Diffusion-based generation of gene regulatory network from scRNA-seq data with DigNet

Chuanyuan Wang and Zhiping Liu

Genome Res. published online December 18, 2024
Access the most recent version at doi:10.1101/gr.279551.124

Diffusion-based generation of gene regulatory network from scRNA-seq data with DigNet

Chuanyuan Wang[1], Zhi-Ping Liu[1,*]

[1]Department of Biomedical Engineering, School of Control Science and Engineering, Shandong University, Jinan, Shandong 250061, China

Abstract

Gene regulatory network (GRN) intricately encodes the interconnectedness of identities and functionalities of genes within cells, ultimately shaping to cellular specificity. Despite decades of endeavors, reverse engineering of GRN from gene expression profiling data remains a profound challenge, particularly when it comes to reconstructing cell-specific GRN that are tailored to precise cellular and genetic contexts. For alternatively approaching network reconstruction from data, we propose a discrete diffusion generation model, called DigNet, capable of generating corresponding GRN from high-throughput single-cell RNA sequencing (scRNA-seq) data. DigNet embeds the network generation process into a multi-step recovery procedure with Markov properties. Each intermediate step has a specific model to recover a portion of the gene regulatory architectures. It thus can ensure compatibility between global network structures and regulatory modules through the unique multi-step diffusion procedure. Furthermore, through the meta-cell integration and non-Euclidean discrete space modeling, DigNet can robustly be resistant to the noise of scRNA-seq data and the sparsity of GRN. Benchmark evaluation results against dozens of state-of-the-art network inference methods demonstrate that DigNet achieves superior performances across various single-cell GRN

reconstruction experiments. Furthermore, DigNet provides unique insights into the immune response in breast cancer, derived from differential gene regulations identified in T cells. As an open-source software, DigNet offers a powerful and effective tool for generating cell-specific GRN from scRNA-seq data.

Keywords: Diffusion generative network model (DigNet); Gene regulatory network generation; non-Euclidean graph embedding; breast cancer biomarkers.

Introduction

The transcriptional state of a cell is intricately controlled by a gene regulatory network (GRN) comprising numerous transcription factors and their target genes. Coordinating with other regulatory elements, GRN modulates the cell's phenotype, identity, and function in a dynamic and specific manner (Levine and Davidson 2005). These complex regulatory architectures are paramount in unraveling precise gene expression pathways and providing crucial insights for elucidating the physiological processes or disease mechanisms of multicellular organisms (Moris et al. 2016). Currently, single-cell RNA sequencing (scRNA-seq) techniques offer more precise methods for profiling high-resolution transcriptional states and delineating differences among diverse cell types. Given the profound importance of cell heterogeneity, it is naturally expected that variations in transcriptional states will correspond to changes in cell state-dependent gene regulatory interactions, which cannot be represented as static networks (Huang et al. 2018). Due to the specificity and dynamics of tissue microenvironments, technical noise, and impacts from other sources at the single-cell resolution, the transcriptomic gene expression levels may be partially decoupled from transcription factor

(TF) regulation, posing significant challenges to exploring the complex cellular landscape (Wagner et al. 2016).

Inferring gene regulations from transcriptomic data stands as a pivotal challenge in computational biology, aiming to reveal the cellular dynamics inherently manipulated by the interplay of genes. Over the past decades, numerous computational methods have emerged to infer GRN from gene expression profiles. These include correlation-based networks (Chan et al. 2017; Specht and Li 2017), Gaussian graphical models (Kotiang and Eslami 2020), tree-based ensemble pipelines (Huynh-Thu et al. 2010; Aibar et al. 2017), dynamic Bayesian models (Liu et al. 2016), and deep learning-based algorithms (Shu et al. 2021). Although these existing methods have achieved some advancements in inferring GRNs from transcriptomic data, they predominantly concentrate on modeling the regulatory relationships between individual genes and their multiple partners. While this approach can capture the local neighborhood influences, it face challenges in simultaneously modeling the interconnected and compatible regulatory relationships among a vast array of genes. Consequently, the derived networks are predominantly constructed from isolated gene interactions, and lack of system-level understanding of complex regulatory mechanisms (Ma et al. 2023; Wang et al. 2023). These limitations undermine the accuracy of reconstructing GRN from scRNA-seq data within specific contexts and hinder the ability of existing methodologies to decipher complex network structures. The architecture of GRN, both globally or locally, is crucial in complex biological systems, revealing essential nodes (e.g., highly interconnected TFs), pivotal regulatory modules, and elucidating how GRN adapt to intercellular variations and environmental stimuli. While cutting-edge algorithms can discern

some linear or nonlinear regulatory relationships, their predominantly focus on coupling individual gene pairs, rarely considering the intricate regulatory interplay among multiple gene simultaneously.

To address these challenges, we introduce DigNet, a deep generative model capable of deriving the underlying cell-specific GRN responsible for the transcriptional state directly from gene expression profiling data. The inspiration for DigNet comes from the booming generative techniques (Ho et al. 2020; Guo et al. 2023). Traditional generative models are usually conducted in an easy-to-understand Euclidean space, but it is difficult to accurately capture the complex regulatory relationships between genes. Therefore, DigNet employs a diffusion model framework, leveraging scRNA-seq data to embed the GRN structure into a non-Euclidean space with broader applicability, thereby generating sparse network architectures with unique structural characteristics. To reduce the complexity of embedding GRN in non-Euclidean space and to enhance the model's interpretability, DigNet further models the network with a binary discrete representation that includes only binary ('on' and 'off') states. This effectively ensures the preservation of both the topological network structure and underlying biological characteristics.

DigNet stands as a generative model that concurrently delves into the intricate regulatory interplay among genes. It emphasizes the global architecture information within GRN and generates corresponding network structure from scRNA-seq data. As a highly adaptable and flexible tool, DigNet necessitates merely single-cell gene expression profiles to iteratively generate a GRN from a random starting point. The extracted structure of GRN allows us for

diverse downstream analyses such as differential regulatory modules, pathways and pivotal factors. A comprehensive benchmark assessment encompassing 13 state-of-the-art GRN inference algorithms underscores the robustness and precision of the proposed network generation algorithm of DigNet. To demonstrate the versatility of DigNet, we applied it to reveal the intricate regulatory landscape of immune responses in human breast cancer. We constructed the immune cell-specific GRN and identified the differential networks across breast cancer samples and normal controls. By rediscovering known key regulatory relationships and prioritizing previously unknown candidate gene regulations, DigNet reveals cellular functional differences in the form of specific network rewiring and proves its utility in exploring network-based biomarkers. These novel differential regulatory associations and interactors offer fresh perspectives on refining the mechanisms underlying breast cancer immune responses and pave the way for the discovery of novel therapeutic targets.

Results

Overview of DigNet

As shown in **Fig. 1**, DigNet generates a cell-specific GRN from scRNA-seq data. Overall, DigNet dissects the network reasoning task into a reversible, multi-step recovery process with Markovian properties, including feature extraction, diffusion-based denoising, and backward inference. Consequently, it allows for the delineation of each temporal stage with a distinctive network model, thereby enhancing its capability to discern and reconstruct network structures with increased granularity. Additionally, graph transformer with the self-attention mechanism is employed to learn the complex data distribution in scRNA-seq data and address challenges

like experimental noise, high dimensionality, and scalability (Please refer to **Supplementary Note 1**). Once its fully trained model parameters are obtained, DigNet can easily generate a GRN given the gene expression profiles for any cells. Specifically, the initial phase involves optimizing gene expression data to mitigate the impact of single-cell dropout events and elevate data quality (**Fig. 1A**). Subsequently, DigNet applies a time-step approach to progressively denoise contaminated networks till achieving a clean network (**Fig. 1B**). During the training phase, DigNet iteratively alternates between "network contamination" and "noise removal" phases until convergence is achieved. DigNet starts with a random network structure for testing and gradually rectifies it using time-step. Both training and testing phases engage network encoding and Bayesian inference processes, which are pivotal to its performance (**Fig. 1C and D**). Finally, DigNet incorporates an ensemble learning strategy to counteract the instability issues stemming from random samplings (**Fig. 1E**). After be trained on single-cell GRN and corresponding transcriptomics data, DigNet can generate an appropriate network for new gene expression profiles, facilitating various downstream analytic tasks such as cellular differential gene expression analysis and biomarker discovery (**Fig. 1F**).

Benefiting from the diffusion generative framework, DigNet is one of the few models that directly generates network architectures at the global scale from scRNA-seq data (See **Supplementary Fig. S1A-B** and **Supplementary Note 2** for details). It emphasizes a holistic network generation process for the entire architecture, placing significant emphasis on ensuring compatibility between global regulatory network structure and gene expression profiles, thereby altering the approach to understanding cellular regulatory mechanisms.

Moreover, it changes the traditional single-step network inference paradigm into a multi-step network generative process. This enables the proposed method to pay more attention to the detailed dynamics in network structure with global architecture corresponding to gene expressions. Moreover, the reversibility of the network generation process allows DigNet to learn precise network architectures, which can be flexibly applied in important reverse operations, underscoring its adaptability and robustness in various analytical contexts.

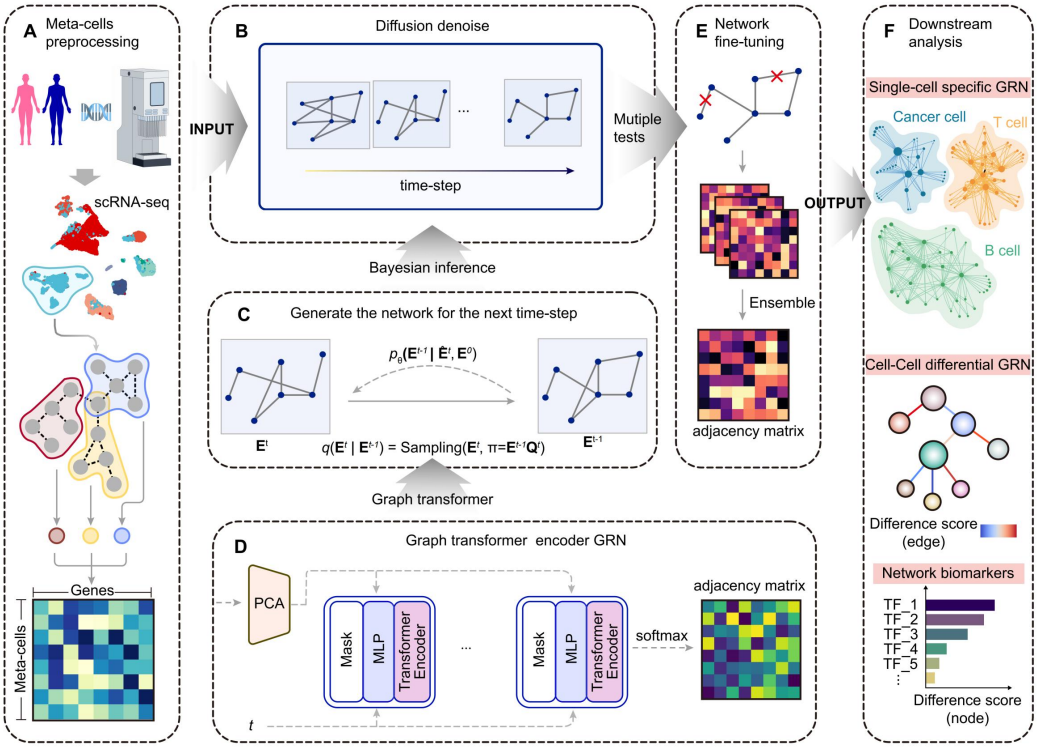

Fig. 1. Overview of DigNet. It generates cell-specific gene regulatory network and extracts differential network structures from single-cell gene expression profiles. (A) Data preprocessing of human tissue scRNA-seq data. (B) Network diffusion denoising across time-steps in DigNet. (C) Transforming the adjacency matrix exported by the Encoder through Bayesian inference to predict the gene regulatory network at the subsequent time-step. (D) Utilizing a transformer to encode GRN for scRNA-seq data based on the current time-step information. The procedures in (C) and (D) are repeated with each time-step to progressively denoise the network structure. (E) Correcting the network linkages (removing incorrect regulatory interactions) and integrating multiple diffusion-generated networks to

produce the final cell-specific network. (F) Once the generated network structure is obtained from scRNA-seq data using DigNet, multiple downstream analytics can be performed to identify key network features driving cell heterogeneity or biomarker signatures indicative of cancerous states.

