Journal of Translational Medicine

RESEARCH

Open Access

Investigation of cell development and tissue structure network based on natural Language processing of scRNA-seq data

Suwen Wei[1†], Yuer Lu[1†], Peng Wang[1,2], Qichao Li[1,2], Jianwei Shuai[1], Qi Zhao[1,3*], Hai Lin[1*] and Yuming Peng[4*]

Abstract

Background Single-cell multi-omics technologies, particularly single-cell RNA sequencing (scRNA-seq), have revolutionized our understanding of cellular heterogeneity and development by providing insights into gene expression at the single-cell level. Investigating the influence of genes on cellular behavior is crucial for elucidating cell fate determination and differentiation, cell development processes, and disease mechanisms.

Methods Inspired by NLP, we present a novel scRNA-seq analysis method that treats genes as analogous to words. Using word2vec to embed gene sequences derived from gene networks, we generate vector representations of genes, which are then used to represent cells by summing gene vectors and subsequently tissues by aggregating cell vectors.

Results Our NLP-based approach analyzes scRNA-seq data by generating vector representations of genes, cells, and tissues. This multi-scale analysis includes mapping cell states in vector space to reveal developmental trajectories, quantifying cell similarity using Euclidean distance, and constructing inter-tissue relationship networks from aggregated cell vectors.

Conclusions This method offers a computationally efficient approach for analyzing scRNA-seq data by constructing embedding representations similar to those used in large language model pre-training, but without requiring high-performance computing clusters. By generating gene embeddings that capture functional relationships, this method facilitates the study of cell development trajectories, the impact of gene perturbations, cell clustering, and the construction and analysis of tissue networks. This provides a valuable tool for single-cell data analysis.

Keywords Single-cell RNA sequencing technology, Natural Language processing, Embedding, Network structure

[†]Suwen Wei and Yuer Lu contributed equally to this work.

*Correspondence:
Qi Zhao
zhaoqi@lnu.edu.cn
Hai Lin
hailin@ucas.ac.cn
Yuming Peng
1584709828@qq.com

[1]Oujiang Laboratory (Zhejiang Lab for Regenerative Medicine, Vision and Brain Health), Wenzhou Institute, University of Chinese Academy of Sciences, Wenzhou 325001, Zhejiang, P. R. China
[2]Postgraduate Training Base Alliance of Wenzhou Medical University, Wenzhou 325001, Zhejiang, P. R. China
[3]School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan 114051, P.R. China
[4]Department of General Practice, Central Hospital of Karamay, Xinjiang 834000, P. R. China

Introduction

The convergence of biology and computer science has established bioinformatics as a crucial interdisciplinary field, employing computational methods and tools to decipher the complexity of biological systems. This field encompasses diverse areas, including genomics, transcriptomics, and proteomics. Progress in these areas is propelled by the development of sophisticated analytical methods and technologies, enabling researchers to investigate various facets of biological tissues and cells [1]. Recent advances in single-cell multi-omics technologies have enabled the simultaneous analysis of multiple biomolecules at single-cell resolution, providing a more comprehensive understanding of cellular heterogeneity and function. Among these technologies, single-cell RNA sequencing (scRNA-seq) has emerged as a transformative technique, enabling the measurement of gene expression at the single-cell level and generating high-resolution transcriptomic profiles [2–4]. These profiles can capture the inherent temporal dynamics of differentiating cells [5–7], offering valuable insights into cellular processes such as growth, development, and differentiation [8, 9]. By facilitating the analysis of functional and expression heterogeneity between individual cells, scRNA-seq allows for a deeper understanding of the diverse roles cells play in biological processes [10–14]. Consequently, scRNA-seq has become an indispensable tool in cell biology research, providing critical insights into cell development processes, cell fate determination, and disease mechanisms, while also driving progress in transcriptomics [15–18].

Natural language processing (NLP) is a crucial field within computer science dedicated to developing computational methods for understanding and processing natural language text and speech [19, 20]. NLP models can generate predicted outputs for test instances by completing the word sequences of input text. Pre-training, where a language model is initially trained on a large dataset and subsequently fine-tuned for a specific task, has proven to be an effective strategy for building robust word prediction models. Among these models, Word2vec employs a simplified shallow neural network to learn distributed word representations and has demonstrated remarkable effectiveness in a variety of NLP tasks [21].

High-dimensional gene expression data from single-cell RNA sequencing (scRNA-seq) requires effective representation methods to capture its structure for better understanding [22–24]. While visualization is important for human interpretation, the core challenge is effectively embedding this high-dimensional data into lower dimensions while preserving crucial biological information [25]. Effective embeddings must retain both local and global structure during dimensionality reduction [26]. Although methods like Principal Component Analysis (PCA), t-Distributed Stochastic Neighbor Embedding (t-SNE) [27], and Uniform Manifold Approximation and Projection (UMAP) are commonly used for dimensionality reduction and embedding [28], they are not always optimal for exploring high-dimensional data [29]. For instance, PCA is sensitive to noise, which cannot be explicitly eliminated, and the embedding can distort the global structure of the data [30]. These traditional methods primarily perform dimensionality reduction based on the geometric or statistical properties of the data, aiming to preserve the distance or local neighborhood relationships between data points, typically using Euclidean distance or similar metrics. In contrast, language model-based embeddings are learned from co-occurrence and semantic similarity of words or sentences within a context, capturing deeper semantic information beyond superficial geometric relationships. Consequently, when data contains complex semantic structures, such as gene regulatory relationships or cell state transition trajectories in gene expression data, language model-based embeddings may be more effective at revealing these hidden patterns. Therefore, to better capture the complex biological information within scRNA-seq data, such as continuous trajectories of cell states or regulatory relationships between genes, novel embedding methods are needed.

We have developed a computationally efficient, lightweight natural NLP method to address the challenges of analyzing high-dimensional single-cell data. Unlike large-scale foundation models for single-cell analysis [31–33], our approach is computationally less demanding, yet generates effective embeddings for efficient analysis and visualization of high-dimensional single-cell data. Using single-cell data for pre-training, our model generates embeddings that enable the following functionalities: (1) More accurately depict the cell development process; (2) Analysis of perturbations and prediction of responses, including inference of cell perturbations, analysis of developmental processes in cell populations at different stages, and investigation of gene sequences underlying transcriptional changes that may drive cell fate decisions, providing insights into the principles of cell growth and development; and (3) Determination of tissue network structure models through partitioning or clustering. These models help understand how cells connect and form tissues, thereby overcoming the challenges that traditional methods face in statistically modeling cell connectivity. This lightweight NLP method offers an efficient solution for analyzing and visualizing single-cell data and holds promise for providing deeper insights into single-cell growth and development.

Datasets

For dimensionality reduction visualization and gene perturbation analysis, we utilized the human embryonic stem cell (hESC)-derived embryoid bodies (EBs) dataset [25] and the Zebrafish Embryos (ZE) dataset [34]. The EBs dataset comprises 16,825 single-cell data samples spanning five time points over 27 days of differentiation. The ZE dataset, generated using Drop-seq technology, includes 38,309 cells from 28 samples covering 12 distinct differentiation stages during zebrafish embryonic development, from the blastula to the somatic period. These two datasets include multiple developmental time points and important gene regulatory information, making them ideal for studying cell differentiation, developmental pathways, and gene perturbations.

