Genome Biology

METHOD

Open Access

scPriorGraph: constructing biosemantic cell–cell graphs with prior gene set selection for cell type identification from scRNA-seq data

Xiyue Cao[1†], Yu-An Huang[1*†], Zhu-Hong You[1*], Xuequn Shang[1], Lun Hu[2], Peng-Wei Hu[2] and Zhi-An Huang[3]

†Xiyue Cao and Yu-An Huang contributed equally to this work.

*Correspondence: yuanhuang@nwpu.edu.cn; zhuhongyou@nwpu.edu.cn

[1] School of Computer Science, Northwestern Polytechnical University, Xi'an, China
[2] Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Sciences, Ürümqi, China
[3] Research Office, City University of Hong Kong (Dongguan), Dongguan 523000, China

Abstract

Cell type identification is an indispensable analytical step in single-cell data analyses. To address the high noise stemming from gene expression data, existing computational methods often overlook the biologically meaningful relationships between genes, opting to reduce all genes to a unified data space. We assume that such relationships can aid in characterizing cell type features and improving cell type recognition accuracy. To this end, we introduce scPriorGraph, a dual-channel graph neural network that integrates multi-level gene biosemantics. Experimental results demonstrate that scPriorGraph effectively aggregates feature values of similar cells using high-quality graphs, achieving state-of-the-art performance in cell type identification.

Keywords: Single-cell RNA sequencing, Cell-type identification, Dual-channel graph neural network, Pathway, Ligand-receptor network, Graph augmentation

Background

Single-cell RNA sequencing (scRNA-seq) is a high-throughput and highly sensitive RNA sequencing method that has revolutionized biological research by increasing the resolution to the individual cell level, thereby significantly enhancing our understanding of cell heterogeneity and the molecular mechanisms that regulate cell behavior [1, 2]. This technology has led to a rapid growth in scRNA-seq data, necessitating the analysis of these datasets to extract various types of information. Single-cell data analysis is a multi-step process encompassing preprocessing and downstream analysis at both the cell and gene levels, with a crucial focus on identifying cell subpopulations, which plays a pivotal role in subsequent analyses. In cancer research, pinpointing diverse cancer cell subpopulations helps us understand tumor development and treatment responses [3], while in immunology research, recognizing various immune cell subtypes enhances our understanding of their distinct roles in immune responses [4]. Due to the high noise and dimensionality of single-cell data, identifying cell types remains a significant challenge. Researchers require a reliable method that leverages established cell labels as a reference

to learn from existing data and assign cell type identities to newly generated datasets rapidly and accurately, prompting the development of various cell type identification and annotation methods.

Currently, single-cell type identification methods fall into three primary categories: those based on annotated gene databases, correlation-based approaches, and supervised machine learning techniques. Approaches relying on annotated gene databases utilize publicly available databases and cell-specific marker genes highly expressed in known cell types to label cell types in test datasets. Such methods, like MarkerCount [5], scCATCH [6], and scTyper [7], have limitations, such as scarce marker genes for rare cell types and potential accuracy issues associated with manually selecting marker gene sets. Correlation-based methods rely on feature selection to identify and eliminate irrelevant or redundant features from expression data, and errors in this process can affect cell classification accuracy. Notable methods in this category include scmap-cell [8], CHETAH [9], and SingleR [10].

Supervised classification methods based on machine learning utilize reference datasets to train classifier models for labeling unannotated datasets' cell types. Typically, these methods build models for cell type distributions using features trained on previously annotated datasets and subsequently employ these models to assign labels to samples in unannotated datasets. Prominent methods in this category include SingleCellNet [11], scPred [12], TOSICA [13], scLearn [14], and Moana [15]. Supervised classification methods offer adaptability, scalability, and transferability but face challenges due to the inherent noise and sparsity in scRNA-seq data [16] and batch effects [17] arising from different operators, experimental protocols, and technical variations that can impact model performance.

While supervised classification methods have value in various scenarios, their capabilities and performance are limited, partly because they extract feature information only from independent cells, disregarding higher-order relationships between cells. Graph convolutional neural networks (GCNs), as graph-based deep learning methods, can effectively capture the topological relationships between cells. The application of GCNs in single-cell data analysis, introducing methods like scGCN [18], GCN-SC [19], and scCDG [20], showcases their capability to learn higher-order cell representations and topological relationships for improved cell feature extraction and model performance. Nevertheless, the construction of high-quality cell graphs heavily depends on gene expression data and lacks support from cell-specific knowledge.

Existing graph-based computational methods, as mentioned above, primarily rely on gene expression as cell features, neglecting biologically significant intercellular communication and intracellular gene relationships. Cellular communication, primarily achieved through ligand-receptor interactions, is fundamental to coordinating organism development, maintaining homeostasis, and sustaining tissue and organ functionality [21]. The primary mode of intercellular information transmission involves ligand-receptor interactions. The binding of ligands to receptors leads to changes in the receptor's conformation or activity, triggering a series of intracellular reactions and signaling pathways. This process gradually amplifies signals, resulting in a comprehensive array of cellular responses. Moreover, pathways, documented in databases such as Humancyc [22], INOH [23], and KEGG [24], encompass gene relationships governing common

biological processes, including intracellular reactions, metabolism, and signal transduction. Literatures underscores the pivotal role of intercellular communication, intracellular reactions, metabolism, and signal transduction pathways in shaping cell functionality. For instance, intercellular communication plays a pivotal role in determining hematopoietic stem cell development, with the communication processes substantially influencing the formation of blood cell types [25].

The expression levels of genes within cells can be used to describe various biological characteristics, such as cell subtypes and whether a cell has undergone malignancy. We refer to this as the "semantic information" of genes. Genes possess multiple semantics, with our primary focus directed towards the semantic information pertaining to cell–cell communication and intracellular biological processes such as genetics and metabolism. By integrating cell–cell communication relationships and pathways with gene expression data, we acquire multi-level biological semantic information about genes. We enhance the interpretative capability of graph convolutional neural networks for scRNA-seq data through the hierarchical organization of gene biological semantics. Considering the high noise in single-cell data and the relatively minor impact of noise on individual genes within a pathway, the integration of pathway-based scRNA-seq data makes the model less susceptible to noise.

Considering the benefits of integrating pathways and diverse biological information into the model, we introduce scPriorGraph, a dual-channel graph convolutional neural network. scPriorGraph incorporates various biological prior knowledge into the construction of cell graphs and enables users to choose gene biological semantics that are suitable for their specific tasks. The model combines gene-level expression data with the chosen gene sets, generating gene set-level expression data, which is used for graph convolution construction, thereby integrating intercellular communication and intracellular reaction information. The sparsity of scRNA-seq data can affect the graph convolutional neural network's aggregation of feature values for adjacent cells. For instance, a cell with a high dropout rate may be forced to associate with other cells with low gene expression in constructed graphs, rather than those that are truly biologically similar. In such cases, the graph convolutional neural network may include lower-quality expression information, affecting the model's prediction capability. To address this challenge, we introduce a graph-augmentation technique based on global cell similarity into the model to enhance feature aggregation, thereby obtaining higher-quality graph embeddings.

Results

scPriorGraph model

scPriorGraph is an automatic cell-type annotation method that fuses various biological prior knowledge using a dual graph convolutional network (see Fig. 1). The model utilizes cell *k*-nearest neighbor graphs to construct graph convolution layers, incorporating the hierarchical information of gene biological semantics during the construction process. One is intercellular communication information, which is obtained by generating a ligand-receptor network from a set of ligand-receptor gene pairs. Using Metapath-based random walks on this network, we acquire ligand-receptor pathway information and generate gene sets based on this pathway information. The other type of prior knowledge involves user-customizable intracellular gene pathways, which typically encompass

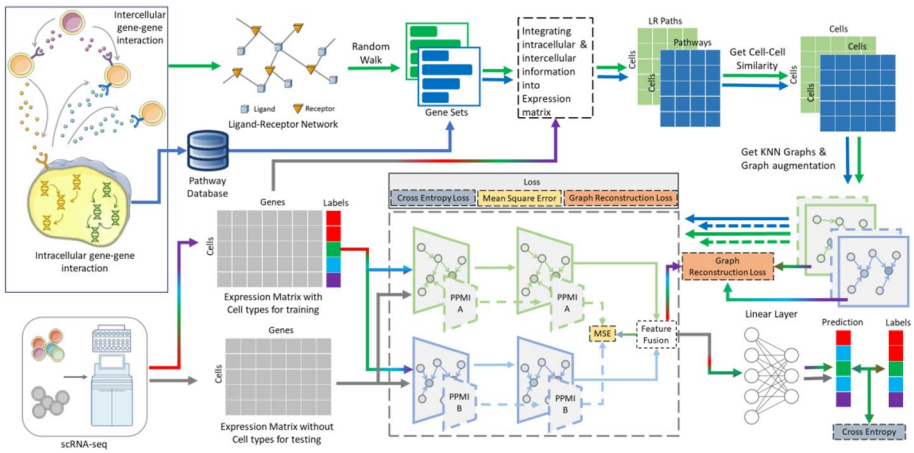

Fig. 1 Overview of scPriorGraph. scPriorGraph is a dual-channel graph neural network that integrates multi-level gene biological semantic information. Initially scPriorGraph extracts intercellular communication information from ligand-receptor network using random walks. Subsequently, it obtains intracellular gene interaction information from a pathway database. These two sets of information are separately integrated with scRNA-seq data, resulting in multi-level gene biological semantics, and two cell *k*-nearest neighbor (KNN) graphs are constructed based on different semantic information. To augment the graphs, scPriorGraph utilizes the Positive Pointwise Mutual Information (PPMI) matrix to capture the global similarity of cells. Integrating various biological prior knowledge, scPriorGraph can provide accurate cell type annotations for unknown cells

information about various genes cooperating to accomplish specific biological processes. Users can choose pathway based on their specific needs. These two types of information are integrated into the generation of cell-*k*-nearest neighbor graphs, forming the foundation of the entire scPriorGraph model.

Due to the presence of challenges like high noise and sparsity in single-cell sequencing data, the generated cell graphs can be influenced by low-quality data. To enhance the cell graphs, we introduced a positive pointwise mutual information (PPMI) matrix [26] in the model to embed global consistency information. After generating two cell *k*-nearest neighbor graphs, we performed random walks on the nodes of each graph, sampled nodes along the paths, and obtained a frequency matrix, F. Based on matrix F, we computed the PPMI matrix. Pointwise mutual information is commonly used to measure the correlation between two entities. In our model, if cells A and B appear together more frequently along a path, we consider them to be more correlated. We generated a PPMI matrix for each cell graph and created corresponding graph convolution layers. During model training, we included the minimization of mean squared errors between the graph convolution layer outputs of the cell graph and the PPMI matrix as part of the loss function. This approach allows our model to consider both local and global similarity information for nodes, introducing additional, previously undiscovered potential relationships among different cells, which cannot be adequately represented by cell graphs alone.

scPriorGraph employs a supervised learning approach, leveraging known cell type datasets for training. It utilizes graph convolutional layers to aggregate neighboring cell expression data, mapping high-dimensional cell expression data to a lower-dimensional space. Following training completion, the model is applied to the expression information

of a new dataset for cell type prediction, facilitating the transfer of cell labels from known datasets to new datasets.

Enhancing cross-batch prediction and mitigating batch effects with scPriorGraph

In single-cell sequencing, batch effects refer to variations in sample outcomes arising from different experimental conditions, operators, reagents from different companies, different batches of reagents, distinct sequencing runs, and other factors. In this context, it specifically denotes the differences between training and testing samples. In single-cell sequencing data analysis, when the training dataset and the testing dataset originate from different sequencing platforms, species, or developmental time points, biases, and distinctions between the training and testing data may emerge, posing challenges to data analysis and interpretation. This situation also demands higher model adaptability. Therefore, we designed three categories of experiments, encompassing cross-platform, cross-species, and cross-development-time scenarios, to evaluate the model's adaptability, generalization capabilities, and its ability to ameliorate batch effects when confronted with data from diverse sources.

In the cross-platform experiments, we selected the human pancreatic dataset and the peripheral blood mononuclear cell (PBMC) dataset [27]. The human pancreatic dataset comprises sequencing data from the Baron Human [28], Muraro [29], Segerstolpe [30], and Xin [31] platforms. Based on this human pancreatic dataset, we designed 12 paired reference-query dataset experimental schemes. These experiments were divided into four groups according to the specific reference dataset used. The PBMC dataset includes sequencing data for PBMCs from the 10Xv2, 10Xv3, CEL-Seq, Drop-Seq, inDrop, Seq-Well, and Smart-Seq2 platforms. Based on the PBMC dataset, we designed 30 paired reference-query dataset experimental schemes, which were categorized into five groups depending on the particular reference dataset employed.

