nature communications

Article

https://doi.org/10.1038/s41467-026-70071-5

Illuminating cell states by a comprehensive and interpretable single cell foundation model

Received: 30 July 2025

Accepted: 13 February 2026

Published online: 16 March 2026

Check for updates

Jue Wang[1,2,4], Cheng Tan[1,2,4], Zhangyang Gao[1,2], Sida Shao[2], Shiping Liu[3] ✉ & Stan. Z. Li[2] ✉

Advances in single-cell sequencing have enabled AI-driven foundation models with powerful data representation. However, their practical use is limited by real-world data sparsity, heterogeneity, and poor interpretability. To overcome these, we introduce CellVQ. To enhance generalizability, we incorporate a large-scale single-cell dataset comprising 68 million cells, model parameters totaling 500 million, and challenging pretraining tasks. Notably, we introduce a Single-Cell Discretization (SCD) module that effectively represents cell embeddings, addressing data heterogeneity. For improved interpretability, the SCD module transforms high-dimensional and sparse single-cell data into a "cell code," facilitating recognition and analysis. Additionally, we also present CellVQ-Graph, a plug-and-play tool that integrates CellVQ's features with multimodal data (genes, cell communication, annotations) to build a knowledge graph for biological discovery. Extensively evaluated, CellVQ outperforms strong baselines in all downstream tasks, and also uncovered intriguing biological phenomena with compelling explanations. CellVQ aspires to serve as a truly applicable and generalizable AI tool for the cell biology community.

Cells, as the fundamental units of life, though tiny, execute a remarkable diversity of essential functions crucial for maintaining human health[1,2]. The discovery of cells by Robert Hooke through his microscope captivated scientists, prompting an enduring fascination with their complexity and intricacies, and initiating a quest to uncover the secrets of cellular life. Historically, during the microscopy era, biologists primarily speculated about cells at the microimaging level, where observations were largely confined to their morphology, activity, and interactions. This limitation impeded deeper investigations into the complexities of cellular states. However, with the advent of the post-genomic era, a paradigm shift has emerged toward a molecular understanding of cell states, highlighting the myriad factors that influence them[3]. A significant international initiative, the Human Cell Atlas Project[4], was launched in 2017 to establish "an open, comprehensive reference map of the molecular state of cells in healthy human tissues." This project employs single-cell RNA sequencing (scRNA-seq)[5,6] techniques and has inspired similar endeavors focusing on well-studied animal models[7,8] and patient-derived tissue samples[9] across a diverse range of diseases. In addition to RNA sequencing, other omics technologies, such as single-cell ATAC sequencing (scATAC-seq)[10,11] for chromatin accessibility and proteomics, have been employed to comprehensively analyze cell states. Concurrently, the availability of sequenced single-cell data has surged to tens of millions, presenting a significant opportunity for the application of AI-based methods in cell biology.

Leveraging large-scale scRNA-seq data, several deep learning-based methods have emerged to learn informative cell representations through pretraining on extensive datasets. For instance, scBERT[12] employs a bidirectional transformer architecture[13] to pretrain on unlabeled scRNA-seq data, subsequently fine-tuning for specific tasks. Similarly, scGPT[14] utilizes a generative pretrained transformer trained on a repository of over 33 million cells, demonstrating robust

[1]Zhejiang University, Hangzhou, China. [2]Westlake University, Hangzhou, China. [3]BGI Research, Hangzhou, China. [4]These authors contributed equally: Jue Wang, Cheng Tan. ✉e-mail: liushiping@genomics.cn; Stan.ZQ.Li@westlake.edu.cn

performance in both generative and representation tasks. scFoundation[15] adopts an asymmetric encoder-decoder transformer to conduct read-depth-aware modeling pretraining on 50 million transcriptomic profiles. Additionally, scPRINT[16] focuses on gene network inference by interpreting the attention matrix generated by its transformer, which is pretrained on 50 million cells. Collectively, these models leverage large-scale single-cell data to learn robust cell representations and exhibit impressive performance across various downstream tasks. Despite these advancements, emerging evidence indicates notable generalizability limitations of these foundation models[17,18]. These limitations can be categorized into three primary aspects: (1) Sparsity of Single-Cell Data[19]: In scRNA-seq and scATAC-seq datasets, only a subset of genes or regions is expressed or captured, resulting in a substantial amount of zero gene expression. This sparsity challenges the model's capacity to extract compact and rich information from high-dimensional data. (2) Heterogeneity of Single-Cell Data[17,18]: Single-cell data are generated by various institutes using different sequencing, processing, and normalization methods. This diversity leads to significant variations in the sequenced genes and expression scales, complicating the generalization of single-cell foundation models. (3) Poor interpretability of Foundation Models: These models contain a vast number of parameters and exhibit high complexity. While they are capable of generating informative gene and cell representations, the high-dimensional and continuous nature of these representations renders them challenging for biological scientists to comprehend.

We present CellVQ, a comprehensive and interpretable single-cell foundation model. CellVQ innovatively integrates a Single-Cell Discretization (SCD) module into the single-cell foundation model. This module utilizes a unified cell codebook to transform the cell representation into a cell code and uses an SCD cell embedding derived from the cell codebook to represent the cell corresponding to the code. CellVQ employs an encoder-SCD-decoder architecture, encompassing 511 million parameters and pretrained on over 68 million single-cell data points. The model incorporates multiple pretraining tasks at both the cell and gene levels. At the cell level, the task involves predicting cell properties such as type, tissue, age, sex, and disease. At the gene level, we implement a cell-gene-prompt (CGP)-based Zero-Inflated Negative Binomial (ZINB) prediction task[20,21], which effectively enhances cell representation and captures gene co-expression patterns, and addresses the high sparsity of the data. Notably, our findings and prior evidence[22,23] indicate that the SCD module effectively mitigates data heterogeneity and batch effects[24]. By utilizing the unified representation and cell codes from the cell codebook, the SCD module represents cells from different batches, leading to significant enhancements in both generalization and interpretability. Additionally, we introduce a lightweight model, CellVQ-Graph, designed to integrate multi-omic and multimodal data to construct a Graph Knowledge Network (GKN). By leveraging the representations generated by CellVQ, the plug-and-play functionality of CellVQ-Graph facilitates the transfer of insights to diverse real-world datasets, enabling the discovery of biological knowledge through the GKN.

We evaluated the performance of our models across multiple tasks, categorized into prediction and analysis tasks. In the realm of prediction tasks, which encompass cell clustering, annotation, property prediction, gene and drug perturbation prediction, as well as drug response prediction, CellVQ demonstrates significant advancements over other foundation models such as scGPT and scFoundation. For analysis tasks, we integrated CellVQ and CellVQ-Graph into four distinct tasks: cell state discovery, gene discovery, gene regulatory network (GRN) analysis[25,26], and multimodal analysis. These analyses highlight the unique advantages of our models in terms of interpretability. We contend that the introduction of CellVQ and CellVQ-Graph effectively bridges the gap between AI models and biological research, thereby enhancing the understanding of cellular and genetic mechanisms.

Results

Overview of CellVQ

Figure 1 illustrates the comprehensive framework of CellVQ. Figure 1a depicts the model architecture of CellVQ. For the encoder input, we first append a [Count] token, representing the sum of gene expression values. We then extract expression values from gene expression profiles and randomly sample genes as input for the encoder. The sampling rates follow a uniform distribution $U(0.5, 0.8)$ to facilitate multi-scale information extraction. These sampled gene expressions are processed through an embedding module comprising three components: gene expression embedding, gene position (gene id) embedding, and gene ontology embedding. The gene expression embedding transforms continuous gene expression scalars into learnable high-dimensional vectors, while the gene position embedding encodes gene identification(gene names, like *DLK1*, *CLU*). The Gene Ontology embedding module incorporates prior knowledge from Gene Ontology[27] to model gene co-expression, which is updated during training. Subsequently, these embeddings are passed through 12 Transformer layers to refine the gene embeddings, leading to the generation of a cell embedding via a pooling method. We input the gene embeddings into an SCD module to obtain SCD gene embeddings, while the cell embedding is quantized to produce an SCD cell embedding and a cell code, which serves as a data-driven identifier for the cell. The SCD cell embedding is utilized for multiple property predictions to train both the encoder and the SCD module. In the decoder component, we initialize all expression values to zero, allowing the embedding module to focus solely on capturing gene position and ontology information. We replace the decoder embeddings at the corresponding gene positions with SCD gene embeddings as prompts, setting the SCD cell embedding at the "Count" token for the cell prompt. These processed decoder embeddings are subsequently input into six Performer[28] layers and a Zero-Inflated Negative Binomial (ZINB) predictor to estimate the ZINB distribution for each gene expression value. We select predictions for non-sampled genes and a subset of non-expressed genes (equal to or fewer than the non-sampled) to compute the ZINB loss against the corresponding gene expression values, facilitating the training of the entire model. Further architectural details are provided in Section 4.

Figure 1b visualizes the six prediction tasks implemented using CellVQ, including cell clustering, annotation, property prediction, gene perturbation prediction, and drug response prediction. All prediction tasks were performed under the zero-shot setting without fine-tuning. Additionally, we provide the comprehensive ablation study of CellVQ in Supplementary Section 1; and the results of batch correction tasks to demonstrate that CellVQ effectively resolves data heterogeneity in Supplementary Section 3. We also showcase the extensive experiments of CellVQ on Cell clustering and cell type annotation tasks, comparing with broader baselines on TS2[29,30] datasets in Supplementary Section 2.

Figure 1c illustrates the integration of CellVQ and CellVQ-Graph. CellVQ-Graph receives representations from CellVQ and combines these with other omics and modal data, such as scATAC-seq and property annotation data, to construct the graph knowledge network. Figure 1d presents four analysis tasks: cell state discovery, gene discovery, GRN analysis, and multimodal analysis. In all analysis tasks, we only utilize the embeddings provided by CellVQ pre-training model and train CellVQ-Graph on the corresponding dataset. Comprehensive details regarding the datasets and experiments are available in the Supplementary Sections 12 and 13.

Representation improvement on cell clustering

The cell clustering task is fundamental to single-cell analysis. By clustering cells based on similar gene expression profiles, biologists can achieve cell type identification, analyze cell heterogeneity, conduct

https://doi.org/10.1038/s41467-026-70071-5

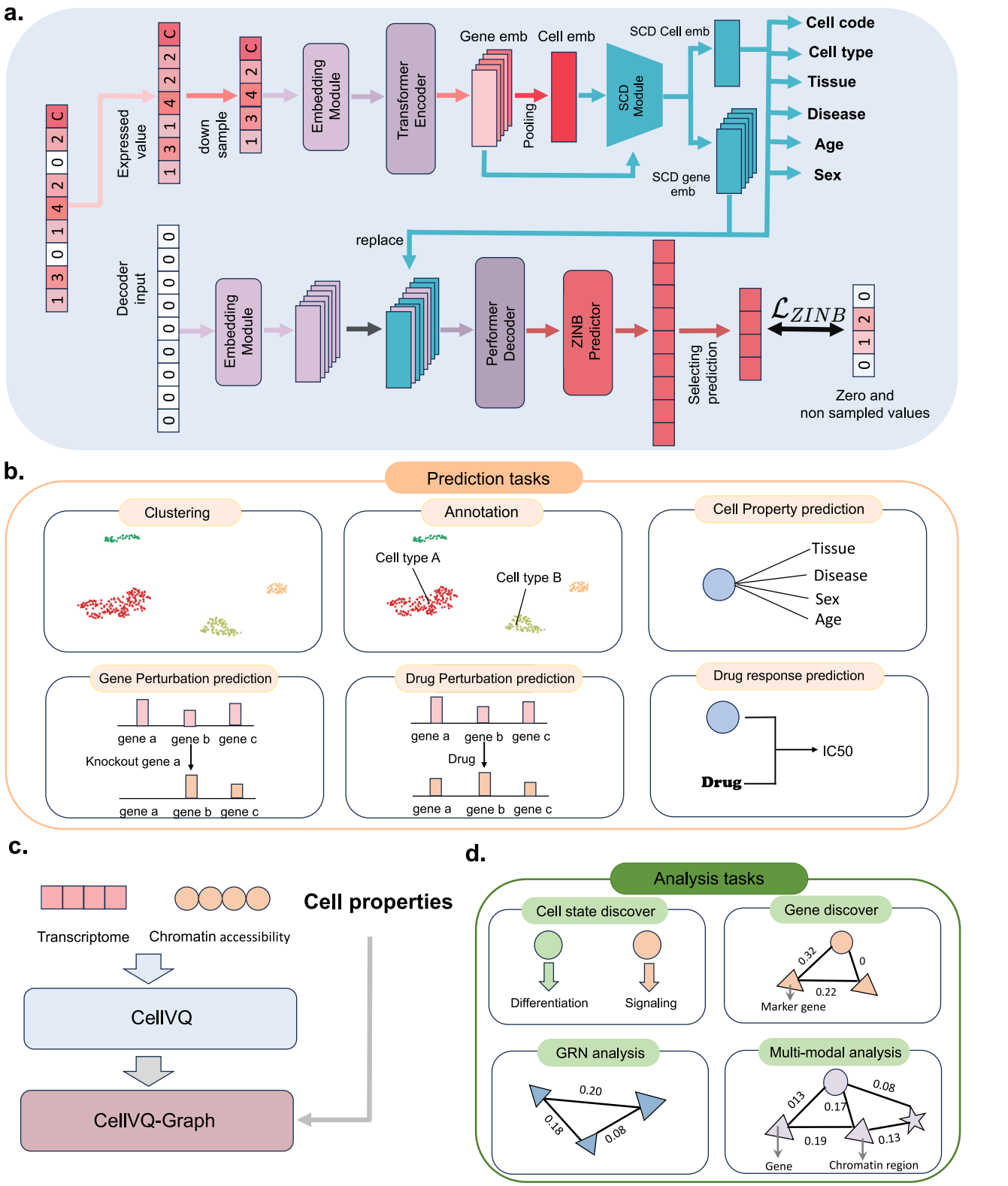

Fig. 1 | The overall framework of CellVQ. a The model architecture of CellVQ. Token c denotes the `All Count', which is the sum of gene expression values. **b** CellVQ is applied in six prediction tasks. **c** The integration of CellVQ and CellVQ-Graph. **d** CellVQ and CellVQ-Graph are applied in four analysis tasks. A circle, triangle, and pentagram represent a cell, gene (or chromatin region), and an attribute, respectively.

drug response studies, and discover biomarkers. Additionally, the effectiveness of the clustering task serves as a rigorous test of representation quality; superior cell representations yield higher accuracy and clearer visualizations.