To investigate gene and tissue network structures, we employed the fetal kidney temporal-spatial immune partition dataset [35] and the Human Lung Cancer dataset [36]. The fetal kidney dataset contains 21,452 cells, partitioned into four compartments: immune, vasculature, developing nephron, and stroma. The Human Lung Cancer dataset consists of single-cell RNA sequencing data from 49 clinical biopsy specimens obtained before and during targeted therapy from 30 patients with metastatic lung cancer. These two datasets represent normal tissue development and disease states, making them suitable for constructing tissue networks and exploring disease mechanisms.

During data preprocessing, we used the raw data directly in the gene network construction phase to better capture the relationships among genes. For constructing cell embeddings, we normalized the gene expression matrix and used the normalized gene expression levels as weights to sum the gene vectors.

Methods

Our natural language processing (NLP)-based model for processing scRNA-seq data begins by calculating the cosine similarity between gene expression profiles from the raw single-cell transcriptome data. This similarity measure is then used to construct a gene distance matrix, which represents a gene network. We perform random walks on this gene network to generate gene sequences. These sequences are treated as text and embedded into a vector space using the word2vec algorithm [37], resulting in gene vectors. Subsequently, gene vectors are aggregated by cell to obtain cell vectors, and cell vectors are further aggregated by tissue to derive tissue vectors. This hierarchical aggregation approach enables dimensionality reduction and allows us to infer pseudo-time trajectories, represent gene perturbations during development, and visualize tissue network structures (Fig. 1).

Dimensionality reduction and embedding of scRNA-seq data

Single-cell RNA sequencing technology has provided unprecedented insights into complex biological systems [38]. Due to the high dimensionality and sparsity of scRNA-seq data, similarity measures such as cosine similarity and Pearson correlation are commonly employed to quantify relationships between gene expression profiles [39]. To capture potential relationships between genes and facilitate downstream analyses, including trajectory inference and visualization, we employ word embedding algorithms to generate gene vectors from the generated gene sequences. Word embedding techniques map high-dimensional elements into a lower-dimensional continuous vector space, representing each element as a vector in the real domain.

To capture potential relationships between genes, we analyzed the raw count matrix from scRNA-seq data. This matrix has dimensions of $m \times n$, where m is the number of cells and n is the number of genes. Each column represents the expression levels of a specific gene across all cells, so the gene expression vector Gi for gene i is an m-dimensional vector. To measure the similarity between gene i and gene j, we calculate the cosine similarity of these two gene expression vectors:

$$\cos[i,j] = \frac{(Gi \bullet Gj)}{(|Gi| * |Gj|)}$$

where|Gi| and|Gj| represent the L2 norms of the respective vectors. We calculated the cosine similarity for all gene pairs, generating an $n \times n$ gene similarity matrix. This matrix served as the adjacency matrix for a gene co-expression network, where nodes represent genes and edge weights are the cosine similarity scores, indicating the strength of gene-gene relationships.

Based on this gene network, we use random walks to generate gene sequences. Random walks are an effective method for simulating gene interaction relationships within a network. The gene sequences generated by random walks capture both local and global structural relationships, reflecting functional associations between genes. Random walks can model co-expression relationships. for example, if two genes frequently appear together, this suggests a potential functional association. Additionally, a gene may appear multiple times in the same sequence, which enhances its representation in the embedding space. This repetition reflects the gene's importance within the network and its involvement in multiple functional modules or regulatory pathways.

In the random walk process, we can either systematically traverse each gene or randomly select genes as starting nodes. In our simulations, we choose to traverse all genes as starting nodes. After establishing the initial

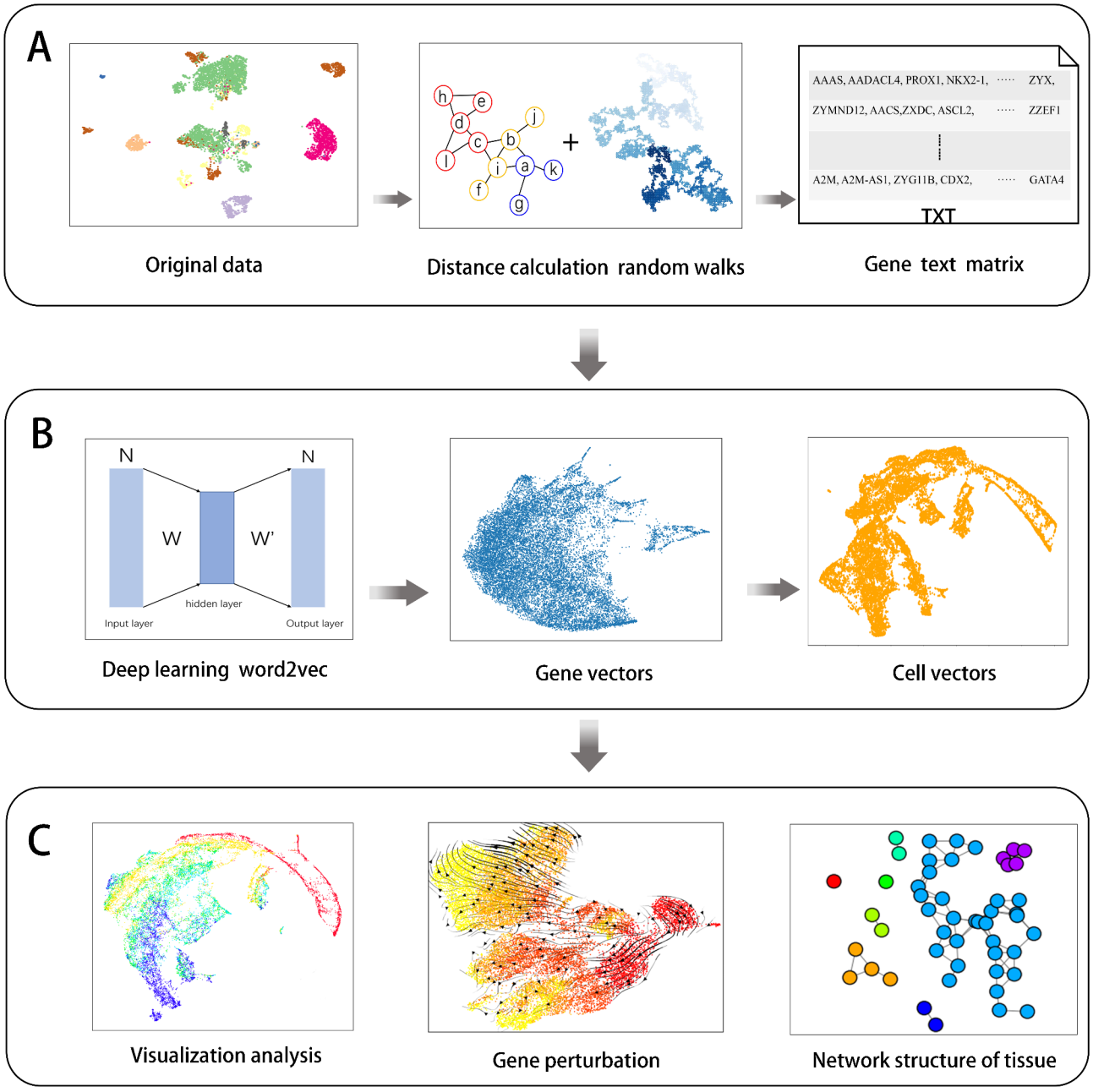

Fig. 1 scRNA-seq data processing and analysis pipeline using natural language processing (NLP). (**A**) Construction of a gene similarity network based on gene expression profiles and generation of gene sequences using random walks. (**B**) Conversion of gene sequences to gene vectors using the word2vec model, and calculation of cell vectors by aggregating gene vectors weighted by gene expression levels. (**C**) Downstream applications of cell vectors, including visualization analysis, gene perturbation analysis, and tissue network structure analysis

node, we perform a random walk by moving to a neighboring node with a probability proportional to the corresponding edge weight. During the random walk, each visited gene is recorded, forming a gene sequence.