We utilized different evaluation metrics, including Accuracy Score (Acc), Weighted F1 Score, Cohen's Kappa, and Cross-Category Average Accuracy, to comprehensively compare the performance across models. In the context of the human pancreas dataset, scPriorGraph consistently demonstrated superior or comparable performance across four evaluation metrics when contrasted with the other eight methods in Fig. 2a and Additional file 1: Table. S1-S4. In experiments employing the Segerstolpe dataset as a reference, enabling cell type predictions for the Baron Human, Muraro, and Xin datasets, scPriorGraph, while not topping the list in individual sub-experiments, ultimately achieved the highest group's mean accuracy (Acc = 0.981) with singleCellNet when the results of these three sub-experiments were considered together. This illustrates the heightened versatility of scPriorGraph relative to other methods across varying datasets. The stability and elevated performance of scPriorGraph were further affirmed through visual analysis, as illustrated in Fig. 2a.

To thoroughly assess the performance of scPriorGraph relative to other methods, we executed a series of cross-platform experiments using the PBMC dataset. Considering that the prediction task for the Human Pancreas dataset is relatively simple, resulting in close comparison results for some methods, we introduced more methods, bringing the total to 20 for this comparison in the series of PBMC experiments, which include sciBet [32], scGPT [33], scANVI [34], expiMap [35] with treeArches [36], scPoli [37], simple

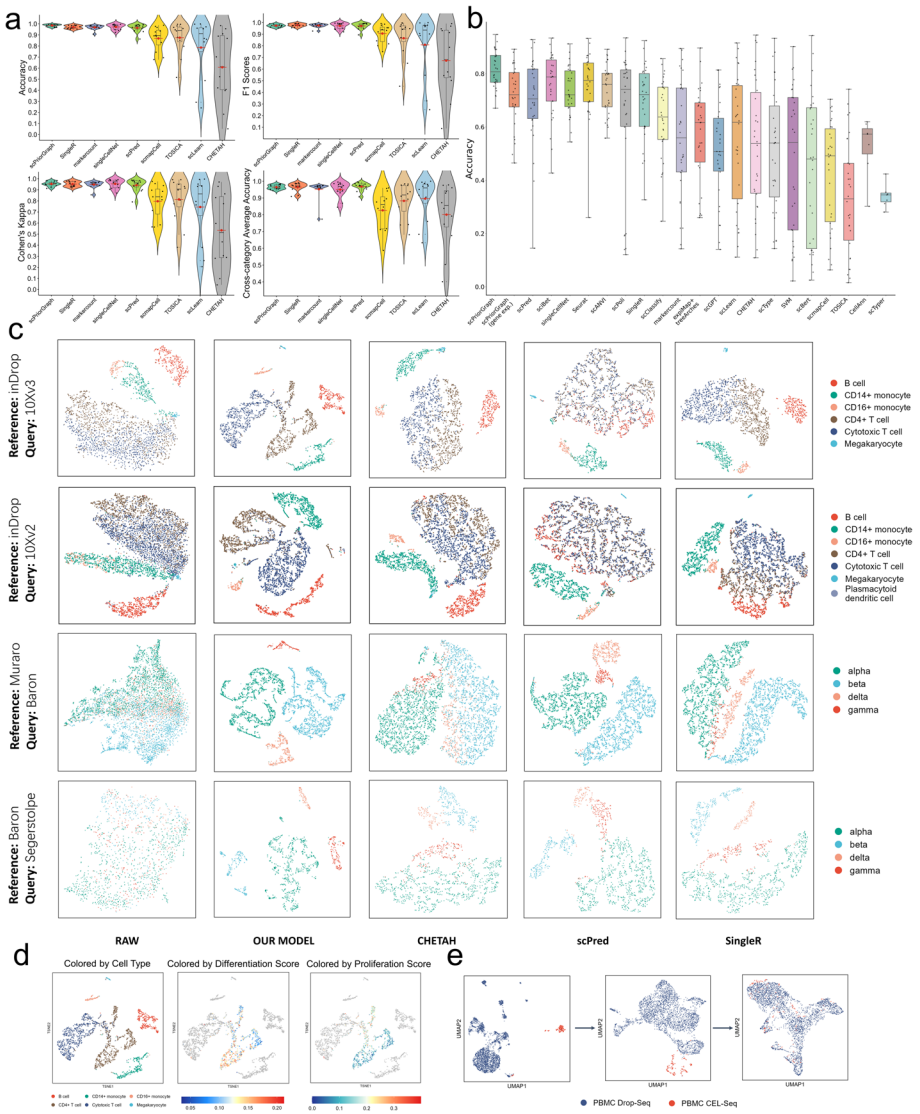

Fig. 2 Comparison of scPriorGraph with other methods in cross-platform experiments. **a** Comparison of the performance of scPriorGraph and other comparative methods on the human pancreatic dataset using Accuracy, F1 Scores, Cohen's Kappa, and Cross-category Average Accuracy. **b** Comparison of the performance of scPriorGraph and other comparative methods on the PBMC dataset using Accuracy. **c** Comparison of our model and three other methods in their classification performance on cross-platform data using t-SNE plots. The t-SNE plots are generated from models' embeddings and are colored based on the true cell types of the query dataset. **d** Analysis of intra-type variability within cells using differentiation and proliferation scores. **e** UMAP projection of the raw data, the first graph convolution layer output from the model, and the second graph convolution layer output from the model when using PBMC Drop-Seq data for training and PBMC CEL-Seq data for prediction

linear SVM, Seurat [38], scBert [39], scClassify [40], scType [41], scTyper [7], CellAnn [42], and the 8 methods previously used in human pancreatic experiments. The evaluation metrics demonstrate in Fig. 2b and the Additional file 1: Table. S5-S8 that scPriorGraph achieved superior prediction performance.

In the five comparative experiments conducted on the PBMC dataset, scPriorGraph achieved top performance in three out of the five categories (10Xv3 platform, Drop-Seq

platform, and inDrop platform), while scPred and Seurat led in the 10Xv2 platform and CEL-Seq platform, respectively. In addition, scPriorGraph demonstrated a higher median accuracy and better stability in the scatter box plot compared to other methods, reflecting the overall superior performance of our approach (see Fig. 2b). It is worth noting that several methods exhibited a substantial drop in predictive performance when faced with certain datasets. For instance, the CHETAH method produced notably inferior results when the reference dataset was derived from the 10xv2, 10xv3, and CEL-Seq platforms. Similarly, the scmapCell method's performance significantly lagged behind other comparative methods when employing the Drop-Seq platform as the reference dataset. In contrast, scPriorGraph showed no substantial performance disparities when compared to other methods. Figure 2b highlights the concentration and consistency of results achieved by scPriorGraph relative to the other methods. We further compared the cross-platform experimental results with those reported in [40] using the same PBMC datasets and found that the performance distribution of our selected comparison methods across different reference-query pairs showed similar patterns.

Figure 2c indicates that, in experiments involving the PBMC dataset, specifically with the inDrop-10xv3 and inDrop-10xv2 reference-query data pairs, scPriorGraph demonstrated excellent cell type identification performance. The training procedure involved reference datasets, and embeddings were obtained by inputting the query datasets into the model. Notably, in both the inDrop-10xv3 and inDrop-10xv2 experiments, scPred failed to effectively differentiate between the categories of CD4 + T cells, cytotoxic T cells, and B cells. SingleR, while capable of distinguishing B cells in the inDrop-10xv3 experiment, still encountered challenges in separating CD4 + T cells from cytotoxic T cells. Similarly, the CHETAH method also struggled to distinguish CD4 + T cells from cytotoxic T cells. In contrast, t-SNE projections generated by scPriorGraph revealed distinct boundaries between all cell types, although a minor overlap of CD4 + T Cells remained within the cytotoxic T cell-enriched region. In the Muraro-Baron and Baron-Segerstolpe experiments, t-SNE plots produced by scPriorGraph effectively discriminated all four cell types. Relative to other comparative methods, scPriorGraph's t-SNE plots exhibited tighter clustering of cells within the same type, with greater separation between cell clusters of different types. This outcome underscores that, despite the disparate origins of the reference and query datasets from distinct sequencing platforms, scPriorGraph successfully captured the distinctive features of various cell types based on their gene expression profiles. Consequently, it achieved precise cell type classification, a finding further validated by accuracy scores. To investigate the association between intra-cell type variability and cell states, we utilized the AUCell method to assess expression levels across individual cells for gene sets pertinent to cell states, as defined by CancerSEA [43]. By visualizing cells based on their AUCell scores in Fig. 2d, we observed significant intra-group variability in the expression of gene sets related to differentiation and proliferation among CD4 + T cells. This suggests that the cell states linked to differentiation and proliferation may play a role in the observed internal variability within this cell type.

The proposed model incorporates two graph convolutional layers: the first layer aggregates features from one-hop neighboring cells to create an embedding, while the second layer extends this aggregation to features from cells at a two-hop distance.

When applied to the PBMC Drop-Seq and PBMC CEL-Seq datasets as a training-test pair, both the original data and the extracted graph embeddings from each graph convolutional layer are projected onto UMAP plots. Figure 2e clearly illustrates that, prior to training, there are well-defined boundaries between the two datasets, indicating significant differences between them. However, as the model aggregates features from both batches, the visualizations in the embeddings obtained from the first graph convolutional layer show a reduction in the distance between the datasets. Furthermore, the embeddings from the second graph convolutional layer visually depict the convergence of the two datasets on the UMAP plot. This observation underscores the effective capacity of scPriorGraph in mitigating batch effects.

The cross-species experiments were divided into two parts: one involved the mutual prediction of human brain cells and mouse brain cells, and the other entailed the mutual prediction of human pancreatic cells and mouse pancreatic cells. The datasets we used were MouseALM (GSE115746 [44]), HumanMTG (phs001790 [45]), and Pancreas (GSE84133 [28]). MouseALM provides a detailed depiction of the diversity of cell types in two regions of the mouse brain: the visual cortex (VISp) and the anterior lateral motor cortex (ALM). HumanMTG focuses on cell types in the human middle temporal gyrus (MTG). We used the MouseALM and HumanMTG datasets to conduct a cross-species experiment on brain tissue cells. For the pancreatic single-cell data, we downloaded the Pancreas dataset to construct separate datasets for human and mouse pancreatic cells. For the cross-species experiments, we designed four paired reference-query datasets, each consisting of one dataset composed of mouse data paired with another composed of human data. The results of these experiments can be found in Additional file 1: Table. S9.

In the cross-temporal experiments, where reference and query datasets were chosen from samples collected at different developmental stages, we utilized data from [46] to construct three distinct datasets. These are designated as CrossTemporal_T1 (GSM3852753, embryonic stage E13.5), CrossTemporal_T2 (GSM3852754, embryonic stage E14.5), and CrossTemporal_T3 (GSM3852755, embryonic stage E15.5). The results can be found in Additional file 1: Table. S10.