To evaluate this, we applied CellVQ to three scRNA-seq datasets, specifically the Pancreas[14], Zheng 68K[31], and Segerstolpe[32] datasets, all of which include cell-type annotation labels. We compared results with raw gene expression, GeneCompass[33], scBERT[12], UCE[34], CellFM[35],

scGPT, and scFoundation. For each method, we derived cell representations and employed the K-means algorithm for clustering based on cell type labels. We utilized several evaluation metrics: Accuracy, Normalized Mutual Information[36] (NMI), Adjusted Rand Index[37] (ARI), and Silhouette Coefficient[38] (SIL). Among these metrics, Accuracy, NMI, and ARI assess whether cells of the same type are clustered together, while SIL measures the closeness of samples to their respective clusters and their separation from other clusters. Thus, these metrics provide a comprehensive evaluation of CellVQ's performance in clustering tasks.

Figure 2a displays the overall performance of the eight methods across the three datasets. The metrics for CellVQ on the Pancreas dataset are as follows: NMI (0.9424), ARI (0.9492), and SIL (0.4636). For the Zheng dataset, the metrics are: Accuracy (0.5127), NMI (0.6000), ARI (0.3904), and SIL (0.2050). The Segerstolpe dataset yields: Accuracy (0.7401), NMI (0.8720), ARI (0.6824), and SIL (0.6288). CellVQ achieves exceptional performance across all datasets and metrics, particularly in the Pancreas dataset, where Accuracy, NMI, and ARI exceed 0.94, with a SIL of 0.4636. In contrast, other methods report NMI below 0.89, ARI below 0.77, and SIL below 0.34. Furthermore, in the Segerstolpe dataset, CellVQ outperforms other methods across all metrics, leading by 0.23 in NMI, 0.30 in ARI, and 0.5 in SIL. These results underscore the superior representation and zero-shot capabilities of CellVQ.

Figure 2b visualizes UMAP[39] representations with cell type labels for the Segerstolpe dataset based on the eight methods. The UMAP representation of raw gene expression appears scattered, indicating that the high dimensionality and sparsity of raw gene expression inhibit the capture of critical information. In contrast, other methods, like scGPT and scFoundation achieve more precise clustering of cells of the same type; however, they still fail to assign all cells into correct and compact clusters. Notably, CellVQ successfully separates all cell types and compresses each cluster into a compact group, effectively capturing the biological relationships between different cell types. For instance, endothelial cells and pluripotent stem cells (PSCs) are positioned closely in the UMAP generated by CellVQ. This proximity reflects the biological understanding that PSCs can differentiate into endothelial cells and co-express certain marker genes, such as *VEGF* (Vascular Endothelial Growth Factor), *CD31*, and *CD34*, which are commonly used to label endothelial and precursor cells. Consequently, CellVQ adeptly identifies the underlying relationship between these two cell types while exhibiting similar representations.

Cell annotation and properties prediction

Cell type annotation is a critical step in the analysis of single-cell data, aimed at assigning biological significance to clustered cell populations by correlating them with known biological cell types or states, as defined by established marker genes. To further validate the zero-shot capability and representation quality of CellVQ, we applied it to three scRNA-seq datasets, similar to those used in the clustering tasks, and compared the results with four baseline methods. We established a linear classification model using scikit-learn[40], employing the representations generated by the eight methods as input to predict cell types. The evaluation of zero-shot performance utilized Accuracy, F1 score, Precision, and Recall as metrics.

Figure 2c presents the overall performance of the eight methods across the three datasets. The metrics for CellVQ are as follows: Pancreas (Accuracy: 0.88, F1 score: 0.87, Precision: 0.88, Recall: 0.88), Zheng (Accuracy: 0.90, F1 score: 0.89, Precision: 0.90, Recall: 0.89), and Segerstolpe (Accuracy: 0.85, F1 score: 0.80, Precision: 0.76, Recall: 0.85). Consistent with the clustering tasks, CellVQ demonstrates significant superiority across all metrics for the three datasets. Specifically, CellVQ outperforms the other methods by margins of 0.20, 0.22, 0.21, and 0.22 in the Pancreas dataset, 0.26, 0.27, 0.26, and 0.25 in the Zheng dataset, and 0.26, 0.33, 0.26, and 0.26 in the Segerstolpe dataset across the four metrics. These results indicate that the representations provided by CellVQ are more informative and exhibit superior linear interpretability.

To evaluate CellVQ's performance on multiple property prediction tasks, we created a test dataset that excluded samples from the pretraining dataset and included cell property annotation labels such as age, disease, sex, and tissue type. We applied the eight methods to this dataset and assessed performance using accuracy as the metric.

Figure 2 d showcases the accuracy of the eight methods on multiple property prediction tasks. Notably, the accuracy of scGPT and scFoundation on these tasks is lower than that of raw gene expression, indicating their inability to capture critical patterns related to these properties, resulting in information loss. In contrast, CellVQ achieves higher accuracy in property predictions, demonstrating its capacity to extract comprehensive insights and transform raw gene expression values into a compact yet informative cell representation.

Figure 2 e illustrates the UMAP visualization of CellVQ on this test dataset, with tissue type labels. The results indicate that CellVQ effectively aggregates cells of the same tissue type into cohesive populations. While some groups of identical tissue types are separated, these distinctions reflect different cell types or states, underscoring that CellVQ provides a more nuanced understanding of cellular patterns.

CellVQ enhances bulk-level RNA-seq data representation

Bulk RNA-seq[41] data is generated by extracting RNA from a mixed population of cells, rather than from individual cells, and subsequently sequencing it. This approach provides the average gene expression profile of all cells in the sample, in contrast to single-cell data, which offers single-cell resolution. Although bulk RNA-seq data lacks the cellular heterogeneity captured by single-cell sequencing, it remains essential for numerous practical applications, such as tumor and drug response research, due to its more established and cost-effective experimental and analytical processes. To evaluate whether CellVQ can enhance the representation of bulk RNA-seq data and further demonstrate its generalization capabilities, we applied CellVQ to two drug response prediction tasks: drug response classification and cancer drug response prediction.

The goal of drug response classification is to predict the sensitivity of a patient or cell to a specific drug based on various biological characteristics, such as gene expression. This task offers valuable insights for precision medicine, optimizing drug development, and identifying potential new therapies. Our approach incorporates both bulk and single-cell data as inputs and outputs the predicted sensitivity for each cell. Utilizing the downstream module SCAD[42], we leverages CellVQ to obtain the embeddings of single-cell and bulked data to train SCAD models for predicting a cell's sensitivity to the drug. We compare CellVQ performance with that of two foundation models: scGPT and scFoundation under the same setting. We also include Linear and Random Forest[43] as additional baselines for more comprehensive comparison. Our analysis focused on four drug test sets which are collected from the Genomics of Cancer Drug Sensitivity database[44]: Sorafenib, PLX4720-451Lu, NVP-TAE684, and Etoposide. We measured the quality of predictions using the AUC and recall metrics.

Figure 3 a presents the ROC curves for the six methods across four drugs. CellVQ demonstrates consistently superior performance, particularly for Etoposide, where it achieves higher accuracy than other baselines at all thresholds. Notably, the baseline SCAD, Linear and Random Forest exhibit poor results for PLX4720 and Sorafenib, performing worse than random predictions, indicating significant challenges in predicting sensitivity for these drugs. In contrast, CellVQ maintains a clear advantage, especially with Sorafenib, where the true positive rate exceeds 0.8 when the threshold is above 0.2, highlighting its effectiveness.

https://doi.org/10.1038/s41467-026-70071-5

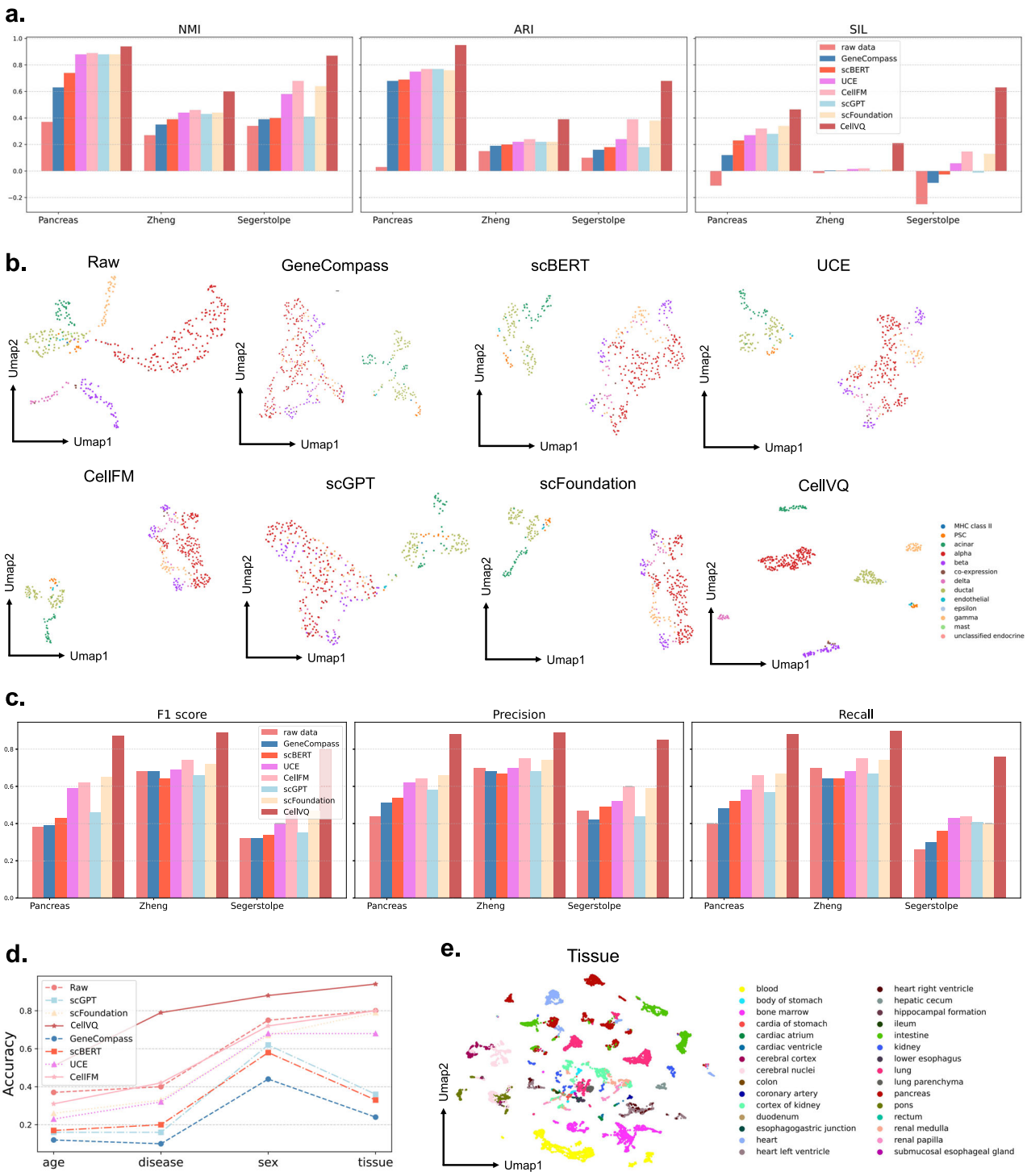

Fig. 2 | Cell clustering, annotation, and properties prediction results of CellVQ and other baselines. a The cell clustering results of eight methods on three scRNA-seq datasets in three metrics. "raw data" represent the raw gene expression values. **b** The UMAP visualization results with cell type as labels for eight methods on Segerstolpe dataset. **c** The cell annotation task results of eight methods on three scRNA-seq datasets in three metrics. **d** The accuracy of four cell properties prediction tasks for eight methods on a sampled dataset. **e** The UMAP visualization results of CellVQ with tissue as labels on a sampled dataset.

Figure 3 b illustrates the average AUC and recall results for the six methods across the four drugs. The average AUC and recall for CellVQ are as follows: Sorafenib (0.88, 0.91), PLX4720-451Lu (0.77, 0.80), NVP-TAE684 (0.83, 0.86), and Etoposide (0.76, 0.79). These results indicate that CellVQ maintains both AUC and recall above 0.75, achieving leading performance across all drugs and metrics compared to other methods. This demonstrates that CellVQ effectively transfers to bulk RNA-seq data, extracting more information to enhance the accuracy of drug response classification.

In addition to drug response classification, we applied CellVQ to the cancer drug response (CDR) prediction task, which examines tumor cell responses to drug interventions. Given the drug information and bulk RNA-seq data, the model predicts the half-maximal inhibitory concentration IC_{50}. To enhance prediction quality, we

https://doi.org/10.1038/s41467-026-70071-5

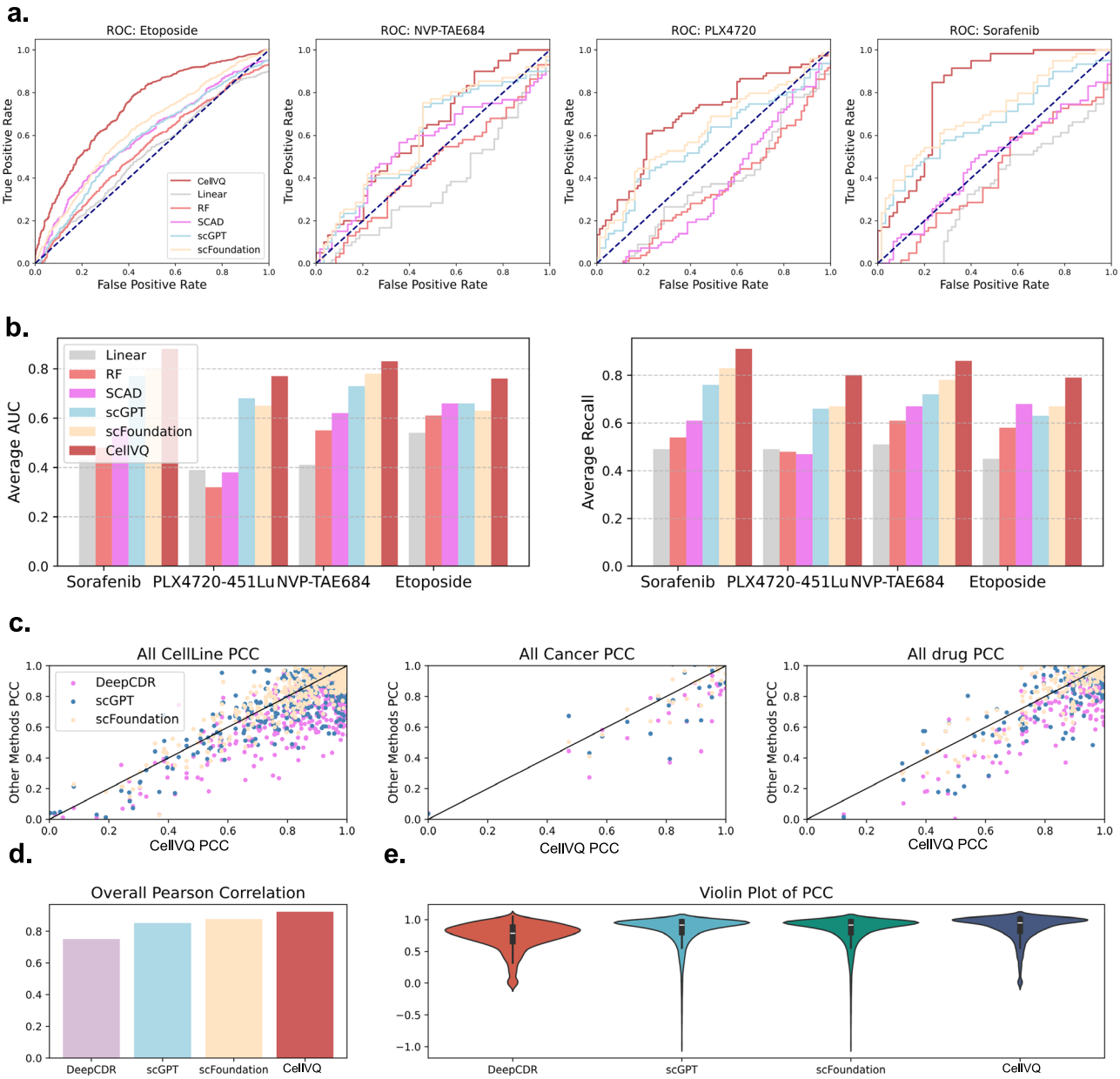

Fig. 3 | Drug Response Prediction Results of CellVQ and Other Baselines. RF denotes the "Random Forest". **a** The ROC curves for the six methods across four drugs are presented. **b** The average AUC and recall for the six methods on these drugs are summarized. **c** A comparison of the Pearson correlation coefficients between CellVQ and the three baseline methods for CDR prediction tasks is illustrated. **d** The overall Pearson correlation coefficients results for all four methods are depicted. **e** The distribution of Pearson correlation coefficients of 4729 cell lines for the four methods is shown. The center line, along with the bottom and top edges of the box in the violin plots, represents the median, first quartile, and third quartile values, respectively. The boundaries of the whiskers extend to 1.5 times the interquartile range from the upper and lower quartiles. Source data are provided as a Source Data file for **a** and **c**.