To explore the gene network, we conducted n random walks, each with n steps, where n is the number of genes in the gene network. This setup ensures an adequate balance between the sampling of each gene and computational efficiency. We observed that the choice of the initial node, whether selected randomly or by traversing the genes, does not significantly affect the results with this parameter setting. Despite the stochastic nature of the random walks, including variations in the random seeds and initial conditions, the generated gene embeddings remain highly consistent across multiple runs. This consistency ensures the stability of subsequent downstream analyses. Furthermore, the n-step walk length captures gene-gene relationships effectively, and extending the

walk length further would significantly increase computational costs with only marginal improvements.

After performing random walks, we obtain a set of gene sequences forming a text corpus, which contains n sentences composed of gene "words", with each sentence also having a length of n. Subsequently, we utilize this gene text corpus and the word2vec model to obtain the embedding of each "word", i.e., each gene.

We used the Gensim library [40] to implement the word2vec model, which is a widely used Python library for efficient word embedding generation. We adopted Gensim's default parameter settings, specifically using the Continuous Bag-of-Words (CBOW) model with a vector dimension of 100, a window size of 5, and a minimum word frequency of 5. As a result, we obtained n gene vectors, each with 100 dimensions, forming a gene vector matrix of size $n \times 100$.

Since each single cell expresses a set of genes, the cell vector is calculated as the weighted sum of its expressed gene vectors, where the weights are the gene expression levels in that cell. The resulting cell vector matrix is $m \times 100$:

$$C_{mk} = \sum_i a_{mi} g_{ik}$$

where C_{mk} represents the k-th dimension of the m-th cell vector, a_{mi} represents the expression level of the i-th gene in the m-th cell, and g_{ik} represents the k-th dimension of the i-th gene vector.

After obtaining cell vectors, we perform further dimensionality reduction and visualization to explore cell relationships within the embedded space. This enables us to identify potential pseudo-time trajectories, observe developmental dynamics, and reveal key cell types and differentiation pathways.

Gene perturbation analysis in embedding space

The rapid advancement of genome sequencing technologies [41, 42] has fueled the need to understand how genomic variations influence organismal phenotypes, linking identified genomic variations to phenotypic variations in health and disease [43, 44]. Analyzing the transcriptional response of cells to genetic perturbations provides crucial insights into cellular functions and regulatory mechanisms [45]. For instance, it helps in understanding how specific genes regulate cell growth and differentiation. While experimental perturbation studies are invaluable, computational methods can complement these approaches by predicting the effects of perturbations in silico, especially in the context of high-dimensional single-cell data. Our embedding-based approach provides an effective method to analyze gene perturbations within the learned latent space.

Unlike methods based solely on transcriptional kinetics or RNA velocity [46], our approach leverages the learned gene embeddings to model the impact of perturbations. Since cell vectors are calculated as the weighted sum of their expressed gene vectors, perturbing a gene's representation directly affects the cell vectors in the embedding space. Specifically, to simulate the effect of a gene perturbation (e.g., overexpression or knockout), we modify the corresponding gene vector and recompute the cell vectors. For example, to simulate gene overexpression, we can add a scaled version of the gene vector to the original gene vector. Conversely, to simulate gene knockout, we can set the corresponding gene vector to zero or remove its contribution from the cell vectors.

The effect of the perturbation on cell m can be quantified by comparing the original cell vector C_m with the perturbed cell vector C'_m. This comparison can be performed by calculating the vector difference (C'_m - C_m). The direction and magnitude of the vector difference reveal the direction and strength of the perturbation's effect on the cell's representation in the embedding space. By visualizing the original and perturbed cell vectors, we can observe how the perturbation shifts the cell's position within the embedded space, providing insights into its potential impact on cell state and developmental trajectory. This approach allows us to predict the positive or negative effect of gene perturbation on cell development.

Gene perturbation vector fields were constructed following the scVelo approach [16]. Briefly, we simulated gene overexpression or knockout by modifying gene vectors and recalculating the corresponding cell vectors. The resulting vector differences, representing pre- and post-perturbation changes, capture shifts in cell states. We then computed the cosine similarity between these differences and neighboring cells to generate a similarity matrix, reflecting the intensity and direction of the perturbation effects. From this matrix, we derived weighted displacement vectors to represent cell movement in latent space. To ensure a smooth and continuous vector field, we applied Gaussian kernel smoothing to interpolate these vectors onto a grid. Finally, we visualized cell state transitions using Matplotlib's quiver or streamplot functions, with arrows or streamlines indicating the direction and magnitude of cell "flow".

Network construction from cell and cell embeddings

Quantitatively characterizing, understanding, and modeling cell-cell interactions within tissues is a key challenge in contemporary biology [47–49]. While constructing gene networks from transcriptome data provides insights into gene relationships [50], analyzing cell-cell interactions requires considering the collective behavior of cells within their tissue context. However, measuring biological distances at a local scale, especially with limited data

from rare diseases or difficult-to-access tissues, can hinder the capture of overall structural information at the tissue level.

To address this, we first construct a cell network based on cell embeddings and use it as a foundation for constructing the tissue network. We leverage the previously generated cell vectors, which capture gene expression patterns within individual cells. By calculating the pairwise distances between cell vectors, we can construct a cell network to gain a preliminary understanding of the relationships between cells and provide context for subsequent tissue-level analysis.

Building upon this, we further construct a tissue network to explore relationships between different tissues. Here, "tissue" refers to a collection of cells, not necessarily a specific anatomical tissue type. For instance, in the context of lung cancer data analysis, we define a "tissue" as the collection of cells from a single patient.

Tissue vectors are calculated by aggregating the cell vectors belonging to the same "tissue". Specifically, we sum all cell vectors within the same "tissue" to obtain the corresponding tissue vector. This aggregation captures the overall gene expression profile of the cell population within that tissue.

Once tissue vectors are obtained, we construct a tissue network to explore relationships between different tissues. This network is constructed by calculating the pairwise distances between tissue vectors, where each node represents a tissue and the connections between nodes represent the proximity or similarity between tissues. By analyzing this network, we can observe relationships between tissues, such as identifying clusters of tissues with similar characteristics, which may reflect shared biological functions or disease states. For example, in lung cancer research, this network can be used to identify patient subgroups with similar gene expression profiles, or discover potential associations between different tissues, thereby better understanding disease development.

Results

Visualization analysis of EBs and ZE datasets

Our model analyzes the development and differentiation of embryonic cells through gene similarity using natural language processing techniques. Utilizing single-cell transcriptome sequencing datasets from human embryonic stem cells and zebrafish embryos, our model extracts informative data on cell development. By constructing a gene embedding space and aggregating gene expression vectors of embryonic cells according to their expression weights, we generate cell vector representations in gene space.