Our method demonstrated outstanding performance across a range of experiments. In the cross-species experiment using Pancreas dataset, we achieved the highest average accuracy (Acc = 0.96) among eight methods. For the cross-species experiment involving MouseALM and HumanMTG datasets, our method ranked second in average accuracy (Acc = 0.53). In the cross-temporal experiment, the result shown in Fig. 3a demonstrates that our method again secured the highest average accuracy (Acc = 0.93). In the cross-species experiment, where human pancreatic data were used to predict murine pancreatic data within Pancreas dataset, scPriorGraph demonstrated the highest performance among all methods (Acc = 0.96). A chord diagram was employed to illustrate the alignment between the predicted data categories by our model and the true categories for murine data. The chord diagram in Fig. 3b demonstrated that our model accurately assigned categories for the majority of cells, reaffirming the precision of our predictions. Notably, in the cross-temporal experiment for the CrossTemporal_T2-CrossTemporal_T3 pair, scPred exhibited performance comparable to our model, as shown in Fig. 3d. However, Fig. 3c, depicted as a

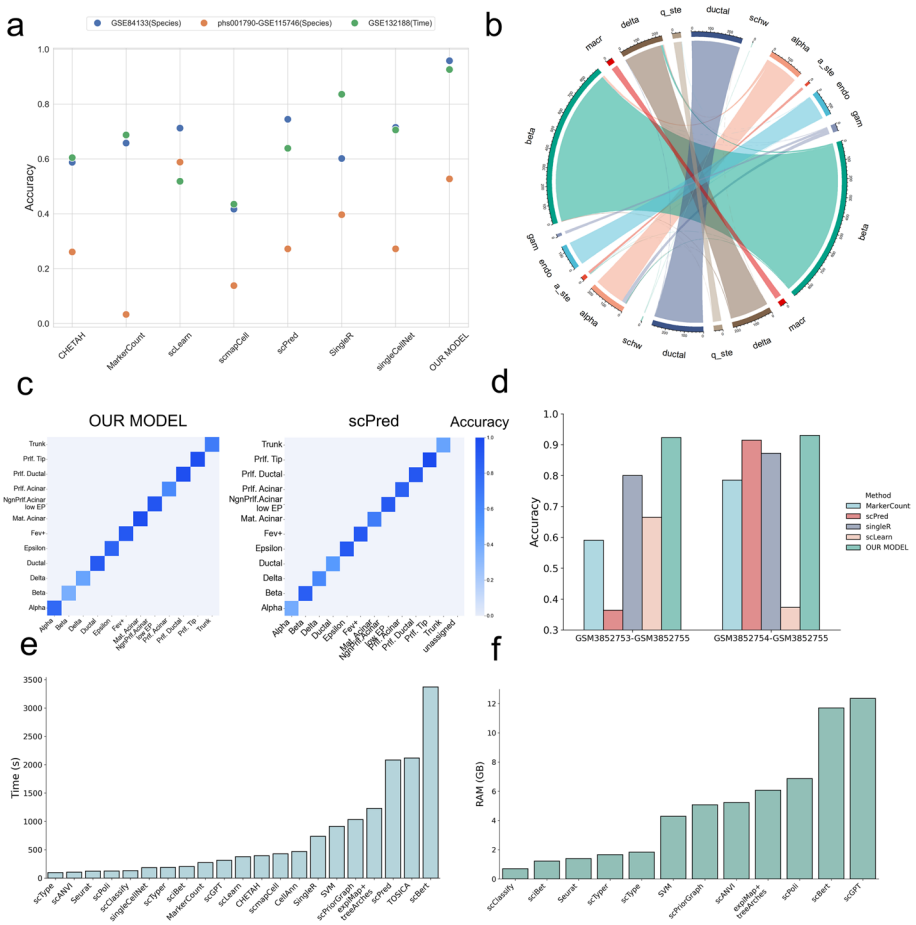

Fig. 3 Comparison of scPriorGraph with other methods in cross-species experiments and cross-temporal experiments. **a** Comparison of the average accuracy of our method with other methods in the experiments GSE84133, phs001790-GSE115746, and GSE132188. **b** Chord diagram illustrating the accuracy of scPriorGraph in predicting different cell types when using human pancreatic data to predict mouse pancreatic data sourced from GSE84133. The proportions of the sectors represent the distribution of cell types in the dataset. The right side of the diagram represents the real data, while the left side represents the predicted data. **c** Comparison of cell type prediction accuracy between scPriorGraph and scPred in the GSM3852754-GSM3852755 experiment using a heatmap. **d** Performance comparison of various methods in the cross-temporal experiments. **e** Comparation of computational time across different methods. **f** Comparison of memory usage across different methods

heatmap, highlighted our method's superior accuracy in cell classification compared to scPred.

To quantify the batch effect correction capability of scPriorGraph, we selected cross-platform datasets of PBMC and human pancreas as benchmarks to assess the ability of scPriorGraph to correct batch effects. The Average Silhouette Width (ASW) was used to evaluate the effectiveness of batch effect correction. Lower batch ASW scores and higher cell type ASW scores indicate better batch correction performance. Specific experimental results are displayed in the tables below. From Table 1, it is evident that scPriorGraph possesses a certain ability to correct batch effects.

To assess the computational efficiency of the proposed method, we conducted performance tests on the PBMC dataset using a consistent hardware setup (NVIDIA

Table 1 Quantitative measurement of batch effect correction

Reference	Query	ASW Score (Cell type)		ASW Score (Batch)	
		RAW	scPriorGraph	RAW	scPriorGraph
10Xv3	10Xv2	0.182	0.223	0.060	−0.011
10Xv3	CEL-Seq	0.230	0.437	0.109	0.165
10Xv3	Drop-Seq	0.112	0.184	0.357	0.171
10Xv3	inDrop	0.090	0.242	0.380	0.140
10Xv3	Seq-Well	0.026	0.091	0.325	0.078
10Xv3	Smart-Seq2	0.390	0.626	0.150	−0.001
Baron Human	Muraro	0.230	0.794	0.444	0.027
Baron Human	Segerstolpe	0.399	0.752	0.326	0.046
Baron Human	Xin	0.294	0.804	0.343	0.016

GeForce RTX 4090 GPU, Intel Core i9-13900KF CPU, and 96 GB of RAM). We compared scPriorGraph with other methods in terms of speed and memory usage, as depicted in Fig. 3e and f. Although scPriorGraph ranked in the middle to lower tier for runtime, it demonstrated lower memory consumption than most Python-based methods in the comparative analysis.

Revealing cell types across different patients and disease states with scPriorGraph

Single-cell sequencing technology has found widespread applications and holds great promise in the medical field. Currently, this technology is being used in the research of various diseases, including cancer, osteoarthritis, and atherosclerosis. In cancer research, for example, researchers utilize the sequencing of cancerous cells to gain a deeper understanding of the genetic evolution and molecular mechanisms underlying clonal diversity within and between cancer subpopulations. This knowledge aids in designing more precise drug combinations to enhance drug effectiveness and reduce toxicity, all while addressing tumor heterogeneity. To assess the model's potential in the context of downstream medical analysis, clinical data was collected for performance evaluation. The clinical data primarily originates from the Human Artery dataset and the Human Bone dataset. Based on these datasets, we designed experiments involving cross-patient and cross-disease-state analyses. Cross-patient experiments entailed reference and test data derived from different patients, while cross-disease-state experiments involved reference and query data from healthy and diseased tissues, respectively. These cross-patient and cross-disease-state experiments enable us to evaluate the model's applicability and reliability when dealing with data samples obtained from diverse conditions. Importantly, they hold significant practical value, especially in clinical applications where it is often challenging to rapidly and accurately determine the subtype information of cells within collected samples due to various constraints. If our model can predict cell types based on annotated samples from different patients or from various regions of the same patient, it would be of immense significance for medical research.

In the Human Artery dataset, we utilized data from GSE159677 [47], comprising six experimental groups. Among them, AtherosclerоticCore-1 (GSM4837523) and ProximalAdjacent-1 (GSM4837524) represented sequenced data from patient 1, with afflicted and non-afflicted tissues. Similarly, AtherosclerоticCore-2 (GSM4837525) and

ProximalAdjacent-2 (GSM4837526) were from patient 2, and AtherosclerotiCore-3 (GSM4837527) and ProximalAdjacent-3 (GSM4837528) were from patient 3, all with afflicted and non-afflicted tissues. We designed six paired reference-query datasets for mutual predictions between healthy and diseased data within the same patient. The Human Bone dataset was derived from GSE152805 [48], focusing on human osteoarthritis data. We selected six datasets, namely, Cartilage-oLT-1 (GSM4626766), Cartilage-oLT-2 (GSM4626767), Cartilage-oLT-3 (GSM4626768), Cartilage-MT-1 (GSM4626769), Cartilage-MT-2 (GSM4626770), and Cartilage-MT-3 (GSM4626771), encompassing 26,192 chondrocytes from three knee osteoarthritis patients. These cells included 11,579 from the inner lesion area and 14,613 from the relatively unaffected lateral tibial plateau. We used healthy cell data from patients to predict cell types in different patients, leading to the design of nine reference-query dataset pairs.

In the cross-patient experiment within the Human Artery dataset, we employed ProximalAdjacent-3 to predict the cell types in ProximalAdjacent-1, both from different patients. Our method outperformed all seven other comparative methods, achieving an accuracy of 0.89. The Sankey diagram (see Fig. 4a) illustrated that scPriorGraph provided accurate results for most cell types, with only a minor misclassification of NK cells as T cells. In contrast, other methods, such as scPred, SingleR, singleCellNet, and MarkerCount, exhibited varying degrees of misclassifications and discrepancies. In the two sets of heatmaps (see Fig. 4b), the left side demonstrated the cross-patient experiment in the Human Bone dataset, utilizing Cartilage-oLT-1 to predict Cartilage-oLT-3. On the right, the cross-disease-state experiment was depicted using Cartilage-oLT-2 to predict Cartilage-MT-2. In both experiments, our method displayed notably accurate predictions. In the cross-patient experiment, CHETAH misclassified numerous preFC cells as FC cells. In the cross-disease-state experiment, scLearn did not accurately predict the HomC cell category, while CHETAH misclassified multiple cell types.

Improving drug response prediction in cancer cell lines with scPriorGraph

Cancer is a disease driven by genetic mutations that regulate cell functions, especially cell growth and division. Current cancer treatments include radiotherapy, chemotherapy, and targeted therapies. While radiotherapy uses radiation to kill tumor cells and chemotherapy employs chemical agents for the same purpose, these approaches can harm healthy cells and cause significant side effects. Targeted therapies, designed to specifically target tumor growth pathways or molecular markers, offer better precision and fewer side effects. However, drug sensitivity varies among patients and cancer cell types. With the growing use of sequencing technologies, we have access to more cancer cell expression data. The Genomics of Drug Sensitivity in Cancer (GDSC) [49, 50] study explores drug responses in cancer cell lines. Using gene expression profiles, we can predict how specific cancer cell lines will respond to drugs, aiding medical analysis and clinical treatment planning.

We collected gene expression and drug response data from the GDSC database through CaDRReS-Sc [51]. This dataset includes gene expression profiles from 1018 cell lines across 17,737 genes and drug response data for 1074 cell lines exposed to 226 different drugs. After data filtering, we selected 985 cell lines and 17,419 genes for further analysis, followed by threefold, fivefold, and tenfold

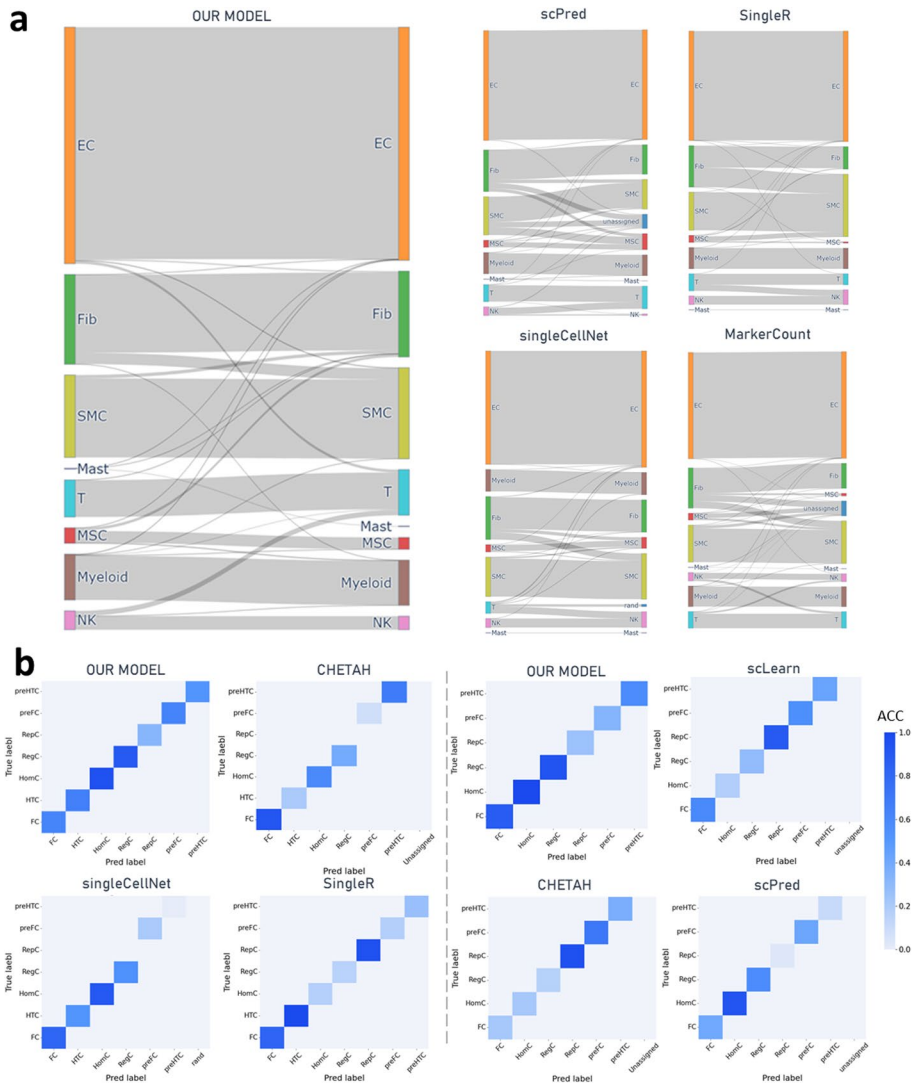

Fig. 4 Comparison of scPriorGraph with other methods in clinical data experiments. **a** In the cross-patient experiments, scPriorGraph and other methods were compared for the accuracy of cell classification using a Sankey diagram. The left side of the diagram represents the true cell types, and the right side represents the predicted cell types. **b** Accuracy of classification displayed using heatmaps, with the *y*-axis representing the true cell types and the *x*-axis representing the predicted cell types. Darker colors indicate a higher number of cells assigned to that type. The data on the left half of the plot is sourced from cross-patient experiment GSM4626766-GSM4626768, while the data on the right half is sourced from cross-disease-state experiment GSM4626767-GSM4626770

cross-validation experiments. The performance evaluation metric for assessing the correlation between predicted values and actual values was the Spearman correlation coefficient. The comparison methods we selected are CaDRReS-Sc, DREEP [52], beyondcell [53], and SCAD [54]. In the testing phase, scPriorGraph achieved the highest Spearman correlation coefficient value of 0.738 (see Fig. 5a and Additional file 1: Table. S11), with an average of 0.670 for the fivefold cross-validation. In Fig. 5b, the Spearman correlation coefficients obtained from threefold, fivefold, and tenfold cross-validation experiments indicate that scPriorGraph outperformed the other methods. The supervised training-based models, including scPriorGraph,

