Briefings in Bioinformatics, 2026, 27, bbag259
https://doi.org/10.1093/bib/bbag259
Published: 27 May 2026
Problem Solving Protocol

OXFORD

PVAED: prior-guided variational autoencoders with diffusion denoising for interpretable single-cell representation learning

Yawei Niu, Yichu Chen, Wenji Ma*

Center for Single-Cell Omics, School of Public Health, Shanghai Jiao Tong University School of Medicine, Shanghai, China
*Corresponding author. Center for Single-Cell Omics, School of Public Health, Shanghai Jiao Tong University School of Medicine, Shanghai, China.
E-mail: wenjimacs@sjtu.edu.cn

Abstract

Single-cell RNA sequencing data rely heavily on dimensionality reduction methods to uncover underlying structures and patterns. Nonlinear dimensionality-reduction methods such as uniform manifold approximation and t-distributed stochastic neighbor embedding have been routinely used for this task. However, gene expression data alone often fails to capture and identify changes in cellular pathways, protein complexes, and TF-targets, which are more enlightening at the regulation level. To address this limitation, we present PVAED, a dimensionality reduction framework that integrates biological prior knowledge into a variational autoencoder (VAE) and further refines its latent embeddings with a diffusion-based denoising module. In addition, we incorporate a neighborhood-preserving loss term to ensure local similarity among cells in the reduced space. Across multiple benchmarks, PVAED achieves an average 43% improvement in low-dimensional cell representation compared to standard VAEs (evaluated across seven clustering metrics) and delivers a 34% (44%) gain in projection quality relative to classical or VAE-based approaches in terms of global (local) structure preservation. Beyond performance, PVAED offers improved interpretability through its prior-based encoder design, enabling prioritization and modeling of interpretable biological variables as hypergraphs. This framework facilitates the identification of critical regulatory factors associated with disease pathology and could also be utilized to predict potentially related genes for known pathways. Finally, we demonstrate the utility of PVAED by revealing pre-committed neuronal subpopulations in the differentiation of migrating neurons within the mammalian cerebral cortex.

Keywords scRNA-seq, dimension reduction, prior knowledge, VAE model, diffusion model

Introduction

The advent of single-cell RNA sequencing (scRNA-seq) has fundamentally transformed our capacity to characterize cellular states, offering unprecedented resolution to dissect the heterogeneity inherent in biological systems [1]. This technology provides a systematic and precise window into organismal development and function [2, 3], generating vast datasets that demand sophisticated computational frameworks for interpretation. Among these, deep generative models, particularly variational autoencoders (VAEs), have emerged as powerful tools for analyzing complex transcriptomic patterns [4]. Notable implementations such as scVI [5], scGen [6], LDVAE [7], and scVAE [8] leverage the flexible architectural design of VAEs to address core challenges in scRNA-seq analysis, including dimensionality reduction [9, 10], clustering [11–13], and the identification of rare cell types [14]. Despite their impressive performance in dedicated modeling tasks, canonical VAEs often function as "black boxes," lacking interpretability and failing to provide biologically meaningful latent representations. For instance, while scGen can extract latent perturbation vectors, these cannot be directly mapped to variations in specific gene modules or regulatory programs.

A central challenge in transcriptomic analysis lies in balancing the capacity to handle noisy, high-dimensional molecular measurements with the imperative to yield actionable biological insights. To truly understand the mechanisms driving cellular state changes, it is essential to move beyond correlative patterns and quantify the underlying biological processes. Cellular pathways exemplify such mechanisms; their structure is largely conserved, but their dynamic activity dictates cellular phenotype and behavior. Consequently, integrating prior knowledge of pathways and other molecular interactions has become a promising strategy to enhance model interpretability. Several models have pioneered this integration. DCell [15] employs a deep neural network structured around hierarchical Gene Ontology (GO) [16], annotations, guiding supervised learning for genotype-to-phenotype predictions. Its interpretability stems from analyzing the activation of neurons representing specific cellular subsystems. However, this approach is confined to supervised settings. The scLVM

Received: December 19, 2025. **Revised:** February 3, 2026. **Accepted:** April 30, 2026

model [17] uses a Bayesian framework to incorporate prior knowledge of cell-cycle stages from GO, enabling the discovery of hidden cell subpopulations. While interpretable, its computational cost and lack of out-of-sample inference limit its broad application. More recently, the Multilevel-GNN model [18] hierarchically integrated multi-omics data with biological priors like gene regulatory networks, using explanation algorithms to identify critical regulatory factors in disease. In the VAE domain, VEGA [19] integrates pathway information from GO and Reactome [20] into its latent space and decoder construction. Yet, its sparse, single-layer decoder may constrain generalizability and robustness. Similarly, siVAE [21] enhances interpretability for downstream tasks via a specialized regularization term but falls short in providing meaningful biological insights during the critical stage of embedding visualization.

Concurrently, latent diffusion models (LDMs) [22] have demonstrated state-of-the-art performance in generative tasks across image, audio, and video domains [23]. Compared to VAEs, LDMs offer more stable training dynamics and superior conditional generation capabilities [24]. Their potential is now being explored in genomics. For example, scDiffusion [25] combines diffusion models with foundation models to generate high-quality scRNA-seq data under controlled conditions, simulating continuous developmental trajectories. ChromoGen [26] employs a diffusion approach to predict single-cell 3D chromatin conformations, and stDiff [27] imputes spatial transcriptomics data by leveraging scRNA-seq relationships during its denoising process.

Despite these parallel advances, a synergistic integration of VAEs and diffusion models for single-cell analysis remains largely unexplored. Such a combination holds the promise of dual benefits: the robust latent representation learning of VAEs coupled with the powerful denoising and generative fidelity of diffusion models. Our work addresses this gap. We propose a framework where a pretrained VAE learns unified representations of gene expression data, establishing a bridge between the original and latent spaces. A diffusion model then acts as a denoising module to refine these latent embeddings, simultaneously enhancing data quality for generation and producing superior low-dimensional representations for downstream analytical tasks like clustering and visualization.

However, purely data-driven architectures, even sophisticated ones, often overlook the rich tapestry of biological knowledge accumulated over decades. To overcome these limitations, we introduce PVAED, a prior-informed interpretable VAE model integrated with a diffusion model for effective single-cell transcriptome dimensionality reduction and analysis. Unlike standard black-box models, PVAED incorporates curated biological knowledge, including pathways from Kyoto Encyclopedia of Genes and Genomes (KEGG) database [28], protein complexes from Comprehensive Resource of Mammalian protein complexes (CORUM) database [29], transcription factor-target interactions from hTFtarget [30], and gene interactions from Reactome [20]. This is achieved by structurally embedding pathways, protein complexes, and transcription factors as neuron nodes in both the encoder and decoder, resulting in a biologically constrained, sparse architecture. To further enhance representation quality, the latent space of the VAE is refined via a diffusion model. Our training regimen involves an initial separate pretraining phase on large-scale datasets to learn complex gene interactions, akin to approaches in scGPT [31], scFoundation [32], and Geneformer [33], followed by a joint fine-tuning of the VAE and diffusion models on specific tissue data.

Crucially, the interpretability of PVAED is achieved by analyzing the trained weights of the prior-based neuron nodes, which we model as a hypergraph—a suitable formalism for characterizing complex prior relationships. This allows us to identify critical regulatory factors that shape the latent space and reveal their changes during pathological progression, as we demonstrate in lung cancer and Alzheimer's disease. Furthermore, we utilize PVAED's latent representations to uncover pre-committed neuronal subpopulations during mammalian cerebral cortex development, showcasing its broad applicability for delivering mechanistic insights from genomic data.

Materials and methods

The architectural design of PVAED is guided by three fundamental principles that address these challenges: (i) systematic integration of curated biological prior knowledge, (ii) leveraging the data compression capabilities of variational autoencoder frameworks, and (iii) exploiting the denoising properties of diffusion models. This tripartite framework enables PVAED to simultaneously accomplish dimensionality reduction of single-cell transcriptomic data, provide interpretable identification of key regulatory factors, and facilitate discovery of novel cellular subpopulations. Specifically, we incorporate structured biological prior knowledge into the VAE architecture to generate biologically informed cell embeddings while establishing a principled approach for model interpretability in deep learning frameworks. The diffusion model component subsequently refines these embeddings through iterative denoising, yielding robust low-dimensional representations optimized for downstream analytical tasks and data visualization. Next, we introduce the methods corresponding to the design of PVAED.

Prior knowledge informed architectural design of PVAED

PVAED integrates prior knowledge of gene–gene interactions, transcription factor regulons, and protein complexes directly into its VAE through a dedicated prior encoder module. This ensures that the derived low-dimensional cell embeddings are biologically grounded. The PVAED framework operates in two stages (Fig. 1). First, in a pretraining phase, the VAE and a denoising diffusion probabilistic model (DDPM) are trained simultaneously on multiple single-cell datasets to learn generalizable features. The VAE encodes cells into a latent space, which the DDPM refines through a forward noising and reverse denoising process to produce robust representations. Second, in a joint fine-tuning phase, the entire model is trained end-to-end on a specific task, adapting the pretrained parameters to particular biological questions.