To visualize the continuous developmental process of human embryonic stem cells from Day 1 to Day 5, we employed three dimensionality reduction methods: PCA, UMAP, and UMAP based on our calculated cell embeddings (NLP-UMAP), with the results shown in Fig. 2A. As shown in the figure, PCA fails to capture the branching structure information in cell development, resulting in a linear trend in the developmental trajectory that fails to reflect the complexity of cell differentiation. While

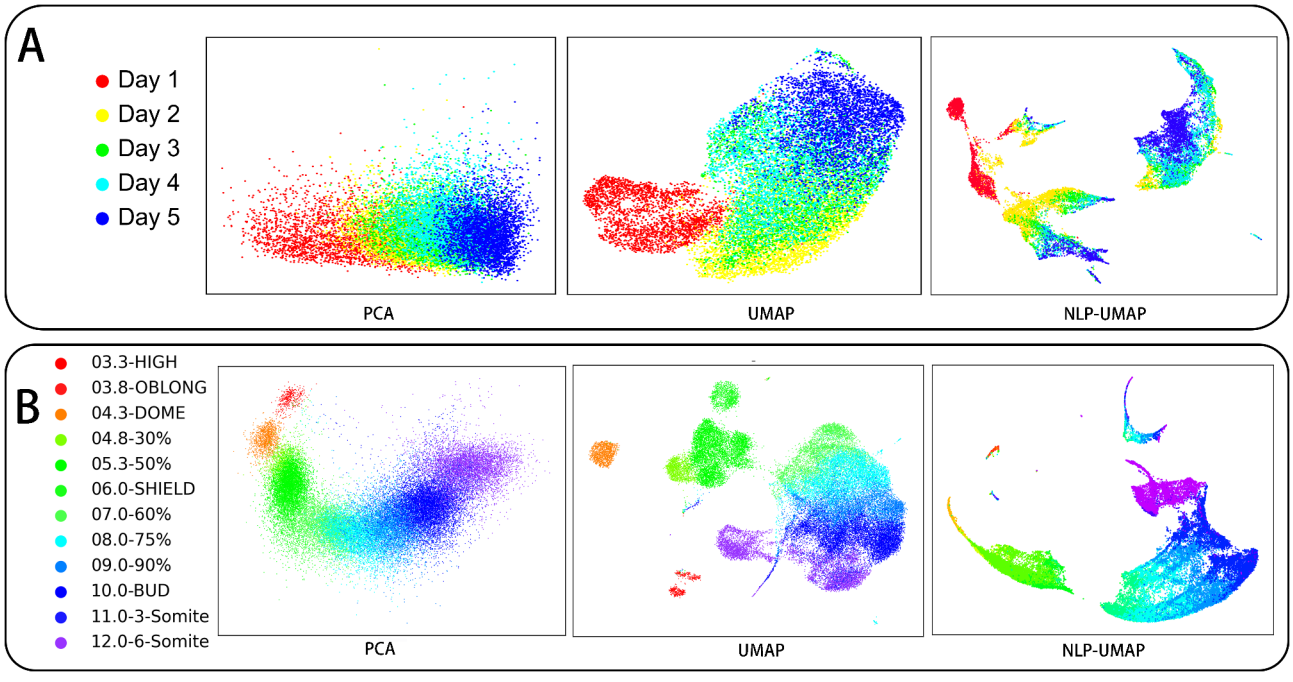

Fig. 2 Comparison of our model (NLP-UMAP) with PCA and UMAP applied to human embryonic stem cell (EBs) and zebrafish embryo (ZE) data. (**A**) EBs developmental trajectory. (**B**) ZE developmental trajectory

UMAP shows some separation of cell populations, its overall structure is rather loose, and the data points are distributed sparsely, making it difficult to clearly reveal the continuous developmental trajectory and branching relationships. In contrast, our NLP-UMAP method better preserves both the global and local structures of the data, clearly revealing branching trajectories in cell differentiation, indicating its improved ability to capture cell differentiation and development.

We also validated our approach on a zebrafish embryo dataset [34]. As shown in Fig. 2B, the NLP-UMAP visualization of cell vectors clearly illustrates the time evolution trajectory of zebrafish embryonic cells in gene space, with distinct color transitions from red to purple signifying different degrees of cell differentiation. Similar to the EB dataset, our method also reveals more branching and differentiation details in this dataset. These results further support the possibility that the gene embedding space constructed through natural language processing techniques may effectively capture semantic relationships between genes, thereby better reflecting the processes of cell development and differentiation.

Gene perturbation of EBs and ZE datasets

Changes in gene vectors lead to corresponding changes in cell vectors. To explore the effect of gene perturbation on cell developmental trajectories, we performed in silico gene perturbation experiments on EBs and ZE single-cell transcriptome sequencing data, observing the resulting changes in cell vectors visualized by NLP-UMAP (Fig. 3).

In the EBs data (Fig. 3A), we focused on genes known to play crucial roles in early development and differentiation. For instance, POU5F1 and OTX2 are key transcription factors known to induce EBs differentiation [25]. As described in [25], differentiation in the extra-embryonic lineage initiates with the induction of the anterior neuroectoderm state, characterized by POU5F1 down-regulation and OTX2 up-regulation. To simulate this process, we performed in silico perturbations by individually up-regulating OTX2, individually down-regulating POU5F1 (0.2-fold), and simultaneously up-regulating OTX2 (2-fold) and down-regulating POU5F1 (0.2-fold). As shown in the UMAP plots in Fig. 3A, individually up-regulating OTX2 and individually down-regulating POU5F1 shifted the cell trajectory in distinct directions. Importantly, simultaneously up-regulating OTX2 and down-regulating POU5F1 resulted in a cell trajectory shift more closely aligned with the expected anterior neuroectoderm state, demonstrating the model's ability to capture synergistic gene effects. Furthermore, we tested that up-regulating GATA3, SATB1, and KLF8, while down-regulating EOMES and FOXA2, can further differentiate into cells expressing the endoderm, with the cell developmental direction consistent with [25].

We also applied the model to study several genes with important roles in ZE development and differentiation (Fig. 3B). We performed in silico up-regulation of the prechordal plate marker GSC, the late notochord marker NTD5 [34], and SOX32 [51], involved in endoderm development, to investigate their effects on cell developmental trajectories. GSC plays a crucial role in regulating embryonic axis formation during early developmental stages, transmitting important embryonic signals that promote normal embryonic development. As shown in Fig. 3B, in silico increasing GSC expression shifted the cell trajectory towards later developmental stages, suggesting that GSC up-regulation may enhance its regulatory role in axis formation, potentially leading to more effective axis development. NTD5 is essential for notochord development [34], and in silico predictions suggest that up-regulating NTD5 expression may impact notochord development, influencing both proper neuronal differentiation and neural structure formation. We also explored branching events driven by single genes, such as the role of SOX32 in endoderm development [51]. As shown in Fig. 3B, in silico up-regulation of SOX32 led to branching in the cell trajectory, demonstrating the model's ability to capture single-gene-driven branching events.

In summary, these results demonstrate that our model can effectively simulate the effects of gene perturbation on cell developmental trajectories, in both EBs differentiation and ZE embryogenesis. The model captures both single-gene and multi-gene synergistic effects and can predict the developmental direction of cells under different perturbation conditions.