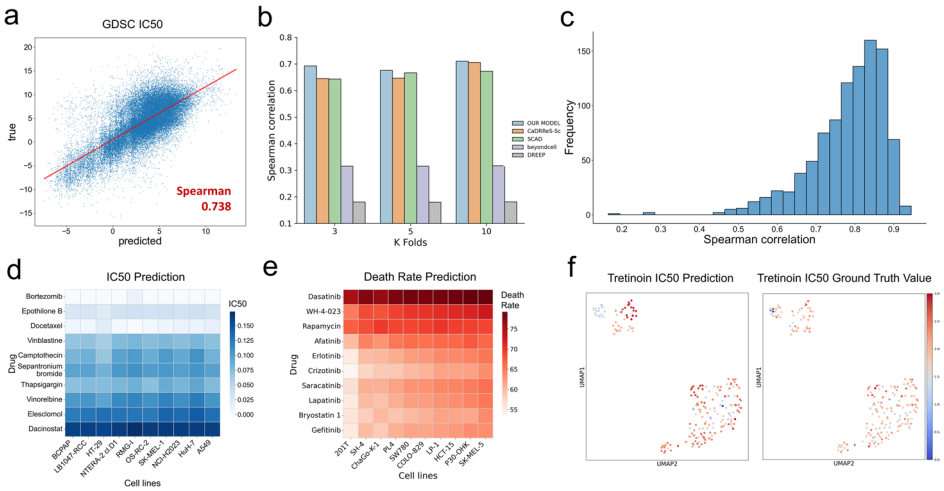

Fig. 5 Performance analysis of scPriorGraph in drug response prediction. **a** Regression analysis using actual drug response data and predicted data. **b** Comparison of average Spearman's correlation coefficient between scPriorGraph, CaDRReS-Sc, SCAD, beyondcell, DREEP in 3-fold, 5-fold, and 10-fold cross-validation experiments. **c** The distribution of Spearman correlation values from the leave-one-out cross-validation results of scPriorGraph on the GDSC dataset. **d** Visualization of the model's predicted Half-Maximal Inhibitory Concentration (IC50) values. **e** Visualization of predicted cell death percentages at a given drug dosage. **f** UMAP projection for comparing the predicted and actual responses of cell lines to the drug using scPriorGraph

CaDRReS-Sc, and SCAD, demonstrated superior performance compared to methods such as beyondCell and DREEP, which are dependent on the enrichment analysis of drug feature sets. To more comprehensively evaluate the performance of the model, we employed leave-one-out cross-validation, with the distribution of predicted Spearman correlation scores presented in Fig. 5c. Most results fell within the 0.7–0.9 range, and the mean Spearman correlation coefficient was 0.78, higher than previous cross-validation results. We further test scPriorGraph on another dataset of sci-Plex3 [55], focusing on the prediction of single-cell drug responses for five drugs at four different dosages. We followed the original dataset's train-test split and used the proliferation index as the prediction target. In the experiments, we obtained an average Spearman correlation value of 0.584 across all experiments involving five drugs at four dosage levels. The detailed results are included in Additional file 1: Table. S12.

We selected cell lines with the top-10 smallest sum of predicted IC50 values or mortality rates, and visualized the predicted IC50 values and corresponding cancer cell mortality rates at specific drug concentrations, as shown in Fig. 5d and e. Furthermore, we utilized Uniform Manifold Approximation and Projection (UMAP) to illustrate the differences between model predictions and actual drug responses, as displayed in Fig. 5f. By comparing the predicted IC50 values with the actual values for cell lines treated with Tretinoin, we observed that scPriorGraph was generally able to accurately classify cells as either sensitive or resistant to the drug. The mean squared error (MSE) between the predicted and actual values presented in Fig. 5f was calculated to be 10.036.

Refining pathway and ligand-receptor network-level analysis with scPriorGraph

Our model integrates hierarchical genetic biological semantic information, acquires intercellular communication data through ligand-receptor interactions, and captures intracellular gene interaction information via pathways. Users can choose pathway database based on their specific needs. In our experiments, we opted for the KEGG pathway database, which encompasses manually curated pathway maps illustrating molecular interactions, reactions, and network relationships. It is categorized into seven major sections: Metabolism, Genetic Information Processing, Environmental Information Processing, Cellular Processes, Organismal Systems, Human Diseases, and Drug Development. In our study, we focused on pathways associated with gene interactions and network relationships that play a role in various physiological activities within cells, such as metabolism and immunity. For instance, in Fig. 6a, the pathway "hsa00010" pertains to the glycolysis and gluconeogenesis of glucose in humans, while the pathway "hsa00640" relates to the metabolism of citric acid in humans. By integrating pathways

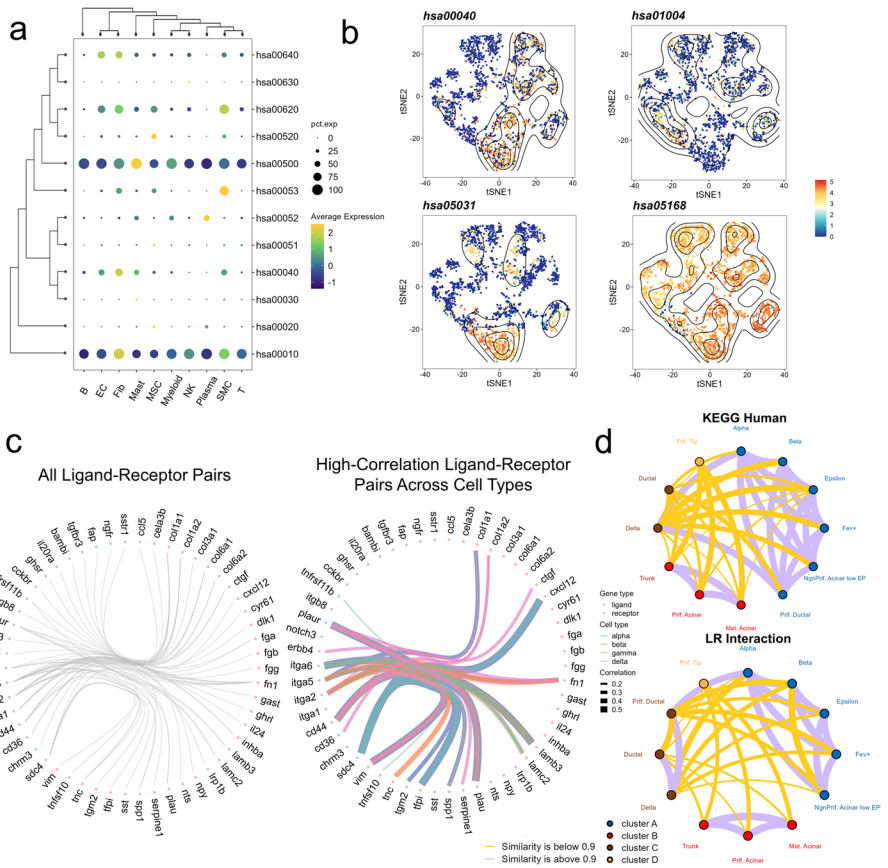

Fig. 6 Single-cell analysis in pathway and ligand-receptor network-level. **a** Dot plot illustrating pathway-level expression variability across different cell clusters. In the plot, lighter colors indicate higher average expression levels, while larger dot sizes represent higher expression proportions in clusters. The pathways utilized are sourced from the KEGG pathway database, and the scRNA-seq data is derived from GSE159677. **b** Feature plots illustrating the expression variability of pathways within the scRNA-seq dataset. **c** Visualization of Ligand-Receptor Interactions Used for Constructing Intercellular Networks. **d** Comparing cell cluster similarity with embedded hierarchical information. The thickness and color of the lines represents the level of similarity

into scRNA-seq data, we gained insights into pathway expression across different cell types. The dot plot in Fig. 6a visualizes pathway expression across different cell types, with dot colors denoting distinct expression levels, and dot sizes indicating expression proportions. Notably, hsa00010 exhibits higher average expression levels in Fib cells, while hsa00640 demonstrates higher average expression levels in EC and Fib cells. Furthermore, while hsa00010 shows consistently high expression proportions across various cell types, hsa00640 exhibits higher expression proportions in EC and Fib cells, and lower proportions in other cell types. We attempted to further confirm the correlation between specific pathways and cell types highlighted in Fig. 6a by consulting existing literature. Notably, pathway hsa00500, which is related to human starch and sucrose metabolism, showed a significantly higher level of expression in Mast cells, as depicted in Fig. 6a and evidenced by [56]. Additionally, elevated expression levels of pathway hsa00020, associated with the human Citric Acid Cycle (TCA Cycle), were observed in mesenchymal stem cells, as confirmed by [57].

Single-cell RNA sequencing data is typically sparse and noisy, and using genes as cell features may negatively impact the model due to low expression levels and proportions. However, utilizing pathways as cell features significantly improves this issue, with noise in individual genes having a lesser impact on the entire pathway. In Fig. 6b, a tSNE plot visually represents pathway expression across the entire dataset. Cell colors and contour lines in the plot depict pathway expression, revealing the distribution of pathways like hsa00040 and hsa05031 and their varying expression levels across the dataset. In Fig. 6c, the left diagram shows a network composed of ligand-receptor gene pairs, while the right diagram displays the top 10 ligand-receptor gene pairs with the highest Pearson correlation coefficients in each cell type. Each color represents a different cell type, and the width of the lines between ligand-receptor genes is based on the Pearson correlation coefficients calculated from gene expression levels.

The scPriorGraph utilizes two distinct biological prior knowledge, derived from intracellular and intercellular sources, to construct two different cell KNN graphs. Based on the same expression data, cell-to-cell similarities vary under different gene semantic information. Figure 6d illustrates the homogeneity and heterogeneity between the intracellular and intercellular networks constructed by scPriorGraph, based on the GSE132188 dataset. This figure displays two similarity networks, each comprising 12 cell types, which are developed using the KEGG Human Pathway Database and the Ligand-Receptor Interaction Database. These networks exhibit a high degree of similarity. Notably, the internal similarity within clusters B and D, as well as their similarity with other cell types, is consistently maintained. The majority of the members of cluster A also show general consistency in their network positioning. However, there are significant changes in local similarity patterns observed. One particular example is the change in the clustering of cells identified as *Prlf. Ductal*, which signifies a shift in how these cells are grouped when comparing intracellular to intercellular network representations.

Evaluating scPriorGraph model robustness with ablation experiments

To validate the model's performance under various parameter settings, we conducted tests with different values of k and α. Here, k represents the proportion of neighbors in the k-nearest neighbor graph relative to all cells, and α denotes the ratio

of cross-entropy loss in the total loss function. Figure 7a illustrates the comparison of accuracy for different k values across various reference-query dataset pairs. It is evident that k values in the range of 0.05 to 0.1 yielded the best results for most experiments. In Fig. 7b, we compare the accuracy of different α values for the reference-query dataset pairs. This analysis reveals that α values between 0.3 and 0.5 generally led to the best performance.