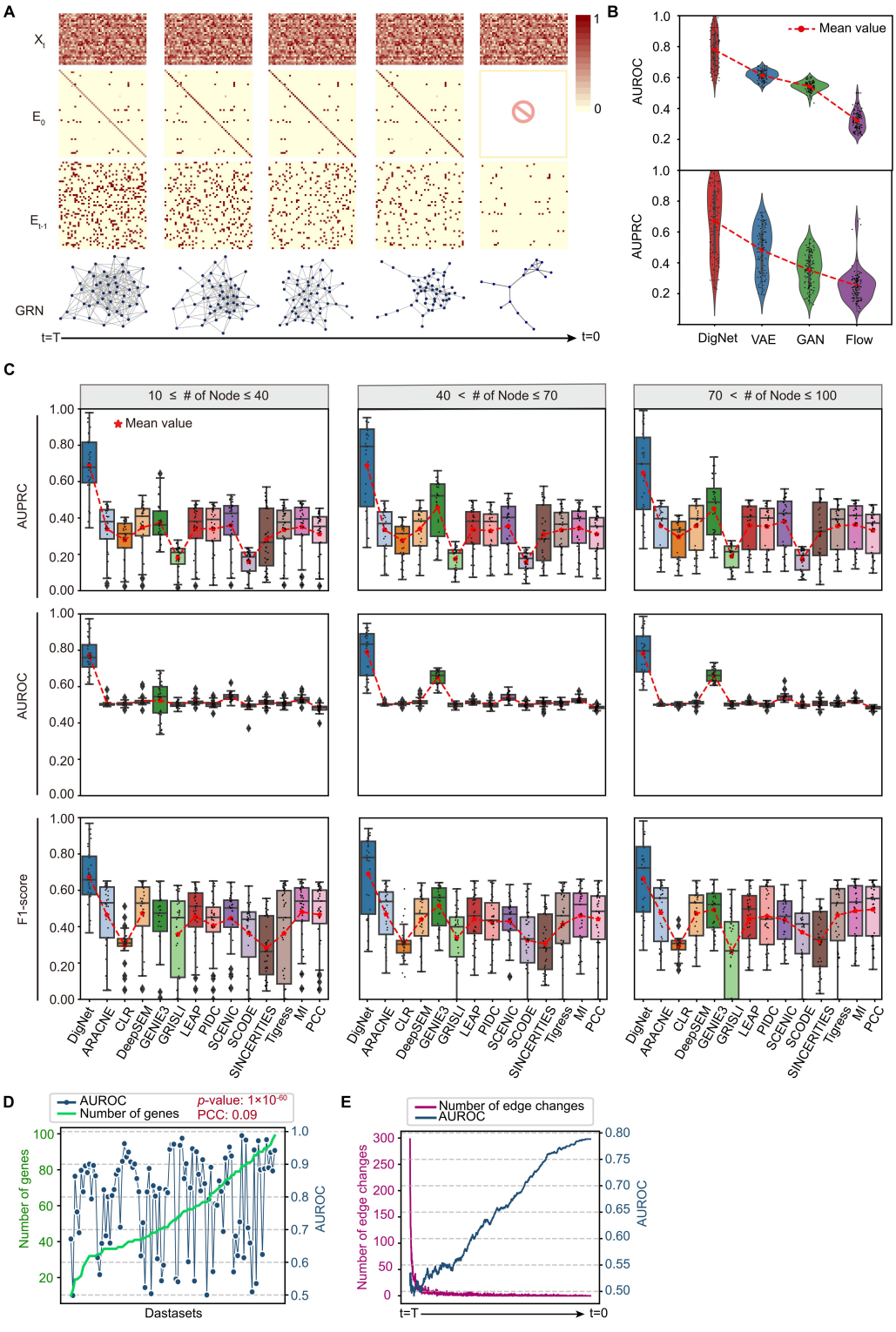

Fig. 2. Benchmark evaluations of DigNet on simulation datasets. (A) The intermediate network illustration in progressively generating a clean network from a random network, where each column represents a snapshot in time, including the input data X^t, the predicted clean adjacency matrix E^0, and the inferred output for the subsequent time-step E^{t-1}. (B) Performance evaluation of DigNet compared to alternative generative models (Diffusion *vs.* VAE, GAN and Flow) in terms of AUROC and AUPRC. (C) Performance evaluation of DigNet and 13 other GRN inference algorithms using 100 synthetic datasets. The red line denotes the mean values of these comparison methods. (D) The evaluation for DigNet network generation capability on the size impact of the number of genes, using AUROC value with the calculation of *t*-test *p*-values and PCCs between them. (E) During the network generation, the comparison of the number of edges corrected by DigNet at different time points during network generation and its subsequent effect on the overall AUROC value.

Extensive benchmark testing on simulation data confirms DigNet efficiency

To evaluate the performance of DigNet in network generation, we develop a simulation scheme for benchmarking gene expression profiles with single-cell GRN. The rationale for employing simulated gene expression data lies in establishing these predefined GRN as the ground truths for assessment. Taking one of the given GRN as an example, **Fig. 2A** illustrates how DigNet starts with a random network wiring and progressively generates a clean network. The gene expression profile serves as the input X^T, with the initial adjacency matrix being randomly generated as E^T. DigNet relies on the previous time-step network adjacency matrix

E^t (where t decreases from T to 0) and employs a Markov stochastic process in conjunction with the gene expression profile to iteratively generate a new adjacency matrix E^{t-1}. Upon repeating this process T times, a clean network will be derived as the final output (see Methods for details).

Based on the diffusion model, DigNet is a gene regulatory network generation method from single-cell gene expression profiles (refer to **Supplementary Note 3**). It decomposes the task of network generation into a series of sequential steps, each refining the current network wiring architecture guided by the previous one. To justify the effectiveness of the multi-step diffusion strategy, we also introduce three other popular generative models, i.e., variational autoencoders (VAE) (Way et al. 2020), generative adversarial networks (GAN) (Wang et al. 2018), and Flow (Stimper et al. 2022), into our algorithmic framework. These models are commonly used in scRNA-seq data analytics for denoising and generating low-dimensional representations. To our knowledge, they also have not been explored in generating gene regulatory networks. Mimicking the network generation pipeline of DigNet, we equip each of them accordingly with a GRN layer by replacing the diffusion strategy of DigNet to fulfill the network generation tasks. The detailed network constructions are described in Supplementary Files (**Supplementary Fig. S1C-E**). For a fair comparison, we train and test each generative method on the same simulation dataset and evaluate them using AUROC (area under the receiver operating characteristic curve), AUPRC (area under the precision-recall curve), and F1-score metrics (**Fig. 2B**; **Supplementary Fig. S1F**). Compared to VAE, GAN, and Flow models, the results demonstrate the superior performance of diffusion-based DigNet, with increasing AUROC values of 16.54%, 23.62%, and 45.81% respectively, and AUPRC

improvements of 18.92%, 31.82%, and 41.90%, respectively. These findings provide direct evidence that DigNet outperforms other generative algorithms in reverse engineering GRN from scRNA-seq data.

Different from traditional network reconstruction algorithms based on gene pair reasoning, DigNet emerges as a network generation method that focuses on wiring global network structure, demonstrating better performance over other generative models (refer to **Supplementary Note 4** and **Supplementary Fig. S2**). Subsequently, we compare DigNet with 13 state-of-the-art traditional GRN inference algorithms, including ARACNE (Margolin et al. 2006), CLR (context likelihood of relatedness) (Faith et al. 2007), DeepSEM (Shu et al. 2021), GENIE3 (Huynh-Thu et al. 2010), GRISLI (Aubin-Frankowski and Vert 2020), LEAP (Lag-based Expression Association for Pseudotime-series) (Specht and Li 2017), PIDC(Chan et al. 2017), SCENIC (Aibar et al. 2017), SCODE (Matsumoto et al. 2017), SINCERITIES (Papili Gao et al. 2018), and Tigress (Haury et al. 2012), along with Mutual information (MI) and Pearson correlation coefficient (PCC) based methods as baselines. The modeling basis of these methods is to infer gene regulatory pairs and then assemble them into a network, which is significantly different from the proposed network generation strategy. For completeness, the functionalities of these comparing methods with their implementation details are available in the Supplementary Files (**Supplementary Note 5**). Totally, we run these algorithms individually across 100 synthetic datasets of varying gene/node sizes and assess their performance using AUROC, AUPRC, and F1-score metrics with the gold-standard prior networks. Based on the number of genes, the datasets are divided into three categories for evaluation: nodes 10~40, 41~70, and 71~100. As shown, the results demonstrate that DigNet

consistently exhibits exceptional GRN reverse inference capabilities across all three-size networks (**Fig. 2C**). In terms of the AUROC value, DigNet achieves significant improvements of 24.80%, 13.89%, and 12.08% over the second-place algorithm (GRISLI) across three different network sizes. Similarly, for the AUPRC metric, DigNet improves over the second-place algorithm with 31.76%, 23.14%, and 19.67% for the three different-size networks, respectively. Regarding the F1-score, DigNet achieves substantial improvements of 23.24%, 17.49%, and 17.09% over the second-place algorithm, respectively. Except for CLR, LEAP, and SCODE, which performed poorly on certain metrics, the evaluation results for these comparing algorithms are relatively consistent and stable. In general, the performance of the diffusion-generation-based DigNet is significantly higher than other GRN inference methods.

Furthermore, we investigate the impact of network size of gene numbers on the performance of DigNet (**Fig. 2D**). As shown, the PCC of 0.09 and the *t*-test *p*-value of 1×10^{-60} indicate a lack of significant correlation between network scale and network generation performance via DigNet. The robustness underscores that the generative capabilities of DigNet are not influenced by the number of genes, allowing it to unbiasedly generate networks of varying size from gene expression data. Moreover, using one gold-standard network as an illustration, we examine the network wiring information corrected by DigNet at each time-step and assess the corresponding individual AUROC values (**Fig. 2E**). Starting from a randomly initialized network, the early diffusion denoising process alters a significant number of regulatory links, yet without a significant improvement in AUROC values. As the diffusion progresses, DigNet gradually capture the core architecture

of the underlying network from the data, where minor edge modifications lead to significant AUROC improvements. Towards the end of the time-steps, the structure of the network generated by DigNet almost cease to change, and corresponding AUROC values stabilized accordingly.