The refined embeddings generated by PVAED serve as powerful features for diverse downstream analyses (Fig. 1). These include cell type annotation, visualization, and clustering, where the denoised representations yield clearer separation of cell states. Furthermore, PVAED facilitates the identification of key regulatory drivers in comparative tasks, such as contrasting diseased and normal tissues. A key advantage of our model is its interpretability: the trained prior encoder allows top-ranked prior nodes to be modeled as a hypergraph, revealing the underlying biological programs. Finally, PVAED's low-dimensional projections are highly effective for identifying subtle cell subpopulations, the distinct identities of which can be validated by differential gene expression analysis.

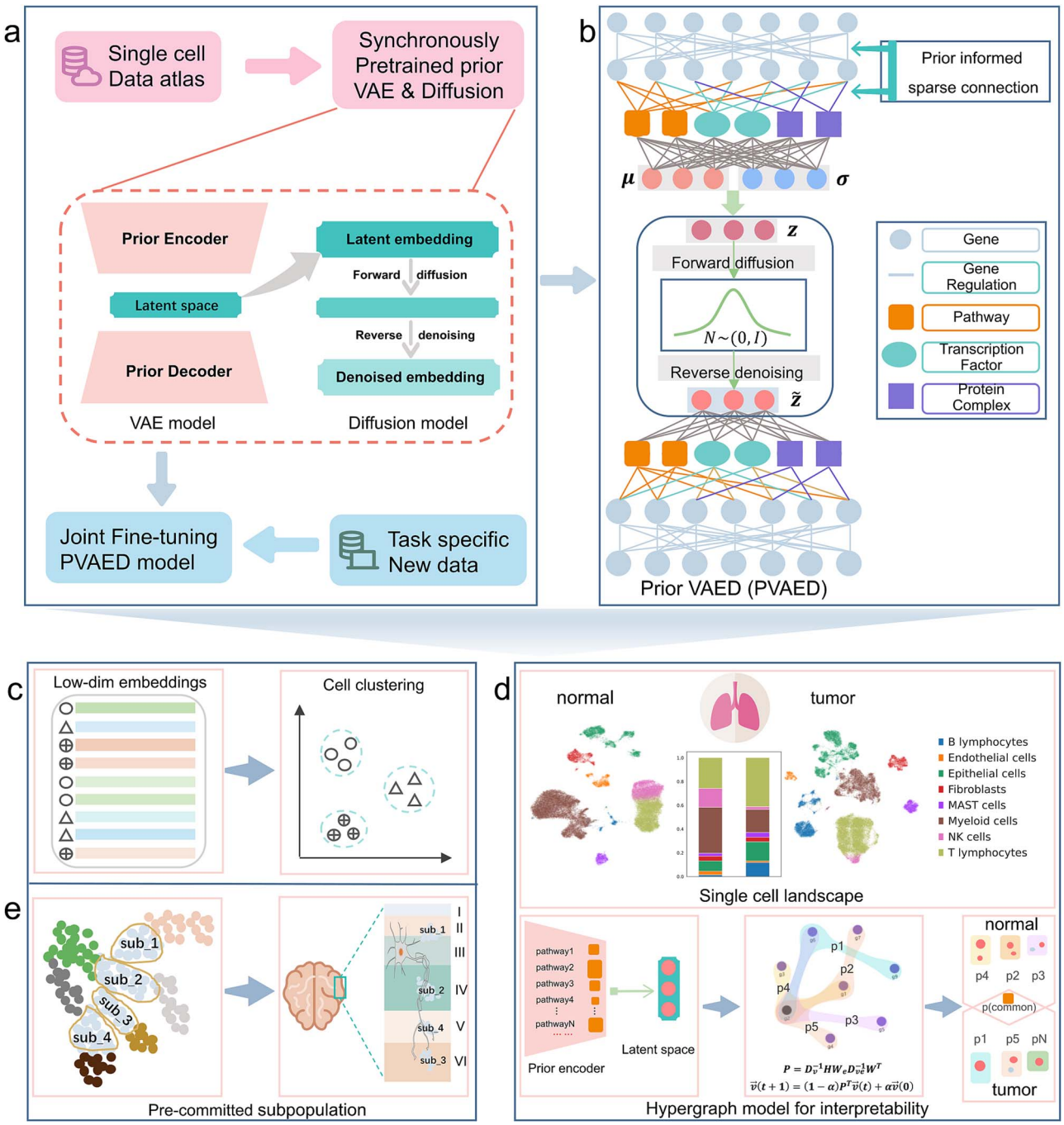


Figure 1 PVAED model schematic. (a) The basic framework of PVAED model comprising a VAE model and a diffusion model. After pretraining, the model undergoes joint fine-tuning with task-specific new data. (b) The detailed construction of PVAED. Prior information regarding pathways, TF-targets, and protein complexes is integrated by modeling them as nodes within the encoder layer. Prior information of pathways, TF-targets and protein complexes are integrated as nodes in encoder layers. The connections between two encoder layers are sparsely connected informed by the gene–gene regulation or gene-prior object constituent relation. (c) Low-dim cell embeddings from PVAED could be used for cell clustering task. (d) PVAED could be conducted for pathology-related comparison task and then for interpretability task based on the trained prior encoder by hypergraph model. (e) PVAED could be instrumental in cell subpopulation discovery, uncovering subtle, pre-committed cellular subgroups.

Prior variational autoencoders

In PVAED model, prior VAE is designed extending the canonical VAE framework by integrating biological prior information into the fundamental structure. We briefly review canonical VAE in S1 part of Supplementary materials. Here, we focus on how we integrate priors in the VAE structure.

The encoder and decoder structures in canonical VAEs typically consist of fully connected neural networks. To leverage well-established biological priors such as pathway memberships, protein complexes, transcription factor-target relationships, and gene interactions, we re-engineer the encoder and decoder architectures by incorporating structured sparse connections. Specifically, we add a

gene layer and a prior layer as hidden layers following the input gene expression layer, where neuronal nodes in the gene layer represent individual genes and neuronal nodes in the prior layer represent pathways, complexes, and transcription factors. Nodes in the gene layer are connected to nodes in the input gene expression layer according to gene relationships from Reactome [20]. Meanwhile, nodes in the prior layer are explicitly connected to nodes in the gene layer based on curated prior databases.

The input of input layer is gene expression data $X \in R^{n\times m}$, n is the cell number and m is the gene/neuron node number, each neuron in this layer represents a gene's normalized expression.

The second layer is gene layer with the same number of neurons as the input layer, the node in this layer connects with the node in the input layer if they are related in Reactome database. The propagation from the input layer to this layer is formed as:

$$H_{gene} = \sigma\left(X \bullet (I + A) \bullet W_{gene}\right) \tag{1}$$

where $A \in (0,1)^{m\times m}$ is the adjacency matrix of the input layer and the gene layer, where $A_{ij} = 1$ if gene node i in the input layer is interacted with gene node j in gene layer. W_{gene} is a trainable weight matrix with the subscript gene meaning current gene layer.

The third layer is prior layer composed of pathway nodes $K_{pathway}$, complex nodes $K_{complex}$ and TF nodes K_{TF}. The node in this layer connects with the node in the gene layer if the gene node is in the prior item. Take pathway item for example, gene node i connects to pathway node j if $i \in$ pathway j. The incidence matrix of this layer is $M \in (0,1)^{G\times K}$, where $K = K_{pathway} + K_{complex} + K_{TF}$ and the activation is formed as:

$$H_{prior} = \sigma\left(H_{gene} \bullet M \bullet W_{prior}\right) \tag{2}$$

W_{prior} is a trainable weight matrix with the subscript prior meaning current prior layer. Specifically, $K_{pathway}$ nodes represent individual pathways obtained from the KEGG collections. $K_{complex}$ denotes protein complex nodes curated from the CORUM database. K_{TF} nodes denote transcription factors derived from hTFtarget database. Each node is connected to its member genes in the second (gene) layer. This design enables the model to integrate pathway, protein complex, and TF level regulatory information in a structured and interpretable manner.

The latent layer μ and σ are the same as canonical VAEs and has h neurons respectively, h is the dimension of latent space. The nodes in latent layer are fully connected with prior layer. The sample layer z has h neurons sampling from the latent layer μ and σ:

$$z \sim \mathcal{N}\left(\mu\left(X, H_{prior}\right), \sigma\left(\mu\left(X, H_{prior}\right)\right)\right) \tag{3}$$

Overall, the input layer of the encoder receives the gene expression data. The second layer (gene layer) has the same number of nodes as the input layer, and each node is connected to its upstream gene only if a regulatory relationship is supported by curated databases. The third layer (prior layer) integrates pathway, protein complex, and transcription factor nodes, allowing both expression-driven and biologically constrained representations. The other layers construct the decoder part with same but reversed structure as above introduced.

Denoising diffusion probabilistic models

DDPM [34] is a latent-variable model consisting of a forward noising process that gradually destroys the structure of the original data and a reverse denoising process that learns to recover the original data from the noisy input. The detailed description is shown in the S2 part of Supplementary materials. In our work, DDPM is utilized to further denoise the latent embeddings from previous VAE latent space in the denoising process, resulting in refined embeddings for downstream analysis tasks. Meanwhile, we explore the optimal choice of back sample step in the denoising process on all the involved datasets as shown in Figure S1–S2.