utilized lCellVQ to encode the bulk data, obtaining cell representations, and subsequently employed the downstream module DeepCDR[45] for the CDR prediction task. The prediction quality was assessed using Pearson correlation coefficients.

Figure 3c compares the results of CellVQ with other baselines across all cell lines, cancer types, and drug data. These results demonstrate that CellVQ achieves stronger Pearson correlation coefficients than other baselines across the three test sets. Furthermore, Fig. 3d indicates that CellVQ attains the highest average Pearson correlation coefficients of 0.923, compared to 0.780 for DeepCDR, 0.852 for scGPT, and 0.878 for scFoundation. Figure 3e displays the overall distribution of Pearson correlation coefficients for the four methods. CellVQ not only achieves the highest average Pearson correlation coefficients but also maintains stable variance across individual samples. These results provide further evidence that CellVQ possesses significant capability in transferring to bulk RNA-seq data and generalizing to unseen datasets and tasks.

CellVQ enhance gene representation for perturb-seq data

Perturb-seq[46] is a high-throughput experimental technique that integrates CRISPR-based gene perturbation[47] with scRNA-seq. This method enables researchers to systematically investigate the effects of gene perturbations on gene expression at the single-cell level, thereby facilitating a deeper understanding of cellular functions, gene interactions, and biological pathways.

To evaluate whether CellVQ enhances gene-level representation and captures gene co-expression during pretraining, we applied CellVQ to three perturb-seq datasets: Adamson[48], Norman[49], and

Dixit[46], focusing on gene perturbation prediction tasks. We combined CellVQ with an advanced model known as GEARS[50] for predicting single-cell-resolution perturbations. The original GEARS model utilized a Gene Ontology knowledge graph to represent unseen gene perturbations by learning from a combination of previously observed gene perturbation nodes. Additionally, it employed a gene co-expression graph, integrated with perturbation information, to predict post-perturbation gene expression.

We compared CellVQ with GEARS, Linear, GeneCompass, CellFM, scGPT, and scFoundation, all under the same experimental conditions. The evaluation of prediction accuracy was conducted using several metrics: Mean Squared Error (MSE), MSE DE (representing the distance between the top 20 differential gene expressions of the ground truth and corresponding predictions), Pearson correlation coefficients, Pearson correlation coefficients DE, and Pearson delta, which calculates the Pearson correlation between the expression changes of prediction and ground truth. MSE measures the overall distance between predictions and ground truth, while Pearson correlation coefficients and Pearson delta assess the correlation between predicted and observed post-perturbation expression changes.

Figure 4a–d presents the overall performance of CellVQ and other baseline models across three perturb-seq datasets, evaluated using four metrics: MSE, MSE DE, Pearson correlation coefficients, and Recall. The performance metrics for CellVQ on the three datasets are as follows: Adamson (0.032, 0.830, 0.990, 0.850), Norman (0.020, 0.200, 0.993, 0.900), and Dixit (0.013, 0.082, 0.990, 0.970). These results indicate that CellVQ outperforms all other models across all datasets and metrics, demonstrating a clear advantage. Notably, the MSE and MSE DE values for CellVQ on the three datasets are consistently below 0.05 and 0.85, respectively, indicating that CellVQ's predictions are substantially close to the ground truth for both overall genes and the top 20 differentially expressed genes. Furthermore, the Pearson correlation coefficients (PCC) and PCC DE values for CellVQ exceed 0.99 and 0.85, respectively, underscoring the substantial similarity of its predictions to the ground truth. In addition, we observe that the linear model performs comparably to some deep learning models. Notably, on the Dixit dataset, the MSE, MSE DE, PCC, and PCC DE are 0.063, 0.174, 0.82, and 0.74, respectively. These results are comparable to, or even better than, those of GEARS (0.145, 0.315, 0.68, 0.53), GeneCompass (0.084, 0.182, 0.7, 0.63), CellFM (0.07, 0.14, 0.85, 0.74), scGPT (0.076, 0.11, 0.73, 0.70), and scFoundation (0.066, 0.15, 0.83, 0.74). However, on the Adamson and Norman datasets, the linear model achieves performance that is only comparable to or lower than that of GEARS, where these two datasets contain more gene perturbation data. This indicates that while the linear model performs well on smaller datasets, foundation models tend to generalize better on larger-scale datasets.

Figure 4e illustrates the Pearson delta comparison of CellVQ with other baselines across the three datasets. The figures demonstrate that CellVQ predicts a greater number of samples with higher Pearson delta values than the other baselines. The average Pearson deltas for CellVQ are 0.631, 0.451, and 0.444 for the Adamson, Norman, and Dixit datasets, respectively. In comparison, the Pearson deltas for the other baselines are as follows: GEARS (0.050, 0.189, 0.090), CellFM(0.483, 0.376, 0.243) scGPT (0.464, 0.258, 0.236), and scFoundation (0.564, 0.388, 0.256). These results indicate that CellVQ achieves superior prediction performance, not only for individual samples but also in terms of overall Pearson delta, highlighting the effectiveness of CellVQ's gene representation and its capability to capture gene co-expression.

Figure 4 f showcases the expression changes of the top 20 genes following the two-gene perturbation with *LYL1* and *CEBPB*. CellVQ's predictions exhibit closer and more consistent delta values compared to the control for almost all genes, indicating that the individual gene representations produced by CellVQ are robust and informative.

CellVQ enhance gene representation for drug perturbation prediction

Drug perturbation prediction aims to elucidate the effects of specific drugs on cellular gene expression, metabolic activity, and other biological characteristics, with a wide range of practical applications that have propelled advances in medical and pharmaceutical research. By predicting the effects of drugs on cells, researchers can identify new drug targets and anticipate an individual's response to specific medications based on genomic and biomarker profiles, thereby enhancing treatment efficacy.

To evaluate whether CellVQ's gene representation can improve drug perturbation prediction, we applied CellVQ to a single-cell perturbation dataset derived from human peripheral blood mononuclear cells (PBMCs), referred to as OP3[43]. Figure 5a illustrates the pipeline for conducting drug perturbation prediction on the OP3 dataset. We utilized a Uni-Mol v1 pretrained model[51] to encode the drug SMILES data, thereby obtaining the drug representation. Concurrently, we employed CellVQ to derive gene representations from the provided scRNA-seq data. These gene embeddings were then concatenated with the drug representation and input into a multi-layer perceptron to predict differential gene expression. We performed this prediction using various methods, including raw gene expression data, scGPT, and scFoundation, under consistent settings. In addition, to more comprehensive comparison, we also include non-foundation models, like scVIDR[52] and CellOT[53]. The evaluation metrics included MSE, MSE DE, Pearson correlation coefficients(PCC), Pearson correlation coefficients DE(PCC DE), and Pearson delta, with Pearson delta DE serving as an additional evaluation metric.

Figure 5 b presents the overall performance of CellVQ and other baseline methods on the OP3 datasets, evaluated using four metrics. The results indicate that relying solely on raw gene expression data yields suboptimal performance, with MSE and MSE DE values exceeding 1.25 and 4, respectively, while the PCC and PCC DE values fall below 0.66 and 0.46. This suggests that predicting using raw gene expression data does not adequately align with the ground truth. Besides, we noticed that the non-foundation models can achieve comparable or better performance over some foundation models. For instance, the MSE, MSE DE, PCC and PCC DE of CellOT are 0.247, 0.763, 0.901 and 0.782, which are superior than scGPT(0.887, 2.333, 0.853, 0.678) and scFoundation(0.312, 1.163, 0.897, 0.800). However, CellVQ still achieves better performance over other methods. The representation derived from CellVQ reduces MSE and MSE DE to 0.087 and 0.233, respectively, while improving PCC and PCC DE to 0.965 and 0.945. These results demonstrate a significant enhancement in prediction quality, with CellVQ outperforming other robust baselines, such as scVIDR and scFoundation, across all metrics.

Figure 5c compares the Pearson delta and Pearson delta DE of CellVQ with five baseline methods on the OP3 dataset. The average Pearson delta and Pearson delta DE for CellVQ are 0.478 and 0.677, respectively. In comparison, raw gene expression yields values of 0.12 and 0.223, while scGPT reports 0.235 and 0.356, scVIDR is 0.212, 0.343, CellOT reports 0.268, 0.391, and scFoundation yields 0.267 and 0.388. CellVQ exhibits superior performance in both Pearson delta metrics, indicating that its gene representation captures more co-expression information, thereby facilitating more accurate predictions.

Figure 5d showcases the expression changes of the top 20 genes following drug perturbation with Clotrimazole in NK cells. The predictions from CellVQ demonstrate closer and more consistent delta values compared to the control across nearly all genes, indicating that the individual gene representation produced by CellVQ is both robust and informative.

Combining CellVQ and CellVQ-graph on analysis tasks

Figure 6a visualizes the pipeline for integrating CellVQ and CellVQ-Graph. Given the molecular census data of a single cell, CellVQ generates

https://doi.org/10.1038/s41467-026-70071-5

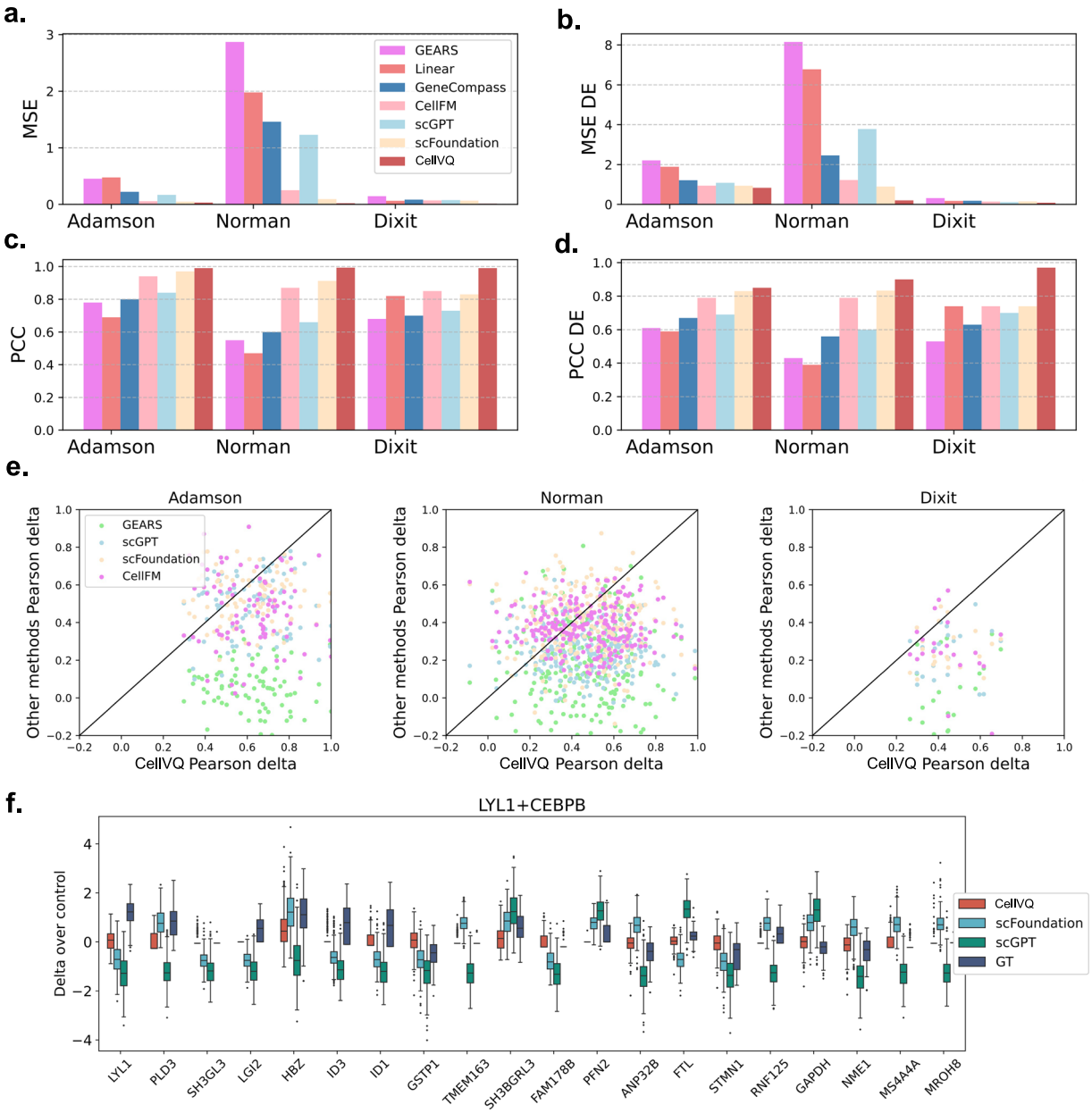

Fig. 4 | Gene Perturbation Prediction Results of CellVQ and Other Baselines. **a**–**d** Present the overall performance of CellVQ and other baseline methods across three perturb-seq datasets, evaluated using four metrics. "PCC" and "DE" denote Pearson correlation coefficients and the top 20 differential gene expressions, respectively. For MSE and MSE DE, lower values indicate better performance, while higher values for PCC and PCC DE reflect improved outcomes. **e** Displays the Pearson delta comparison between CellVQ and the three baseline methods across the three datasets for each gene perturbation. **f** Illustrates the predicted gene expression levels for the top 20 most differentially expressed (DE) genes following a combinatorial perturbation with *LYL1* and *CEBPB*. A total of 300 cells were assessed for each gene. The edges of the box represent the interquartile range, while the horizontal lines within the box indicate the median of all values. The whiskers extend to 1.5 times the interquartile range (IQR) beyond the quartiles. Source data are provided as a Source Data file for (**e**, **f**).

both cell and gene embeddings for CellVQ-Graph. CellVQ-Graph constructs a knowledge graph in which each cell, molecule, and property is represented as a node within a unified framework, connecting all factors associated with the cell. This approach facilitates the embedding of these properties into a shared latent space. Subsequently, CellVQ-Graph employs a graph attention module[54] to update cell representations and compute attention weights for each neighboring node. This module is designed to incorporate multi-modal information into the cell representation, thereby enhancing the interpretability of the model.