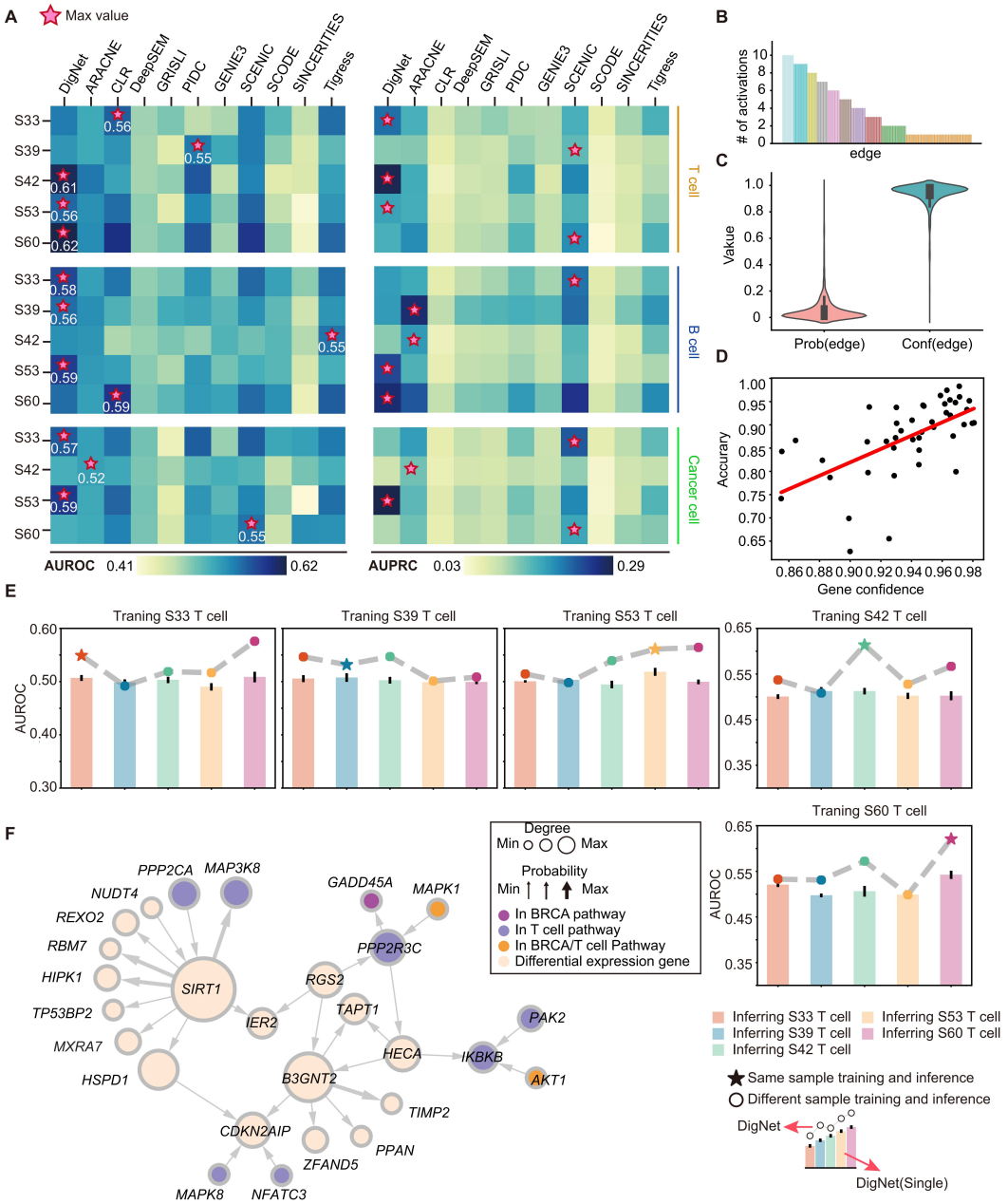

Fig. 3. Performance evaluation and network analysis of DigNet in breast cancer case study. (A) AUROC and AUPRC results for DigNet applied to breast cancer single-cell data.

The horizontal axis denotes individual cell types, and the vertical axis compares DigNet with 10 other alternative algorithms. (B) DigNet generates specific gene regulatory networks for T cells of Patient S42 ten times. We have conducted a statistical summary of all the activated regulatory links, where the horizontal axis represents the existing regulatory relationships, and the vertical axis indicates their frequency occurrence. (C) *Prob* and *Conf* represent the probability and confidence scores of all regulatory relationships across the ten-time generated networks, respectively. (D) Visualization of the relationship between accuracy and gene confidence, where gene confidence is the aggregated confidences of generation edges associated with each gene. (E) Benchmark testing of DigNet across diverse individual cell environments. The dotted line shows the AUROC evaluation results for DigNet on different individual cells after a single training session, while the bar graph shows results from a single run of DigNet (excluding the ensemble network aggregation module). (F) Specific GRN generated for T cells of Patients S33-S60, showcasing only subnetworks with relatively high node degrees for clarity.

DigNet generates reliable GRN in specific single cells

Subsequently, we verify the gene regulatory network generation capabilities of DigNet in real single-cell contexts of breast cancer. We compile the gene expression profiles of T cells, B cells, and cancer cells from five breast cancer patients (S33, S39, S42, S53, and S60, as specified in **Supplementary Table S1**) (Qian et al. 2020). Given the availability limitation of the gold standard networks, benchmarking GRN generation under real single-cell settings remains a significant challenge. Consequently, we first develop specific gene reference

networks for each cell type within every individual samples based on prior knowledge-bases and gene expression profiles. To reduce computational demands and topological complexity, we further divide the complete gene network into numerous subnetworks, each representing distinct biological pathways and functions according to the KEGG database. To comprehensively evaluate the model performance, the proposed DigNet method is rigorously compared against the former 10 baselines, including ARACNE, CLR, DeepSEM, GENIE3, GRISLI, PIDC, SCENIC, SCODE, SINCERITIES, and Tigress. In numerical experiments, we conduct a detailed evaluation of GRN in the gene set of the breast cancer KEGG pathway (Pathway ID: hsa05224).

Overall, DigNet outperforms the other GRN inference algorithms with the highest AUROC and AUPRC values (**Fig. 3A**; **Supplementary Fig. S1G**). Regarding AUROC evaluation, DigNet achieves optimal performance in 8 out of 14 evaluations, ranking within the top 5 in 13 evaluations, and only one exception in the cancer cell context. Moreover, DigNet exhibits the best average performance across all three cell types. In benchmark tests, the performance of DeepSEM, GRISLI, GENIE3, SCODE, and SINCERITIES (which DigNet surpasses by 6.5% to 11%) is under random predictions, suggesting their inability to accurately infer gene networks from breast cancer scRNA-seq data. The ARACNE, CLR, PIDC, GRNboost2, and Tigress algorithms can infer some proper regulatory relationships within single-cell environments. However, their performance is still slightly inferior to DigNet, with a decrease ranging from 1.7% to 3.4%. Observed, the several better-performing algorithms are mostly based on correlation or regression methods, which may be effective in navigating network inference in intricate cellular environments. When considering AUPRC

values, DigNet achieves optimal performance in 6 out of 14 evaluations and ranks among the top 5 in 12 evaluations, sharing a comparable optimal average performance with SCENIC. Generally, DigNet consistently demonstrates remarkable and stable capabilities in directly generating appropriate network architectures for single-cell gene expression profiles.

Furthermore, we explore the stability of network generation by DigNet. Utilizing T cells from patient S42 as a case study, we execute DigNet ten times and count the number of activations for each regulatory relationship (**Fig. 3B**, displaying only edges activated more than once for clarity). To facilitate the observation of edge permutations, we define $Prob$ and $Conf$ to represent the probability and confidence of *Gene* a regulating *Gene* b, respectively.

$$Prob(\text{a,b}) = \sum_{i=1}^{l} G_{a,b}^{i} \Big/ l, \tag{1}$$

$$Conf(a,b) = 1 - \text{std}(\{G_{a,b}^{1}, G_{a,b}^{2}, \ldots, G_{a,b}^{l}\}), \tag{2}$$

where $G_{a,b}^{i}$ denotes the regulatory relationship between *Gene* a and *Gene* b in the i-th iteration's network, and "std" represents the standard deviation. **Fig. 3C** compiles the *Prob* and *Conf* for all regulatory relationships, indicating the most gene pairs do not have a regulatory relationship and only a few gene pairs have regulatory information transmission. This result is consistent with the expected distribution of biological network regulations. Observed, the confidence of the majority of edges remains high across multiple repetitions, indicating that the networks generated by DigNet are quite stable. Furthermore, we calculate the node confidence (the cumulative confidence of incoming and outgoing edges) and the accuracy of the corresponding edges (**Fig. 3D**). Based on the results of regression analysis, we find that nodes with higher confidence within the network also exhibit higher accuracy.

Therefore, we have reason to conclude that the networks generated by multiple DigNet repetitions not only contain high confidence but also facilitate the selection of more credible gene regulatory relationships based on confidence values.

DigNet demonstrates notable generalization capabilities in network generation performance across novel environments for models trained under varying conditions. To address the scenario of network generation with limited training datasets, DigNet can be trained on samples from the same type of cells across different individual tissues, and generate a network structure for new samples. To align the new samples with the feature distribution of the training environment, we utilize PCA to project the new sample data into the principal component space defined by the training data, significantly enhancing the model's generalizability. Taking T cells as an illustration, DigNet uses multiple trained models to generate a network structure for five distinct samples (**Fig. 3E**). Additionally, under a cross-sample testing scenario, the AUROC values for samples S33, S39, and S53 demonstrate improvements compared to when the same models are trained and tested on identical datasets: 0.5491 (*vs.* 0.5759), 0.5318 (*vs.* 0.5470), 0.5630 (*vs.* 0.5643), respectively. In cross-sample testing experiments, networks generated by DigNet exhibit considerable and promising performance across most datasets, with network structures generated across different samples outperforming single iterations on certain datasets. Furthermore, the performance of an ensemble DigNet approach has been proven to surpass of a single iteration of run.

GRN reveals key regulatory pathways in breast cancer T cells

To further verify the efficiency of DigNet in the real application scenario, we apply it to generate an appropriate cell-specific GRN using the gene expression profiles of breast cancer T cells. Initially, we merge the gene count matrices from breast cancer patients S33 to S60 into a comprehensive dataset. Then, we construct a curated gene set of differentially expressed genes, T cell signaling pathways, and KEGG breast cancer pathways (**Supplementary Note 6; Supplementary Fig. S3**). DigNet is executed to generate a T cell-specific gene regulatory network on the integrated data (**Fig. 3F**; **Supplementary Fig. S4**).

DigNet reveals unprecedented regulatory connections among multiple pivotal genes, providing a new perspective for our in-depth understanding of the T cell immune response in breast cancer. Moreover, our analysis using DigNet reveals that *SIRT1*, a member of the Sirtuin protein family renowned for bridging transcriptional regulation with intracellular energy metabolism, exerts its influence on T cell immune responses via its downstream targets *HIPK1* and *MAP3K8*. This discovery underscores a novel, yet underappreciated, function of *SIRT1* in modulating breast cancer immunity. While *SIRT1* is not explicitly listed as a key player in breast cancer pathway according to the KEGG database (absent from hsa05224), it is recognized as a tumor suppressor due to its protective role against DNA damage and oxidative stress, safeguarding genomic stability during tumor progression (Sung et al. 2010; Elangovan et al. 2011; Bajrami et al. 2021). Moreover, *HIPK1*, another gene in our analysis, has been implicated in aberrant states during inflammatory responses and tumor

development (Liu et al. 2018; Zhang et al. 2021). Although there is no direct evidence of *HIPK1* being regulated by *SIRT1*, its homolog *HIPK2* has shown potential crosstalk with *SIRT1* in DNA damage response, leading to the phosphorylation of *SIRT1* at serine 682 after lethal damage (Puca et al. 2009; Conrad et al. 2016; Choi et al. 2017). Moreover, *SIRT1*'s inhibitory effect on the pro-inflammatory cytokine *TNF* during immune responses is well-documented, and *MAP3K8*, another key player in our findings, is crucial for *TNF* production (Gantke et al. 2011; Huang et al. 2012; Chen et al. 2020; Wang et al. 2020). *MAP3K8* participates in T helper cell differentiation and intracellular gene regulation, and under certain conditions, it activates the MAPK/ERK pathway, leading to *TNF* production, thereby playing a pivotal role in immune responses (Tsatsanis et al. 2008).

