationships among $\mathcal{V}$.

At the second stage, the proposed sorting algorithm will filter out the zero-expression genes. Specifically, suppose an input cell i is composed of an expression values sequence $X_i = [x_{i1}, x_{i2}, \ldots, x_{iG}]$ and a gene sequence $G_i = [g_{i1}, g_{i2}, \ldots, g_{iG}]$. Only the genes with non-zero expression values will be retained; hence the remaining genes and expression values can be formulated as:

$$\mathrm{G'_i} = \left[\mathrm{g'_{i1}}, \mathrm{g'_{i2}}, \ldots, \mathrm{g'_{ij}}\right], \quad \mathrm{X'_i} = \left[\mathrm{x'_{i1}}, \mathrm{x'_{i2}}, \ldots, \mathrm{x'_{ij}}\right], \qquad \text{for } \forall \mathrm{j}: \mathrm{x_{ij}} \neq 0$$

Further, we map the remaining non-zero expression genes to each node in the regulatory DAG $\mathcal{G}(\mathcal{V}, \mathcal{E})$ and extract the sub-graph. In a more formal formulation: for each cell (the cell index i has been omitted for simplicity here), we map the remaining genes into node set $\mathcal{V}' = v \in \mathcal{V} | v \in G'$. Then, we assign the edge set among the remaining gene nodes $\mathcal{V}'$ as $\mathcal{E}' = (u, v) \in \mathcal{E} | u, v \in \mathcal{V}'$. Finally, a sub-graph $\mathcal{G}' = (\mathcal{V}', \mathcal{E}')$ denoting regulatory relationships within one cell has been constructed.

Topological Sorting: Topological sorting is achieved through a systematic procedure that organizes the vertices of a directed acyclic graph (DAG) in a linear ordering. In the scenario of single-cell sequences, the topological sorting algorithm guarantees that the resulting ordering satisfies the prerequisite relationships between genes, making it particularly useful for capturing the regulatory structure of expressed genes in a cell.

We illustrated the detailed flow of topological sorting in Figure 2(a). We sort the gene sequence G' through the topological sorting of the corresponding sub-graph $\mathcal{G}'(\mathcal{V}', \mathcal{E}')$. The resultant sorted gene sequence can be formulated as:

$$\mathrm{G}'' = \mathrm{TopoSort}\big(\mathcal{G}'(\mathcal{V}', \varepsilon')\big) = \left[\mathrm{g}''_1, \mathrm{g}''_2, \ldots, \mathrm{g}''_\mathrm{j}\right]$$

Correspondingly, the expression value sequence will be rearranged into $X'' = \left[x''_1, x''_2, \ldots, x''_j\right]$.

Pretraining

Recipe for generative training: In the field of natural language processing, the objective of generative pre-training (e.g., GPT and SSM) is to mimic the pattern of human speech, where a complete sentence is gradually constructed one word at a time:

$$\mathrm{P}(\mathrm{Tokent{+}1} | \mathrm{Token1{:}t}) = \mathrm{GenerativeModel}(\mathrm{Token}_{1:\mathrm{t}}), \quad \text{for } \mathrm{t} \in 1,2,\ldots,\mathrm{j}$$

However, in single-cell RNA sequencing (scRNA-seq), there is no predetermined order among different genes, which poses great challenges to their analysis and study. Thanks to our proposed sorting algorithm, gene sequences can now reflect their roles in the regulatory system of the cell. More specifically, the genes in the front part of the sequence can globally regulate the genes in the rear part of the sequence. This sequential nature paves the way for the deployment of generative models.

As illustrated in Figure 2(c), we use two pretraining schemes (Expression Value Prediction and Regulatory Role Prediction) for pretraining.

Expression Value Prediction: In the field of natural language processing, the masked token prediction task holds a significant position, as it enables the capture of semantic information and relationships among vocabulary in sentences. In pre-training for natural language processing, models such as BERT and RoBERTa extensively employ the Masked Language Model (MLM) task. The core concept of this task involves randomly masking some vocabulary in the input sentences, after which the model is required to predict what these masked-out words are.

In the field of single-cell research, a similar paradigm holds true. The masked value prediction task plays a pivotal role, facilitating the extraction of semantic information and relationships among features in single-cell data. In pre-training methodologies for single-cell analysis, such as those utilizing autoencoders or variational autoencoders, the concept of masked value prediction is widely applied. This task involves concealing certain features in the input single-cell data and tasking the model with predicting the masked-out values.

Specifically, for a cell with expression values set X_i, we randomly select a subset of values and replace them with a special [MASK] token. During training, the model aims to minimize the discrepancy between the predicted value $\widehat{X_i}$ and the target value X_i:

$$\widehat{X}_i = \mathrm{Mamba}\big(\mathrm{Mask}(X_i)\big)$$
$$\mathcal{L}_i = \frac{1}{\mathcal{M}} \sum j \in \mathcal{M} \mathrm{Loss}\big(\widehat{X}_{ij}, \widehat{Y}_{ij}\big)$$

where Loss is a suitable loss function, such as cross-entropy loss, calculated between the predicted and ground truth tokens.

Regulatory Role Prediction: We have sorted the gene sequences, where different positions in the sequence reflect the diverse roles genes play in the regulatory system. To enable the model to learn these insights, we adopted a common approach in natural language processing known as "generative training" to optimize RegFormer.

Different from self-attention, the scan-operation in Mamba doesn't need a causal mask when doing generative training or inference due to the recurrent nature of SSM. In each iteration, RegFormer predicts the next gene token $Gene_{t+1}$ *using previous gene tokens* $Gene_{1:t}$ and, in turn, uses $Gene_{1:t+1}$ *to predict* $Gene_{t+2}$. Hence, the equation for generative pretraining can be rewritten as:

$$\mathrm{P}(g_{t+1}|g_{1:t}) = \mathrm{Mamba}(g_{1:t}), \qquad \text{for} \quad t \in 1,2,\ldots,j$$

$$\mathrm{L} = -\frac{1}{T}\sum_{t=1}^{T} \log\big(p(g_{t+1}|g_1, g_2, \ldots, g_t)\big)$$

where $p(g_{t+1}|g_1, g_2, \ldots, g_t)$represents the predicted probability distribution over tokens for position $(t+1)$ given the previous input gene sequence $(g_1, g_2, \ldots, g_t)$. The goal during training is to minimize this cross-entropy loss, typically using optimization algorithms like stochastic gradient descent (SGD) or variants such as Adam.

Evaluation of Downstream Tasks

Cell Embedding

To accommodate the specific requirements of each foundational model, the input data underwent tailored preprocessing. For Geneformer and scGPT, which support sequence lengths of 2048 and 1200 respectively, a selection of 3000 highly variable genes was made as input features. In contrast, other models operated

directly on the complete gene sequences, bypassing any feature selection. This preprocessing was carefully designed to align with the pretraining protocols of each model, ensuring optimal performance.

For models like scBERT, scFoundation, and scGPT, gene expression data was transformed using a log1p operation, while Geneformer processed raw count data without any normalization. In generating cell embeddings, RegFormer offers three pooling strategies for scBERT: CLS, mean, and sum. The CLS approach uses the embedding of the CLS token to represent a cell, while mean and sum pooling compute the average or total of token embeddings generated by the encoder, respectively.