Marker gene discover and analysis with CellVQ

Figure 6 b presents the UMAP visualization of raw gene expression and CellVQ embeddings, labeled by cell types in the pancreas dataset. The UMAP representation of raw gene expression fails to distinctly separate the various cell types, particularly for smooth muscle, endothelial, and myeloid cells. In contrast, CellVQ effectively segregates these cells into distinct populations based on their respective cell types, demonstrating the interpretability of the CellVQ representation. We utilized CellVQ to encode the scRNA-seq data from the pancreas dataset and employed max pooling to derive cell embeddings. An importance metric was defined to represent the significance of each gene for the corresponding cell types, calculated as the frequency with which a gene's channel is extracted during max pooling. This frequency was normalized by dividing the number of times each gene is extracted by its maximum value.

https://doi.org/10.1038/s41467-026-70071-5

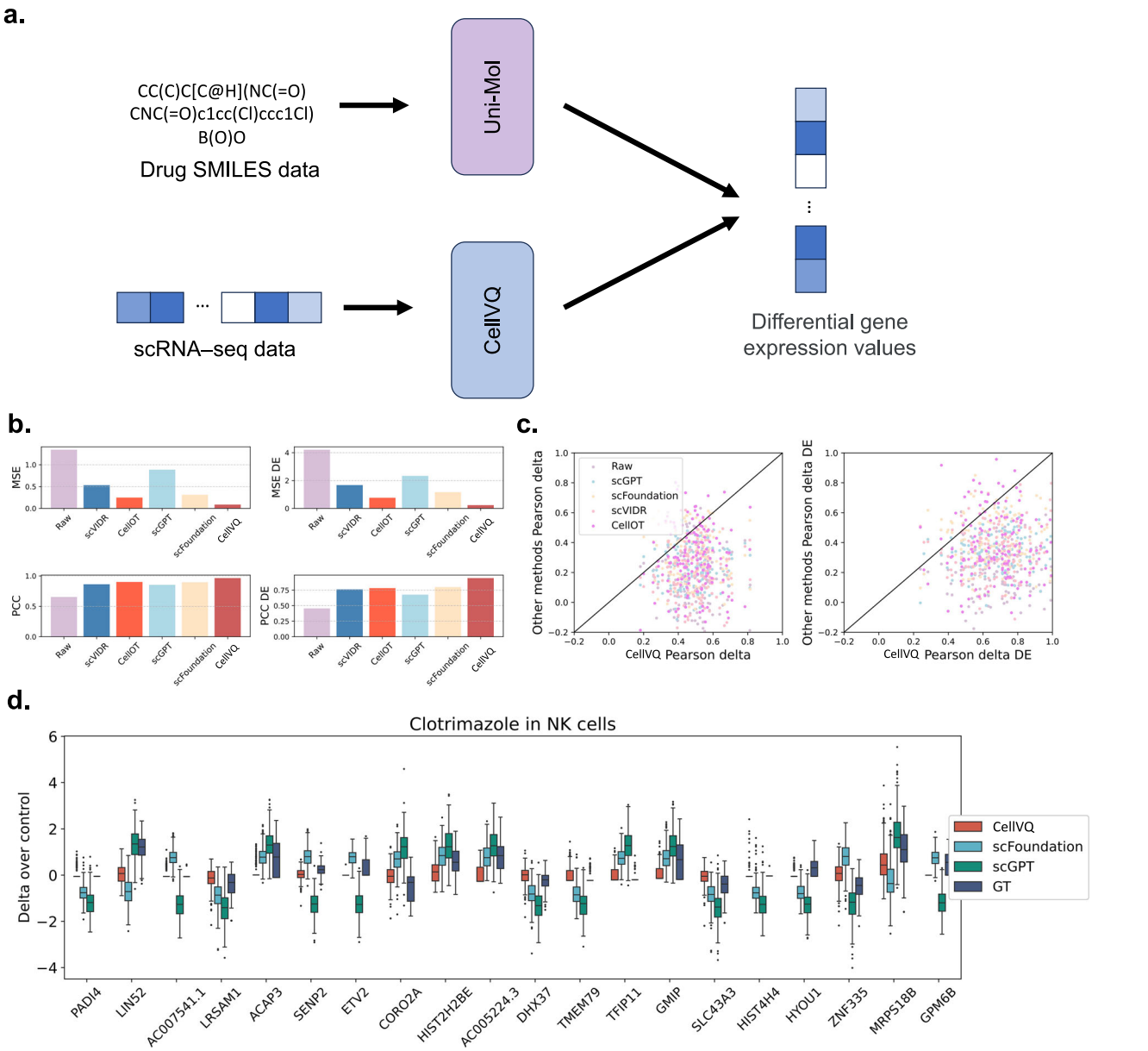

Fig. 5 | Drug Perturbation Prediction Results of CellVQ and Other Baselines. **a** Illustrates the pipeline for the drug perturbation prediction task. **b** Presents the overall performance of CellVQ and other baseline methods across three drug perturbation datasets, evaluated using four metrics. "PCC" and "DE" refer to Pearson correlation coefficients and the top 20 differentially expressed genes, respectively. For MSE and MSE DE, lower values indicate better performance, while higher values for PCC and PCC DE are preferable. **c** Compares the Pearson delta and Pearson delta DE of CellVQ with three baseline methods on the OP3 dataset. **d** Analyzes the predicted gene expression levels for the top 20 most differentially expressed (DE) genes following combinatorial perturbations with LYL1 and CEBPB. A total of 300 cells were assessed for each gene. The edges of the box represent the interquartile range, while the horizontal lines within the box indicate the median of all values. The whiskers extend to 1.5 times the interquartile range (IQR) beyond the quartiles. Source data are provided as a Source Data file for (**c**).

Figure 6c and d illustrate the importance of the top 20 genes in Acinar and Endocrine cells, as predicted by CellVQ. Based on the importance rankings provided by CellVQ, we successfully identified several marker genes for both Acinar and Endocrine cells. For instance, the top 10 genes, including *CEL*[55], *CTRB2*[56], *CPA2*[57], *CRTB1*[56], *SPINK1*[58], *CLPS*[59], and *SYCN*[60], are recognized as marker genes for Acinar cells. These genes are highly expressed in Acinar cells and are primarily involved in the synthesis and secretion of digestive enzymes, lipid metabolism, and cytoprotective mechanisms. Notably, *CTRB2*[56] and *CRTB1*[56] act as precursors for digestive enzymes, while *CPA2*[57] and *CEL*[55] contribute to protein and fat digestion, working synergistically to ensure adequate availability of digestive enzymes in the gut. In the case of Endocrine cells, the top 20 genes identified by CellVQ, including *CHGA*[61], *INS*[62], *FTL*[63], *SST*[64], and *GCG*[65], serve as crucial marker genes. These genes are integral to hormone synthesis, storage, and secretion, thereby influencing metabolic and physiological homeostasis throughout the body. Additionally, genes such as *CADM1*[66] play significant roles in cell aggregation and interactions, contributing to the structural integrity and functionality of pancreatic endocrine cells by promoting intercellular adhesion, which affects hormone secretion and cell growth.

Figure 6 e displays the top 100 gene importance scores for pancreatic ductal cells. CellVQ not only highlights the significance of important genes or marker genes, such as *KRT19*[67], *SOX9*[68],

https://doi.org/10.1038/s41467-026-70071-5

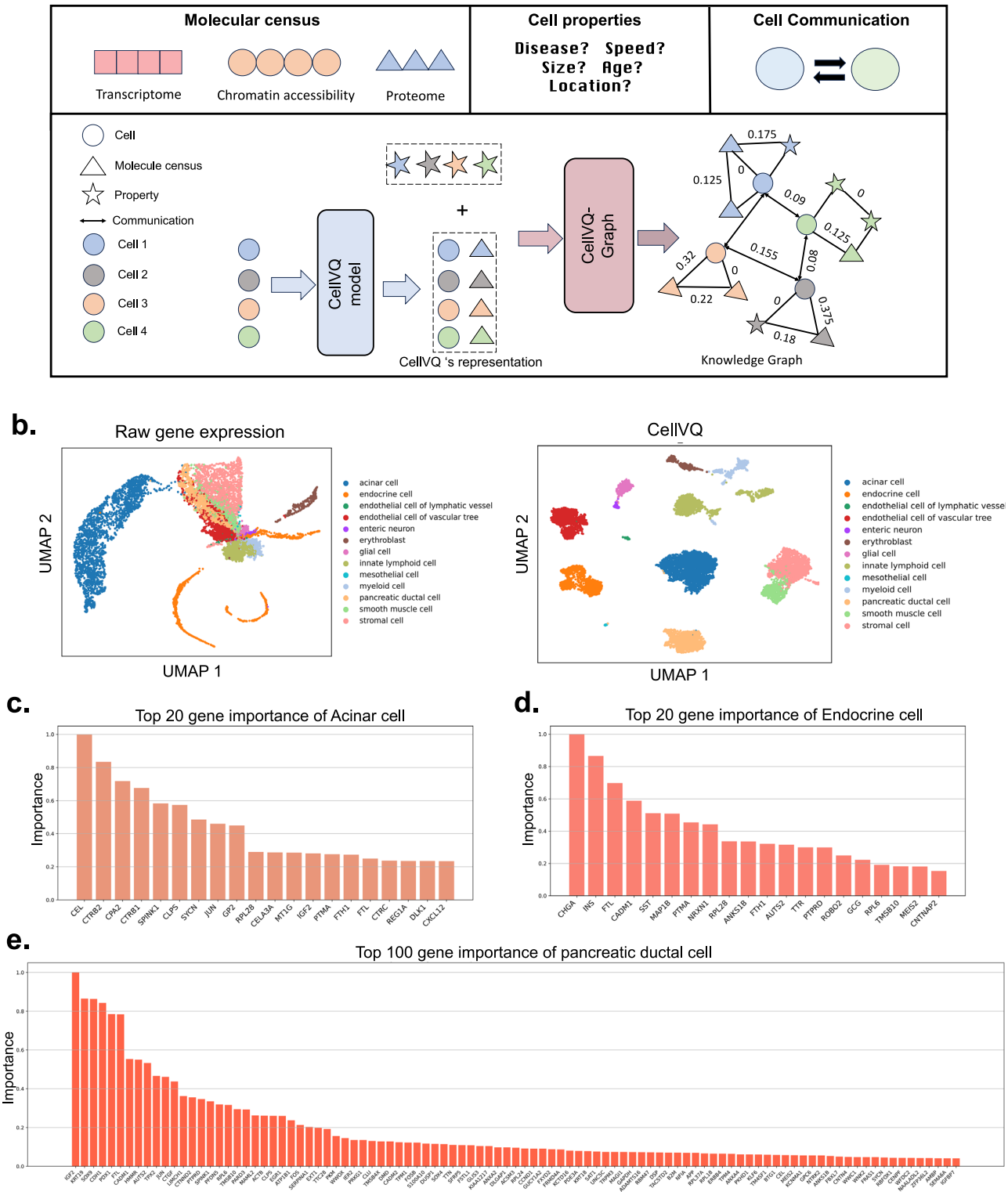

Fig. 6 | The pipeline of combining CellVQ and CellVQ-Graph and the analysis results for marker gene discover. The three up boxes in **a** illustrate three acceptable inputs for CellVQ-Graph, including molecular census, cell properties, and cell communication. The middle box in **a** visualizes the architecture of CellVQ-Graph. **b** The UMAP visualization of raw gene expression and CellVQ with the cell types as labels on the pancreas dataset. **c**, **d** The top 20 gene importance for acinar and endocrine cells. **e** The top 100 gene importance of Myeloid cell.

CDH1[69], and *PDX1*[70], but also captures unique markers for specific subtypes of ductal cells, including *HMMR*[71] and *TPX2*[72]. Furthermore, emphasizes a wide range of genes whose functions are essential for myeloid cells.