PVAED loss function

The standard VAE loss function consists of two primary components: a reconstruction loss that ensures accurate data generation, and a Kullback–Leibler (KL) divergence term that regularizes the latent space to approximate Gaussian distribution. However, this formulation does not explicitly constrain the preservation of local geometric structure in the latent space, which may result in distorted latent representations where semantically similar inputs are not mapped to nearby points, thereby undermining the interpretability and utility of the learned embeddings. Trustworthiness [35] is a quantitative indicator in evaluating how well the structure preservation is in projection space. Consequently, we enhance VAEs with trustworthiness metric that could guide and constrain the preservation of local similarities during the encoding process by integrating trustworthiness into the VAE loss function with a weight factor λ as follows:

$$\widetilde{\mathcal{L}}_{VAE}(\boldsymbol{x};\theta,\phi) = \left\|\hat{\boldsymbol{x}} - \boldsymbol{x}\right\|^2 + \beta \bullet \mathbb{KL}\left(q(\boldsymbol{z}|\boldsymbol{x};\theta) \mid |p(\boldsymbol{z})\right) - \lambda T(k) \tag{4}$$

$$T(k) = 1 - \frac{2}{N \bullet k\left(2N - 3k - 1\right)} \sum_{i=1}^{N} \sum_{j\in U_k(i)} \left(r\left(i,j\right) - k\right) \tag{5}$$

where $\left\|\hat{\boldsymbol{x}} - \boldsymbol{x}\right\|^2$ is the reconstruction loss between the original cell profile $\boldsymbol{x}$ and restructured profile $\hat{\boldsymbol{x}}$, $\mathbb{KL}\left(q(\boldsymbol{z}|\boldsymbol{x};\theta) \mid |p(\boldsymbol{z})\right)$ is the KL divergence, which measures the discrepancy between the approximate posterior distribution $q(\boldsymbol{z}|\boldsymbol{x};\theta)$ and the prior distribution $p(\boldsymbol{z})$, acting as a regularization component that encourages the learned latent distribution to remain close to the chosen prior, typically a standard normal distribution$\mathcal{N}(0, I)$. N is the total number of single cells, k is the top k considered closest neighbors for a cell, $U_k(i)$ is the set of cells that are among the k nearest neighbors of target cell i in the reduced space but not in the original space, and $r\left(i,j\right)$ is the rank of cell j in the original space's neighborhood of cell i. $2/N \bullet k\left(2N - 3k - 1\right)$ scales the trustworthiness values between zero and one, where one indicates perfect preservation of local neighborhoods and zero indicates that neighborhoods are completely disrupted.

After training the prior VAEs, we obtain the original cell embedding $\boldsymbol{x}_0$ as the start of DDPM, then $\boldsymbol{x}_0$ becomes a noised embedding $\boldsymbol{x}_T$ after the forward process. Finally, the loss of PVAED becomes:

$$\mathcal{L}_{scPVAED} = \widetilde{\mathcal{L}}_{VAE}(\boldsymbol{x};\theta,\phi) + \mathcal{L}_{DDPM} \#(6) \tag{6}$$

$\mathcal{L}_{DDPM}$ is described in the S2 part of Supplementary materials.

Random walk on PVAED encoder-based hypergraph

Generally, a hypergraph is a generalization of a graph in which an edge, termed a hyperedge, can connect any number of vertices (two or more) representing the collective relationship or group interaction between a group of nodes. In this work, in order to explore multi-way relationships among gene sets, we construct a hypergraph model, where prior neurons like pathways are represented as hyperedges

and genes associated with pathways are represented as nodes. More specifically, after model training, each prior node (representing biological entities such as pathways, transcription factors, or protein complexes) acquires an optimized weight parameter, reflecting its regulatory contribution to latent representation learning. In the case of pathways, we denote these trained weights as $w_{pathway}$. The interpretability analysis proceeds as follows:

Firstly, we extract the matrix of learned prior weights $w_{pathway}$ from the encoder and pick the top K ranked pathways.

Then, using the curated incidence matrix $M \in \{0,1\}^{K\times K}$, which results from the gene membership of the corresponding pathway items. Pathways are connected if they share at least one same gene.

Finally, a hypergraph is constructed based on M, in which prior nodes act as hyperedges and their associated genes act as vertices. On this hypergraph, we perform random-walk and spectral clustering analysis to identify key regulatory hubs and to predict potential gene associations contributing to disease-relevant processes.

Detailed description is in part S3 of Supplementary materials.

Clustering metrics

In this work, in order to explore the quality of dimensionality reduction, we cluster the data using the latent embeddings of PVAED and make use of some classic metrics, including Adjusted Rand Index (ARI), Mutual Information (MI), Normalized Mutual Information (NMI), Adjusted Mutual Information (AMI), V-measure, Completeness, and Homogeneity. Briefly, ARI assesses the overall agreement in pairwise sample assignments between two clustering results. MI quantifies the mutual dependency between the two clusterings, measuring how much knowing one reduces uncertainty about the other. NMI normalizes MI to a 0–1 scale, making it comparable across different datasets and numbers of clusters. AMI is a chance-corrected version of MI. The V-measure is the harmonic mean of two complementary metrics: Homogeneity, which requires that each cluster contains only members of a single class, and Completeness, which requires that all members of a given class are assigned to the same cluster. The detailed formulations are presented in part S4 of Supplementary materials.

Projection quality metrics

Shepard goodness (SG) [36] and mean relative rank error (MRRE) [37] metrics are employed to assess the reliability of dimensionality reduction methods in preserving the global and local structures of high-dimensional data, respectively. Shepard goodness is a nonparametric measure that quantifies how well pairwise distances among data points are preserved when mapped from the original high-dimensional space to a lower-dimensional embedding, and a higher SG value indicates a more faithful global projection. In contrast, MRRE captures distortions in neighborhood ranking that occur during the projection process, where a lower MRRE value reflects better preservation of local neighborhood relationships. Accordingly, Shepard goodness can be regarded as a global metric evaluating overall geometric consistency, whereas MRRE serves as a local metric quantifying neighborhood retention. These two complementary metrics are particularly useful in scRNA-seq analyses to ensure that biologically meaningful similarities, such as shared cell-type identities or gene co-expression patterns, are maintained after dimensionality reduction, which is essential for accurate downstream interpretation. Unlike adversarial or contrastive learning approaches that depend on complex auxiliary networks or carefully curated positive and negative sample pairs, SG and MRRE provide a computationally efficient and theoretically grounded means to evaluate and enforce both global and local structural consistency. The detailed description is in part S5 of Supplementary materials.

Results

PVAED achieves superior dimensionality reduction through integration of biological priors and diffusion-based denoising

The accurate recovery of intrinsic biological features from single-cell RNA sequencing data is essential for robust downstream interpretation. We first establish the model's performance on the well-known PBMC 3k dataset, where PVAED generates embeddings with enhanced population separation. The model clearly delineates transcriptionally similar subsets, such as FCGR3A$^+$ and CD14$^+$ monocytes, and formed a compact, distinct B cell cluster, demonstrating a fine-grained resolution superior to standard UMAP and naïve VAE (Fig. 2a). We further demonstrate the specific advantage of the diffusion process in preserving complex local structures by extending our visualization analysis to the large-scale GSE204684 dataset [38]. As shown in Fig. 2b, the diffusion module significantly enhances the representation of continuous subtypes: whereas PVAE fragments IN-fetal and OPC cells into disjointed clusters, PVAED preserves their local cohesion and better maintains biological continuity.

Incorporating a trustworthiness loss for local structure preservation improves clustering metrics by 35.24% over the baseline VAE, outperforming a graph-based L2 norm loss (27.41% improvement), and is thus adopted in PVAED (Fig. 2c). Systematic benchmarking across seven clustering metrics shows that integrating biological priors in PVAE yields a 32.27% average improvement over naïve VAE. The ablation experiment on DDPM in PVAED model shows that adding the diffusion module in PVAED further increases performance, achieving a 43.16% overall gain over naïve VAE and an 8.39% improvement over PVAE (Fig. 2d).

Ablation studies on prior sources reveals that integrated multi-prior modeling performed best. Among individual priors, pathway information contributes most (35.05% average improvement), followed by transcription factor targets (32.59%) and protein complexes (32.34%), underscoring the value of diverse biological knowledge (Fig. 2e).

Comprehensive benchmarking of clustering metrics across seven datasets shows that PVAED achieves the highest average performance across all clustering measures (∼11% improvement over standard UMAP, ∼15% improvement over siVAE, and ∼7% improvement over VEGA, Fig. 2f). Furthermore, PVAED outperforms established methods across seven datasets in projection quality on 2-dim projection, improving average Shepard goodness by 39.89% over standard UMAP, 30.99% over siVAE, and 30.45% over VEGA, and enhancing MRRE by 47.62%, 55.26%, and 30.52%, respectively (Fig. 2g and h). Additionally, we provide the projection quality on latent embeddings of all methods in Table S1 in Supplementary materials, where PVAED also shows best performance. These results confirm PVAED as a robust framework that effectively leverages biological knowledge and diffusion-based refinement for interpretable single-cell analysis.