For scGPT, cell embeddings are derived directly from the CLS token's embedding. In the case of Geneformer, the native mean_nonpadding_embs function averages the embeddings of non-padded tokens. scFoundation, on the other hand, applies a max pooling operation to the token embeddings to generate the final cell embeddings.

This individualized preprocessing and embedding generation strategy takes full advantage of each model's strengths, ensuring accurate cellular data representation for downstream tasks.

Cell Type Annotation

Cell type annotation was performed in both intra-dataset and inter-dataset contexts to evaluate the foundational models' effectiveness. Each dataset was split into training and testing sets with an 80:20 ratio to ensure robust performance evaluation. Furthermore, the training data was divided into training and validation subsets with the same 8:2 split.

During preprocessing, the full gene expression profiles of all datasets were retained, and no highly variable genes were selected. This ensured that the models could leverage the entire spectrum of gene expression data, enabling them to perform well on comprehensive datasets.

In training the models, two crucial hyperparameters—number of epochs and learning rate (lr)—were fine-tuned. To maintain consistency, all models were trained for 20 epochs. The learning rates were set according to the default recommendations for each model: 0.0001 for scGPT and scFoundation, 0.001 for scBERT, and 0.00005 for Geneformer.

To assess how hyperparameters impacted cell type annotation, the Zheng68k dataset was used. This allowed for an in-depth exploration of how variations in epoch count, learning rates, and the inclusion of highly variable genes affected annotation accuracy. The results helped identify the most optimal configurations for precise cell type classification.

Gene Regulatory Network Analysis

For Gene Regulatory Network (GRN) analysis, the human newborn lung dataset was processed using foundational models integrated within the RegFormer framework. The data underwent quality control measures: genes with fewer than three counts were excluded, and the dataset was narrowed to include only the top 1200 most variable genes. After preprocessing, the data was passed through the models to generate gene embeddings, which were then used to calculate Euclidean distances, resulting in an adjacency matrix. This matrix captured the relationships between genes, with each entry reflecting the similarity between a gene pair, thereby uncovering potential regulatory connections. The adjacency matrix was used to construct the GRN, where genes acted as nodes, and regulatory interactions were represented as edges between them.

genetic perturbation

In the context of single-cell foundational models, each gene undergoes representation learning, allowing for the retrieval of its corresponding embedding via the gene identifier. For the genetic perturbation prediction task, two datasets from GEARS, Norman and Adamson, were utilized. GEARS encodes gene representations by constructing a Genetic Relationship Graph based on co-expression information.

During the experiments, the initial feature vectors of the graph nodes were replaced with gene embeddings derived from single-cell foundational models using gene IDs. The GEARS model parameters were subsequently fine-tuned, maintaining consistent hyperparameter settings across all single-cell foundational models. Specifically, the number of training epochs was set to 20, and the batch size was set to 32. Additionally, the dimensionality of gene features was matched to the embedding dimensions of the foundational models to ensure uniformity and comparability of input representations.

Drug Response

To evaluate the predictive performance of the five foundational single-cell models on drug sensitivity, two datasets from DeepCDR were utilized. The Genomics of Drug Sensitivity in Cancer (GDSC) dataset

provides IC50 values for drug-cell interactions, while the Cancer Cell Line Encyclopedia (CCLE) dataset offers detailed gene expression, mutation, and methylation profiles for various cancer cell lines.

DeepCDR employs two primary modules for feature extraction. The first module is a Graph Convolutional Network (GCN) that processes drug features and adjacency matrices. Concurrently, three specialized sub-networks extract latent biological features from the gene expression, mutation, and methylation data of cancer cell lines. These outputs are integrated and passed through a Convolutional Neural Network (CNN) to predict IC50 values.

When integrating single-cell foundational models into DeepCDR, the gene expression data from CCLE was processed by models such as RegFormer to derive cell-specific embeddings. These embeddings were incorporated into DeepCDR's sub-networks, enhancing the input quality for IC50 predictions by the CNN.

To assess the models' performance, a leave-one-drug-out blind test was conducted. The Pearson Correlation Coefficient (PCC) were used to evaluate the alignment between the predicted IC50 values and the actual observed values. This evaluation method provided valuable insights into each model's predictive accuracy regarding drug responses.

Evaluation Metrics

Cell Embedding Metrics

To assess the quality of cell embeddings, we focused on the clustering of cell types and the presence of batch effects within the embedding space. The primary evaluation metric used was the Average Silhouette Width (ASW), which measures the balance between intra-cluster cohesion and inter-cluster separation. ASW scores were computed for both cell types and batch effects using a modified approach from scBI.

For cell type clusters, ASW scores range from 0 to 1. A value of 0 indicates that the clusters are tightly grouped, 0.5 suggests partial overlap between clusters, and 1 signifies well-separated clusters. Higher ASW values are indicative of better-defined cell clusters, demonstrating more accurate model performance. The formula for ASW calculation in the context of cell types is expressed as follows:

$$ASW_{celltype} = \frac{ASW_C + 1}{2}$$

where C represents the set of cell type labels.

For evaluating batch effects, the ASW was computed based on batch labels. A score of 0 implies optimal mixing between batches, while deviations from 0 point to the presence of batch effects. To ensure that higher scores reflect improved mixing, the ASW values were inverted by subtracting them from 1, giving a final scale between 0 and 1. A score of 0 corresponds to complete batch separation, while a score of 1 suggests perfect batch integration.

Cell Type Annotation Metrics

The effectiveness of the cell type annotation process was assessed by comparing predicted labels against true labels, using four common evaluation metrics: accuracy, precision, recall, and macro F1 score. To address class imbalances, the macro F1 score was calculated in two stages. First, the F1 score was computed for each class as the harmonic mean of precision and recall:

$$\mathrm{F1_i} = 2 \times \frac{\mathrm{Precision_i} \times \mathrm{Recall_i}}{\mathrm{Precision_i} + \mathrm{Recall_i}}$$

where $F1_i$ represents the F1 score for a specific class, and $Precision_i$ and $Recall_i$ refer to the precision and recall for that class i. Next, the macro F1 score is computed by averaging the F1 scores across all n classes:

$$\mathrm{MacroF1} = \frac{1}{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{F1_i}$$

Gene Regulatory Network Metrics

To evaluate the performance of foundational models in constructing Gene Regulatory Networks (GRNs), we processed the single-cell data from human newborn lungs. The data was preprocessed according to the specific requirements of each model to ensure efficient embedding generation. These gene embeddings were then used to build a neighborhood graph representing the connectivity between gene. For the genes involved in the neighborhood graph, a Functional Similarity Index was calculated to measure the average degree of functional similarity in Gene Ontology (GO) Biological Processes (BP) between genes within the gene regulatory network. The Functional Similarity Index is defined as:

$$\text{Functional Similarity Index} = \frac{1}{|\mathrm{E(G)}|} \sum_{(\mathrm{u,v}) \in \mathrm{E(G)}} \mathrm{J(u,v)}$$

Where $|E(G)|$ represents the total number of edges in the network, and $\mathrm{J}(\mathrm{u},\mathrm{v})$ is the Jaccard Index of the gene nodes u and v connected by an edge. The Jaccard Index is a metric that quantifies the similarity between sets. In this context, $\mathrm{J}(\mathrm{u},\mathrm{v})$ measures the functional similarity between gene pairs u and v and is defined as:

$$\mathrm{J}(\mathrm{u},\mathrm{v}) = \frac{|\mathrm{S_u} \cap \mathrm{S_v}|}{|\mathrm{S_u} \cup \mathrm{S_v}|}$$

where $|S_u \cap S_v|$ is the intersection size of the sets of functional items enriched by genes u and v, and $|S_u \cup S_v|$ is the size of their union.