Cell state discover and analysis with CellVQ-Graph

To evaluate the efficacy of CellVQ-Graph in analyzing scRNA-seq data for cell state discover, we conducted an experimental analysis using a well-established brain dataset. Fig. 7a, b illustrate the UMAP

https://doi.org/10.1038/s41467-026-70071-5

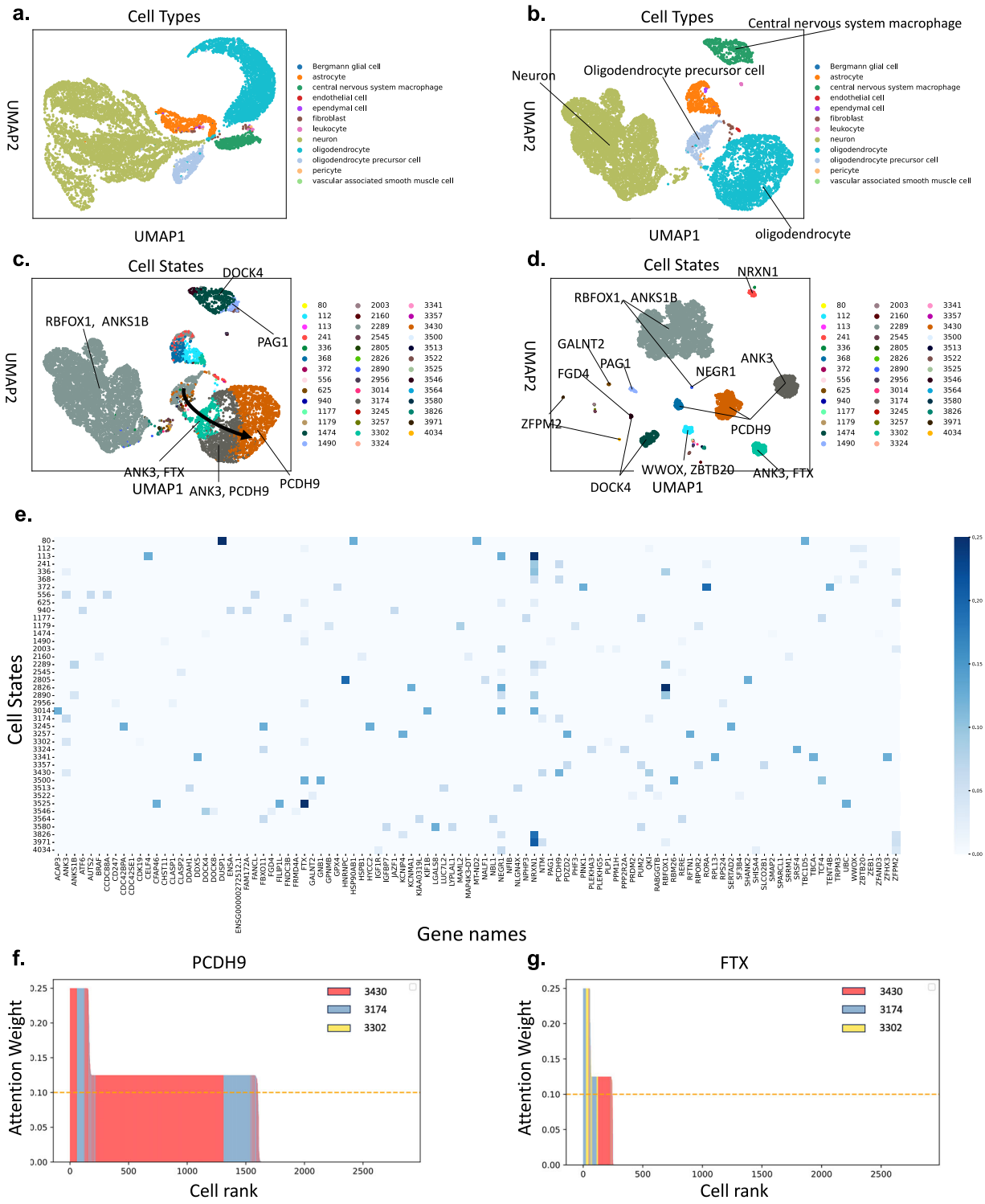

Fig. 7 | Experimental analysis of CellVQ-Graph on the human brain cell dataset. **a** and **b** illustrate example UMAP plots: **a** displays the raw gene expression matrix, while **b** shows the representation provided by CellVQ-Graph, with labels indicating the cell types. **c** presents a UMAP visualization using CellVQ-Graph representation, where the labels correspond to the cell states generated by CellVQ-Graph, and the arrow represents the direction of cell state transition. **d** features a UMAP visualization generated by CellVQ-Graph, displaying vectors that correspond to the respective codes. Finally, **e** provides a heatmap based on the weights of each gene as determined by CellVQ-Graph's graph attention module. The genes in **d** are annotated according to those with high attention weights corresponding to each code shown in **e**. **f** and **g** plot the cell rank of the Attention weight in *PCDH9* and *FTX*, respectively.

visualization of the brain dataset, depicting both raw gene expression and the CellVQ-Graph representation, respectively. The CellVQ-Graph representation distinctly separates major cell categories, such as neurons, astrocytes, central nervous system macrophages, oligodendrocytes, and oligodendrocyte precursor cells, offering clearer differentiation compared to the raw gene expression. Moreover, CellVQ-Graph effectively disentangles smaller cell categories, including leukocytes, endothelial cells, ependymal cells, and astrocytes. We also provide further analytical results of CellVQ-Graph on cell-type-specific datasets, along with ablation studies, in silico perturbation experiments, integration with other foundation models, and a comparison of GRN analysis with CollecTRI[73] in Supplementary Sections 4–8, respectively.

Figure 7c presents the UMAP visualization of the CellVQ-Graph representation, with labels corresponding to the cell states. This representation enables a more precise state identification of cells based on known cell type categories. For Fig. 7d, we obtain the corresponding SCD embeddings of each cell, performing the UMAP visualization. Figure 7e displays a heatmap illustrating the relationship between cells with different cell states and genes, utilizing the average attention weights derived from the graph attention module.

According to the heatmap in Fig. 7e, we identify crucial genes with high attention weights for each cell state, which are annotated in Fig. 7d and annotated some state genes in Fig. 7c for the following analysis examples. We denote these genes as state genes. This figure showcases the UMAP visualization based on the SCD cell embeddings, annotated with these significant genes. Subsequent analyses reveal several notable findings related to cell states.

According to the results from Fig. 7c and d, we argue that CellVQ-Graph's characterization assists biological researchers in swiftly identifying crucial genes essential for understanding cell states. For instance, by comparing Fig. 7b and c, it exhibits that cells categorized under state 2289 predominantly fall within neuron and oligodendrocyte precursor cell (OPC) types, with *RBFOK1* and *ANKS1B* identified as high-attention-weight genes. Notably, both genes are recognized as vital for neuronal function. *RBFOX1*[74] (RNA binding protein, Fox-1 homolog 1) is an RNA-binding protein involved in splicing regulation, playing a critical role in neuronal development and function by regulating neuron-specific gene splicing[74]. *ANKS1B*[75] (ankyrin repeat and SAM domain-containing protein 1B) contributes to cell signaling and intercellular interactions, influencing synapse formation and plasticity, and is associated with various neurological disorders, including autism spectrum disorders, thereby impacting neurodevelopment and function. Furthermore, these genes also provide essential functions for oligodendrocyte precursor cells. In OPCs, *RBFOX1* may influence RNA splicing and post-transcriptional regulation, affecting their proliferation and differentiation. *ANKS1B* is involved in cell signaling and neural development, particularly in interactions with neurons. Consequently, CellVQ-Graph classifies neurons and OPCs under the same cell state, with corresponding proximity in the visualization of Fig. 7c.

The second example involves three distinct states of oligodendrocytes, identified by states 3430, 3302, and 3174 by comparing Fig. 7b, c. We also annotate state genes corresponding to these three states in Fig. 7c by referring to Fig. 7d. In the cells corresponding to state 3430, the *PCDH9*[76] gene exhibits a significant attention weight. *PCDH9*[76] (Protocadherin 9) is recognized as a critical gene in oligodendrocytes, belonging to the protocadherin family of proteins that facilitate cell-cell adhesion and interactions, particularly within the nervous system. The expression of *PCDH9* is closely associated with oligodendrocyte maturation and myelin formation, indicating that cells with state 3430 are likely in a mature oligodendrocyte state[77]. In the cells designated by state 3302, we observe high attention weights for the *ANK3*[78] and *FTX*[79] genes. *FTX*[79] (Ftx RNA) is a long non-coding RNA primarily involved in the nervous system and oligodendrocytes; it may regulate myelin formation as well as cell proliferation and differentiation. *ANK3*[78] (Ankyrin 3) is a critical cytoskeletal protein that plays a central role in the nervous system, contributing to the construction and maintenance of the cytoskeleton. It connects the cell membrane to intracellular structures and is expressed in both neurons and glial cells, facilitating signaling and intercellular interactions with a closer relationship to neurons. The prominence of these two genes in state 3302 suggests that these cells are engaged in myelin formation and neural development, closely associated with neuronal activity[80]. Notably, Fig. 7c shows that state 3302 cells are positioned closer to neurons and oligodendrocyte precursor cells (OPCs) in the UMAP visualization compared to the other two oligodendrocyte states. For cells classified under state 3174, both *PCDH9*[76] and *ANK3*[78] exhibit prominent expression. Given that *ANK3* is a key gene in state 3302 and *PCDH9* is crucial in state 3430, this finding suggests that cells in state 3174 represent a transitional state between states 3302 and 3430. This classification allows us to differentiate a broad group of oligodendrocytes into three distinct and meaningful states. Figure 7f and g provide further validation of our analysis. In terms of the *PCDH9* gene, cells in state 3430 show the highest attention weight, while cells in state 3174 also exhibit a relatively high attention weight, albeit lower than that of state 3430. Conversely, state 3302 cells show minimal attention weight for the *PCDH9* gene. Regarding the *FTX* gene, most cells in state 3302 display relatively high attention weights, while state 3174 cells possess somewhat lower weights. In contrast, state 3430 cells do not exhibit significant attention weights for the *FTX* gene, with many not even considering its weight.

The third example involves two distinct states, 1490 and 1474, within central nervous system macrophages in Fig. 7b and c. In cells classified under state 1474, the *DOCK4*[81] gene shows high attention weight, whereas in state 1490 cells, *PAG1*[82] exhibits a similar prominence. Although both *DOCK4* and *PAG1* are highly expressed in immune cells, they fulfill different functions depending on the context. *PAG1*[82] (Phosphoprotein Associated with Glycosphingolipid-enriched Microdomains 1) is a membrane protein that modulates cellular signaling by regulating tyrosine kinase activity, playing a crucial role in immune cell activation and signaling, particularly within T and B cells. In contrast, *DOCK4*[81] (Dedicator of Cytokinesis 4) regulates small GTPases (e.g., Rac1) and orchestrates cytoskeletal reorganization and cell migration through the activation of these GTPases, thus playing a key role in cell motility, development, and immune responses. The functional divergence between these two genes indicates that the primary roles of the cells associated with the two codes differ, with state 1490 cells actively engaging in immune functions[82], while state 1474 cells are primarily involved in migration[81].

These findings and examples indicate that CellVQ-Graph is capable of learning the intricate relationships and differences between cells in a data-driven manner, effectively identifying discrete cell states using the SCD module. Moreover, the incorporation of the attention module effectively enhances the interpretability of each cell state, highlighting crucial genes that are pivotal for recognizing the characterization of specific cell states. By providing a robust framework for understanding cellular heterogeneity, CellVQ-Graph holds transformative potential for the discovery of cell states and their specific functions that were previously challenging to characterize.

We calculated the Mean Square Distance[83] between the vectors corresponding to different states and visualized this information in the distance confusion matrix presented in Supplementary Fig. 6a. This visualization supports our earlier conclusions. Specifically, states 3302, 3430, and 3174 exhibit the closest distances among all states, while states 1474 and 1490 also display relatively close distances to one another. Notably, the distance between states 3174 and 3302 is closer

https://doi.org/10.1038/s41467-026-70071-5

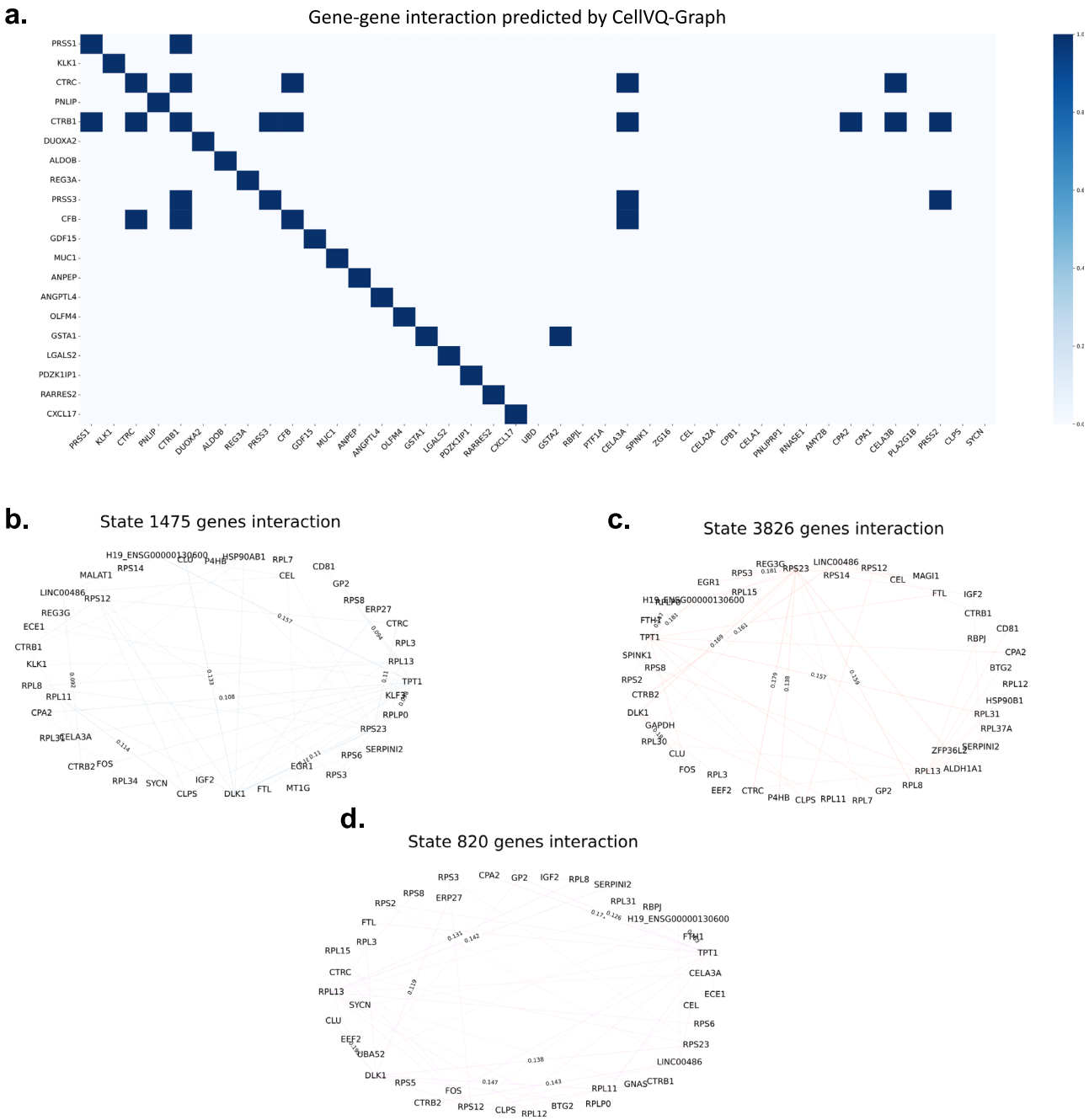

Fig. 8 | The gene regulatory network analysis by CellVQ and CellVQ-Graph. a visualize the gene-gene interaction predicted by the CellVQ model. **b**–**d** plot the state 1475, 3826, and 820 gene-gene interaction generated by CellVQ-Graph, respectively.

than that between either of these states and 3430, corroborating the findings illustrated in Fig. 7c. We posit that this characterization provided by CellVQ-Graph offers a more detailed and precise quantization method for measuring the distances between cell states.

Gene Regulatory Network Analysis with CellVQ-Graph

Studying gene-gene interactions is a critical step in understanding GRN[26]. GRNs are complex networks in which genes interact through regulatory elements, such as transcription factors, regulatory RNAs, and signaling pathways. These networks regulate gene expression levels and timing, thereby influencing cellular function and behavior. Investigating GRNs is essential for elucidating biological processes, understanding disease mechanisms, and developing therapeutic strategies.