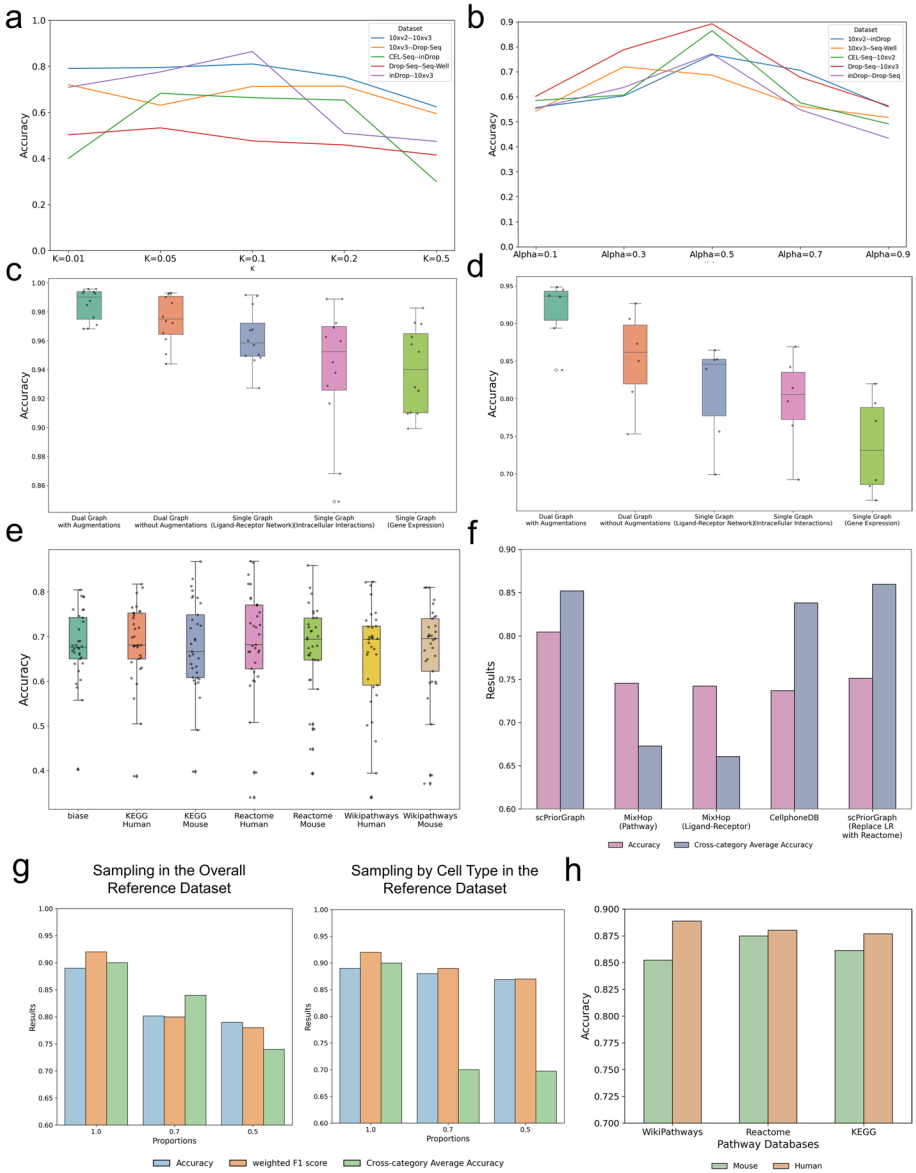

Fig. 7 Validating model robustness by adjusting model parameters and structure. **a** Differences in model performance under various neighbor node ratios. **b** Comparative performance of the model under different loss function weights. **c** Comparative performance of the model after removal of various components on the Human Pancreas dataset. **d** Comparative performance of the model after removal of various components on the Human Artery dataset. **e** Comparative performance of the model after substituting with different pathways. **f** Performance comparison of different strategies for consider intercellular communication. **g** Performance comparison on difference sizes of reference datasets. **h** Comparison of prediction accuracies using pathway databases across different species

In order to assess the contributions of individual components of our model, we conducted ablation experiments. Ablation experiments were conducted using the Human Pancreas dataset and the Human Artery dataset. Figure 7c and d displays the performance of five models: the full model, a model without graph enhancement, a single-graph model that solely uses *k*-nearest neighbor graphs generated from pathway information, another single-graph model using *k*-nearest neighbor graphs generated from ligand-receptor pathway information, and model that build cell graphs using only gene expression data. As the model is simplified by removing components, the results show a corresponding decrease in performance.

Our model employs two graphs, each incorporating pathway and ligand-receptor pathway information. In the field of biology, there exist various pathway databases from different sources such as KEGG, Reactome [58], Wikipathways [59], and more. These databases encompass pathways from diverse species and biological processes, and selecting different pathways can yield different model results. Figure 7e demonstrates the performance of the model with the use of different databases, including de novo pathway [60] database generated from Biase [61] data. The data indicate that Reactome Human pathways generally produce superior results compared to other pathways.

To assess the effectiveness of the graph neural network implemented in scPriorGraph, we conducted a series of experiments where we substituted the original cell graph, which incorporated ligand-receptor information, with a cell graph generated by CellphoneDB [62]. Additionally, we explored the performance of a higher-order graph neural network, MixHop [63], by constructing two variants: MixHop (pathway) and MixHop (Ligand-Receptor), using the same graphs as those used in scPriorGraph. To evaluate the impact of the choice of gene set, we replaced the original ligand-receptor gene set with Reactome. Figure 7f presents the comparative performance of these four models alongside scPriorGraph. The results indicate that the cell graphs developed using the ligand-receptor gene set are of high quality. Furthermore, scPriorGraph demonstrated effective utilization of inter-cellular relationships to accurately predict cell types, showcasing its robustness and the potential advantages of integrating detailed molecular interactions in cell type classification.

To investigate the impact of reference size and coverage on model performance, we conducted experiments using the Human Artery dataset. We employed random sampling at three different proportions—100, 70, and 50%—both on the overall training set and internally within different cell types of the training set. Predictions were then made on the same query dataset. We observed varying degrees of change in various performance metrics as the sampling ratio was altered, as depicted in Fig. 7g.

To evaluate the impact of pathway databases from different species on the performance of scPriorGraph, we conducted multiple experiments using the Human Artery dataset. We evaluated model performance by taking the average accuracy of predictions across different pathway databases. As shown in Fig. 7h, human pathways from WikiPathways demonstrated a significant improvement compared to mouse pathways. This discrepancy in performance likely stems from the species-specific nature of the gene expression data used in the model.

Discussion

In this study, we developed scPriorGraph, a dual-channel graph neural network that integrates multi-level gene biological semantic information, providing accurate, batch-effect-insensitive cell type identification across platforms, species, and temporal dimensions. scPriorGraph combines biological prior knowledge with scRNA-seq data to establish high-quality cell relationships. To obtain cell relationships that cannot be inferred solely from the expression matrix due to self-expression sparsity, scPriorGraph employs mutual information matrices to calculate global cell similarities for graph augmentation. Leveraging the incorporation of biological prior knowledge and graph augmentation techniques, scPriorGraph achieves high-quality graph embeddings for tasks such as cell type identification and drug response prediction. scPriorGraph demonstrates excellent performance in cell type identification and drug response prediction compared to existing methods in the field.

Obtaining high-quality graphs is paramount for graph neural networks. In prior research, graph neural networks have been applied to the field of single-cell data analysis, encompassing tasks such as cell classification, clustering, and dropout imputation. However, current methods predominantly rely on expression matrices to generate graphs, with limited consideration given to incorporating prior biological knowledge into graph construction to enhance graph quality, consequently improving model performance. Given the complementarity of intra-cellular and inter-cellular information, scPriorGraph simultaneously accounts for intra-cellular gene interactions and inter-cellular communication. By incorporating multi-level gene biological semantic information into the model, the predictive capability of the model has been significantly enhanced. The model acquires intra-cellular gene interaction information through pathways, enabling users to customize gene collaboration information by switching between different pathways according to the specific task.

In our experiments, we conducted various tests to validate the ability of scPriorGraph in batch-effect correction across platforms, species, and temporal dimensions, as well as its potential applications in clinical data processing. By making minor modifications to the model, we enabled it to predict drug responses in cancer cell lines. Our model achieved superior results compared to state-of-the-art methods in the field, indicating that the graph embeddings generated by scPriorGraph through the dual-channel graph convolution layers are of high quality, and scPriorGraph holds the potential for application in other single-cell analysis domains.

Our experiments reveal that the selection of the reference dataset and pathway database significantly impacts the predictive performance of scPriorGraph. Specifically, when we use human scRNA data for testing (see Fig. 7h), pathways related to humans from databases such as WikiPathways, Reactome, and KEGG generally outperform those related to mice. It is advisable to choose a pathway database that matches the species of the dataset to achieve better model performance. Since scPriorGraph is based on supervised learning, it cannot predict cell types that are not present in the reference dataset. Therefore, selecting a comprehensive reference dataset that includes all relevant cell types is crucial. If the reference dataset lacks certain cell types, unless a threshold is set, scPriorGraph will classify them as one of the known types from the reference, which can affect the accuracy of the predictions. Additionally, as evidenced by the results in

Fig. 7g, the amount of data is also critical. To ensure robustness, each cell type in the reference dataset should have at least 50 samples. Furthermore, the sequencing data should generally contain more than 2000 genes to provide a rich set of pathway-level features. An imbalance in cell types or insufficient gene numbers can reduce the effectiveness of the model. For rare cell types, the reference dataset should have at least 20 cells to support model training. Finally, selecting reference data from patients who share the same gender, are of a similar age, are at comparable stages of the disease, and have had their samples sequenced using the same platform as the query data can effectively enhance prediction accuracy.

In this study, for the first time, we introduced intra-cellular interactions and inter-cellular information propagation into graph neural network for cell type identification. To address the sparsity inherent in scRNA-seq data, we incorporated the graph augmentation technique based on global cell similarity into this domain. Currently, the landscape of cell type identification is diverse, encompassing various aspects such as temporal dynamics, diseases, and developmental states. In the future, we aim to expand our datasets to cover a wider range of cell type identification scenarios, making our model applicable to any situation where cells can be discretely categorized, including cell subtypes or customized cell states. We will also continue to expand our model into more single-cell data analysis domains, further facilitating various downstream analyses with biological significance.

Conclusions

Selecting an appropriate subset from the high-dimensional gene list is a critical task in the computational realm of single-cell cell type prediction. We firmly believe that knowledge of gene interactions, both within and between cells, from both intracellular and intercellular perspectives, represents valuable and complementary data that can significantly enrich the pursuit of this computational goal. In an effort to explore this challenge, our study introduces scPriorGraph, a novel dual-channel graph neural network seamlessly integrating multi-layered gene biological semantic information. It equips us with an efficient, robust, and batch-effect-insensitive capability for cell type identification. Acknowledging the complementary nature of intracellular and intercellular information, scPriorGraph leverages ligand-receptor networks to embark on random walks to gather intercellular communication insights. Simultaneously, it taps into pathway data to extract intracellular gene interaction information, harmoniously amalgamating these facets into a dual-channel graph structure. Moreover, users have the flexibility to selectively integrate gene biological semantic information into the model to align it precisely with the unique requirements of diverse classification tasks. To supplement cellular relationships that may elude conventional cell graphs, scPriorGraph employs advanced graph-enhancement techniques based on mutual information matrices. We conducted a comprehensive series of experiments, encompassing cross-platform, cross-species, cross-development stages, cross-patient, and cross-disease state analyses, to rigorously evaluate the model's performance. In comparisons with a range of methodologies, scPriorGraph emerged as a superior performer in terms of stability and accuracy. Furthermore, we enhanced our model by substituting the classifier layer with a fully connected regression layer, enabling it to

process IC50 values for predicting drug responses in cancer cell lines. The promising results illuminate the model's potential for expansion into diverse domains. Our future objectives encompass a broader application of the model in an array of cell type identification scenarios and a deeper exploration of its capabilities in the analysis of scRNA-seq data.