The Leiden clustering algorithm was used to identify gene groups based on their connectivity patterns. By adjusting the resolution parameter from 0.1 to 0.6 in steps of 0.1, we explored clustering outcomes at varying levels of granularity. This approach helped reveal distinct subgroups, offering valuable insights into cellular heterogeneity and diversity.

Clusters with fewer than 10 genes were excluded from further analysis, and Gene Ontology (GO) enrichment analysis was performed on the remaining clusters. This analysis provided an understanding of the biological processes associated with each cluster. Categories with adjusted P-values below 0.0001 were retained, providing a meaningful biological interpretation of the clusters.

Genetic Perturbation Metrics

The definitions and evaluation metrics from the GEARS studies were adopted to ensure consistency in benchmarking. Model performance was assessed by calculating the mean squared error (MSE) and Pearson correlation between predicted and experimentally observed differential gene expression values.

The mean squared error (MSE) is calculated using the following equation:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

where:

y_i represents the experimentally observed gene expression value for the i-th gene,

$\hat{y}_i$ is the predicted gene expression value for the i-th gene,

n is the total number of genes evaluated.

The Pearson correlation coefficient (r) is computed as:

$$r = \frac{\sum_{i=1}^{n}(y_i - \overline{y})(\hat{y}_i - \overline{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}\sqrt{\sum_{i=1}^{n}(\hat{y}_i - \overline{\hat{y}})^2}}$$

where:

$\overline{y}$ is the mean of the experimentally observed values,

$\overline{\hat{y}}$ is the mean of the predicted values.

Drug Response Metrics

To assess the accuracy of drug sensitivity predictions, we used preprocessed data from DeepCDR and evaluated it using two primary metrics: Pearson Correlation Coefficient (PCC) and Spearman's Rank Correlation Coefficient (SRCC). These metrics measure the correlation between predicted and actual IC50 values for each cell line, providing a comprehensive evaluation of the foundational models' predictive performance. The formula for the Pearson Correlation Coefficient is:

$$\mathrm{PCC}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

where x_i and y_i are the predicted and actual IC50 values for the i-th drug-cell pair, and $\bar{x}$ and $\bar{y}$ represent the mean values of the predicted and observed IC50 values, respectively. The Spearman Rank Correlation Coefficient is calculated as:

$$\mathrm{SRCC}=1-\frac{6\sum_{i=1}^{n}d_i^2}{n(n^2-1)}$$

Where d_i is the rank difference between the predicted and observed IC50 values for the i-th drug-cell pair, and n is the total number of pairs analyzed.

Data availability

All data used in the research were collected from published sources, with detailed descriptions provided in Supplementary Materials.

Code availability

RegFormer is packaged, and distributed as an open-source, publicly available repository at https://github.com/BGIResearch/RegFormer.

Funding

This work is supported by the National Natural Science Foundation for Young Scholars of China (32300526) and National Key R&D Program of China (2022YFC3400400).

Acknowledgement

We acknowledge the Stomics Cloud platform (https://cloud.stomics.tech/) for providing GPU computational resources. We thank the colleagues in our research group for inspiring discussion and their contributions.

Author contributions

Y.Z., Y.L. and S.F. conceptualized the study. L.H. and P.Q. were responsible for the model design and tool implementation. L.H., H.Q. and P.Q. performed data analysis and model evaluation. S.F., Y.L., W.J., T.X. and L.C. contributed key ideas and advice. L.H., wrote the manuscript. Y.L., S.F. and Y.Z. supervised the study.

Figure Legends

Figure 1. RegFormer Framework for Single-Cell RNA Sequencing Data Analysis.

(A) Overview of the RegFormer framework, comprising Pre-training Data (22 million human cells), RegFormer Pre-training Module (Graph Topological Sorting and Pretraining Scheme), Task-specific Data (cell-gene matrices and labels), and Downstream Tasks (Data Integration, Gene Regulatory Network Analysis, Cell Type Annotation, Perturbation Prediction, and Drug Response Prediction).

(B) Extraction of gene regulatory networks from a knowledge base to construct a directed acyclic graph (DAG). Genes with non-zero expression in each cell are filtered to form a subgraph.

(C) Dual pretraining strategy including Expression Value Prediction, which masks and reconstructs gene expression values, and Regulatory Role Prediction, which predicts the next gene in the topologically sorted sequence to capture regulatory relationships.

(D) Mamba Block architecture integrating Multi-Layer Perceptrons (MLP), Convolutional Layers (Conv), and State Space Models (SSM) to efficiently capture dependencies in gene expression data.

Figure 2. ASW and UMAP Visualizations across Different Datasets.

(A) ASW scores and UMAP projection of cell types in the human lung dataset.

(B) ASW scores and UMAP projection of cell types in the human bone marrow dataset.

(C) ASW scores and UMAP projection showing cell type and batch information in the human DC dataset.

(D) ASW scores and UMAP projection showing cell type and batch information in the human pancreas dataset.

Figure 3. Evaluation of Cell Annotation Performance.

(A) Annotation accuracy and Macro-F1 scores across multiple datasets for foundational models.

(B) UMAP projections showing cell clustering results for different models.

(C) Confusion matrices illustrating cell type classification performance across various models.

(D) Comparison of accuracy and Macro-F1 scores between pretraining + fine-tuning and from-scratch training approaches.

(E) Confusion matrix and classification performance for specific datasets, highlighting model performance variations.

Figure 4. Evaluation of Gene Regulatory Network Analysis.

(A) Workflow for constructing Gene Regulatory Networks (GRNs) using gene expression data, Mamba pretrained models, and SCENIC. Gene embeddings are used to infer transcription factor (TF) interactions and target genes.

(B) UMAP visualization of cell types in the human lung dataset, showing distinct clustering of cell populations based on cell type annotations.

(C) Biological index scores across foundational models, reflecting their ability to capture biologically relevant features.

(D) Number of enriched Gene Ontology (GO) pathways identified at different Leiden clustering resolutions, highlighting variation in pathway enrichment.

(E) Heatmaps showing regulon scores (left) and transcription factor (TF) expression (right) for various cell types in the human lung dataset.

(F) UMAP plots of selected regulons (FOSB and) and their corresponding TF expression levels, with inferred regulatory networks illustrating TF-target relationships.

Figure 5. Evaluation of Modeling Genetic Perturbation Effects.