We propose that CellVQ-Graph serves not only as a valuable tool for studying gene interactions through its graph attention module but also visualizes these interactions at the cell state level via the CellVQ SCD module. To validate this, we applied the CellVQ-Graph model to the human pancreas cell dataset, focusing on Acinar cells.

Figure 8a displays the gene-gene interactions predicted by the CellVQ-Graph for an acinar cell. We encoded the genes from the acinar cell and obtained their corresponding codes by SCD module, as SCD module can also embed the gene embeddings into a code. Two genes are defined as connected if they share the same code. We selected highly expressed genes in acinar cells to visualize their interactions, as shown in Fig. 8a. The gene regulatory network predicted by CellVQ-Graph is biologically meaningful. For instance, CellVQ-Graph predicts an interaction between the *PRSS1* and *CTRB1* genes. *PRSS1* encodes

trypsinogen, a major digestive enzyme in the pancreas, while *CTRB1* encodes a trypsin inhibitor that regulates the activity of pancreatic digestive enzymes and prevents their premature activation in alveolar cells. The expression and function of these genes are interdependent; overexpression of *PRSS1* may result in insufficient inhibition by *CTRB1*, thereby increasing the risk of cellular injury. The importance of these genes in the Gene Ontology database is rated at 0.5, indicating a significant biological relevance.

Another example is the interaction between *CTRC* and *CFB* genes. *CTRC* encodes a regulator of chymotrypsinogen, a pancreatic protease involved in the activation and regulation of digestive enzymes. *CFB* encodes complement factor B, which plays a role in the immune response and may influence the inflammatory response in the pancreas. In pathological states such as pancreatitis, *CTRC* expression may be modulated, while *CFB* activity can affect the extent of the inflammatory response. Therefore, the interaction between these genes may contribute to the response and repair mechanisms of the pancreas. Their importance in the Gene Ontology database is rated at 0.363, indicating a critical co-expression relationship. These examples demonstrate that the CellVQ-Graph effectively captures the co-expression relationships of genes, serving as a valuable analytical tool for discovering gene regulatory networks in individual cells.

Figure 8b–d illustrates the gene interactions across three primary states. Fig. 8b presents the overall gene-gene interactions within state 1475 cells. Notably, we found several genes that exhibit interactions with numerous other genes. The first example is the *TPT1* gene[84], also known as Tumor Protein 1 or *TCTP* (Translationally Controlled Tumor Protein). *TPT1* encodes a widely expressed intracellular protein involved in various biological processes, including cell proliferation, survival, and stress response. It interacts with genes that regulate the cell cycle and influence proliferation rates[84]. *TPT1* is connected to ribosomal protein family genes, such as *RPL31*, *RPLP0*, and *RPS8*, which are essential components of the ribosome and play critical roles in protein synthesis, displaying high attention weights[84]. Furthermore, TPT1 exhibits a significant attention weight of 0.157 in relation to the *H19_ENSG00000130600*[85] gene. This non-coding RNA gene, located on human chromosome 11, is crucial for embryonic development and cellular proliferation and is hypothesized to be part of an imprinted gene[85]. Thus, *H19_ENSG00000130600* and *TPT1* are interconnected in processes related to cell proliferation, tumorigenesis, and stress response.

The second example focuses on the *DLK1* gene[86]. *DLK1* (Delta-like 1) encodes a cell surface protein that is crucial for embryonic development, stem cell fate determination, and cell differentiation. It is closely associated with the Notch signaling pathway, growth factors, transcription factors, and tumor-related genes[86]. As shown in Fig. 8b, *DLK1* exhibits a high attention weight of 0.133 in relation to the *CLU* gene. The *CLU* gene (Clusterin) encodes a widely expressed glycoprotein that primarily mediates cellular stress responses, apoptosis, and immune responses[87]. *CLU* plays a significant role in various physiological and pathological processes, including neuroprotection and tumorigenesis[87]. Both *CLU* and *DLK1* are implicated in the cellular response to environmental stress and are aberrantly expressed in numerous cancers, linking them to tumorigenesis and progression. Their interaction may involve the regulation of cell proliferation and apoptosis, with *DLK1* potentially modulating the expression or function of *CLU* via the Notch signaling pathway, suggesting a mutually dependent relationship in cellular processes. Consequently, CellVQ-Graph assigns a high attention weight to the interaction between these two genes. Additionally, *DLK1* is connected to several ribosomal protein family genes, including *RPS23*, *RPL11*, and *RPLP0*, which are integral to cell development and differentiation[86].

The final example pertains to the intrinsic interactions among ribosomal protein family genes. For instance, genes such as *RPS23*, *RPL8*, *RPL13*, and *RPLP0* are linked within the CellVQ-Graph framework and exhibit high attention weights. This suggests that these ribosomal protein genes work together to form ribosomal complexes and influence each other's expression, leading to increased attention weights.

Figure 8c illustrates the gene-gene interactions in state 3826 cells. We observe that the gene-gene interactions in state 3826 cells do not significantly differ from those in state 1475 cells, suggesting that the gene-gene interactions within specific cell types remain relatively stable. However, we note that ribosomal protein family genes, such as *RPS23* and *RPL13*, demonstrate more pronounced interactions with other genes in this state. This phenomenon is likely linked to the active differentiation or developmental processes characterizing state 3826 cells.

Figure 8d presents the gene-gene interactions in state 820 maintain a similar pattern to those observed in the other two states. This further supports our earlier conclusion that, despite differences in cell states, the gene-gene interactions within a specific cell type adhere to gene regulatory network principles[88].

The gene regulatory network analysis conducted using CellVQ-Graph reveals its capability to capture gene-level knowledge and connections, identifying crucial regulatory genes such as *DLK1* and *TPT1*. Moreover, CellVQ-Graph provides interaction analysis results at the cell state level, allowing users to explore not only state-gene interactions but also to uncover gene-gene interactions. This dual capability further enhances our understanding of cellular mechanisms.

Multi-modal analysis with CellVQ-Graph

For cell state identification, manual annotations such as cell types and tissue characteristics are critical informational factors that contribute to the construction of cell states. However, prior research has predominantly concentrated on analyzing scRNA-seq or scATAC-seq data, often neglecting the influence of these properties. To address this gap, we introduce a scalable approach that employs a language tokenizer which introduced by BioGPT[89]. It include 42,383 tokens and can convert the text of single cell properties describtion into one or a seires of tokens, effectively embedding these properties within a unified graph framework. This enables a comprehensive examination of how these annotated properties influence cell states.

To evaluate the significance of these properties, we validated CellVQ-Graph using a sampled pretraining dataset comprising 10,000 cells, categorized into 167 cell types, sourced from 30 different tissues and involving 465 donors. This dataset enables us to consider various properties, including tissue type, disease state, age, and sex, and to investigate their respective influences on cell states.

Figure 9a–d depict the distribution of attention weights for Tissue, Disease, Age, and Sex across nine distinct states. Our analysis reveals that tissue annotations are the most influential factors among these properties, with attention weights ranging from 0 to 0.25 and an average attention weight of 0.12 across the states. Additionally, Fig. 9e illustrates a strong correlation between cell distribution and tissue types, indicating that cells sharing the same tissue tend to cluster together in specific areas of the visualization.

Following tissue annotations, disease annotations represent the second most significant factor influencing cell states. While one might intuitively expect disease to play a major role, the predominance of normal cells within our pretraining dataset (as depicted in Fig. 9f) diminishes the perceived importance of disease, resulting in average attention weights that span from 0 to 0.14. Furthermore, the attention weights associated with age and sex are comparatively lower, averaging only 0.03 and 0.01, respectively. Figure 9g and h provide UMAP visualizations of CellVQ-Graph representations labeled by Age and Sex, revealing a relatively random distribution. This observation further substantiates the minimal contribution of age and sex to the construction of cell states.

In summary, CellVQ-Graph offers a valuable tool for biological researchers, enabling the analysis of multimodal properties and facilitating the identification of crucial factors that contribute to cell states.

https://doi.org/10.1038/s41467-026-70071-5

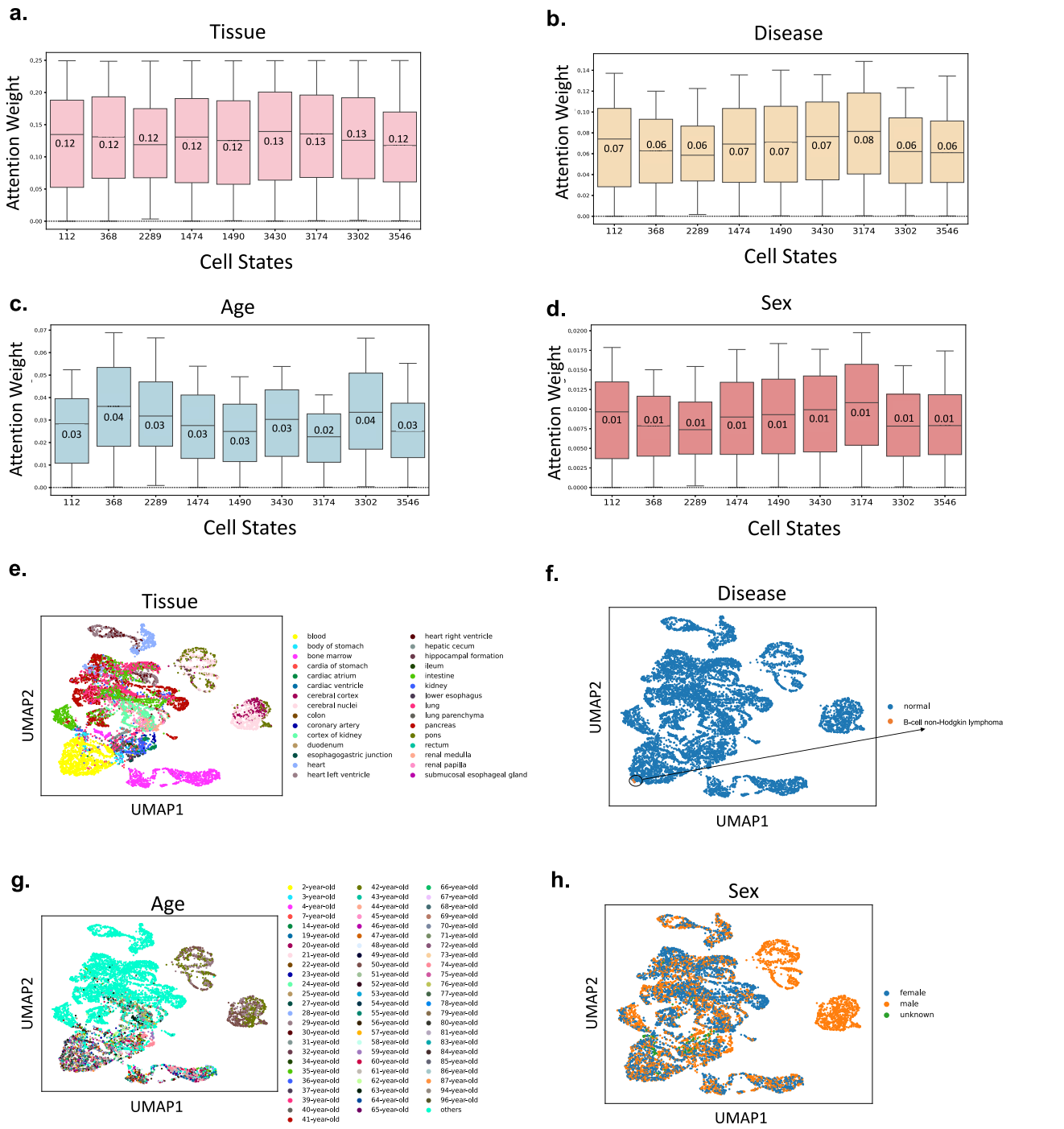

Fig. 9 | Properties analysis of CellVQ-Graph on a sampled pretraining dataset with 10,000 cells. a–d visualize attention weight distribution of Tissue, Disease, Age and Sex for nine states. The center line, along with the bottom and top edges of the box, represents the median, first quartile, and third quartile values, respectively. The boundaries of the whiskers extend to 1.5 times the interquartile range from the upper and lower quartiles. **e–h** plot the UMAP visualization of CellVQ-Graph representation with Tissue, Disease, Age and Sex as labels.

By incorporating these annotations, CellVQ-Graph enhances our understanding of the complexities underlying cellular behavior and state identification.

Discussion

In this paper, we address the significant limitations of current single-cell foundation models, namely their restricted generalizability for heterogeneous and sparse data, as well as their lack of interpretability. These limitations severely hinder the application of single-cell foundation models in real-world scenarios. We introduce our models, termed CellVQ and CellVQ-Graph. CellVQ is a single-cell foundation model comprising 513 million parameters, integrating a transformer encoder, a vector quantization module, and a performer decoder. It is pretrained on 68 million single-cell data points across multiple pre-training tasks. CellVQ-Graph serves as a lightweight and plug-and-play algorithm for CellVQ, effectively leveraging its representations for biological analysis and discovery within single-cell multi-omic and multi-modal datasets.

We applied CellVQ and CellVQ-Graph across various tasks, categorized into prediction and analysis. In the domain of prediction tasks,

which include cell clustering, annotation, property prediction, gene and drug perturbation prediction, as well as drug response prediction, CellVQ demonstrates significant advancements over other foundation models, such as scGPT and scFoundation.

For analysis tasks, we integrated CellVQ and CellVQ-Graph into four distinct objectives: cell state discovery, gene discovery, GRN analysis, and multimodal analysis. These analyses underscore the unique advantages of our models in terms of interpretability. We assert that the introduction of CellVQ and CellVQ-Graph effectively bridges the gap between AI models and biological research, thereby enhancing our understanding of cellular and genetic mechanisms while facilitating drug discovery.

Looking ahead, we aim to expand CellVQ to support multi-omics pretraining. Currently, due to limitations in data scale, CellVQ has only been pretrained on a large-scale scRNA-seq dataset and fine-tuned on a scATAC-seq dataset. In the future, we plan to establish a comprehensive dataset that integrates scRNA-seq, scATAC-seq, and multiple annotations for each single cell. Subsequently, we will design a model that can effectively leverage all modal data to achieve improved representational capacity and enhanced interpretability.

Methods

Preprocess

To preprocess gene expression and scATAC data, we begin by applying normalization and log1p transformation. These steps are essential for addressing scale disparities and mitigating the impact of missing values within the expression matrix. For the property data, which are typically represented as textual annotations, we utilize the BioGPT tokenizer[89] to convert these textual properties into a sequence of integers. This method enables us to leverage advanced language modeling techniques, allowing us to effectively represent cellular properties as embeddings.