Moreover, an important downstream application of DigNet lies in prioritizing key regulatory genes to elucidate the mechanisms underlying disease progression when cellular processes become dysregulated. Specifically, by applying DigNet, we meticulously compare the regulatory networks of T cells between normal and cancerous breast tissue samples, pinpointing crucial nodes that underlie the regulatory discrepancies driving disease progression. As a reference, we construct a knowledge-based gene regulatory network for normal breast T cell samples (GEO ID: GSE195665). Based on the breast cancer-specific T cell GRN and the normal breast T cell GRN generated by DigNet, we quantify the regulatory rewiring differences for each gene within the normal/cancer networks (**Fig. 4A**). To this end, we introduce a gene difference score (GDS) to evaluate the variation in node connectivity between two gene networks:

$$GDS_{g1} = \sum \left| A^{a}_{g1,*} - A^{b}_{g1,*} \right|, \quad (3)$$

where A^a and A^b represent the adjacency matrices of the respective GRN, and g1 denotes the gene node. A GDS greater than 0 suggests an active regulatory function of the gene in normal breast T cells. Consequently, we select the top 3 genes with the highest GDS in both normal and cancer contexts (normal: *JUN*, *MYC*, *SP1*; cancer: *IGFBP4*, *FABP5*, *ZC3H12A*) and assemble a differential gene regulatory network for breast cancer (**Fig. 4B**). Among these genes, *MYC* and *SP1* belong to the KEGG breast cancer pathway, while *JUN* is featured in both the breast cancer and T cell receptor pathways. The differential regulatory network analysis also highlights substantial variations in the regulatory interactions medicated by these genes in normal versus cancerous T cell environments. Furthermore, most downstream target genes of these prioritized genes exhibit abnormal activation in cancerous conditions, evident by their tendency towards Magenta color in **Fig. 4B**).

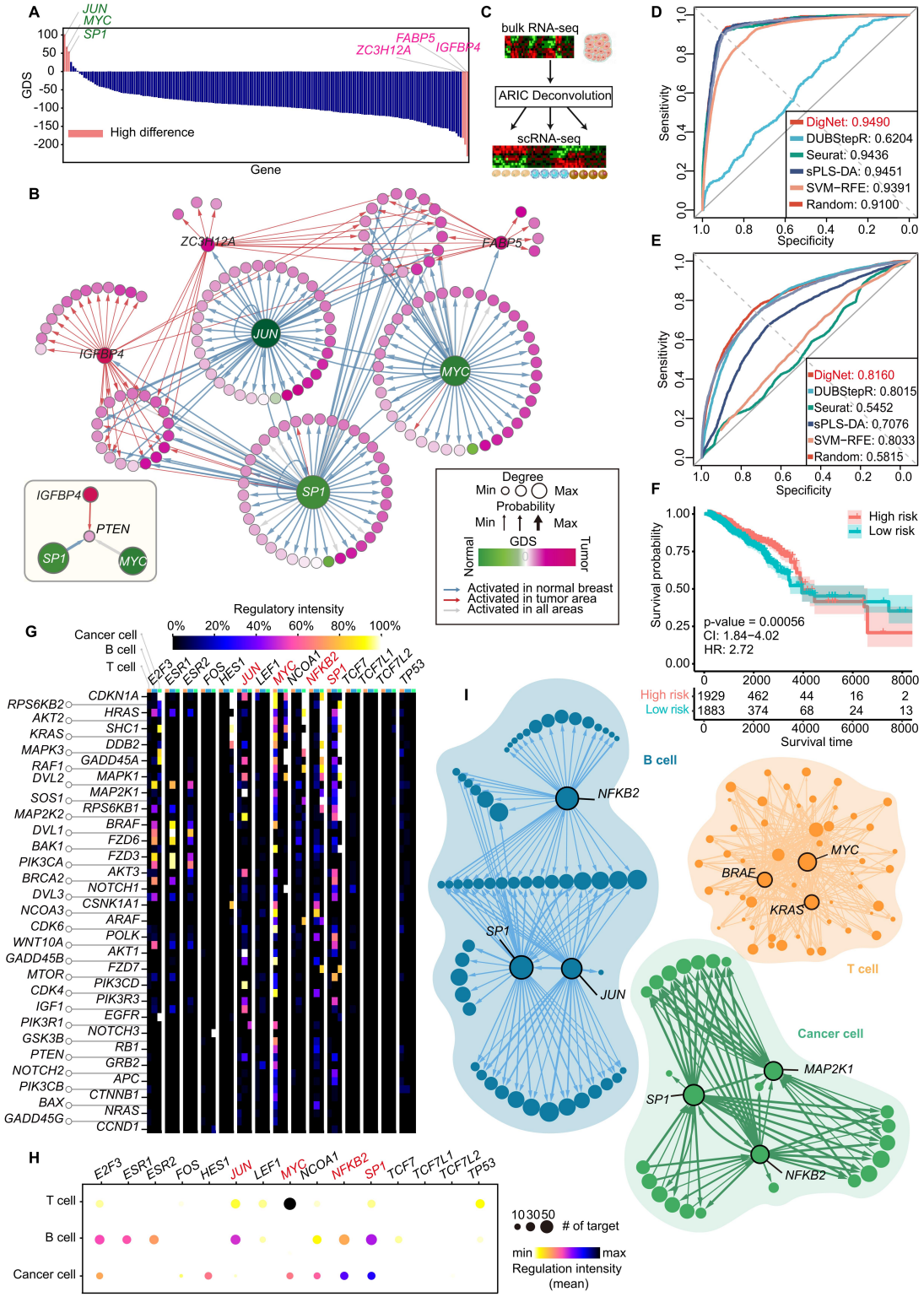

Fig. 4. Breast cancer cell-specific GRN reveals key genes and regulatory architectures. (A) Gene differential scores are calculated based on the difference between T cell GRN in normal breast tissue and cancerous samples, highlighting the top 3 genes with the highest

scores. (B) Extraction and analysis of T cell differential gene regulations associated with high differential genes, with a focus on the regulatory heterogeneity of the *PTEN* gene under normal conditions (bottom left). (C) Schematic illustration of deconvolving TCGA breast cancer bulk RNA-seq data into scRNA-seq data using the ARIC method. (D) Classification AUROC values for the top 10 biomarker genes identified by DigNet and other methods in TCGA BRCA T cells, with optimal values highlighted in red. (E) Classification AUROC values for the top 20 biomarker genes selected by DigNet and alternative methods in normal/cancerous breast T cells. (F) Kaplan-Meier (K-M) curves for the top 10 biomarker genes chosen by DigNet in TCGA breast cancer T cells, analyzed through a multifactorial Cox survival analysis, where CI represents the confidence interval, and HR stands for the hazard ratio. (G) Regulatory landscape of transcription factors-target gene interactions in T cells, B cells, and cancer cells of Patient S53, with regulatory strength inferred from DigNet probability values. (H) Cell type-specific summary of each transcription factor's target genes in subfigure (G), where dot size reflects target gene frequency, and color indicates the average probability regulatory strength). (I) Top 3 genes (or TFs) ranked by "Var-score" for each cell type, along with their associated regulatory relationships, with TF genes highlighted in red in (G) and (H).

In addition to these three hotspot genes, we focus on *IGFBP4*, *FABP5*, and *ZC3H12A*, which exhibit associations with cell proliferation, inflammatory responses, or immune regulation. For instance, *IGFBP4*, swiftly induced by estrogen and exhibiting abnormal expression across diverse tumor tissues, emerges as a crucial prognostic indicator for breast cancer patients (Ryan et al. 2009; Flynn and Houston 2022; Chen et al. 2023). Moreover,

IGFBP4 plays a pivotal signaling role in the differentiation of select T cell subtypes, maintaining a delicate balance between T helper 17 and regulatory T cells (Miyagawa et al. 2017; DiToro et al. 2020). Similarly, *FABP5* and *ZC3H12A* genes are intriguing in breast cancer progression and T cell immune responses, potentially representing novel therapeutic targets and prognostic markers (Matsushita et al. 2009; Liu et al. 2011; Levi et al. 2013; Lu et al. 2016; Senga et al. 2018; Li et al. 2021). Furthermore, the derived differential gene regulatory network implies regulatory functions in both normal and cancerous tissues. For instance, *PTEN* is known to be regulated by *MYC* and *SP1* in normal conditions, yet DigNet suggests an additional regulatory link with *IGFBP4* in cancerous tissue. This novel regulatory interaction is not documented in existing knowledgebases, and studies suggest have shown that they could affect tumor proliferation through signaling pathways (such as the AKT pathway), with other members of the *IGFBP* family also having some interactions with *PTEN* (Baxter 2014; Lee et al. 2018; Ruan et al. 2023). In normal tissues, estrogen or progesterone receptors regulate *IGFBP4* levels via *SP1*, but in breast cancer T cells, DigNet found that this regulatory relationship is not active. Although no studies support or refute the latter, we propose this hypothesis from the differential network.

To validate the differential gene regulatory network generated by DigNet from multiple perspectives, we attempt to distinguish T cell data from breast cancer samples in TCGA using the top 10 key genes identified by GDS (**Supplementary Table S2**). The T cell data from TCGA breast cancer cases is obtained through ARIC deconvolution (**Fig. 4C** and **Supplementary Note 7**) (Zhang et al. 2022; Zhang et al. 2023). We benchmark several prominent feature selection algorithms: DUBSTepR (Ranjan et al. 2021), Seurat (Hao et al.

2023), sPLS-DA (Lê Cao et al. 2011), and SVM-RFE (Duan et al. 2005), by comparing their performance against a randomly selected gene set. These algorithm are employed to select key biomarker signature genes within our breast cancer/normal single-cell dataset, which are then evaluated on the TCGA deconvoluted T cell data for their effectiveness. We conduct T cell classification experiments under an SVM classifier based on five-fold cross-validations to ensure robustness (**Fig. 4D**). To further strengthen the validation of these biomarker genes, we replicate the experiments on a single-cell dataset of normal/cancerous breast tissue sourced from the GEO database (ID: GSE114725) (**Fig. 4E**), where all algorithms use the top 20 key genes by GDS (**Supplementary Table S2**). Both sets of results underscore the remarkable discriminative capacity of the key network nodes identified by DigNet. Furthermore, survival analysis of TCGA patients affirms that the biomarkers pinpointed by DigNet exhibit statistically significant differences (**Fig. 4F**). Furthermore, we perform survival analysis specifically on the top 10 genes with high GDS in the TCGA breast cancer cohort (**Fig. 4F**). Multivariable Cox regression analysis is utilized to estimate coefficients, alongside confidence intervals (CI) and hazard ratios (HR) for these genes (**Supplementary Note 8**). Based on the derived coefficients, we assigned weighted scores to the top 10 genes and classified patients into high/low-risk groups accordingly. Subsequently, we estimated the survival probability for these groups using the Kaplan-Meier (K-M) method, which revealed a statistically significance difference (p-value = 0.00056). These findings from the survival analysis indicate that the biomarkers identified by DigNet exhibit significant differences in survival time estimation for breast cancer.

Overall, the unique network structure generation approach of DigNet enables us to construct cell-specific regulatory networks, discern normal-disease differential networks, and uncover pivotal network-based biomarkers. The essential modules, regulatory interactions, and biomarkers embedded within these networks reflect the internal carcinogenic mechanisms within cells, holding immense potential to revolutionize cancer diagnosis, prognosis, and therapy.

Differential analysis of single-cell gene networks reveals the heterogeneity among breast cancer cells

Cell-specific GRN serve as potent instruments to decipher variations in cellular functions, as they orchestrate the expression of gene products and shape the unique developmental pathways of individual cells (Karlebach and Shamir 2008). In contrast, public, non-specific reference networks, rooted in general knowledge-bases, do not adequately capture the intricacies of these phenomena. However, DigNet offers a solution by generating cell-specific GRN that facilitates an intricate analysis of the nodal interactions and network architectures responsible for functional disparities among diverse cell types. Specifically, we harness DigNet to construct cell-specific networks for T cells, B cells, and cancer cells from breast cancer patient sample S53 (**Supplementary Table S3-S5**). Furthermore, we quantify the regulation influence from TFs to target genes across these three cell types to reflect changes in TF activity within each cellular context (**Fig. 4G**). Moreover, significant disparities in TF activity are observed among these cells (**Fig. 4H**). For instance, TFs such as *E2F3*, *ESR1*, and *ESR2*, which influence tissue growth and development, exhibit heightened activity in B cells,

regulating multiple target genes (Zhu et al. 2001; Xiao et al. 2008; Langendonk et al. 2022). In contrast, the TF *MYC* is markedly silent in B cells but active in T cells, with extensive prior research underscoring the consequences of *MYC* overexpression in T cells (Zimmerli et al. 2022). These pivotal TFs demonstrate the superiority of specific GRN over public networks and elucidate the paramount importance of prioritizing TFs associated with immune cells in breast cancer progression.