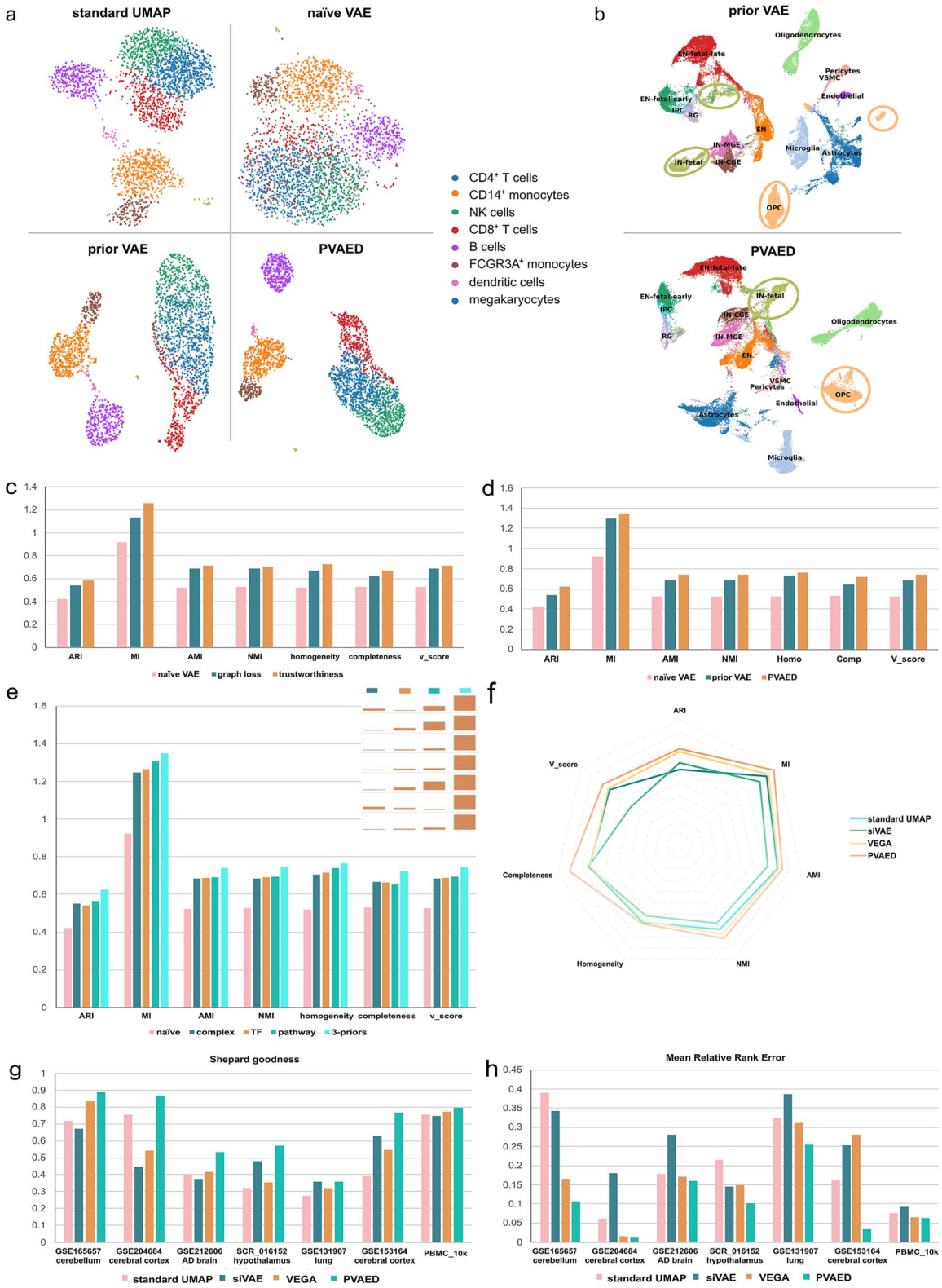


Figure 2 Evaluation and benchmark of PVAED. (a) The visualization of different methods on PBMC 3k data, including standard UMAP, naïve VAE, PVAE, and PVAED. (b) The visualization of PVAE and PVAED on large-scale dataset of GSE204684 about developing cerebral cortex to reveal the visualization contribution of the diffusion module. PVAED preserves better maintained local structure of EN-fetal and OPC cell types. (c) Incorporation of trustworthiness loss term yields better clustering results than naïve VAE and graph loss term. (d) Ablation study underscores the synergistic effect of prior integration and diffusion. (e) Ablation study demonstrates that the combination of all three kinds of prior information leads to the best overall performance. (f) The benchmark of PVAED with standard UMAP and the other two existing VAE-like models on clustering metrics across seven different datasets. (g, h) The benchmark results on projection metrics of Shepard goodness and MRRE across all datasets.

All experiments were performed on an NVIDIA H100 GPU with 80 GB of HBM3 memory. The PVAED model was optimized using a mean squared error (MSE) reconstruction loss, with the weighting factor of trustworthiness term λ set to 1. In the fine-tuning process, we use the intersection of task-specific data and pretrained large dataset. For the single-cell RNA-seq dataset comprising 10 804 cells and 11 160 genes, the model was trained with a 100-dimensional latent space. The pretraining phase, conducted for 80 epochs, required ∼585 seconds including 30 s diffusion time. This was followed by a fine-tuning phase of 60 epochs, which was completed in ∼400 seconds including 21 s diffusion time. When the genes were filtered to 5000, the pretraining and fine-tuning time was reduced to 350 s and 270 s, respectively. The diffusion module operates only in the 100-dimensional VAE latent space, each diffusion epoch requires just 0.3–0.4 seconds, and its cost is independent of input size, which ensures the scalability of PVAED for large atlases (Fig. S4 in Supplementary materials).

Interpretable pathway-centric master regulator analysis in lung adenocarcinoma reveals critical gene sets via hypergraph inference

Lung adenocarcinoma (LUAD), the most common subtype of non-small cell lung cancer (NSCLC), accounts for ∼40% of all lung cancer cases [39]. To decipher the dysregulated transcriptional programs in LUAD, we apply the PVAED framework for master regulator analysis and critical gene set prediction. The model is firstly pretrained on the Human Lung Cell Atlas (HLCA), then fine-tuned using the GSE131907 dataset [40], which includes 208 506 cells from normal and tumor lung tissues of 44 patients.

Visualization of the learned embeddings reveals distinct cellular landscapes in normal versus tumor samples (Fig. 3a). PVAED demonstrates robust performance across a suite of clustering metrics (including ARI and NMI) in both normal (Fig. 3b) and tumor (Fig. 3c) subsets. Furthermore, when compared with the reference projections from the original study (Kim et al.), PVAED yields superior embedding fidelity, achieving higher scores on projection quality metrics such as SG and MRRE (Fig. 3d).

Leveraging PVAED's interpretable prior layer, we analyze pathway importance shifts between conditions. Tumor samples are dominated by oncogenic pathways ("Pathways in cancer", "MAPK", "Endocytosis", "PI3K-Akt") and immune activation signals ("Cytokine-cytokine receptor interaction", "TNF signaling"), while normal tissue emphasized homeostasis pathways ("Focal adhesion", "Neuroactive ligand–receptor interaction") (Fig. 3e). Additionally, ranking genes by their influence in PVAED's regulation layer identified the top 50 tumor-associated genes (Fig. 3f). Functional enrichment confirms their significant roles in cancer-relevant pathways, validating the model's capacity to pinpoint critical regulatory elements in LUAD (Fig. 3g).

Based on the prioritized pathways in normal and tumor samples, we construct a hypergraph model to characterize pathway-gene relationships (Fig. 3h). Random Walk with Restart (RWR) analysis on this hypergraph prioritizes key regulatory genes in the tumor context (Fig. 3i), with 4 out of the top 5 candidates (e.g. *RPA3*, *DHAH3*) being literature-confirmed in lung cancer (Fig. 3j, left). Spectral clustering [41] identifies a densely connected 8-gene community representing core components of the *MAPK/ERK* and *PI3K-Akt* oncogenic pathways (Fig. 3j, right). The GO Biological Process enrichment analysis of these eight identified genes reveals four lung cancer-related processes (Fig. 3k), among which EGFR shows the most significant enrichment and is well known to be associated with lung tumorigenesis. Furthermore, RWR successfully predicted novel gene associations for the HIF-1 signaling pathway, including the literature-confirmed *SOCS* gene family [42] (Fig. 3l).

To quantitatively assess the reproducibility and stability on biological insights, we trained PVAED with four random seeds (0, 1, 10, 1024) on the GSE131907 dataset and compared the resulting pathway ranking vectors across runs. The Kendall's τ rank correlations among all pairs of runs averaged 0.95 ± 0.03, with $P < 1e\text{-}129$, demonstrating highly consistent ranking orders (Fig. S5 in Supplementary materials). Moreover, the top 50 pathways were nearly identical across runs, showing 98% overlap (49 of 50 pathways shared among all trials). These results confirm that PVAED's interpretability outputs are robust to random initialization, and that the biological conclusions derived from the ranked pathways are stable and reproducible.

Application of PVAED interpretability for critical pathway and gene identification in Alzheimer's disease

Alzheimer's disease (AD) is a progressive neurodegenerative disorder presenting with both amnestic and non-amnestic cognitive deficits. While single-cell transcriptomics has opened new avenues for dissecting cell-type-specific dynamics in the brain, many previous studies have been constrained by limited sample sizes and technical noise, often missing rare AD-associated cell states and isoform-level expression changes. Here, we apply the PVAED model to the GSE212606 dataset [44], comprising 118 240 single-cell transcriptomes from AD and wild-type (WT) human brain samples, to identify pathology-related regulatory programs and candidate genes.