CellVQ model architecture

Input embeddings. We utilize an embedding module, which consists of three components: gene expression, gene position (gene id), and gene ontology embeddings, to embed the normalized gene expression values. Given a series of gene expression values $E = \{e_1, e_2, e_3 \ldots e_n\}$ where $e_i \in \mathbb{R}$, and a series of gene position indexes $P_{\text{input}} = \{p_1, p_2, p_3 \ldots p_n\}$ where $p_i \in \mathbb{N}$. We utilize a gene expression embedding module f_E and a gene position embedding module f_P to embed these two inputs:

$$E_{input} = f_E(e_1, e_2, e_3 \ldots e_n) \; where \; E_{input} \in R^{n \times d} \tag{1}$$

$$P_{input} = f_P(p_1, p_2, p_3 \ldots p_n) \; where \; P_{input} \in R^{n \times d} \tag{2}$$

where d is the embedding dimension of the model input.

Besides, we utilize the prior gene ontology knowledge to construct a symmetric importance matrix A, where $A \in R^{n,n}$ and $A_{i,j}$ represents the importance of gene i for gene j. We utilize a learnable matrix $W_A \in R^{n,d}$ to embed the gene ontology matrix:

$$A_{input} = AW_A \; where \; A_{input} \in R^{n,d} \tag{3}$$

Notably, we make the gene ontology matrix A learnable, meaning the gene ontology updates during training. Subsequently, we add these three embeddings as the model input:

$$X_{input} = E_{input} + P_{input} + A_{input} \tag{4}$$

Encoder. To refine the gene features to a greater extent, we utilize multiple transformer layers to incorporate the information of expressed genes. Each transformer layer consists of two modules: the Self-Attention Mechanism and the Feed-Forward Neural Network.

Self-Attention Mechanism Given the encoder input $X_{input} = \{x_1, x_2, x_3 \ldots x_n\}$7D1$where X_{input} \in R^{n \times d}$, the self-attention operation computes attention scores between all pairs of elements using query (Q), key (K), and value (V) matrices. These matrices are linear transformations of the input embeddings:

$$Q = X_{input} W_Q, \; K = X_{input} W_K, \; V = X_{input} W_V. \tag{5}$$

The attention weights are calculated as:

$$\widehat{X} = \text{softmax}\left(\frac{QK^T}{\sqrt{D}}\right) V. \tag{6}$$

This mechanism allows each gene to capture information from other genes.

Feed-Forward Neural Network After getting the output of Self-Attention Mechanism, we apply a Feed-Forward Neural Network $\mathcal{F}$, which is a multi-layer perception function. What's more, we add a residue connection:

$$X_{output} = \mathcal{F}(\widehat{X}) + X_{input}. \tag{7}$$

We use the X_{output} as the next transformer layer's input to get the final encoder output. Finally, we use a max polling to obtain the cell embeddings:

$$X_{cell} = MaxPooling(X_{output}). \tag{8}$$

The max pooling can capture the significant feature among all genes for each dimension.

Single-cell discretization module. Following the Foldtoken[90], we adopt the Vector Quantization method called SoftCVQ, which enables the transformation of continuous embeddings into discrete codes. Unlike traditional approaches that directly map embeddings to tokens, SoftCVQ employs a more nuanced mechanism to capture the relationships between embeddings. The process involves two key steps: (1) mapping pre-defined binary vectors (b_j) to continuous token embeddings (v_j) and (2) performing soft alignment between these token embeddings and latent embeddings (h_s).

Given a decimal integer z and a codebook size m, the integer z is represented as a binary vector b_j with a length of $\log_2(m)$. For example, if $m = 4$, the binary vectors are $b_1 = [0, 0]$, $b_2 = [0, 1]$, $b_3 = [1, 0]$, and $b_4 = [1, 1]$. The quantization operation converts continuous embeddings h_s into discrete tokens z through the following process:

$$\begin{cases} a_{sj} = \frac{\exp(f_s^T v_j / T)}{\sum_{j=1}^{m} \exp(f_s^T v_j / T)} \\ z = \arg\max_j a_{sj} \end{cases} \tag{9}$$

where v_j represents the j-th token embedding, which also serves as the de-quantization component. During training, the binary vector b_s is derived using bitwise operations:

$$\begin{cases} b_s = \text{Bit}(z, \log_2(m)) \\ \widehat{f}_s = \text{ConditionNet}(b_s) \end{cases} \tag{10}$$

The temperature parameter T in Equation (9) controls the softness of the attention mechanism. When T is large, the attention weights approximate uniformity, whereas smaller T values result in weights closer to one-hot distributions. To stabilize training, T is gradually reduced from 1.0 to 10^{-8}. The conditional network, ConditionNet, is a multi-layer perceptron (MLP) that maps high-dimensional binary

vectors to latent embeddings of dimension d. For a codebook size of 2^{16}, this MLP transforms 65536 16-dimensional binary vectors into 65536 d-dimensional embeddings.

The core process of the aforementioned discretization encoding involves mapping continuous biomolecular data to a finite discrete binary encoding space, achieving efficient feature compression. This universal single-cell compilation method based on binary discrete encoding ensures that the model provides uniformly granular discrete representations for different genes and single-cell batches. It fully captures key biomolecular features, enabling standardized representation of heterogeneous data. Through a unified codebook, heterogeneous data can be compared, analyzed, and modeled within the same space, enabling seamless integration of data across different batches. We provide the discussion about how SCD module resolve the batch effect in Supplementary Section 11.

Decoder. After getting the encoder output, we concatenate the non-expressed, non-sampled gene embeddings, the encoder output, and the SCD Embeddings as the decoder input.

$$X_{dec} = [X_{ne}, X_{ns}, X_{SCD\_gene}, X_{SCD\_cell}]. \tag{11}$$

Given that the decoder input consists of nearly 20,000 genes—too many for a standard Transformer to handle—we opted to use a Performer as the decoder. The Performer is a Transformer-based variant designed to enhance the computational efficiency of the self-attention mechanism. It achieves this goal by utilizing kernelized attention, which reduces computational complexity and enhances efficiency when handling long sequences. Given the decoder input X_{dec}, a performer layer perform:

$$Q = X_{dec}W_Q,\ K = X_{dec}W_K,\ V = X_{dec}W_V. \tag{12}$$

$$\hat{X}_{dec} = \sum_{j=1}^{n} \phi(Q, K)V \tag{13}$$

where ϕ is a kernel function, which is used for approximating the attention matrix. Then, we also perform a Feed-Forward Neural Network $\mathcal{F}$

$$X_{dec\_output} = \mathcal{F}(\hat{X}_{dec}), \tag{14}$$

After passing through six performer layers, we get the decoder outputs.

ZINB predictor. To predict the Zero-Inflated Negative Binomial, we deploy a ZINB predictor.

Given input features X_{dec_output}, the model output can be represented as:

$$Z \sim \text{Bernoulli}(\pi) \tag{15}$$

where π is the probability of the zero-inflated part, typically predicted by logistic regression:

$$\pi = \sigma(W_\pi X_{dec\_output} + b_\pi) \tag{16}$$

Next, the parameters of the negative binomial distribution r and p are predicted by another regression model:

$$Y \sim \text{NB}(r, p) \tag{17}$$

$$r = f_r(X),\ p = f_p(X_{dec\_output}) \tag{18}$$

where the NB denotes the negative binomial distribution. The final ZINB output can be expressed as:

$$Y = \begin{cases} 0 & \text{with probability } \pi \\ \text{NB}(r, p) & \text{with probability } (1-\pi) \end{cases} \tag{19}$$

ZINB loss funtion

We utilize the ZINB distribution predicted by the ZINB predictor and the ground truth to calculate the ZINB loss to train our models. Notably, we select the genes that are not sampled for the encoder input and a number of non-expressed genes, where the number of non-expressed genes is less than or equal to the non-sampled genes, to avoid the unbalanced distribution. Given observed data Y and predicted parameters π, r, p, the ZINB loss can be expressed as:

$$\mathcal{L}(Y, \pi, r, p) = -\sum_{i=1}^{N} \Big[Y_i \log(1-\pi_i) + (1-Z_i) \Big(\log(\pi_i) + \log\Big(\frac{\Gamma(Y_i + r)}{\Gamma(r)\Gamma(Y_i+1)}\Big) \tag{20}$$

$$+ \frac{r}{p}\log\Big(\frac{p}{p+1}\Big) + \frac{Y_i}{p+1}\log\Big(\frac{1}{p+1}\Big)\Big)\Big] \tag{21}$$

Where: - Y_i is the observed count for the i-th observation. - Z_i is an indicator variable that is 1 if $Y_i = 0$ (zero-inflated), and 0 otherwise. - Γ is the gamma function.

How ZINB predictor and loss function overcome the data sparsity? As prior works supported[16,91], ZINB predictor effectively addresses data sparsity by combining a zero-inflation model with a negative binomial distribution. It first models additional zeros through the zero-inflation component, enabling the model to capture the high proportion of zeros in real data rather than treating them solely as outcomes of the negative binomial distribution. The negative binomial component models the variability and discreteness of non-zero count data, providing a more accurate description of sparse data. This combination makes ZINB particularly well-suited for handling highly sparse data in fields such as gene expression, thereby enhancing the model's fitting capability and predictive accuracy.

CellVQ-graph model architecture

Graph construction. To identify the neighboring cells, we compute the mean squared distance between the gene expression profiles of a target cell and other cells within the same batch. We then select the top 15 closest cells based on this distance as its neighbors. To optimize computational efficiency, we first reduce the dimensionality of our analysis by focusing on the top 500 highly expressed genes, determined through covariance analysis of the expression data.

Furthermore, we conceptualize each gene and property as nodes within a network. For each cell, we isolate the top 200 genes based on expression levels, alongside chromatin peaks characterized by the highest 1000 peak values. This strategy establishes connections between the cellular representations and the most expressive genes and chromatin regions, thereby enhancing our model's ability to capture relevant biological interactions.

Graph attention module. We introduce the graph attentional layer[54] designed for modeling relationships in graph-structured data, which is a fundamental component in various graph-based machine learning models. This layer is versatile and can be integrated into different Graph Attention Network (GAT) architectures, offering flexibility without assuming any specific attention mechanism.

Input and Transformation: The layer processes a set of node features $\mathbf{h} = \{h_1, h_2, \ldots, h_N\}$, where each node i has features $h_i \in \mathbb{R}^F$. The transformation begins with a shared linear layer using a weight matrix $\mathbf{W} \in \mathbb{R}^{F' \times F}$, projecting each node's features to a new space $\mathbf{W}h_i \in \mathbb{R}^{F'}$.

Graph Attention Mechanism: The core of the layer is the attention mechanism, which computes coefficients e_{ij} representing the importance of node *j*'s features for node *i*:

$$e_{ij} = a(\mathbf{W}h_i, \mathbf{W}h_j) \tag{22}$$

These coefficients are normalized using the softmax function to ensure they sum to 1 for each node (i):

$$\alpha_{ij} = \mathrm{softmax}_j(e_{ij}) = \frac{\exp(e_{ij})}{\sum_{k \in \mathcal{N}_i} \exp(e_{ik})}. \tag{23}$$

where $\mathcal{N}_i$ denotes node *i*'s neighborhood.

Feature Transformation: The new features for each node are computed as a weighted sum of its neighbors' transformed features:

$$h_i = \sigma\left(\sum_{j \in \mathcal{N}_i} \alpha_{ij} \mathbf{W}h_j\right). \tag{24}$$

Here, σ introduces nonlinearity, enhancing the model's ability to capture complex patterns.

Multi-Head Attention: To enhance stability and model capacity, the layer employs multi-head attention. (K) independent attention mechanisms compute their own coefficients α_{ij}^k and apply their own transformations with weight matrices $\mathbf{W}^k$. The outputs are concatenated:

$$h_i = \Big\Vert_{k=1}^{K} \sigma\left(\sum_{j \in \mathcal{N}_i} \alpha_{ij}^k \mathbf{W}^k h_j\right) \tag{25}$$

This increases the model's capacity by capturing diverse aspects of the data.

Final Layer Considerations: When applied to the final layer of a network, concatenation may not be optimal. Instead, averaging across heads is preferred:

$$h_i' = \sigma\left(\frac{1}{K}\sum_{k=1}^{K}\sum_{j \in \mathcal{N}_i} \alpha_{ij}^k \mathbf{W}^k h_j\right) \tag{26}$$

This approach prevents over-reliance on any single head, improving generalization.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The pretraining dataset can be assessed in (https://github.com/bowang-lab/scGPT/blob/main/data/cellxgene/). We provided all test datasets used in this paper in https://doi.org/10.6084/m9.figshare.28787489.v2, including Pancreas dataset (https://github.com/bowang-lab/scGPT/blob/main/data/cellxgene/), Zheng68K dataset (https://www.dropbox.com/sh/w3yg2nucnng5v1u/AAAM8Ym_KU9XF4z51RT81eNEa?dl=0), Segerstolpe dataset (https://zenodo.org/records/3357167), CDR dataset (https://github.com/kimmo1019/DeepCDR), CDR dataset (https://github.com/kimmo1019/DeepCDR), Single cell drug response classification dataset (https://github.com/CompBioT/SCAD), Perturbation dataset (https://github.com/snap-stanford/GEARS) and the drug perturbation dataset (https://www.kaggle.com/competitions/open-problems-multimodal/data). Source data are provided with this paper in https://doi.org/10.6084/m9.figshare.30294007[92], respectively. Source data are provided with this paper.

Code availability

We provide the source code in (https://github.com/A4Bio/CellVQ/) under MIT license. This repository also contains instructions and a tutorial for applying CellVQ for multiple downstream tasks. The specific version of the code associated with this publication is archived in Zenodo and is accessible via https://doi.org/10.5281/zenodo.18145133[92].