Furthermore, we aimed to delineate cell type-specific genes and network architectures to address the heterogeneity in cellular functions. To accomplish this, we devised a multicellular gene differential score metric to evaluate the gene quality specific to each cell type. Specifically, we transformed the adjacency matrices generated by DigNet into binary (0-1) matrices, denoted as A. Subsequently, we formulated the following equation to capture the differential score of gene g1 in cell type a:

$$\text{Var-Score}_{g1}^{a} = \sum |A_{g1,*}^{a} - (A_{g1,*}^{b} \odot A_{g1,*}^{c})|, \quad (4)$$

where $A_{g1,*}^{a}$ signifies the comprehensive set of target regulatory interactions involving gene $g1$ in cell type a. $\odot$ is the Hadamard product. Furthermore, we quantified the Var-Score for genes across the specific networks of all cells (see **Supplementary Table S6**), and extracted the network structures corresponding to the top three gene with the highest "Var-Score" (**Fig. 4I**). The pivotal gene sets for the three groups of cells are as follows: *MYC*, *BRAF*, *KRAS* for T cells; *SP1*, *NFKB2*, *JUN* for B cells, and *SP1*, *MAP2K1*, *NFKB2* for cancer cells. Although, B cells and cancer cells share two crucial genes, but their target genes are entirely different. For example, the gene *mTOR*, likely regulated by *SP1*, controls the growth, development, and proliferation of B cells (Astrinidis et al. 2010; Iwata et al. 2017). In cancer cells, SP1 robustly

regulates genes such as *SOS1*, *NFKB2*, *HRAS*, and *AKT2*, and elucidating these aberrant regulatory mechanisms can provide insights into cancer cell invasion and metastasis. It is noteworthy that the key genes in T cells exhibit marked differences from those of other cell types. Besides the previously discussed *MYC*, the specific responses generated by mutant *KRAS* and *BRAF* in T cells are also prime targets for anti-tumor immune therapy (Wilmott et al. 2012; Tran et al. 2016). The partial regulatory actions of these genes demonstrate heterogeneity across diverse cell types, facilitating the understanding of cell functions within the tumor microenvironment and identifying crucial therapeutic targets. In essence, the single-cell GRN generated using DigNet from data enables the discovery of significant cell-specific regulatory relationships and pivotal network nodes via downstream analyses. This heightened resolution and specificity in gene regulatory dynamics, alongside with the intricate molecular details, surpass the limitations of average gene expression profiles and publically available gene networks.

Discussion

In this paper, we introduced a network generation method called DigNet for deriving cell-specific GRN from scRNA-seq data. DigNet uses Bayesian inference and graph transformer techniques by iteratively refining an initial random network to constrct a comprehensive and detailed GRN for individual cells. The non-Euclidean discrete diffusion modeling enables DigNet to generate a global network architecture rich in structural features. Meanwhile, the progressive generation process and reversibility enable DigNet to capture structural details within the entire network, ensuring that the overall structure of the generated

network remains consistent with the input gene expression profiles. The uniqueness of DigNet can be summarized in three pivotal aspects: the generation of GRN from gene expression data with discrete diffusion models; multi-time-step diffusion techniques for noise reduction and network refinement; and the integration of generative deep learning with a hybrid model architecture. Through rigorous benchmark tests across diverse biological contexts and datasets, we demonstrated the efficiency, robustness, and superiority of DigNet, particularly in terms of reproducing cell-type gene regulatory specificity. Moreover, DigNet achieves single-cell-specific gene regulatory network inference from scRNA-seq data, identifying crucial regulatory network nodes and causal modules leading to cell-type specificities. DigNet introduces a novel generation network model for GRN reverse engineering, enabling it to respond to single-cell gene expression profiles with a more suitable network architecture through a progressive denoising procedure rather than assembling isolated regulatory signals.

Recovering GRN architectures through generative models offers a novel reverse engineering paradigm and alternative for gene expression data, presenting multiple challenges. A critical challenge for DigNet is that simple random sampling can result in slight variations in outcomes at the same time-step, which may inadvertently introduce novelty-driven rewiring and unwarraented randomness into the network. Unlike conventional diffusion models, DigNet incorporates no specific conditional controller to determine which networks are more suitable, primarily due to the absence of clear criteria or justifications for filtering specific network architectures across diverse cellular environments. To address this, our solution strategy revolves around statistically estimating the probability of regulatory events by counting the activation frequencies of regulatory signals across multiple networks, offering

a straightforward yet effective learning approach. Compared to other GNN (graph neural network)-based methods, DigNet eliminates the requirement for pre-constructed initial graphs by utilizing a diffusion model-based generation strategy (**Supplementary Note 9**). This approach enhances both the adaptability and accuracy of GRN inference. A potential future direction for DigNet involves incorporating cell developmental trajectories to model dynamic GRN throughout continuous cellular developmental stages. Furthermore, the integration of multi-omics data, which encompasses genomic sequence information, chromatin accessibility data, TF activity, and protein-protein interaction networks (Badia-i-Mompel et al. 2023), emerges as crucial future direction for advancing the capabilities of DigNet. By utilizing these diverse multi-omics data, we foresee a significant enhancement in the accuracy and precision of reconstructing dynamic GRN from intricate datasets. Moreover, the involvement of TF information will be substantially elevated through this comprehensive integration. For more detailed expansions and limitations, please refer to **Supplementary Note 10**

Methods

Framework

The emergence of generative techniques such as GAN, VAE, Flow, and Diffusion models, has revolutionized data generation and diversification (Kingma and Welling 2013; Goodfellow et al. 2014; Ho et al. 2020; Austin et al. 2021; Ramesh et al. 2022). Currently, the diffusion model, a typical generative model, has demonstrated its powerful generative capabilities across text, image, and video domains (Ho et al. 2022; Saharia et al. 2022). In this work, DigNet leverages the diffusion model framework for discrete space modeling to capture

complex gene regulatory interactions within non-Euclidean biological systems. It transforms the input gene expression profiling data into vector representations in a high-dimensional space, which is then utilized to refine gene regulatory relationships in wiring-contaminated networks. Given a GRN $G = (V, E)$, where V denotes genes and E signifies their regulatory relationships between regulators and targets, we consider the GRN to contain binary attributes, 1 or 0, corresponding to the activation or inhabitation of gene switches, respectively. DigNet primarily comprises two components: the forward diffusion (noise addition) of gene network E and the backward denoising stage (employing a neural network). The relevant parameter settings are discussed in **Supplementary Note 11** and **Supplementary Fig. S5A-C**. To ensure the efficiency and effectiveness of the diffusion model, DigNet embodies the following three properties for both forward diffusion and backward denoising processes:

1. $q(E^t|E^0)$ possess a closed-form solution to ensure its stability across varying time-step t.
2. $p_\theta(E^{t-1}|E^t)$ is an expression with a closed-form solution, which empowers the neural network with parameters θ in learning and targeting the original network E^0.
3. As the diffusion time T approaches infinity, the network structure should converge to a marginal distribution related solely to noise values, independent of E^0, denoted as $q(E^T) \propto q(E^T|E^0)$.

Forward diffusion

The noise diffusion process for GRN is based on a Markov chain framework, where the generation of subsequent networks with noise values progresses step by step along a

predetermined direction and noise level. For the network E^t at time t, its derivation solely depends on E^{t-1}, and it can further induce E^{t+1} based on the predetermined noise values. Due to the Markov property, given the preset noise parameters and the initial network E^0, we can derive the joint prior distribution of networks at any given time as:

$$q(E^1, ..., E^T|E^0) = q(E^1|E^0)\prod_{t=2}^{T} q(E^t|E^{t-1}). \quad (5)$$

When T is sufficiently large, the model aligns with a Markov jump process under a discrete space distribution. The forward noise levels are predetermined as a state transition matrix Q, which contains two states, 0 and 1, to map the GRN. At $t = 0$, Q is initialized such that the probability of self-transition is 1. As t progresses from 1 to T, the state transition probabilities evolve, gradually transforming the original GRN towards a random network. Let α denote the network noise coefficient, ranging from 0 to 1. Then, $Q_{i,j}^t$ represents the probability of transitioning from state i to state j at time t, which can be mathematically described as follows:

$$Q^t = (1 - \alpha^t)I + \alpha^t \begin{bmatrix} M_{0,0} & M_{0,1} \\ M_{1,0} & M_{1,1} \end{bmatrix}. \quad (6)$$

As the noise coefficient α approaches 1, the network converges to a random distribution based on a fixed value M^t.

Furthermore, based on the noise values, we can infer the network for the subsequent time-step as follows:

$$q(E^t|E^{t-1}) = Sampling(E^t; \pi = E^{t-1}Q^t), \quad (7)$$

where $Sampling(E, \pi)$ refers to a state distribution encoded in a one-hot scheme, derived from simple random sampling with a probability value of π. Furthermore, since the noise values are predetermined in a closed form, we can combine and simplify the probability

distributions up to $t-1$ similar to the approach in DDPM (Ho et al. 2020), to derive the marginal distribution network at time t conditioned on the initial network E^0:

$$q(E^t|E^0) = Sampling(E^t; \pi = E^0\overline{Q}^t), \text{ where } \overline{Q}^t = \prod_{i=1}^{t} Q_i. \quad (8)$$

Given that the noise value Q is fixed, E^t can be directly derived from E^0, enabling the diffusion model to be trained from any arbitrary time-step. Based on Equations (6) and (8), we can generate the noised network at any specified time-step.

Backward denoising

During the reverse process, DigNet accomplishes the task of regenerating network linkages from E^t to E^{t-1} through trained neural networks. By training the deep neural network under specific parameters, the clean network can be iteratively denoised and obtained from the noisy network at time t (**Supplementary Note 12**). Intuitively, E^{t-1} can be inferred by decoding E^t through the neural network. However, this iterative process may lead to error accumulation, making model training extremely challenging. Based on the Bayesian theorem, we derive a posterior probability inference related to E^t, E^0, and Q:

$$q\left(E^{t-1}\middle|E^t, E^0\right) = \frac{q\left(E^t|E^{t-1},E^0\right)q\left(E^{t-1},E^0\right)}{q\left(E^t|E^0\right)} = Sampling\left(E^{t-1}; \pi = \frac{E^t(Q^t)^{'} \odot E^0\overline{Q}^{t-1}}{E^0\overline{Q}^t(E^t)^{'}}\right), \quad (9)$$

where E^t is known, while E^0 and Q are constants. Thus, Equation (9) admits a closed-form solution. Analogous to the likelihood estimation in continuous spaces, the integral of the evidence lower bound (ELBO) in discrete space is formulated as:

$$l_{vb} = \mathbb{E}_{q(E^0)}\left[D_{KL}[q(E^T|E^0)\|p(E^T)] + \mathbb{E}_{q\left(E^1\middle|E^0\right)}[\log p_\theta(E^0|E^1)] + \sum_{t=2}^{T} \mathbb{E}_{q(E^t|E^0)}[D_{KL}[q(E^{t-1}|E^t, E^0)\|p_\theta\left(E^{t-1}|E^t\right)]]\right]. \quad (10)$$

Each component can be estimated as follows: $D_{KL}[q(E^t|E^0)\|p(E^0)]$ (prior loss) does not require optimization since it contains no trainable parameters, and E^t is predefined as a stochastic distribution network when T is sufficiently large; $\mathbb{E}_{q\left(E^1|E^0\right)}[\log p_\theta(E^0|E^1)]$ (reconstruction loss) is derived from the clean gene network based on the final noise-free network; $\sum_{t=2}^{T}\mathbb{E}_{q(E^t|E^0)}[D_{KL}[q(E^{t-1}|E^t,E^0)\|p_\theta\left(E^{t-1}|E^t\right)]]$ (diffusion loss) enforces consistency matching between predictions and noise-adding processes at each intermediate step, ensuring that the prediction network align with the noise-adding network. Our optimization objective is to train $p_\theta(E^{t-1}|E^t)$ to closely match $q(E^{t-1}|E^t,E^0)$.