Tissue network structure in fetal kidneys and human lung Cancer

We applied our model to construct tissue-level networks based on cell type composition and gene similarity. Using a spatiotemporal immune compartmentalization dataset of the fetal kidney, comprising 24 cell types and 4 spatiotemporal immune compartments [35], we established a baseline tissue network (Fig. 4). The construction process begins with dimensionality reduction and visualization of the original data using methods like UMAP (Fig. 4A). Subsequently, the gene expression information is transformed into gene vectors using natural language processing (Fig. 4B), which are then converted into cell vectors (Fig. 4C). To assess the spatial relationships between cells and inform the network structure, we analyzed intercellular distances (Fig. 4F). Tissue vectors were then calculated by summing the cell vectors within each cell category. Connecting these cell category nodes based on gene similarity generated the tissue network (Fig. 4D). By color-coding cell types according to their spatiotemporal immune compartments in the kidney, we visualized the

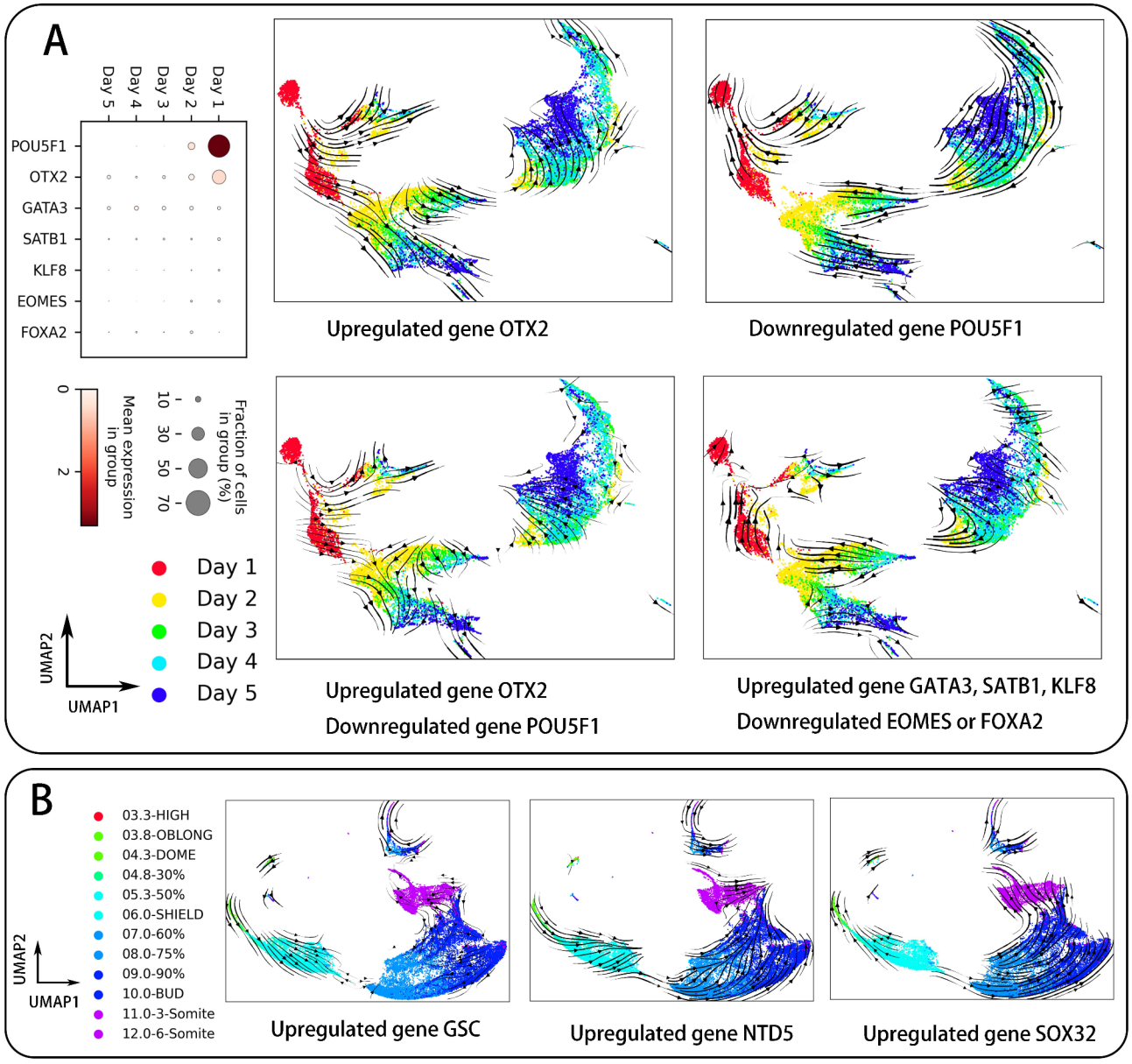

Fig. 3 In silico gene perturbation analysis. (**A**) Human Embryonic Stem Cells (EBs): Individual and combined perturbations of OTX2 and POU5F1, and multi-gene co-perturbations. The left matrix plot shows the expression changes of key genes at different time points. (**B**) Zebrafish Embryos (ZE): Up-regulation of GSC, NTD5, and SOX32

coexistence of different cell types within these compartments. The resulting color patterns showed a trend of agreement with the immune compartments defined by our network structure, suggesting the potential of our tissue vectors for spatiotemporal immune compartment classification.

To explore patient-specific network structures in human lung cancer, we further applied this approach to a human lung cancer dataset (Fig. 5). As shown in the figure, we first performed UMAP dimensionality reduction on the single-cell data (Fig. 5A), and then transformed gene expression information into gene vectors using NLP (Fig. 5B), further deriving cell vectors (Fig. 5C). Based on the similarity between cell vectors, we constructed a patient network (Fig. 5D), where each node represents a patient. To gain a deeper understanding of the cellular organization within individual patients, we took patient TH103 as an example and constructed a cell network based on their cell vectors (Fig. 5E). By comparing cell distance distributions across different patients (Fig. 5F), we can analyze differences in cellular organization between patients. This network-based approach can connect patients with shared disease characteristics, even in cases of significant variations in disease progression. By