(A) Workflow illustrating the prediction of perturbed gene expression states. Gene embeddings from unperturbed states and perturbation embeddings are integrated using the Gears model to predict the perturbed gene expression levels.

(B,) Model performance evaluated on the Adamson datasets.

(C) Model performance evaluated on the Norman datasets.

Figure 6. Evaluation of Drug Sensitivity Prediction.

(A) Workflow of DeepCDR integrated with foundational models for drug sensitivity prediction. Drug structure and gene expression data are processed through graph convolutional networks and subnetworks, followed by IC50 prediction.

(B) Pearson Correlation Coefficient (PCC) and Spearman Rank Correlation Coefficient (SRCC) for various models on drug sensitivity prediction.

(C) Correlation of PCC values across all drugs.

(D) Correlation of PCC values across cancer types.

(E) Leave-drug-out blind test evaluating model generalization. Positive PCC values indicate better predictions compared to baseline.

(F) Gene set enrichment analysis for TNF-alpha signaling via NF-kB and p53 pathways, showing enrichment scores and ranked gene metrics.

Supplementary Figure 1. Overview of RegFormer Pretraining Data

(A) Distribution of cell numbers across different tissue categories.

(B) Distribution of cell numbers across different technical platforms.

(C) Distribution of donors across human developmental stages.

(D) Distribution of cell numbers across different disease states.

Supplementary Figure 2. Gene Regulation Network Biological Prior Knowledge

(A) Log-transformed number of nodes and edges in the network.

(B) Proportion of transcription factors (TFs) and target genes.

(C) Distribution of TF depths in the network.

(D) Transcription factor with maximum depth and its corresponding regulatory path.

(E) Top 10 transcription factors with the most descendant nodes.

(F) Network diagram of the top 10 transcription factors and their descendant nodes.

Supplementary Figure 3. Ablation of the Graph Topological Sorting in Single-cell Representation

(A) ASW scores for cell type in the human lung dataset.

(B) UMAP projection of cell type in human lung dataset.

(C) ASW scores for cell type classification in the human bone marrow dataset.

(D) UMAP projection of cell type in human bone marrow dataset.

(E) ASW scores for cell type and batch information in the human DC dataset.

(F) UMAP projection of cell type in human DC dataset.

(G) ASW scores for cell type and batch information in the human pancreas dataset.

(H) UMAP projection of cell type in human pancreas dataset.

Supplementary Figure 4. Evaluation of Cell Annotation Performance.

UMAP projections are shown for cell clustering results, accompanied by confusion matrices illustrating cell type classification performance in several datasets: (A) blood, (B) small intestine (SI), (C) mouse (ms), (D) kidney, and (E) liver.

Supplementary Figure 5.. Evaluation of Gene Regulatory Network Analysis.

UMAP plots are presented for selected regulons: (A) SPI1, (B) TBX21, (C) FOXA2, and (D) NR4A1. Each plot shows the corresponding transcription factor (TF) expression levels, with inferred regulatory networks illustrating the TF-target relationships for each regulon.

Supplementary Figure 6.. Evaluation of Drug Sensitivity Prediction.

(A) Boxplots of Pearson Correlation Coefficient (PCC) and Spearman Rank Correlation Coefficient (SRCC) for various models across all drugs.

(B) Boxplots of Pearson Correlation Coefficient (PCC) and Spearman Rank Correlation Coefficient (SRCC) for various models across cancer types.