Although $p_\theta(E^{t-1}|E^t)$ can be predicted directly, according to the experience of Ho *et al.*, the prediction of $p_\theta(E^{t-1}|E^t)$ with inherent noise may lead to model instability due to the noise level of uncertainty (Ho et al. 2020; Austin et al. 2021). Recognizing that the neural network model can learn the intrinsic data distribution, a feasible solution is to predict $\hat{p}_\theta(\widehat{E}^0|E^t)$ and subsequently derive $p_\theta(E^{t-1}|E^t)$ based on the Bayesian theorem:

$$p_\theta(E^{t-1}|E^t) \propto \sum_{\widehat{E}^0} q(E^{t-1}|E^t,\widehat{E}^0)\,\hat{p}_\theta(\widehat{E}^0|E^t). \tag{11}$$

When the estimated $\hat{p}_\theta(\widehat{E}^0|E^t)$ perfectly aligns with the data distribution of E^0, the Kullback-Leibler (KL) divergence $D_{KL}[q(E^{t-1}|E^t,E^0)\|p_\theta E^{t-1}|E^t)]$ approaches 0. The parameterization of E^0 not only enhance the stability and performance of model training but also simplifies the learning task of deep neural network model.

Therefore, optimizing the ELBO problem essentially involves training a neural network to predict clean GRN from an arbitrarily contaminated GRN. To train the neural network ϕ_θ,

we optimize the cross-entropy loss $loss_{CE}$ between the predicted probability edges $\phi_\theta(\widehat{E}^0|E^t)$ and the true network E^0:

$$loss_{CE} = \sum_{1<i,j<n} cross - entropy(\phi_\theta(\widehat{E}^0|E^t)_{i,j}, E^0_{i,j}). \quad (12)$$

Distinct from other generative models, $loss_{CE}$ is exclusively focuses on generating cleaner network architectures and does not encompass other tasks. The training process framework for our proposed diffusion model is illustrated in **Supplementary Fig. S6**.

Denoising transformer network

Our task is to predict the distribution of clean networks conditioned on noisy networks and gene expression profiles, which involves detailed scoring of potential interactions between TF and target gene pairs. To achieve this, we employ the graph transformer network with a self-attention mechanism (AGTN) (Dwivedi and Bresson 2020), capable of leveraging existing regulatory edge features to amplify the scores of the implicit attention mechanism and infer richer feature information.

To better learn the complex distributions of scRNA-seq data and accurately capture the intricate network structures within GRN, we improved the original AGTN method. Specifically, we substituted the Laplacian features and positional embeddings in AGTN with dimensionality-reduced features obtained through PCA. Furthermore, we integrated two feature learning modules, FiLM and PNA, to refine the modulation of nodes and edges features respectively (Perez et al. 2018). The definitions of these two feature learning layers are detailed below:

$$FiLM(x_1, x_2) = x_1W_1 + (x_1W_2 \odot x_2) + x_2, \quad (13)$$

$$PNA(x) = \{max(x), min(x), \text{mean}(x), std(x)\}W_3, \quad (14)$$

where W_1 , W_2 and W_3 are learnable weight matrices. $\odot$ denotes element-by-element multiplication. $\{max(x), min(x), \text{mean}(x), std(x)\}$ represents the operations of taking the maximum, minimum, mean, and standard deviation of matrix x by rows respectively, and horizontally concatenating the results. Recognizing that the core of the diffusion model lies in the denoising of networks across different time-steps, we incorporated a time-step module into AGTN. This module is conditioned equally by node information, edge information, and time information, enabling it to effectively capture temporal dynamics during the denoising process.

In addition, we defined and weighted the self-attention module for edge learning. Specifically, DigNet obtains the regulatory scores between TF and target gene pairs through self-attention learning of low-dimensional embeddings of the gene expression matrix. This is formulated as:

$$X\text{-}Score(i, j) = X_{i,*} \cdot X_{j,*}. \quad (15)$$

Subsequently, the current network E^t is modified by the regulatory scores $X\text{-}Score$, and the network information evolves along the time dimension, i.e.,

$$XE\text{-}Score^t = FiLM(FiLM(X\text{-}Score, E^t), t), \quad (16)$$

where the $XE\text{-}Score^t$ represents the final outcome obtained through the weighted self-attention mechanism. It seamlessly integrates gene expression profiling, input network topology, and temporal information, fully utilizing multi-dimensional information for the learning of network edge weights. Please refer to **Supplementary Notes 13-14 and**

Supplementary Fig. S5D-E for details on the algorithm runtime environment and initialization.

Marginal probability and noise presets

The selection of the Markov transition matrix that defines the network regulatory probabilities within the network is inherently subjective, and there is often uncertainty regarding which noise model would optimally capture the dynamics of the diffusion process. Under most conditions, the state transition probabilities are assigned indiscriminately, adhering to a uniform probability distribution. However, given that GRN are inherently sparse, it becomes evident that a uniform distribution model inadequately represents the natural state of GRN. To address this limitation and enhance the realism of transition, we propose reducing the probability of regulatory relationship activation in random networks. In our experiments, the probabilities for $M_{0,1}$ and $M_{0,0}$ are constrained using the number of nodes N_{node} and edges N_{edge}, i.e., $\max(N_{edge}/(N_{node}(N_{node}-1)),0.1)$. We empirically set a lower limit of 0.1 on the probability values to maintain the stability of model training. Furthermore, the noise coefficient in the transition matrix is set to be the commonly used cosine schedule, and the values are determined according to the following formula with $\alpha^t = \bar{\alpha}^t/\bar{\alpha}^{t-1}$ (**Fig. 5A**):

$$\bar{\alpha}^t = \frac{f(t)}{f(0)}, f(t) = \cos\left(\frac{\pi(t/T+0.008)}{2(1+0.008)}\right)^2. \quad (17)$$

Feature enhancement by Meta-cell

Currently, the scRNA-seq techniques, despite their capability of revealing high-resolution biological landscapes that traditional bulk sequencing cannot achieve, inherently introduce

noise and other stochastic effects due to cellular heterogeneity. This phenomenon frequently manifests as dropouts in gene expression measurement. Even among cells of the same type, the capture of low-abundance mRNA can be compromised by technical limitations, resulting in incomplete data. To reduce technical noise and optimize cell representation, we devised a Bagging-based cell ensemble algorithm, called Meta-cell (**Fig. 5B** and **Algorithm 1**). Ideally, cells of the same type would conform to a uniform distribution, but this premise is often challenged by the different stages of differentiation and functional states exhibited by most tissue cells (Baran et al. 2019). Consequently, we modeled the cell population as a dynamic ensemble of distinct cell subsets, each representing a unique state (**Supplementary Note 15**; **Algorithm 1**; **Supplementary Fig. S7A-B**).

Datasets and preprocessing

Simulated single-cell gene expression profiles are used to evaluate model performance

In the simulation-data experiments, SERGIO (Dibaeinia and Sinha 2020) is used to generate scRNA-seq gene expression profiles for synthetic GRN. Specifically, we constructed 100 random GRN, each containing 10 to 100 genes, for performance testing. Additionally, we synthesized 200 further random GRN for model training. Utilizing the stochastic differential equation algorithm embedded in SERGIO, we simulated the gene expressions of 100 cells for each GRN, ensuring that the GRN in these simulation datasets are cell-specific, as they correspond to a homogeneous cell type. Representative networks and their corresponding gene expression profiles are displayed in **Figs 5C and D**. A comprehensive overview of our

simulation framework is outlined in Supplementary Files (**Supplementary Note 16**; **Supplementary Fig. S7C**; **Supplementary Fig. S8**). Furthermore, we conducted an analysis of the distribution of gene nodes within these networks (**Fig. 5E**). The simulated data and evaluation results from the DREAM challenge are also discussed in the Supplementary Files (**Supplementary Note 17**; **Supplementary Fig. S9**).

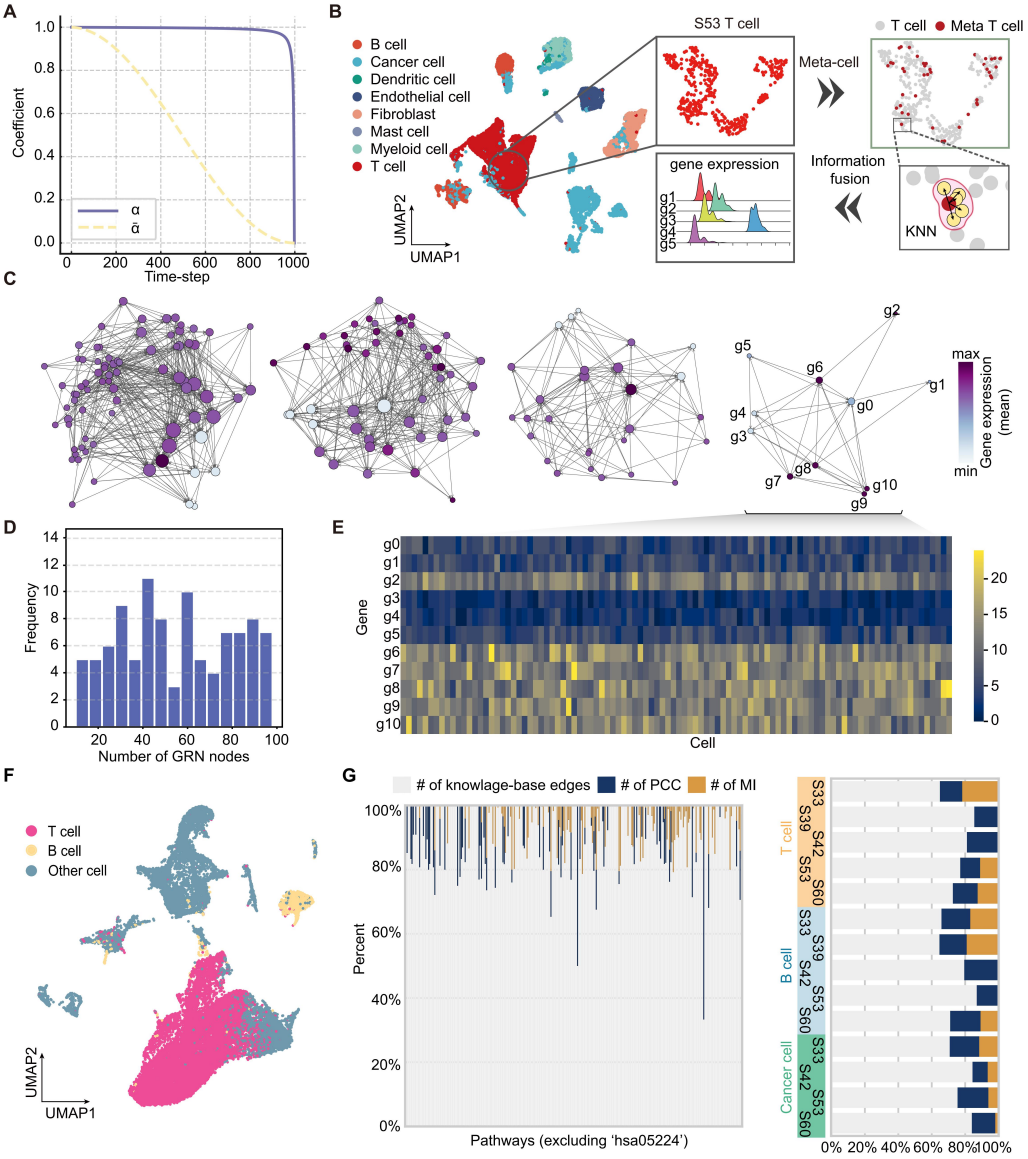

Fig. 5. Details of the DigNet method and benchmark datasets. (A) The decay curve of the noise coefficient alpha as the time-step increases within DigNet. (B) Cell UMAP plots of breast cancer Patients S33-S60 (left) and the detailed process of constructing Meta-cells. (C) Partially synthesized gene network structures (with varying gene numbers of 80, 51, 28, and

11, respectively), with colors indicating their mean gene expression values. (E) Statistics on the number of genes in the synthetic networks. (D) Generating gene expression profiles for networks of different sizes using SERGIO. (F) UMAP of single-cell data for normal and cancerous breast tissues, with merged and labeled distributions of T cells and B cells. (G) Network composition information for a subset of benchmark data used in the DigNet training set (left, the benchmark network for B cells of Patient S53). The KEGG breast cancer pathway hsa05224 is used for performance testing (right, the benchmark network for all cells).