We first visualize cellular organization in WT and AD brains using UMAP embeddings derived from PVAED (Fig. 4a). PVAED generates informative embeddings as quantified by seven clustering metrics (Fig. 4b and c). PVAED also achieved better SG score and better MRRE score than Sziraki *et al.*, the authors who released the dataset [44] (Fig. 4d).

Pathway analysis reveals that PVAED's pathway prioritization is not biased by gene set size ($R = 0.074$, $P = .23$), confirming the biological relevance of identified pathways (Fig. 4f). AD samples specifically show elevated importance of "Polycomb repressive complex" and "mRNA surveillance pathway" both mechanistically linked to AD pathobiology (Fig. 4e).

From PVAED's gene-regulation layer, we extract top-weighted genes, all 10 of which are validated as AD-relevant by ssREAD database [45] (Fig. 4g). Notably, only 4 of these 10 key regulatory genes show significant differential expression in conventional analysis (Fig. 4h), demonstrating PVAED's unique ability to identify functionally critical elements that escape detection by standard differential expression methods.

Biologically informed visualization of cortical development benefits pre-committed subpopulation discovery

The developing mammalian cerebral cortex exhibits precisely orchestrated cellular differentiation and diversification, providing an ideal model system for evaluating the PVAED framework's ability to capture intrinsic topological structures in single-cell data. We apply PVAED to the GSE153164 dataset [46], which comprehensively profiles all

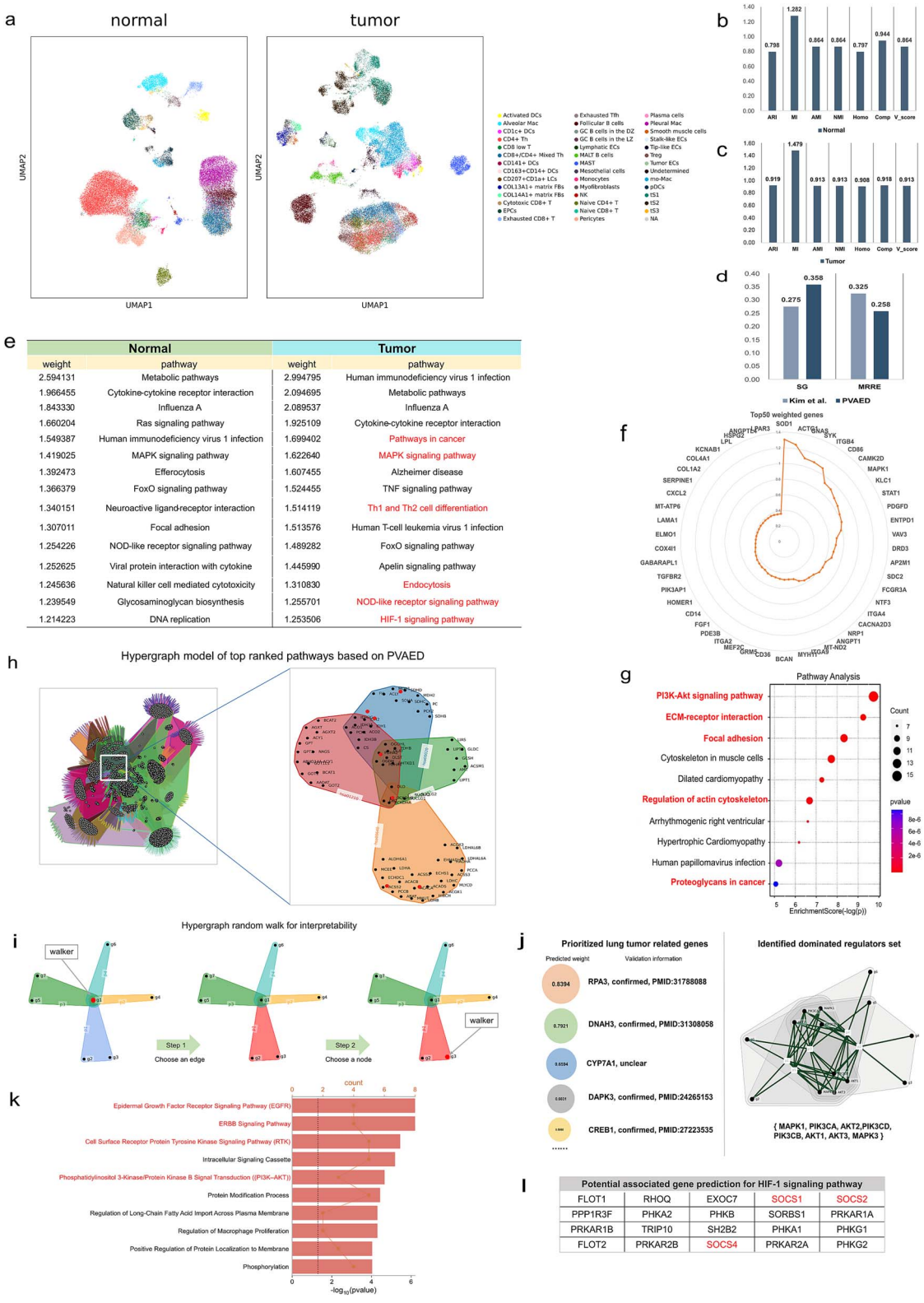
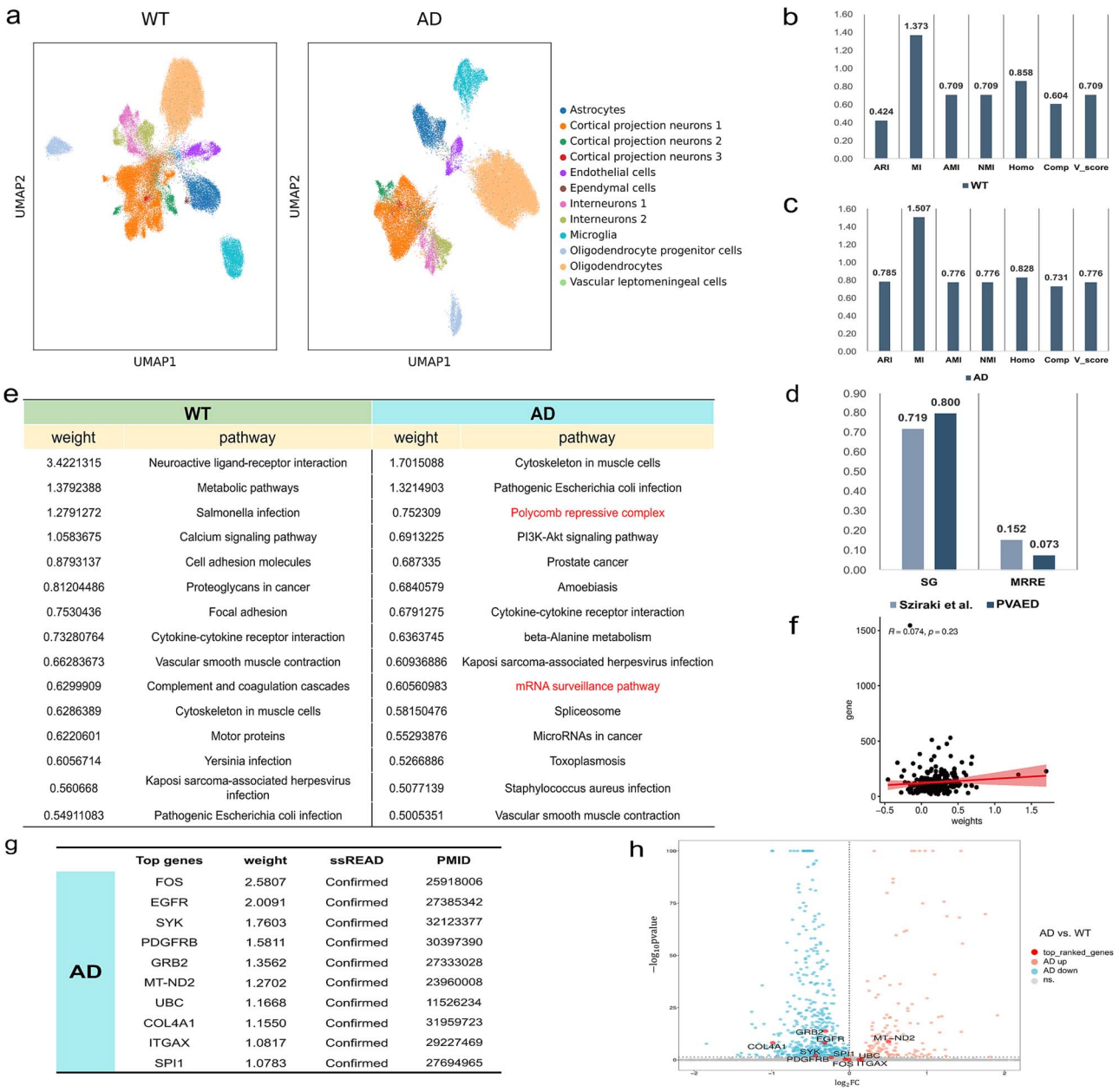