Figure 1

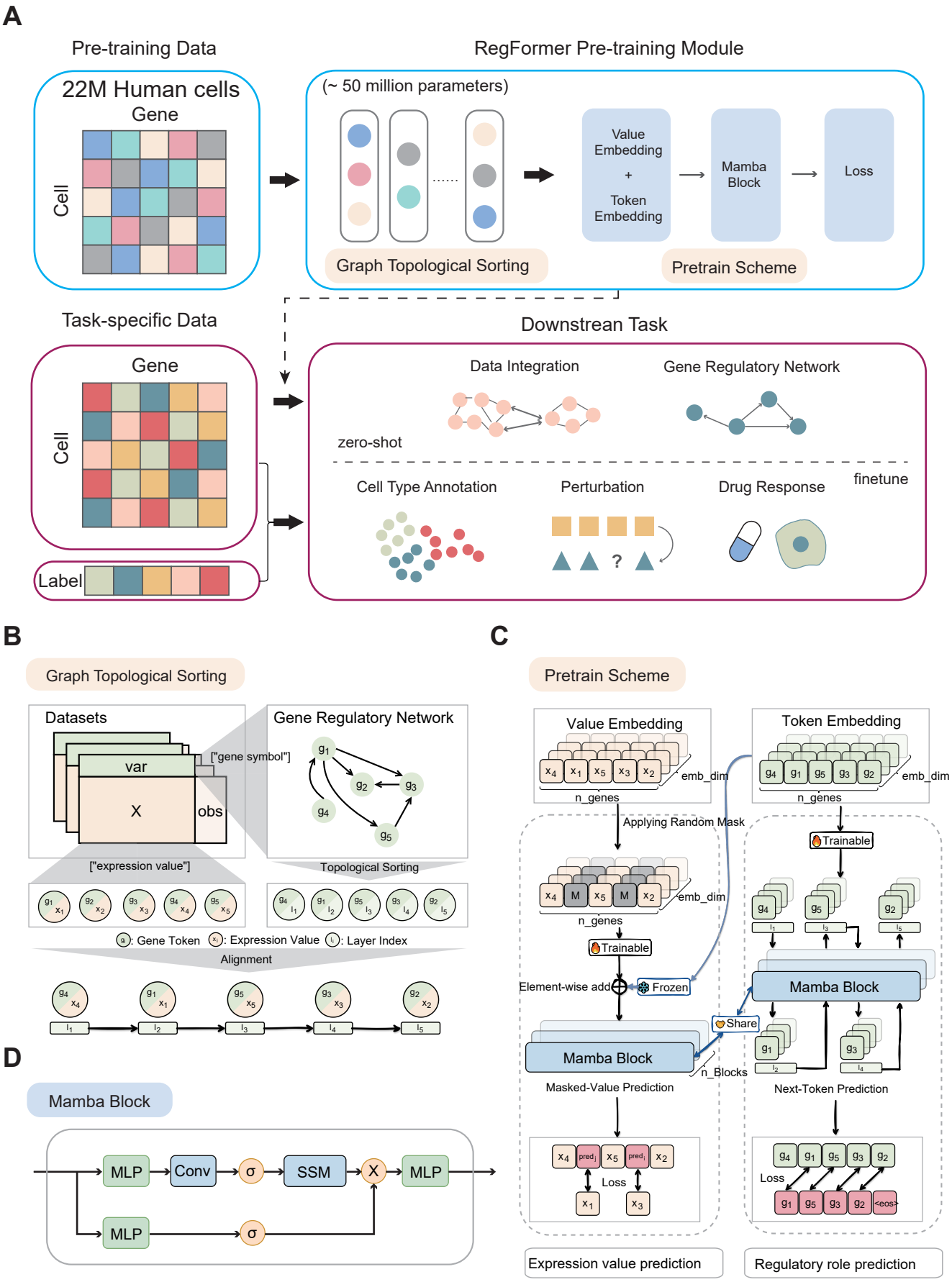

A
Pre-training Data
22M Human cells
Gene
Cell
RegFormer Pre-training Module
(~ 50 million parameters)
Value Embedding
+
Token Embedding
Mamba Block
Loss
Graph Topological Sorting
Pretrain Scheme
Task-specific Data
Gene
Cell
Label
Downstream Task
Data Integration
Gene Regulatory Network
zero-shot
finetune
Cell Type Annotation
Perturbation
Drug Response
B
Graph Topological Sorting
Datasets
var
X
obs
["gene symbol"]
["expression value"]
Gene Regulatory Network
Topological Sorting
: Gene Token
: Expression Value
: Layer Index
Alignment
C
Pretrain Scheme
Value Embedding
Token Embedding
emb_dim
n_genes
Applying Random Mask
Trainable
Element-wise add
Frozen
Share
Mamba Block
n_Blocks
Masked-Value Prediction
Next-Token Prediction
Loss
Expression value prediction
Regulatory role prediction
D
Mamba Block
MLP
Conv
σ
SSM
X
MLP
MLP
σ

Figure 2

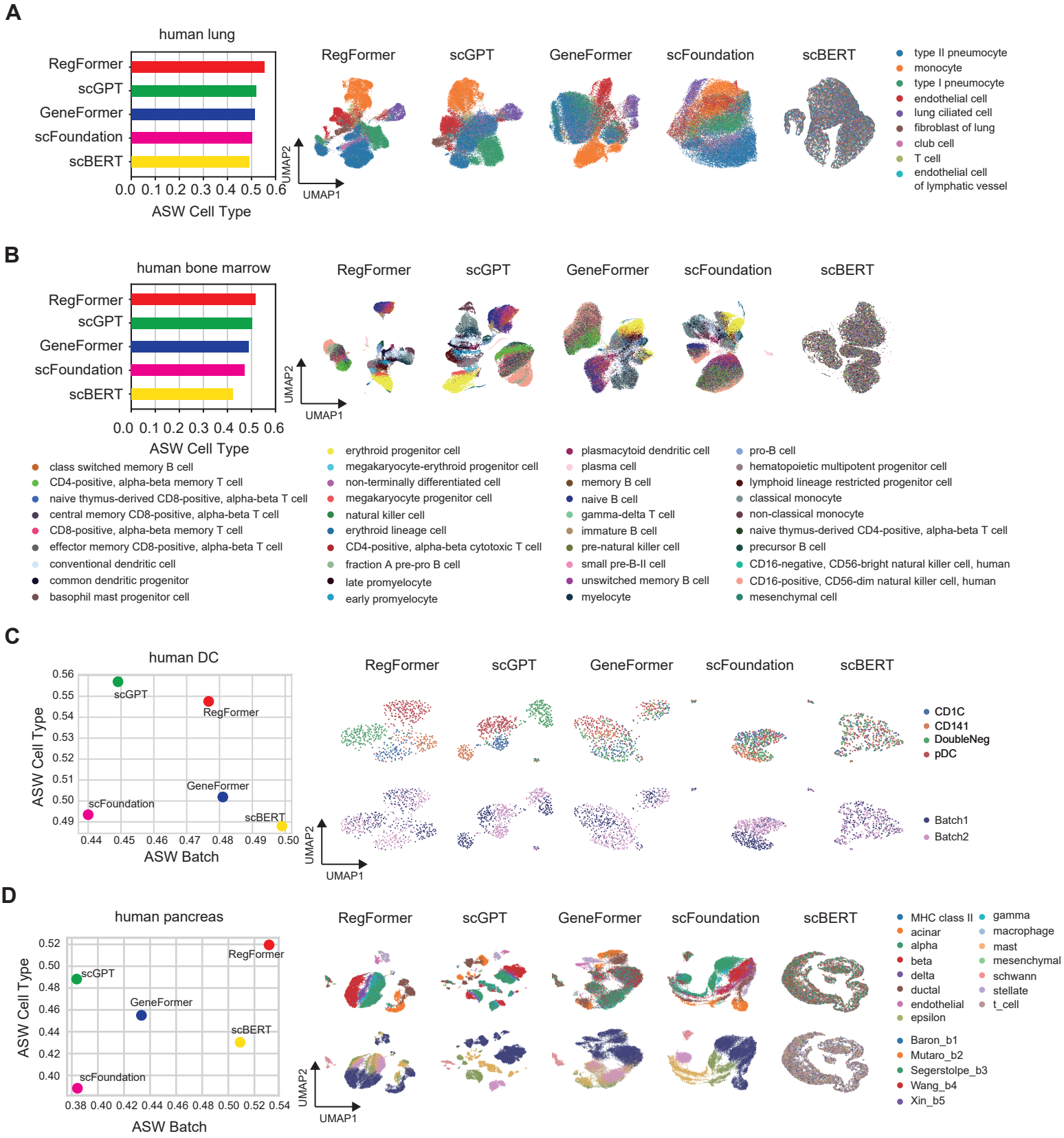
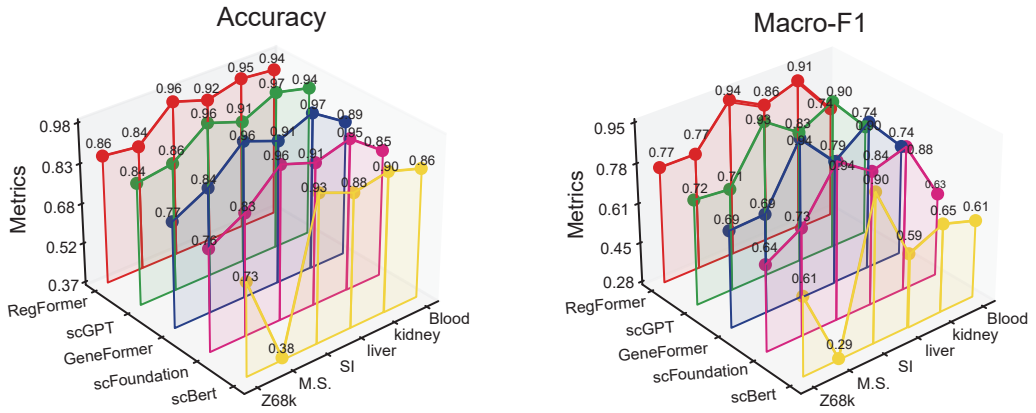
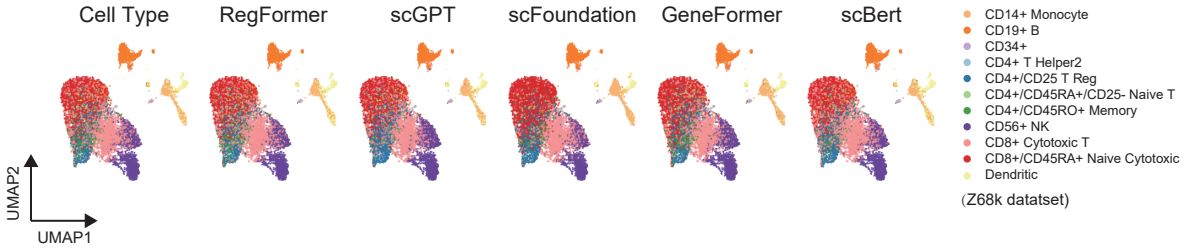