Breast cancer single-cell gene expression profiling data

To demonstrate the capability of DigNet, we use breast cancer as a representative example and expand our case study. We employ real BRCA scRNA-seq data sourced from Qian *et al.* (Qian et al. 2020). For detailed processing steps, please refer to **Supplementary Note 18.**

Immune cell sequencing data of human breast cancer tumor and normal tissues

We downloaded the scRNA-seq data curated by Azizi *et al.* (Azizi et al. 2018), comprising both tumor and control samples, to validate the efficacy of DigNet. Unlike the data provided by Qiao *et al.*, this collection includes original sequencing information about both normal human breast tissue and cancerous conditions. The gene expression profiles are normalized using the "*LogNormalize*" method with a default scale factor based on Seurat v4 (Hao et al. 2021). Additionally, we performed SAVER data imputation to enhance data quality. This comprehensive dataset encompasses sequencing results for diverse immune cell subtypes. To facilitate a broader analysis, we merged cell subtypes according to the annotations provided

by the processed data (**Fig. 5F**). This approach enables a more streamlined comparison and interpretation of the immune landscape across normal and cancerous tissues.

Construct single-cell-specific reference gene regulatory networks

For constructing gold-standard gene regulatory networks, we combined universally recognized reference networks with data-driven specific regulatory relationships recorded in knowledge-bases. Specifically, we utilized an updated version of RegNetwork (Liu et al. 2015) to build the original reference network and extracted edges with high PCC and high MI values from the gene expression profiles, thereby forming the cell-specific gene reference networks (**Supplementary Note 19**). Moreover, in numerical experiments, DigNet was tested on the hsa05224 breast cancer pathway, while utilizing the remaining pathways for training (**Fig. 5G**).

Data sets

The public datasets used in this paper are freely available. The scRNA-seq data of breast cancer can be downloaded from EMBL-EBI ArrayExpress (www.ebi.ac.uk/arrayexpress) under accession number E-MTAB-8107. The scRNA-seq data of normal breast tissue can be downloaded from NCBI GEO database with the accession number GSE195665. Sequencing data of normal and diseased breast immune cells can be downloaded from NCBI GEO database under the accession number GSE114725. Furthermore, the TCGA breast cancer RNA-seq data (Level 3) is downloaded from the UCSC Xena public database (https://xena.ucsc.edu), along with the corresponding clinical information.

Software availability

The source code and pre-trained models used in this study have been distributed across multiple platforms for easy access. The comprehensive source code is included in the Supplemental Code file, which has been uploaded as supplemental material and can also be found on GitHub (https://github.com/zpliulab/DigNet) and Zenodo (https://doi.org/10.5281/zenodo.10907470). Additionally, the Supplemental Material file, encompassing the datasets and pre-trained models, is available in the supplemental material section, as well as on GitHub and Zenodo.

Acknowledgements

This work was partially supported by the Key Program of National Natural Science Foundation of China [Nos. 92374107, 62373216]; the National Key Research and Development Program of China [No. 2020YFA0712402]; the Fundamental Research Funds for the Central Universities [No. 2022JC008].

Author contributions

Z.L. conceived the project. Z.L. and C.W. designed the framework. C.W. collected the data and conducted the experiments. C.W. and Z.L. wrote the manuscript. All authors read and approved the final manuscript.

Competing interest statement

The authors declare no competing interests.

References

Aibar S, González-Blas CB, Moerman T, Huynh-Thu VA, Imrichova H, Hulselmans G, Rambow F, Marine J-C, Geurts P, Aerts J. 2017. SCENIC: single-cell regulatory network inference and clustering. *Nature Methods* **14**: 1083-1086.

Astrinidis A, Kim J, Kelly CM, Olofsson BA, Torabi B, Sorokina EM, Azizkhan-Clifford J. 2010. The transcription factor SP1 regulates centriole function and chromosomal stability through a functional interaction with the mammalian target of rapamycin/raptor complex. *Genes, Chromosomes Cancer* **49**: 282-297.

Aubin-Frankowski P-C, Vert J-P. 2020. Gene regulation inference from single-cell RNA-seq data with linear differential equations and velocity inference. *Bioinformatics* **36**: 4774-4780.

Austin J, Johnson DD, Ho J, Tarlow D, Van Den Berg R. 2021. Structured denoising diffusion models in discrete state-spaces. In *Advances in Neural Information Processing Systems*, Vol 34, pp. 17981-17993.

Azizi E, Carr AJ, Plitas G, Cornish AE, Konopacki C, Prabhakaran S, Nainys J, Wu K, Kiseliovas V, Setty M. 2018. Single-cell map of diverse immune phenotypes in the breast tumor microenvironment. *Cell* **174**: 1293-1308. e1236.

Badia-i-Mompel P, Wessels L, Müller-Dott S, Trimbour R, Ramirez Flores RO, Argelaguet R, Saez-Rodriguez J. 2023. Gene regulatory network inference in the era of single-cell multi-omics. *Nature Reviews Genetics* **24**: 739-754.

Bajrami I, Walker C, Krastev DB, Weekes D, Song F, Wicks AJ, Alexander J, Haider S, Brough R, Pettitt SJ. 2021. Sirtuin inhibition is synthetic lethal with BRCA1 or

BRCA2 deficiency. *Communications Biology* **4**: 1270.

Baran Y, Bercovich A, Sebe-Pedros A, Lubling Y, Giladi A, Chomsky E, Meir Z, Hoichman M, Lifshitz A, Tanay A. 2019. MetaCell: analysis of single-cell RNA-seq data using K-nn graph partitions. *Genome Biology* **20**: 1-19.

Baxter RC. 2014. IGF binding proteins in cancer: mechanistic and clinical insights. *Nature Reviews Cancer* **14**: 329-341.

Chan TE, Stumpf MP, Babtie AC. 2017. Gene regulatory network inference from single-cell data using multivariate information measures. *Cell Systems* **5**: 251-267. e253.

Chen M, Chen Z, Huang D, Sun C, Xie J, Chen T, Zhao X, Huang Y, Li D, Wu B. 2020. Myricetin inhibits TNF-α-induced inflammation in A549 cells via the SIRT1/NF-κB pathway. *Pulmonary Pharmacology Therapeutics* **65**: 102000.

Chen W, Hu L, Lu X, Wang X, Zhao C, Guo C, Li X, Ding Y, Zhao H, Tong D. 2023. The RNA binding protein MEX3A promotes tumor progression of breast cancer by post-transcriptional regulation of IGFBP4. *Breast Cancer Research Treatment* **201**: 353-366.

Choi J-R, Lee S-Y, Shin KS, Choi CY, Kang SJ. 2017. p300-mediated acetylation increased the protein stability of HIPK2 and enhanced its tumor suppressor function. *Scientific Reports* **7**: 16136.

Conrad E, Polonio-Vallon T, Meister M, Matt S, Bitomsky N, Herbel C, Liebl M, Greiner V, Kriznik B, Schumacher S. 2016. HIPK2 restricts SIRT1 activity upon severe DNA damage by a phosphorylation-controlled mechanism. *Cell Death & Differentiation* **23**: 110-122.

Dibaeinia P, Sinha S. 2020. SERGIO: A Single-Cell Expression Simulator Guided by Gene Regulatory Networks. *Cell Systems* **11**: 252-271.e211.

DiToro D, Harbour SN, Bando JK, Benavides G, Witte S, Laufer VA, Moseley C, Singer JR, Frey B, Turner H. 2020. Insulin-like growth factors are key regulators of T helper 17 regulatory T cell balance in autoimmunity. *Immunity* **52**: 650-667. e610.

Duan K-B, Rajapakse JC, Wang H, Azuaje F. 2005. Multiple SVM-RFE for gene selection in cancer classification with expression data. *IEEE Transactions on Nanobioscience* **4**: 228-234.

Dwivedi VP, Bresson X. 2020. A generalization of transformer networks to graphs. *arXiv preprint arXiv:09699*.

Elangovan S, Ramachandran S, Venkatesan N, Ananth S, Gnana-Prakasam JP, Martin PM, Browning DD, Schoenlein PV, Prasad PD, Ganapathy V. 2011. SIRT1 is essential for oncogenic signaling by estrogen/estrogen receptor α in breast cancer. *Cancer Research* **71**: 6654-6664.

Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, Cottarel G, Kasif S, Collins JJ, Gardner TS. 2007. Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biology* **5**: e8.

Flynn KL, Houston KD. 2022. Insulin-like growth factor binding protein 4 exerts differential effects on breast cancer cell proliferation based on the expression status of pregnancy-associated plasma protein A. *Cancer Research* **82**: 108-108.

Gantke T, Sriskantharajah S, Ley SC. 2011. Regulation and function of TPL-2, an IκB

kinase-regulated MAP kinase kinase kinase. *Cell Research* **21**: 131-145.

Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, Vol 27.

Guo Z, Liu J, Wang Y, Chen M, Wang D, Xu D, Cheng J. 2023. Diffusion models in bioinformatics and computational biology. *Nature Reviews Bioengineering*: 1-19.

Hao Y, Hao S, Andersen-Nissen E, Mauck WM, Zheng S, Butler A, Lee MJ, Wilk AJ, Darby C, Zager M. 2021. Integrated analysis of multimodal single-cell data. *Cell* **184**: 3573-3587. e3529.

Hao Y, Stuart T, Kowalski MH, Choudhary S, Hoffman P, Hartman A, Srivastava A, Molla G, Madad S, Fernandez-Granda C. 2023. Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nature Biotechnology*: 1-12.

Haury A-C, Mordelet F, Vera-Licona P, Vert J-P. 2012. TIGRESS: trustful inference of gene regulation using stability selection. *BMC Systems Biology* **6**: 1-17.

Ho J, Chan W, Saharia C, Whang J, Gao R, Gritsenko A, Kingma DP, Poole B, Norouzi M, Fleet DJ. 2022. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:02303*.

Ho J, Jain A, Abbeel P. 2020. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, Vol 33, pp. 6840-6851.

Huang JK, Carlin DE, Yu MK, Zhang W, Kreisberg JF, Tamayo P, Ideker T. 2018. Systematic evaluation of molecular networks for discovery of disease genes. *Cell Systems* **6**: 484-495. e485.

Huang W, Shang W-l, Wang H-d, Wu W-w, Hou S-x. 2012. Sirt1 overexpression protects murine osteoblasts against TNF-α-induced injury in vitro by suppressing the NF-κB signaling pathway. *Acta Pharmacologica Sinica* **33**: 668-674.

Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P. 2010. Inferring regulatory networks from expression data using tree-based methods. *PloS One* **5**: e12776.