Figure 3 Interpretability of PVAED on phenotype related regulation analysis for lung cancer. (a) UMAP visualizations of human normal and tumor samples in GSE139107 dataset built on PVAED latent representations. (b, c) The quantification of clustering metrics on normal and tumor subsets shows the strong performance of PVAED. (d) The projection quality comparison between PVAED and the original data provider's embeddings. (e) Top 15 ranked pathways by exploring the pathway prior nodes in the prior layer of PVAED encoder. (f, g) Top 50 ranked genes from PVAED encoder and the corresponding enrichment analysis shows important pathways associated with lung cancer highlighted in red. (h) Hypergraph model of top-ranked pathways, plotted using HNX [43]. The partial zoom-in visualization showcases the construction details and the red colored genes are related to lung cancer. (i) RWR is performed on hypergraph model. Basically, a walker randomly chooses an incident hyperedge and then randomly chooses a node in the hyperedge with a probability for returning to the original node. (j) Left: prioritized lung tumor–related genes ranked according to the predicted weights obtained from the RWR process; Right: dominant regulator set identified from the hypergraph by spectral clustering. (k) GO biological process enrichment analysis based on identified eight dominant genes. (l) The predicted genes that could be potentially associated with HIF-1 signaling pathway by RWR on hypergraph model.



e

WT		AD	
weight	pathway	weight	pathway
3.4221315	Neuroactive ligand-receptor interaction	1.7015088	Cytoskeleton in muscle cells
1.3792388	Metabolic pathways	1.3214903	Pathogenic Escherichia coli infection
1.2791272	Salmonella infection	0.752309	Polycomb repressive complex
1.0583675	Calcium signaling pathway	0.6913225	PI3K-Akt signaling pathway
0.8793137	Cell adhesion molecules	0.687335	Prostate cancer
0.81204486	Proteoglycans in cancer	0.6840579	Amoebiasis
0.7530436	Focal adhesion	0.6791275	Cytokine-cytokine receptor interaction
0.73280764	Cytokine-cytokine receptor interaction	0.6363745	beta-Alanine metabolism
0.66283673	Vascular smooth muscle contraction	0.60936886	Kaposi sarcoma-associated herpesvirus infection
0.6299909	Complement and coagulation cascades	0.60560983	mRNA surveillance pathway
0.6286389	Cytoskeleton in muscle cells	0.58150476	Spliceosome
0.6220601	Motor proteins	0.55293876	MicroRNAs in cancer
0.6056714	Yersinia infection	0.5266886	Toxoplasmosis
0.560668	Kaposi sarcoma-associated herpesvirus infection	0.5077139	Staphylococcus aureus infection
0.54911083	Pathogenic Escherichia coli infection	0.5005351	Vascular smooth muscle contraction

g

	Top genes	weight	ssREAD	PMID
AD	FOS	2.5807	Confirmed	25918006
	EGFR	2.0091	Confirmed	27385342
	SYK	1.7603	Confirmed	32123377
	PDGFRB	1.5811	Confirmed	30397390
	GRB2	1.3562	Confirmed	27333028
	MT-ND2	1.2702	Confirmed	23960008
	UBC	1.1668	Confirmed	11526234
	COL4A1	1.1550	Confirmed	31959723
	ITGAX	1.0817	Confirmed	29227469
	SPI1	1.0783	Confirmed	27694965

Figure 4 PVAED's interpretability enables the identification of risk pathways and genes in Alzheimer's disease. (a) UMAP visualizations of GSE212606 dataset built on PVAED latent representations. (b, c) The clustering metrics for both WT and AD samples. (d) Projection quality comparison between the original data provider's embeddings and PVAED. (e) Top 15 ranked pathways from PVAED encoder. (f) Pearson correlation between pathway prior node weight and gene count. (g) Top 10 weighted genes from gene–gene regulation layer. This confirms the biological relevance and existing evidence for these identified genes in the context of AD. (h) Gene expression volcano plot for top-ranked genes shows that genes identified by PVAED are not all significantly up- or down-expressed, which means they will not be discovered by general analysis pipeline.

known cell types and differentiation trajectories in the developing mouse cortex.

The model identifies four previously uncharacterized subpopulations of migrating excitatory neurons (sub_1 to sub_4) in low-dimensional embeddings (Fig. 5a), achieving good results on clustering metrics, as well as higher SG score and better MRRE score than Di Bella *et al.*, the authors of the original dataset [46] (Fig. 5b and c). Meanwhile, the identified subpopulations could not be founded identified by standard workflow like UMAP (Supplementary materials Fig. S6). Differential expression analysis using Scanpy [47] identifies distinct molecular signatures for each subpopulation and the top-ranked genes are shown in Fig. 5d–f by the test statistic scores with filtering thresholds as log2FC >1 and adjust *P* value <.05. We also provide all the top 50 differential expression gene lists of subpopulations as Table S2 in Supplementary materials. The Venn diagram of differential gene lists illustrates both shared and unique expression signatures among subpopulations (Fig. 5g). Among the 200 genes analyzed, 144 are identified as subpopulation-specific, underscoring transcriptional heterogeneity across the four cell groups. The heatmap based on differentially expressed genes provide detailed views of gene expression patterns across these groups (Fig. 5h). These fundamental analyses methods consistently indicate that there were differences

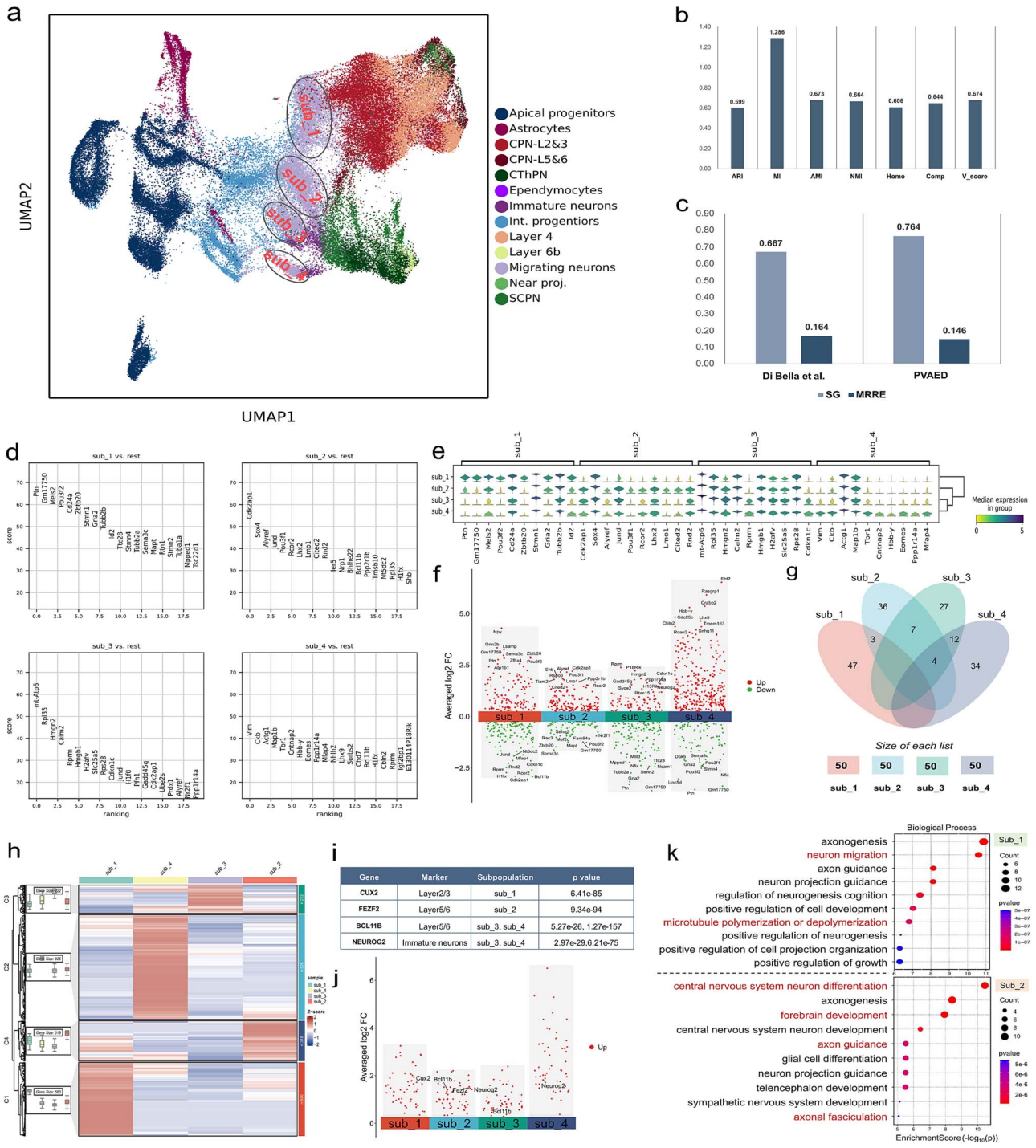


Figure 5 PVAED facilitates the identification of pre-committed cell subpopulations. (a) UMAP visualization of excitatory neurons in the GSE153164 dataset and the identified four subpopulations of migrating neurons based on PVAED latent embeddings. (b) Clustering metrics on dataset. (c) Projection quality comparison of SG and MRRE, where PVAED shows the better result. (d–f) Show top-ranked differentially expressed genes of the four subpopulations based on standard Scanpy analysis pipeline. (g) Venn diagrams of the top 50 ranked differential expression genes of the subpopulations. (h) Expression heatmap for identified subpopulations. (i) The markers of cortical layers and the expression significance in subpopulations. (j) Shows the up regulated marker genes in every subpopulation. (k) GO enrichment analyses of sub_1 and sub_2 are conducted based on the top 50 ranked differentially expressed genes by scores resulting from Scanpy [47] pipeline ($P < .05$) and application of Enrichr [48].

between these cell subpopulations, inspiring us to further explore the biological significance hidden beneath the surface of visualization.