Methods

Obtaining intercellular communication information through random walks

scPriorGraph is a dual-graph neural network that combines both intra-cellular and intercellular information. The model incorporates two k-nearest neighbor (KNN) graphs, each capturing intercellular ligand-receptor pathway information and intracellular pathway information, which are referred to as $A1$ and $A2$, respectively. To extract ligand-receptor path information, we constructed a ligand-receptor heterograph from a series of ligand-receptor gene pairs, and collected path information through random walks within the network. The ligand-receptor heterograph is defined as $G = (V, E, T)$, where V represents the nodes of ligands and receptors in the graph, E represents the edges connecting the ligand and receptor nodes, and T is a set encompassing the categories of nodes and edges. In order to effectively collect path information on the heterograph, we designed a meta-path-based random walk. The meta-path schema is defined as $V_1 \overset{R_1}{\rightarrow} V_2 \overset{R_2}{\rightarrow} \cdots \overset{R_{l-2}}{\rightarrow} V_{l-1} \overset{R_{l-1}}{\rightarrow} V_l$. The probability of moving at the ith step is as follows:

$$p(v_{i+1}|v_i) = \begin{cases} \frac{1}{|N(v_i)|}, (v_{i+1}, v_i) \in E \\ 0, (v_{i+1}, v_i) \notin E \end{cases}, \tag{1}$$

where $N(v_i)$ represents the neighbors of node v_i. We initiated the random walker from ligand nodes, collecting 707 path information by random walking on the ligand-receptor heterograph. These paths encapsulate the communication processes between intercellular ligands and receptors, and will serve as the foundation for constructing cell graph $A1$.

The pathway information utilized in our method is extracted from databases such as KEGG [24], Reactome [58], WikiPathways [59], and others, which encompass gene pathway information related to various cellular responses and biological activities. In addition to the aforementioned three databases, we also provide a pathway database generated using algorithms. The pathway information retrieved from the pathway database will form the basis for constructing cell graph $A2$.

Integrating expression data with biological information

Once we have acquired ligand-receptor paths and pathways, the next step is to integrate this data with expression data for the generation of cell graphs $A1$ and $A2$. To assess the enrichment of gene sets (ligand-receptor path information or pathways) in single-cell RNA sequencing data, we employed the AUCell [64] method. The AUCell method utilizes the area under the curve (AUC) to determine whether a gene set is enriched in the gene expression of each cell. By examining the distribution of AUC scores across cells, we can obtain the expression patterns of features associated with the gene set in the cells.

Integrating features at the gene level and features at the gene set level

To merge the cell-level features and gene set-level features during the generation of cell k-nearest neighbor graphs, we employed the Similarity Network Fusion (SNF) [65] method. SNF is an effective approach for integrating different networks. In order to integrate patient similarity networks from different data sources, such as amalgamating patient similarity graphs derived from mRNA expression data and those derived from DNA methylation, SNF constructs corresponding similarity networks for each data type and utilizes a non-linear fusion approach to integrate these networks into a unified similarity network. Given SNF's capability to merge data with different metric, we applied it to integrate gene-level expression information and gene set-level expression information. Let $W(i,j)$ denote the similarity matrix between cell i and cell j, and the definition of the normalized weight matrix P is as follows:

$$P(i,j) = \begin{cases} \frac{W(i,j)}{2\sum_{k\neq i} W(i,k)}, j \neq i \\ \frac{1}{2}, j = i \end{cases}. \tag{2}$$

The local similarity matrix represented as a KNN graph is expressed as follows:

$$S(i,j) = \begin{cases} \frac{W(i,j)}{\sum_{k\in N_i} W(i,k)}, j \in N_i \\ 0, otherwise \end{cases}. \tag{3}$$

where N_i represents the set of neighboring cells for cell i. SNF sets the similarity of cell i to its non-neighboring cells to zero in this manner. SNF employs the following formula to iteratively update the similarity matrices corresponding to each data type:

$$P_{t+1}^{(1)} = S^{(1)} \times P_t^{(2)} \times (S^{(1)})^T \tag{4}$$

$$P_{t+1}^{(2)} = S^{(2)} \times P_t^{(1)} \times (S^{(2)})^T. \tag{5}$$

The matrix output by SNF after iteration is as follows:

$$P^{(c)} = \frac{P_t^{(1)} + P_t^{(2)}}{2}. \tag{6}$$

We constructed k-nearest neighbor graphs, A1 and A2, based on the similarity matrices output by the SNF method and employed them for the creation of graph-based convolution, $Conv_{A1}$ and $Conv_{A2}$.

Graph augmentation based on positive pointwise mutual information

The KNN graphs $A1$ and $A2$ aggregate expression information of similar cells, but they still have limitations. For instance, when the gene expression information is too sparse, it may result in low-quality graphs. To overcome this, we introduce positive pointwise mutual information (PPMI) matrix in our model to capture potential relationships between cells that were not captured by the KNN graph and enhance the graph. We first calculate the frequency matrix F. The frequency matrix F is derived by computing the frequency of two nodes appearing in the paths produced by random walks on the graph. When the random

walker is located at node x_i at a specific time t, we denote this state as $s(t) = x_i$, and the probability of transitioning from the current node x_i to an adjacent node x_j is defined as:

$$p\big(s(t+1) = x_j | s(t) = x_i\big) = \frac{A_{i,j}}{\sum_j A_{i,j}}. \tag{7}$$

By sequentially setting each node in the graph as the root node and performing random walks, we obtain multiple paths. We then sample cell pairs for each path. For each sampled cell pair (x_n, x_m), in the frequency matrix F, the corresponding values Fn, m and Fm, n are incremented by 1. Using the frequency matrix F as a basis, we compute the Positive Pointwise Mutual Information (PPMI) matrix $P \in R^{n \times n}$ as:

$$p_{i,j} = \frac{F_{i,j}}{\sum_{i,j} F_{i,j}}; \tag{8}$$

$$p_{i,*} = \frac{\sum_j F_{i,j}}{\sum_{i,j} F_{i,j}}; \tag{9}$$

$$p_{*,j} = \frac{\sum_i F_{i,j}}{\sum_{i,j} F_{i,j}}; \tag{10}$$

$$P_{i,j} = \max\left\{pmi_{i,j} = \log\left(\frac{p_{i,j}}{p_{i,*}p_{*,j}}\right), 0\right\}. \tag{11}$$

The probability $p_{i,j}$ represents the likelihood of cell x_i occurring in context c_j. Similarly, $p_{i,*}$ represents the probability of cell x_i in general, and $p_{*,j}$ represents the likelihood of context c_j in general.

Following the statistical concept of independence, if x_i is independent of c_j, then $p_{i,j}$ equals the product of $p_{i,*}$ and $p_{*,j}$, and in such cases, $P_{i,j}$ equals 0. However, when there is a connection or association between x_i and c_j, $p_{i,j}$ will be greater than the product of $p_{i,*}$ and $p_{*,j}$. Since we are interested in the case where x_i and c_j are related, we use non-negative pmi values. For each KNN graph, we generated a corresponding PPMI matrix for constructing graph-structured convolutions $Conv_{P1}$ and $Conv_{P2}$.

Model construction

The scPriorGraph proposed in this study is a dual-graph neural network based on Graph Convolutional Networks (GCN). The fundamental concept of GCN is to combine node feature information with network topology for propagation. The model typically consists of two steps: local neighborhood feature aggregation and feature transformation. In the local neighborhood feature aggregation step, for each node i, GCN aggregates and transforms the features of its neighboring nodes to obtain a new feature representation for that node. In the feature transformation step, GCN employs a learnable linear transformation matrix to convert the aggregated feature representation into a new feature representation, facilitating aggregation and transformation in the subsequent layer.

The primary structure of scPriorGraph consists of $Conv_{A1}$ and $Conv_{A2}$ based on graphs $A1$ and $A2$, and $Conv_{P1}$ and $Conv_{P2}$ for graph augmentation using PPMI matrices $P1$ and $P2$. The cell KNN graph is represented by an adjacency matrix $A \in \mathbb{R}^{N*N}$, and the expression matrix is denoted as $X \in \mathbb{R}^{N\times D}$, where N represents the number of cells, and D signifies the dimension of cell features. The ith hidden layer $H^{(i)}$ for $Conv_{A1}$ and $Conv_{A2}$ is defined as follows:

$$H^i = Conv_A^i(X) = \sigma(\widetilde{D}^{-\frac{1}{2}}\widetilde{A}\widetilde{D}^{-\frac{1}{2}}H^{i-1}W^i), \tag{12}$$

where W^i is the weight matrix for the ith layer, and $\sigma(\bullet)$ represents a non-linear activation function like ReLU. To account for the self-influence of nodes, we augmented the adjacency matrix A with an identity matrix $I_N \in R^{N\times N}$, resulting in a self-connected adjacency matrix $\widetilde{A} = A + I_N$. The $\widetilde{D}$ matrix is defined as $\widetilde{D} = \sum_j \widetilde{A}_{i,j}$, and $\widetilde{D}^{-\frac{1}{2}}\widetilde{A}\widetilde{D}^{-\frac{1}{2}}$ represents the normalized adjacency matrix. H^{i-1} is the output of the i-1 layer, with $H^0 = X$. W^i denotes the trainable parameters of the network. The ith hidden layer $H^{(i)}$ for $Conv_{P1}$ and $Conv_{P2}$, based on PPMI matrices $P1$ and $P2$, is defined as follows:

$$H^i = Conv_P^i(X) = \sigma(D^{-\frac{1}{2}}PD^{-\frac{1}{2}}H^{i-1}W^i), \tag{13}$$

where P is the PPMI matrix, and $D = \sum_j P_{i,j}$. To integrate the graph embeddings that aggregate both intra-cellular and inter-cellular information in the model, namely the outputs of $Conv_{A1}$ and $Conv_{A2}$, we used a linear layer. Let Z_{A1} and Z_{A2} represent the outputs of $Conv_{A1}$ and $Conv_{A2}$, and the integrated output Z_A is defined as follows:

$$Z^A = softmax(\sigma(Linear([Z_{A1}, Z_{A2}]))), \tag{14}$$

where $[\bullet, \bullet]$ denotes the concatenation operation, the dimensions of Z_A are the same as Z_{A1} and Z_{A2}. The softmax activation function is defined as $softmax(x_i) = \frac{1}{\mathcal{Z}}\exp(x_i)$, with $\mathcal{Z} = \sum_i \exp(x_i)$. The model employs cross-entropy to define the supervised training loss function. Assuming there are c cell type labels for prediction, the dimensions of the softmax output Z_A are $Z_A \in \mathbb{R}^{N\times c}$. For multi-class problems, the cross-entropy is defined as follows:

$$\mathcal{L}_{CE} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{c} y_{ij} log\left(p_{ij}\right), \tag{15}$$

where N is the number of cells, c is the number of categories, $\mathrm{y_{ij}} \in \{0, 1\}$, where y_{ij} is equal to 1 if the ith sample belongs to category j, and equal to 0 otherwise. p_{ij} is the probability assigned by the model for sample i to belong to category j, with $\sum_{j=1}^{C} p_{ij} = 1, i = 1,2,\cdots,N$.

In order to incorporate graph-enhanced information into the model, the model is trained in a supervised manner for $Conv_{A1}$ and $Conv_{A2}$, while also using the outputs of $Conv_{A1}$ and $Conv_{A2}$ to train $Conv_{P1}$ and $Conv_{P2}$. $\mathcal{L}_{MSE}(Conv_{A1}, Conv_{P1})$ and $\mathcal{L}_{MSE}(Conv_{A2}, Conv_{P2})$ represent the loss functions obtained by training $Conv_{P1}$ and $Conv_{P2}$ in an unsupervised manner. The model minimizes the above functions to integrate graph-enhanced information into the final model output. The definition of $\mathcal{L}_{MSE}$ is as follows:

$$\mathcal{L}_{MSE} = \frac{\sum_{i=1}^{n} ||Z_{i,:}^{A} - Z_{i,:}^{P}||^2}{n}, \tag{16}$$

where Z^A represents the output of the merged graph embeddings generated by $Conv_{A1}$ and $Conv_{A2}$, and Z^P represents the output of $Conv_{P1}$ or $Conv_{P2}$. In addition to the two aforementioned loss functions, to ensure that the graph embeddings of the model output maintain the relationships between cells in the original graph structure, we incorporated a graph reconstruction loss in the model. The definition of the graph reconstruction loss is as follows:

$$\mathcal{L}_{Reconstruction} = MSE\Big(KNNGraph(Z^A), A1\Big) + MSE\Big(KNNGraph(Z^A), A2\Big), \tag{17}$$

where the MSE function is defined as $MSE = \frac{1}{n}\sum_{i=1}^{n}(A_i - R_i)^2$, where A_i represents the original graph, and R_i represents the reconstructed graph. We partition the loss function into two parts based on whether it is a graph reconstruction loss. The model's final loss function is defined as the weighted sum of the two loss functions:

$$Loss = \alpha\Big(\mathcal{L}_{CE}\Big(Z^A\Big) + \mathcal{L}_{MSE}(Conv_{A1}, Conv_{P1}) + \mathcal{L}_{MSE}(Conv_{A2}, Conv_{P2})\Big) + (1-\alpha)(\mathcal{L}_{Reconstruction}) \tag{18}$$

In this way, the model is able to consider both intracellular and intercellular information during training, along with the PPMI matrix information used for augmenting the graph structure.