Iwata TN, Ramírez-Komo JA, Park H, Iritani BM. 2017. Control of B lymphocyte development and functions by the mTOR signaling pathways. *Cytokine Growth Factor Reviews* **35**: 47-62.

Karlebach G, Shamir R. 2008. Modelling and analysis of gene regulatory networks. *Nature Reviews Molecular Cell Biology* **9**: 770-780.

Kingma DP, Welling M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:02303*.

Kotiang S, Eslami A. 2020. A probabilistic graphical model for system-wide analysis of gene regulatory networks. *Bioinformatics* **36**: 3192-3199.

Langendonk M, de Jong MR, Smit N, Seiler J, Reitsma B, Ammatuna E, Glaudemans AW, van den Berg A, Huls GA, Visser L. 2022. Identification of the estrogen receptor beta as a possible new tamoxifen-sensitive target in diffuse large B-cell lymphoma. *Blood Cancer Journal* **12**: 36.

Lê Cao K-A, Boitard S, Besse P. 2011. Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. *BMC Bioinformatics* **12**: 1-17.

Lee Y-Y, Mok MT, Kang W, Yang W, Tang W, Wu F, Xu L, Yan M, Yu Z, Lee S-D et al. 2018. Loss of tumor suppressor IGFBP4 drives epigenetic reprogramming in hepatic

carcinogenesis. *Nucleic Acids Research* **46**: 8832-8847.

Levi L, Lobo G, Doud MK, Von Lintig J, Seachrist D, Tochtrop GP, Noy N. 2013. Genetic ablation of the fatty acid–binding protein FABP5 suppresses HER2-induced mammary tumorigenesis. *Cancer Research* **73**: 4770-4780.

Levine M, Davidson EH. 2005. Gene regulatory networks for development. *Proceedings of the National Academy of Sciences* **102**: 4936-4942.

Li J-N, Chen P-S, Chiu C-F, Lyu Y-J, Lo C, Tsai L-W, Wang M-Y. 2021. TARBP2 suppresses ubiquitin-proteasomal degradation of HIF-1α in breast cancer. *International Journal of Molecular Sciences* **23**: 208.

Liu B, Du R, Zhou L, Xu J, Chen S, Chen J, Yang X, Liu D-x, Shao Z-m, Zhang L. 2018. miR-200c/141 regulates breast cancer stem cell heterogeneity via targeting HIPK1/β-catenin axis. *Theranostics* **8**: 5801.

Liu F, Zhang S-W, Guo W-F, Wei Z-G, Chen L. 2016. Inference of gene regulatory network based on local Bayesian networks. *PLoS Computational Biology* **12**: e1005024.

Liu R-Z, Graham K, Glubrecht DD, Germain DR, Mackey JR, Godbout R. 2011. Association of FABP5 expression with poor survival in triple-negative breast cancer: implication for retinoic acid therapy. *The American Journal of Pathology* **178**: 997-1008.

Liu Z-P, Wu C, Miao H, Wu H. 2015. RegNetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse. *Database* **2015**: bav095.

Lu W, Ning H, Gu L, Peng H, Wang Q, Hou R, Fu M, Hoft DF, Liu J. 2016. MCPIP1 selectively destabilizes transcripts associated with an antiapoptotic gene expression

program in breast cancer cells that can elicit complete tumor regression. *Cancer Research* **76**: 1429-1440.

Ma A, Wang X, Li J, Wang C, Xiao T, Liu Y, Cheng H, Wang J, Li Y, Chang Y et al. 2023. Single-cell biological network inference using a heterogeneous graph transformer. *Nature Communications* **14**: 964.

Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Favera RD, Califano A. 2006. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* **7**: 1-15.

Matsumoto H, Kiryu H, Furusawa C, Ko MS, Ko SB, Gouda N, Hayashi T, Nikaido I. 2017. SCODE: an efficient regulatory network inference algorithm from single-cell RNA-Seq during differentiation. *Bioinformatics* **33**: 2314-2321.

Matsushita K, Takeuchi O, Standley DM, Kumagai Y, Kawagoe T, Miyake T, Satoh T, Kato H, Tsujimura T, Nakamura H. 2009. Zc3h12a is an RNase essential for controlling immune responses by regulating mRNA decay. *Nature* **458**: 1185-1190.

Miyagawa I, Nakayamada S, Nakano K, Yamagata K, Sakata K, Yamaoka K, Tanaka Y. 2017. Induction of regulatory T cells and its regulation with insulin-like growth factor/insulin-like growth factor binding protein-4 by human mesenchymal stem cells. *The Journal of Immunology* **199**: 1616-1625.

Moris N, Pina C, Arias AM. 2016. Transition states and cell fate decisions in epigenetic landscapes. *Nature Reviews Genetics* **17**: 693-703.

Papili Gao N, Ud-Dean SM, Gandrillon O, Gunawan R. 2018. SINCERITIES: inferring gene regulatory networks from time-stamped single cell transcriptional expression profiles.

Bioinformatics **34**: 258-266.

Perez E, Strub F, De Vries H, Dumoulin V, Courville A. 2018. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, Vol 32.

Puca R, Nardinocchi L, Sacchi A, Rechavi G, Givol D, D'Orazi G. 2009. HIPK2 modulates p53 activity towards pro-apoptotic transcription. *Molecular Cancer* **8**: 1-14.

Qian J, Olbrecht S, Boeckx B, Vos H, Laoui D, Etlioglu E, Wauters E, Pomella V, Verbandt S, Busschaert P. 2020. A pan-cancer blueprint of the heterogeneous tumor microenvironment revealed by single-cell profiling. *Cell Research* **30**: 745-762.

Ramesh A, Dhariwal P, Nichol A, Chu C, Chen M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:06125* **1**.

Ranjan B, Sun W, Park J, Mishra K, Schmidt F, Xie R, Alipour F, Singhal V, Joanito I, Honardoost MA. 2021. DUBStepR is a scalable correlation-based feature selection method for accurately clustering single-cell data. *Nature Communications* **12**: 5849.

Ruan P, Wang S, Yang C, Huang X, Sun P, Tan A. 2023. m6A mRNA methylation regulates the ERK/NF-κB/AKT signaling pathway through the PAPPA/IGFBP4 axis to promote proliferation and tumor formation in endometrial cancer. *Cell Biology Toxicology* **39**: 1611-1626.

Ryan AJ, Napoletano S, Fitzpatrick PA, Currid CA, O'Sullivan NC, Harmey JH. 2009. Expression of a protease-resistant insulin-like growth factor-binding protein-4 inhibits tumour growth in a murine model of breast cancer. *British Journal of Cancer* **101**: 278-286.

Saharia C, Chan W, Saxena S, Li L, Whang J, Denton EL, Ghasemipour K, Gontijo Lopes R, Karagol Ayan B, Salimans T. 2022. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems*, Vol 35, pp. 36479-36494.

Senga S, Kobayashi N, Kawaguchi K, Ando A, Fujii H. 2018. Fatty acid-binding protein 5 (FABP5) promotes lipolysis of lipid droplets, de novo fatty acid (FA) synthesis and activation of nuclear factor-kappa B (NF-κB) signaling in cancer cells. *Biochimica et Biophysica Acta -Molecular Cell Biology of Lipids* **1863**: 1057-1067.

Shu H, Zhou J, Lian Q, Li H, Zhao D, Zeng J, Ma J. 2021. Modeling gene regulatory networks using neural network architectures. *Nature Computational Science* **1**: 491-501.

Specht AT, Li J. 2017. LEAP: constructing gene co-expression networks for single-cell RNA-sequencing data using pseudotime ordering. *Bioinformatics* **33**: 764-766.

Stimper V, Schölkopf B, Hernández-Lobato JM. 2022. Resampling base distributions of normalizing flows. In *International Conference on Artificial Intelligence and Statistics*, pp. 4915-4936. PMLR.

Sung JY, Kim R, Kim JE, Lee J. 2010. Balance between SIRT1 and DBC1 expression is lost in breast cancer. *Cancer Science* **101**: 1738-1744.

Tran E, Robbins PF, Lu Y-C, Prickett TD, Gartner JJ, Jia L, Pasetto A, Zheng Z, Ray S, Groh EM. 2016. T-cell transfer therapy targeting mutant KRAS in cancer. *New England Journal of Medicine* **375**: 2255-2262.

Tsatsanis C, Vaporidi K, Zacharioudaki V, Androulidaki A, Sykulev Y, Margioris AN, Tsichlis

PN. 2008. Tpl2 and ERK transduce antiproliferative T cell receptor signals and inhibit transformation of chronically stimulated T cells. *Proceedings of the National Academy of Sciences* **105**: 2987-2992.

Wagner A, Regev A, Yosef N. 2016. Revealing the vectors of cellular identity with single-cell genomics. *Nature Biotechnology* **34**: 1145-1160.

Wang C, Gao Y, Zhang Z, Chi Q, Liu Y, Yang L, Xu K. 2020. Safflower yellow alleviates osteoarthritis and prevents inflammation by inhibiting PGE2 release and regulating NF-κB/SIRT1/AMPK signaling pathways. *Phytomedicine* **78**: 153305.

Wang J, Chen Y, Zou Q. 2023. Inferring gene regulatory network from single-cell transcriptomes with graph autoencoder model. *PLoS Genetics* **19**: e1010942.

Wang X, Ghasedi Dizaji K, Huang H. 2018. Conditional generative adversarial network for gene expression inference. *Bioinformatics* **34**: i603-i611.

Way GP, Zietz M, Rubinetti V, Himmelstein DS, Greene CS. 2020. Compressing gene expression data using multiple latent space dimensionalities learns complementary biological representations. *Genome Biology* **21**: 1-27.

Wilmott JS, Long GV, Howle JR, Haydu LE, Sharma RN, Thompson JF, Kefford RF, Hersey P, Scolyer RAJCcr. 2012. Selective BRAF inhibitors induce marked T-cell infiltration into human metastatic melanoma. *Clin Cancer Res* **18**: 1386-1394.

Xiao P, Chen Y, Jiang H, Liu YZ, Pan F, Yang TL, Tang ZH, Larsen JA, Lappe JM, Recker RR. 2008. In vivo genome-wide expression study on human circulating B cells suggests a novel ESR1 and MAPK3 network for postmenopausal osteoporosis. *Journal of Bone Mineral Research* **23**: 644-654.

Zhang F, Qi L, Feng Q, Zhang B, Li X, Liu C, Li W, Liu Q, Yang D, Yin Y. 2021. HIPK2 phosphorylates HDAC3 for NF-κB acetylation to ameliorate colitis-associated colorectal carcinoma and sepsis. *Proceedings of the National Academy of Sciences* **118**: e2021798118.

Zhang W, Xu H, Qiao R, Zhong B, Zhang X, Gu J, Zhang X, Wei L, Wang X. 2022. ARIC: accurate and robust inference of cell type proportions from bulk gene expression or DNA methylation data. *Briefings in Bioinformatics* **23**: bbab362.

Zhang W, Zhang X, Liu Q, Wei L, Qiao X, Gao R, Liu Z, Wang X. 2023. Deconer: A comprehensive and systematic evaluation toolkit for reference-based cell type deconvolution algorithms using gene expression data. *bioRxiv* doi:10.1101/2023.12.24.573278: 2023.2012.2024.573278.

Zhu JW, Field SJ, Gore L, Thompson M, Yang H, Fujiwara Y, Cardiff RD, Greenberg M, Orkin SH, DeGregori J. 2001. E2F1 and E2F2 determine thresholds for antigen-induced T-cell proliferation and suppress tumorigenesis. *Molecular and Cellular Biology* **21**: 8547-8564.

Zimmerli D, Brambillasca CS, Talens F, Bhin J, Linstra R, Romanens L, Bhattacharya A, Joosten SE, Da Silva AM, Padrao N. 2022. MYC promotes immune-suppression in triple-negative breast cancer via inhibition of interferon signaling. *Nature Communications* **13**: 6579.