A
human lung
RegFormer
scGPT
GeneFormer
scFoundation
scBERT
0.0 0.1 0.2 0.3 0.4 0.5 0.6
ASW Cell Type
UMAP2
UMAP1
type II pneumocyte
monocyte
type I pneumocyte
endothelial cell
lung ciliated cell
fibroblast of lung
club cell
T cell
endothelial cell of lymphatic vessel
B
human bone marrow
class switched memory B cell
CD4-positive, alpha-beta memory T cell
naive thymus-derived CD8-positive, alpha-beta T cell
central memory CD8-positive, alpha-beta T cell
CD8-positive, alpha-beta memory T cell
effector memory CD8-positive, alpha-beta T cell
conventional dendritic cell
common dendritic progenitor
basophil mast progenitor cell
erythroid progenitor cell
megakaryocyte-erythroid progenitor cell
non-terminally differentiated cell
megakaryocyte progenitor cell
natural killer cell
erythroid lineage cell
CD4-positive, alpha-beta cytotoxic T cell
fraction A pre-pro B cell
late promyelocyte
early promyelocyte
plasmacytoid dendritic cell
plasma cell
memory B cell
naive B cell
gamma-delta T cell
immature B cell
pre-natural killer cell
small pre-B-II cell
unswitched memory B cell
myelocyte
pro-B cell
hematopoietic multipotent progenitor cell
lymphoid lineage restricted progenitor cell
classical monocyte
non-classical monocyte
naive thymus-derived CD4-positive, alpha-beta T cell
precursor B cell
CD16-negative, CD56-bright natural killer cell, human
CD16-positive, CD56-dim natural killer cell, human
mesenchymal cell
C
human DC
ASW Cell Type
ASW Batch
CD1C
CD141
DoubleNeg
pDC
Batch1
Batch2
D
human pancreas
MHC class II
acinar
alpha
beta
delta
ductal
endothelial
epsilon
gamma
macrophage
mast
mesenchymal
schwann
stellate
t_cell
Baron_b1
Mutaro_b2
Segerstolpe_b3
Wang_b4
Xin_b5