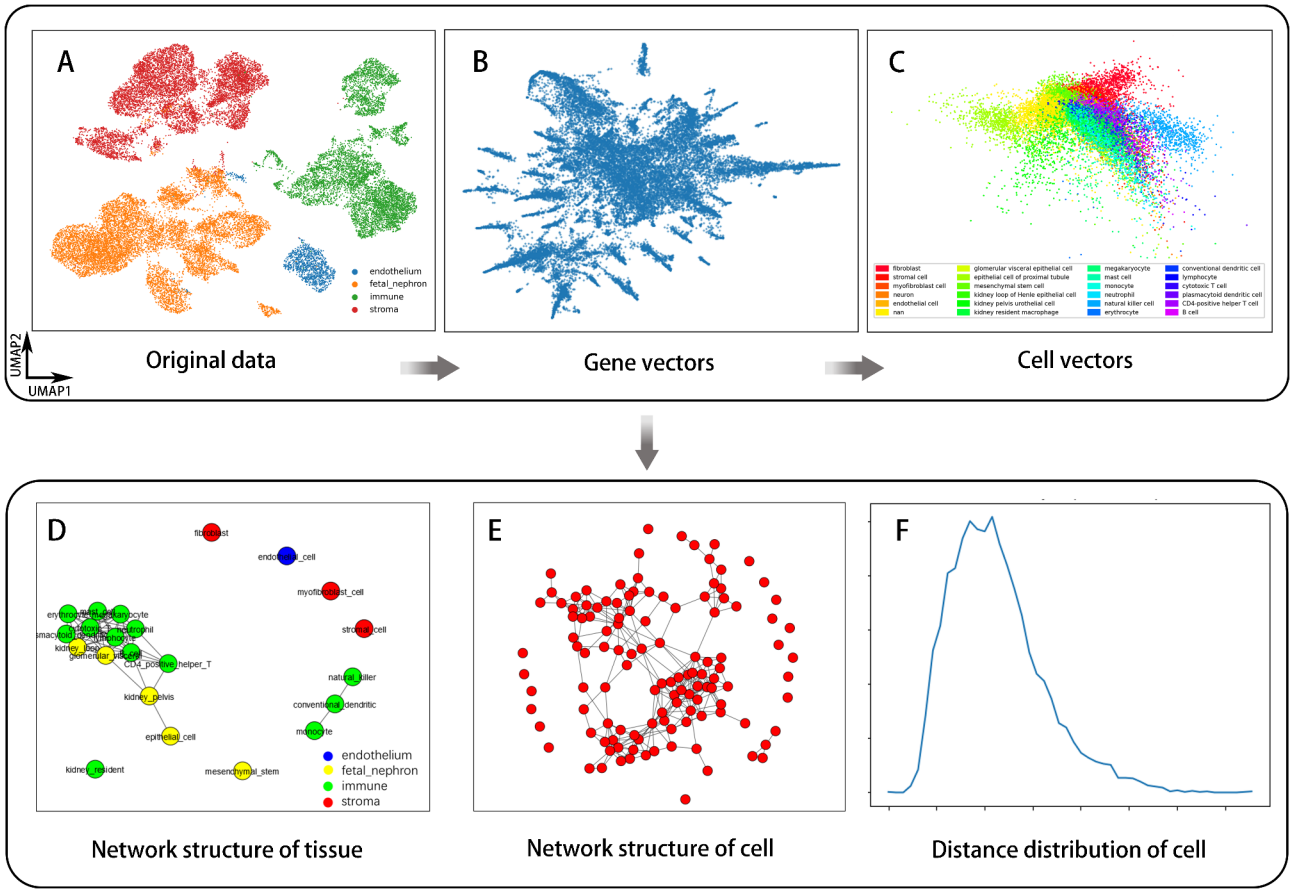

Fig. 4 Construction of a fetal kidney tissue network from scRNA-seq Data. (**A**) Original data (UMAP). (**B**) Gene vectors (NLP embedding). (**C**) Cell vectors. (**D**) Tissue network based on cell categories. (**E**) Henle's loop epithelial cell network. (F) Inter-cellular distance distribution of Henle's loop epithelial cells

analyzing the network structure, we can infer differences in genetic similarity between patients with the same lung cancer.

Although detailed patient-specific descriptions in lung cancer literature are limited, some information can be gleaned from case studies [36]. For example, patient TH226, with samples taken from the same primary tumor site at three different treatment time points, exhibited a standard EGFR exon 19 deletion oncogenic mutation. Compared to other patients in the scRNA-seq dataset, TH226 showed significantly higher expression of many genes related to squamous cell differentiation. Additionally, patient TH266 showed a decreased proportion of macrophages in two tumor biopsies. A network model based on genetic similarity can reveal relationships and categories between different patients and identify potential genetic mutations (Fig. 5D).

Discussion

We have developed an effective model for analyzing scRNA-seq data. Our approach leverages natural language processing (NLP) to represent genes as embedded word vectors in a gene space. These gene vectors are then aggregated to create cell vectors, representing individual cells in a lower-dimensional embedding space. This framework enables visualization and analysis of scRNA-seq datasets at both the cellular and tissue levels.

At the cellular level, cell vectors facilitate the inference of pseudo-time series and the simulation of genetic perturbations. Our model demonstrated promising results in analyzing human embryonic stem cell (EBs) data, effectively illustrating the processes of embryonic cell development and differentiation, thus providing a valuable foundation for further investigations into growth and developmental mechanisms. By capturing the continuous and branching structures of these trajectories, we identify key regulatory genes and pathways that govern cell fate decisions. For instance, in silico perturbations of POU5F1 and OTX2 in EBs revealed critical insights into how these genes influence the anterior neuroectoderm state, contributing to the understanding of the molecular mechanisms underlying early developmental stages and cell fate determination.

At the tissue level, we construct tissue vectors by aggregating cell vectors within defined cell categories. These tissue vectors can then be used to establish tissue

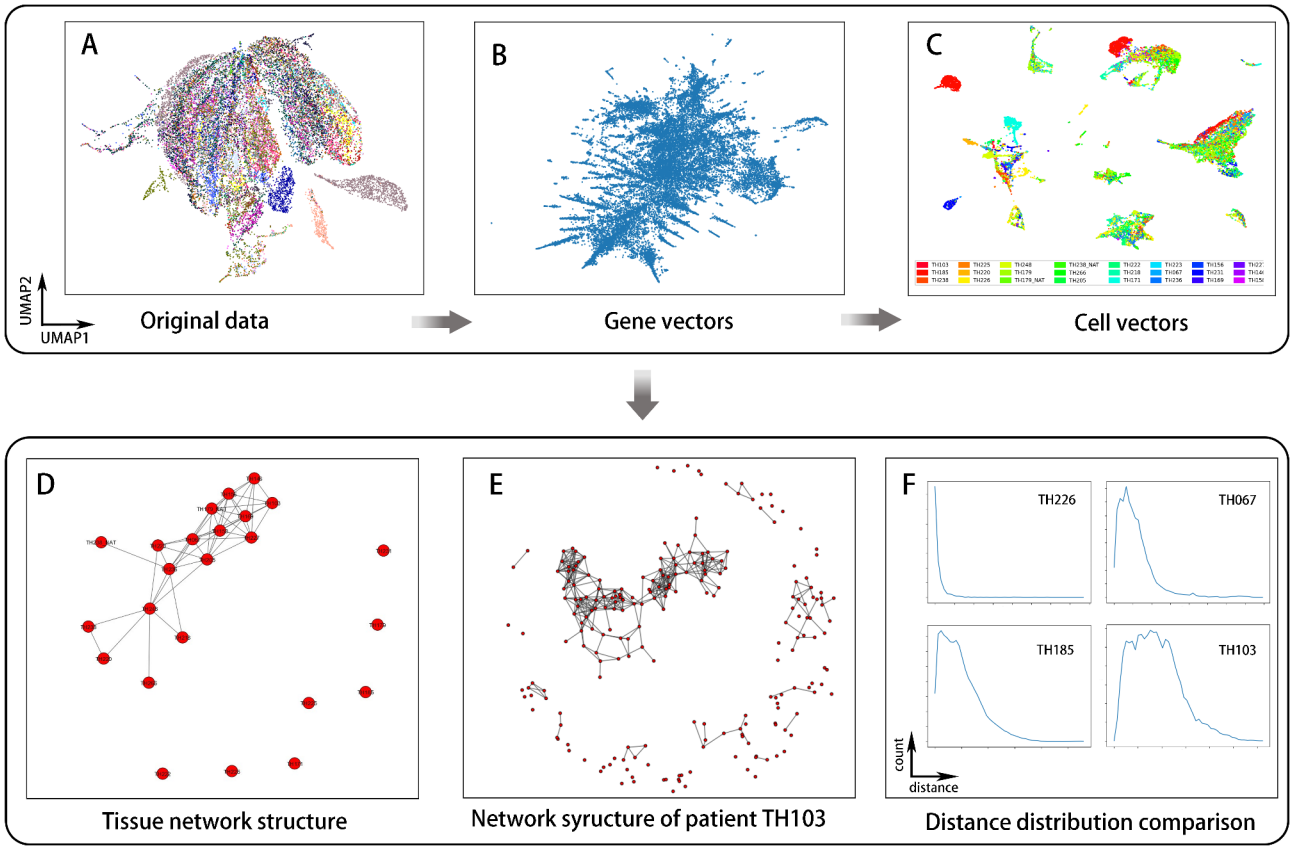

Fig. 5 Patient-Specific Network Analysis in Human Lung Cancer. (**A**) UMAP of single-cell data. (**B**) Gene vectors (NLP embedding). (**C**) Cell vectors. (**D**) Patient network based on gene similarity. (**E**) Cell network for patient TH103. (**F**) Comparison of cell distance distributions within individual patients