To characterize the cellular identity of the identified subpopulations, we analyze the expression of established marker genes across different cortical layers. The PVAED-identified subpopulations show a strong correlation with specific cortical layer markers, including Cux2 (Layers 2/3), Fezf2 (Layers 5/6), and the immature neuron marker Neurog2 (Fig. 5i and j). Furthermore, GO enrichment analysis

of the top 50 differentially expressed genes, conducted by Enrichr [48], confirms that each subpopulation is associated with biologically distinct functional terms. For instance, the sub_1 (Layer 2/3) and sub_2 (Layer 5/6) subpopulations exhibit functional profiles consistent with their respective cortical identities (Fig. 5k). Specifically, sub_1 shows significant enrichment in terms related to neuronal migration and dynamic cytoskeletal reorganization (highlighted in red), suggesting an active state of positioning and structural plasticity. In contrast, sub_2 is characterized by terms associated with advanced neuronal differentiation and circuit formation. Enrichment results for sub_3 and sub_4 are detailed in the Figure S3 of Supplementary Part S8. These functional distinctions validate PVAED's ability to resolve developmentally meaningful cellular states and provide interpretable insights into cortical development.

Discussion

The application of deep generative models to single-cell omics data continues to expand rapidly. Established methods such as scVI [5] and scArches [49] have demonstrated the utility of VAEs in batch correction and cross-condition transfer learning, enabling robust integration of diverse datasets. More recently, foundation models including scGPT [31], scFoundation [32], and Geneformer [33] have leveraged transformer architectures and large-scale pretraining to learn generalizable representations from millions of single-cell profiles. These approaches position deep generative models not only as analytical tools but also as reusable backbones adaptable to specific biological tasks. Concurrently, diffusion-based frameworks such as scDiffusion [25] and stDiff [27] are reshaping trajectory inference and spatial transcriptomic imputation. Integrating structured biological knowledge with these advanced architectures offers a promising pathway to enhance both accuracy and interpretability in single-cell analysis, particularly for addressing the high noise and sparsity inherent in transcriptomic data.

Building on these advances, we introduce PVAED, a generative framework that synergistically combines VAEs and diffusion models with multi-scale biological priors, including transcription factor–target interactions, signaling pathway annotations, and protein complex associations. These priors are embedded into the VAE encoder via structured sparse connections, guiding the model to learn latent representations consistent with established molecular mechanisms. A subsequent diffusion module further refines these embeddings through a forward noising and reverse denoising process, substantially improving the resolution of rare cell states and visualization fidelity. Beyond dimensionality reduction, PVAED enables dynamic identification of key regulators during cellular differentiation by interpreting biologically grounded latent variables, such as pathways networks, thereby integrating mechanistic insight with computational performance.

A key advantage of PVAED is its flexibility in accommodating diverse forms and quantities of prior knowledge. The model adaptively constructs prior nodes during encoding, supporting broad applicability across biological contexts. However, on ultra-large datasets, PVAED incurs non-trivial computational cost; we therefore recommend its use within a pretraining–fine-tuning pipeline. By bridging mechanistic biology with generative AI, PVAED offers a scalable and interpretable platform for therapeutic target discovery and cross-species data alignment, advancing precision analytics in developmental and disease contexts.

Looking forward, several directions offer promising extensions to PVAED. First, the framework can be extended to multimodal single-cell data, accommodating the rapid growth of joint profiling technologies. Second, rather than treating priors as static inputs, future versions could incorporate dynamic knowledge graph embeddings that encode complex relationships among genes, pathways, drugs, and diseases, enabling richer interpretability and cross-domain inference. Lastly, integration with existing foundation models is a compelling direction; initializing PVAED's encoder with pretrained embeddings from models like scGPT or Geneformer could marry general-purpose representation learning with domain-aware interpretability.

In summary, PVAED represents a step toward a new paradigm in single-cell computational biology, one that unifies domain-specific knowledge with state-of-the-art generative modeling. As the field transitions from technical benchmarking to clinical translation, we expect that interpretable, scalable, and biologically grounded frameworks such as PVAED will play an essential role in bridging machine learning innovation with mechanistic discovery.

Key Points

- Integration of biological priors: PVAED incorporates known regulatory structures (pathways, protein complexes, and TF–target networks) into a variational autoencoder to achieve biologically meaningful dimension reduction of scRNA-seq data.
- Enhanced latent representation: a diffusion-based denoising module refines VAE embeddings, while a neighborhood-preserving loss item maintains local cell similarity, improving the quality of the reduced space.
- Superior performance: PVAED outperforms standard VAE and classical nonlinear dimension reduction methods like UMAP, siVAE, and VEGA, achieving roughly 43% better representation result and 34% (44%) higher projection quality across benchmark on multiple real datasets in terms of global (local) structure preservation.
- Interpretability and regulatory insight: through its prior node-based encoder and hypergraph modeling, PVAED allows interpretation for domain regulatory factors identification and enables prediction for underlying gene-pathway associations based on disease-relevant gene hypergraph.
- Biological discovery: application of PVAED uncovers pre-committed neuronal subpopulations during cortical neuron differentiation, demonstrating its utility in revealing hidden cellular heterogeneity.

Acknowledgements

This project is supported by National Natural Science Foundation of China (No. 32470682 and No. 32100512 to W.M.). We gratefully acknowledge the Bioinformatics Core in Center for Single-Cell Omics (CSCOmics), Shanghai Jiao Tong University School of Medicine for providing high-performance computing services.

Author contributions

Conceptualization, methodology, supervision, W.M.; software, validation, formal analysis, and visualization, Y.N., Y.C. and W.M.; Writing – original draft, writing – review & editing, Y.N. and W.M. All authors read and approved the final manuscript.

Supplementary data

Supplementary material is available at *Briefings in Bioinformatics* online.

Conflicts of interest

None declared.

Data availability

PBMC 3k and 10k data are downloaded from 10x genomics website under the corresponding keywords at https://www.10xgenomics.com/. Up to now, all the GEO data used in this work are freely accessed in GEO database under the corresponding accession number, which are GSE165657 (human cerebellar development), GSE204684 (human cerebral cortex development), GSE212606 (AD samples in mammalian brain), GSE131907 (LUAD samples), and GSE153164 (mouse cerebral cortex diversification). The NEMO data SCR_016152 (human hypothalamus development) is downloaded from https://assets.nemoarchive.org/dat-jx4eu3g.

Code availability

The current version of source code and software implementation for PVAED is available at: https://github.com/MaLab-scGenomics/PriorVAED. Basic tutorial is available at Read the Docs website: https://pvaed.readthedocs.io/en/master/index.html.