Figure 3

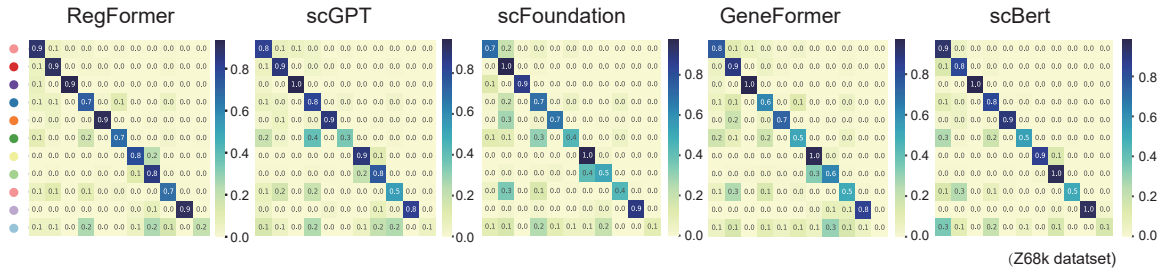
A



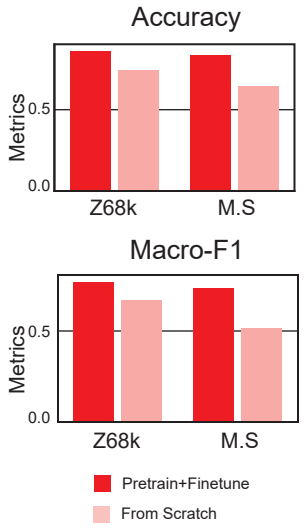
B



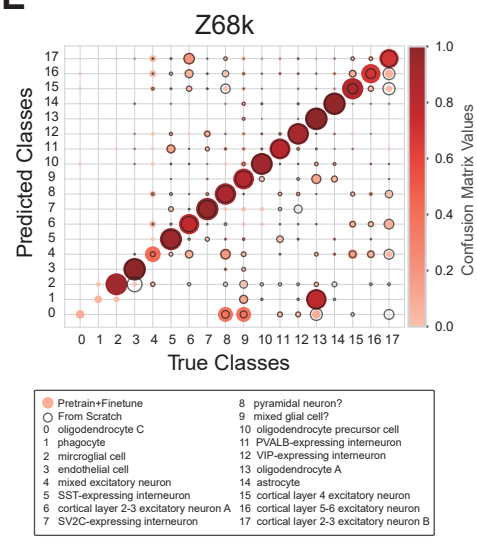
C



D



E



F

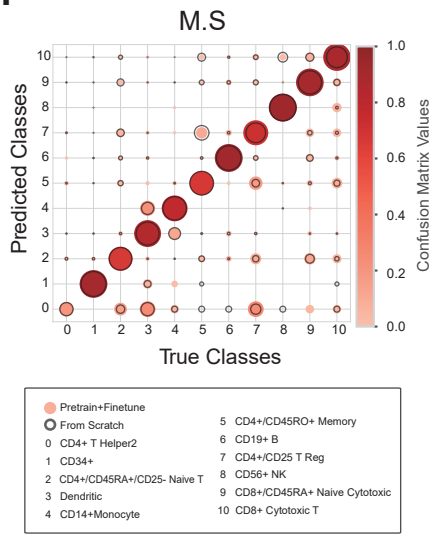


Figure 4

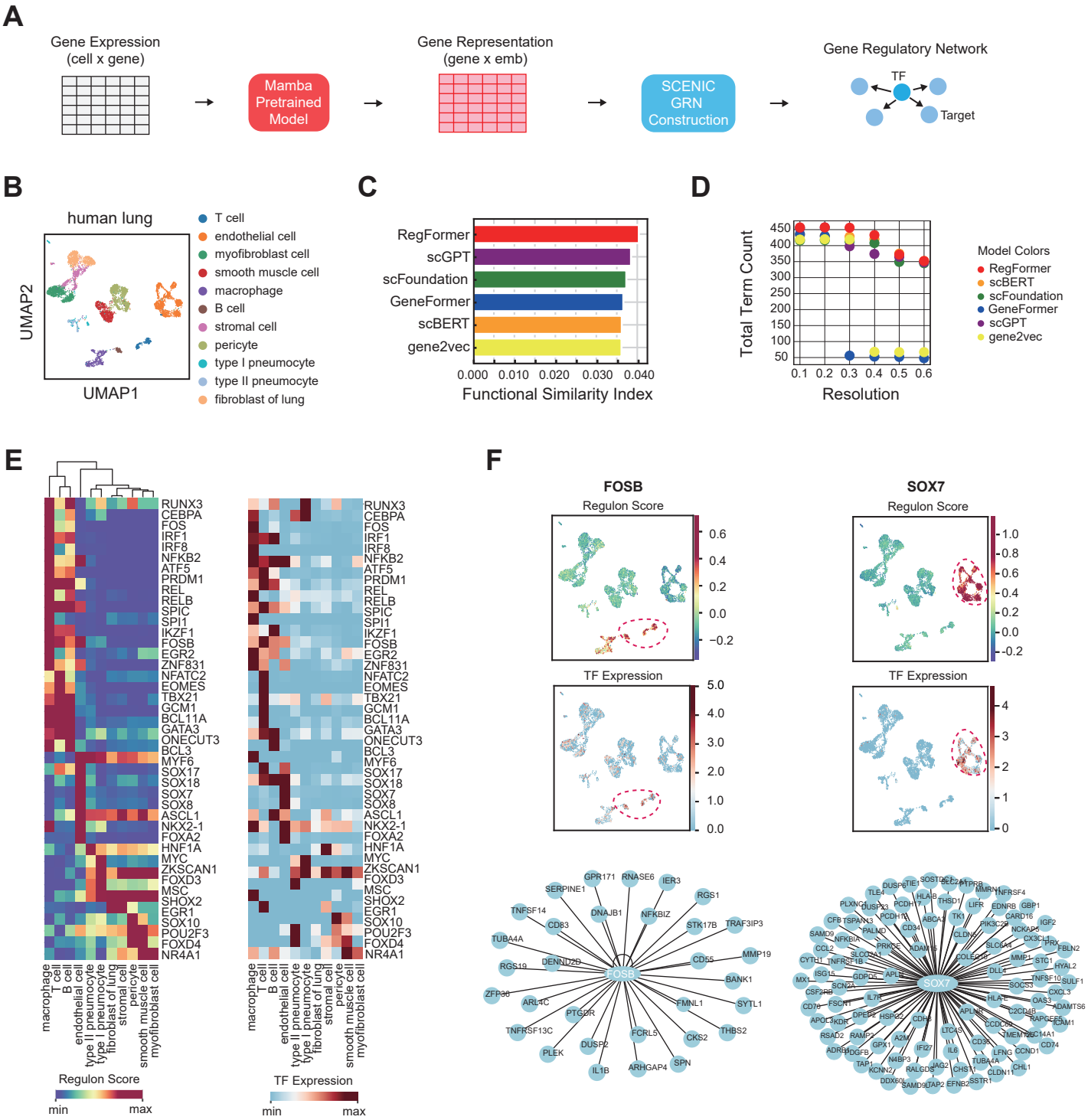

A
Gene Expression
(cell x gene)
Mamba
Pretrained
Model
Gene Representation
(gene x emb)
SCENIC
GRN
Construction
Gene Regulatory Network
TF
Target
B
human lung
UMAP2
UMAP1
T cell
endothelial cell
myofibroblast cell
smooth muscle cell
macrophage
B cell
stromal cell
pericyte
type I pneumocyte
type II pneumocyte
fibroblast of lung
C
RegFormer
scGPT
scFoundation
GeneFormer
scBERT
gene2vec
0.000 0.010 0.020 0.030 0.040
Functional Similarity Index
D
Total Term Count
450
400
350
300
250
200
150
100
50
0.1 0.2 0.3 0.4 0.5 0.6
Resolution
Model Colors
RegFormer
scBERT
scFoundation
GeneFormer
scGPT
gene2vec
E
RUNX3
CEBPA
FOS
IRF1
IRF8
NFKB2
ATF5
PRDM1
REL
RELB
SPIC
SPI1
IKZF1
FOSB
EGR2
ZNF831
NFATC2
EOMES
TBX21
GCM1
BCL11A
GATA3
ONECUT3
BCL3
MYF6
SOX17
SOX18
SOX7
SOX8
ASCL1
NKX2-1
FOXA2
HNF1A
MYC
ZKSCAN1
FOXD3
MSC
SHOX2
EGR1
SOX10
POU2F3
FOXD4
NR4A1
macrophage
T cell
B cell
endothelial cell
type II pneumocyte
type I pneumocyte
fibroblast of lung
stromal cell
pericyte
smooth muscle cell
myofibroblast cell
Regulon Score
min
max
TF Expression
min
max
F
FOSB
Regulon Score
0.6
0.4
0.2
0.0
−0.2
TF Expression
5.0
4.0
3.0
2.0
1.0
0.0
SOX7
Regulon Score
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
TF Expression
4
3
2
1
0
FOSB
GPR171
RNASE6
IER3
SERPINE1
RGS1
TNFSF14
DNAJB1
NFKBIZ
CD83
STK17B
TRAF3IP3
TUBA4A
MMP19
DENND2D
CD55
RGS19
BANK1
ZFP36
ARL4C
FMNL1
SYTL1
PTGDR
TNFRSF13C
THBS2
FCRL5
CKS2
DUSP2
PLEK
SPN
IL1B
ARHGAP4
SOX7