networks, providing insights into tissue organization and inter-cellular relationships within tissues. Our analysis of the spatiotemporal immune compartmentalization dataset of the fetal kidney demonstrated the effectiveness of this approach in capturing known tissue structures. This method can also be applied to the study of other tissue types, facilitating a deeper understanding of the interactions between different cell populations and contributing to analysis-based disease classification research.

By constructing tissue-level networks based on cell type composition and gene similarity, we have demonstrated an effective approach to analyze both normal tissue development (fetal kidney) and disease progression (lung cancer). This method allows for the integration of diverse data types and provides insights into the complex relationships between cells, tissues, and patient-specific characteristics. For instance, a multi-layered network analysis from lung cancer patient TH226 may help identify potential therapeutic targets for EGFR-mutant patients. Additionally, examining intercellular distances allows us to infer interactions within the tumor microenvironment, enhancing our understanding of cancer heterogeneity and progression. These insights help identify key genes and pathways involved in both normal development and cancer progression, offering valuable perspectives for early prediction and treatment discovery.

In this study, we chose word2vec as the gene embedding method to balance computational efficiency and model complexity. The stability and effectiveness of word2vec have been validated in gene network analysis [52]. As a lightweight method, word2vec efficiently processes large-scale single-cell data, even in resource-limited settings. For example, processing the EBs dataset took approximately 38 min on a laptop with an AMD Ryzen 5 4600 CPU, a GTX 1650 GPU, and 32GB RAM. Our results demonstrate that word2vec effectively captures gene co-occurrence patterns and semantic similarities by mapping genes into an embedding space.

While more advanced contextual embedding methods based on Transformer architectures, such as scBERT [31], Geneformer [32], and scGPT [33], have emerged in recent years, they typically require higher computational resources and more complex model structures. Given the data scale and the goal of developing a lightweight model, word2vec is sufficiently effective for this study. Future research could incorporate attention mechanisms to address word2vec's limitations in capturing long-range

dependencies, thereby enhancing the model's ability to represent more complex biological relationships.

Abbreviations

scRNA-seq	Single-cell RNA sequencing technology
UMAP	Uniform Manifold Approximation and Projection
PCA	Principal Component Analysis
t-SNE	t-Distributed Stochastic Neighbor Embedding
EBs	Embryoid bodies
ZE	Zebrafish Embryos

Acknowledgements
Not applicable.

Authors' contributions
S.W.W. and Y.L.: Conceptualization, Methodology, Data curation, Software, Validation, Formal analysis, Writing - Original Draft. P.W. and Q.C.L.: Data curation, Software, Validation. J.W.S.: Conceptualization, Resources, Funding acquisition. Q.Z.: Conceptualization, Project administration, Writing - Review & Editing. H.L.: Conceptualization, Methodology, Supervision, Writing - Review & Editing. Y.M.P.: Conceptualization, Methodology, Funding acquisition, Project administration, Supervision, Writing - Review & Editing.

Funding
This work is supported by National Natural Science Foundation of China (Grant Nos. 12090052 and U24A2014), National Key Research and Development Program of China (Grant No. 2020YFC2005605), Xinjiang Tianshan Talents Project (Grant No. TSYC202301A076), and Xinjiang Regional Collaborative Innovation Project (Grant No. 2021E02080), Natural Science Foundation of Liaoning Province (Grant No. 2023-MS-288), Fundamental Research Funds for the Liaoning Universities (Grant No. LJ212410146026).

Data availability
The codes are available online at https://github.com/WSW997/random-walk. The human embryonic stem cell (hESC)-derived embryoid bodies (EBs) dataset [25] is available via the Mendeley Data repository (https://doi.org/10.17632/v6n743h5ng.1). The Zebrafish Embryos (ZE) dataset [34] can be accessed through NCBI GEO under Accession Number GSE106587. The Human Lung Cancer dataset [36] is hosted on NCBI BioProject (ID: PRJNA591860). The fetal kidney temporal-spatial immune partition dataset (Stewart_Fetal.h5ad) [35] is available for download from the Cell Blast repository at https://cblast.gao-lab.org/Stewart_Fetal/Stewart_Fetal.h5ad.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Competing interests
The authors declare that they have no competing interests.

Received: 28 December 2024 / Accepted: 14 February 2025
Published online: 04 March 2025