References

1. Kolodziejczyk AA, Kim JK, Svensson V *et al.* The technology and biology of single-cell RNA sequencing. *Mol Cell* 2015;**58**:610–20. https://doi.org/10.1016/j.molcel.2015.04.005
2. Jovic D, Liang X, Zeng H *et al.* Single-cell RNA sequencing technologies and applications: a brief overview. *Clin Transl Med* 2022;**12**:e694. https://doi.org/10.1002/ctm2.694
3. Gohil SH, Iorgulescu JB, Braun DA *et al.* Applying high-dimensional single-cell technologies to the analysis of cancer immunotherapy. *Nat Rev Clin Oncol* 2021;**18**:244–56. https://doi.org/10.1038/s41571-020-00449-x
4. Kingma DP, Welling M. Auto-encoding variational bayes. *The 2nd International Conference on Learning Representations (ICLR)*. Banff, Canada, 2014. https://doi.org/10.48550/arXiv.1312.6114
5. Lopez R, Regier J, Cole MB *et al.* Deep generative modeling for single-cell transcriptomics. *Nat Methods* 2018;**15**:1053–8. https://doi.org/10.1038/s41592-018-0229-2
6. Lotfollahi M, Wolf FA, Theis FJ. scGen predicts single-cell perturbation responses. *Nat Methods* 2019;**16**:715–21. https://doi.org/10.1038/s41592-019-0494-8
7. Svensson V, Gayoso A, Yosef N *et al.* Interpretable factor models of single-cell RNA-seq via variational autoencoders. *Bioinformatics* 2020;**36**:3418–21. https://doi.org/10.1093/bioinformatics/btaa169
8. Grønbech CH, Vording MF, Timshel PN *et al.* scVAE: Variational auto-encoders for single-cell gene expression data. *Bioinformatics* 2020;**36**:4415–22. https://doi.org/10.1093/bioinformatics/btaa293
9. Wang D, Gu J. VASC: dimension reduction and visualization of single-cell RNA-seq data by deep variational autoencoder. *Genom Proteom Bioinform* 2018;**16**:320–31. https://doi.org/10.1016/j.gpb.2018.08.003
10. Jiang J, Xu J, Liu Y *et al.* Dimensionality reduction and visualization of single-cell RNA-seq data with an improved deep variational autoencoder. *Brief Bioinform* 2023;**24**:bbad152. https://doi.org/10.1093/bib/bbad152
11. Geddes TA, Kim T, Nan L *et al.* Autoencoder-based cluster ensembles for single-cell RNA-seq data analysis. *BMC Bioinformatics* 2019;**20**:660. https://doi.org/10.1186/s12859-019-3179-5
12. Yushan Q, Lingfei Y, Hao J *et al.* scTPC: a novel semisupervised deep clustering model for scRNA-seq data. *Bioinformatics* 2024;**5**:5.
13. Liu T, Jia C, Bi Y *et al.* Li F: scDFN: enhancing single-cell RNA-seq clustering with deep fusion networks. *Brief Bioinform* 2024;**25**:bbae486. https://doi.org/10.1093/bib/bbae486
14. McFarland JM, Paolella BR, Warren A *et al.* Multiplexed single-cell transcriptional response profiling to define cancer vulnerabilities and therapeutic mechanism of action. *Nat Commun* 2020;**11**:4296. https://doi.org/10.1038/s41467-020-17440-w
15. Ma J, Yu MK, Fong S *et al.* Using deep learning to model the hierarchical structure and function of a cell. *Nat Methods* 2018;**15**:290–8. https://doi.org/10.1038/nmeth.4627
16. Ashburner M, Ball CA, Blake JA *et al.* Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nat Genet* 2000;**25**:25–9. https://doi.org/10.1038/75556
17. Buettner F, Pratanwanich N, McCarthy DJ *et al.* f-scLVM: scalable and versatile factor analysis for single-cell RNA-seq. *Genome Biol* 2017;**18**:212. https://doi.org/10.1186/s13059-017-1334-8
18. Yan H, Weng D, Li D *et al.* Prior knowledge-guided multilevel graph neural network for tumor risk prediction and interpretation via multi-omics data integration. *Brief Bioinform* 2024;**25**:bbae184. https://doi.org/10.1093/bib/bbae184
19. Seninge L, Anastopoulos I, Ding H *et al.* VEGA is an interpretable generative model for inferring biological network activity in single-cell transcriptomics. *Nat Commun* 2021;**12**:5684. https://doi.org/10.1038/s41467-021-26017-0
20. Jassal B, Matthews L, Viteri G *et al.* The reactome pathway knowledgebase. *Nucleic Acids Res* 2020;**48**:D498–D503. https://doi.org/10.1093/nar/gkz1031
21. Choi Y, Li R, Quon G. siVAE: Interpretable deep generative models for single-cell transcriptomes. *Genome Biol* 2023;**24**:29. https://doi.org/10.1186/s13059-023-02850-y
22. Rombach R, Blattmann A, Lorenz D *et al.* High-resolution image synthesis with latent diffusion models. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 2022;10674–85. https://doi.org/10.1109/CVPR52688.2022.01042
23. Croitoru FA, Hondru V, Ionescu RT *et al.* Diffusion models in vision: a survey. *IEEE Trans Pattern Anal Mach Intell* 2023;**45**:10850–69. https://doi.org/10.1109/TPAMI.2023.3261988
24. Bond-Taylor S, Leach A, Long Y *et al. Deep generative modelling: a comparative review of VAEs*, GANs, normalizing flows, energy-based and autoregressive models. *IEEE Trans Pattern Anal Mach Intell* 2021;**44**:7327–47. https://doi.org/10.1109/TPAMI.2021.3116668
25. Luo E, Hao M, Wei L *et al.* scDiffusion: conditional generation of high-quality single-cell data using diffusion model. *Bioinformatics* 2024;**40**:btae518. https://doi.org/10.1093/bioinformatics/btae518
26. Schuette G, Lao Z, Zhang B. ChromoGen: diffusion model predicts single-cell chromatin conformations. *Sci Adv* 2025;**11**:eadr8265. https://doi.org/10.1126/sciadv.adr8265
27. Li K, Li J, Tao Y *et al.* stDiff: a diffusion model for imputing spatial transcriptomics through single-cell transcriptomics. *Brief Bioinform* 2024;**25**:bbae171. https://doi.org/10.1093/bib/bbae171

28. Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 2000;**28**:27–30. https://doi.org/10.1093/nar/28.1.27
29. Ruepp A, Waegele B, Lechner M *et al.* CORUM: the comprehensive resource of mammalian protein complexes—2009. *Nucleic Acids Res* 2010;**38**:D497–501. https://doi.org/10.1093/nar/gkp914
30. Zhang Q, Liu W, Zhang HM *et al.* hTFtarget: a comprehensive database for regulations of human transcription factors and their targets. *Genom Proteom Bioinform* 2020;**18**:120–8. https://doi.org/10.1016/j.gpb.2019.09.006
31. Cui H, Wang C, Maan H *et al.* scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nat Methods* 2024;**21**:1470–80. https://doi.org/10.1038/s41592-024-02201-0
32. Hao M, Gong J, Zeng X *et al.* Large-scale foundation model on single-cell transcriptomics. *Nat Methods* 2024;**21**:1481–91. https://doi.org/10.1038/s41592-024-02305-7
33. Theodoris CV, Xiao L, Chopra A *et al.* Transfer learning enables predictions in network biology. *Nature* 2023;**618**:616–24. https://doi.org/10.1038/s41586-023-06139-9
34. Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. *Proceedings of the 34th International Conference on Neural Information Processing Systems 2020*, pp. 6840–51. Red Hook, NY, USA: Curran Associates Inc., Article 574. https://doi.org/10.48550/arXiv.2006.11239
35. Kaski S, Nikkilä J, Oja M *et al.* Trustworthiness and metrics in visualizing similarity of gene expression. *BMC Bioinformatics* 2003;**4**:48. https://doi.org/10.1186/1471-2105-4-48
36. Shepard RN. The analysis of proximities: multidimensional scaling with an unknown distance function. *I Psychometrika* 1962;**27**:125–40. https://doi.org/10.1007/BF02289630
37. John AL, Michel V. Quality assessment of dimensionality reduction: rank-based criteria. *Neurocomputing* 2009;**72**:1431–43. https://doi.org/10.1016/j.neucom.2008.12.017
38. Zhu K, Bendl J, Rahman S *et al.* Multi-omic profiling of the developing human cerebral cortex at the single-cell level. *Sci Adv* 2023;**9**:adg3754. https://doi.org/10.1126/sciadv.adg3754
39. Hendriks LEL, Remon J, Faivre-Finn C *et al.* Non-small-cell lung cancer. *Nat Rev Dis Primers* 2024;**10**:71. https://doi.org/10.1038/s41572-024-00551-9
40. Kim N, Kim HK, Lee K *et al.* Single-cell RNA sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma. *Nat Commun* 2020;**11**:2285. https://doi.org/10.1038/s41467-020-16164-1
41. Contisciani M, Battiston F, De Bacco C. Inference of hyperedges and overlapping communities in hypergraphs. *Nat Commun* 2022;**13**:7229. https://doi.org/10.1038/s41467-022-34714-7
42. Abolghasemi S, Bikhof Torbati M, Pakzad P *et al.* Gene expression analysis of SOCS, STAT and PIAS genes in lung cancer patients. *Pathol Res Pract* 2023;**249**:154760. https://doi.org/10.1016/j.prp.2023.154760
43. Praggastis B, Aksoy S, Arendt D *et al.* HyperNetX: a python package for modeling complex network data as hypergraphs. *J Open Source Softw* 2024;**9**:6016. https://doi.org/10.21105/joss.06016
44. Sziraki A, Lu Z, Lee J *et al.* A global view of aging and Alzheimer's pathogenesis-associated cell population dynamics and molecular signatures in human and mouse brains. *Nat Genet* 2023;**55**:2104–16. https://doi.org/10.1038/s41588-023-01572-y
45. Wang C, Acosta D, McNutt M *et al.* A single-cell and spatial RNA-seq database for Alzheimer's disease (ssREAD). *Nat Commun* 2024;**15**:4710. https://doi.org/10.1038/s41467-024-49133-z
46. Di Bella DJ, Habibi E, Stickels RR *et al.* Molecular logic of cellular diversification in the mouse cerebral cortex. *Nature* 2021;**595**:554–9. https://doi.org/10.1038/s41586-021-03670-5
47. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol* 2018;**19**:15. https://doi.org/10.1186/s13059-017-1382-0
48. Xie Z, Bailey A, Kuleshov MV *et al.* Gene set knowledge discovery with Enrichr. *Curr Protoc* 2021;**1**:e90. https://doi.org/10.1002/cpz1.90
49. Lotfollahi M, Naghipourfar M, Luecken MD *et al.* Mapping single-cell data to reference atlases by transfer learning. *Nat Biotechnol* 2022;**40**:121–30. https://doi.org/10.1038/s41587-021-01001-7

Briefings in Bioinformatics, 2026, **27**, bbag259
https://doi.org/10.1093/bib/bbag259
Problem Solving Protocol