Datasets

scRNA-seq data

We selected the human pancreas dataset and peripheral blood mononuclear cell (PBMC) dataset, which comprise a collection of multi-platform sequencing data for cross-platform experiments. The human pancreas dataset comprises sequencing data from the Baron Human [28], Muraro [29], Segerstolpe [30], and Xin [31]. The PBMC [27] dataset includes sequencing data for PBMCs from the 10Xv2, 10Xv3, CEL-Seq, Drop-Seq, inDrop, Seq-Well, and Smart-Seq2 platforms.

In the context of cross-species experiments concerning the cerebral cortex, the human and mouse datasets we employed originated from GSE115746 [44] and phs001790 [45], respectively. GSE115746 is a study that involves gene expression profiling to explore the diversity of cell types in the mouse cerebral cortex. On the other hand, phs001790 is a genomics study that focuses on the cell types within the human middle temporal gyrus (MTG). In the experiment involving pancreatic cells, human pancreatic cells and mouse pancreatic cells were sourced from GSE84133 [28], a single-cell transcriptome study investigating the various cell types in both human and mouse pancreases.

The cross-temporal developmental experiment data were sourced from GSE132188 [46]. This dataset consists of scRNA-seq data from uniform mouse pancreatic epithelial cells at four different embryonic stages, obtained using the 10X platform. We utilized the samples GSM3852753-GSM3852755 [46], representing the embryonic stages E13_5, E14_5, and E15_5, respectively.

Within the clinical data, the Human Artery dataset includes six sets of experimental data from GSE159677 [47]. Specifically, GSM4837523 and GSM4837524 [47] represent sequencing data from the diseased and non-diseased tissues of patient 1, GSM4837525 and GSM4837526 [47] from patient 2, and GSM4837527 and GSM4837528 [47] from patient 3. The Human Bone dataset was derived from the human osteoarthritis dataset GSE152805 [48]. We focused our selection on six specific datasets: GSM4626766, GSM4626767, GSM4626768, GSM4626769, GSM4626770, and GSM4626771 [48].

Drug-response data

In the drug response experiment, we employed drug response data for 1074 cancer cell lines across 226 drugs, which were acquired from the GDSC database [49, 50] through CaDRReS-Sc [51], as our experimental dataset. We employed the Half-Maximal Inhibitory Concentration (IC50) as the metric for assessing drug responses.

Criteria for benchmark selection

The criteria for selecting benchmarks in the comparative experiments between scPriorGraph and other models are as follows. (1) The reference dataset and the query dataset should not have too large a disparity in cell types. (2) The number of genes in the sequencing data should generally be greater than 2000. The genes in the expression data used for benchmarks should also overlap significantly with the Pathways. (3) The vast majority of cell type in the dataset must have a sufficient number of cells to support model training. Based on experience, the number of cell samples for rare cell types present in the reference dataset should be at least 20. We have listed all the reference-query data pairs we used in this work in Additional file 1: Table. S13.

Data preprocessing

In our comparative experiments, if a method can operate independently of log-normalization or other data preprocessing procedures, we use raw count data as input for that method. If data preprocessing is essential for certain methods, where its absence would impact the method's performance, or if the preprocessing steps are embedded within the method, then we follow the tutorials provided by the authors to preprocess the data. We

Table 2 Setting details for method comparison

Method	Version	Input data	Method	Version	Input data
CHETAH	1.14.0	Raw count	scANVI	1.1.2	Log-normalized
scPoli	0.6.1	Raw count	expiMap + treeArches	0.6.1	Log-normalized
scLearn	1.0	Raw count	MarkerCount	0.6.6	Log-normalized
scmapCell	1.20.2	Raw count	scGPT	0.2.1	Log-normalized
scClassify	1.5.1	Raw count	Seurat	5.0.1	Log-normalized
SingleR	2.0.0	Raw count	scBert	1.0.0	Log-normalized
singleCellNet	0.1.0	Raw count	scPred	1.9.2	Log-normalized
TOSICA	1.0.0	Raw count	scType	Releases Jan 13, 2022	Log-normalized
sciBet	0.1.0	Raw count	scTyper(a2bi)	0.1.0	Log-normalized
simple linear SVM	1.3.0	*z*-score	CellAnn	1.0	Log-normalized

have organized the version numbers of the methods used in our comparisons as well as the data used, and have included this information in Table 2.

Approaches to obtain cell embeddings

In Fig. 2c, we compare the clustering performance of scPriorGraph with scPred, CHETAH, and SingleR. Regarding scPred, after training on the reference dataset, the undimension-reduced query dataset is input into the scPredict() method, which returns a Seurat object. Accessing $scpred@cell.embeddings retrieves a 50-dimensional reduced expression matrix for all cells, which we use as the embedding output of scPred. For CHETAH, after inputting the reference and query datasets into the CHETAHclassifier() method, assuming the output from CHETAHclassifier() is stored in output, accessing output$int_colData$CHETAH$correlations retrieves the correlations of cells with respect to various classification nodes produced by the CHETAH method. We use this matrix as the embeddings for the CHETAH method. Regarding SingleR, its output includes probability scores for each cell type. We have extracted these scores and used them as the embeddings for SingleR.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-024-03357-w.

Additional file 1. In the article provides comprehensive results from the scPriorGraph and comparative methods across various datasets and metrics. It includes performance evaluations for the Human Pancreas and PBMC datasets across metrics like Accuracy, F1 Score, Cohen's Kappa, and Cross Category Average Accuracy, as well as accuracy assessments for cross-species and developmental time datasets. Also featured are results from drug response prediction experiments on the GDSC and sci-Plex3 datasets, and a list of used reference-query dataset pairs.

Additional file 2. Review history.

Review history
The review history is available as Additional file 2.

Peer review information
Andrew Cosgrove was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Authors' contributions
YAH conceived of the study. YAH and XYC analyzed and interpreted the data, wrote the manuscript, and wrote the code for scPriorGraph. XYC ran the experiments on benchmarking the deconvolutional accuracy. ZHY and YAH supervised the project. XQS, LH, PWH, and ZAH provided resources for the research. All authors approved the final manuscript.

Funding
This work is supported in part by the National Science Fund for Distinguished Young Scholars of China (62325308), the Fundamental Research Funds for the Central Universities (D5000230199), Natural Science Foundation of Guangdong Province of China (2024A1515011984), and Specific Research Project of Guangxi for Research Bases and Talents (2021AC19354).

Availability of data and materials
The datasets used during the current study were obtained from publicly available repositories or data portals and are described in the "Methods" section. The Human and Mouse pancreatic islet datasets from Baron et al. are available in the GEO database under the accession code GSE84133 [66]. The Human pancreatic islet dataset from Muraro et al. is available in the GEO database under the accession code GSE85241 [67]. The Human pancreatic islet dataset from Segerstolpe et al. is available in the ArrayExpress database under the accession code E-MTAB-5061 [68]. The Human pancreatic islet dataset from Xin et al. is available in the GEO database under the accession code GSE81608 [69]. scRNA-seq dataset for PBMC is available in the GEO database under the accession code GSE132044 [70]. scRNA-seq datasets for human and mouse cerebral cortex are available in the GEO database under the accession code GSE115746 [71] and database of GenoFtypes and Phenotypes (dbGaP) under the accession code phs001790 [72]. scRNA-seq dataset for mouse pancreatic epithelial cells is available in the GEO database under the accession code GSE132188 [73]. scRNA-seq dataset for human atherosclerosis is available in the GEO database under the accession code GSE159677 [74]. scRNA-seq dataset for human osteoarthritis is available in the GEO database under the accession code GSE152805 [75]. Sequencing data and

drug response data of cancer cell lines in GDSC database is available in CaDRReS-Sc [76]. The sci-Plex3 dataset we used was obtained from [77].
scPriorGraph source codes have been deposited at the GitHub repository (https://github.com/ChrisOliver2345/scPriorGraph) [78]. The repository is licensed under the open-source GPL-3.0. Source code for the software release used in the paper has been placed into a DOI-assigning repository (https://zenodo.org/records/10981089) [79].

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

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

Received: 26 October 2023 Accepted: 29 July 2024
Published online: 05 August 2024

References

1. Potter SS. Single-cell RNA sequencing for the study of development, physiology and disease. Nat Rev Nephrol. 2018;14:479–92.
2. Kolodziejczyk AA, Kim JK, Svensson V, Marioni JC, Teichmann SA. The technology and biology of single-cell RNA sequencing. Mol Cell. 2015;58:610–20.
3. González-Silva L, Quevedo L, Varela I. Tumor functional heterogeneity unraveled by scRNA-seq technologies. Trends in cancer. 2020;6:13–9.
4. Chen Y-P, Yin J-H, Li W-F, Li H-J, Chen D-P, Zhang C-J, Lv J-W, Wang Y-Q, Li X-M, Li J-Y. Single-cell transcriptomics reveals regulators underlying immune cell diversity and immune subtypes associated with prognosis in nasopharyngeal carcinoma. Cell Res. 2020;30:1024–42.
5. Kim H, Lee J, Kang K, Yoon S. MarkerCount: a stable, count-based cell type identifier for single-cell RNA-seq experiments. Computational Structural Biotechnology Journal. 2022;20:3120–32.
6. Shao X, Liao J, Lu X, Xue R, Ai N, Fan X: scCATCH: automatic annotation on cell types of clusters from single-cell RNA sequencing data. *Iscience* 2020, 23.
7. Choi J-H, In Kim H, Woo HG. scTyper: a comprehensive pipeline for the cell typing analysis of single-cell RNA-seq data. BMC Bioinformatics. 2020;21:1–8.
8. Kiselev VY, Yiu A, Hemberg M. scmap: projection of single-cell RNA-seq data across data sets. Nat Methods. 2018;15:359–62.
9. De Kanter JK, Lijnzaad P, Candelli T, Margaritis T, Holstege FC. CHETAH: a selective, hierarchical cell type identification method for single-cell RNA sequencing. Nucleic Acids Res. 2019;47:e95–e95.
10. Aran D, Looney AP, Liu L, Wu E, Fong V, Hsu A, Chak S, Naikawadi RP, Wolters PJ, Abate AR. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. Nat Immunol. 2019;20:163–72.
11. Tan Y, Cahan P. SingleCellNet: a computational tool to classify single cell RNA-Seq data across platforms and across species. Cell Syst. 2019;9:207-213. e202.
12. Alquicira-Hernandez J, Sathe A, Ji HP, Nguyen Q, Powell JE. scPred: accurate supervised method for cell-type classification from single-cell RNA-seq data. Genome Biol. 2019;20:1–17.
13. Chen J, Xu H, Tao W, Chen Z, Zhao Y. Han J-DJ: Transformer for one stop interpretable cell type annotation. Nat Commun. 2023;14:223.
14. Duan B, Zhu C, Chuai G, Tang C, Chen X, Chen S, Fu S, Li G, Liu Q. Learning for single-cell assignment. Sci Adv. 2020;6:eabd0855.
15. Wagner F, Yanai I: Moana: a robust and scalable cell type classification framework for single-cell RNA-Seq data. BioRxiv. 2018:456129.
16. Jia C, Hu Y, Kelly D, Kim J, Li M, Zhang NR. Accounting for technical noise in differential expression analysis of single-cell RNA sequencing data. Nucleic Acids Res. 2017;45:10978–88.
17. Haghverdi L, Lun AT, Morgan MD, Marioni JC. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Nat Biotechnol. 2018;36:421–7.
18. Song Q, Su J, Zhang W. scGCN is a graph convolutional networks algorithm for knowledge transfer in single cell omics. Nat Commun. 2021;12:3826.
19 Gao H, Zhang B, Liu L, Li S, Gao X, Yu B. A universal framework for single-cell multi-omics data integration with graph convolutional networks. Briefings in Bioinformatics. 2023;24:bbad081.
20. Wang H-Y, Zhao J-P, Su Y-S, Zheng C-H. scCDG: a method based on DAE and GCN for scRNA-seq data analysis. IEEE/ACM Transactions on Computational Biology Bioinformatics. 2021;19:3685–94.
21. Nakahama K-i. Cellular communications in bone homeostasis and repair. Cellular Molecular Life Sciences. 2010;67:4001–9.
22. Trupp M, Altman T, Fulcher CA, Caspi R, Krummenacker M, Paley S, Karp PD. Beyond the genome (BTG) is a (PGDB) pathway genome database: HumanCyc. Genome Biol. 2010;11:1–1.
23 Yamamoto S, Sakai N, Nakamura H, Fukagawa H, Fukuda K, Takagi T. INOH: ontology-based highly structured database of signal transduction pathways. Database. 2011;2011:bar052.

24. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 2017;45:D353–61.
25. Drevon C, Jaffredo T. Cell interactions and cell signaling during hematopoietic development. Exp Cell Res. 2014;329:200–6.
26. Zhuang C, Ma Q: Dual graph convolutional networks for graph-based semi-supervised classification. In *Proceedings of the 2018 world wide web conference*. 2018: 499–508.
27. Ding J, Adiconis X, Simmons SK, Kowalczyk MS, Hession CC, Marjanovic ND, Hughes TK, Wadsworth MH, Burks T, Nguyen LT. Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. Nat Biotechnol. 2020;38:737–46.
28. Baron M, Veres A, Wolock SL, Faust AL, Gaujoux R, Vetere A, Ryu JH, Wagner BK, Shen-Orr SS, Klein AM. A single-cell transcriptomic map of the human and mouse pancreas reveals inter-and intra-cell population structure. Cell Syst. 2016;3:346-360. e344.
29. Muraro MJ, Dharmadhikari G, Grün D, Groen N, Dielen T, Jansen E, Van Gurp L, Engelse MA, Carlotti F, De Koning EJ. A single-cell transcriptome atlas of the human pancreas. Cell Syst. 2016;3:385-394. e383.
30. Segerstolpe Å, Palasantza A, Eliasson P, Andersson E-M, Andréasson A-C, Sun X, Picelli S, Sabirsh A, Clausen M, Bjursell MK. Single-cell transcriptome profiling of human pancreatic islets in health and type 2 diabetes. Cell Metab. 2016;24:593–607.
31. Xin Y, Kim J, Okamoto H, Ni M, Wei Y, Adler C, Murphy AJ, Yancopoulos GD, Lin C, Gromada J. RNA sequencing of single human islet cells reveals type 2 diabetes genes. Cell Metab. 2016;24:608–15.
32. Li C, Liu B, Kang B, Liu Z, Liu Y, Chen C, Ren X, Zhang Z. SciBet as a portable and fast single cell type identifier. Nat Commun. 1818;2020:11.
33. Cui H, Wang C, Maan H, Pang K, Luo F, Duan N, Wang B: scGPT: toward building a foundation model for single-cell multi-omics using generative AI. Nature Methods. 2024:1–11.
34. Xu C, Lopez R, Mehlman E, Regier J, Jordan MI, Yosef N. Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. Mol Syst Biol. 2021;17:e9620.
35. Lotfollahi M, Rybakov S, Hrovatin K, Hediyeh-Zadeh S, Talavera-López C, Misharin AV, Theis FJ. Biologically informed deep learning to query gene programs in single-cell atlases. Nat Cell Biol. 2023;25:337–50.
36 Michielsen L, Lotfollahi M, Strobl D, Sikkema L, Reinders MJ, Theis FJ, Mahfouz A. Bioinformatics: Single-cell reference mapping to construct and extend cell-type hierarchies. NAR Genomics. 2023;5:lqad070.
37. De Donno C, Hediyeh-Zadeh S, Moinfar AA, Wagenstetter M, Zappia L, Lotfollahi M, Theis FJ. Population-level integration of single-cell datasets enables multi-scale analysis across samples. Nat Methods. 2023;20:1683–92.
38. Hao Y, Stuart T, Kowalski MH, Choudhary S, Hoffman P, Hartman A, Srivastava A, Molla G, Madad S, Fernandez-Granda C: Dictionary learning for integrative, multimodal and scalable single-cell analysis. Nature Biotechnology. 2023:1–12.
39. Yang F, Wang W, Wang F, Fang Y, Tang D, Huang J, Lu H, Yao J. scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. Nature Machine Intelligence. 2022;4:852–66.
40. Lin Y, Cao Y, Kim HJ, Salim A, Speed TP, Lin DM, Yang P, Yang JYH. scClassify: sample size estimation and multi-scale classification of cells using single and multiple reference. Mol Syst Biol. 2020;16:e9389.
41. Ianevski A, Giri AK, Aittokallio T. Fully-automated and ultra-fast cell-type identification using specific marker combinations from single-cell transcriptomic data. Nat Commun. 2022;13:1246.
42 Lyu P, Zhai Y, Li T, Qian J. Cell Ann: a comprehensive, super-fast, and user-friendly single-cell annotation web server. Bioinformatics. 2023;39:btad521.
43. Yuan H, Yan M, Zhang G, Liu W, Deng C, Liao G, Xu L, Luo T, Yan H, Long Z. CancerSEA: a cancer single-cell state atlas. Nucleic Acids Res. 2019;47:D900–8.
44. Tasic B, Yao Z, Graybuck LT, Smith KA, Nguyen TN, Bertagnolli D, Goldy J, Garren E, Economo MN, Viswanathan S. Shared and distinct transcriptomic cell types across neocortical areas. Nature. 2018;563:72–8.
45. Hodge RD, Bakken TE, Miller JA, Smith KA, Barkan ER, Graybuck LT, Close JL, Long B, Johansen N, Penn O. Conserved cell types with divergent features in human versus mouse cortex. Nature. 2019;573:61–8.
46. Bastidas-Ponce A, Tritschler S, Dony L, Scheibner K, Tarquis-Medina M, Salinno C, Schirge S, Burtscher I, Böttcher A, Theis FJ. Comprehensive single cell mRNA profiling reveals a detailed roadmap for pancreatic endocrinogenesis. Development. 2019;146:dev173849.
47. Alsaigh T, Evans D, Frankel D, Torkamani A. Decoding the transcriptome of calcified atherosclerotic plaque at single-cell resolution. Commun Biol. 2022;5:1084.
48. Chou C-H, Jain V, Gibson J, Attarian DE, Haraden CA, Yohn CB, Laberge R-M, Gregory S, Kraus VB. Synovial cell cross-talk with cartilage plays a major role in the pathogenesis of osteoarthritis. Sci Rep. 2020;10:10868.
49. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, Wilson CJ, Lehár J, Kryukov GV, Sonkin D. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature. 2012;483:603–7.
50. Iorio F, Knijnenburg TA, Vis DJ, Bignell GR, Menden MP, Schubert M, Aben N, Gonçalves E, Barthorpe S, Lightfoot H. A landscape of pharmacogenomic interactions in cancer. Cell. 2016;166:740–54.
51. Suphavilai C, Bertrand D, Nagarajan N. Predicting cancer drug response using a recommender system. Bioinformatics. 2018;34:3907–14.
52. Pellecchia S, Viscido G, Franchini M, Gambardella G. Predicting drug response from single-cell expression profiles of tumours. BMC Med. 2023;21:476.
53. Fustero-Torre C, Jiménez-Santos MJ, García-Martín S, Carretero-Puche C, García-Jimeno L, Ivanchuk V, Di Domenico T, Gómez-López G, Al-Shahrour F. Beyondcell: targeting cancer therapeutic heterogeneity in single-cell RNA-seq data. Genome Medicine. 2021;13:187.
54. Zheng Z, Chen J, Chen X, Huang L, Xie W, Lin Q, Li X, Wong KC. Enabling Single-Cell Drug Response Annotations from Bulk RNA-Seq Using SCAD. Adv Sci. 2023;10:2204113.
55. Srivatsan SR, McFaline-Figueroa JL, Ramani V, Saunders L, Cao J, Packer J, Pliner HA, Jackson DL, Daza RM, Christiansen L. Massively multiplex chemical transcriptomics at single-cell resolution. Science. 2020;367:45–51.

56. Folkerts J, Stadhouders R, Redegeld FA, Tam S-Y, Hendriks RW, Galli SJ, Maurer M. Effect of dietary fiber and metabolites on mast cell activation and mast cell-associated diseases. Front Immunol. 2018;9:380022.
57. Costello LC, Franklin RB, engineering t: A review of the important central role of altered citrate metabolism during the process of stem cell differentiation. J Regenerative Med. 2013;2.
58. Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P, Haw R, Jassal B, Korninger F, May B. The reactome pathway knowledgebase. Nucleic Acids Res. 2018;46:D649–55.
59. Slenter DN, Kutmon M, Hanspers K, Riutta A, Windsor J, Nunes N, Mélius J, Cirillo E, Coort SL, Digles D. WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. Nucleic Acids Res. 2018;46:D661–7.
60. Ji Z, Ji H. TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. Nucleic Acids Res. 2016;44:e117–e117.
61. Biase FH, Cao X, Zhong S. Cell fate inclination within 2-cell and 4-cell mouse embryos revealed by single-cell RNA sequencing. Genome Res. 2014;24:1787–96.
62. Efremova M, Vento-Tormo M, Teichmann SA, Vento-Tormo R. Cell PhoneDB: inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes. Nat Protoc. 2020;15:1484–506.
63. Abu-El-Haija S, Perozzi B, Kapoor A, Alipourfard N, Lerman K, Harutyunyan H, Ver Steeg G, Galstyan A: Mixhop: Higher-order graph convolutional architectures via sparsified neighborhood mixing. In international conference on machine learning. PMLR; 2019:21–29.
64. Aibar S, González-Blas CB, Moerman T, Huynh-Thu VA, Imrichova H, Hulselmans G, Rambow F, Marine J-C, Geurts P, Aerts J. SCENIC: single-cell regulatory network inference and clustering. Nat Methods. 2017;14:1083–6.
65. Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, Haibe-Kains B, Goldenberg A. Similarity network fusion for aggregating data types on a genomic scale. Nat Methods. 2014;11:333–7.
66. A single-cell transcriptomic map of the human and mouse pancreas reveals inter- and intra-cell population structure. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE84133. Accessed 10 Oct 2023.
67. A Single-Cell Transcriptome Atlas of the Human Pancreas. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE85241. Accessed 10 Oct 2023.
68. Single-cell RNA-seq analysis of human pancreas from healthy individuals and type 2 diabetes patients. https://www.ebi.ac.uk/biostudies/arrayexpress/studies/E-MTAB-5061. Accessed 10 Oct 2023.
69. RNA Sequencing of Single Human Islet Cells Reveals Type 2 Diabetes Genes. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE81608. Accessed 10 Oct 2023.
70. Systematic comparative analysis of single cell RNA-sequencing methods. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE132044. Accessed 10 Oct 2023.
71. Shared and distinct transcriptomic cell types across neocortical areas. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE115746. Accessed 10 Oct 2023.
72. NIMH Human Middle Temporal Gyrus (MTG) Cell Types. https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001790.v2.p1. Accessed 10 Oct 2023.
73. Comprehensive single-cell mRNA profiling reveals a detailed roadmap for pancreatic endocrinogenesis. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE132188. Accessed 10 Oct 2023.
74. Decoding the transcriptome of calcified atherosclerotic plaque at single-cell resolution. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE159677. Accessed 10 Oct 2023.
75. Synovial Cell Cross-talk with Cartilage Plays a Major Role in the Pathogenesis of Osteoarthritis. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE152805. Accessed 10 Oct 2023.
76. CSB5/CaDRReS-Sc. https://github.com/CSB5/CaDRReS-Sc. Accessed 10 Oct 2023.
77. Lotfollahi M, Klimovskaia Susmelj A, De Donno C, Hetzel L, Ji Y, Ibarra IL, Srivatsan SR, Naghipourfar M, Daza RM, Martin B. Predicting cellular responses to complex perturbations in high-throughput screens. Mol Syst Biol. 2023;19:e11517.
78. Cao XY, Huang YA, You ZH, Shang XQ, Hu L, Hu PW, Huang ZA.scPriorGraph: Constructing Biosemantic Cell-Cell Graphs with Prior Gene Set Selection for Cell Type Identification from scRNA-seq Data.Github. https://github.com/ChrisOliver2345/scPriorGraph(2024).
79. Cao XY, Huang YA, You ZH, Shang XQ, Hu L, Hu PW, Huang ZA.scPriorGraph: Constructing Biosemantic Cell-Cell Graphs with Prior Gene Set Selection for Cell Type Identification from scRNA-seq Data.Zenodo. 10.5281/zenodo.10981088(2024).

Publisher's Note