References

1. Saliba A-E, Westermann AJ, Gorski SA, Vogel J. Single-cell RNA-seq: advances and future challenges. Nucleic Acids Res. 2014;42:8845–60.
2. Luecken MD, Theis FJ. Current best practices in single-cell RNA-seq analysis: a tutorial. Mol Syst Biol. 2019;15:e8746.
3. Gulati GS, Sikandar SS, Wesche DJ, Manjunath A, Bharadwaj A, Berger MJ, et al. Single-cell transcriptional diversity is a hallmark of developmental potential. Science. 2020;367:405–11.
4. Hu H, Feng Z, Shuai XS, Lyu J, Li X, Lin H, et al. Identifying SARS-CoV-2 infected cells with ScVDN. Front Microbiol. 2023;14:1236653.
5. Travaglini KJ, Nabhan AN, Penland L, Sinha R, Gillich A, Sit RV, et al. A molecular cell atlas of the human lung from single-cell RNA sequencing. Nature. 2020;587:619–25.
6. Schelker M, Feau S, Du J, Ranu N, Klipp E, MacBeath G, et al. Estimation of immune cell content in tumour tissue using single-cell RNA-seq data. Nat Commun. 2017;8:2032.
7. Hu H, Feng Z, Lin H, Cheng J, Lyu J, Zhang Y, et al. Gene function and cell surface protein association analysis based on single-cell multiomics data. Comput Biol Med. 2023;157:106733.
8. Qu R, Cheng X, Sefik E, Stanley JS III, Landa B, Strino F et al. Gene trajectory inference for single-cell data by optimal transport metrics. Nat Biotechnol. 2024;1–11.
9. Sha Y, Qiu Y, Zhou P, Nie Q. Reconstructing growth and dynamic trajectories from single-cell transcriptomics data. Nat Mach Intell. 2024;6:25–39.
10. Chung W, Eum HH, Lee H-O, Lee K-M, Lee H-B, Kim K-T, et al. Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. Nat Commun. 2017;8:15081.
11. Roohani Y, Huang K, Leskovec J. Predicting transcriptional outcomes of novel multigene perturbations with GEARS. Nat Biotechnol. 2024;42:927–35.
12. Papalexi E, Satija R. Single-cell RNA sequencing to explore immune cell heterogeneity. Nat Rev Immunol. 2018;18:35–45.
13. Lin H, Hu H, Feng Z, Xu F, Lyu J, Li X, et al. SCTC: inference of developmental potential from single-cell transcriptional complexity. Nucleic Acids Res. 2024;52:6114–28.
14. Wang W, Zhang L, Sun J, Zhao Q, Shuai J. Predicting the potential human lncRNA–miRNA interactions based on graph Convolution network with conditional random field. Brief Bioinform. 2022;23:bbac463.
15. Huang D, Ma N, Li X, Gou Y, Duan Y, Liu B, et al. Advances in single-cell RNA sequencing and its applications in cancer research. J Hematol Oncol. 2023;16:98.
16. Bergen V, Lange M, Peidli S, Wolf FA, Theis FJ. Generalizing RNA velocity to transient cell States through dynamical modeling. Nat Biotechnol. 2020;38:1408–14.
17. Hu H, Feng Z, Lin H, Zhao J, Zhang Y, Xu F, et al. Modeling and analyzing single-cell multimodal data with deep parametric inference. Brief Bioinform. 2023;24:bbad005.
18. Yang X, Sun J, Jin B, Lu Y, Cheng J, Jiang J, et al. Multi-task aquatic toxicity prediction model based on multi-level features fusion. J Adv Res. 2025;68:477–89.
19. Chowdhary KR. Natural Language processing. Fundamentals of artificial intelligence. New Delhi: Springer India; 2020. pp. 603–49.
20. Nadkarni PM, Ohno-Machado L, Chapman WW. Natural Language processing: an introduction. J Am Med Inform Assoc. 2011;18:544–51.
21. Mikolov T. Efficient estimation of word representations in vector space. arXiv preprint arXiv:13013781. 2013;3781.
22. Saelens W, Cannoodt R, Todorov H, Saeys Y. A comparison of single-cell trajectory inference methods. Nat Biotechnol. 2019;37:547–54.
23. Haghverdi L, Büttner M, Wolf FA, Buettner F, Theis FJ. Diffusion pseudotime robustly reconstructs lineage branching. Nat Methods. 2016;13:845–8.
24. Yin S, Xu P, Jiang Y, Yang X, Lin Y, Zheng M, et al. Predicting the potential associations between circrna and drug sensitivity using a multisource feature-based approach. J Cell Mol Med. 2024;28:e18591.
25. Moon KR, Van Dijk D, Wang Z, Gigante S, Burkhardt DB, Chen WS, et al. Visualizing structure and transitions in high-dimensional biological data. Nat Biotechnol. 2019;37:1482–92.
26. Chung NC, Storey JD. Statistical significance of variables driving systematic variation in high-dimensional data. Bioinformatics. 2014;31:545–54.
27. Van der Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res. 2008;9:2579–605.
28. Becht E, McInnes L, Healy J, Dutertre C-A, Kwok IW, Ng LG, et al. Dimensionality reduction for visualizing single-cell data using UMAP. Nat Biotechnol. 2019;37:38–44.
29. Marx V. Seeing data as t-SNE and UMAP do. Nat Methods. 2024;21:930–3.
30. Tenenbaum JB, de Silva V, Langford JC. A global geometric framework for nonlinear dimensionality reduction. Science. 2000;290:2319–23.
31. Yang F, Wang W, Wang F, Fang Y, Tang D, Huang J, et al. ScBERT as a large-scale pretrained deep Language model for cell type annotation of single-cell RNA-seq data. Nat Mach Intell. 2022;4:852–66.
32. Theodoris CV, Xiao L, Chopra A, Chaffin MD, Al Sayed ZR, Hill MC, et al. Transfer learning enables predictions in network biology. Nature. 2023;618:616–24.

33. Cui H, Wang C, Maan H, Pang K, Luo F, Duan N, et al. ScGPT: toward Building a foundation model for single-cell multi-omics using generative AI. Nat Methods. 2024;21:1470–80.
34. Farrell JA, Wang Y, Riesenfeld SJ, Shekhar K, Regev A, Schier AF. Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. Science. 2018;360:eaar3131.
35. Stewart BJ, Ferdinand JR, Young MD, Mitchell TJ, Loudon KW, Riding AM, et al. Spatiotemporal immune zonation of the human kidney. Science. 2019;365:1461–6.
36. Maynard A, McCoach CE, Rotow JK, Harris L, Haderk F, Kerr DL, et al. Therapy-induced evolution of human lung cancer revealed by single-cell RNA sequencing. Cell. 2020;182:1232–51.
37. Wu F, Yang R, Zhang C, Zhang L. A deep learning framework combined with word embedding to identify DNA replication origins. Sci Rep. 2021;11:844.
38. Rosen Y, Brbić M, Roohani Y, Swanson K, Li Z, Leskovec J. Toward universal cell embeddings: integrating single-cell RNA-seq datasets across species with SATURN. Nat Methods. 2024;21:1492–500.
39. Serra A, Coretto P, Fratello M, Tagliaferri R. Robust and sparse correlation matrix Estimation for the analysis of high-dimensional genomics data. Bioinformatics. 2018;34:625–34.
40. Řehůřek R, Sojka P. Software Framework for Topic Modelling with Large Corpora. Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. Valletta, Malta: University of Malta; 2010:45–50. Available from: http://is.muni.cz/publication/884893/en
41. Lotfollahi M, Wolf FA, Theis FJ. ScGen predicts single-cell perturbation responses. Nat Methods. 2019;16:715–21.
42. Liu L, Wei Y, Zhang Q, Zhao Q. SSCRB: predicting circRNA-RBP interaction sites using a sequence and structural feature-based attention model. IEEE J Biomedical Health Inf. 2024;28:1762–72.
43. Schubert M, Klinger B, Klünemann M, Sieber A, Uhlitz F, Sauer S, et al. Perturbation-response genes reveal signaling footprints in cancer gene expression. Nat Commun. 2018;9:20.
44. Zhu F, Niu Q, Li X, Zhao Q, Su H, Shuai J. FM-FCN: a neural network with filtering modules for accurate vital signs extraction. Research. 2024;7:0361.
45. Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat Biotechnol. 2014;32:381–6.
46. La Manno G, Soldatov R, Zeisel A, Braun E, Hochgerner H, Petukhov V, et al. RNA velocity of single cells. Nature. 2018;560:494–8.
47. Barabasi A-L, Oltvai ZN. Network biology: Understanding the cell's functional organization. Nat Rev Genet. 2004;5:101–13.
48. Zhang L, Yang P, Feng H, Zhao Q, Liu H. Using network distance analysis to predict lncRNA–miRNA interactions. Interdisciplinary Sciences: Comput Life Sci. 2021;13:535–45.
49. Meng R, Yin S, Sun J, Hu H, Zhao Q. ScAAGA: single cell data analysis framework using asymmetric autoencoder with gene attention. Comput Biol Med. 2023;165:107414.
50. Xu F, Hu H, Lin H, Lu J, Cheng F, Zhang J, et al. ScGIR: Deciphering cellular heterogeneity via gene ranking in single-cell weighted gene correlation networks. Brief Bioinform. 2024;25:bbae091.
51. Kikuchi Y, Agathon A, Alexander J, Thisse C, Waldron S, Yelon D, et al. Casanova encodes a novel Sox-related protein necessary and sufficient for early endoderm formation in zebrafish. Genes Dev. 2001;15:1493–505.
52. Du J, Jia P, Dai Y, Tao C, Zhao Z, Zhi D. Gene2vec: distributed representation of genes based on co-expression. BMC Genomics. 2019;20:82.

Publisher's note