References

1. Alison, M. R., Poulsom, R., Forbes, S. & Wright, N. A. An introduction to stem cells. *J. Pathol.: A J. Pathological Soc. Gt. Br. Irel.* **197**, 419–423 (2002).
2. Woese, C. R. On the evolution of cells. *Proc. Natl. Acad. Sci.* **99**, 8742–8747 (2002).
3. Rafelski, S. M. & Theriot, J. A. Establishing a conceptual framework for holistic cell states and state transitions. *Cell* **187**, 2633–2651 (2024).
4. Atlas, H. C. et al. Human Cell Atlas. *Human cell atlas sequences first 250K developmental cells* (2018).
5. Kharchenko, P. V. The triumphs and limitations of computational methods for scRNA-seq. *Nat. methods* **18**, 723–732 (2021).
6. Chen, G., Ning, B. & Shi, T. Single-cell RNA-seq technologies and related computational data analysis. *Front. Genet.* **10**, 317 (2019).
7. Li, Z. et al. scrna-seq of ovarian follicle granulosa cells from different fertility goats reveals distinct expression patterns. *Reprod. Domest. Anim.* **56**, 801–811 (2021).
8. Yan, Y. et al. Advances in single-cell transcriptomics in animal research. *J. Anim. Sci. Biotechnol.* **15**, 102 (2024).
9. Liu, C. et al. Patient-derived tumor organoids combined with function-associated scRNA-seq for dissecting the local immune response of lung cancer. *Adv. Sci.* 2400185 (2024).
10. Lu, C. et al. Application of single-cell assay for transposase-accessible chromatin with high throughput sequencing in plant science: Advances, technical challenges, and prospects. *Int. J. Mol. Sci.* **25**, 1479 (2024).
11. Baek, S. & Lee, I. Single-cell atac sequencing analysis: from data preprocessing to hypothesis generation. *Comput. Struct. Biotechnol. J.* **18**, 1429–1439 (2020).
12. Yang, F. et al. scbert as a large-scale pretrained deep language model for cell type annotation of single-cell rna-seq data. *Nat. Mach. Intell.* **4**, 852–866 (2022).
13. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 4171–4186 (2019).
14. Cui, H. et al. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature Methods* 1–11 (2024).
15. Hao, M. et al. Large-scale foundation model on single-cell transcriptomics. *Nature Methods* 1–11 (2024).
16. Kalfon, J., Samaran, J., Peyré, G. & Cantini, L. scprint: pre-training on 50 million cells allows robust gene network predictions. *Nat. Commun.* **16**, 3607 (2025).
17. Kedzierska, K. Z., Crawford, L., Amini, A. P. & Lu, A. X. Zero-shot evaluation reveals limitations of single-cell foundation models. *Genome Biol.* **26**, 101 (2025).
18. Boiarsky, R. et al. Deeper evaluation of a single-cell foundation model. *Nat. Mach. Intell.* **6**, 1443–1446 (2024).
19. Zhuang, J. et al. A streamlined scrna-seq data analysis framework based on improved sparse subspace clustering. *IEEE Access* **9**, 9719–9727 (2021).

20. Yau, K. K., Wang, K. & Lee, A. H. Zero-inflated negative binomial mixed regression modeling of over-dispersed count data with extra zeros. *Biometrical J.: J. Math. methods Biosci.* **45**, 437–452 (2003).
21. Weng, J., Ge, Y. E. & Han, H. Evaluation of shipping accident casualties using zero-inflated negative binomial regression technique. *J. Navigation* **69**, 433–448 (2016).
22. Su, J. et al. Saprot: Protein language modeling with structure-aware vocabulary. *OpenReview* https://openreview.net/forum?id=6MRm3G4NiU (2023).
23. van Kempen, M. et al. Foldseek: fast and accurate protein structure search. *Biorxiv* https://doi.org/10.1101/2022.02.07.479398 2022–02 (2022).
24. Li, X. et al. Deep learning enables accurate clustering with batch effect removal in single-cell RNA-seq analysis. *Nat. Commun.* **11**, 2338 (2020).
25. Schlitt, T. et al. From gene networks to gene function. *Genome Res.* **13**, 2568–2576 (2003).
26. Karlebach, G. & Shamir, R. Modelling and analysis of gene regulatory networks. *Nat. Rev. Mol. Cell Biol.* **9**, 770–780 (2008).
27. Consortium, G. O. The gene ontology (go) database and informatics resource. *Nucleic acids Res.* **32**, D258–D261 (2004).
28. Choromanski, K. et al. Rethinking attention with performers. OpenReview https://openreview.net/forum?id=Ua6zuk0WRH (2021).
29. Consortium, T. T. S. et al. The tabula sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science* **376**, eabl4896 (2022).
30. Quake, S. R. & Consortium, T. S. Tabula sapiens reveals transcription factor expression, senescence effects, and sex-specific features in cell types from 28 human organs and tissues. *bioRxiv* https://doi.org/10.1101/2024.12.03.626516 2024–12 (2024).
31. Zheng, G. X. et al. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
32. Segerstolpe, Å et al. Single-cell transcriptome profiling of human pancreatic islets in health and type 2 diabetes. *Cell Metab.* **24**, 593–607 (2016).
33. Yang, X. et al. Genecompass: deciphering universal gene regulatory mechanisms with a knowledge-informed cross-species foundation model. *Cell Res.* **34**, 830–845 (2024).
34. Rosen, Y. et al. Universal cell embeddings: A foundation model for cell biology. *bioRxiv* https://doi.org/10.1101/2023.11.28.568918 (2023).
35. Zeng, Y. et al. Cellfm: a large-scale foundation model pre-trained on transcriptomics of 100 million human cells. *Nat. Commun.* **16**, 4679 (2025).
36. McDaid, A. F., Greene, D. & Hurley, N. Normalized mutual information to evaluate overlapping community finding algorithms. *arXiv preprint arXiv:1110.2515* (2011).
37. Santos, J. M. & Embrechts, M. *On the use of the adjusted rand index as a metric for evaluating supervised classification*. In *International conference on artificial neural networks*, 175–184 (Springer, 2009).
38. R^ezanková, H. et al. *Different approaches to the silhouette coefficient calculation in cluster evaluation*. In *21st international scientific conference AMSE applications of mathematics and statistics in economics*, 1–10 (2018).
39. McInnes, L., Healy, J. & Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).
40. Kramer, O. & Kramer, O. Scikit-learn. *Machine learning for evolution strategies*. 45–53 (Springer, 2016).
41. Thind, A. S. et al. Demystifying emerging bulk rna-seq applications: the application and utility of bioinformatic methodology. *Brief. Bioinforma.* **22**, bbab259 (2021).
42. Zheng, Z. et al. Enabling single-cell drug response annotations from bulk RNA-seq using scad. *Adv. Sci.* **10**, 2204113 (2023).
43. Luecken, M. D. et al. Defining and benchmarking open problems in single-cell analysis. *Nat. Biotechnol.* **43**, 1035–1040 (2025).
44. Iorio, F. et al. A landscape of pharmacogenomic interactions in cancer. *Cell* **166**, 740–754 (2016).
45. Liu, Q., Hu, Z., Jiang, R. & Zhou, M. Deepcdr: a hybrid graph convolutional network for predicting cancer drug response. *Bioinformatics* **36**, i911–i918 (2020).
46. Dixit, A. et al. Perturb-seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *cell* **167**, 1853–1866 (2016).
47. Barrangou, R. & Doudna, J. A. Applications of CRISPR technologies in research and beyond. *Nat. Biotechnol.* **34**, 933–941 (2016).
48. Adamson, B. et al. A multiplexed single-cell CRISPR screening platform enables systematic dissection of the unfolded protein response. *Cell* **167**, 1867–1882 (2016).
49. Norman, T. M. et al. Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. *Science* **365**, 786–793 (2019).
50. Roohani, Y., Huang, K. & Leskovec, J. Predicting transcriptional outcomes of novel multigene perturbations with gears. *Nat. Biotechnol.* **42**, 927–935 (2024).
51. Zhou, G. et al. Uni-mol: A universal 3d molecular representation learning framework. In *The Eleventh International Conference on Learning Representations*. (2022).
52. Kana, O. et al. Generative modeling of single-cell gene expression for dose-dependent chemical perturbations. *Patterns* **4**, 100817 (2023).
53. Bunne, C. et al. Learning single-cell perturbation responses using neural optimal transport. *Nat. methods* **20**, 1759–1768 (2023).
54. Veličković, P. et al. Graph attention networks. *arXiv preprint arXiv:1710.10903* (2017).
55. Johansson, B. B. et al. The role of the carboxyl ester lipase (cel) gene in pancreatic disease. *Pancreatology* **18**, 12–19 (2018).
56. Jermusyk, A. et al. A 584 bp deletion in ctrb2 inhibits chymotrypsin b2 activity and secretion and confers risk of pancreatic cancer. *Am. J. Hum. Genet.* **108**, 1852–1865 (2021).
57. Yu, J. & Langridge, W. H. A plant-based multicomponent vaccine protects mice from enteric diseases. *Nat. Biotechnol.* **19**, 548–552 (2001).
58. Räsänen, K., Itkonen, O., Koistinen, H. & Stenman, U.-H. Emerging roles of spink1 in cancer. *Clin. Chem.* **62**, 449–457 (2016).
59. Chastanet, A., Prudhomme, M., Claverys, J.-P. & Msadek, T. Regulation of Streptococcus pneumoniae clp genes and their role in competence development and stress survival. **183**, 295–307 (2001).
60. Day, J. B. & Plano, G. V. A complex composed of sycn and yscb functions as a specific chaperone for yopn in yersinia pestis. *Mol. Microbiol.* **30**, 777–788 (1998).
61. Hendy, G. N. et al. Targeted ablation of the chromogranin a (chga) gene: normal neuroendocrine dense-core secretory granules and increased expression of other granins. *Mol. Endocrinol.* **20**, 1935–1947 (2006).
62. Melloul, D., Marshak, S. & Cerasi, E. Regulation of insulin gene transcription. *Diabetologia* **45**, 309–326 (2002).
63. Lesche, R., Peetz, A., van der Hoeven, F. & Rüther, U. Ftl, a novel gene related to ubiquitin-conjugating enzymes, is deleted in the fused toes mouse mutation. *Mamm. genome* **8**, 879–883 (1997).
64. Liu, Y. et al. Methylation of serum sst gene is an independent prognostic marker in colorectal cancer. *Am. J. Cancer Res.* **6**, 2098 (2016).
65. Zhang, L. et al. Association of tcf7l2 and gcg gene variants with insulin secretion, insulin resistance, and obesity in new-onset diabetes. *Biomed. Environ. Sci.* **29**, 814–817 (2016).

66. Zhiling, Y. et al. Mutations in the gene encoding cadm1 are associated with autism spectrum disorder. *Biochem. biophys. Res. Commun.* **377**, 926–929 (2008).
67. Xu, J., Deng, Y. & Li, G. Keratin 19 (krt19) is a novel marker gene for epicardial cells. *Front. Genet.* **15**, 1385867 (2024).
68. Yan, Y.-L. et al. A zebrafish sox9 gene required for cartilage morphogenesis (2002).
69. Berx, G., Becker, K.-F., Höfler, H. & Van Roy, F. Mutations of the human e-cadherin (cdh1) gene. *Hum. Mutat.* **12**, 226–237 (1998).
70. Tanihara, F. et al. Generation of viable pdx1 gene-edited founder pigs as providers of nonmosaics. *Mol. Reprod. Dev.* **87**, 471–481 (2020).
71. Ma, X. et al. Hmmr associates with immune infiltrates and acts as a prognostic biomaker in lung adenocarcinoma. *Comput. Biol. Med.* **151**, 106213 (2022).
72. Chang, H. et al. The tpx2 gene is a promising diagnostic and therapeutic target for cervical cancer. *Oncol. Rep.* **27**, 1353–1359 (2012).
73. Müller-Dott, S. et al. Expanding the coverage of regulons from high-confidence prior knowledge for accurate estimation of transcription factor activities. *Nucleic acids Res.* **51**, 10934–10949 (2023).
74. Gehman, L. T. et al. The splicing regulator rbfox1 (a2bp1) controls neuronal excitation in the mammalian brain. *Nat. Genet.* **43**, 706–711 (2011).
75. Younis, R. M., Taylor, R. M., Beardsley, P. M. & McClay, J. L. The anks1b gene and its associated phenotypes: focus on cns drug response. *Pharmacogenomics* **20**, 669–684 (2019).
76. Xiao, X. et al. The gene encoding protocadherin 9 (pcdh9), a novel risk factor for major depressive disorder. *Neuropsychopharmacology* **43**, 1128–1137 (2018).
77. Chen, X. et al. A brain cell atlas integrating single-cell transcriptomes across human brain regions. *Nat. Med.* **30**, 2679–2691 (2024).
78. Peters, L. L. et al. Ank3 (epithelial ankyrin), a widely distributed new member of the ankyrin gene family and the major ankyrin in kidney, is expressed in alternatively spliced forms, including forms that lack the repeat domain. *J. cell Biol.* **130**, 313–330 (1995).
79. Sheykhi-Sabzehpoush, M. et al. Emerging roles of long non-coding RNA ftx in human disorders. *Clin. Transl. Oncol.* **25**, 2812–2831 (2023).
80. Ding, X. et al. Age-dependent regulation of axoglial interactions and behavior by oligodendrocyte AnkyrinG. *Nat. Commun.* **15**, 10865 (2024).
81. Yajnik, V. et al. Dock4, a gtpase activator, is disrupted during tumorigenesis. *Cell* **112**, 673–684 (2003).
82. Yang, Z. et al. Pag1, a cotton brassinosteroid catabolism gene, modulates fiber elongation. *N. phytologist* **203**, 437–448 (2014).
83. Besson, O., Dobigeon, N. & Tourneret, J.-Y. Minimum mean square distance estimation of a subspace. *IEEE Trans. Signal Process.* **59**, 5709–5720 (2011).
84. Bae, S.-Y. et al. Tpt1 (tumor protein, translationally-controlled 1) negatively regulates autophagy through the becn1 interactome and an mtorc1-mediated pathway. *Autophagy* **13**, 820–833 (2017).
85. Singh, N. et al. The long noncoding RNA h19 regulates tumor plasticity in neuroendocrine prostate cancer. *Nat. Commun.* **12**, 7349 (2021).
86. Masoudzadeh, S. et al. Dlk1 gene expression in different tissues of lamb. *Iran. J. Appl. Anim. Sci.* **10**, 669–677 (2020).
87. Rizzi, F., Coletta, M. & Bettuzzi, S. Clusterin (clu): from one gene and two transcripts to many proteins. *Adv. cancer Res.* **104**, 9–23 (2009).
88. Davidson, E. & Levin, M. Gene regulatory networks. *Proc. Natl. Acad. Sci.* **102**, 4935–4935 (2005).
89. Luo, R. et al. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Brief. Bioinforma.* **23**, bbac409 (2022).
90. Gao, Z. et al. Foldtoken: Learning protein language via vector quantization and beyond. *arXiv preprint arXiv:2403.09673* (2024).
91. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nat. methods* **15**, 1053–1058 (2018).
92. Jue, W. et al. Illuminating cell states by a comprehensive and interpretable single cell foundation model. *Genome Biol.* **26**, 334 (2025).

Acknowledgements

This work was supported by National Science and Technology Major Project (No. 2022ZD0115101), National Natural Science Foundation of China Project (No. 624B2115, No. U21A20427), Project (No. WU2022A009) from the Center of Synthetic Biology and Integrated Bioengineering of Westlake University, Project (No. WU2023C019) from the Westlake University Industries of the Future Research Funding.

Author contributions

J.W. conceived the idea and developed the framework. Z.G. provided the crucial technology for the vector quantization algorithm. Jue Wang drafted the manuscript, and C.T. helped in editing the abstract and introduction. C.T., Z.G., S.S., and S.L. helped in revising the manuscript. Jue Wang and Cheng Tan prepared codes and released them on GitHub. S.Z.L. supevised the project.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-026-70071-5.

Correspondence and requests for materials should be addressed to Shiping Liu or Stan. Z. Li.

Peer review information *Nature Communications* thanks Yu-An Huang and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.