Figure 5

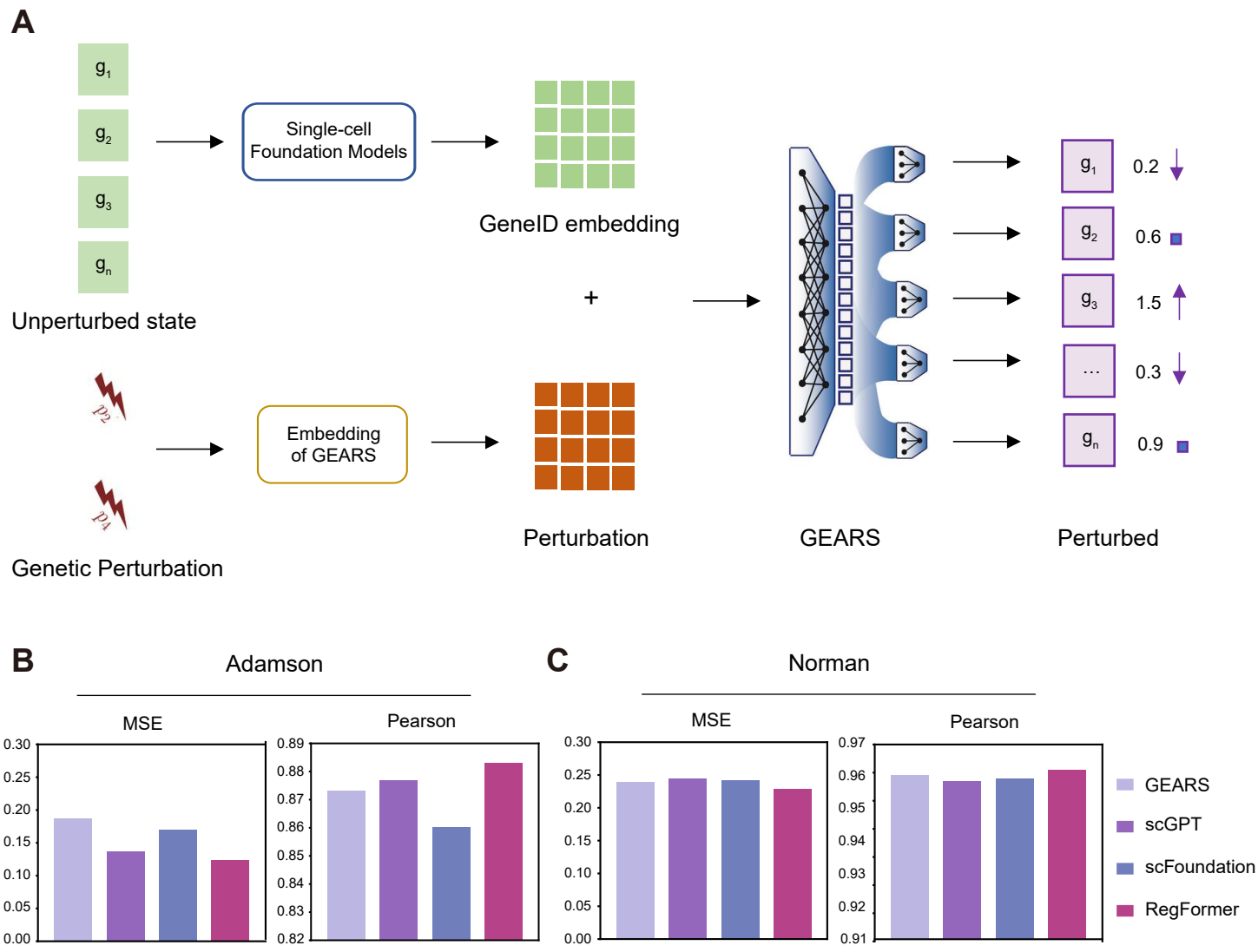
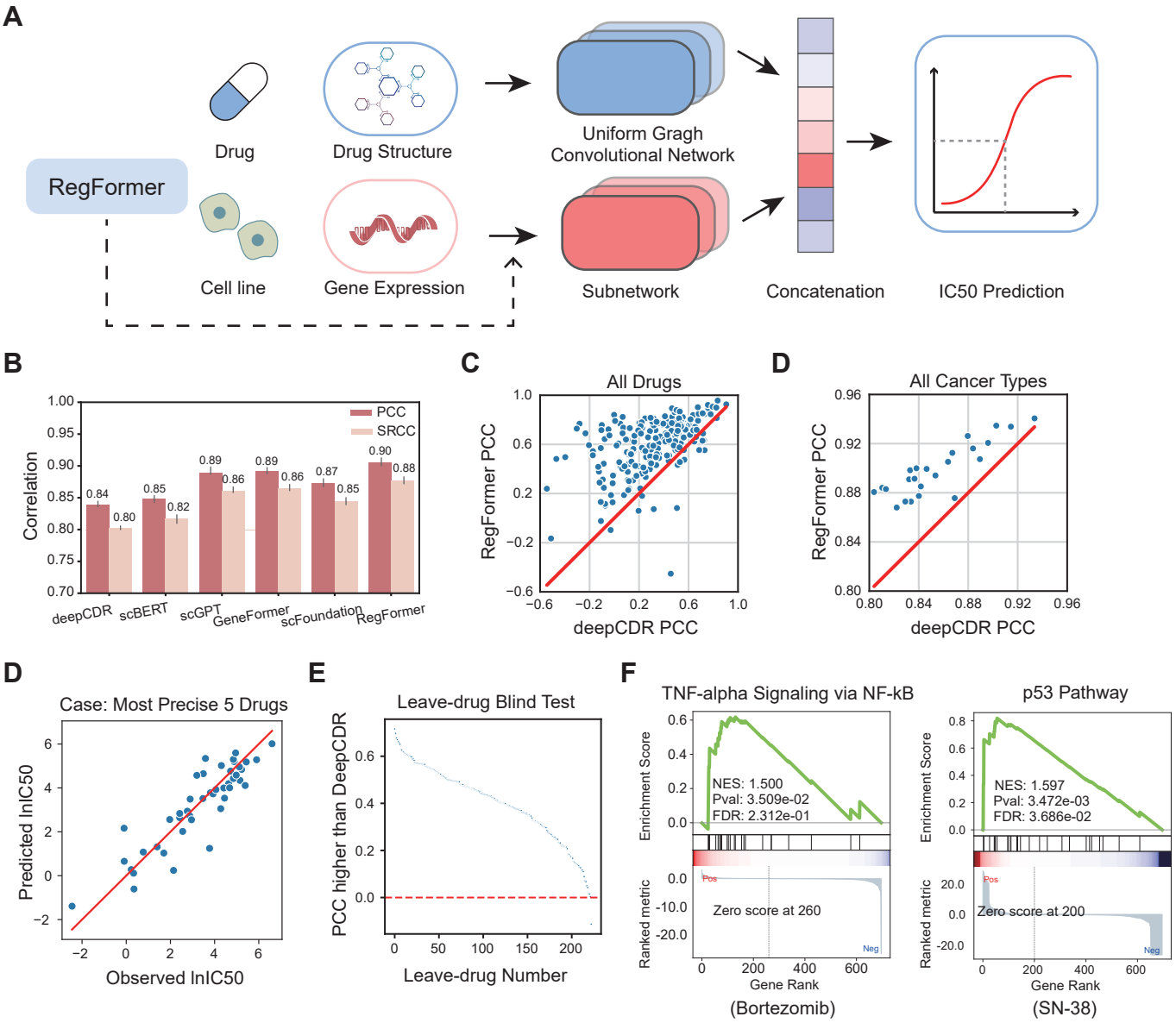

A
g1
g2
g3
gn
Unperturbed state
Single-cell Foundation Models
GeneID embedding
+
p2
p4
Genetic Perturbation
Embedding of GEARS
Perturbation
GEARS
Perturbed
0.2
0.6
1.5
0.3
0.9
B
Adamson
MSE
Pearson
C
Norman
MSE
Pearson
GEARS
scGPT
scFoundation
RegFormer

Figure 6


A
RegFormer
Drug
Drug Structure
Uniform Gragh Convolutional Network
Cell line
Gene Expression
Subnetwork
Concatenation
IC50 Prediction
B
Correlation
PCC
SRCC
0.84
0.80
0.85
0.82
0.89
0.86
0.89
0.86
0.87
0.85
0.90
0.88
deepCDR
scBERT
scGPT
GeneFormer
scFoundation
RegFormer
C
All Drugs
RegFormer PCC
deepCDR PCC
D
All Cancer Types
RegFormer PCC
deepCDR PCC
D
Case: Most Precise 5 Drugs
Predicted lnIC50
Observed lnIC50
E
Leave-drug Blind Test
PCC higher than DeepCDR
Leave-drug Number
F
TNF-alpha Signaling via NF-kB
NES: 1.500
Pval: 3.509e-02
FDR: 2.312e-01
Zero score at 260
Enrichment Score
Ranked metric
Gene Rank
(Bortezomib)
p53 Pathway
NES: 1.597
Pval: 3.472e-03
FDR: 3.686e-02
Zero score at 200
(SN-38)